Enhancing Sampling of Water Rehydration upon Ligand Binding using variants of Grand Canonical Monte Carlo

Yunhui Ge,[†] Oliver J. Melling,[‡] Weiming Dong,[†] Jonathan W. Essex,[‡] and David L. Mobley[*,†,¶]

[†]*Department of Pharmaceutical Sciences, University of California, Irvine, CA 92697, USA*

[‡]*School of Chemistry, University of Southampton, Southampton, SO17 1BJ, United Kingdom*

[¶]*Department of Chemistry, University of California, Irvine, CA 92697, USA*

E-mail: dmobley@mobleylab.org

1 ABSTRACT

Water plays an important role in mediating protein-ligand interactions. Water rearrangement upon a ligand binding or modification can be very slow and beyond typical timescales used in molecular dynamics (MD) simulations. Thus, inadequate sampling of slow water motions in MD simulations often impairs the accuracy of the accuracy of ligand binding free energy calculations. Previous studies suggest grand canonical Monte Carlo (GCMC) outperforms normal MD simulations for water sampling, thus GCMC has been applied to help improve the accuracy of ligand binding free energy calculations. However, in prior work we observed protein and/or ligand motions impaired how well GCMC performs at water rehydration, suggesting more work is needed to improve this method to handle water sampling. In

this work, we applied GCMC in 21 protein-ligand systems to assess the performance of GCMC for rehydrating buried water sites. While our results show that GCMC can rapidly rehydrate all selected water sites for most systems, it fails in 5 systems. In most failed systems, we observe protein/ligand motions, which occur in the absence of water, combine to close water sites and block instantaneous GCMC water insertion moves. For these 5 failed systems, we both extended our GCMC simulations and tested a new technique named grand canonical nonequilibrium candidate Monte Carlo (GCNCMC). GCNCMC combines GCMC with the nonequilibrium candidate Monte Carlo (NCMC) sampling technique to improve the probability of a successful water insertion/deletion. Our results show that GCNCMC and extended GCMC can rehydrate all target water sites for three of the five problematic systems and GCNCMC is more efficient than GCMC in 2 out of the 3 systems. In one system, only GCNCMC can rehydrate all target water sites, while GCMC fails. Both GCNCMC and GCMC fail in one system. This work suggests this new GCNCMC method is promising for water rehydration especially when protein/ligand motions may block water insertion/removal.

2 INTRODUCTION

Water plays an important role in mediating protein-ligand interactions. When estimating water binding thermodynamics using molecular dynamics (MD) simulations, accuracy hinges partly on adequate sampling of water motions. However, water rearrangement in a binding site upon a ligand binding/modification can be beyond the typical timescale of MD simulations (e.g., ns or μs).[1,2] Thus, the accuracy of binding free energy calculations of ligands is often limited by insufficient water sampling.[3,4]

Grand canonical Monte Carlo (GCMC)[5–8] is a rigorous technique to accelerate water sampling in MD simulations. In the grand canonical ensemble, the chemical potential (μ) of water molecules, the volume and the temperature is constant while the number of particles

can fluctuate. The water molecules can therefore be considered to be in equilibrium with an ideal gas reservoir allowing them to be exchanged between the reservoir and the simulated system. In practice, the reservoir does not need to be simulated. Previous studies showed GCMC can be coupled with ligand binding free energy calculations.[9–13]

GCMC has shown success at enhancing water sampling during binding free energy simulations for ligand perturbations that disrupt buried water.[13] In our previous study, we compared different simulation techniques for water rehydration and we found GCMC in general was essentially the most robust method among those tested.[14] However, in our work we also observed protein/ligand motions that affected the performance of GCMC on water sampling. For example, when the water site is not occupied, a protein side chain may fill the site, blocking water insertion. We showed that restraining protein/ligand atoms to maintain the crystallographic pose was helpful to prevent the binding site from collapsing but the helpfulness of restraints was system dependent.[14] Alternatively, a longer simulation can be performed so the protein/ligand conformation changes and no longer blocks water insertions attempted by GCMC moves. But the required timescale for such conformational change is normally not known in advance so such simulations can be very computationally demanding. Thus, we are interested in alternative techniques that are able to handle protein/ligand motions during water insertion/removal.

In previous work,[15] we explored a water hopping method to accelerate water sampling in MD simulations. It is a nonequilibrium candidate Monte Carlo (NCMC) based technique that gradually turns on/off a water molecule and hops it between bulk solvent and a binding site. The advantage of the water hopping method over GCMC is that it allows the environment (e.g., protein, ligand) relax once a water molecule is inserted (or removed) in to (or out of) the binding site. However, relative to GCMC simulations, each insertion or deletion move is much slower so it may take longer for water hopping to reach convergence, except when GCMC acceptance rates are very low or when it does not sample the relevant motions.

In our previous work,[14] we compared the performance of water hopping, GCMC and

normal MD on water rehydration. We found both water hopping and GCMC outperformed normal MD but GCMC was more efficient than water hopping in most studied cases. In the water hopping method, a spherical region is defined that must include both the water binding region and some bulk solvent to provide waters for translation. This leads to a larger volume of sampling in water hopping compared to GCMC. Besides, since bulk solvent is involved, some moves are accepted for transferring waters between regions of bulk instead of between bulk and a binding site. Since only one move is proposed at a time, the effort is thus wasted for such an accepted move, lowering the efficiency of this method compared to GCMC. Even though previous work[14,16] suggests water hopping is less efficient than GCMC, we believe inserting/removing water molecules in an NCMC fashion could be useful when protein/ligand blocks instantaneous moves for water insertions like GCMC does. Moreover, if understanding protein, ligand or other water motions for inserting/removing a water into/from the site is important, instantaneous attempts like GCMC moves do not serve well in such cases. So we are interested in exploring alternatives for water sampling especially in cases where GCMC fails to rapidly insert/remove water due to protein/ligand motions.

Recently, Melling et al. developed a new method, grand canonical nonequilibrium candidate Monte Carlo (GCNCMC).[16] This method gradually turns on/off a water in an NCMC fashion and the water is exchanged between the reservoir and the region of interest (e.g., binding site). Thus, GCNCMC takes advantage of GCMC by avoiding sampling a large volume as water hopping does and the advantage of NCMC by allowing the environment to relax during water insertions/removals. We are interested in testing this new method in cases where protein/ligand motions may affect water insertion/removal and comparing this method to GCMC.

In this work, we focus on buried waters in protein-ligand systems. Based on the crystal structures, these waters mediate appear to mediate protein-ligand interactions. We removed these target water molecules prior to our simulations and tested if these hydration sites can be rehydrated in our simulations. We first simulated 21 systems using GCMC moves during

both the equilibration and production phases with two independent trials (more details in Section 3.2). Then we identified systems in which, in at least one simulation trial, some of the target water sites were not successfully rehydrated. We consider these systems to be challenging cases for GCMC. Then we performed longer GCMC and GCNCMC production runs for these challenging systems and assessed the performance of each simulation technique for water rehydration. We will discuss lessons we learned from these simulations in this paper.

3 METHODS

3.1 Selected targets.

Figure 1 shows examples of the binding sites of studied systems, along with crystallographic water molecules and the relevant Protein Data Bank (PDB) IDs. The complete list of selected systems with their PDB IDs is deposited in the SI and is also available at `https://github.com/MobleyLab/GCNCMC_GCMC`. We chose from studies which either focused on using enhanced sampling for water motions to improve the accuracy of binding free energy calculations[13,17] or focused on calculating binding free energies of buried water molecules.[18] The selected protein-ligand complexes include several proteins: Protein Tyrosine Phosphatase 1B (PTP1B), Heat Shock Protein 90 (HSP90), Bruton's Tyrosine Kinase (BTK), transcription initiation factor TFIID subunit 2 (TAF1(2)), thrombin, trypsin, HIV-1 protease, and Factor Xa (FXa). We aim to include considerable diversity and cover a broad range of systems that differ in the binding site position or the occupancy of water sites between congeneric ligands. We did not aim to recover all ordered water molecules for each system. Instead, we only focused on selected buried water sites that mediate the protein-ligand interaction.

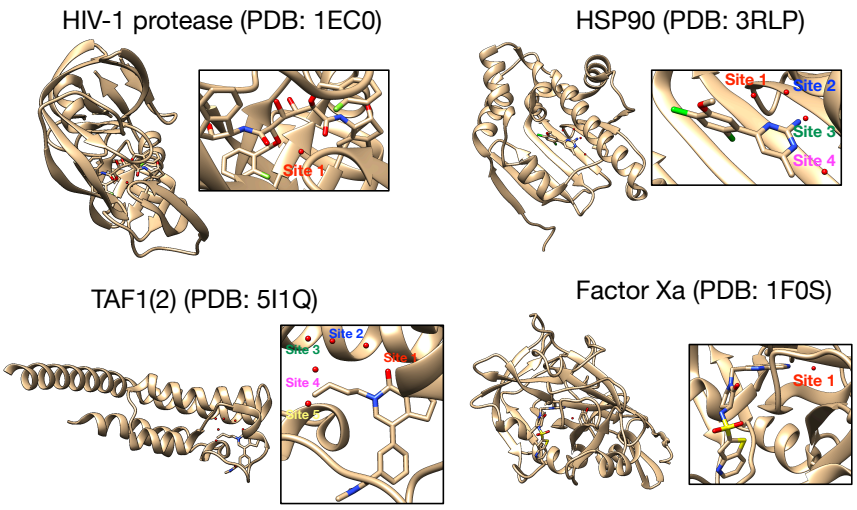

Figure 1: Examples of protein-ligand systems studied in this work and their hydration sites (red spheres) and their PDB IDs. The full set of systems studied was 21 protein-ligand complexes spanning 10 different proteins. The full set of complexes is available at `https://github.com/MobleyLab/GCNCMC_GCMC`.

3.1.1 Electron Density Calculations.

To compare modeling results with experiment, we felt it best to compare directly with experimental electron density maps rather than with the discrete waters deposited in crystal structures, since these waters are the result of a refinement process which aims in part to reproduce the electron density. Thus, we calculated the electron density in the same way as described in previous work.[14] We used a python script (`xtraj.py`, distributed in the LUNUS open source software) for processing, analysis, and modeling of diffuse scattering[19] (`https://github.com/lanl/lunus`). We used `xtraj.py` along with functions in the Computational Crystallography Toolbox (CCTBX)[20] and the MDTraj library for MD trajectory analysis[21] to compute the electron density from simulations. We visualized both calculated and experimental electron density maps using Coot molecular graphics[22,23] (v0.9.4). To be consistent with our previous study,[14] we used a contour level of 3 σ for calculated water electron density maps and 1.5 σ for experimental protein/water maps across all systems. We then visually examined overlap between the calculated and experimental electron density in Coot to determine whether the target sites were successfully recovered in our simulations. In

the rest of this paper, we use the averaged electron density calculated from each simulation block to determine if the hydration site is occupied in the simulation block (more details in Section 3.2).

3.2 GCMC/GCNCMC Simulations.

The *grand* package[24] was used to perform GCMC and GCNCMC moves with MD sampling using the OpenMM simulation engine (version 7.4.2).[25] The simulation set-up was very similar to our previous work using *grand* for GCMC simulations.[14] To set up GCMC or GCNCMC simulations, a region of interest needs to be defined first. To do that, we selected two atoms (e.g., Cα) on the receptor so that the midpoint between them was used as the center of a spherical region for enhanced water sampling (see `https://github.com/MobleyLab/GCNCMC_GCMC` for a detailed list of selected atoms). All target hydration sites were within this defined spherical region and the radius of this region varied between systems and was dependent on the binding site size of the protein target.

The AMBER ff14SB force field[26] was used for protein parameterization in conjunction with TIP3P water model.[27] The ligand was parameterized using Open Force Field version 1.2.1 (codenamed "Parsley").[28,29] The Langevin BAOAB integrator[30] was used with a time step of 2 fs, and a friction constant of 1 ps^{-1}. Long-range electrostatics were calculated using Particle Mesh Ewald (PME)[31,32] with nonbonded cutoffs of 10 Å. Each system was simulated at the experimental temperature listed on the PDB website (`https://www.rcsb.org`). We used pdbfixer 1.6 (`https://github.com/openmm/pdbfixer`) to add the missing heavy atoms to the receptor. The PROPKA algorithm[33,34] on PDB2PQR web server[35] was used to protonate the receptors' residues at experimental pH values. The pK_a values of ligands were calculated using Chemicalize (ChemAxon, `https://www.chemaxon.com`) and then were used to determine protonation states of ligands based on the experimental pH conditions for the crystal structure.

There are two additional key parameters used in *grand* simulations that affect the accep-

tance rate of GCMC/GCNCMC moves:[24] the excess chemical potential (μ') of bulk water and the standard state volume of water (V°). As suggested by prior work,[14,24] the former was calculated as the hydration free energy of water, and the latter as the average volume per water molecule. The details of these calculations can be found in prior work[24] and the calculated results at different temperatures are listed in Table S1.

For GCNCMC simulations, we need to determine the switching time for the nonequilibrium switching process to turn on/off the target water interactions. This parameter can affect the acceptance rate and thus is important to the efficiency of this method. If the switching time is too short, the acceptance rate can be low. However, if the switching time is too long, the computational cost of this method also increases. Ideally, we can find the point at which further doubling of the switching time no longer results in doubled acceptance rate.

However, this ideal switching time is system dependent and is not known in advance. So we decided to run test simulations with multiple switching times and find the ideal switching time based on the acceptance rate. We first tested in bulk water simulations with a series of switching times (0.2, 0.4, 0.8, 1.6, 3.2, ..., 102.4 ps). We doubled the switching time and checked if the acceptance rate was also doubled. When the acceptance rate no longer at least doubled with the doubled switching time, we took the last tested switching time before this one to be ideal. We used the same simulation set-up and input files as used in this work[16] (available at `https://github.com/essex-lab/gcncmc-paper`). We then tested a series of switching times (3.2, 6.4, 12.8, 25.6 ps) on selected protein-ligand systems (one system for each protein target). The simulation set-up was described below. All of these test simulations were performed with 2 replicates. We will discuss our results from these tests in Section 4.

In our previous work,[14] we used the number of force evaluations to compare the efficiency between different techniques. In this work we keep the total number of force evaluations in the production phase the same for GCMC and GCNCMC simulations which provides us

an opportunity to compare the efficiency between these two methods. The total simulation timescale (in nanoseconds) differs between these two methods (see below) because of the extra NCMC switching process in GCNCMC simulations which is not needed in GCMC simulations.

As mentioned, we are interested in testing if GCNCMC simulations can serve better than GCMC in systems where protein/ligand motions affect GCMC performance. To identify such systems, we applied GCMC moves during the equilibration phase and performed a short production run with GCMC/MD simulations (details below). For those systems where all target water sites were rehydrated after these simulations, we did not run further GCMC or GCNCMC simulations, as GCMC sampling was apparently adequate. However, when any of the target sites were not successfully rehydrated in at least one simulation trial (of two trials in total) we performed both GCNCMC and longer GCMC simulations to check if these techniques can rehydrate all the target sites on longer simulation timescales.

We also checked GCNCMC success during production simulations by performing two GCNCMC simulation trials on systems in which GCMC simulations successfully rehydrate all target sites to verify that GCNCMC simulations could also achieve success.

For simulation set up, we first removed all ordered water molecules prior to simulations. Then the systems were minimized until forces were below a tolerance of 10 kJ/mol using the L-BFGS optimization algorithm[36] implemented in OpenMM. Then we performed the equilibration process as following: We first started the simulations by performing 10000 GCMC moves. Then we did 100 iterations in which each iteration consists of 5 MD steps followed by 1000 GCMC moves. Then we performed 500 ps NPT simulation to equilibrate the system volume. In the final stage, we performed 100k GCMC moves over 500 ps MD simulations. After equilibration, we performed production GCMC run which involved 2.5 ns of GCMC/MD (50 GCMC moves carried out every 1 ps of MD).

For systems where any target sites were not successfully rehydrated in at least one simulation trial, we performed GCNCMC and GCMC simulations with an extended length (in

the rest of this manuscript, we refer to them as "longer GCMC simulations"). We performed two replicates for both GCNCMC and longer GCMC simulations and each replicate started from the same structure for GCMC and GCNCMC. The two starting structures though were different since they were equilibrated independently (see above). Due to the walltime limit of our cluster, we divided our simulation into multiple blocks. Each simulation block was continued from the last point of the previous block. For GCMC simulations, the production run involved 2.5 ns of GCMC/MD (50 GCMC moves carried out every 1 ps of MD) for each single simulation block (1.4 million force evaluations) and the simulation was extended to 25 ns (10 blocks, $\sim$ 14 million force evaluations) in total. These simulation block lengths were selected in consideration of our cluster's wallclock time limit. For GCNCMC simulations, the production run involved 0.5 ns of GCNCMC/MD (1 GCNCMC move carried out every 1 ps of MD) for each single simulation block and we used an NCMC switching time of 8 ps (more details in Section 4) for each GCNCMC move (2.25 million force evaluations). We extended GCNCMC simulations to 3 ns (6 blocks, $\sim$ 14 million force evaluations) eventually. The total number of force evaluations was the same for both GCMC and GCNCMC simulations.

4 RESULTS

4.1 NCMC switching time determination.

We first needed to determine the NCMC switching time to be used in GCNCMC simulations. This parameter affects both the acceptance rate and efficiency of this method. An ideal switching time is beyond which the increased switching time is no longer compensated for by the increase in acceptance rate. This value is not known in advance and requires tests, ideally for each simulated system. In practice, one can run GCNCMC simulations with different switching times and identify this parameter. However, this can be very expensive and not feasible when simulating a large number of systems as we did in this work. But we still wanted to test how this protocol works in practice for different protein-ligand systems

because it has not been well explored yet.

We started with a simple system and did this test in bulk water simulations and used a series of switching time: 0.2, 0.4, 0.8, 1.6, ..., 102.4 ps, where each switching time doubles the last one. We checked the ratio between acceptance rates for adjacent switching times and found beyond a switching time of 3.2 ps, a doubled switching time did not return a doubled acceptance rate (Figure S1). Thus, we identified 3.2 ps as an ideal NCMC switching time for GCNCMC simulations of bulk water.

We were not able to perform this test for all 21 protein-ligand system because the total computational cost would have been prohibitive. Instead, we did the test on selected protein-ligand systems. We picked one system for each protein target and used four different switching times (3.2, 6.4, 12.8, 25.6 ps). We started with the switching time of 3.2 ps based on our results in bulk water simulations. Since water insertions/deletions might be more challenging in protein-ligand systems than bulk water so a longer switching time could be necessary. So we doubled the switching time to get another three longer switching times (6.4, 12.8, 25.6 ps), which we then tested.

We performed our tests with two replicates for each system. We checked the averaged acceptance rates for each tested switching time (Figure S2). Unlike our results from bulk water simulations, we obtained very noisy acceptance rates from replicates for each studied system. For some systems (e.g., HSP90, PTP1B) the acceptance rate was zero or close to zero for all tested switching times. For other systems, we did not observe a clear trend to help us identify an ideal switching time for each system as we did in bulk water simulations (Figure S1). The acceptance rate varied significantly between replicates and systems. This is due to the extra complexity in protein-ligand simulations compared to bulk water simulations. In bulk solvent simulations, water can be inserted/deleted anywhere in the box with a reasonable chance of acceptance, given enough relaxation. However, water can only bind to a few locations in the protein binding site and the chance of a water insertion move being accepted is sensitive to the protein/ligand motions. Because of that, the acceptance rate

could vary between simulation trials and is highly system dependent. Of course, with very long simulations, acceptance rates would be converge, but the computational cost of doing so would be prohibitively high.

We decided to use a uniform switching time (8 ps) for all GCNCMC simulations of protein-ligand systems. This switching time falls into the time range (7-9 ps) recommended by the author of the GCNCMC method[16] for protein-ligand systems. Given the literature precedent, this seemed to be a reasonable choice, though future work could generate system-specific switching times if more computational resources are available. Our results show that GCNCMC simulations using the selected switching time successfully rehydrated all target water sites in most systems we studied, suggesting the switching time we used is reasonable.

4.1.1 Case studies

We simulated 21 systems in total and examined the rehydration of target water sites by comparing the calculated electron density from GCMC simulations to the experimental density. Among all 21 studied systems, we found our initial GCMC simulations successfully rehydrated all target sites in 16 systems. As we described in Section 3.2, we also performed short GCNCMC simulations on these 16 systems and confirmed all target sites were successfully rehydrated with GCNCMC as well. The results (electron density maps) for both GCMC and GCNCMC simulations are available at `https://github.com/MobleyLab/GCNCMC_GCMC`.

We observed 5 systems (PDB: 1F0S, 1EC0, 1EC1, 5I1Q, 3RLP) where at least one GCMC simulation trial failed to rehydrate at least one of the target water sites in the initial short GCMC simulations (2.5 ns, 1.4 million force evaluations). We were interested in testing whether GCNCMC and longer GCMC simulations could rehydrate all target sites in these systems. In these simulations, both GCMC and GCNCMC simulations for each trial used the same initial structure. In the following sections, we will discuss lessons we learned from GCNCMC and GCMC simulations of the 5 challenging systems, which are our main focus here.

HIV1-protease (PDB: 1EC1) In this HIV-1 protease system (PDB: 1EC1), there is only one target water site. This site was not rehydrated in either trial of the initial GCMC simulations (see Section 3.2). In longer GCMC simulations (25 ns, 14 million force evaluations, 1250000 GCMC moves in total), we still did not see any successful water insertions to this target site either simulation trials.

In one GCNCMC replicate, the water site was refilled after the first simulation block (0.5 ns, 2.25 million force evaluations, 500 GCNCMC moves in total) (Figure 2). The calculated electron density map overlaps well with the experimental map (simulation block 2-6 in Figure 2). In another GCNCMC simulation trial, though, we did not observe any successful water insertions.

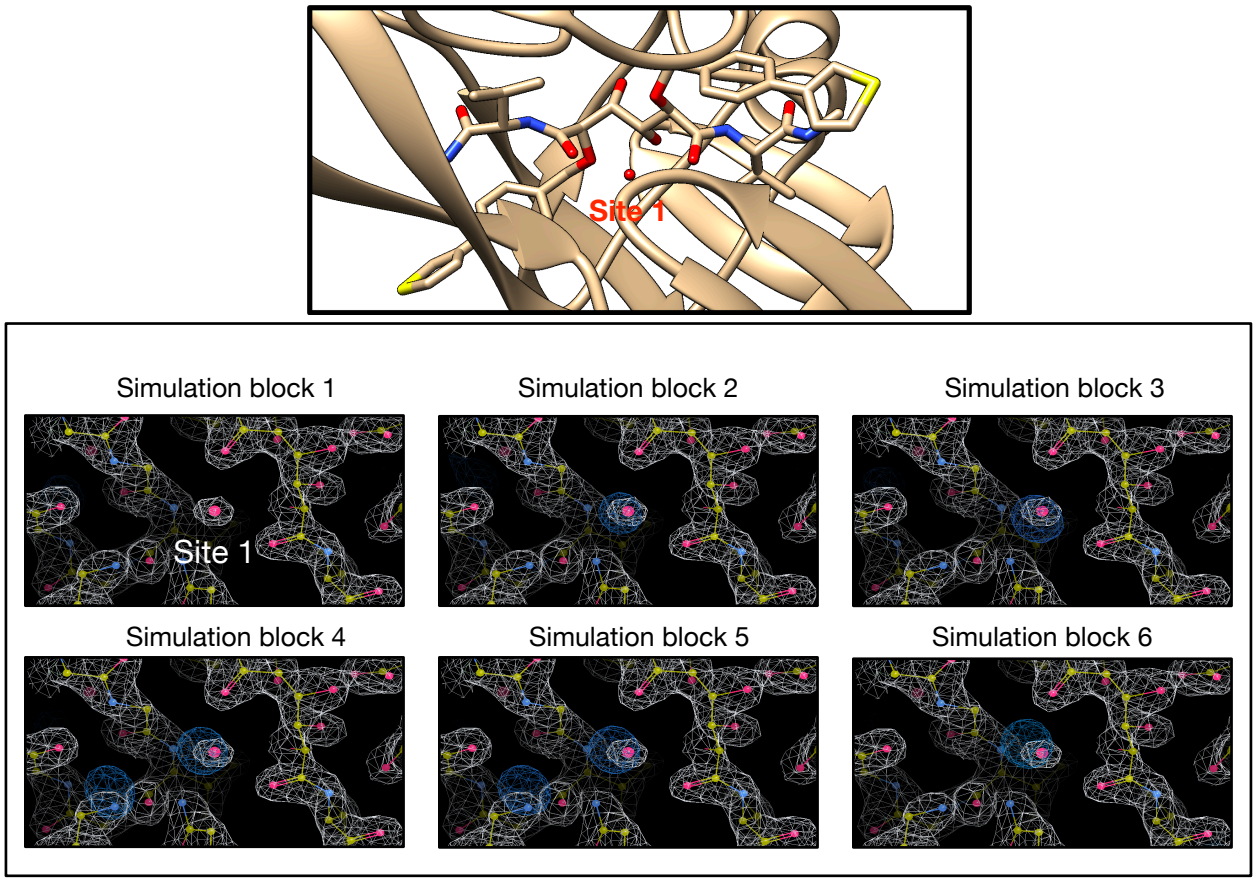

Figure 2: Our GCNCMC simulation rehydrated the target water site in a HIV-1 protease system (PDB: 1EC1) whereas our GCMC simulations failed to do so. The electron density maps calculated from simulations are shown in blue and the experimental maps are shown in white.

Based on the crystal structure, this target water molecule is stabilized by multiple hydrogen bonds with protein/ligand atoms (Figure S3). ILE50 is located on a loop near the binding site and the amide hydrogen atoms on ILE50 and two oxygen atoms on the ligand are important hydrogen bond donors and acceptors for this water molecule. Our hypothesis is failure to form these hydrogen bonds, such as due to binding site rearrangements that leave the environment unfavorable for forming these bounds, affects the opportunity to insert the water to this site.

To look for this effect, we took a snapshot from GCNCMC simulations where this site was successfully rehydrated. When we compare it to the crystal structure we found the loop we mentioned above is in a similar position as the crystallographic binding mode (Figure S4). However, in a snapshot taken from GCMC simulations where water insertions all failed, the loop is closer to the binding site than it is in the crystallographic binding mode (Figure S4), suggesting the binding site partially collapses. After this collapse, water insertion is more difficult because of the limited space in the binding site. Moreover, the amide hydrogen is not in the position to form the key hydrogen bond with the water to help stabilize this water after being inserted. Then the opportunity to have an insertion move accepted ought to be low.

It seems likely that binding site collapse will be a common outcome in some binding sites, since maintaining empty cavities near binding sites can often be highly unfavorable. Here removing the ordered water from the binding site prior to simulations creates a cavity, and apparently this is unfavorable, so the protein responds by moving the loop.

We calculated the distance between ILE50(N) and ALA28(Cα) (Figure S5) over the course of the GCNCMC simulations. We wanted to confirm that before water can be successfully inserted, it is necessary to have this loop move back to its original position as in the crystal structure. We used ALA28 as an anchor to compute this distance because this ALA28 was stable in simulations and did not change its conformation when we compared these snapshots. This protein has two identical chains and is a functional dimer so we

monitored this distance change for ILE50(N) and ALA28(Cα) on both chains. A distance decrease indicates the loop moves closer to the binding site. When the loop is closer, it is unlikely for water insertion moves to be accepted.

In Figure 3, we can see the distance between ILE50 and ALA28 in one chain (Figure 3A) is not changing much and centers around the reference distance in the crystal structure (orange line). However, in the other chain (Figure 3B), the distance is below the reference distance indicating this loop is closer to the binding site earlier in the simulation compared to the crystallographic pose. But later (after 0.5 ns, during the second simulation block), the distance increases sharply and gets above the reference distance followed by a successful water insertion in the NCMC simulation. The reference distance is 0.975 nm and the distance in the frame before and after the successful water insertion are 0.937 and 1.015 nm, respectively.

In the NCMC portion of this successful water insertion (Figure 3C), the distance gradually increases as the water is gradually inserted. This represents a motion which the loop is first close to the binding site and filling the cavity. Then during the NCMC portion the water is slowly inserted to this site and the loop is pushed out of the site, making room for this water. Once this loop is back to its original position, the inserted water is further stabilized by the hydrogen bond formed between the amide hydrogen on ILE50 and the inserted water. We do not observe such motions in GCMC simulations as GCMC does not allow the protein to relax upon water insertion.

In another GCNCMC simulation trial and both GCMC simulation trials, this loop is close to the binding site. No water insertion attempts succeed in pushing the loop back to the original position. Not surprisingly, all of these simulations failed to insert a water molecule to the target site. We believe the alternate loop conformation is critical for successful water insertions. Thus, to improve the opportunity to rehydrate this site, it is important to apply methods that can handle this protein backbone motion. Based on our results, GCNCMC outperforms GCMC and is an ideal method in this case.

Another way to handle this protein motion is to apply position restraints to maintain

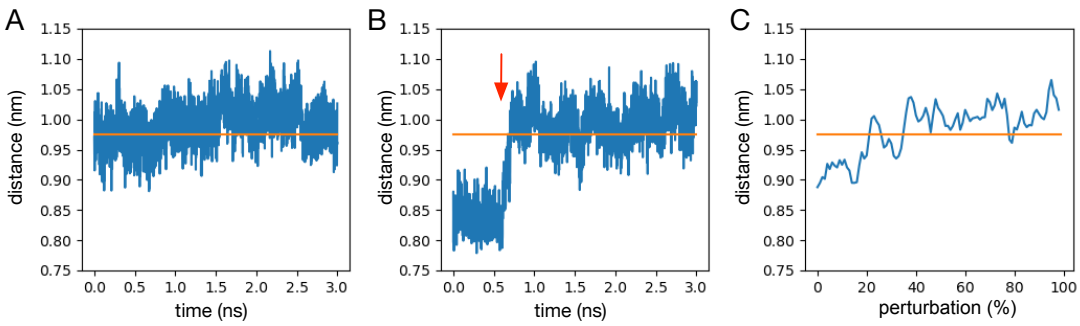

Figure 3: A key loop in the HIV-1 protease system (PDB: 1EC1) must move back to its original position, as in the crystal structure, before a water can be successfully inserted. (A)-(B) The distance between ILE50(N) and ALA28(Cα) on chain A and B during GCNCMC simulations. (C) The distance change between ILE50(N) and ALA28(Cα) on chain B during the NCMC simulation in which the water was slowly inserted. The orange line shows the distance between the two atoms in the crystal structure. The red arrow in (B) shows where the successful water insertion happened.

the protein to its crystallographic binding mode even when the target site is not occupied, allowing any protein motions to be treated in a separate stage of the calculations as needed. We have used such restraints in GCMC simulations in our previous work[14] and found it helped water insertion in some studied systems. We applied position restraints of 10 kcal mol^{-1} Å^{-2} on the heavy atoms of the protein and ligand to maintain the crystallographic pose in GCMC simulations. The simulation protocol is the same as what we described in Section 3.2. We performed two simulation trials and each has ten simulation blocks (25 ns in total for each trial). Figure S6 shows that GCMC successfully rehydrated the target site when used with position restraints on the protein and ligand heavy atoms. This confirms that water inserts more effectively when the protein binding site remains open, as in the crystal structure.

We did not perform GCNCMC simulations with restraints on protein and/or ligand atoms because the point of inserting water molecules in an NCMC fashion is that the environment is allowed to relax upon the attempt. Restraining the protein would reduce the amount of relaxation which could occur during NCMC moves, potentially making instantaneous moves in GCMC more efficient than GCNCMC. Because of this, in this work, we only

tested applying restraints on the protein and ligand atoms in GCMC simulations but not in GCNCMC simulations.

HIV-1 protease (PDB: 1EC0) For another HIV-1 protease system (PDB: 1EC0), GCMC and GCNCMC simulations both failed to rehydrate the target water site (Figure 1). Based on what we learned from another HIV-1 protease system (discussed above), we speculate the loop motion is the reason for these failures. To confirm that, we calculated the distance between ILE50(N) and ALA28(Cα) in both GCMC and GCNCMC simulations to check if the binding site was partially collapsed when the water site was not occupied.

The calculated distance (Figure S7) shows that when the target water site is not occupied, the loop is closer to the binding site compared to the crystallographic pose in both GCNCMC and GCMC simulations (orange line in Figure S7). This is similar to what we observed in the HIV-1 protease system (PDB: 1EC1) discussed above.

For a period of simulation time in Trial 1 of GCNCMC simulations (Figure S7A), the calculated distance is slightly above the reference distance measured in the crystal structure (orange line). However, we did not observe any successful water insertions in this simulation segment.

We also tested applying position restraints on heavy atoms of the protein and ligand in GCMC simulations similar to what we did in another HIV-1 protease system (PDB: 1EC1). The calculated electron densities overlap well with the experimental densities (Figure S8) confirming the successful rehydration of this target site in simulations. In simulation block 3 of Trial 1 (Figure S8A), the target water was removed and then re-inserted in simulation block 4, indicating that simulations allowed reversible water insertion/removal for this target site.

HSP90 (PDB: 3RLP) In this HSP90 system, there are 4 target water sites (Figure 1). In our previous work,[14] we also studied this system and found Site 1 was more challenging than other sites for rehydration. Here, all four target sites were successfully rehydrated in one trial

of the initial GCMC simulations (Section 3.2). In another trial, Site 1 was not rehydrated. So we performed GCNCMC and longer GCMC simulations using the equilibrated structure from Trial 2 as the starting structure. In GCNCMC simulations, Site 1 was only recovered in one simulation block but was unoccupied in other simulation blocks (Figure 4B).

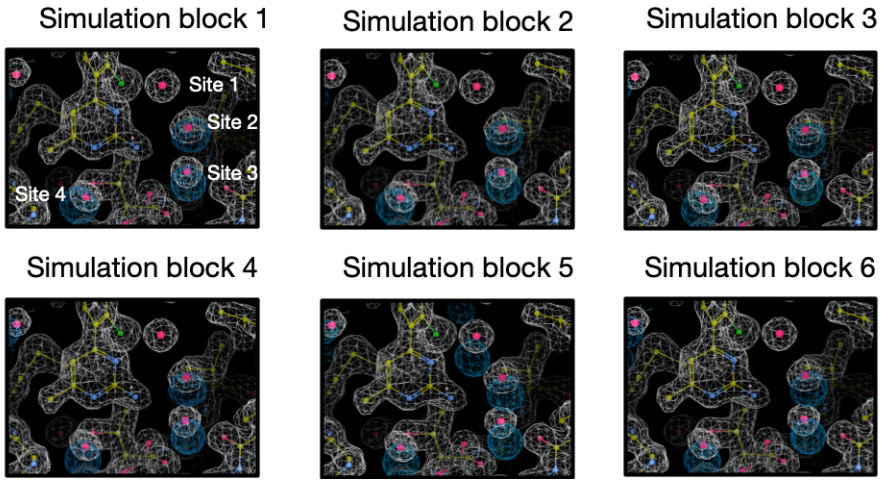

Figure 4: Each of the target water sites in a HSP90 system (PDB: 3RLP) was recovered in at least one simulation block of GCNCMC simulations. The electron density maps calculated from simulations are shown in blue and the experimental maps are shown in white.

We removed all ordered waters in the starting structure prior to simulations. This removal of water from Site 1 creates a cavity. Consequently, the ASN42 side chain changes its orientation to fill the cavity during simulations (Figure 5B). Even though Site 1 is not completely occupied by this side chain, it becomes more challenging to insert a water molecule in to this site.

In GCNCMC simulations, after the water is successfully inserted in Site 1, the side chain of ASN42 rotates closer to the crystallographic pose orientation (Figure 5A, around 2 ns). But the ASN42 side chain is still not in the same orientation as the crystallographic binding mode (Figure S9). So the inserted water in Site 1 is close to but not exactly in the same position as when it is in Site 1 in the crystal structure. The resulting water network in the binding site (Site 1-4) is thus slightly different from that in the crystal structure (Figure 4). As we discussed in the previous work,[14] those discrete waters deposited in the crystal

structure are based on interpretations of the underlying data. These interpretations may introduce the potential for human bias and/or errors.[37–39] Additionally, water occupancies are typically deposited as 100% to avoid overfitting during the refinement of the crystal structure. Thus, we still considered our simulations successfully recovered these target sites in this HSP90 system since a reasonable overlap was obtained between the simulated and crystallographic water pattern.

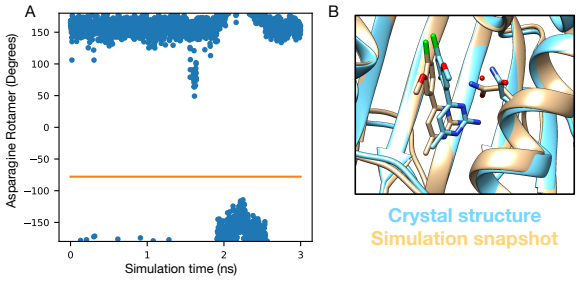

Figure 5: ASN42 rotamer sampling and representative snapshots for GCNCMC simulations of a HSP90 system (PDB: 3RLP). When Site 1 is not occupied, the ASN42 side chain is in a different orientation from the crystal structure. (A) Calculated $\chi 1$ angle of ASN42 as a function of GCNCMC simulation time (ns) for GCNCMC simulations. The orange line shows the angle calculated from the crystal structure. (B) The crystallographic binding mode (blue) and a simulated snapshot (tan) from GCNCMC simulations in which inserting a water to Site 1 (red blob) was challenging.

Our GCMC simulation results show that all target sites are rehydrated within the first simulation block (Figure 6). Similar to what we observed in GCNCMC simulations, ASN42 side chain is in a different orientation from the crystal structure (Figure S10A) and partially occupies Site 1 when this site is empty. After water is inserted in this site, the side chain of ASN42 moves back to the same position as in the crystal structure (Figure S10B).

HSP90 (PDB: 3RLQ) There are three target sites in this HSP90 system (PDB: 3RLQ, Figure S11). We did not observe any protein/ligand motions that affected water insertion. But in both GCMC and GCNCMC simulations, only Site 2 and 3 were successfully rehydrated, while Site 1 was not recovered (Figure S12 and S13).

We studied this system in our previous work[14] in which we tested several different en-

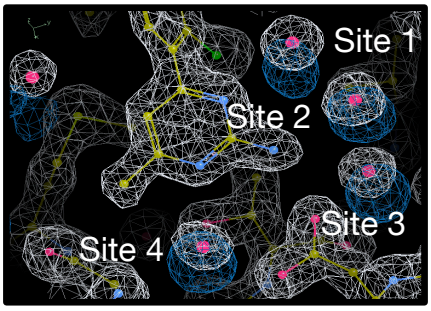

Figure 6: All target water sites in a HSP90 system (PDB: 3RLP) were successfully recovered in the first simulation block of GCMC simulation. The electron density maps calculated from simulations are shown in blue and the experimental maps are shown in white.

hanced sampling techniques (but did not compute water binding free energies). In this work we used the same force field as that used in our previous work. None of our simulations could rehydrate Site 1, neither here nor in the prior study. We also ran MD simulations begun from initial structures which retained all ordered water molecules. In these simulations, the water in Site 1 escaped quickly, suggesting this site was unfavorable.

As noted previously,[14] the experimental electron density for water in Site 1 is weaker than that in Site 2 and 3. Experimental refinement of the second (otherwise equivalent) protein chain in the crystal structure deposited no waters in Site 1, further supporting the idea that this site might not be well occupied. Indeed, had we chosen to compare our simulations with the second chain in the crystal structure, we would have concluded that Site 1 ought not to be occupied.

In the present work, our focus is not on computing water occupancies but rather on assessing effectiveness of water insertion. Still, in terms of occupancy, our simulations (and those from our previous work[14]) return consistent results, further confirming Site 1 is not favorable with the present force field. Further work is needed to determine whether the apparent difference from the crystal structure is due to limitations of the force field used for simulations, or problems with refinement or interpretation of the experimental crystallographic data.

FXa (PDB: 1F0S) There is only one target site in this FXa system (Figure 1). We did not observe any sampling issues in our simulations that affected water insertion. Both GCNCMC and GCMC simulations can rehydrate this target site but the occupancy of this site is generally low in our simulations. As shown in Figure S14 and S15, we only observed calculated electron density in the target site in a few simulation blocks (GCMC: block 3 and 6 in Trial 1, block 7 in Trial 2; GCNCMC: block 1-3 and 6 in Trial 1). These results seem to indicate that our chosen energy model predicts a low (but perhaps not negligible) water occupancy for this site.

A previous study[18] reported a positive binding free energy of this water molecule in this system, suggesting a lower than 50% occupancy of this site. The experimental electron density is also weak for this water site compared to other deposited water molecules in the crystal structure, suggesting the probability of observing a water molecule in this site ought to be low.

TAF1 (PDB: 5I1Q) There are five target sites in this TAF1 system (Figure 1). GCNCMC simulations successfully rehydrated all five target sites in both trials (Figure S16) whereas GCMC only worked in one trial (Figure 7).

Our system preparation created a cavity in the target site by removing ordered water molecules prior to simulations. In GCMC simulations (Trial 2) the ligand moved in to and occupied the space of Site 1 and 2, blocking water insertions in to the two sites (Figure 8). The target sites (Site 1 and 2) were not rehydrated before the ligand filled the cavity, and ligand rearrangement then made it very difficult to insert water molecules to these sites. We did not observe this ligand motion in GCNCMC simulations because all water sites were successfully rehydrated quickly in simulations, prior to the ligand moving to fill the cavities.

We were interested in testing whether GCMC and GCNCMC could insert water in to Site 1 and 2 even in cases where the ligand might initially block those sites (Figure 8). So we performed both GCMC and GCNCMC simulations with the sites initially blocked,

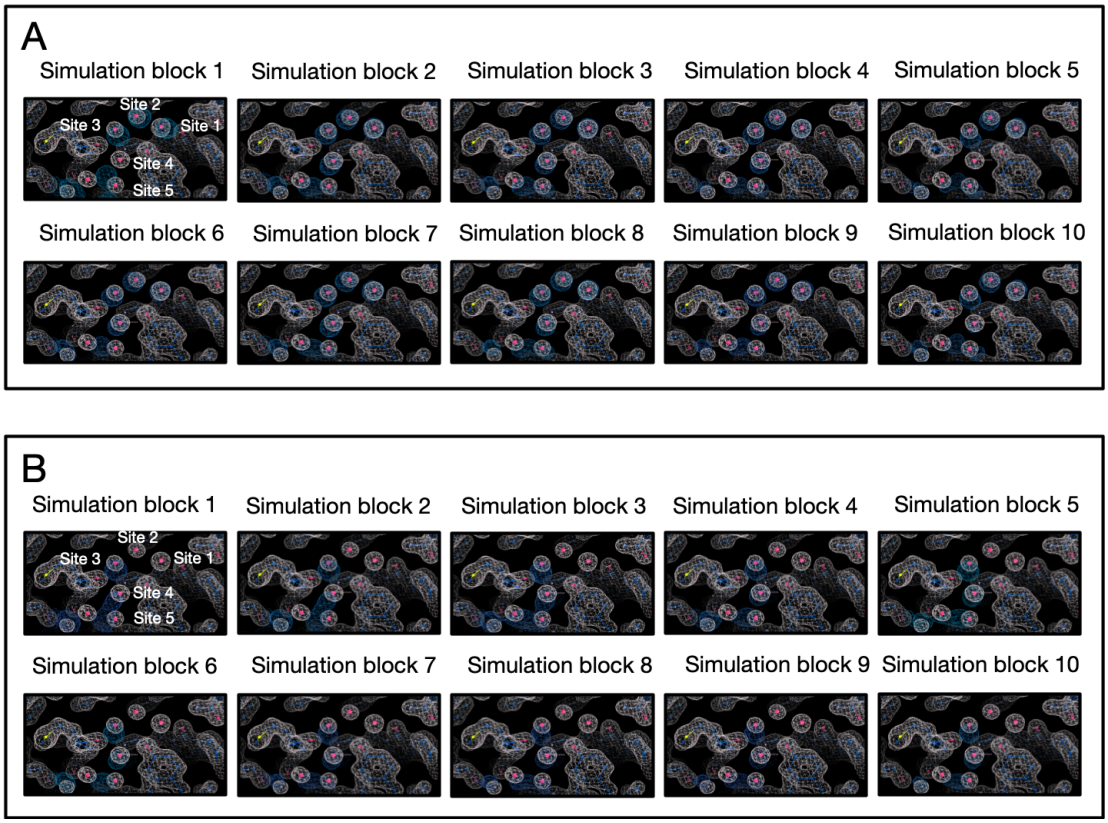

Figure 7: GCMC simulations successfully rehydrated all five target sites in a TAF1 system (PDB: 5I1Q) in Trial 1 (A) but failed to do so for Site 1 and 2 in Trial 2 (B). The electron density maps calculated from simulations are shown in blue and the experimental maps are shown in white.

using the starting structure shown in Figure 8 (yellow). We performed two trials for each simulation technique, with each trial including 14 simulation blocks in total (GCMC: 35 ns in total, 19 million force evaluations; GCNCMC: 7 ns in total, 32 million force evaluations). However, neither method could rehydrate either site (Figure S17, S18) when the ligand already occupied both sites. It is possible that there are other slow rearrangements in the system which need to occur to allow the ligand to move back to the crystallographic pose, but if these are present our analysis has not revealed them.

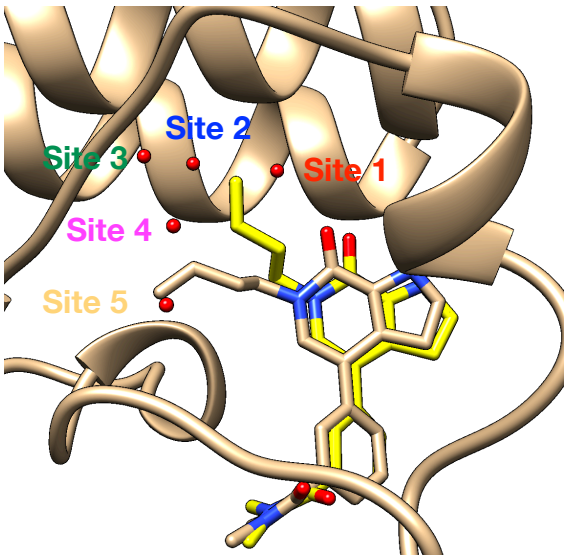

Figure 8: Ligand motion blocks insertion of water molecules in the target site (Site 1 and 2) of a TAF1(2) system (PDB: 5I1Q). The extracted snapshot from GCMC simulations is shown in yellow and the crystal structure is shown in tan. Motion of the alkyl tail on the yellow structure brings it close to Sites 1 and 2, blocking water insertion in to those sites.

5 DISCUSSION

In this work, we performed GCMC simulations of water rehydration for selected sites in 21 different systems across 10 protein targets. Our results show that GCMC is able to rehydrate all target sites after equilibration phases with GCMC moves and a short production run (2.5 ns, 1.4 million force evaluations) for most studied systems (16 out of 21 systems). These results, along with our previous work,[14] suggest GCMC is a relatively robust approach for simulating water rehydration. However, for 5 studied systems, in at least one trial of the initial short GCMC simulations, some of the target water sites were not successfully rehydrated.

We also tested a newly developed method, GCNCMC, which combines NCMC and GCMC techniques so that water insertion/removal is performed in an NCMC fashion instead of instantaneous attempts as in GCMC. For each attempt, GCNCMC is more expensive than GCMC but the environment (e.g., protein, ligand, water) is allowed to relax upon the water insertion/removal which was expected to improve the acceptance rate and perhaps allow

some types of insertions which could be very rare with GCMC alone.

To test whether GCNCMC provides benefits that GCMC does not, we revisited the five systems which initial short GCMC simulations had failed to rehydrate. We then performed GCNCMC and longer GCMC simulations on these systems. Our results show that GCNCMC can rehydrate water sites in both a HIV-1 protease system (PDB: 1EC1) and a HSP90 system (PDB: 3RLP) even when protein motions in the binding site are observed (e.g., binding site collapse).

However, GCNCMC did not always work when protein/ligand motions were observed that blocked water insertion in simulations. For example, in simulations of the HIV-1 protease system (PDB: 1EC0), neither GCNCMC simulation trial could rehydrate the water site when the binding site was slightly collapsed. In simulation of the TAF1 system (PDB: 5I1Q), when the ligand occupied the target water sites, GCNCMC failed to rehydrate those sites.

In simulations of a HSP90 system (PDB: 3RLP), both GCMC and GCNCMC can rehydrate the target water site even though a protein side chain blocks water insertion to the site. GCMC successfully rehydrated all target sites more rapidly than GCNCMC simulations. We found the protein side chain rotated and left the target site before a water was inserted in both GCMC and GCNCMC simulations. Potentially GCMC could rehydrated the site more rapidly than GCNCMC simply because the sidechain rearrangement happened earlier in the GCMC simulations than in the GCNCMC simulations due to statistical fluctuations.

Our results suggest this newly developed GCNCMC method is a better choice to handle protein/ligand motions for water rehydration. However, compared to GCMC, GCNCMC has an additional parameter, the NCMC switching time, which needs to be determined in advance. This parameter affects the efficiency of this method. More work is needed to develop a protocol for determining an efficient NCMC switching time in GCNCMC simulations for protein-ligand systems. Our results show that it is straightforward to determine such a value in bulk water simulations. However, when we tried to use the same protocol to determine an efficient switching time for simulations of protein-ligand systems, our results were very

noisy. Currently, the best approach for choosing a switching time seems to be to rely on prior work which tested several switching times and chose one for efficiency.

Based on what we learned from this work and our previous studies, we have following suggestions on water sampling in MD simulations. For cases where water site locations are known but they are challenging to adequately sample with normal MD (e.g., in relative binding free energy calculations, morphing from one ligand to another displaces water molecules and such rearrangement is slow), GCMC provides a fairly general and robust approach to recover those water sites. If the crystal structure is available, using position restraints on protein and ligand heavy atoms to maintain the crystallographic pose is helpful for water insertion in most systems we studied. In cases where GCMC performance is affected due to protein/ligand motions or it is important to understand concerted motions related to water insertion/removal, GCNCMC can be useful. In cases where it is unclear about the water site locations, we suggest using both GCMC and GCNCMC along with other available methods to provide multiple predictions and to cross validate the results for a better confidence in the predicted water locations and occupancies. In cases where diverse methods agree about water locations and occupancies, it likely means results are converged and the results are clear; in cases where there are discrepancies, results indicate sampling problems or other issues that require further investigation.

6 CONCLUSION

In this work we assess GCMC and GCNCMC performance in water sampling using 21 protein-ligand complexes. Our results suggest GCMC is in general relatively robust but can be adversely affected by protein/ligand motions. In such cases, GCNCMC can be more useful, as it overcomes some of the limitations of GCMC. We found that GCNCMC provides substantial but system dependent benefits. We hope this work can be useful for future developments of techniques for handling water motions sampling in simulations.

7 ACKNOWLEDGEMENTS

D.L.M. appreciates financial support from the National Institutes of Health (R01GM108889 and R01GM132386). DLM and YG also appreciate financial support from XtalPi. We appreciate the Open Force Field Consortium for its support of the Open Force Field Initiative, which provided software infrastructure used in this work.

8 ASSOCIATED CONTENT

Supporting Information Available

Supporting information is available free of charge via the Internet at `http://pubs.acs.org`.

Supporting tables of parameters used in *grand* simulations; supporting figures of experimental/calculated electron density maps.

Input files for simulations and scripts for analysis are freely available at `https://github.com/MobleyLab/GCNCMC_GCMC`.

Simulations were performed using the *grand* (`https://github.com/essex-lab/grand`).

Analysis was performed using Mdtraj (v1.9.4, `https://github.com/mdtraj/mdtraj`), *grand* (`https://github.com/essex-lab/grand`), CCTBX (v2021.1, `https://github.com/cctbx/cctbx_project`), LUNUS (`https://github.com/mewall/lunus`), Phenix (v1.91.1, `https://www.phenix-online.org`), Coot (v0.9.4, installed with Phenix), CCP4 (v7.1, `https://www.ccp4.ac.uk`).

9 Notes

D.L.M. is a member of the Scientific Advisory Boards of OpenEye Scientific Software and Anagenex and is an Open Science Fellow with Roivant.

References

(1) Laage, D.; Elsaesser, T.; Hynes, J. T. Water Dynamics in the Hydration Shells of Biomolecules. *Chemical Reviews* **2017**, *117*, 10694–10725, PMID: 28248491.

(2) Maurer, M.; De Beer, S. B. A.; Oostenbrink, C. Calculation of Relative Binding Free Energy in the Water-Filled Active Site of Oligopeptide-Binding Protein A. *Molecules* **2016**, *21*.

(3) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization. *Journal of the American Chemical Society* **2009**, *131*, 15403–15411, PMID: 19778066.

(4) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *Journal of Chemical Information and Modeling* **2017**, *57*, 2911–2937, PMID: 29243483.

(5) Adams, D. Chemical Potential of Hard-Sphere Fluids by Monte Carlo Methods. *Mol. Phys.* **1974**, *28*, 1241–1252.

(6) Adams, D. Grand Canonical Ensemble Monte Carlo for a Lennard-Jones Fluid. *Mol. Phys.* **1975**, *29*, 307–311.

(7) Mezei, M. A Cavity-Biased (T, V, μ) Monte Carlo Method for the Computer Simulation of Fluids. *Mol. Phys.* **1980**, *40*, 901–906.

(8) Mezei, M. Grand-Canonical Ensemble Monte Carlo Study of Dense Liquid: Lennard-Jones, Soft Spheres and Water. *Mol. Phys.* **1987**, *61*, 565–582.

(9) Ross, G. A.; Bodnarchuk, M. S.; Essex, J. W. Water Sites, Networks, And Free Energies with Grand Canonical Monte Carlo. *J. Am. Chem. Soc.* **2015**, *137*, 14930–14943.

(10) Ross, G. A.; Bruce Macdonald, H. E.; Cave-Ayland, C.; Cabedo Martinez, A. I.; Essex, J. W. Replica-Exchange and Standard State Binding Free Energies with Grand Canonical Monte Carlo. *J. Chem. Theory Comput.* **2017**, *13*, 6373–6381.

(11) Bruce Macdonald, H. E.; Cave-Ayland, C.; Ross, G. A.; Essex, J. W. Ligand Binding Free Energies with Adaptive Water Networks: Two-Dimensional Grand Canonical Alchemical Perturbations. *J. Chem. Theory Comput.* **2018**, *14*, 6586–6597.

(12) Bodnarchuk, M. S.; Packer, M. J.; Haywood, A. Utilizing Grand Canonical Monte Carlo Methods in Drug Discovery. *ACS Med. Chem. Lett.* **2020**, *11*, 77–82.

(13) Ross, G. A.; Russell, E.; Deng, Y.; Lu, C.; Harder, E. D.; Abel, R.; Wang, L. Enhancing Water Sampling in Free Energy Calculations with Grand Canonical Monte Carlo. *J. Chem. Theory Comput.* **2020**, *16*, 6061–6076.

(14) Ge, Y.; Wych, D. C.; Samways, M. L.; Wall, M. E.; Essex, J. W.; Mobley, D. L. Enhancing Sampling of Water Rehydration on Ligand Binding: A Comparison of Techniques. *J. Chem. Theory Comput.* **2022**, *18*, 1359–1381.

(15) Bergazin, T. D.; Ben-Shalom, I. Y.; Lim, N. M.; Gill, S. C.; Gilson, M. K.; Mobley, D. L. Enhancing Water Sampling of Buried Binding Sites Using Nonequilibrium Candidate Monte Carlo. *J Comput Aided Mol Des* **2021**, *35*, 167–177.

(16) Melling, O.; Samways, M.; Ge, Y.; Mobley, D.; Essex, J. Enhanced Grand Canonical Sampling of Occluded Water Sites Using Nonequilibrium Candidate Monte Carlo. *ChemRxiv* **2022**,

(17) Ben-Shalom, I. Y.; Lin, Z.; Radak, B. K.; Lin, C.; Sherman, W.; Gilson, M. K. Accounting for the Central Role of Interfacial Water in Protein–Ligand Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2020**, *16*, 7883–7894.

(18) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. Classification of Water Molecules in Protein Binding Sites. *J. Am. Chem. Soc.* **2007**, *129*, 2577–2587.

(19) Wall, M. E. In *Micro and Nano Technologies in Bioanalysis. Methods in Molecular Biolog^{TM} (Methods and Protocols)*; Lee, J. W., Foote, R. S., Eds.; Humana Press: Totowa, NJ., 2009; pp 269–279.

(20) Grosse-Kunstleve, R. W.; Sauter, N. K.; Moriarty, N. W.; Adams, P. D. The Computational Crystallography Toolbox: Crystallographic Algorithms in a Reusable Software Framework. *J. Appl. Crystallogr.* **2002**, *35*, 126–136.

(21) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532.

(22) Emsley, P.; Cowtan, K. *Coot* : Model-Building Tools for Molecular Graphics. *Acta Crystallogr D Biol Crystallogr 60*, 2126–2132.

(23) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and Development of *Coot*. *Acta Crystallogr D Biol Crystallogr 66*, 486–501.

(24) Samways, M. L.; Bruce Macdonald, H. E.; Essex, J. W. Grand: A Python Module for Grand Canonical Water Sampling in OpenMM. *J. Chem. Inf. Model.* **2020**, *60*, 4436–4441.

(25) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput Biol* **2017**, *13*, e1005659.

(26) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys* **1983**, *79*, 926–935.

(28) Qiu, Y.; Smith, D. G. A.; Boothroyd, S.; Jang, H.; Hahn, D. F.; Wagner, J.; Bannan, C. C.; Gokey, T.; Lim, V. T.; Stern, C. D.; Rizzi, A.; Tjanaka, B.; Tresadern, G.; Lucas, X.; Shirts, M. R.; Gilson, M. K.; Chodera, J. D.; Bayly, C. I.; Mobley, D. L.; Wang, L.-P. Development and Benchmarking of Open Force Field v1.0.0—the Parsley Small-Molecule Force Field. *J. Chem. Theory Comput.* **2021**, *17*, 6262–6280.

(29) Wagner, J.; Thompson, M.; Dotson, D.; hyejang,; Rodríguez-Guerra, J. openforcefield/openforcefields: Version 1.2.1 "Parsley" Update. `https://doi.org/10.5281/zenodo.4021623`.

(30) Leimkuhler, B.; Matthews, C. Rational Construction of Stochastic Numerical Methods for Molecular Sampling. *Applied Mathematics Research eXpress* **2012**, abs010.

(31) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N -log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys* **1993**, *98*, 10089–10092.

(32) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(33) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of p K_{a} Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284–2295.

(34) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K_{a} Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

(35) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: An Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations. *Nucleic Acids Res.* **2004**, *32*, W665–W667.

(36) Liu, D. C.; Nocedal, J. On the Limited Memory BFGS Method for Large Scale Optimization. *Math. Program.* **1989**, *45*, 503–528.

(37) Fields, B. A.; Bartsch, H. H.; Bartunik, H. D.; Cordes, F.; Guss, J. M.; Freeman, H. C. Accuracy and precision in protein crystal structure analysis: two independent refinements of the structure of poplar plastocyanin at 173 K. *Acta Crystallogr D Biol Crystallogr* **1994**, *50*, 709–730.

(38) Ohlendorf, D. H. Accuracy of refined protein structures. II.Comparison of four independently refined models of human interleukin 1beta. *Acta Crystallogr D Biol Crystallogr* **1994**, *50*, 808–812.

(39) Samways, M. L.; Taylor, R. D.; Bruce Macdonald, H. E.; Essex, J. W. Water molecules at protein–drug interfaces: computational prediction and analysis methods. *Chem. Soc. Rev.* **2021**, *50*, 9104–9120.

10 Supplementary Information

11 Supporting Tables

Table S1: The excess chemical potential and standard state volume of water used in *grand* simulations at different temperatures for different systems.

Temperature (K)	Excess chemical potential (kcal/mol)	Standard state volume (Å^3)	Systems and PDB IDs
278	-6.34	29.823	PTP1B (2QBS) HSP90 (2XAB, 2XJG) Thrombin (2ZFF) BTK (4ZLZ) TAF1(2) (5I1Q, 5I29) HIV-1 protease (1EC1, 1EC0, 1EBW, 1EBY)
286	-6.19	30.035	HSP90 (3RLP, 3RLQ, 3RLR)
298	-6.09	30.345	Trypsin (1C5T, 1GI1, 1F0U, 1O2J) FXa (1EZQ, 1LPG, 1LPZ, 1F0S)

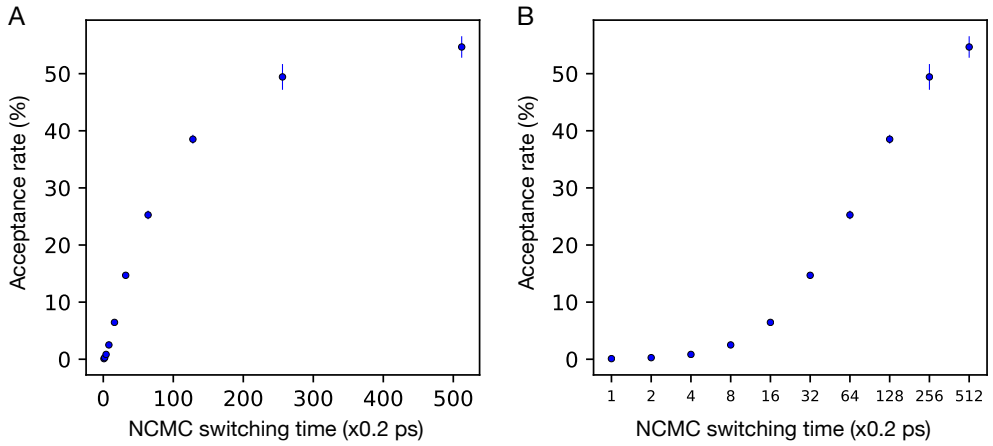

Figure S1: Acceptance rates as a function of different NCMC switching time in bulk water simulations show that beyond a switching time of 3.2 ps, doubling the switching time does not return a doubled acceptance rate. The x-axis is shown in a (A) linear scale (B) log scale. The error bar error bar was calculated using the standard deviation value from two simulation replicates.

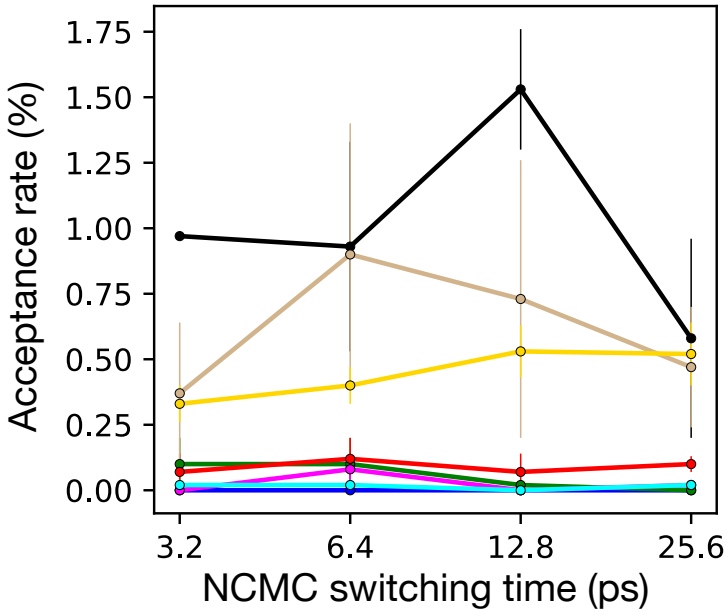

Figure S2: Acceptance rates as a function of different NCMC switching time in protein-ligand complex simulations are very noisy and show no clear trends to identify an optimal switching time. Different color are results from simulations of different protein target systems: HSP90 (blue, PDB: 2XAB), HIV-1 protease (green, PDB: 1HPX), trypsin (red, PDB: 1AZ8), FXa (magenta, PDB: 1EZQ), thrombin (black, PDB: 2ZFF), BTK (tan, PDB: 4ZLZ), PTP1B (cyan, PDB:2QBS), TAF1 (gold, PDB: 5I1Q). The error bar error bar was calculated using the standard deviation value from two simulation replicates.

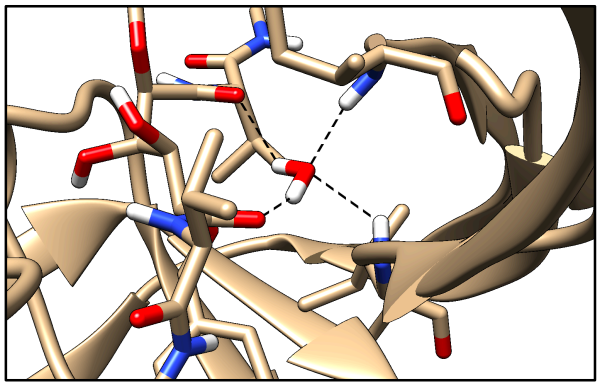

Figure S3: Hydrogen bonds between the target water molecule and the protein/ligand atoms stabilize the water molecule in the binding site of a HIV-1 system (PDB: 1EC1).

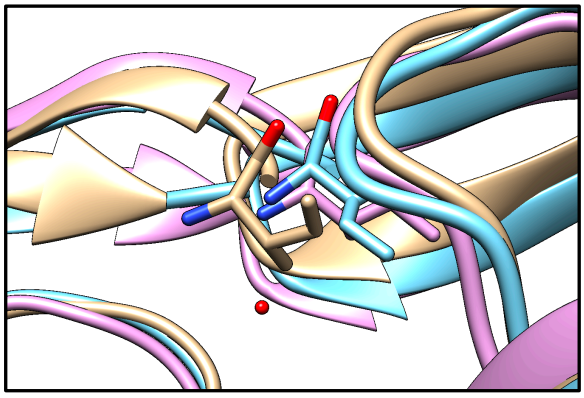

Figure S4: In simulations of a HIV-1 system (PDB: 1EC1) where the water insertion failed, we found the ILE50 backbone on the loop is closer to the binding site than that in the crystal structure and simulation snapshot taken from GCNCMC simulations where water can be inserted.

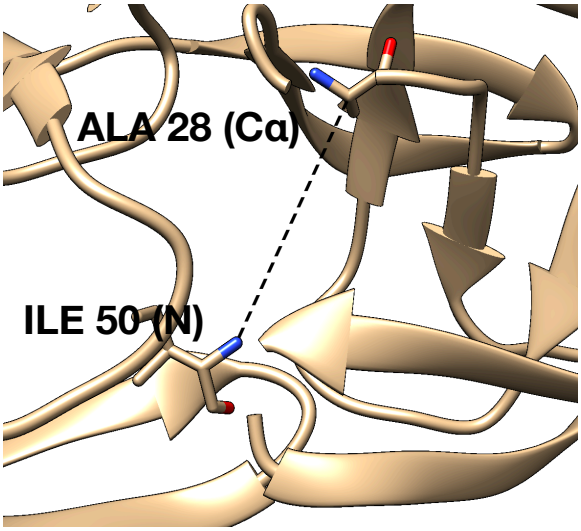

Figure S5: We monitored the distance between ILE50(N) and ALA28(Cα) to check if the loop moved closer to the binding site during simulations of a HIV-1 system (PDB: 1EC1).

Figure S6: The target water site in a HIV-1 protease system (PDB: 1EC1) is successfully rehydrated in both replicates (panel A and B) of GCMC simulations where the protein and ligand heavy atoms are restrained to the crystallographic pose. The electron density maps calculated from simulations are shown in blue and the experimental maps are shown in white. Each simulation block is continued from the previous one.

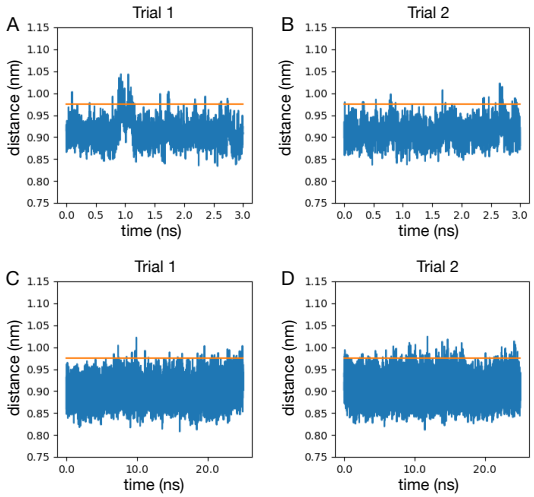

Figure S7: The loop backbone of the HIV-1 protease (PDB: 1EC0) moves closer to the binding site when the target hydration site is not occupied. The distance change between ILE50(N) and ALA28(Cα) during (A-B) GCNCMC simulations and (C-D) GCMC simulations. The orange line shows the distance between the two atoms in the crystal structure.

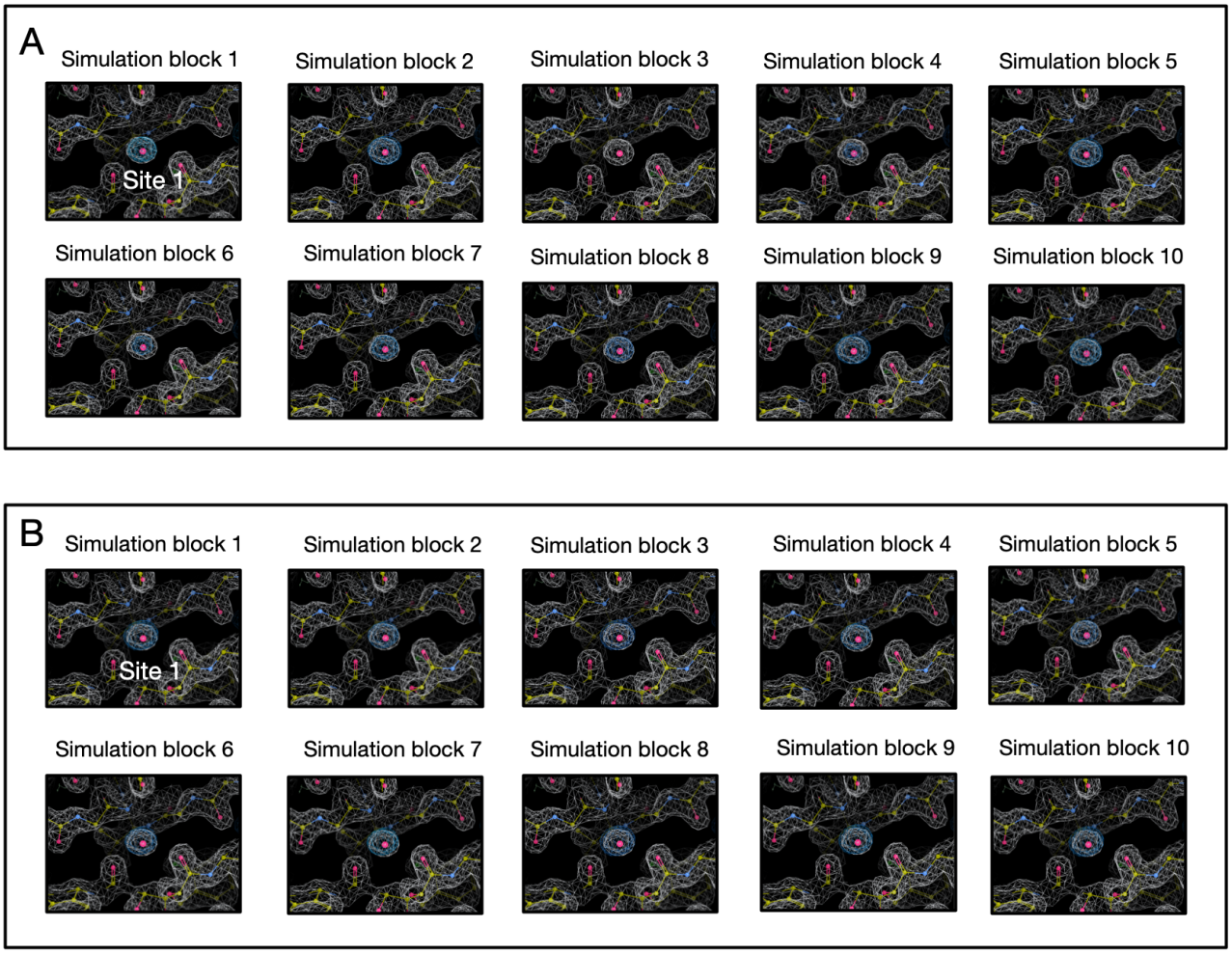

Figure S8: The target water site in a HIV-1 protease system (PDB: 1EC0) is successfully rehydrated in both replicates (panel A and B) of GCMC simulations where the protein and ligand heavy atoms are restrained to the crystallographic pose. The electron density maps calculated from simulations are shown in blue and the experimental maps are shown in white. Each simulation block is continued from the previous one.

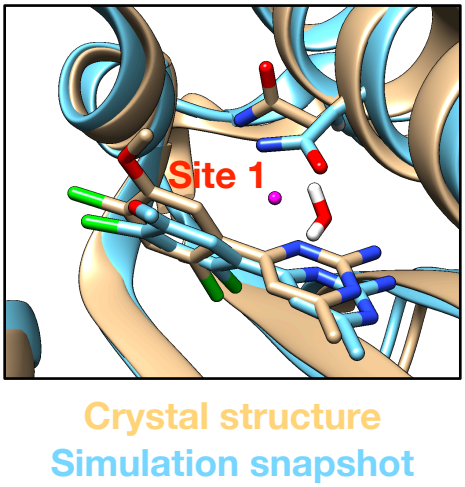

Figure S9: A comparison between a simulation snapshot extracted from GCNCMC simulations (block 5 from Trial 2 in Figure 4B) and the crystal structure of a HSP90 system (PDB: 3RLP). The inserted water molecule is close to Site 1 in the crystal structure (magenta blob).

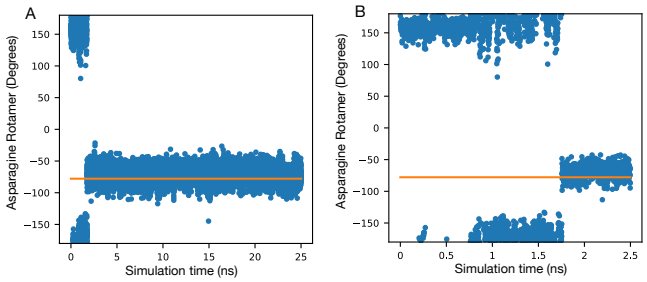

Figure S10: ASN42 rotamer data for GCMC simulations of a HSP90 system (PDB: 3RLP). (A) Calculated $\chi 1$ angle of ASN42 as a function of GCMC simulation time (ns). The orange line shows the angle calculated from the crystal structure. (B) Calculated $\chi 1$ angle of ASN42 as a function of GCMC simulation time (ns) for simulation block 1.

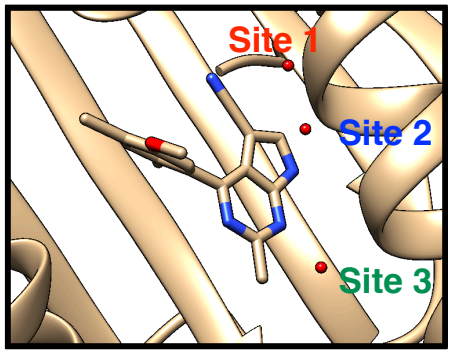

Figure S11: The target water site in a HSP90 system (PDB: 3RLQ).

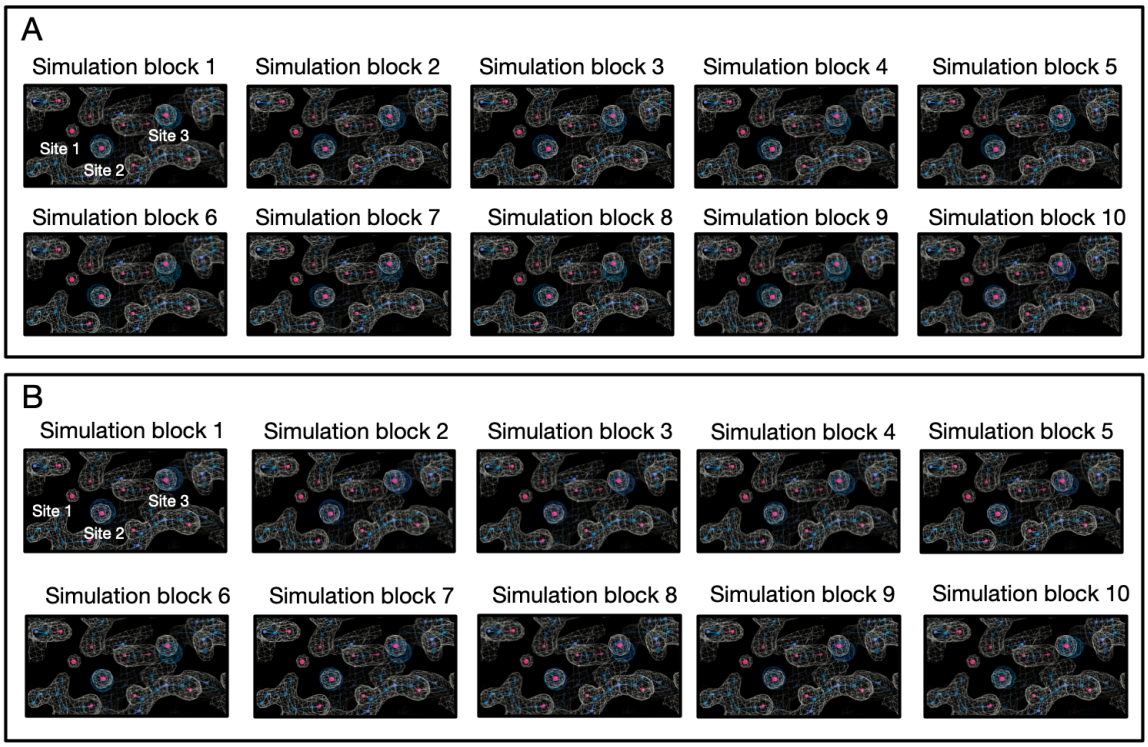

Figure S12: Only two target water sites in a HSP90 system (PDB: 3RLQ) were successfully recovered in both GCMC simulation trials (A-B). The electron density maps calculated from simulations are shown in blue and the experimental maps are shown in white.

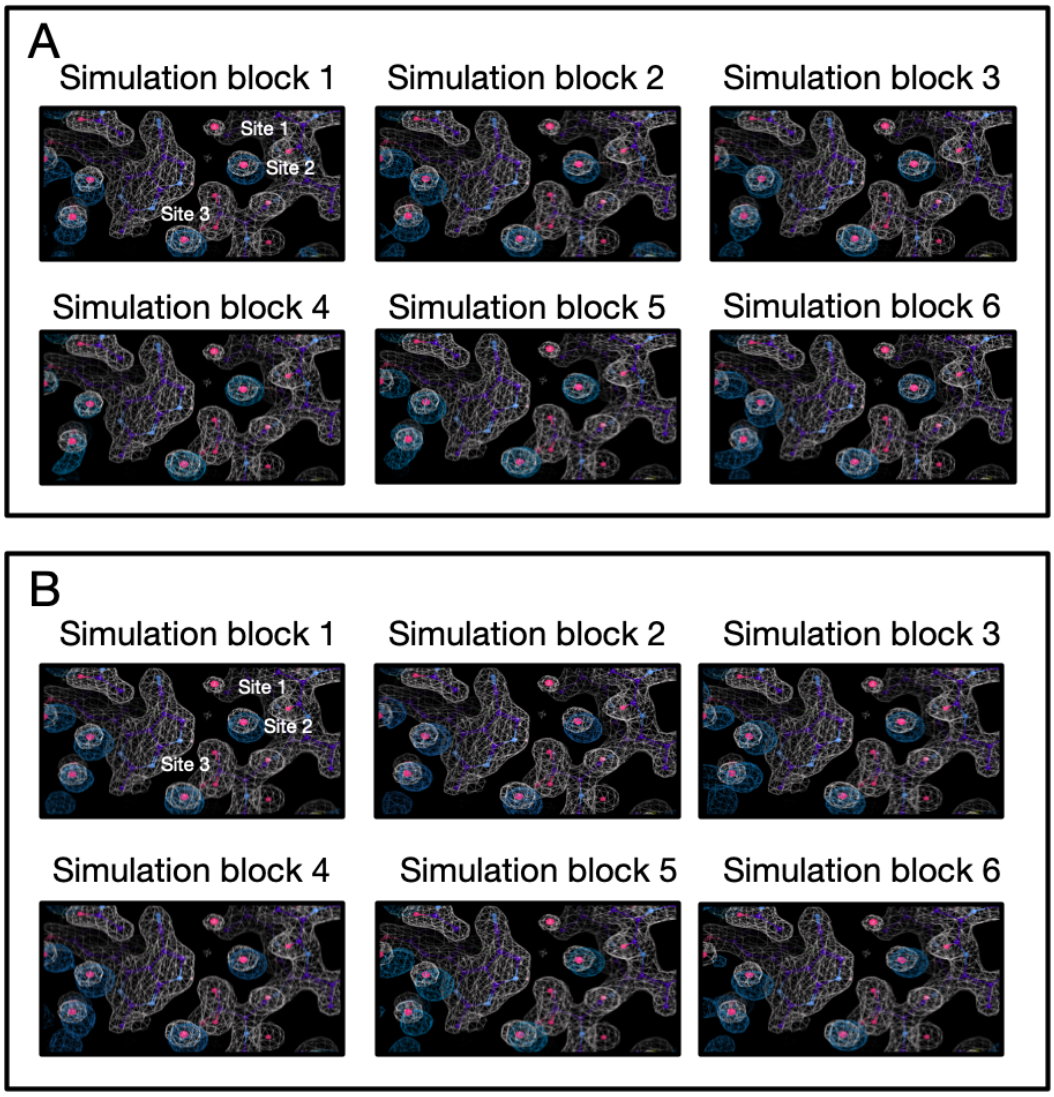

Figure S13: Only two target water sites in a HSP90 system (PDB: 3RLQ) were successfully recovered in both GCNCMC simulation trials (A-B). The electron density maps calculated from simulations are shown in blue and the experimental maps are shown in white.

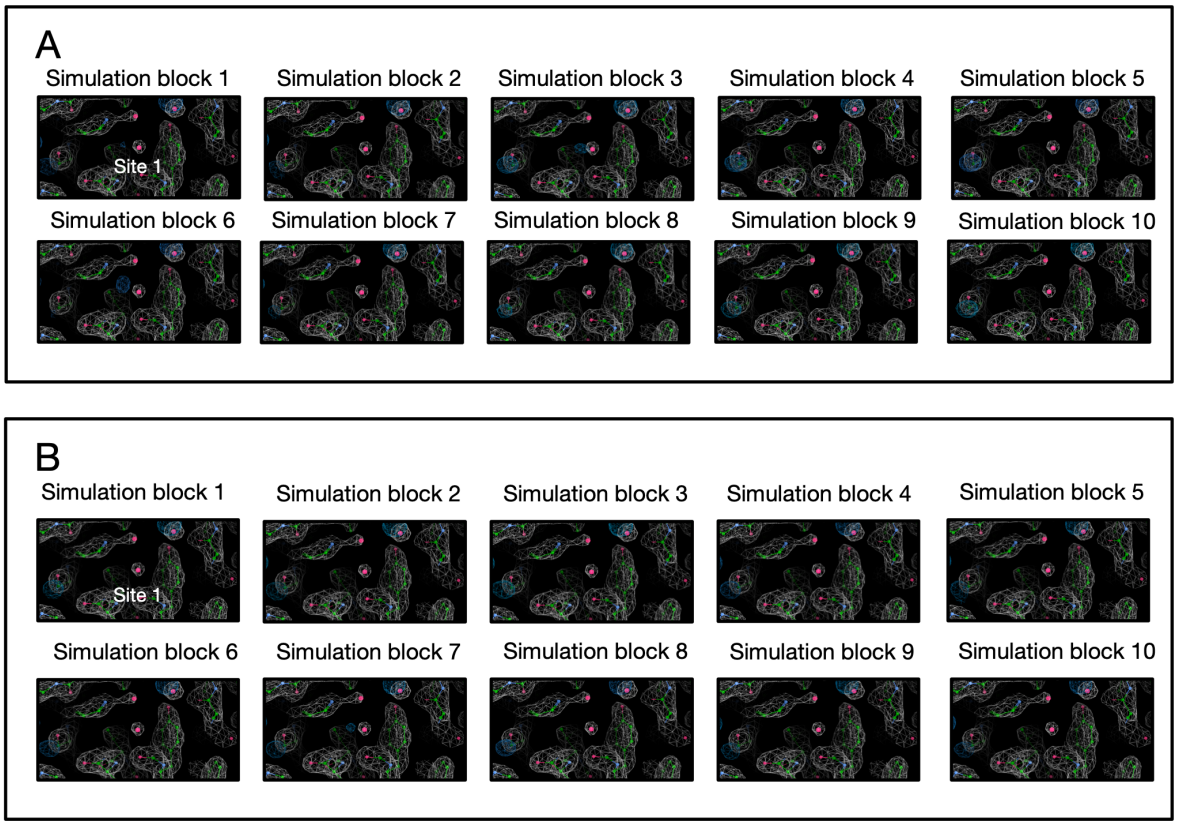

Figure S14: The target water site in a FXa system (PDB: 1F0S) is not occupied in most GCMC simulation blocks. (A-B) The electron density maps calculated from GCMC simulations (two trials) are shown in blue and the experimental maps are shown in white. Each simulation block is continued from the previous one.

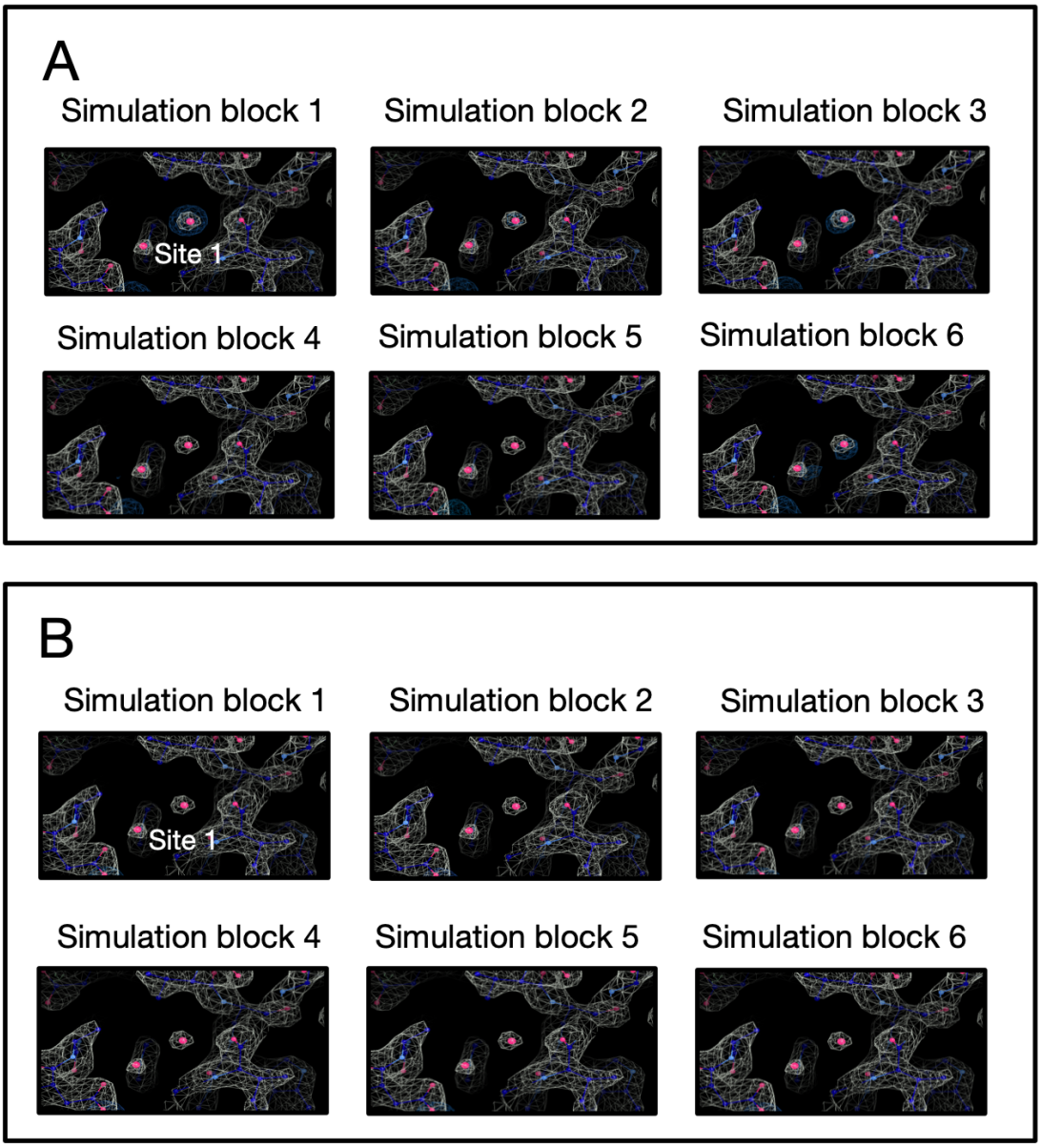

Figure S15: The target water site in a FXa system (PDB: 1F0S) is not occupied in most GCNCMC simulation blocks. (A-B) The electron density maps calculated from GCMC simulations (two trials) are shown in blue and the experimental maps are shown in white. Each simulation block is continued from the previous one.

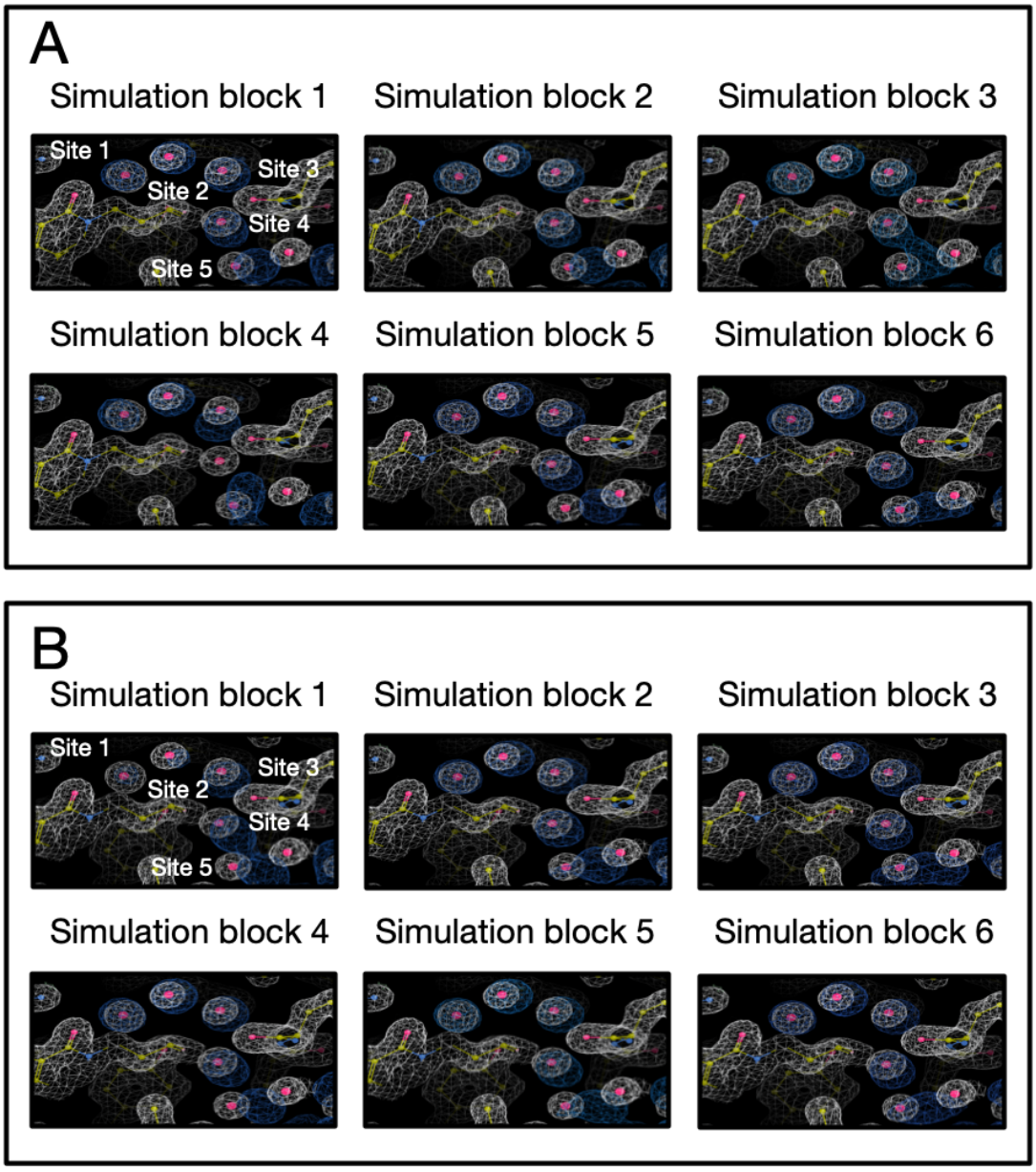

Figure S16: All five target sites in a TAF1 system (PDB: 5I1Q) were successfully rehydrated in both GCNCMC simulation trials (A-B). The electron density maps calculated from simulations are shown in blue and the experimental maps are shown in white.

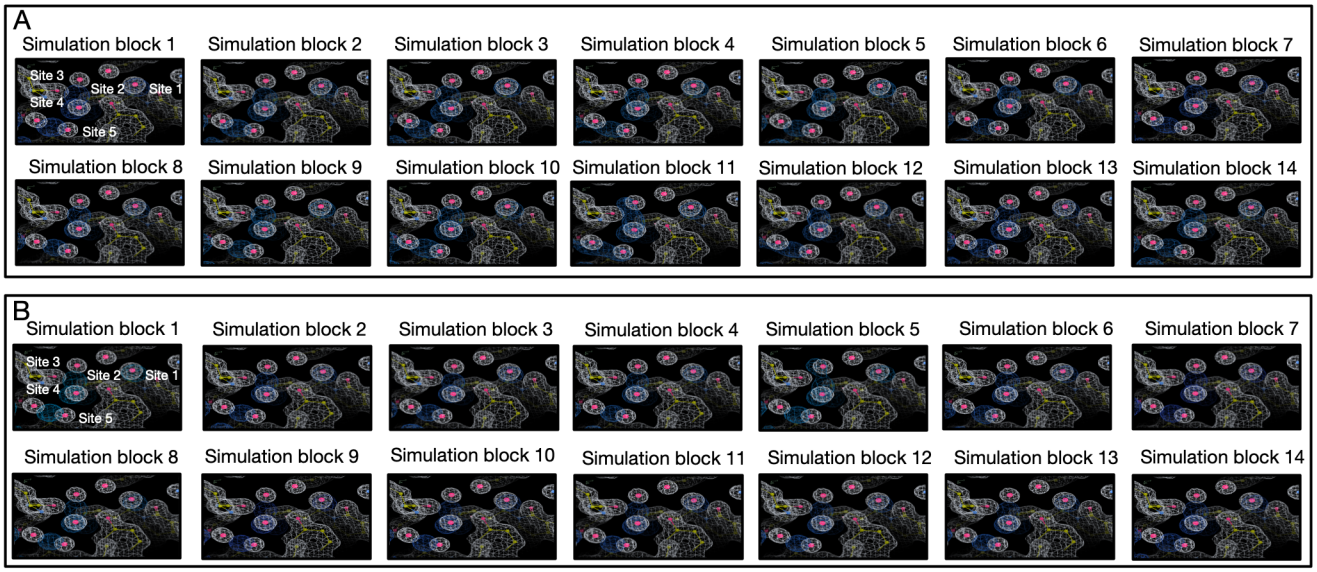

Figure S17: (A-B) Both GCMC simulation trials using the starting structure as shown in Figure 8 (yellow) failed to rehydrate Site 1 and 2 of a TAF1 system (PDB: 5I1Q). The electron density maps calculated from simulations are shown in blue and the experimental maps are shown in white.

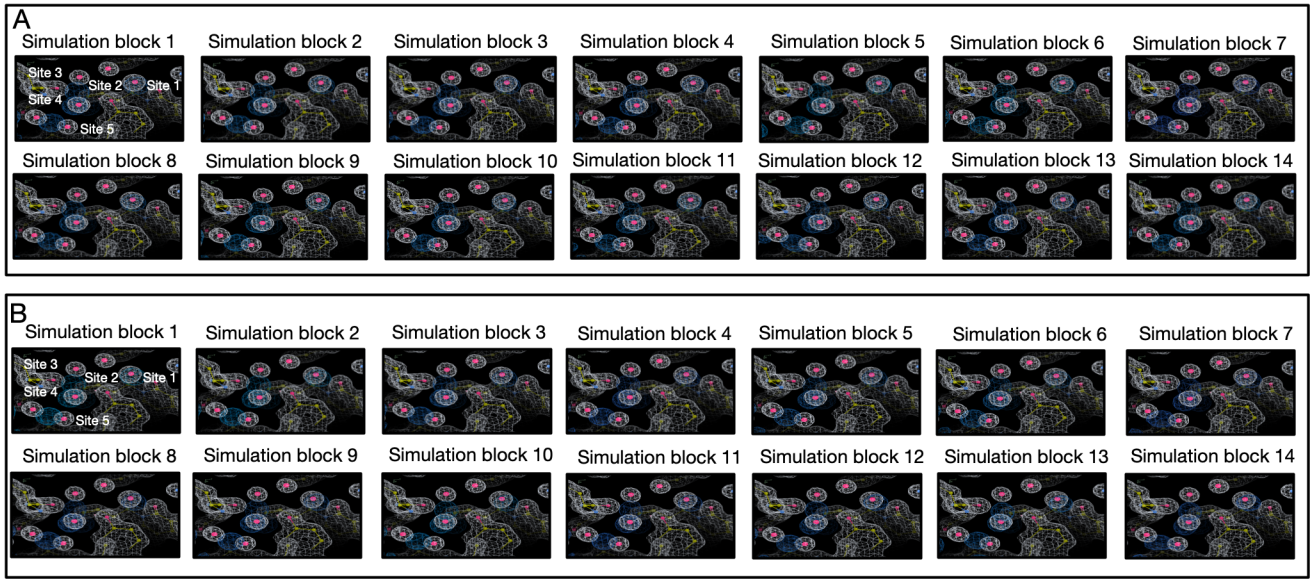

Figure S18: (A-B) Both GCNCMC simulation trials using the starting structure as shown in Figure 8 (yellow) failed to rehydrate Site 1 and 2 of a TAF1 system (PDB: 5I1Q). The electron density maps calculated from simulations are shown in blue and the experimental maps are shown in white.