Water sites, networks, and free energies with grand canonical Monte Carlo

Gregory A. Ross,[*,†] Michael S. Bodnarchuk,[‡,¶] and Jonathan W. Essex[*,†]

School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K. , and School of Mechanical Engineering, Imperial College London, Exhibition Road, London, SW1 2AZ, U.K.

E-mail: g.a.ross@soton.ac.uk; j.w.essex@soton.ac.uk

Abstract

Water molecules play integral roles in the formation of many protein-ligand complexes, and recent computational efforts have been focused on predicting the thermodynamic properties of individual waters and how they may be exploited in rational drug design. However, when water molecules form highly coupled hydrogen bonding networks, there is, as yet, no method that can rigorously calculate the free energy to bind the entire network, or assess the degree of cooperativity between waters. In this work, we report theoretical and methodological developments to the grand canonical Monte Carlo simulation technique. Central to our results is a rigorous equation that can be used to calculate efficiently the binding free energies of water networks of arbitrary size and complexity. Using a single set of simulations, our methods can locate waters, estimate their binding affinities, capture the cooperativity of the water network, and evaluate the hydration free energy of entire protein binding sites. Our techniques have been applied to multiple test systems and compare favourably to thermodynamic integra-

[*] To whom correspondence should be addressed

[†] School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K.

[‡] School of Mechanical Engineering, Imperial College London, Exhibition Road, London, SW1 2AZ, U.K.

[¶] Previous address: School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K.

tion simulations and experimental data. The implications of these methods in drug design are discussed.

Introduction

Recent years have witnessed the maturation of our understanding of water in biomolecular association,[1] such that present day structure-based drug design efforts often consider the influence of water on the ligand and protein.[2] A typical concern is whether water molecules within binding sites should be targeted for displacement to improve affinity and specificity.[3]

Computational techniques have played an essential role in this advancement, and can suggest ways to exploit the waters that mediate bound complexes,[4,5] predict the effect of water on protein-ligand binding kinetics,[6] and quantify the energetic cost of removing a water molecule from locations in a binding site.[7,8] While the targeted displacement of water can lead to more selective and strongly bound compounds,[9,10] expelling water from particular locations can worsen affinity,[11,12] and new compounds may even fail in dislodging strongly bound water.[13,14] Although the binding free energy of water at a particular location can determine the success or failure of a lead compound, it is not yet possible to measure this property experimentally.

Atomistic simulations underpin many of the most rigorous, albeit time consuming, techniques to calculate the properties of bound water molecules. Alchemical methods, such as double decoupling, can calculate the binding free energy of water at a location using multiple simulations, during which the interactions of the water molecule in question are gradually switched off.[7,8] Inhomogeneous solvation theory, which has been used by numerous groups,[15–18] can also calculate the binding free energy of individual water molecules, with the added benefit of decomposing the binding free energies into enthalpies and entropies. Simple models and scoring-based methods have also been utilised to inform which waters should be targeted for displacement[4,19,20] .

While such approaches have proved insightful, they are limited to discerning the energetics of hydration sites in the presence of a fixed distribution of surrounding water molecules. Water molecules often form hydrogen bonded networks in a binding site; removing one water will affect the free energies and structure of the remainder, and methods that predict the binding thermodynamics and locations of individual waters within a given network, such as WaterMap[16,21] and 3D-RISM,[22] will require additional simulations to account for these changes. In addition, knowing the individual binding free energy of each water site is insufficient to understand the stability of the network as a whole, and it is not possible to elucidate the degree of cooperativity between neighbouring waters without further analysis. There are, as yet, no techniques that can fully interrogate this web of highly coupled interactions.

In addition to the binding free energy of individual water molecules, the calculation of the hydration free energy of an entire protein cavity is technically very demanding to carry out with traditional alchemical methods, and is seldom attempted.[23] Instead, more approximate methods, such as Poisson-Boltzmann surface area techniques,[24] cellular automata models,[25] and grid cell theory[26] are typically employed. There is an unmet need for a technique that can rigorously and efficiently calculate the hydration free energy of protein binding sites using atomistic models, as such a method could facilitate investigations on the driving forces of biomolecular association,[27–29] help validate faster and more approximate techniques, as well as indicate forcefield inadequacies in projects where the length of simulation is not a constraint.

Underpinning any method that predicts how water will affect the binding of a protein and ligand is an accurate map of the locations of water around and within a binding site. While there are numerous computationally inexpensive methods to locate hydration sites, as with free energy estimates, atomistic simulations are the most rigorous. Molecular dynamics represents the most popular simulation technique, and the waters that appear in a given trajectory can be used to form continuous water density maps, or clustered to produce discrete locations.[30] However, without excessively long simulations, sampling may be poor in cavities that are occluded from the bulk solvent, due to prohibitively long diffusion times. Grand canonical Monte Carlo (GCMC) techniques provide alternative simulation methods which allow for the creation and annihilation of waters in a given region, completely bypassing any physical barriers to the bulk solvent.[31,32]

In contrast to molecular dynamics, Monte Carlo simulations do not generate dynamic trajectories, but instead produce probable system configurations via trialling random moves, such as translation and rotation. In grand canonical Monte Carlo (GCMC), "insertion" and "deletion" moves are added to the set of possible move types. The probability to accept or reject an insertion or deletion move is controlled by the chemical potential, and – in a manner directly analogous to energy fluctuations at a constant temperature – GCMC allows for the number of particles to fluctuate at a constant chemical potential. In GCMC, the chemical potential is a user defined parameter; previously, it has been unknown which value of the chemical potential yielded the water occupancy when the system is at equilibrium with bulk water. Consequently, the application of GCMC to solvated proteins has been confined to cases where the occupancy of a cavity was known *a priori*.[33–35] This work develops GCMC to allow the determination of optimal water occupancy from the simulations themselves, thereby expanding the utility of this technique.

In this paper, we revisit the GCMC simulation method, and derive new theoretical relations that can, among other uses, calculate the total binding free energy of a chemical species to a biomolecule, even in highly cooperative binding processes. When applied to simulations of proteins and water, we calculate the binding free energies of *entire* networks of water molecules using a set of simulations no more complicated than an alchemical decoupling calculation of a *single* water molecule. Our theoretical results are validated using Monte Carlo simulations, thermodynamic integration, and high quality experimental structure data. We expect these developments will be of benefit to rational drug design and optimization, molecular simulations, and will provide insight into the nature of water in protein cavities.

Theoretical Results

Free energies via titration

Throughout, we focus on Adams's formulation of GCMC,[31] where the parameter B – known as the Adams value – is used instead of the chemical potential:

$$B = \mu'(N)\beta + \ln(N), \tag{1}$$

where μ' is the excess chemical potential, $\beta = 1/k_B T$ with temperature T and Boltzmann's constant k_B, and N is the average number of water molecules in the GCMC volume. An extensive background on this formulation and others is provided in Section 1 of the Supplementary Information, where we show that B is completely equivalent to using the chemical potential in the acceptance test for the GCMC insertion and deletion moves. A benefit of using the Adams formulation is that many of the constants that would otherwise appear in the Monte Carlo acceptance test are absorbed into a single term. Adams originally introduced B to facilitate the calculation of the excess chemical potential from a GCMC simulation.[31]

One may consider B (or the chemical potential) as specifying the size of a molecule reservoir from which water molecules are transferred. This reservoir does not need to be simulated (as in the Gibbs ensemble[43]), as it is implicit in the GCMC method. With phases of a pure chemical species, such as liquid or gas, thermodynamic equilibrium is established when the chemical potentials of both phases are equal. However, as discussed in Section 2.1 of the Supplementary Information, when considering binding reactions, such as water binding to a protein, the condition for equilibrium is more complicated. One can either calibrate B so that the average number of inserted molecules matches a known amount,[34,35,38] or, one can run many GCMC simulations with different B values.[42] The latter method can rank order water molecules by affinity but cannot predict the absolute binding free energies; this work provides methods to do so.

We show in Section 1 of the Supplementary Information that for phase equilibrium $B = \ln(N_{\text{res}})$, where N_{res} is the number of molecules in the ideal gas reservoir to which the GCMC region is coupled. From this reservoir, molecules can be inserted into the simulated system; the larger the reservoir, the greater the chance of accepting an insertion move and rejecting a deletion move. Therefore, B can be considered as the GCMC equivalent of a logarithmic amount of titrant, and many GCMC simulations with different B values are equivalent to performing a virtual titration. Our first result is that when using GCMC to titrate molecules into a cavity that can have a maximum capacity of one molecule,

$$N(B) = \frac{1}{1 + \exp\left(\beta \Delta F_{\text{trans}} - B\right)}, \tag{2}$$

where ΔF_{trans} is the free energy to transfer a *single* molecule from the ideal gas reservoir to the system and $N(B)$ is the average number of inserted waters at a given B. The point of half maximum of $N(B)$ is equal to $\beta \Delta F_{\text{trans}}$, which can be accurately determined by fitting equation 2 to GCMC titration data. This equation is of the same form of logistic equation found in protein-ligand binding, where B represents the amount of free ligand, and is also encountered in a version of the Henderson–Hasselbalch equation for acid-base titrations.[44]

Equation 2 represents a generalization of a relation used by Clark et al., and later by Bodnarchuk et al., to calculate the binding free energies of small molecules to proteins from GCMC data.[45,46] Notably, both studies found that Clark's equation only yielded reliable free energies when B was low. This was previously thought to be due to sampling problems.[46] However, in Section 2.6 of the Supplementary Information, we derive equation 2 from kinetic arguments to show that the equation used by Clark et al. implicitly assumes that the system to which a molecule is being coupled is of a thermodynamic size, which is not the case in small protein cavities.

Grand Canonical Integration

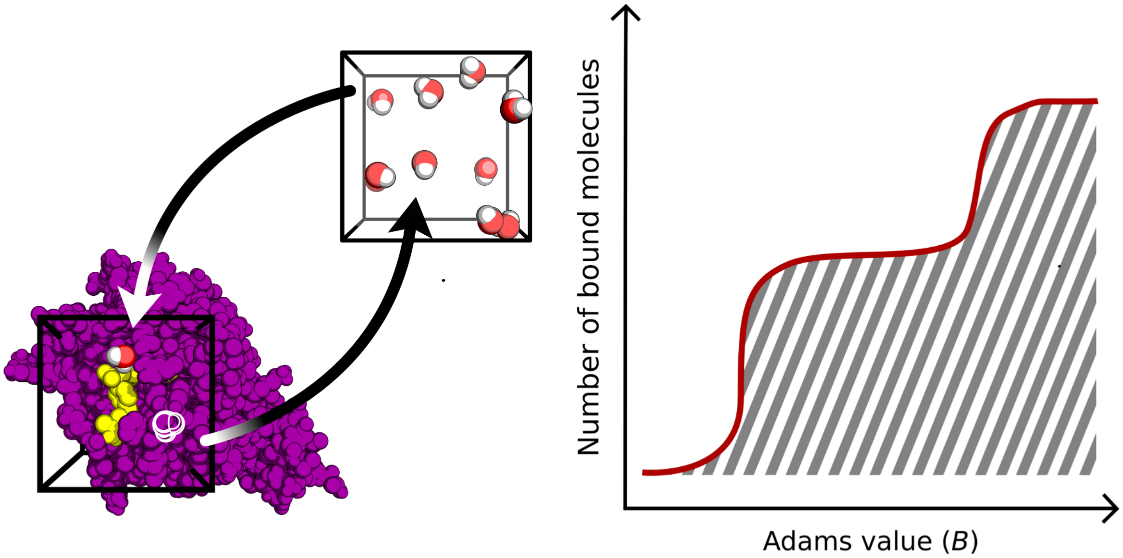
For a single-water binding site, the previous section showed that the free energy of a binding event can be determined directly from the GCMC titration curve. When running GCMC on a site that can bind multiple molecules, a closed form expression for the shape of the curve may be difficult, if not impossible to obtain without approximations. To fully realize the potential of GCMC in molecular modeling, a general method is needed that can calculate binding free energies no matter how many molecules are inserted over the course of a simulation.

In Section 2.2 of the Supplementary Information, we show that the Helmholtz free energy to transfer water molecules from an ideal gas reservoir with a volume equal to the sampled GCMC region is given by

$$\beta \Delta F_{\text{trans}}(N_i \rightarrow N_f) = N_f B_f - N_i B_i + \ln\left(\frac{N_i!}{N_f!}\right) - \int_{B_i}^{B_f} N(B)\,\mathrm{d}B \tag{3}$$

from an initial N_i to a final N_f waters in the system, and where B_k denotes the Adams value that produces an average of N_k molecules at equilibrium.

The integral over $N(B)$ is the area under the GCMC titration curve, as shown in Figure 1. Previously, Peterson and Gubbins integrated over the average number of particles in a series of GCMC simulations, but the resultant isotherms were not related to Helmholtz free energy changes.[47] We refer to the method presented here as grand canonical integration (GCI), owing to its similarity to thermodynamic integration.[48,49] For a single-water binding site, where the explicit titration curve is given by equation 2, this integral can be evaluated analytically. Crucially, we prove in Section 2.7 of the Supplementary Information that, for this case, the free energy calculated with equation 3 yields the same free energy as in equation 2. Any discrepancies between the calculated free energies of either method will, therefore, arise as a result of the different numerical techniques that are required to evaluate the equations.



(a)

(b)

Figure 1: (a) Schematic showing water molecules in an ideal gas reservoir coupled to a region within the protein. The Adams parameter, which is proportional to the chemical potential, is used in the acceptance test to insert or delete a molecule within the protein. The value of the Adams parameter determines the number of molecules that are coupled to the system. (b) Varying the applied Adams parameter and measuring the average number of inserted molecules at equilibrium can be considered as a type of titration.[42] When the maximum capacity of the protein region is one molecule, the shape of the GCMC titration curve is given by equation 2. More generally, equation 3 reveals that the area under the titration curve is related to the free energy to couple N molecules from the ideal gas phase.

Equation 3 was derived from both thermodynamic *and* statistical mechanical perspectives. As a consequence, it remains exact even when applied to calculating the coupling free energy for low numbers of molecules, where the factorial terms account for the number of ways to count identical molecules. It is therefore more general than previous GCMC free energy methods, which have been derived solely from thermodynamic considerations.[46,50] Indeed, we show in Section 2.3 of the Supplementary Information that a recent approach by Fan et al.[50] requires a correction factor at low numbers of molecules. The generality of equation 3 means that it can be of use in a range of cases: from investigating the effects of a small number of water molecules on the affinity of a compound,[6] to calculating the hydration free energy of large protein binding sites.[29]

To apply equation 3, one requires multiple GCMC simulations over a range of B values. For each

B value, the average number of inserted molecules needs to be recorded, such that the integral in equation 3 can then be calculated. To reduce the variance that can occur when computing the integral, it is prudent to smooth the GCMC titration data. While an explicit expression for the curve would be ideal, it is system dependent, and non-trivial to derive. Instead, based on the logistic equation for a single water molecule as in equation 2, for multiple sites, we propose that to a good approximation

$$N(B)_{\text{approx}} = \sum_{i=1}^{m} \frac{n_i}{1 + \exp\left(w_{0i} - w_i B\right)}, \tag{4}$$

where m, n_i, w_{0i} and w_i are free parameters to be fitted to titration data. The parameter m may be estimated from the approximate number of "steps" that appear in the titration data; n_i represents the number of waters that are coupled in step i, whilst w_{0i} and w_i are the point of inflection of the logistic curve i and its steepness, respectively. By ensuring that n_i and w_i are greater than zero, one can impose monotonicity on $N(B)_{\text{approx}}$. The sum of logistic functions has also been applied to describe pKa titration data[51] and constant pH simulations[44] for coupled proton sites.

By coincidence, equation 4 describes a simple artificial neural network known as a "single layer perceptron", a popular model in machine learning.[52] Details of the fitting of equation 4 are in Section 3.4 of the Supplementary Information.

Water binding and equilibrium

When integrating from a B_i value where the corresponding $N_i = 0$, equation 3 yields the free energy to couple N_f molecules to a cavity from the ideal gas reservoir. As the hydration free energy of a single water molecule, denoted μ'_{hyd}, is a known quantity, the *binding* free energy of a network of molecules, $\Delta F_{\text{bind}}(N)$ can be evaluated using the thermodynamic cycle shown in Figure 2(a), so that

$$\Delta F_{\text{bind}}(N) = \Delta F_{\text{trans}}(N) - \Delta F_{\text{hyd}}(N). \tag{5}$$

where $\Delta F_{\text{trans}}(N) = \Delta F_{\text{trans}}(0 \rightarrow N)$ and $\Delta F_{\text{hyd}}(N) = N\mu'_{\text{hyd}}$. With equations 3 and 5, one can rigorously determine how many water molecules are required to hydrate the GCMC volume, when that cavity can exchange molecules with bulk water at an identical temperature. As illustrated in Figure 2, the optimal number of bound water molecules, denoted N^*, is the amount that minimizes equation 5. At N^*, $\Delta F_{\text{trans}}(N^*)$ corresponds to the thermodynamic hydration free energy of the volume that was simulated with GCMC.

Previous GCMC treatments required that the number of bound water molecules had to be known *a priori*, after which a series of sequential or iterative simulations would adjust the chemical potential to reproduce this number.[34,35,38] Instead, locating the minimum of equation 5 requires no prior knowledge, and the occupancy of a cavity can be determined as accurately as the applied forcefield and sampling permit.

Finding which N minimizes the Helmholtz free energy state can be greatly simplified in the thermodynamic limit, which is when $N \rightarrow \infty$. As shown in Section 2.5 of the Supplementary Information, the lowest thermodynamic free energy state occurs when

$$\mu'_{\text{prot}}(N^*) = \mu'_{\text{hyd}}, \tag{6}$$

where $\mu'_{\text{prot}}(N^*)$ is the excess chemical potential of the GCMC region at the optimal occupancy. As shown in equation 1, calculating $\mu'_{\text{prot}}(N)$ only requires an estimate of the average number of water molecules at a given B, so that the optimal occupancy can be determined without calculating free energies with GCI. The notable feature of equation 6 is that it specifies that equilibrium is

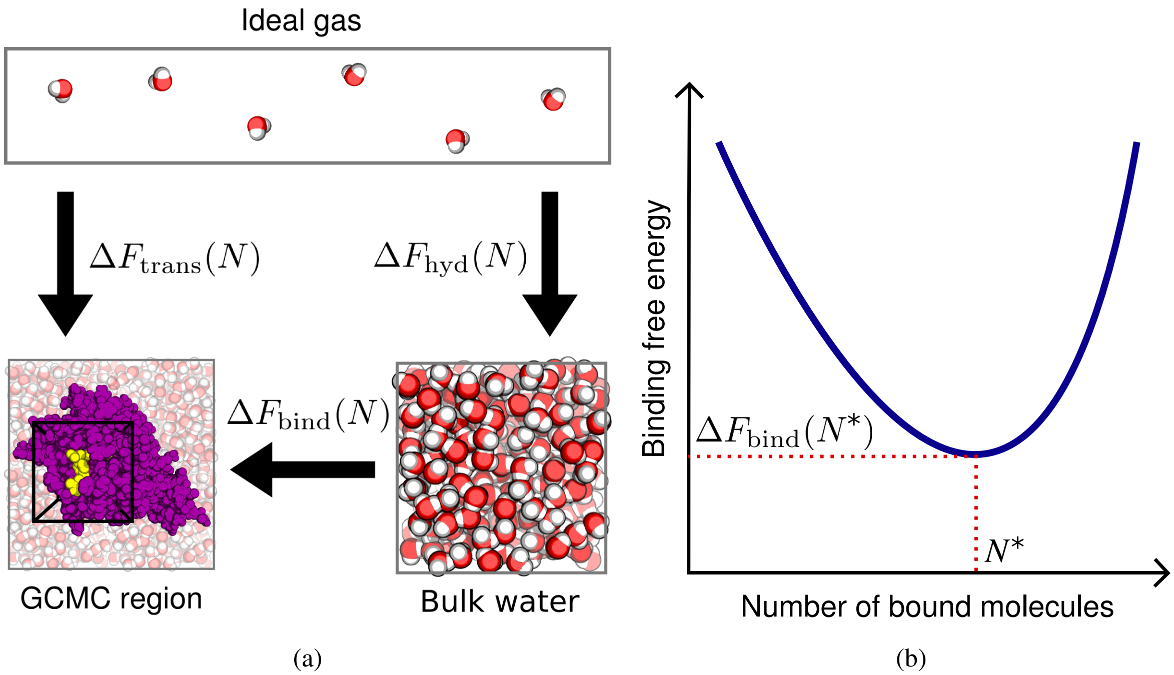

Ideal gas
$\Delta F_{\mathrm{trans}}(N)$
$\Delta F_{\mathrm{hyd}}(N)$
$\Delta F_{\mathrm{bind}}(N)$
GCMC region
Bulk water
(a)
Binding free energy
$\Delta F_{\mathrm{bind}}(N^*)$
N^*
Number of bound molecules
(b)

Figure 2: (a) Schematic representation of the thermodynamic cycle used to calculate the binding free energy of N water molecules to a GCMC region. Molecule and volume sizes are not to scale. The free energy to transfer molecules from ideal gas to the GCMC region, $\Delta F_{\mathrm{trans}}(N)$, is given by equation 3, and the hydration free energy of N water molecules, $\Delta F_{\mathrm{hyd}}(N)$, is trivial to calculate. Based on this cycle, the binding free energy of N water molecules is given by equation 5. (b) The equilibrium occupancy of molecules simulated with GCMC can be found via the application of equations 3 and 5, or more simply with equation 6 in the thermodynamic limit. As more molecules are bound to the system, the binding free energy decreases to its minimum value, denoted $\Delta F_{\mathrm{bind}}(N^*)$, as the number of bound molecules approaches N^*. Adding more molecules than N^* to the system increases the binding free energy, so that these occupancies are less likely to occur.

established when the *excess* chemical potentials of water in the protein and bulk water are equal, as opposed to the chemical potentials. Unequal chemical potentials, but equal excess chemical potentials, implies that the water density in a binding site is different to that of bulk water. Section 2.1 of the Supplementary Information also derives equation 6 in a separate manner to Section 2.5 of the Supplementary Information to show that standard state concentrations are implicitly considered, and to highlight the difficulty in setting the chemical potential without prior knowledge of the water occupancy of the binding site.

An apparent drawback to the theoretical results presented thus far is that, as with previous GCMC treatments, the B parameters must be chosen before running the simulations. If the free energy minimum is not present in the interval between B_i and B_f, additional GCMC simulations must be executed. However, the following result, found by combining equations 1 and 6, demonstrates that the B value that produces the optimal occupancy of a cavity, denoted B^* is given by

$$B^* = \beta\mu'_{\text{hyd}} + \ln(N^*). \tag{7}$$

While N^* may not be known *a priori*, this relation shows that B^* has a slowly varying dependence on N^*, so that a rough guess of N^* will be sufficient to approximate B^*. For instance, if it is suspected that a binding site contains between 10 and 30 molecules, one should choose a B parameter between -8.3 and -7.2 to obtain the optimal water occupancy, assuming $\mu'_{\text{hyd}} = -6.3$ kcal/mol.[41] A more rigorous approach would be to simulate a range of B parameters between upper and lower estimates, and use either equations 5 or 6 to calculate the location of the free energy minimum.

Previously, Deng and Roux used an equation similar to equation 7 to set B before performing a GCMC simulation.[39] However, they erroneously had N^* as the number of waters that would be found in bulk water in the same volume of the GCMC simulation, rather than the expected number of grand canonical molecules in the system of interest.

Two independent methods for determining the equilibrium occupancy of a protein cavity with GCMC have therefore been presented. The first involves explicitly calculating binding free energies with GCI (equation 3) and applying equation 5 to find the the occupancy with the lowest binding free energy. The second is represented by equation 6, which involves finding the Adams value where the excess chemical potential of the cavity equals that of bulk water. This result is conceptually simple and can, in principle, be applied in a single simulation – for instance, via a rough guess of the occupancy or an iterative procedure. However, it was derived in the thermodynamic limit so that its validity is questionable when the cavity can only bind a small number of waters. Calculating equilibrium with GCI has a wider domain of applicability than equation 6 but necessitates GCMC titration data around the optimal Adams value. While computationally more expensive to apply, additional benefits of utilizing GCI include using the values of the calculated free energies to investigate water binding thermodynamics, and using the shape of the binding free energy curve as a function of N to infer the stability of the water network. The optimal Adams values and occupancies predicted with equations 6 and 5 will be compared in the Results section.

Methods

Simulations

The software package ProtoMS, versions 2.3 and 3.0,[53] were used for the Monte Carlo simulations and data analysis. Proteins were modelled with the Amber FF99[54] forcefield, and small molecules with the GAFF[55] forcefield. To speed up the simulations, protein residues that were further than between 16 Å and 20 Å away from a binding site were removed, with the exact distance chosen so as to retain whole residues. The TIP4P water model was used,[56] and systems were solvated up to a radius of 30 Å around the binding sites. The simulations were carried out at 298 K and non-bonded interactions were cut off at 10 Å.

Prior to the Monte Carlo sampling, all crystallographically derived structures were subject to a very short minimization of 100 steps via steepest decent using AMBER 12[57] to remedy any overlapping contacts and energetically unfavourable bond lengths. Unless otherwise stated, the AMBER minimized structures were subject to a further 5 million (M) steps of equilibration with ProtoMS. Monte Carlo moves were generated in proportion to the number of solvent, protein residues and solutes, approximately according to the ratio 1:5:5 respectively. When GCMC moves were allowed, half of all Monte Carlo moves were dedicated to inserting, deleting and moving the grand canonical water molecules.

Our GCMC protocol is as described previously.[45] A cuboidal subregion was defined within each system, with the specific location and dimensions for each described in the Supplementary Information. Bulk water was not permitted to diffuse into this region, and water could only enter and leave via the random insertion and deletion protocol pioneered by Adams.[31] The Adams value is kept constant in a given GCMC simulation. Before the start of a simulation, no water was present in the GCMC region, so – with the exception of cavities that could contain a maximum of 1 water molecule – an additional 10M moves of equilibration were performed, where the only allowed moves were the insertion, deletion, translation and rotation of the grand canonical water.

Determination of bulk water density with GCMC

The theory developed in this work allows for the determination of the water occupancies of systems that are coupled to bulk water with GCMC. These occupancies can be predicted by finding the minimum binding free energy state via the evaluation of equation 3, or by equating the excess chemical potential of the system with that of bulk water, as stated by equation 6. The latter method is exact only in the thermodynamic limit.

The density of bulk water is both widely known and well reproduced by the TIP4P water model,[58] such that a given sub volume within bulk water will have a known occupancy of water molecules. As a basic initial validation of equations equations 3 and 6, we sought to predict the number of water molecules found within a sub-volume of bulk water.

A cubic box with periodic boundary conditions was filled with 1051 TIP4P water molecules and simulated within the NPT ensemble at a pressure of 1 atm. The water was equilibrated for 20M moves, followed by 85M production moves. Moves were generated according to the ratio of 997:3 between the solvent and volume respectively.

In agreement with experiment and Jorgensen's original study,[56] the average density during the production run was 999.5 kg/m^3 with a standard deviation and standard error of 8.9 kg/m^3 and 1.0 kg/m^3 respectively, which were based on 85 snapshots spaced 1M moves apart. A representative simulation snap-shot with a density of 999.8 kg/m^3 was selected for GCMC simulations, for which the volume is kept constant. Water molecules were removed from a cubic sub-volume at the centre of the bulk water box to create a vacuum. The dimensions of the sub-volume was 6.5×6.5×6.5 Å^3, which at bulk density will contain an average of 9 water molecules. Thirty-two GCMC 200M move production simulations with B values at every integer between and inclusive of -22 and +9 were performed, with insertions and deletions being carried out within the vacated volume. As described, bulk water molecules cannot diffuse from the surrounding bulk into the GCMC region, and water molecules can only be inserted into this region from the ideal gas. Equations 3 and 5 were used to calculate the free energy to fill the sub-volume with water, and the occupancy corresponding to the minimum binding free energy state was evaluated. Separately, equation 6 was also used to estimate the optimum occupancy of the sub-volume.

Free energy calculations

This paper presents new free energy calculation techniques based on GCMC that can efficiently calculate the binding free energies of many water molecules, which is technically demanding to do using other methods. To balance the ability to perform complex calculations with the need to validate using a well established free energy method, we selected three systems that could bind a maximum of one water molecule, and two systems that could bind three water molecules, for detailed study using conventional double decoupling methods.

For the single water binding systems, we selected a small buried site in bovine pancreatic trypsin inhibitor (BPTI) based on the protein data bank (PDB) structure 5PTI, and a small cavity adjacent to the binding site of scytalone dehydratase (SD), with two different ligands with structures modelled on the PDB structure 3STD. The SD ligands were named L1 and L3 to match a previous study.[59] Both BPTI and SD were chosen due to their use as model systems in previous studies.[59,60] For the three-water cavities, we selected Chk-1 bound to a ligand (private communication) and a cavity in BPTI, different to the single-water case, using the same structure as before. The coordinates of the water molecules located in this work are listed in Tables S3 and S6.

The GCMC protocol follows that of the simulated annealing of the chemical potential technique.[42] A number of B values were chosen between 0 and -12 and GCMC simulations with 40M production moves were run. By inspecting the equilibrium number of inserted water molecules, new B values were chosen for additional runs, such that there was a range of simulations where the average number of inserted waters was between 0 and the maximum number. As detailed in the Results, this provides the data from which binding free energies, as well as water locations, can be computed. The B values used in all the simulations are listed in the Supplementary Information. The average number of inserted water molecules for each B value constitutes the GCMC titration data for a given system. Free energies were calculated for GCI by fitting equation 4 to the titration data and using the resultant curve to evaluate equation 3. Errors were calculated by repeating this procedure on 1000 bootstrap samples of the titration data. Similarly for single-water cavities, free energies were calculated by fitting equation 2 to titration data, and errors were estimated by fitting to 1000 bootstrap samples.

Free energies computed with GCMC were compared to two replica exchange thermodynamic integration[61] (RETI) methods: single topology and dual topology.[62] Sixteen evenly spaced λ replicas between 0 and 1 were used. In the single topology calculations, a water was decoupled from a site in two steps: the first turned the partial charges of TIP4P to zero (decoupling the electrostatic interactions), and the second step shrunk the water molecule to nothing (decoupling the van der Waals interactions). In the dual topology simulations, a water molecule was transformed in a single transformation to a “dummy” molecule that did not interact with the system. Both single and dual topology techniques were evaluated to place the GCMC predictions in the context of the variations that occur when using different free energy methods.

There were three repeats for each RETI simulation. Each calculation was equilibrated for 5M moves at each λ value, and was succeeded by a 40M move production run. Exchanges between adjacent λ windows was attempted every 100,000 moves. When computing the coupling free energy with RETI, a symmetry correction of $-k_BT\log(2)$ was applied in the cases where no rotation about a water's axis of symmetry was observed.[63] No symmetry correction is required for GCI, as a water molecule has equal probability to be inserted and deleted in either orientation.

Decoupling molecules with thermodynamic integration, as with other free energy techniques, requires the specification of where the molecule will be decoupled, as well as the careful consideration of constraints or restraints to prevent decoupled molecules from sampling outside the region of interest. The locations, restraints and constraints, and the order in which the water molecules were decoupled were based on the hydration sites observed in the GCMC simulations. A schematic that summarises how this was done is presented in Figure S2. In the first set of RETI simulations, harmonic restraints were applied to water molecules, and the calculated free energies were corrected to remove the effect of the restraint.[7] If bulk water was found to have occupied the decoupled site during a simulation, the simulation was repeated with a hard-wall constraint applied to the decoupled water molecule as previously described.[8,59] The final set of restraints and constraints for each water molecule are shown in Tables S7 and S8.

From the occupancy of the water locations during the GCMC titrations, the water molecules in Chk-1 and BPTI were rank ordered by affinity. This ordering, shown in Tables S7 and S8, was then used when sequentially decoupling the waters from the cavities using both RETI techniques. With BPTI, two water sites had equivalent occupancies, so two routes to decoupling the three waters were performed. The free energies reported in the results for this system are averaged over these pathways.

To reduce the statistical noise of the free energy calculations, and thus facilitate the comparison between the different methods, the GCMC and RETI simulations of BPTI, SD and Chk-1 were run with a fixed protein backbone. Side chains were sampled over the angles and dihedrals.

Locating waters in protein and protein-ligand systems

Water locations can be determined from the GCMC simulations whose average number of water molecules corresponds to the binding free energy minimum. To test this method as a rigorous water placement tool, proteins whose binding site solvation structure was well understood were chosen for analysis. For all systems in this test, sixteen evenly spaced B values between -1 and -16 were selected for the running of the GCMC simulations. Crystallographic water molecules were removed from the GCMC region prior to starting the simulations.

First, MUP-I protein was used as a negative control, as it is expected that its mainly hydrophobic cavity has a low occupancy of water.[64] To investigate the effect of fixing the protein backbone on the water binding free energies, two sets of simulations were run, with four repeats each, of free and fixed protein backbone. The PDB structure 1QY1 was used to set-up the system before simulation. An initial 10M moves were dedicated to sampling GCMC, followed by 20M moves where all of the system was also sampled as equilibration. The production runs had 20M moves.

Second, five protein and protein-ligand structures from the validation set of WaterDock, a fast placement method,[4] had water locations and occupancies predicted using our GCMC method. The systems are shown in Table 1. Two systems from the WaterDock study were omitted because of the presence of metal ions in the binding sites. Each system has been resolved at least twice using X-ray crystallography and/or neutron diffraction, allowing the comparison of the water predictions from the GCMC simulations to experimentally well validated water sites. The GCMC box coordinates were the same as in the WaterDock study, with each box being $15 \times 15 \times 15$ Å^3. To prepare the experimental data, the crystal structures were structurally aligned in Pymol,[65] and the crystallographic water oxygen positions were hierarchically clustered with average linkage with a 1 Å cut-off using SciPy.[66] Clusters that encompassed more than one crystallographic water molecule from different structures were classified as “consensus” sites, and these were deemed to be the most well defined positions. All clusters were retained to assess the false positive rate of the predicted water positions. Placement accuracy for each system was compared to WaterDock.[4]

Table 1: The systems used to validate the water sites predicted by GCMC using the free energies calculated by GCI. Water molecules that overlapped in at least two structures were deemed to be well resolved and labelled as “consensus” sites. The first PDB code listed was the one used for the set-up of the simulations, while all listed PDB codes were used to determine consensus sites.

System	PDB codes	Ligand bound
HIV1-Protease	2ZYE, 3FX5, 1HPX	KNI-272
Ribonuclease A	1KF5, 1FS3, 5RSA	None
GluR2	1FTM, 1MY2	AMPA
Trypsin	1S0Q, 1UTQ, 1TPO	None
Glutathione S-T	1K3Y, 1K3L	S-hexyl glutatione

To facilitate the comparison to crystallographically determined water sites, the systems in Table 1 were simulated with a fixed protein backbone. Only one repeat per structure was carried out. For each system, an initial equilibration of 10M moves was carried out only on the grand canonical water molecules, followed by 60M move production simulation where the protein side chains, ligands and bulk water were also allowed to move.

Results

Bulk water density

Multiple GCMC simulations at a range of Adams values were performed on an empty sub-volume in bulk water to test the consistency and validity of equations 1, 3, 5 and 6. For a known occupancy of a GCMC region, one can calculate the excess chemical potential using equation 1, and separately, the free energy to transfer a molecule from ideal gas with GCI (equation 3). Alternatively, if the excess chemical potential of the molecular reservoir is known (in this case bulk water), one can predict the occupancy of the GCMC volume either by applying equation 6 or locating the minimum of the binding free energy, given by equation 5. Each of these approaches should predict equivalent occupancy and excess chemical potentials as one approaches the large occupancy limit.

The size of the cubic GCMC volume was chosen so that there would be 9 water molecules at bulk water density. The GCMC titration data, shown in Figure S1, was fitted with equation 4 with $m = 3$. This fit was used to estimate the B that produced an average of exactly 9 water molecules in the sub-volume. Inputting this B value in equation 1 predicts the excess chemical potential to be -6.4 ± 0.2 kcal/mol. The determination of excess chemical potential in this manner is similar in spirit to Adams's original work.[31] To compare, GCI calculated $\Delta F_{\text{trans}}(9 \rightarrow 10)$ to be -6.2 ± 0.2 kcal/mol, as shown in Figure 3. Both predictions are consistent, and agree with the experimentally measured hydration free energy of a single water molecule at -6.3 kcal/mol.[41] As GCI explicitly calculates the free energy to transfer a single water molecule from ideal gas to bulk water, the hydration free energy value of -6.2 kcal/mol will henceforth be used for internal consistency.

For the sub-volume where GCMC moves were performed, the "binding" free energy refers to the free energy to hydrate the otherwise empty cubic region, and contains a significant cavitation contribution. The negative of the binding free energy equals the free energy required to empty the cubic cavity of water. Binding free energies were calculated with equation 5 and are shown in Figure S1 (b). The minimum free binding energy occurred at 9 water molecules to the nearest integer, which is agreement with the expected number at bulk density. As implied by the previous paragraph, the minimum binding free energy also coincides with the equal excess chemical potentials of the GCMC region and bulk water. Therefore, while being strictly valid only in the thermodynamic limit, equation 6 is accurate despite being applied to so few molecules. This is explained in Section 2.3 of the Supplementary Information.

As described in the Methods, a hard-wall constraint is used to separate solvent water from GCMC water. At the minimum binding free energy state, the GCMC region and bulk water are in thermodynamic equilibrium. At this point, the effect of the hard-wall vanishes because water is indistinguishable either side of the barrier. While it is clear from the above discussion that the average density within the GCMC region matches that of bulk water, visual inspection revealed that water forms a continuous water density profile across the hard-wall.

In general, if the water density is continuous over the boundary of the GCMC region, such as in

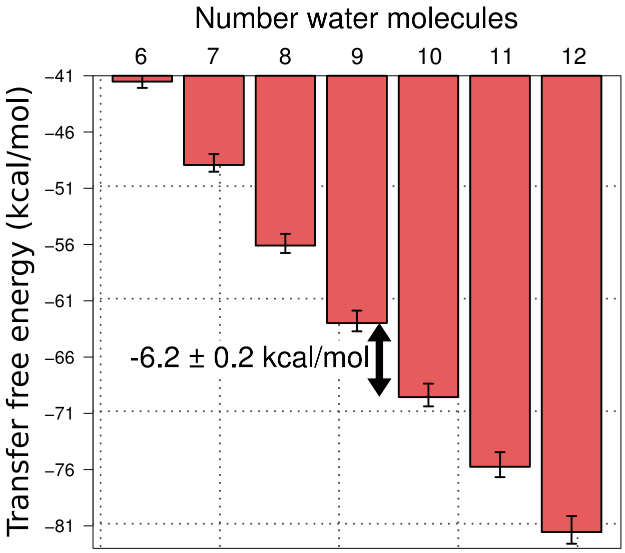


Figure 3: Barplot showing cumulative free energies to transfer water molecules from ideal gas to the GCMC sub-volume within bulk water as calculated with GCI (equation 3). Bulk water density occurs when 9 water molecules are present in the GCMC sub-volume. The calculated free energy to hydrate a single water molecule with bulk density is indicated, and the value agrees with excess chemical potential calculated with equation 1 and an experimental value for the hydration free energy of water.[41]

these bulk water simulations, the ideal gas transfer free energies – calculated with equations 3 – will be dependent on the volume of the box. For instance, increasing the volume of the GCMC region in bulk water will increase the calculated free energy because more water molecules will be able to fit inside. If water density is discontinuous over the GCMC boundary, such as within buried protein cavities, changing the size or shape of the sub-volume will not affect the calculated free energies so long as the hydration sites remain the same.

Free energy calculations

Single-water cavities

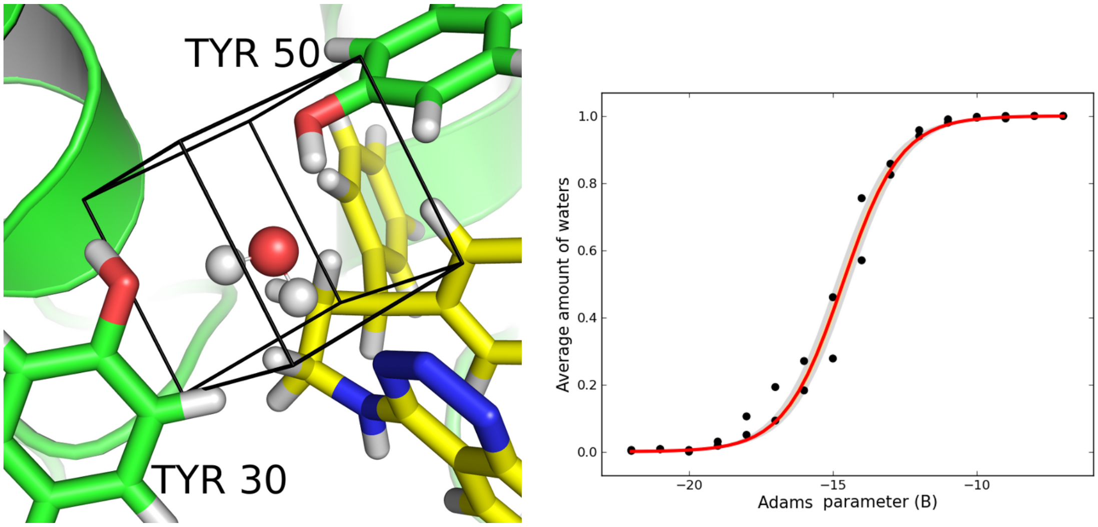
Table 2 shows how the binding free energies calculated for the single-water cavities compare between RETI, the fitting of the GCMC logistic function (equation 2), and GCI (equation 3). All the techniques are in satisfactory agreement, although the GCMC calculations predict more positive binding free energies than both thermodynamic integration methods. The effect is more pronounced in SD L1 (shown in Figure 4), where the GCMC methods are approximately 1.2 kcal/mol more positive than the mean of the RETI calculations. We hypothesized this was due to the protein not fully relaxing around the grand canonical water molecule during the simulation as half of all moves were dedicated to inserting, deleting and translating the grand canonical water molecule. To test this, we repeated the GCMC simulations for each water molecule starting with the final frame of one of the dual topology RETI simulations in which the water was fully coupled to the system. These frames were assumed to be better equilibrated around the water molecule. After re-calculating the free energies using both GCMC methods, as expected, SD L1 system showed a significant decrease in the predicted binding free energy, with GCI and the logistic fit both predicting -9.5 ± 0.1 kcal/mol, which are in closer agreement with the single topology calculations.

Table 2: Free energies to transfer a single water molecule from an ideal gas to small cavities. *As discussed in the main text, these calculated free energies both changed to -9.5 ± 0.1 kcal/mol when using a different starting structure. The errors quoted for the GCI and logistic fit methods are the standard deviations calculated from bootstrap sampling, whereas for both TI methods the standard deviation arising from three repeats is shown.

	Calculated free energies (kcal/mol)			
Water	TI single topology	TI dual topology	GCI	Logistic fit
BPTI	-12.4 ± 0.0	-12.6 ± 0.0	-12.2 ± 0.1	-12.2 ± 0.1
SD L1	-9.6 ± 0.1	-10.3 ± 0.3	-8.8* ± 0.1	-8.7* ± 0.1
SD L3	-4.3 ± 0.1	-4.1 ± 0.0	-3.3 ± 0.1	-3.5 ± 0.1

Table S4 details the restraints/constraints used for the RETI simulations. Only a harmonic restraint was required for the BPTI single-water cavity, whereas the SD systems required the application of a hard-wall constraint, as it was found that the cavity could be accessed by bulk waters. The hard-wall constraint has the effect of restricting the protein movement, even when the water molecule is fully decoupled. We expect this slight difference in the end states of GCMC simulations and the RETI simulations for the SD systems to be responsible for the additional discrepancies between the calculated free energies.

Given the variability between the predictions of different methods, it is clear that the error bars in Table 2 underestimate the "true" error of the calculations, with some calculated errors being less than 0.05 kcal/mol. The errors bars for both RETI methods are the standard deviations over three repeats, where each repeat had the same starting configuration but a different random seed. Thus, these errors reflect the statistical consistency of the methods, rather than how reproducible the calculation would be if a different method and an uncorrelated starting structure were to be used. On the other hand, the error bars calculated for GCMC are from bootstrap re-sampling of the GCMC titration data; instead of statistical consistency, these errors estimate the effect that the uncertainty in the data has on the calculations. With regard to these considerations, the free energies calculated using GCMC and TI are consistent within sampling limits.



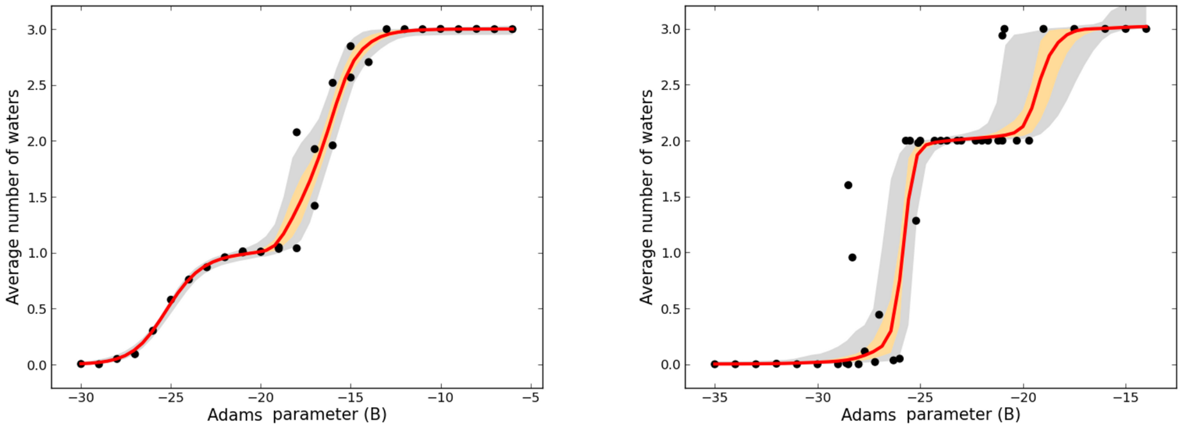
(a)

(b)

Figure 4: (a) The water molecule bound to scytalone dehydratase (SD) (green cartoon). The ligand L1 is shown as yellow sticks and TYR 30 and TYR 50 are highlighted in green sticks. The black lines mark out the volume where insertion and deletion moves of water were attempted. (b) GCMC titration plot for SD in complex with L1, showing how the average occupancy of the cavity varies with the applied B parameter. The line of best fit (in red) is obtained by least squares fitting of equation 2 to the GCMC data (black dots), which yields the free energy to insert a single water molecules from the ideal gas phase to the protein cavity. The error of the insertion free energy was estimated with bootstrap sampling, and the grey shaded area indicates where 90% of the bootstrap fits lay.

Three-water cavities

Figure 5 shows how the occupancy of the Chk-1 and BPTI cavities increases with B in the GCMC simulations. In both systems, the occupancy increases by two within a single step as B is increased, and visual inspection of the structures revealed that in each case two water molecules were binding to adjacent sites. In particular, the coupled two water sites in BPTI have equivalent occupancies during the simulations where $N \leq 2$, indicating that water molecules bind to these sites as a *dimer*. In rational drug design, it is typical to consider the energetic cost of displacing a single water molecule by a chemical group, without regarding the cooperativity between sites. These results imply that the removal of a single water can destabilise other sites, unless the ligand is modified in such a way to recover the lost interactions. GCMC simulation data, such as in Figure 5, can be used to predict where these instances will occur.



(a)

(b)

Figure 5: Chk-1 (a) and BPTI (b) GCMC titration for the three-water cavities. The red lines are the artificial neural network models (equation 4) that were fitted with $m = 3$. The GCI free energies shown in Table 3 and Table 4 were obtained by integrating over these lines from $N = 0$ to $N = 1$, 2 and 3. The light grey and light orange areas mark, respectively, where 90% and 50% of the bootstrapped fits lay.

The binding energies of the water molecules in the Chk-1 and BPTI cavities are shown in Table 3 and Table 4. In the RETI calculations, each water molecule was coupled sequentially and constrained/restrained to occupy a predefined location. In contrast, GCMC simulations can sample all water sites simultaneously within a predefined volume. With GCMC, therefore, an integer number of total water molecules can refer to multiple partially occupied sites. In the case of BPTI and Chk-1, the binding free energies for 3 water molecules correspond to the same state in the RETI and GCMC simulations as the cavities are fully occupied. In contrast, when 1 and 2 water molecules are present, the free energies calculated by RETI and GCI refer to different, albeit similar, states. Nevertheless, there is broad quantitative agreement between the RETI and GCI predictions. For Chk-1, the average difference between the GCI predictions and the mean of the dual and single topology RETI free energies is 0.9 kcal/mol, compared to 0.2 kcal/mol between the RETI predictions. The average difference between the GCI and RETI predictions is larger for BPTI, at 1.6 kcal/mol compared to 0.3 kcal/mol between the RETI predictions. This larger difference is partly due to the noisier variation of N with B in BPTI, shown in Figure 5. GCMC sampling could be improved using cavity-biased insertion moves.[67,68] However, the large outlier at $B = -28.5$ indicates that poor deletion rates of the grand canonical water have also hindered sampling. Encouragingly, the fitted neural network model, shown as the red line in Figure 5, has mitigated these factors by smoothing over the data in a physically meaningful way. Also, the bootstrap error estimates, shown in Table 3 and Table 4, correctly capture the larger uncertainty in the calculated free energies of BPTI that result from the poorer quality data.

Table 3: Free energy to transfer a given number of waters from the ideal gas into the Chk-1 cavity. The standard deviations after three repeats are shown for the TI calculations, and the bootstrap standard deviation is shown for GCI. As the second and third water molecules are coupled together in the simulated annealing of the chemical potential (see Figure 5), only the transfer free energies to insert three waters are directly comparable.

	Calculated free energies (kcal/mol)		
# Water(s)	TI single topology	TI dual topology	GCI
1	-15.5 ± 0.1	-15.5 ± 0.0	-14.9 ± 0.1
2	-24.6 ± 0.5	-24.9 ± 0.0	-25.9 ± 0.4
3	-34.7 ± 0.5	-34.5 ± 0.7	-35.4 ± 0.5

Table 4: Free energy to transfer a given number of waters from the ideal gas into the BPTI three-water cavity. The standard deviations after three repeats are shown for the TI calculations, and the bootstrap standard deviation is shown for GCI.

	Calculated free energies (kcal/mol)		
# Water(s)	TI single topology	TI dual topology	GCI
1	-14.2 ± 0.1	-14.3 ± 0.5	-15.4 ± 0.5
2	-33.0 ± 0.3	-33.5 ± 0.5	-31.0 ± 0.6
3	-41.8 ± 0.3	-42.2 ± 0.5	-43.5 ± 0.8

Discrepancies between the free energies calculated with RETI and GCMC can also be attributed to the applied restraints or constraints. The way these were chosen and the manner in which they were applied is included in the Supplementary Information. Although a harmonic restraint was preferred – and used in the Chk-1 calculations – a spherical hard-wall constraint had to be applied on a few of the water molecules during the RETI BPTI calculations. Constraints were used to stop adjacent water molecules moving into the decoupled water location and to plug the entrance to the binding site to prevent the bulk water molecules accessing the cavity. While the effect of the constraint on the sampling of the water molecule was corrected analytically, it is not possible to fully account for the effect of the hard-wall on the sampling of the protein, as protein side chains cannot occupy the volume defined by the constraint. The only constraint used in the GCMC simulations was to prevent water molecules diffusing in or out of the predefined grand canonical volume, such that there are no equivalent effects on the sampling of the protein. Protein sampling with GCMC and the harmonic restraint should be identical.

It is important to highlight that many of the RETI simulations had to be re-run because of set-up errors, or due to the fact that the restraints were insufficient to prevent the filling of the cavities with bulk water. Alchemical methods also require preliminary simulations to locate the binding sites of the water molecules. The GCMC methodology presented here is thus much easier to implement than conventional free energy methods, requiring only the specification of a volume of interest, and the B values to simulate. Additionally, the order of annihilation does not need to be specified with GCMC, and the cooperativity within the network is explicitly captured, as demonstrated by

Figure 5.

There were a total of 48 production simulations to decouple three waters in Chk-1, as there were 16 λ windows per water molecule. Not including the simulations that had to be re-run, there was a total of 64 RETI simulation windows in BPTI, as extra simulations were required to change the ordering of the first two waters. The transformations for these calculations are listed in tables S7 and S8. This is compared to 32 and 41 B values used in the GCMC simulations for Chk-1 and BPTI, respectively. Using 2.6 GHz Intel Sandy Bridge processors, the time to decouple all three waters in Chk-1 was approximately 363 CPU hours per repeat, compared to a total of 212 CPU hours required to run all the Chk-1 GCMC simulations. With BPTI, the RETI simulations required 595 CPU hours per repeat when computing both decoupling routes, and approximately 357 CPU hours per repeat if only one decoupling route was computed. In contrast, the total time to run all the GCMC simulations with BPTI was 209 CPU hours. Therefore, not only were the GCMC simulations easier to set-up, they were less computationally expensive. The cost effectiveness of GCMC relative to conventional free energy methods will increase as the number of water molecules increases: to calculate the hydration free energy of a binding site, TI and other alchemical methods require at least as many sets of decoupling simulations as there are water molecules (exponentially more if all decoupling routes are explored), whereas GCMC requires only one set of simulations to decouple an arbitrary number of waters.

Water placement

The previous sections established that the free energies calculated with GCI compare favourably to those calculated with RETI, and that GCI naturally predicts the bulk density of water, as well as the hydration free energy of water. In this section, we use free energy profiles computed with GCI and equation 6 to predict the water occupancies and locations in protein binding sites.

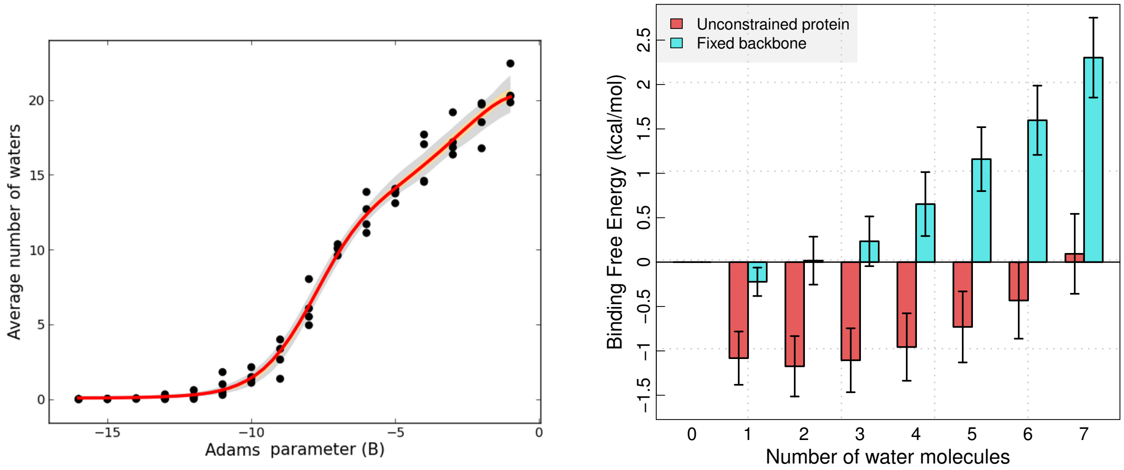
MUP-I

To test the ability of GCI to predict the water occupancies in protein cavities, we first applied it to MUP-I, which has a hydrophobic binding site. In addition, we sought to understand the effect fixing the protein backbone has on the binding free energy of water. The average number of inserted water molecules for each B was used to produce the titration plots, such as Figure 6 (a). The free energy to insert up to twenty water molecules was calculated using equation 3, which required evaluating the area under the titration curves.

Figure 6 (b) shows the binding free energies for the first seven waters calculated with equation 5, using our predicted value of -6.2 kcal/mol as the hydration free energy of water. Looking solely at the minimum binding free energies, the unconstrained simulations and the fixed-backbone simulations predict similar occupancies of 2 and 1 water respectively. However, the shape of the profiles are very different, with the binding free energy for the fixed-backbone simulations increasing faster than that of the unconstrained backbone simulations, indicating that the constrained protein is less able to adapt as more waters are added to the cavity. As the binding free energies are small in magnitude, with -1.2 ± 0.3 kcal/mol as the most negative value, the ability of MUP-I to bind water is sensitive to the flexibility of the protein.

As well as having small binding free energies, the binding free energy profile for the unconstrained protein, shown in Figure 6 (b), is sufficiently shallow such that any number of bound waters between 0 and 6 waters are thermally accessible. While the minimum free energy state for MUP-I in the unconstrained simulations is to have 2 waters bound in any particular configuration, the shape of the binding free energy profile ensures that at any one time, the number of water molecules in the cavity can fluctuate significantly. A previous study, using two 9.5 ns molecular dynamics simulations and the TIP3P water model, estimated there to be between 3 and 4 waters bound in MUP-I cavity.[64]

The optimum occupancy of a cavity can also be calculated without evaluating binding free energies by applying equation 6. Doing so on the unconstrained MUP-I simulations predicted an average



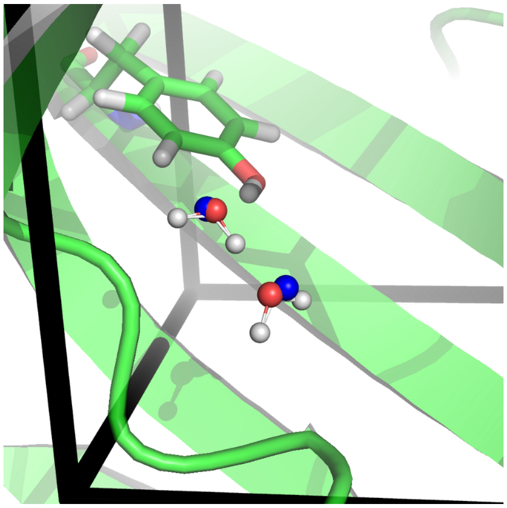
(a)

(b)

Figure 6: (a) The GCMC titration plot showing how the average number of bound water molecules varies with B for MUP-I. The line of best fit is shown in red, and was obtained by fitting equation 4, with $m = 2$. The error was determined using bootstrap sampling, and the grey area indicates where 90% of the lines of best fit resided. The data were obtained from unconstrained MUP-I GCMC simulations at 16 B values, each with 4 repeats. (b) The binding free energy of a given number of water molecules to MUP-I for simulations where the protein backbone was kept fixed, compared to when it was unconstrained. Error bars represent the standard deviations from 1000 bootstrap samples. The free energy profile for the unconstrained simulations indicate that the cavity contains between 0 and 6 water molecules. By definition, the binding free energy for zero waters is zero, and the free energies have been calculated using this reference state.

of 1 ± 0 bound water to the nearest integer, which is in agreement with 2 ± 1 bound waters to the nearest integer predicted by minimizing the binding free energy. A discrepancy between the two estimates is to be expected because of statistical noise and the fact the equation 6 is strictly valid in the thermodynamic limit.

To study the water locations in MUP-I, we used the unconstrained simulations at $B = 10$ as these simulations produce an average occupancy close to the optimum value of 2. Water positions from these simulations were clustered using the average linkage hierarchical clustering method with a 2 Å distance cut-off in SciPy.[66] Figure 7 shows the top two most occupied clusters. These water sites correspond to the two conserved water molecules in the X-ray crystal structures of the apo protein (PDB code 1I04), and holo structures with sec-butyl-thiazoline (PDB code 1I06), hydroxy-methyl-heptanone (PDB code 1I05) and glycerol (PDB code 1QY0). A more extensive validation of this rigorous water placement method is discussed in the next section.



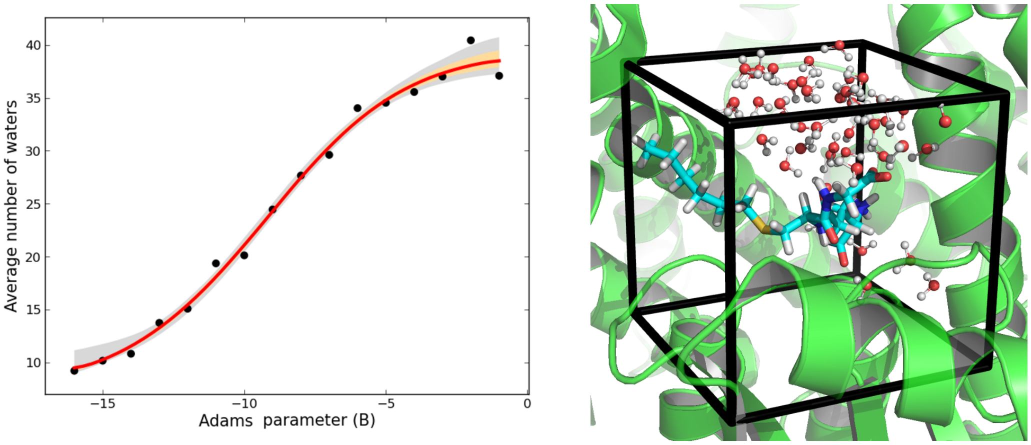
(a)

Figure 7: The GCMC region (delimited by black lines) of MUP-I (green cartoon) with the top two most-occupied water clusters from the unconstrained GCMC simulations at $B = -10$ (red spheres). In the binding site, TYR 138 is shown in green sticks and the blue spheres are the water molecules that bridge the interaction between the protein and hydroxy-methyl-heptanone, taken from PDB 1I05.

WaterDock data set

Five protein and protein-ligand systems taken from the validation of WaterDock[4] were used to further assess the ability of our GCMC methodology to correctly place and predict water occupancies. For each system, the relationship between B and the average number of water molecules was modeled using equation 4. Exemplified by Figure 8 (a), all titration data were well reproduced by fitting equation 4 with $m = 1$.

With the fitted models, the equilibrium number of bound water molecules was calculated independently by minimizing equation 5 and using equation 6. Both methods predicted equal optimal occupancies to the nearest integer for all five systems. These optimal number of water molecules, denoted N^*, are shown in Table 5, and correspond to the number of water molecules that would be present within the GCMC volume if it were in contact with an infinitely sized water bath at an identical temperature. While this equilibrium number could be determined with molecular dynamics, the insertion and deletion moves of GCMC means that water positions in occluded cavities can be sampled more efficiently.



(a)

(b)

Figure 8: (a) The variation of number of bound water molecules with Adams value, with line of best fit for glutathione S-transferase. The light orange area shows where 50% of the bootstrap samples produced the line of best fit, and the light grey where 90% of the fits lay. (b) The predicted water locations for glutathione S-transferase, which were obtained using hierarchical average linkage clustering with a 2 Å distance cut-off. Most of the predicted sites are in a bulk accessible region, and are not fully resolved in the X-ray crystal structure.

The water positions in the simulations whose average occupancies most closely matched N^* were clustered in the same way as with MUP-I. Table 5 shows how well the predicted water sites correspond to crystallographic data. Owing to the vagaries of crystallographically determined water molecules,[69,70] water sites predicted by GCMC were not compared to a single structure, but instead, multiple structures were used to determine consensus sites (see Table 1).

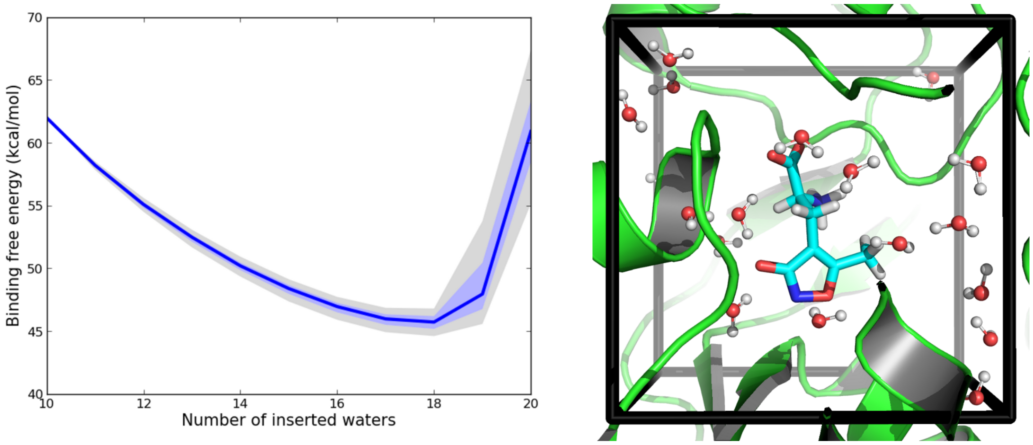
Clustered water positions from GCMC that were within 2 Å of a crystallographic water molecule were classified as correct predictions. This cut-off is larger than the 1 Å cut-off used when clustering the crystallographic waters to account for the thermal motion that occurred during the simulations. False positives resulted from a site being predicted more than once, and from predictions that were over 2 Å away from an experimentally determined water molecule. In Table 5, there are discrepancies between the total number of experimentally determined sites, the number of water molecules predicted, and false positives, owing to the fact that experimentally determined water molecules immediately outside the GCMC volume did not count towards the total, but could be within 2 Å of a predicted water.

Table 5 shows that all consensus water molecules were within 2 Å of a predicted water site. For comparison, the rapid placement tool WaterDock predicted 93% of consensus sites.[4] In these cases, the rigor and additional computational expense of the GCMC method presented here was necessary to capture all these most well resolved sites. Visual inspection of the clustered water sites and crystal structures revealed that the false positives were predominantly found in areas where the GCMC regions were accessible to bulk water. As an example, the predicted water sites for Glutathione S-transferase are shown in Figure 8 (b). The bulk-water accessible regions of the crystal structures would have the most poorly resolved water molecules due to their inherent freedom. A faster method that does not fully treat water-water interactions, such as WaterDock, is unable to capture bulk waters. In short, the ability of equations 6 and 3 to determine the correct occupancy of a cavity is supported by these data.

Table 5: The number of distinct and consensus water sites resolved by crystallography that lie within the volume of interest. The crystallographic summary shows the number of consensus crystallographic water molecules and the total number of distinct crystallographic sites within the GCMC volume. Consensus waters have been resolved at least twice in two different experimentally determined structures. In the GCMC summary, the optimum number of water molecules, N^*, is shown, as well as the corresponding Adams parameter, B^*. Predicted water sites that were within 2 Å of a crystallographic site that was outside the GCMC volume did not contribute to the false positives count.

	Crystallographic summary		GCMC summary			
System	# consensus	# total	N^*	B^*	# consensus predicted	# false positives
HIV1 protease	10	15	13	-7.9	10	2
Trypsin	14	20	37	-6.9	14	8
GluR2	14	14	18	-7.6	14	0
Ribonuclease A	4	10	11	-8.1	4	3
Glutathione S-T	14	20	30	-7.1	14	13

With GCI (equation 3), the relative free energies of different water occupancies could also be calculated. Figure 9 shows the water binding free energy profile for GluR2 (a), and the predicted water locations (b), which are in complete agreement with X-ray crystallography. Figure 9 (a) is similar to the schematic in Figure 2, and the minimum free energy state can be clearly seen at 18 bound water molecules. The binding free energies are only calculated up to an additive constant, as 10 water molecules were still bound at the lowest simulated *B*. This additive constant is different for each cavity, and to calculate the absolute hydration free energy of the site, as in Chk-1 or BPTI, GCMC simulations where the average number of water molecules is zero are required.



(a) (b)

Figure 9: (a) Free energy to bind water molecules into the volume, computed only up to an unknown additive constant. Errors were calculated using bootstrap sampling of the GCMC titration data; the light blue area shows the top 50% of the bootstrapped free energies, and the light grey areas shows the top 90%. (b) The GluR2-AMPA complex with the GCMC volume is delimited by the black lines. Clustered water molecules at the optimal *B* parameter are also shown.

Discussion

We have expanded the theoretical underpinnings of the grand canonical Monte Carlo method (GCMC), such that in the case of water and biomolecules, it can be used to determine relative binding free energies of water molecules, the hydration free energy of protein cavities, and be used as a rigorous water placement tool. Our central theoretical result, which we refer to as grand canonical integration (GCI), has the primary benefit of efficiently calculating the binding free energies of *multiple* molecules – calculations which, using other methods such as replica exchange thermodynamic integration (RETI), require technically demanding sequential simulations, including additional simulations to determine the appropriate order for water annihilation.

We applied GCI to determine the binding free energy of water molecules in three cavities that could bind one water molecule, and two cavities that could bind three. The free energies calculated with GCI were in good agreement with the free energies calculated with RETI using both single and dual topology transformations. Notwithstanding, the RETI calculations were significantly more difficult to carry out, requiring the careful application of constraints/restraints, and pre-determined water positions. Both these considerations require preliminary simulations, which in this study were the GCMC simulations themselves. The ability to calculate free energies directly from the GCMC simulations will, therefore, greatly simplify the rigorous calculation of water binding affinities. The ease with which the hydration free energy of protein cavities can be calculated with GCI may prove fruitful in cases where one wishes to de-construct the hydration contribution to the binding free energy of a ligands, particularly in rational drug design.

In addition to developing GCI, we solved an outstanding problem with GCMC simulations: how to determine at which chemical potential (or Adams value) one should run a GCMC simulation without any *a priori* knowledge. Previous attempts determined the chemical potential when the occupancy of a cavity was already known.[34–38] In our treatment, the most likely occupancy of a cavity that is in contact with bulk water emerges as that which minimizes the total binding free energy, which can be calculated explicitly using GCI (equation 3), or estimated by finding where the excess chemical potential of the cavity matches that of bulk water (equation 6). A benefit of calculating the binding free energy as a function of occupancy is that is trivial to determine the free energy changes of adding or removing a given number of waters.

We expect these developments will encourage a more widespread use of GCMC. To this end, we investigated the ability of GCI to determine the water locations and occupancies in pharmaceutically relevant proteins, and protein-ligand systems whose structures had been experimentally resolved multiple times. There was an excellent correspondence between the most well validated experimental water sites and sites predicted by GCMC using GCI. In these cases, the additional computational expense required by GCMC was required to improve on the accuracy of the much faster WaterDock.[4] A particularly stringent test of water placement with GCI was its application to MUP-I, as the binding cavity was expected to be have a low occupancy despite its size.[71] In agreement with a previous study,[72] we predicted the occupancy of the MUP-I cavity to be between zero and six molecules. Interestingly, the binding free energy was a shallow function of N, indicating that the number of water molecules in the cavity is a highly fluctuating quantity. Further work is needed to ascertain whether this feature is common among other hydrophobic binding sites.

While there exist a number of other water placement tools, GCI and long time-scale molecular dynamics are the only truly rigorous water placement methods, as they will yield the lowest free energy water distributions for a given forcefield and sufficient sampling. However, unlike molecular dynamics, GCMC sampling is unhindered by occluded cavities, and, when simulations are analyzed with GCI, one obtains information of the stability of a whole network of water molecules via the binding free energy curve. In contrast to methods that predict the binding free energies of individual waters, such as WaterMap[16,21] and 3D-RISM,[22] one also obtains information on the correlations and cooperative effects between waters via water titration plots. For instance, in BPTI it was revealed that a pair of water molecules bind as a dimer. We note that methods based on inhomogeneous fluid solvation theory, including WaterMap, may require GCMC to sample waters in buried cavities. Fast, more approximate water placement methods such, as WaterDock[4] and SZMAP[73] are often validated against rigorous free energy calculations. Thus, GCMC as presented here may prove useful in validating future techniques.

Scripts to automate the set-up, running, and analysis of GCMC and GCI are freely available to download with the Monte Carlo package ProtoMS.[53]

Supporting Information Available

Background and a summary of prior research on GCMC; detailed exposition on all the theoretical results, including derivations of all the equations presented in this work, an exploration of how the thermodynamic limit impacts the computed free energies, and a demonstration that equation 2 generalizes the method proposed by Clark et al.;[46] details on how the RETI simulations were set-up, and the sizes and locations of the applied restraints; the *B* values simulated with GCMC for each complex and GCMC box dimensions; a comparison of free energies calculated with equation 2 and the method proposed by Clark et al.; details on the fitting procedure and loss functions used for the artificial neural network shown in equation 4. All data and simulation materials are available on request. This material is available free of charge via the Internet at `http://pubs.acs.org`.

Acknowledgement

We thank the Vernalis Chk-1 project team for sharing specific date about this system, Mihaly Mezei for his helpful discussions, and Michelle Lamb, Richard Ward, Marcel Verdonk, and Paul Mortenson for their insights throughout the project. We gratefully acknowledge the EPSRC via the CCPBioSim project (grant codes EP/J010189/1 and EP/J010588/1) and AstraZeneca for funding.

References

(1) Ball, P. *Nature* **2011**, *478*, 467.

(2) Huggins, D. J.; Sherman, W.; Tidor, B. *J. Med. Chem.* **2012**, *55*, 1424.

(3) Bissantz, C.; Kuhn, B.; Stahl, M. *J. Med. Chem.* **2010**, *53*, 5061.

(4) Ross, G. A.; Morris, G. M.; Biggin, P. C. *PLoS ONE* **2012**, *7*, e32036.

(5) Kumar, A.; Zhang, K. Y. J. *J. Chem. Inf. Model.* **2013**, *53*, 1880.

(6) Bortolato, A.; Tehan, B. G.; Bodnarchuk, M. S.; Essex, J. W.; Mason, J. S. *J. Chem. Inf. Model.* **2013**, 1700.

(7) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 7683.

(8) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. *J. Am. Chem. Soc.* **2007**, *129*, 2577.

(9) Lam, P. Y.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Et, A. *Science* **1994**, *263*, 380.

(10) Liu, C.; Wrobleski, S. T.; Lin, J.; Ahmed, G.; Metzger, A.; Wityak, J; Gillooly, K. M.; Shuster, D. J.; McIntyre, K. W.; Pitt, S.; Shen, D. R.; Zhang, R. F.; Zhang, H.; Doweyko, A. M.; Diller, D.; Henderson, I.; Barrish, J. C.; Dodd, J. H.; Schieven, G. L.; Leftheris, K. *J. Med. Chem.* **2005**, *48*, 6261.

(11) Mikol, V.; Papageorgiou, C.; Borer, X. *J. Med. Chem.* **1995**, *38*, 3361.

(12) Nasief, N. N.; Tan, H.; Kong, J.; Hangauer, D. *J. Med. Chem.* **2012**, *55*, 8283.

(13) Kadirvelraj, R.; Foley, B. L.; Dyekjær, J. D.; Woods, R. J. *J. Am. Chem. Soc.* **2008**, *130*, 16933.

(14) Vollmuth, F.; Geyer, M. *Angew. Chem. Int. Ed.* **2010**, *49*, 6768.

(15) Li, Z.; Lazaridis, T. *J. Am. Chem. Soc.* **2003**, *125*, 6636.

(16) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. *Proc. Natl. Acad. Sci.* **2007**, *104*, 808.

(17) Nguyen, C. N.; Cruz, A.; Gilson, M. K.; Kurtzman, T. *J. Chem. Theory Comput.* **2014**, *10*, 2769.

(18) Huggins, D. J.; Marsh, M.; Payne, M. C. *J. Chem. Theory Comput.* **2011**, *7*, 3514.

(19) García-Sosa, A. T.; Mancera, R. L.; Dean, P. M. *J. Mol. Model.* **2003**, *9*, 172.

(20) Amadasi, A.; Surface, J. A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. *J. Med. Chem.* **2008**, *51*, 1063.

(21) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2008**, *130*, 2817.

(22) Sindhikara, D. J.; Hirata, F. *J. Phys. Chem. B* **2013**, *117*, 6718.

(23) Helms, V.; Wade, R. C. *Proteins* **1998**, *32*, 381.

(24) Stoica, I.; Sadiq, S. K.; Coveney, P. V. *J. Am. Chem. Soc.* **2008**, *130*, 2639.

(25) Setny, P.; Zacharias, M. *J. Phys. Chem. B* **2010**, *114*, 8667.

(26) Michel, J.; Henchman, R. H.; Gerogiokas, G.; Southey, M. W. Y.; Mazanetz, M. P.; Law, R. J. *J. Chem. Theory Comput.* **2014**, *10*, 4055.

(27) Homans, S. W. *Drug Discov. Today* **2007**, *12*, 534.

(28) Baron, R.; Setny, P.; Andrew McCammon, J. *J. Am. Chem. Soc.* **2010**, *132*, 12091.

(29) García-Sosa, A. T. *J. Chem. Inf. Model.* **2013**, *53*, 1388.

(30) Henchman, R. H.; McCammon, J. A. *J. Comp. Chem.* **2002**, *23*, 861.

(31) Adams, D. J. *Mol. Phys.* **1974**, *28*, 1241.

(32) Marrone, T. J.; Resat, H.; Hodge, C. N.; Chang, C. H.; McCammon, J. A. *Protein Sci.* **1998**, *7*, 573.

(33) Resat, H.; Mezei, M. *Biophys. J.* **1996**, *71*, 1179.

(34) Speidel, J. A.; Banfelder, J. R.; Mezei, M. *J. Chem. Theory Comput.* **2006**, *2*, 1429.

(35) Malasics, A.; Gillespie, D.; Boda, D. *J. Chem. Phys.* **2008**, *128*, 124102.

(36) Rutledge, G. C. *Phys. Rev. E* **2001**, *63*, 2001.

(37) Wilding, N. B. *J. Chem. Phys.* **2003**, *119*, 12163.

(38) Lakkaraju, S. K.; Raman, E. P.; Yu, W.; MacKerell, A. D. *J. Chem. Theory Comput.* **2014**, *10*, 2281.

(39) Deng, Y.; Roux, B. *J. Chem. Phys.* **2008**, *128*, 115103.

(40) McQuarrie, D. A. *Statistical Mechanics*, 1st ed.; University Science Books, 1976.

(41) Naim, A. B.; Marcus, Y. *J. Chem. Phys.* **1984**, *81*, 2016.

(42) Guarnieri, F.; Mezei, M. *J. Am. Chem. Soc.* **1996**, *118*, 8493.

(43) Panagiotopoulos, A. Z. *Molecular Physics* **1987**, *61*, 813.

(44) Goh, G. B.; Hulbert, B. S.; Zhou, H.; Brooks, C. L. *Proteins* **2014**, *82*, 1319.

(45) Bodnarchuk, M. S.; Viner, R.; Michel, J.; Essex, J. W. *J. Chem. Inf. Model.* **2014**, *54*, 1623.

(46) Clark, M.; Meshkat, S.; Wiseman, J. S. *J. Chem. Inf. Mod.* **2009**, *49*, 934.

(47) Peterson, B. K.; Gubbins, K. E. *Mol. Phys.* **1987**, *62*, 215.

(48) Shirts, M.; Mobley, D.; Chodera, J. *Annu. Rep. Comput. Chem.* **2007**, *3*, 41.

(49) Michel, J.; Foloppe, N.; Essex, J. W. *Mol. Inf.* **2010**, *29*, 570.

(50) Fan, C.; Do, D. D.; Nicholson, D.; Ustinov, E. *Mol. Phys.* **2013**, *112*, 60.

(51) Onufriev, A.; Case, D. A.; Ullmann, G. M. *Biochemistry* **2001**, *40*, 3413.

(52) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*, Springer: USA, 2009.

(53) *ProtoMS*, Bodnarchuk, M.; Bradshaw, R.; Cave-Ayland,; Genheden, S.; Martinez, A. C.; Michel, J.; Ross, G. A.; Woods, C. J. www.protoms.org.

(54) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.

(55) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comp. Chem.* **2004**, *25*, 1157.

(56) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(57) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comp. Chem.* **2005**, *26*, 1668.

(58) Jorgensen, W. L.; Jenson, C. *J. Comput. Chem.* **1998**, *19*, 1179.

(59) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2009**, *131*, 15403.

(60) Olano, L. R.; Rick, S. W. *J. Am. Chem. Soc.* **2004**, *126*, 7991.

(61) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13703.

(62) Michel, J.; Verdonk, M. L.; Essex, J. W. *J. Chem. Theory Comput.* **2007**, *3*, 1645.

(63) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Phys.* **2006**, *125*, 084902.

(64) Barratt, E.; Bingham, R. J.; Warner, D. J.; Laughton, C. A.; Phillips, S. E. V.; Homans, S. W. *J. Am. Chem. Soc.* **2005**, *127*, 11827.

(65) *The PyMOL Molecular Graphics System*, version 1.6 Schrödinger, New York, 2013.

(66) Oliphant, T. E. *Comput. Sci. Eng.* **2007**, *9*, 10.

(67) Mezei, M. *Mol. Phys.* **1980**, *40*, 901.

(68) Woo, H.-J.; Dinner, A. R.; Roux, B. *J. Chem. Phys.* **2004**, *121*, 6392.

(69) Carugo, O.; Bordo, D. *Acta Crystallogr. Sect. D* **1999**, *55*, 479.

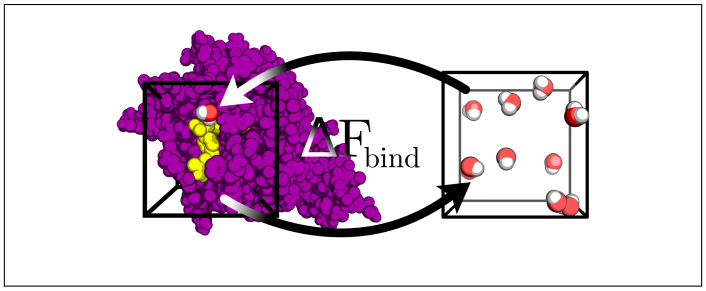
(70) Nucci, N. V.; Pometun, M. S.; Wand, A. J. *Nat. Struct. Mol. Biol.* **2011**, *18*, 245.

(71) Bingham, R. J.; Findlay, J. B.; Hsieh, S.-Y. Y.; Kalverda, A. P.; Kjellberg, A.; Perazzolo, C.; Phillips, S. E.; Seshadri, K.; Trinh, C. H.; Turnbull, W. B.; Bodenhausen, G.; Homans, S. W. *J. Am. Chem. Soc.* **2004**, *126*, 1675.

(72) Malham, R.; Johnstone, S.; Bingham, R. J.; Barratt, E.; Phillips, S. E.; Laughton, C. A.; Homans, S. W. *J. Am. Chem. Soc.* **2005**, *127*, 17061.

(73) Bayden, A. S; Moustakas, D. T.; Joseph-McCarthy, D.; Lamb, M. L. *J. Chem. Inf. Model.* **2015**, *55*, 1552.

Graphical TOC Entry


ΔF_{bind}