Enhancing water sampling in free energy calculations with grand canonical Monte Carlo

Gregory A. Ross,* Ellery Russell, Yuqing Deng, Chao Lu, Edward D. Harder, Robert Abel, and Lingle Wang*

Schrödinger, Inc., 120 West 45th Street, New York, New York 10036, United States

E-mail: gregory.ross@schrodinger.com; lingle.wang@schrodinger.com

Abstract

The prediction of protein-ligand binding affinities using free energy perturbation (FEP) is becoming increasingly routine in structure-based drug discovery. Most FEP packages use molecular dynamics (MD) to sample the configurations of proteins and ligands, as MD is well-suited to capturing coupled motion. However, MD can be prohibitively inefficient at sampling water molecules that are buried within binding sites, which has severely limited the domain of applicability of FEP and its prospective usage in drug discovery. In this paper, we present an advancement of FEP that augments MD with grand canonical Monte Carlo (GCMC), an enhanced sampling method, to overcome the problem of sampling water. We accomplished this without degrading computational performance. On both old and newly assembled data sets of protein-ligand complexes, we show that the use of GCMC in FEP is essential for accurate and robust predictions for ligand perturbations that disrupt buried water.

Introduction

In structure-based lead optimization, a popular strategy to enhance ligand selectivity and affinity is to design compounds with chemical groups that occupy sites which would otherwise be filled with water. Free energy perturbation (FEP) methods, particularly the FEP+ implementation,[1] are becoming increasingly routine in prospective drug discovery projects. FEP calculations typically use molecular dynamics (MD) to efficiently capture the coupled and concerted motion of atomistic configurations.[2,3] However, water molecules that are buried deep within the interiors of proteins can have residence times as long as hundreds of microseconds[4] – well beyond the typical sampling time of FEP calculations. As a result, FEP transformations that modify or displace buried water have remained inaccurate as well as sensitive to the initial placement of water inside the protein.

If a water molecule is expected to be displaced after a ligand modification, one rigorous treatment, known as "double decoupling", calculates the free energy to decouple a water molecule from the hydration site and from bulk water.[5,6] These free energies are then combined with the free energy change from the ligand transformation, calculated in the absence of the water molecule, to make the final prediction. However, the double decoupling method is computationally expensive, difficult to set up, and involves the complex application of restraints or constraints. Other techniques, such as those based on inhomogenious fluid solvation theory, can also predict the binding thermodynamics of water at the cost of additional simulations and calculations.[7–10] Due to these difficulties, FEP practitioners instead often experiment (often through trial and error) with including or excluding the "displaced" water molecule in the starting structure, or use a fast algorithm to predict the occupancy of the hydration site prior to running FEP.[11]

Recently, grand canonical Monte Carlo (GCMC) has shown great promise as a simulation framework to automatically incorporate the thermodynamics of buried water in ligand FEP calculations.[12–14] Unlike pure MD methods, the number of atoms and molecules in a GCMC simulation is variable. Molecules are added or removed using specialized moves that are accepted or rejected using a Metropolis-Hastings criterion, which is constructed to maintain a valid thermodynamic ensemble. GCMC is particularly well suited to sampling water in buried cavities as the specialized moves completely bypass any physical barriers that slow the binding and unbinding of water. Importantly, GCMC requires no prior knowledge of the number and location of water molecules in the binding site. These benefits can also be achieved in simulation methods that have a fixed number of molecules by allowing waters to make discontinuous jumps across barriers.[15,16] However, these types of moves are likely to have lower acceptance rates than GCMC because each move requires a simultaneous insertion and deletion, whereas a GCMC move requires only a single insertion or deletion.

Questions of computational efficiency and reliability abound for any new technique in free energy calculations, particularly so in prospective drug discovery projects where calculations are conducted at scale and can be subject to strict time constraints. Simulation methods that involve the addition or removal of molecules have notoriously low acceptance rates, which has prompted the development of numerous acceleration schemes, such as biasing insertions into free space[17,18] and the use of nonequilibrium trial moves.[19,20] One avenue that has remained as yet unexplored for GCMC-type methods has been the massive parallelism afforded by modern graphical processing units (GPUs). As GPUs are now routinely used to conduct MD simulations, new sampling techniques could benefit by exploiting this pre-existing and powerful infrastructure.

In this work we report a GPU-accelerated implementation of GCMC water sampling in the Schrödinger FEP+ workflow.[21] We begin by discussing the theory behind GCMC and its validity for use in ligand free energy calculations. Our GCMC protocol is thoroughly validated, such as by comparing GCMC free energy predictions to double decoupling calculations and hydration free energies, as well as demonstrating that the location of important hydration sites can be sampled efficiently. We assemble a large data set of proteins and ligands that focuses on transformations that displace or disrupt buried water molecules. We show that GCMC lowers the prediction error irrespective of whether or not the disrupted water molecules are included in the starting structure. The propensity for GCMC to increase the convergence of the free energy predictions is also investigated and discussed. We expect that expanding the domain of applicability of FEP+ to reliably include modifications that involve buried water will be of significant benefit to structure-based drug discovery.

Theory

The interpretation of the grand canonical ensemble

Most biomolecular simulations are performed in the isobaric-isothermal (NPT) ensemble, in which the number of particles (N), pressure (P), and temperature (T) are fixed. In the grand canonical ensemble (μVT), the number of particles and the pressure fluctuates, but the chemical potential of the fluctuating chemical species (μ), the volume (V), and the temperature remain constant. The term *grand* comes from the French for large and was first introduced to signify that the grand canonical ensemble is a superset of smaller, *petite*, ensembles that have a fixed number of particles.[22] Grand canonical Monte Carlo (GCMC) is a simulation technique that is used to sample from the grand canonical ensemble.

One practical difficulty encountered when trying to implement GCMC is deciding on what chemical potential to select. The chemical potential is the free energy to add or remove the molecules from some predefined reservoir. In our GCMC simulations we use the chemical potential of bulk water. This ensures an inserted water molecule is, in effect, being transferred *from* bulk water and any deleted water is being transferred *to* bulk water. By construction, this choice of chemical potential ensures that a GCMC simulation of pure water has the same density as a MD NPT simulation of pure water. We discuss how we calibrated the chemical potential for bulk water in section 4 of the Supporting Information (SI).

The μVT and NPT ensembles both deal with fluctuations in the particle *density*. NPT simulations sample the density by varying the volume for a fixed number of particles, whereas

μVT simulations sample the density by varying the number of particles for a fixed volume. A useful byproduct of having the equal densities in NPT and μVT is that an equivalence between the density fluctuations can be established. Proved in section 1 of the SI, the variance of the density, denoted $\mathrm{Var}(\rho)$, in the NPT and μVT ensembles are related via

$$\frac{\langle V \rangle_{NPT}}{\langle \rho \rangle^2_{\mathrm{NPT}}} \mathrm{Var}(\rho)_{NPT} = \frac{V_{\mu\mathrm{VT}}}{\langle \rho \rangle^2_{\mu\mathrm{VT}}} \mathrm{Var}(\rho)_{\mu\mathrm{VT}}, \tag{1}$$

where angular brackets with subscripts represent ensemble averages. This – to our knowledge – new relation shows that the density fluctuations in NPT and μVT ensembles are equal when the mean densities are the same and the fixed volume in the μVT ensemble is equal to the mean of the volume in NPT. Thus, GCMC, particularly when applied to water in solvated systems, has a similar effect to a barostat as the density of a system can equilibrate and fluctuate with the same magnitude as in NPT.

Of particular interest is the relationship between free energies in NPT and μVT. We show in section 2 of the SI that when perturbations are made to the potential energy of a system, like in FEP, the same free energy estimators (such as thermodynamic integration, Bennetts Acceptance Ratio method etc.) can be derived for μVT. Pertinently, we also demonstrate how these estimators will yield predictions that are approximately equal to those in NPT. Critical to the establishment of the latter is the equivalence of the density in μVT and NPT, and the approximate equivalence of the variance of the density between the two ensembles.

The thermodynamics of water displacement

This section uses a toy model to quantify the thermodynamic contribution a buried water molecule has on the binding affinity of a ligand. The aim is to help interpret the results that are presented later in this paper and to illustrate the benefits that GCMC provides in FEP when ligand transformations displace bound water molecules.

Consider two hypothetical ligands, a and b, that can bind to a receptor. Ligand a is

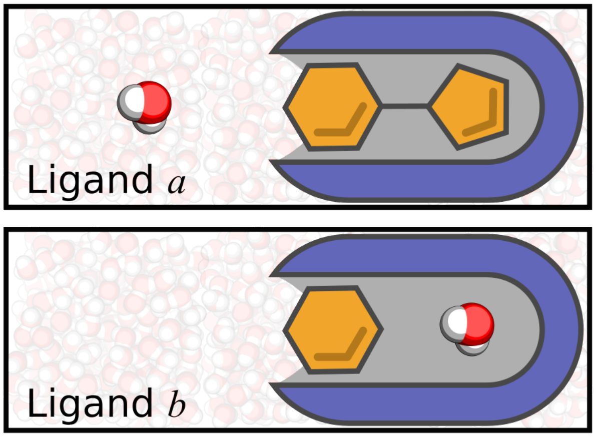

Figure 1: A schematic diagram of a receptor that can bind 2 ligands. Ligand a (top panel) fills the entire binding site of the receptor, whereas ligand b (bottom panel) is smaller and supports a buried water molecule when it is bound. Relative to ligand b, ligand a can be said to "displace" a buried water molecule. Irrespective of the nature of the hydration site, this water molecule contributes favorably to the affinity of ligand b such that ligand a will incur a penalty to its affinity by displacing it. The goal of applying GCMC water sampling to ligand FEP calculations is to be able to capture the thermodynamic cost of transferring water molecules to and from bulk water for different ligands.

larger than ligand b such that ligand b can accommodate a single buried water molecule. This water molecule resides in the bulk solvent when ligand a is bound. Figure 1 shows schematic representations of these complexes. We are interested in the relative binding free energy between these two ligands, and, in particular, the contribution of the buried water molecule.

Derived in section 3 of the SI, the relative binding free energy between ligands a and b, denoted $\Delta\Delta G_{a\to b}$ can be expanded into the relative free energy when hydration site is unoccupied with both ligands, i.e. *dry*, plus a term that corrects for solvation of the cavity in ligand b, denoted ΔG_{solv}:

$$\Delta\Delta G_{a\to b} = \Delta\Delta G^{\mathrm{dry}}_{a\to b} + \Delta\Delta G_{\mathrm{solv}}, \tag{2}$$

where

$$\Delta\Delta G_{\text{solv}} = -k_B T \ln\left(1 + e^{-\beta \Delta G_{\text{water}}^{\text{bind}}}\right). \tag{3}$$

Here, $\Delta G_{\text{water}}^{\text{bind}}$ denotes the standard free energy to transfer a water molecule from the bulk solvent to the hydration site. The derivation in section 3 of the SI fully accounts for the indistinguishably of water. Because the water molecule is bound with ligand b and not with ligand a, $\Delta\Delta G_{\text{solv}}$ is always less than or equal to zero and, crucially, this property is irrespective of the details of the receptor and the ligands. Thus, the hydration of a site near a ligand contributes favorably to the binding affinity. For a new ligand to displace a water molecule and have a lower (i.e. more favorable) binding free energy, the new ligand must be able to compensate for loss of hydration, for instance, by increasing the strength of the interaction with the receptor.

Also in section 3 of the SI, an alternative form of the solvent correction is shown to be

$$\Delta\Delta G_{\text{solv}} = k_B T \ln(1 - p), \tag{4}$$

where $p \in [0, 1]$ is the fractional occupancy of the hydration site with ligand b. In a relative binding free energy calculation, this relation shows that it may not be sufficient to place a water molecule in the starting receptor structure. Instead, to correctly capture the thermodynamic contribution of a hydration site, it is important that the hydration site is occupied by a water molecule with the correct *frequency* during the simulations. The primary goal of using GCMC in a ligand FEP calculation is for the occupancy of all hydration sites to be accurately sampled in order to fully capture their thermodynamic contribution to relative binding free energy calculations.

Simulation details

The relative binding free energy between two ligands is computed by alchemically transforming one to another when they are bound to the complex and when they are solvated in bulk water. All solvent FEP calculations are conducted in NPT ensemble. For the calculations in the complex, this work explores different water sampling methods.

General MD and FEP details

The OPLS3e forcefield[23] was used throughout along with the SPC water model. The Nosé-Hoover chains integrator[24] was used in the MD stages to maintain a constant temperature of 300 K in all simulations. Hydrogen mass repartitioning was used along with the RESPA multiple time-stepping algorithm[25] with timesteps of 4 fs for bonded interactions, 4 fs for nonbonded interactions within the distance cutoff, and 8 fs for electrostatic interactions in reciprocal space. No salt or neutralizing counterions were added to the systems. When simulating from the NPT ensemble, the Martyna-Tobias-Klein barostat was used to maintain pressure at 1 atm.[26]

As previusly described,[27] FEP+ uses replica exchange solute-tempering (REST). Replicas are exchanged every 1.2 ps. The data collected during REST is used to calculate free energies. Unless otherwise stated, 16 lambda windows (replicas) were used for core-hopping transformations and all others used 12 lambda windows - no charge-changing transformations were used in this study. The default simulation length used in this study is 20 ns (i.e. 20ns per lambda window). The Bennett Acceptance Ratio[28] method is used to calculate the relative free energy between neighboring lambda windows, and the estimated free energies are summed to give the relative free energy of the whole transformation. The cycle-closure algorithm[29] is used to combine the relative binding free energies over a FEP+ map into consistent predictions.

GCMC specific details

Our simulation protocol iterates between a GCMC stage, where attempts to insert and delete water are made, and an NVT MD stage. The combination of these two stages ensures the simulations sample from a type of grand canonical ensemble; although the simulations have a fixed number of solute atoms, this ensemble will be referred to as μVT for brevity. In the production μVT FEP protocol, each MD stage lasts for 5ps. In the GCMC stage, a "local"[30] routine is performed with a probability of 90%. In this mode, 34,000 insertion or deletion attempts are made within an orthorhombic box with faces that are a minimum of 4 Å away from the ligand. With a probability of 10%, a "globalâĂİ routine is performed in which 102,000 insertions and deletion attempts are made over the entire simulation volume. The combination of these two modes ensures that water molecules near the alchemical perturbation region are well sampled and that density fluctuations of the whole system do not occur on detrimentally slow timescales. After the final attempt is made in either the local or global modes, velocities for all atoms in the system are drawn from the Maxwell-Boltzmann distribution before restarting the MD stage.

Using the scheme of Woo et al.,[18] insertions are biased into unoccupied space by using a 3D grid that tracks cavities within the system. The length of each cell is 0.22 Å, and a cell is deemed to be "occupied" if the entire volume of the cell is within 1.5 Å of an atom. An insertion move is attempted by first, selecting an "unoccupied" cell, and second, selecting a position within the cell for the water oxygen atom with a uniform probability. The orientation of the proposed water molecule is randomly drawn uniformly over the unit sphere. As described by Woo et al., the use of the occupancy grid requires a particular Metropolis-Hastings criteria to ensure the algorithm maintains detailed balance.

GCMC equilibration

Before starting REST, but after the minimization and relaxation stages, GCMC is used to equilibrate water around the ligand and equilibrate the system density for each lambda window. To accelerate the water placement and density equilibration, the GCMC equilibration stage occurs after every 25 fs of MD and global attempts are made with a probability of 75% . For the first 20 ps of this equilibration scheme, heavy atom position restraints are applied with a force constant of 50 kcal/mol/Å^2. This is followed by another 20 ps without any position restraints.

The implementation and performance of the GCMC GPU code

Our implementation of GCMC has been highly optimized on GPUs and achieves a simulation performance that is comparable with pure MD with a barostat. The GCMC energy evaluations use the same force field and treatment of long range interactions as MD.

GCMC insertion and deletion attempts are typically made in sequence within a for-loop. In our implementation, this sequence of attempts is "unrolled" onto the GPU and attempts are evaluated simultaneously within fixed-size batches. If a move is accepted within a batch, the remainder of the attempts are discarded, the system configuration is updated, and a new batch is evaluated. No more than one move is accepted per batch. Batched attempts are made on the GPU until the specified total amount of attempts has been completed. This approach is ideally suited to GCMC because of its low acceptance rate, which ensures that the number of discarded attempts is low. In our implementation, the probability to accept an insertion or deletion attempt is roughly 0.5 %. Thus, batching attempts on a GPU is much faster than the sequential for-loop approach as the number of batches can be made to be much smaller than the number of attempts.

To evaluate the performance of the GCMC GPU code, a protein-ligand complex was taken from the bace1 data set from a previous publication.[1] The system was solvated and minimized using the default FEP+ protocol for NPT and simulated for 5 ns. Using the final structure, the system was simulated for 5 ns in NPT as well as with the GCMC-MD simulation protocol described above on the same GPU and CPU. It was found that the GCMC-MD simulation was 11 % faster than the simulation in NPT, with the relative speed-up due to the fact that in our workflow, the barostat in NPT is updated more frequently than GCMC is performed.

Comparing FEP protocols

FEP predictions made with three different water sampling protocols are compared throughout this manuscript. These protocols are summarized in Table 1. The protocol where GCMC water sampling is used during both the equilibration and production REST stages is referred to as μVT. When no GCMC water sampling is performed either during the equilibration or the REST stages and a barostat is applied, the protocol is referred to NPT.

In this work we also explore using GCMC to equilibrate water around the ligand prior to launching REST in the NPT ensemble. This protocol is referred as “NPT pre-solvate”. As this ensemble requires that each lambda window has the same number of atoms in the system, water molecules that are furthest from the protein and ligand are removed until this condition is met. This removal step is followed by a brief minimization on each lambda window..

Table 1: The complex leg sampling protocols that are evaluated and compared in this study.

Protocol name	Description
μVT	GCMC is performed during equilibration and the REST stage.
NPT	There is no GCMC water sampling.
NPT pre-solvate	GCMC sampling is only performed during equilibration and the REST stage is in NPT.

The prediction of conserved hydration sites

Critical to the utility of the combination of GCMC water sampling and MD is the ability to predict the location of conserved water molecules, particularly those that are relevant for ligand optimization. It is of interest how quickly and efficiently the GCMC-MD protocol can populate hydration sites.

Methods

Well characterized hydration sites were selected from four protein-ligand systems: HIV1 protease (PDB entry 3FX5), PTP1B (PDB entry 2QBS), HSP90 (PDB entry 3RLP), and Brd4(1) (PDB entry 3JVK). The hydration sites that will be used to assess the speed and accuracy of the GCMC protocol have been either displaced or considered for displacement via ligand modifications in previous work[31–34] and are shown in Figure 2.

A single hydration site was used from HIV1 protease. This water molecule forms hydrogen bonds with the ligand as well as ILE 50 from protein chain A and ILE 150 from protein chain B. The two hydration sites from PTP1B are buried in a subpocket between the ligand and the protein residues near ALA 217 and ARG 221. The four hydration sites from HSP90 are found between the ligand and the protein, near ASN 51, SER 52, and ASP 93. The four conserved hydration sites in Brd4(1) lie between acetylated lysine and protein residues GLY 85, MET 105, MET 132.

All crystallographic water molecules within 4 Å of the ligands were removed from the protein-ligand structures, which included the conserved waters of interest. As the purpose is to investigate the placement of water in known structures, 50 kcal/mol/Å^2 harmonic position restraints were placed on the proteins and ligands. The systems were solvated within an orthorhombic solvent buffer that was at least 8 Å from the protein. To be consistent with the current FEP+ workflow for neutral transformations, no counterions or salt was added to the systems. The systems were relaxed using the standard FEP+ minimization and thermalizing protocol.

Each system was simulated 20 times for 1 ns with a different random seed using the local mode of GCMC-MD. Configurations were recorded every picosecond. A hydration site was determined to be “occupied” by a water molecule if it was within 1.5 Å of the crystallographic position and no closer to any other hydration site.

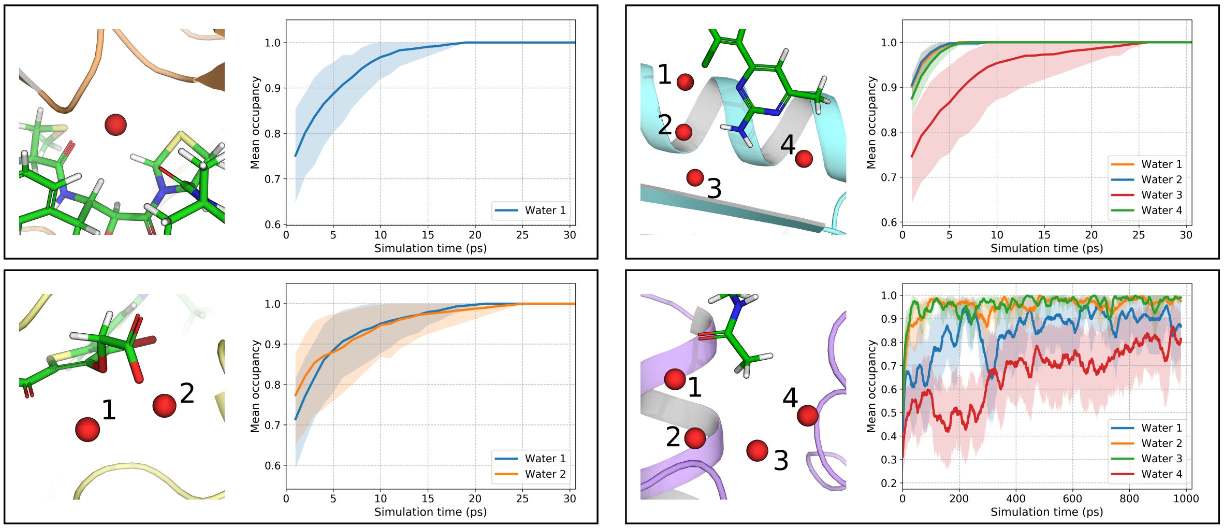

Figure 2: The average water occupancy of conserved hydration sites in 4 protein-ligand systems. Top left, HIV1 protease (PDB code 3FX5); bottom left PTP1B (PDB code 2QBS); top right HSP90 (PDB code 3RLP); bottom right Brd4(1) (PDB code 3JVK). The conserved water molecules (red spheres) were removed from the systems before running 20 repeats of 1 ns-long GCMC-MD simulations. The plots show 10 ps moving-averages of the hydration site occupancies averaged over twenty repeats (lines) and their corresponding 95% confidence intervals (transparent colors). The confidence intervals were calculated by bootstrap sampling over the simulation repeats.

Results

Figure 2 shows the conserved hydration sites and their occupancy, averaged over all repeats, as a function of simulation time. The hydration sites in HIV1 protease, PTP1B, and HSP90 are rapidly populated, with all hydration sites in all twenty repeats being fully occupied within 25 ps (125000 insertion/deletion attempts). Once occupied, these sites remain fully hydrated for the remainder of the simulations. The four hydration sites in Brd4(1) on the other hand demonstrate partial occupancy, with average occupancy in hydration site 4 probably requiring more than 1 ns to fully equilibrate. While the average occupancies in Brd4(1) are slower to converge than the other systems, there are multitudinous instances when all 4 sites are simultaneously occupied; the longest first time for this to occur in one of the repeats is 177 ps (885000 insertion/deletion attempts).

Validation of the μVT free energy calculations

As discussed in the Theory section, binding free energies calculated in the μVT ensemble should deviate minimally from free energies calculated in the NPT ensemble. However, direct comparisons between binding free energies calculated in μVT and NPT can be hindered by the long autocorrelation times of buried water molecules that can occur in the NPT ensemble. Two model scenarios were used to probe the expectation that free energies in μVT and NPT are sufficiently close and thereby validate our implementation of GCMC water sampling in FEP+. In the first model scenario, absolute hydration free energies were calculated and compared using μVT and NPT. Unlike deep within the interior of a protein, there are no barriers that hinder water sampling, so hydration free energies that are calculated in NPT should be directly comparable to those calculated in μVT. In the second model scenario, the relative binding free energy between two ligands was computed in μVT and NPT. The ligand transformation involved the displacement of a buried water molecule. While this is automatically handled in μVT, alchemical decoupling is required in NPT to account for the buried solvation contribution to the relative binding free energy. The predictions from both protocols were expected to be in close agreement.

Comparing hydration free energies in μVT and NPT

Two hundred small molecules were randomly selected from Schrödinger's small molecule test set[23] and their absolute hydration free energies were calculated in NPT, μVT, and – as a negative control – NVT. It was expected that the μVT hydration free energies would be in much closer agreement with NPT than the NVT free energies. The small molecules were solvated in a cubic box whose faces were at least 15 Å from the solute. As a trivial source of error in NVT free energy calculations comes from having unequilibrated system densities, all simulations were first equilibrated with GCMC. For consistency, the same GCMC equilibration method was performed before the μVT and NPT calculations. Before launching into the REST stages of the calculations in NVT and NPT, it was ensured that each lambda window had the same number of water molecules. Hydration free energies were calculated using 12 lambda windows, each of which were run for 10 ns in the REST stage.

Results

Both relative and absolute hydration free energies calculated in the μVT and NPT ensembles were found to be very close agreement. The root-mean-square deviation (RMSD) between the absolute hydration free energies calculated in μVT and NPT was found to be 0.082 ± 0.004 kcal/mol. In contrast, the RMSD between the absolute hydration free energies calculated in NPT and NVT was found to be 0.213 ± 0.004 kcal/mol. Significant differences between NVT and NPT can arise because the density of the surrounding water in NVT can be significantly perturbed by the addition or removal of a solute. In contrast, the water density in the NPT and μVT can relax via volume fluctuations and molecule number fluctuations, respectively.

Comparing relative binding free energies in μVT and NPT

Relative binding free energy calculations in which a perturbation changes the occupancy of a buried water molecule require care in pure NPT protocols. If a water molecule cannot adapt to the perturbation – for instance, because of steric hindrance – the water molecule must be transferred to the bulk solvent prior to carrying out the perturbation in a separate set of calculations. The change in the free energy for this transfer (the standard binding free energy of water for a given site) can be calculated using double decoupling[5] – so called because it involves alchemically decoupling the water molecule from a particular hydration site and by decoupling it from bulk solvent. The free energy to decouple a water molecule from bulk solvent is the hydration free energy of water, which has already been calculated for the SPC water model as part of the calibration of the chemical potential.

In choosing a test system to compute relative binding free energies in μVT and in NPT with double decoupling, there are two primary criteria. The first is that the hydration site should be sufficiently buried so that no other water molecules can occupy the site when a water molecule is being decoupled from it. This is because in MD, unlike pure Monte Carlo packages,[6,13] it is extremely difficult to apply restraints that prevent other water molecules from entering a hydration site. The second criteria is that the perturbation between the two ligands must be sufficiently small to ensure the calculation can be reasonably converged in a short simulation time. Two HSP90 ligands, shown in Figure 3, taken from a study by Woodhead et al., fulfill these two criteria.[33] The transformation between the two ligands involves an addition of a methyl group that displaces a buried water that is completely occluded from bulk. Following the naming scheme by Woodhead et al., the ligand containing a methanol group is referred to as ligand 1, and the ligand with the added methyl – transforming the methanol to a methoxy group – that displaces a water molecule is referred to as ligand 2.

Methods

The relative binding free energy between ligand 1 and ligand 2 was calculated in μVT using 3 repeats, where each repeat comprised 24 lambda windows at 10 ns each. PDB entry 2XJJ was used as the starting structure. Averaged over the 3 repeats, the relative binding free energy was predicted to be 2.04 ± 0.08 kcal/mol, which was in good agreement with the experimental value of 1.98 kcal/mol.[33] The two buried water molecules – shown bound with ligand 1 in Figure 3 – were removed from the starting structure and the relative binding free energy was computed in the NPT ensemble using the same simulation length and number of repeats as with the μVT calculations. The contribution that the water molecules have to the relative binding free energy in the NPT ensemble was computed in separate decoupling calculations. In the presence of ligand 1, both water molecules were decoupled, and in the presence of ligand 2, the other water molecule was decoupled. Decoupling both water molecules, rather than just the one that was sterically displaced, accounts for the possibility that the remaining hydration site has a partial occupancy; *a priori* the water occupancy of this site in the presence of ligand 2 was not known for certain, despite the fact that no water is present in the cyrstal structure (PDB entry 2XJG).

Restraints on a water molecule are required when decoupling it from a hydration site, otherwise – to the detriment of the convergence – it will drift away from the site when it is partially decoupled. Unlike previous studies, which used absolute position restraints or volume constraints,[5,6] we use relative position restraints to ligand atoms. We restrain only the oxygen atom in the water molecule so that it is free to explore different orientations during the decoupling. Inspired by the restraints used by Boresch et al.,[35] three atoms are chosen from the ligand to define one distance restraint, one angle restraint, and one dihedral restraint. The ligand atoms, shown in Figure 2 of the SI (section 5), were chosen to reduce the possibility of collinear geometries of the four atoms, which breaks the definition of a dihedral angle. The equilibrium distance, angle, and dihedral angle used in the restraints were informed by short equilibrium simulations. Force constants were selected by trial and error to be as small as possible to minimise the perturbation to system, but large enough to ensure the water molecule did not move into a collinear geometry when in the decoupled state. The restraints were applied to the water molecule during the same alchemical schedule that decoupled the water molecule, such that, at λ=0, the water(s) were fully interacting without any restraints and, at λ=1, the water molecule(s) were fully decoupled and fully restrained. In the fully decoupled state, the free energy to apply to restraints can be analytically derived and estimated using the rigid rotor approximation, which is shown in section 5 the SI.

Results

The free energies to decouple the water molecules in the presence of ligands 1 or 2 are shown in Figure 3 and the raw data is contained in Table 1 of the SI. There is excellent agreement between the relative free energy calculated in μVT and the total NPT free energy difference. The disparity in the free energy between the two approaches (labelled as “hysteresis” in Figure 3) is 0.05 ± 0.14 kcal/mol, which is smaller than the statistical variance encountered in typical relative free energy calculations. This result, along with the hydration free energy

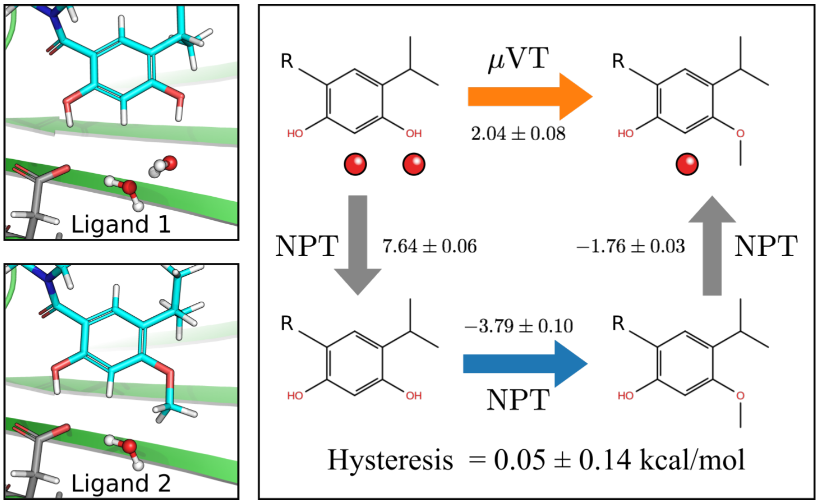

Figure 3: Comparing the relative binding free energies of two ligands computed in the μVT and NPT ensemble. Left: The crystallographic binding modes of two ligands (cyan sticks) bound to HPS90 (green cartoon). ASP 93 is also shown in grey sticks. Relative to ligand 1, ligand 2 displaces a buried water molecule. Right: the relative free energy between ligands 1 and 2 has been calculated in the μVT ensemble and the NPT ensemble. The latter requires the thermodynamic contribution of the buried waters to be calculated using the double decoupling technique. The left vertical leg shows the contribution to remove both water molecules and the right vertical leg shows the free energy to add back the remaining water molecule. The sum of the NPT free energies (shown in kcal/mol) should equal the relative free energy calculated with μVT. The overall discrepancy between the two approaches – the hysteresis – does not differ from zero by a statistically significant amount.

calculations, demonstrates that our implementation of μVT delivers predictions that are practically equivalent to well sampled NPT calculations.

In the absence of the water molecules, the relative free energy to add a methyl to the ligand is predicted to be very favorable using NPT. This is because the transformation does not account for the displacement of water which, as shown by equation 3 in the Theory section, is an unfavorable contribution to the free energy.

The accuracy of GCMC-FEP on water disrupting transformations

In cases where the occupancy of buried hydration sites is affected by ligand transformations, we expect that FEP that is enhanced with GCMC water sampling will have a lower average error compared to FEP that does not used GCMC. We also expect FEP with GCMC to be less dependent on whether or not crystallographic water molecules were included in the protein structure.

Replication of Wahl and Smieško's study

Previously, Wahl and Smieško constructed a data set of ten ligand perturbations across four different proteins to assess the accuracy of a GCMC equilibration protocol that was previously developed by ourselves.[36] X-ray crystal structures exist for all ligands in the study and transformations were made between matched pairs of ligands if the perturbation displaced or interacted with buried water but left the protein and ligand conformations largely unaffected. To explore the sensitivity of FEP on the initial solvation in the binding site, they repeated the FEP transformations starting from both crystal structures of each ligand.

We replicated the ten transformations in the Wahl and Smieško data set using the same two starting structures (which were supplied by Wahl and Smieško in the supporting information[36]) and included the same protein side chains in the REST region as before. Each transformation was run using both starting structures with the μVT, NPT, and NPT pre-solvate protocols. The results are summarized in Table 2 and the raw data is shown in Table 2 of the SI.

Both protocols that utilize GCMC water sampling (μVT and NPT pre-solvate) have lower errors than the pure NPT protocol. The μVT predictions are the most consistent between the two starting structures, with the RMSD being significantly lower compared to

Table 2: The root-mean squared errors (RMSE) (in kcal/mol) of the three FEP protocols on the water disrupting transformations that were assembled by Wahl and Smieško. The root-mean squared deviation (RMSD) of the predictions in the different structures is also shown. Statistical uncertainties show the standard error as calculated by sampling over the ten ligands using bootstrap sampling.

	Edgewise RMSE		
	Without overlapping water	With overlapping water	RMSD
μVT	1.71± 0.24	1.77 ± 0.31	0.41 ± 0.06
NPT	2.39 ± 0.68	2.54 ± 0.60	3.03 ± 1.01
NPT pre-solvate	1.77 ± 0.35	1.80 ± 0.33	1.06 ± 0.26

the other protocols.

Expanding the water displacement data set

To more fully assess the impact that GCMC has on the accuracy of FEP, we expanded on Wahl and Smieško's data to include more proteins and ligands. Ligand pairs were included in our data set if 1) they had the same scaffold or binding mode, 2) a small difference between them was expected to alter the stability (e.g. via steric displacement) of at least one buried water molecule, 3) there was crystallographic evidence for the existence of the buried water molecule(s) in question, 4) a crystal structure existed for a ligand that was in the same congeneric series as the pair, and 5) the affinity of the ligands had been measured with the same method and by the same group. While there was a preference for using high quality direct measurements of affinity (such as isothermal titration calorimetry), this was not always possible for all protein-ligand data sets. Figures 4 and 5 illustrates the data sets that were selected for this study. Two starting structures were used for all of proteins; one structure contained all known crystallographic water molecules, whereas the other had any water molecules that overlapped with at least one of the ligands removed. The FEP transformations were run for 20ns in the μVT ensemble, NPT ensemble, and the NPT ensemble with GCMC pre-solvation. All of the structures and binding affinities used in this study are provided with the supplementary information.

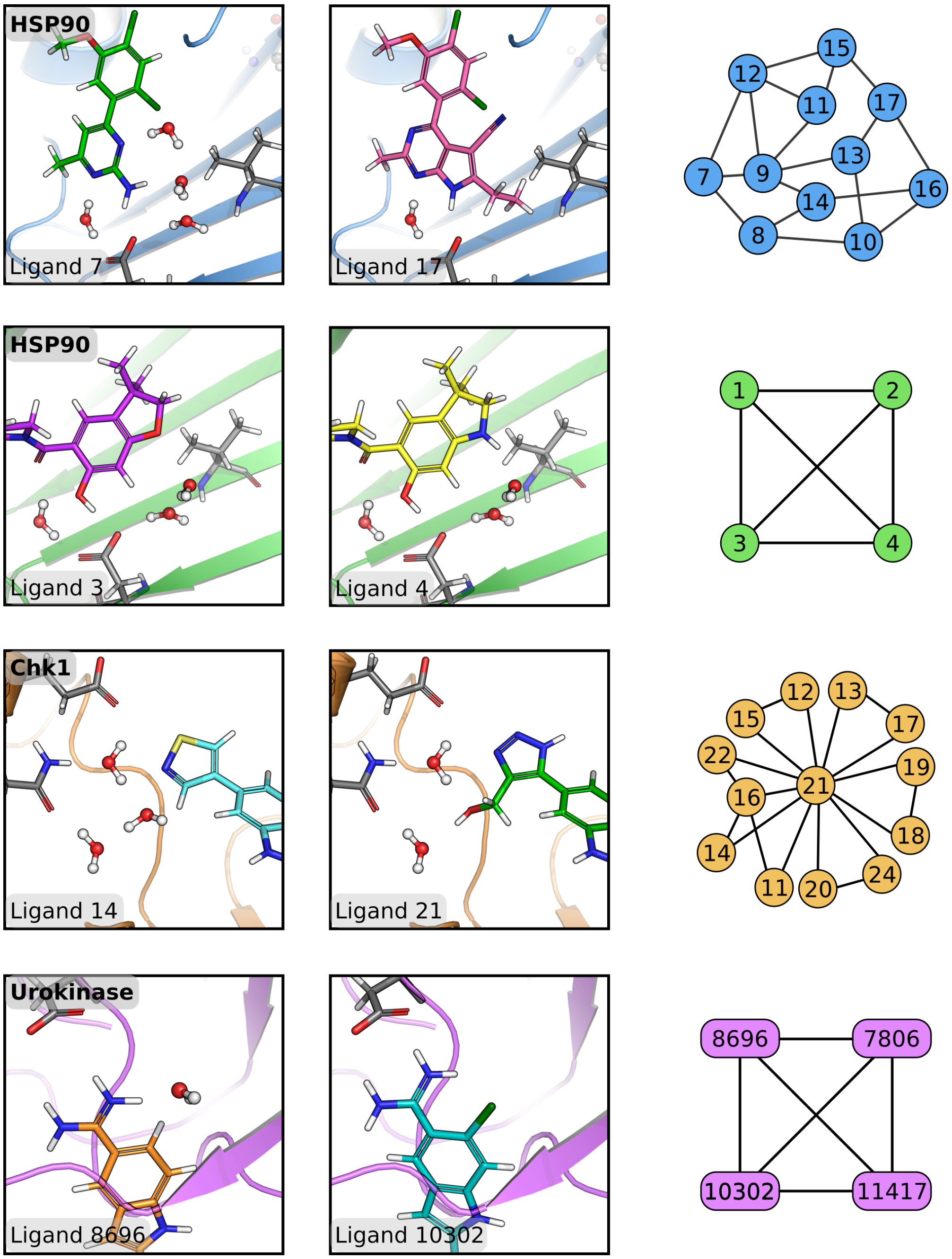

Figure 4: Four out of the eight data sets used to assess the accuracy of GCMC water sampling in FEP+. Examples are shown of ligands that either interact with or displace buried water molecules in each data set, along with the the topology of each FEP map. The ligands names are taken from the original publications. From top to bottom, the data sets are HSP90 from Kung et al.,[37] HSP90 from Woodhead et al.,[33] Chk1 from Fraley et al.,[38] and urokinase from Katz et al.[39,40]

HSP90

HSP90 has four highly conserved water molecules that are buried deep within the binding site in the *apo* state. Three of these water molecules, (those that are adjacent to ASN51, SER52, and ASP93), are displaced or disrupted by the two congeneric series considered in this study.

The first congeneric series for HSP90 was taken from the study by Kung et al. in which the growth of a fused ring and the addition of substituents gradually displace the 3 waters.[37] The experimental binding affinities were taken from a high-throughput competition assay. These measurements should be interpreted with some caution; the binding free energies were measured with ITC on four compounds and these values had an RMSD of 0.86 kcal/mol compared to the faster technique. The protein structure was taken from PDB entry 3RLP (which contains the 3 buried water molecules). The ligand binding modes were based on the binding poses found in PDB entries 3RLP, 3RLQ, and 3RLR.

In the second congeneric series for HSP90, four ligands were taken from Woodhead et al.,[33] two of these ligands are the same as those used for the comparison of GCMC with alchemical decoupling (see Figure 3). All ligands in this series displace the buried water that is closest to ASN51, and the ligands used in this study have different interactions with the two remaining waters molecules, with one ligand sterically displacing one of them. The ligands had their affinities measured with ITC. The binding poses of the ligands was taken from PDB entries 2XAB, 2XJG, and 2XJJ; the protein structure was taken from 2XJJ.

Chk1

Ligands were taken from Fraley et al. in which different rings and ring substitutions interact with a buried pocket that contains three water molecules.[38] While some ligands do not sterically displace the waters, some or all of the waters can be displaced by the modifications in this series. The ligands were aligned to the crystallographic ligand in the structure that accompanied the publication (PDB entry 2HOG). In that structure, one of the buried water molecules is displaced. Despite knowing the overall binding modes of the ligands, the orientations of the ligand perturbations required further modelling. The orientation of one of the ligand modifications (compound 24) was taken from PDB entry 2C3K with the rest predicted using Glide.[41]

Some of the ligands contained aromatic nitrogens whose tautomeric state was uncertain, for instance, ligand 21 in Figure 4. Tautomers for all ligands were generated with Epik.[42] After using Epik, uncertainty remained for ligands 13 and 21 which was resolved with Schrödinger's macro-pKa prediction protocol.[43]

As 2HOG is missing residues 44 to 50 as well as the side chain for TYR 20, PDB entry 2E9V (resolved at 1.9 Å, 0.65 Å backbone RMSD from 2HOG) was used as the starting protein structure for FEP. The oxygen positions of the unperturbed three-water network were also taken from 2E9V.

Urokinase

Four ligands were taken Katz et al.[39,40] to explore the displacement of the buried water (in the vicinity of ASP 189) by F and Cl. Although more chemical substituents were synthesized and assayed by Katz et al., the binding affinities of the four ligands selected for this study were measured more rigorously than many others.

The protein structures and ligand binding modes were taken from PDB entries 1GJ7 and 1GJB. PDB entries 1GJ7 and 1GJB show that the addition of Cl to the scaffold shifts the binding mode of the ligand towards ASP 189 and the buried hydration site water by approximately 1 Å. As a result, the ligand pose in 1GJB was used as the basis for the FEP map in the presence of the buried water molecule, and pose in IGJ7 was used for the FEP map in the absence of the water molecule. ASN 192 was added to the REST region as both 1G7J and 1G7B resolve this residue in 2 conformations. Experimental evidence accumulated by Katz et al. indicates that HIS 57 and the phenol in the ligand scaffold form a salt bridge at pH 7. Thus, HIS 57 was prepared in the biprotonated state and the phenol oxygen was deprotonated.

Thrombin

Baum et al. explored 26 small modifications to a single scaffold where all except one all consist of small additions to a phenyl ring.[44] A buried water molecule (adjacent to ASP 189) that is present with the unadorned phenyl group (PDB entry 2ZFF) is displaced by the addition of CH3, F, or Cl in the *meta* position (PDB entries 2ZF0, 2ZDZ, and 2ZC9 respectively). The binding affinities of all the ligands were measured using a kinetic competition assay and a subset of 12 ligands also had their affinities measured with ITC. The absolute difference of the measured affinities between the two techniques range from 0.02 kcal/mol to 0.80 kcal/mol with an RMSD of 0.50 kcal/mol. The 12 ligands that were assayed with ITC were used previously by a previous FEP+ study.[1] The experimental affinities were taken from the competition assay given its greater coverage of ligands. Only the neutral ligands that had unambiguous stereochemistry (a total of 24) were used in this work.

PDB entry 2ZFF was used for the protein structure and as the basis for the binding modes of the ligands. Restraints were placed on a Na+ ion adjacent to ASP 189 to prevent its diffusion. As indicated by Figure 5, ligand 5 was placed at the center of the map with all ligands connected to it. The orientation of most of the ligands with substitutes in the *ortho* and *meta* position was unknown and highly unlikely to change during the course of the simulation. Two orientations were added to the FEP map, both with an edge to ligand 5 and edge between them. The relative free energies between these modes were combined into a single relative free energy prediction using the same scheme as established previously.[47,48]

Scytalone dehydratase

The ligands developed by Chen et al.[45] for scytalone dehydratase are commonly used for validating protocols that displace water molecules (see, for instance Michel et al.[13,49]). This series explores nitrogen substitutions to quinoline derivatives as well as the growth of a

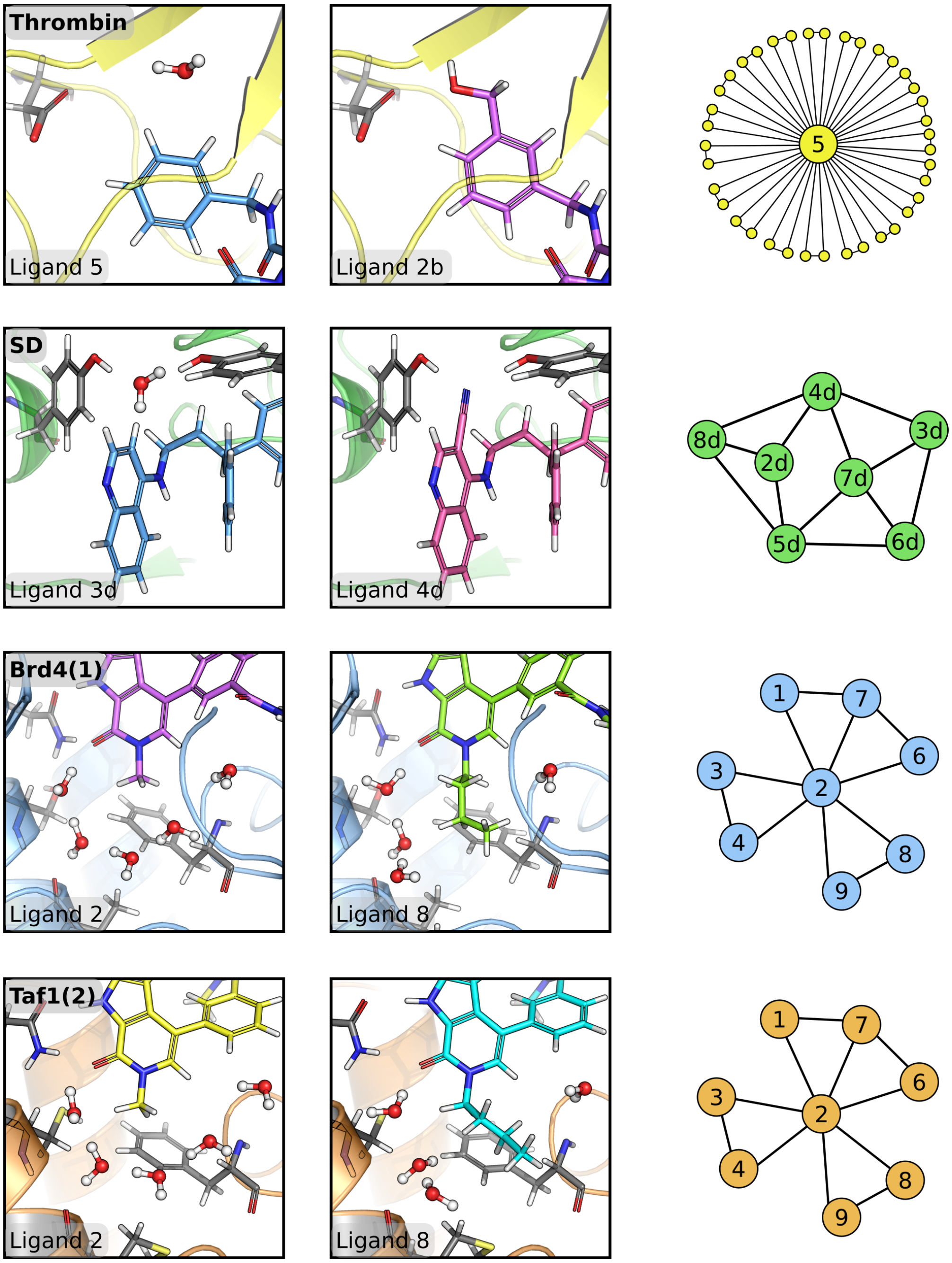

Figure 5: Four out of the eight data sets used to assess the accuracy of GCMC water sampling in FEP+. Examples are shown of ligands that either interact with or displace buried water molecules in each data set, along with the the topology of each FEP perturbation map. The ligands names are taken from the original publications. From top to bottom, the data sets are thrombin from Baum et al.,[44] scytalone dehydratase from Chen et al.,[45] Brd4(1) and Taf1(2) from Crawford et al.[46] The thrombin map (top, yellow cartoon) explores two orientations for ligands that have substitutions in the *ortho* and *meta* position. These orientations are represented by two connected nodes in the map.

cyano-nitrile group into a conserved hydration site. PDB entry 3STD – in which the buried water has been displaced – was used as the protein model for FEP and as the basis for binding modes of the ligands. The location of the water oxygen atom was taken from PDB entry 5STD.

Based on an experimental measurement of a close chemical analogue,[50] the quinoline nitrogen in ligand 3d of this set (shown in Figure 5) is expected to have a pKa of 8.0. As the experimental assay was conducted at pH 7,[45] this ligand will be protonated in solvent. A preliminary FEP+ calculation between the protonated and neutral form of ligand 3d when bound to the protein predicted that the pKa lowers by roughly 9 units. Based on this result, it was assumed that the remaining ligands bind in their neutral forms. Schrödinger's method to determine macro-pKas was run on the remaining ligands to estimate the pKas of the aromatic nitrogens.[43] These predictions, shown in Table 4, have an expected error of about 1 pKa unit, making the protonation state of ligands 2d, 6d, and 8d in solvent uncertain. A correction, described previously,[51] was applied to predicted relative free energies to account for the pKa of the ligands in solvent. The uncertainty in the macro-pKa predictions – particularly for 2d, 6d, and 8d – adds a degree of uncertainty to the scytalone dehydratase FEP predictions that is not present in the other data sets.

Bromodomains Brd4(1) and Taf1(2)

There are four hydration sites that are present in all *apo* bromodomains whose propensity for displacement has been previously investigated by GCMC.[52] Eight ligands were taken from a study by Crawford et al. in which an alkyl chain was grown into the buried water pocket for a large number of bromodomains.[46] Crystallographic structures that accompanied the work show some of the four waters are displaced for Brd4(1) and Taf1(2). The same FEP ligand map was used for both Brd4(1) and Taf1(2).

For Brd4(1) and Taf1(2), PDB entries 5I80 and 5I29 were used as the structure that contained all four buried molecules, respectively, whereas PDB 5I88 and 5I1Q were used as the structure whose water molecules had been displaced and disrupted. The ligand binding modes were based on 5I88 and 5I1Q for Brd4(1) and Taf1(2), respectively.

Results

Table 3: The mean unsigned error (MUE) and root-mean squared errors (RMSE) in kcal/mol of the three FEP protocols on the water disrupting maps shown in Figures 4 and 5. The errors for the protocols marked with $*$ have been calculated with the highest confidence data sets - the Kung HSP90 has been excluded due to the high experimental uncertainty and the scytalone dehydratase data has been omitted owing to uncertainty of the ligand pKas. The root-mean squared deviation (RMSD) between the FEP predictions when using different initial structure is also shown. The standard errors have been calculated by bootstrap sampling over all ligand perturbation.

	Without overlapping water		With overlapping water		
	MUE	RMSE	MUE	RMSE	RMSD
μVT	1.11 ± 0.09	1.43 ± 0.11	1.15 ± 0.10	1.47 ± 0.11	0.43 ± 0.04
NPT	1.40 ± 0.17	2.23 ± 0.36	1.52 ± 0.15	2.12 ± 0.19	2.32 ± 0.30
NPT pre-solvate	1.11 ± 0.10	1.42 ± 0.11	1.15 ± 0.12	1.51 ± 0.12	0.59 ± 0.05
μVT*	0.89 ± 0.09	1.18 ± 0.13	0.98 ± 0.10	1.29 ± 0.13	0.46 ± 0.05
NPT*	1.06 ± 0.11	1.41 ± 0.14	1.22 ± 0.13	1.66 ± 0.19	1.35 ± 0.23
NPT pre-solvate*	0.94 ± 0.09	1.22 ± 0.11	0.96 ± 0.11	1.33 ± 0.14	0.61 ± 0.06

Sampling water with grand canonical Monte Carlo, either throughout the entire FEP calculation (μVT) or as an equilibration step (NPT pre-solvate), has a significantly lower error than the protocol without GCMC (NPT). Table 3 and Figure 6 aggregate the root-mean-squared error (RMSE) from all 139 perturbations from the eight protein data sets. The difference between the RMSEs of the μVT and pre-solvate protocols over all the data sets is not statistically significant.

Figure 6 shows that when using the NPT protocol, it is possible to pick a starting arrangement of water that results in a lower error in FEP compared to the other arrangement. For instance, the NPT protocol that has the lowest RMSE with urokinase is when the starting structure does not contain the conserved water molecule (see Figure 4). The opposite is true with HSP90 data set from Woodhead et al., as the lowest error occurs when the

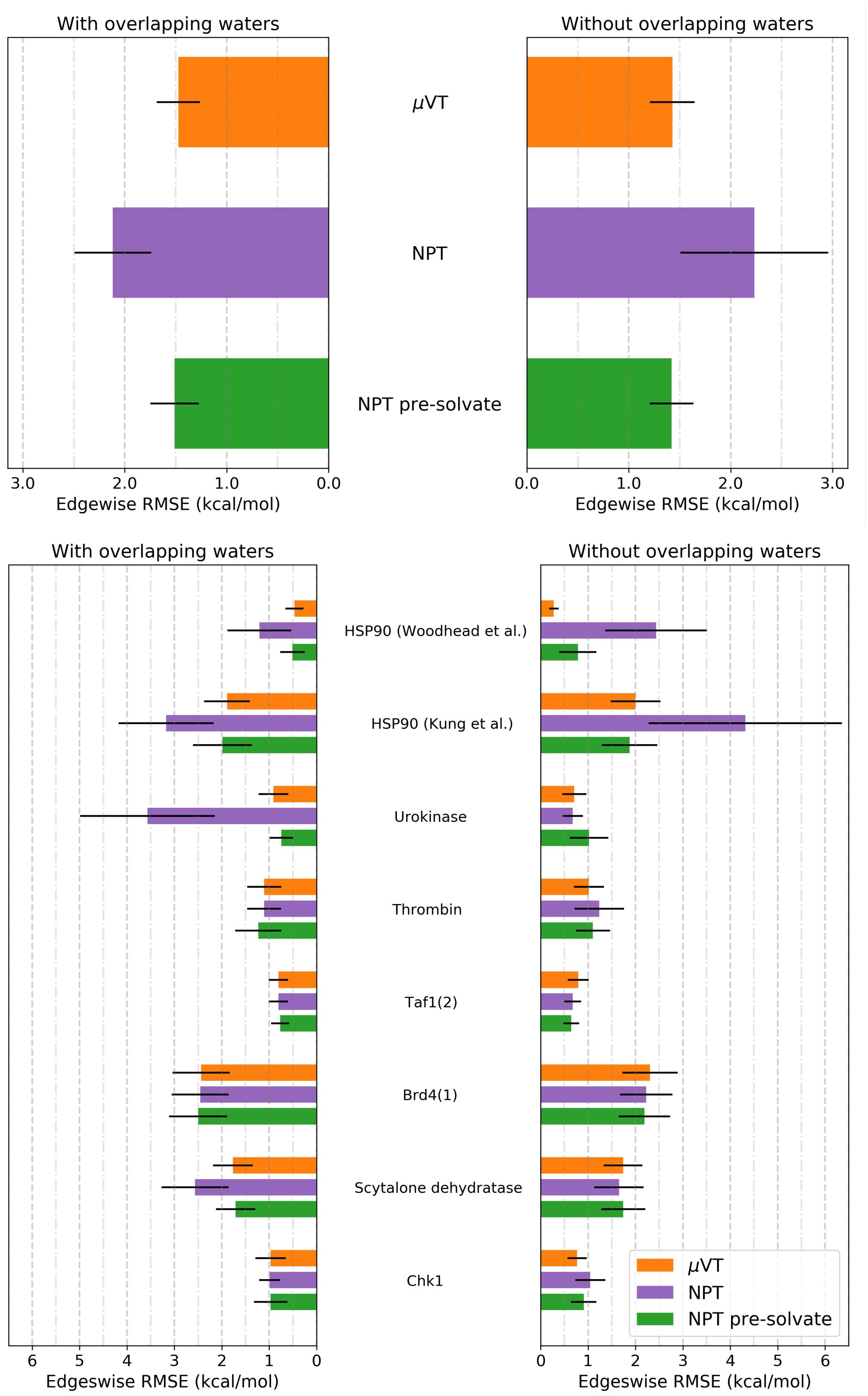

Figure 6: The edgewise error for all data sets. Top panel: the RMSE aggregated over all pairs of compounds in each map. Lower panel: the RMSE for each data set. Error bars reflect standard errors which have been calculated by bootstrap sampling over all perturbations.

overlapping conserved water (seen in figures 3 and 4) is included in the starting structure. For some systems, such as Taf1(2), NPT is equally accurate irrespective of whether the starting structure contains the conserved overlapping waters or not. Prospectively however, one would not know whether starting with or without overlapping water would be detrimental to the results.

Table 3 and Figure 4 of the SI show the root-mean squared difference (RMSD) between predictions when calculations are started with and without overlapping water molecules. With an RMSD of 2.32 kcal/mol, the NPT protocol is completely unreliable for perturbations that disrupt or displace buried water molecules. The μVT protocol drastically reduces the dependence of the free energy predictions on the starting placement of water relative to the NPT protocol. The lowest RMSD of 0.43 kcal/mol between predictions is found with the μVT protocol. The higher consistency of the GCMC predictions indicates that accelerated water sampling, run either before or during the production free energy calculation, is essential in a prospective setting. Other sampling metrics – discussed in below – provide further evidence that the μVT protocol is the most robust compared to the NPT and pre-solvate protocols.

With either the μVT or pre-sovlation NPT protocol, the lowest RMSE occurs with the structures that do not contain water molecules that overlap with the ligand. This is to be expected, as GCMC is first run after the systems have been minimized; water molecules that initially overlap with ligands have the potential to distort the binding modes during the minimization. The most robust strategy is therefore to not include conserved water molecules in a FEP map if they sterically overlap with the ligands. As exhibited in Figure 2, any missing water molecules in a structure will be rapidly replaced with GCMC.

Despite the improved accuracy from GCMC water sampling, Figure 6 shows that three systems, namely the Kung et al. HSP90 set, Brd4(1), and scytalone dehydratase have RMSEs that are at least 1.5 kcal/mol when using either the μVT and NPT pre-solvation protocols. As discussed below, the higher error in the scytalone dehydratase data set appears to stem from the uncertainty in the predicted pKas of the ligands. A higher error in the

Kung et al. HSP90 data set is also to be expected given the higher experimental error in this data set. Table 3 shows the RMSEs of the methods when these two data sets are removed from the analysis. Despite the higher error, Figure 7 shows that the rank ordering of the predictions with the μVT protocol are still appreciable with HSP90 (Kung et al.) and scytalone dehydratase. A contributing factor to the error of the μVT protocol may be an underestimation of the free energy to desolvate cavities, which is analyzed and discussed below.

pKa correction in scytalone dehydratase

Unlike the other data sets, the relative binding free energy predictions of scytalone dehydratase included a pKa correction to account for the protontion state of the ligands in solvent. This correction was calculated from the estimated pKas of the ligands using a QM macro-pKa protocol.[43] The predicted pKas are shown in Table 4. It is expected that the uncertainty in the predicted pKas contributes to the RMSE of this data set, which is 1.74 ± 0.21 kcal/mol with μVT.

To investigate the sensitivity of the relative binding free energy predictions on the estimated pKas of ligands 2d, 6d, and 8d, the pKa correction for these ligands was recalculated 5000 times by drawing the pKas from a normal distribution centered on the values shown in Table 4 with a standard deviation of 1 pKa unit – the estimated error of the pKa prediction. These pKas were applied to the results from the μVT protocol when the starting structure did not contain the overlapping water molecule. Ninety-five percent of the results had RMSEs between 1.49 kcal/mol and 1.99 kcal/mol, indicating a strong dependence of the error on the predicted pKas.

In order to find the macro-pKa of ligands 2d, 6d, and 8d that achieved the lowest error, the RMSE was minimized by optimizing the pKas of ligands 2d, 6d, and 8d subject to the constraint that the solutions could be no more than 1 pKa unit away from the QM macropKa prediction. With μVT, the lowest RMSE achievable was 1.48 ± 0.16 kcal/mol.

This procedure was repeated on the NPT and NPT pre-solvate predictions, with the resultant pKas shown in Table 4. The direction of pKa change was consistent over the three FEP protocols, suggesting these changes are not solely due to statistical noise and may reflect an error in predicted pKas.

Table 4: The estimated pKas of the compounds in scytalone dehydratase data set. The pKa of ligand 3d was taken from a close chemical analogue, whereas all others were predicted using a QM macro-pKa protocol.[43] The ligand names have been taken from Chen et al.[45] The expected error for these predictions is expected to be roughly ±1 pKa unit. The 'Optimized to FEP' multi-column refers to the pKas that have been optimized to minimize the RMSEs of the FEP predictions in the μVT, NPT, and NPT pre-solvate protocols on the structures that start without overlapping water. The optimization was constrained such that the prediction could differ by no more than 1 unit from the QM macro-pKa prediction.

		Optimized to FEP map		
Ligand	Estimated macro-pKa	μVT	NPT	NTP pre-solvate
2d	6.33	6.02	5.50	5.57
3d	8.0	-	-	-
4d	5.45	-	-	-
5d	4.06	-	-	-
6d	7.09	8.08	8.07	8.09
7d	5.42	-	-	-
8d	7.13	7.69	7.52	7.65

The cost of desolvation

The Theory section of this manuscript describes how the free energy to grow into a buried pocket can be erroneously predicted to be too favorable when buried waters do not have a high enough occupancy during the simulation. This feature may be present in the thrombin and bromodomain data sets.

In the thrombin data set, phenyl decorations that occupy a buried hydration site are predicted to bind more strongly than experiment. Averaging over all the perturbations that displace buried water, the mean difference between the μVT predictions (starting with a structure without buried water) and experiment is -0.60 [-0.96, -0.26] kcal/mol. Here, square brackets show the 95% confidence intervals that have been estimated by bootstrap

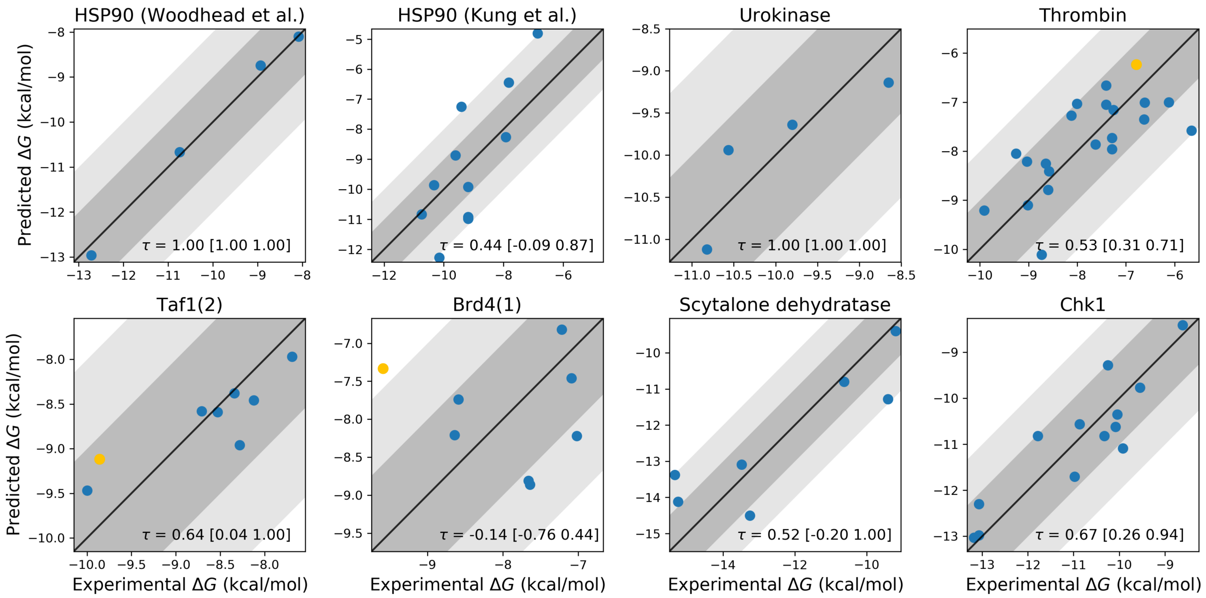

Figure 7: The correlation between the experimental binding free energies and the predicted binding free energies for each test system using the μVT protocol. The starting protein structures for these results did not contain any overlapping water molecules. The dark and light grey regions are within 1 kcal/mol and 2 kcal/mol of $y = x$ line, respectively. Kendall's tau is shown for each data set with 95 % confidence intervals that have been estimated by boostrap sampling the ligands. The affinity of the ligands highlighted in orange in the thrombin, taf1(2), and brd4(1) plots may be affected by an underestimation of the binding strength of water and are discussed in the section titled "The cost of desolvation".

sampling the edges. With NPT, the difference between the water displacing predictions and experiment is -0.73 [-1.18, -0.35] kcal/mol on average. The predicted over-stabilization of the phenyl decorations from these calculations could come from an underestimation of the binding affinity of water at that site, as illustrated by equation 3.

In the bromodomain data sets Brd4(1) and Taf1(2), the binding pockets "without overlapping water" are not devoid of water like with the thrombin data set. Nevertheless, there is an indication that the affinity of the buried water is also underestimated in these data, particularly so with Brd4(1). Collecting all the predictions that grow the hydrophobic chain from the smallest ligand (ligand 2), the μVT predictions differ from experiment by -0.84 [-1.11, -0.56] kcal/mol in Taf1(2) and by -2.58 [-3.17, -1.96] kcal/mol with Brd4(1). These values indicate that the affinity of ligand 2 is being underestimated relative to the compounds that displace or disrupt the buried water network. This underestimation could be ameliorated using a water model in which the water network was bound more strongly to the protein. The erroneously low predicted affinity for ligand 2 could also be a result of unexplored binding modes, where additional, stable binding modes would lower the predicted affinity.

The underestimation of the water stability in these cases could be due to the SPC water model that was used, and may be alleviated with a different water model. The effect of different water models on the accuracy of FEP predictions will be investigated in future work.

Comparing the μVT and NPT pre-solvation protocols

Although the prediction accuracy of the μVT and the pre-solvate NPT protocols are statistically equivalent over the whole data set, the quality of the sampling is better with μVT. Two features of the FEP+ workflow were used to probe the FEP sampling quality: the hysteresis of the thermodynamic cycles in the perturbation maps and the degree of convergence of the solute-tempering replica exchange method.

Comparing Hysteresis

The hysteresis of a closed perturbation cycle is measured by the sum of the relative free energies within the cycle. Because free energy is a state function, this sum should be zero in the infinite sampled limit. The lower the hysteresis, the more consistent the configurations and states sampled by the transformation edges within the cycle appear to be. All of the FEP maps used for the water disrupting data set contain a number of closed perturbation cycles. The hystereses for all of the cycles in the water disrupting data set are aggregated and shown in Figure 8, where it can be seen that the μVT protocol has a lower cycle closure error than the NPT pre-solvate one. Calculating the differences of the hystereses for the same cycles of the NPT pre-solvate and the μVT protocols and averaging over both starting structures reveals that the hysteresis in the μVT protocol is better by a statistically significant degree. Specifically, the hysteresis of μVT protocol is 0.21 [0.13, 0.29] kcal/mol lower on average than the hysteresis for the pre-solvate NPT protocol.

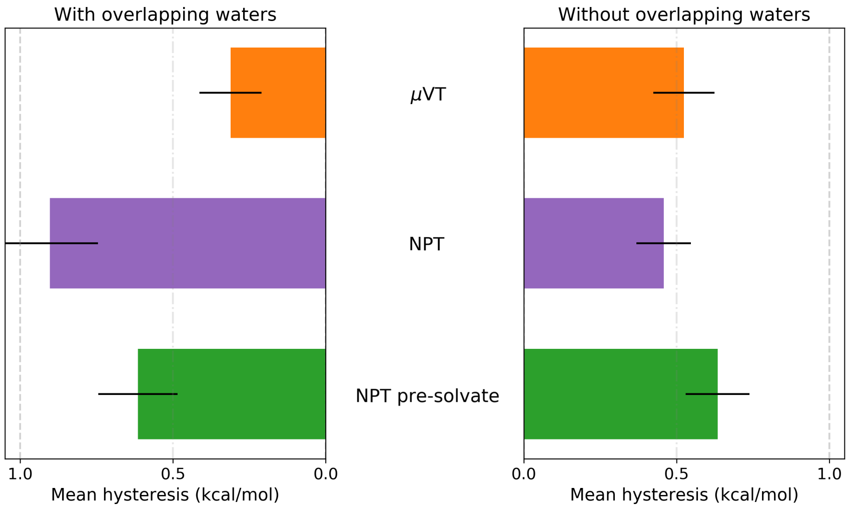

Figure 8: The average hysteresis over all closed cycles in the maps.

Comparing the quality of replica exchange

A score was devised to quantify and compare the sampling efficiency of replica exchange in the three sampling protocols. Replica exchange allows the different replicas to perform a random walk over the lambda windows. In the infinite sampling limit, each replica will have occupied all of the lambda windows with the same frequency. The deviation of the normalized lambda window sampling frequencies from a uniform distribution is therefore a measure of sampling quality. Owing to its simplicity, the Kullback-Leibler divergence was used to measure the deviation of the lambda window sampling frequencies, denoted f, from a uniform distribution, denoted u:

$$D(f||u) = \sum_{i=1}^{N} f_i \ln\left(\frac{f_i}{u_i}\right), \tag{5}$$

where N is the number of lambda windows and $u_i = 1/N$. A "better" sampled replica has a score, i.e. $D(f||u)$, that is closer to zero. This direction of the Kullback-Leibler divergence, as opposed to $D(u||f)$, the ensures that the score remains finite even if some of the sampling frequencies are zero.

Figure 9 shows the replica exchange scores for every replica in every perturbation edge in the water displacing data set. The NPT protocol with a pre-solvation stage has a noticeably worse score (higher deviation from uniform) than the μVT and NPT protocols. To ensure that the improvement displayed in Figure 9 is statically significant, the scores were averaged over each edge. These edgewise scores are statistically independent of each other, which facilitates a statistical test for significance. Calculating the difference between the scores of the same edge reveals that on average, the μVT protocol score is 24.8 [17.7, 32.1] % lower (i.e. better) than the pre-solvate NPT score. Compared to NPT, the μVT scores are 13.6 [7.6, 20.1] % lower. The square brackets denote 95% confidence intervals that have been calculated using bootstrap sampling of the edges.

Evaluating the performance of GCMC-FEP on a previous benchmark

Previously, we applied the FEP+ workflow on a data set comprised of eight proteins, 199 ligands, and 330 perturbations.[27] Using OPLS 2.1 and 5 ns of simulation time per lambda window, an RMSE of 1.1 kcal/mol was obtained across the perturbations. To further evaluate the performance of GCMC water sampling in protein-ligand free energy calculations, the μVT, NPT, and NPT pre-solvate protocols were applied to this data set.

Methods

Seven out of the eight of the protein-ligands structures and FEP maps were taken from a previous FEP study of ours.[1] These proteins were bace1, cdk2, jnk1, mcl1, p38, ptp1b, and

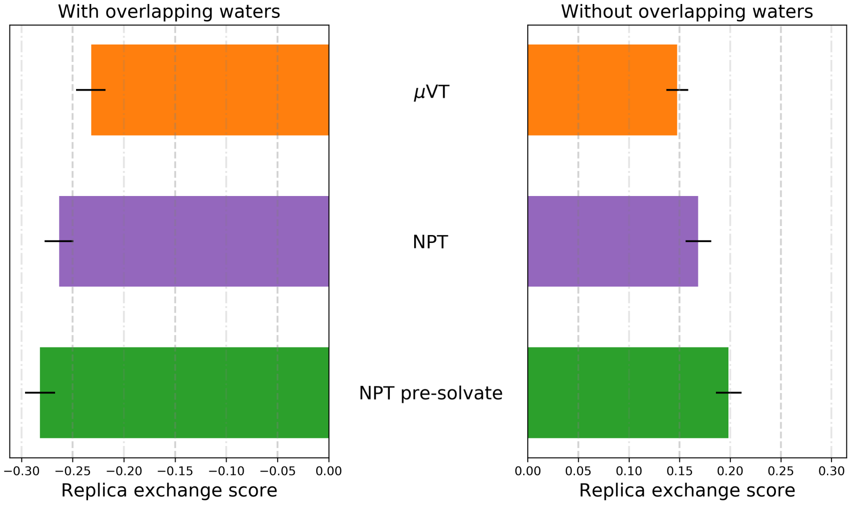

Figure 9: The average replica exchange score (calculated using equation 5) over all edges and replicas in the water disruption data set. Error bars indicate the standard errors that have been calculated using bootstrap sampling. The score measures the deviation the sampling distribution over the lambda windows has from a uniform distribution. On average, the μVT protocol produces samples that have a more uniform sampling distribution than the NPT pre-solvate method. The μVT replica exchange scores are better than the solvate scores by a statistically significant degree when the scores from the same edge are compared. Replica exchange is more efficient when structures do not start with water molecules that directly overlap with ligands.

tyk2. A thrombin FEP was also included in that study, but this data set was omitted as all of the ligands within it are already part of the thrombin data set that was analyzed earlier in this manuscript. Over three-quarters of the ligands in the bace1 sterically overlap with a buried water molecule that is present in the starting structure. Notably, the starting structures for cdk2, jnk1, mcl1, and tyk2 were completely absent of any buried water molecules. With the exclusion of thrombin, there are a total of 314 FEP edges in the seven protein data sets in this collection. Each edge was simulated for 20 ns using the μVT, NPT, and NPT pre-solvate protocols using the OPLS3e forcefield.

Results

Table 5 shows the total error of the μVT, NPT, and NPT pre-solvate protocols over the seven protein data sets that we previously studied.[1] The three methods have RMSEs that are within statistical uncertainty of each other. Figure 10 shows how the RMSEs are distributed for each individual system.

As demonstrated earlier in this manuscript, the FEP predictions from NPT simulations that do not use any form of GCMC sampling can be very dependent on the starting placement of water. The data set where this sensitivity should be most evident is bace1, where a number of ligands displace a buried water molecule. Under the expectation that the μVT predictions are more converged than the NPT predictions as a result of the buried water molecule, the bace1 FEP map was repeated in NPT and run for 50 ns per edge.

Bace1 convergence analysis

The bace1 structure was taken from the PDB entry 4DJW which contains two buried water molecules in the volume enclosed by the violet mesh shown in the leftmost panel of Figure 11. Twenty-eight out of the 36 the ligands in the FEP map have chemical groups, such as Cl, cyano-nitrile, and methoxy groups, that sterically overlap with one of the two buried water molecules. The center panel of Figure 11 demonstrates that when the pure NPT simulations

Table 5: The mean unsigned error (MUE) and root-mean-squared error (RMSE) in kcal/mol of each FEP protocol across the 314 FEP edges taken from our previous study.[1] The RMSEs for the individual protein data sets are shown in Figure 10. The errors presented here were obtained after simulating for 20 ns per lambda window. The NPT RMSE increases to 1.02 ± 0.04 kcal/mol when including the final 20 ns of bace1 simulations after they were extended to 50 ns per lambda window.

	MUE	RMSE
μVT	0.81 ± 0.04	1.03 ± 0.05
NPT	0.80 ± 0.03	1.00 ± 0.04
NPT pre-solvate	0.81 ± 0.04	1.03 ± 0.04

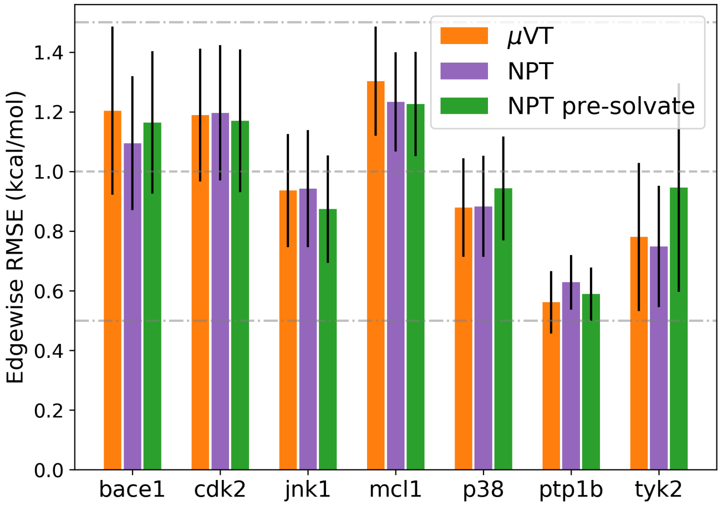

Figure 10: The edgewise RMSE of the μVT, NPT, and NPT pre-solvate protocols on data sets that were collected by a previous publication of ours.[1]

are extended up to 50 ns, the $\Delta\Delta G$ predictions converge on the predictions from the 20 ns μVT predictions. Table 6 shows that convergence to the μVT predictions is accompanied by an increase in the RMSE; when the first 30 ns of the extended NPT simulations are discarded, the RMSE very close to the RMSE of μVT.

The slow convergence of the NPT predictions is a result of the slow diffusion of a single water molecule from a buried pocket. The rightmost panel in Figure 11 shows the water occupancy of the subpocket takes at least 50 ns to fully equilibrate with the water displacing ligands using the NPT protocol, but is fully equilibrated from the start of the μVT simulations.

If the presence of the clashing water did indeed hinder convergence in NPT, then removing that water from the starting protein structure should result in NPT predictions that are closer to the μVT predictions. To test this, the overlapping water was removed from the protein and the FEP map was re-run with NPT for 20 ns. The RMSE of these predictions, shown in Table 6 are closer to the RMSE of the μVT protocol. The RMSD of the μVT predictions and the NPT predictions after 20ns initially *with* the overlapping water was 0.39 ± 0.04 kcal/mol. *Without* the overlapping water molecule, the NPT predictions have an RMSD of 0.28 ± 0.03 kcal/mol relative to the μVT predictions. Thus, removing the overlapping water results in predictions that are closer to the μVT predictions.

As established with the water disrupting data sets, the bace1 predictions made with μVT are more robust with respect to the initial placement of water than the NPT predictions. When the FEP map was re-run in μVT without the overlapping water, the RMSD between the μVT predictions made with and without the initially clashing water is 0.22 ± 0.03 kcal/mol compared to an RMSD of 0.35 ± 0.04 kcal/mol with NPT.

Table 6: The RMSE (in kcal/mol) of the μVT and NPT protocols – both run for 20 ns – on the bace1 system. The starting protein structure in our original study contained a water molecule that overlapped with many of the ligands.[1] The FEP simulations were extended to 50 ns per lambda window in NPT using the structure that contained the overlapping water molecule to assess the level of convergence. The last 20 ns of these extended simulations produce an RMSE that is closer to the μVT predictions, as do the NPT predictions when the starting structure does not contain the overlapping water molecule.

	With overlapping water	Without overlapping water
μVT	1.20 ± 0.14	1.20 ± 0.13
NPT	1.11 ± 0.12	1.18 ± 0.12
NPT after 50 ns	1.13 ± 0.13	
NPT from 30 ns to 50 ns	1.19 ± 0.12	

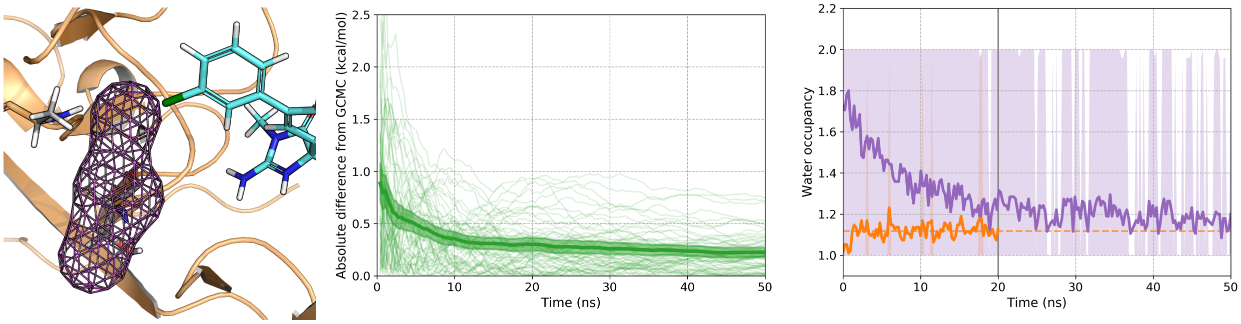

Figure 11: Evaluating the convergence of the NPT FEP predictions on the bace1 map. The starting structure contains 2 buried water molecules, one of which sterically overlaps with 28 out of the 36 ligands. Left: the starting structure of ligand 13k in which the position of the Cl atom (in green) overlaps with a buried water that is present in the starting structure (modeled on PDB entry 4DJW). The violet mesh envelops a volume in which the water occupancy was tracked. Middle: The absolute difference of all NPT $\Delta\Delta G$ predictions over 50 ns to the μVT predictions at 20 ns. The differences from each edge are shown as light green lines, the thick green line indicates the mean absolute difference, and the associated 95 % confidence region is also colored in green. On average, the NPT predictions tend towards the μVT predictions as time progresses. Right: the mean water occupancy of the violet volume as a function for time for the 20 ns μVT calculations (in orange) and for the 50 ns NPT calculations (in purple). The thick lines indicate the mean over all physical state simulations for the ligands that overlap with the starting buried water. The shaded regions represent the 68th percentile regions of the water occupancy samples – approximately equivalent to one standard deviation in a normal distribution. The orange dotted line indicates the mean water occupancy of the violet volume during the μVT simulations.

Conclusion

Despite the increasing success of free energy perturbation techniques to predict the relative binding free energies of ligands, a reliable protocol for transformations that displace buried water molecules has remained elusive. In this work, we have shown that sampling water with grand canonical Monte Carlo within FEP calculations is a robust solution to this issue and computationally efficient enough for large scale lead optimization projects.

To assess the impact of using GCMC in FEP calculations, we assembled a data set consisting of 8 proteins and 139 ligand transformations that displace or disrupt conserved, buried water molecules. To model prospective FEP applications in which practitioners are uncertain of the placement of water, the protein structures were prepared in the presence and absence of the important buried water molecules. GCMC, run only at the start or throughout the production simulation, significantly reduced the error of the predicted relative affinities with respect to the experiment. Without GCMC, the FEP predictions were extremely sensitive to the initial water placement (with an RMSD of 2.32 kcal/mol). In contrast, running GCMC throughout the entirety of the FEP simulations resulted in far more reliable predictions (with an RMSD of 0.43 kcal/mol).

FEP calculations with GCMC can be reliably applied in cases when one has no prior knowledge of the buried and conserved water molecules within binding sites. Any buried hydration sites that are unoccupied at the start of the simulation will be rapidly populated by GCMC (see Figure 2). The most accurate FEP predictions were obtained when the clashing, buried water molecules were *not* present in the starting protein structure, as these waters had the potential to distort the ligand binding mode during the energy minimization stages. As a result of this observation, the FEP+ worklow automatically removes water molecules that overlap with ligands before running simulations in the μVT ensemble.

GCMC sampling of water has expanded the domain of applicability of ligand FEP and, as a result, has highlighted other issues that affect accuracy. For instance, with scytalone dehydratase, we found that the predictions were very sensitive to the estimated pKas of the ligands. Although not investigated in this work, we anticipate GCMC to reveal even larger sensitivities to pKa in systems where there is strong coupling between hydration and proton affinity. Our results also suggest that the SPC water used in FEP+ could be enhanced to more accurately account for the desolvation free energy of buried water molecules, although we leave the systematic analysis of the accuracy of different water models for future research. Because GCMC greatly accelerates sampling of water, it is important to note that FEP calculations are likely to be more sensitive to the water model than before.

In spite of its reported promise for FEP calculations,[12–14] GCMC has unfairly garnered a reputation that it is conceptually difficult as well as slow and cumbersome to use. In the theoretical work that underpins our simulation methodology, we show how GCMC equilibrates and varies the system density in a similar way as a barostat would. Ultimately, this results in relative binding free energies that are approximately equivalent between the μVT and NPT ensembles for aqueous systems. Regarding GCMC's ease of use, our implementation exploits the parallelism afforded by GPUs, which results in a computational performance comparable to that of a barostat. In addition, GCMC sampling has been fully integrated in the FEP+ workflow.

As the benefits of using GCMC to accelerate water sampling can be achieved with no additional overhead to users, we recommend using GCMC whenever possible in protein-ligand FEP calculations.

Acknowledgement

The authors thank the FEP+ team at Schrödinger Inc. for their assistance with the underlying software infrastructure of this work, particularly Yujie Wu and Dong Zhou for updating the trajectory storage format to handle simulations with a variable number of molecules.

Supporting Information Available

The supporting information includes derivations of equations 1, 2, 3, and 4, a detailed discussion on the theoretical equivalence of free energy differences in μVT and NPT, a description of how the chemical potential was calibrated, the methods used to alchemically decouple the two buried water from HSP90 (the results of which are shown are in Figure 3), and additional FEP results from the water displacement data sets. Additionally, all of the inputs files for the expanded water displacement set are supplied, as well the output files for the μVT calculations when overlapping water is not included in the starting protein structure.

References

(1) Wang, L. et al. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(2) Pan, A. C.; Xu, H.; Palpant, T.; Shaw, D. E. *J. Chem. Theory Comput.* **2017**, *13*, 3372–3377.

(3) Kuhn, M.; Firth-Clark, S.; Tosco, P.; Mey, A. S. J. S.; Mackey, M.; Michel, J. *J. Chem. Inf. Model.* **2020**, *60*, 3120–3130.

(4) Denisov, Vladimir P.; Halle, Bertil; Peters, Jorg; Horlein, H. D. *Biochemistry* **1995**, *34*, 9046–9051.

(5) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.

(6) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. *J. Am. Chem. Soc.* **2007**, *129*, 2577–2587.

(7) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2008**, *130*, 2817–2831.

(8) Cao, S.; Konovalov, K. A.; Unarta, I. C.; Huang, X. *Adv. Theory Simul.* **2019**, *123*, 1900049.

(9) Irwin, B. W.; Vukovic, S.; Payne, M. C.; Huggins, D. J. *J. Phys. Chem. B* **2019**, *123*, 4220–4229.

(10) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2009**, *113*, 13337–13346.

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

(12) Bruce Macdonald, H. E.; Cave-Ayland, C.; Ross, G. A.; Essex, J. W. *J. Chem. Theory Comput.* **2018**, *14*, 6586–6597.

(13) Ross, G. A.; Bodnarchuk, M. S.; Essex, J. W. *J. Am. Chem. Soc.* **2015**, *137*, 14930–14943.

(14) Deng, Y.; Roux, B. *J. Chem. Phys.* **2013**, *115103*, 115103.

(15) Panagiotopoulos, A. Z. *Mol. Phys.* **1987**, *61*, 813–826.

(16) Ben-Shalom, I. Y.; Lin, C.; Kurtzman, T.; Walker, R. C.; Gilson, M. K. *J. Chem. Theory Comput.* **2019**, *15*, 2684–2691.

(17) Mezei, M. *Mol. Phys.* **1980**, *40*, 901–906.

(18) Woo, H.-j.; Dinner, A. R.; Roux, B. *J. Chem. Phys.* **2004**, *121*, 6392–6400.

(19) Ross, G. A.; Rustenburg, A. S.; Grinaway, P. B.; Fass, J.; Chodera, J. D. *J. Phys. Chem. B* **2018**, *122*, 5466–5486.

(20) Gill, S. C.; Lim, N. M.; Grinaway, P. B.; Rustenburg, A. S.; Fass, J.; Ross, G. A.; Chodera, J. D.; Mobley, D. L. *J. Phys. Chem. B* **2018**, *122*, 5579–5598.

(21) FEP+, Schrödinger Release 2020-2. Schrödinger, New York, NY.

(22) Gibbs, J. W. *Elementary Principles in Statistical Mechanics*; Charles Scribner's Sons, 1902.

(23) Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D. *J. Chem. Theory Comput.* **2019**, *15*, 1863–1874.

(24) Martyna, G. J.; Klein, M. L.; Tuckermana, M. *J. Chem. Phys.* **1992**, *97*, 2635–2643.

(25) Tuckerman, M; Berne, B. J.; Martyna, G. J. *J. Phys. Chem.* **1992**, *97*, 1990–2001.

(26) Martyna, G. J.; Tobias, D. J.; Klein, M. L. *J. Chem. Phys.* **1994**, *101*, 4177–4189.

(27) Wang, Lingle; Friesner, R. A. B. B. J. *J. Phys. Chem. B* **2011**, *115*, 9431–9438.

(28) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.

(29) Wang, L.; Lin, T.; Abel, R. Cycle closure estimation of relative binding affinities and errors, Patent US2014278295. 2014.

(30) Papadopoulou, A.; Becker, E. D.; Lupkowski, M.; Van Swol, F. *J. Chem. Phys.* **1993**, *98*, 4897–4908.

(31) 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. *Science* **1994**, *263*, 380–384.

(32) Bleasdale, J. E. et al. *Biochemistry* **2001**, *40*, 5642–5654.

(33) Woodhead, A. J. et al. *J. Med. Chem.* **2010**, *53*, 5956–5969.

(34) Vollmuth, F.; Geyer, M. *Angew. Chemie Int. Ed.* **2010**, *49*, 6768–6772.

(35) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J. Phys. Chem. B* **2003**, *107*, 9535–9551.

(36) Wahl, Joel; Smieško, M. *J. Chem. Inf. Model.* **2019**, *59*, 754–765.

(37) Kung, P. P.; Sinnema, P. J.; Richardson, P.; Hickey, M. J.; Gajiwala, K. S.; Wang, F.; Huang, B.; McClellan, G.; Wang, J.; Maegley, K.; Bergqvist, S.; Mehta, P. P.; Kania, R. *Bioorganic Med. Chem. Lett.* **2011**, *21*, 3557–3562.

(38) Fraley, M. E. et al. *Bioorganic Med. Chem. Lett.* **2006**, *16*, 6049–6053.

(39) Katz, B. A.; Sprengeler, P. A.; Luong, C.; Verner, E.; Elrod, K.; Kirtley, M.; Janc, J.; Spencer, J. R.; Breitenbucher, J. G.; Hui, H.; McGee, D.; Allen, D.; Martelli, A.; Mackman, R. L. *Chem. Biol.* **2001**, *8*, 1107–1121.

(40) Katz, B. A.; Elrod, K.; Verner, E.; Mackman, R. L.; Luong, C.; Shrader, W. D.; Sendzik, M.; Spencer, J. R.; Sprengeler, P. A.; Kolesnikov, A.; Tai, V. W.; Hui, H. C.; Breitenbucher, J. G.; Allen, D.; Janc, J. W. *J. Mol. Biol.* **2003**, *329*, 93–120.

(41) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. *J. Med. Chem.* **2006**, *49*, 6177–6196.

(42) Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. *J. Comput. Aided. Mol. Des.* **2010**, *24*, 591–604.

(43) Philipp, D. M.; Watson, M. A.; Yu, H. S.; Steinbrecher, T. B.; Bochevarov, A. D. *Int. J. Quantum Chem.* **2018**, *118*, 1–8.

(44) Baum, B.; Mohamed, M.; Zayed, M.; Gerlach, C.; Heine, A.; Hangauer, D.; Klebe, G. *J. Mol. Biol.* **2009**, *390*, 56–69.

(45) Chen, J. M.; Xu, S. L.; Wawrzak, Z.; Basarab, G. S.; Jordan, D. B. *Biochemistry* **1998**, *37*, 17735–17744.

(46) Crawford, T. D. et al. *J. Med. Chem.* **2016**, *59*, 5391–5402.

(47) Kaus, J. W.; Harder, E.; Lin, T.; Abel, R.; McCammon, J. A.; Wang, L. *J. Chem. Theory Comput.* **2015**, *11*, 2670–2679.

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

(49) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2009**, *131*, 15403–15411.

(50) Kaschula, C. H.; Egan, T. J.; Hunter, R.; Basilico, N.; Parapini, S.; Taramelli, D.; Pasini, E.; Monti, D. *J. Med. Chem.* **2002**, *45*, 3531–3539.

(51) De Oliveira, C.; Yu, H. S.; Chen, W.; Abel, R.; Wang, L. *J. Chem. Theory Comput.* **2019**, *15*, 424–435.

(52) Aldeghi, M.; Ross, G. A.; Bodkin, M. J.; Essex, J. W.; Knapp, S.; Biggin, P. C. *Commun. Chem.* **2018**, *1*.

Graphical TOC Entry