

Replica exchange and standard state binding free energies with grand canonical Monte Carlo

Gregory A. Ross,[†] Hannah E. Bruce Macdonald,[‡] Christopher Cave-Ayland,[‡] Ana
I. Cabedo Martinez,[‡] and Jonathan W. Essex^{*,‡}

*Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan
Kettering Cancer Center, New York, New York, USA., and Department of Chemistry,
University of Southampton, Southampton, SO17 1BJ, UK*

E-mail: J.W.Essex@soton.ac.uk

Abstract

The ability of grand canonical Monte Carlo (GCMC) to create and annihilate molecules in a given region greatly aids the identification of water sites and water binding free energies in protein cavities. However, acceptance rates without the application of biased moves can be low, resulting in large variations in the observed water occupancies. Here, we show that replica-exchange of the chemical potential significantly reduces the variance of the GCMC data. This improvement comes at a negligible increase in computational expense when simulations comprise of runs at different chemical potentials. Replica-exchange GCMC is also found to substantially increase the precision of standard state water binding free energies as calculated with grand canonical integration, which has allowed us to address a missing standard state correction.

*To whom correspondence should be addressed

[†]Memorial Sloan Kettering Cancer Center.

[‡]University of Southampton

Introduction

Aided by community efforts such as the Drug Design Data Resource (D3R) and Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) challenges,¹⁻³ rigorous binding free energy calculations are becoming increasingly feasible and reliable for use in drug discovery projects.⁴ Recent studies using these methods have reported promising results in predicting and explaining the selectivity of small molecule inhibitors to protein targets.^{5,6} In addition to predicting the binding affinity of small molecules to proteins, free energy calculations can also help validate hypotheses in structure-based drug design. In particular, medicinal chemists often seek to exploit the existence of water molecules in binding sites by designing compounds that will displace water upon binding.^{7,8} While there are a number of cases where the targeted displacement of water molecules has resulted in more potent and/or specific compounds,⁹⁻¹¹ there are also examples where new compounds have failed to displace a water molecule,^{12,13} or in doing so have reduced the affinity of the compound.^{14,15} Binding free energy calculations using all-atom simulations have the potential to help medicinal chemists in this endeavor, as the calculated binding free energies for water at particular sites have been shown to be indicative of how ‘displaceable’ a water molecule is.¹⁶

One of the most popular ways to calculate the absolute binding free energy of water is via double decoupling (DD).¹⁷⁻²⁰ This method uses two sets of simulations that gradually reduce the interaction energy between a chosen water molecule and the rest of the system over a series of non-physical, *alchemical*, intermediate states. One simulation alchemically decouples a water molecule from a particular location in a protein and the other decouples a water molecule from bulk solvent. When performing DD calculations on water molecules, the bulk solvent calculation need only be performed once for a given water model and set of simulation parameters (such as the length of the non-bonded cutoff). The decoupling calculations in the protein require the careful application of constraints and/or restraints to keep the water molecule in question bound to a particular location and to prevent other water

molecules drifting into the site after decoupling. Applying DD calculations to a bonded network of buried water molecules would not only require cumbersome constraints, but would also require a separate decoupling simulation for each water in the network. The water molecules should be decoupled in the order of weakest to tightest bound, but establishing the correct order will require multiple simulations and the choice of pathway may increase the variance of the results.

Using binding free energy calculations of water to drive rational drug design is predicated on knowing the probable locations of water within binding sites. Whilst X-ray crystallography is the most well used method of structure determination, locating water sites within X-ray structures is fraught with difficulty. For instance, water molecules may be artifactual,²¹ may not appear at a given level of resolution^{22–24} or may require special techniques to find them.²⁵ There can even be poor agreement between high resolution structures, with studies finding that an observed water has approximately only a 50% chance of being within one Ångstrom of an independently resolved structure of the same protein.^{26,27} Molecular simulations with explicit water offer the possibility of predicting the locations of water in protein cavities that are consistent with the forcefield used in the free energy calculation. However, despite the ever-increasing timescales accessible by molecular dynamics (MD) simulations—brought about by specialized processors and GPU accelerated code^{28,29}—water molecules that are buried within proteins are systematically under sampled in MD. Nuclear magnetic resonance studies have indicated that the residence times of water in cavities that are inaccessible to bulk water range from tens to hundreds of microseconds,^{30,31} which is beyond the timescales typically explored by MD studies.

We recently reported theoretical and methodological improvements to the sampling technique known as grand canonical Monte Carlo (GCMC) applied to water.³² In essence, GCMC can be considered as an enhanced sampling method, as molecules (in this case water

molecules) can be created and annihilated within a given cavity, circumventing the kinetic barriers that would be encountered in MD and, thus, aiding the determination of water sites. The recent improvements to GCMC allowed for the determination of the absolute binding free energy of highly coupled networks of water as well as the relative stabilities of individual water molecules, all in a single set of simulations. Key to the latest developments in GCMC was the exploitation of simulations performed at a range of different chemical potentials as part of a technique we refer to as grand canonical integration (GCI). When applied to calculating the binding affinity of networks of water molecules, GCI was both easier to execute and computationally less intensive than sequential alchemical DD calculations. However, the statistical uncertainty with GCI was found to be significantly larger than for DD.

Although GCMC sampling of water is orders of magnitude more time efficient than MD in sampling buried sites, GCMC suffers from its own sampling difficulties that affect the precision of the free energy calculations. In GCMC, Metropolis-Hastings acceptance rates for insertions and deletions are low, often less than 1%.³³ This is due to the possibility of steric clashes when inserting molecules into dense systems, and large energy deficits when removing waters from systems. An early solution to this issue is the cavity-bias algorithm by Mezei,³³ where before every insertion, a random set of points is uniformly sampled within the GCMC simulation region, and an insertion is attempted on a randomly selected point that does not clash with the system. This idea was later developed by Woo et al.,³⁴ who used a dynamically updated grid that kept track of the free space. Orientational biases have also been applied to improve insertion rates.^{34,35} A recurring issue with such Monte Carlo biasing schemes is whether the increased time of computing and applying a bias does not nullify the improvement in sampling efficiency.

Based on from the benefits in sampling seen with temperature,³⁶ Hamiltonian,³⁷⁻³⁹ and constant pH replica-exchange methods,⁴⁰ here we apply replica-exchange (RE) to the chem-

ical potential in GCMC simulations. While biasing schemes can suffer from increased computing time as a payoff for their enhanced sampling, the addition of RE to the GCI protocol achieves enhanced sampling with a negligible effect on the run-time. The variance of these calculated free energies significantly decreased with RE, which allowed for comparisons between GCI and double decoupling calculations at a higher level of precision than was previously possible. This facilitated a detailed examination of the GCI binding free energy equation and prompted the development of an improved equation that, unlike before, computes standard state binding free energies. Therefore, the GCMC-based method reported here is more accurate and precise than the previous,³² with a statistical uncertainty that is comparable to double decoupling. Thus, absolute binding free energy calculations with GCI may be an ‘all in one’ solution for water-focused medicinal computational chemistry, as a single set of simulations can determine a large number of properties of the water molecules, pertinently indicating which of the waters, or sets of waters, are most displaceable. GCMC is able to determine locations of water molecules, binding free energies of networks and cooperatively in a single simulation.

Theory

Grand canonical Monte Carlo

Grand canonical Monte Carlo (GCMC) is the technique that allows for the creation and annihilation of molecules within a simulation. By construction, simulation statistics in GCMC are consistent with the grand canonical ensemble, where it is imagined that the system of interest can exchange molecules with a *reservoir*. The reservoir exists in a predetermined thermodynamic state that is specified by its temperature and chemical potential, which accounts for the average density and interaction energy of the molecular species in the reservoir. This work is concerned with performing GCMC on water molecules within protein cavities,

such that the natural choice of the reservoir is bulk water at room temperature.

In Adams’s formulation of GCMC,⁴¹ the probability to insert one molecule into the simulated system from the reservoir is given by

$$P_{\text{insert}} = \min \left[1, \frac{1}{N+1} \exp(B) \exp(-\beta \Delta E_{\text{insert}}) \right] \quad (1)$$

and the probability to remove one molecule from the system and add it to the reservoir is given by

$$P_{\text{remove}} = \min [1, N \exp(-B) \exp(-\beta \Delta E_{\text{remove}})], \quad (2)$$

Where N is the instantaneous number of molecules of the chemical species in the system, ΔE is the change in potential energy of inserting or removing the water molecule, β is the inverse temperature, and B is the applied Adams parameter,⁴¹ a term that accounts for the chemical potential of the reservoir and the volume of the simulated system. Originally defined by Adams in terms of the excess chemical potential and average occupancy of the system,⁴¹ B is related to the applied chemical potential, μ and volume of the GCMC region, V , via

$$B = \mu\beta + \ln \left(\frac{V}{\Lambda^3} \right), \quad (3)$$

where Λ^3 is the thermal wavelength of the GCMC molecule.

Replica-exchange GCMC

Guarnieri and Mezei pioneered the technique in which many independent GCMC simulations of water are performed at a range of different B values—equivalently, different chemical potentials.⁴² By viewing how and where the occupancy of water changed as a function of B , they obtained a semi-quantitative map of the regions of high water affinity. By analogy with ligand-protein binding assays, we use the term ‘titration’ to refer to a set of GCMC simulations of the same system at different B values. This titration technique was extended in our previous study where the average number of water molecules as a function of B could be integrated to predict the binding free energies of water networks.³² This method is referred to as grand canonical integration (GCI) and is discussed in more detail below. Given the utility of performing GCMC simulations at different chemical potentials, and the sampling improvements gained by Hamiltonian replica-exchange in alchemical free energy calculations,^{37,38} replica-exchange of the B values in GCMC simulations has the potential to improve the accuracy of the free energies calculated with GCI.

When running concurrent GCMC simulations at a range of chemical potentials, the probability of accepting a Metropolis-Hastings move that swaps the B values of the i th and j th simulations is given by

$$P_{\text{swap}} = \min \left[1, \exp \left((B_i - B_j)(N_j - N_i) \right) \right], \quad (4)$$

where N_i and N_j are the number of water molecules in the i th and j th replicas. This equation is similar to the Metropolis-Hastings criteria in constant-pH replica-exchange,⁴⁰ as well as parallel tempering and Hamiltonian replica-exchange³⁶ and is derived in the Supplementary Information by constructing an expanded ensemble that allows for chemical potentials to vary. One would expect replica-exchange of the chemical potentials to enhance the sam-

pling in GCMC titrations, because simulations at high chemical potentials – where insertions are more likely – can mix with simulations at low chemical potentials – where deletions are more likely.

Binding free energies with grand canonical integration

Grand canonical integration (GCI) is a technique that integrates over the average number of water molecules as a function of the B value to predict the relative binding free energies of water molecules.³² If the lowest B value has no waters bound, GCI can predict absolute binding free energies. When applied to calculating the binding free energy of water networks, the GCI method is more computationally efficient and easier to implement than traditional double-decoupling techniques,³² which require multiple simulations for each water molecule in the network as well as the careful application of restraints or constraints.

Given that GCI uses the information from simulations run at different B values, one can expect the statistical uncertainty of the calculated free energies to improve with replica-exchange. Indeed, as shown in the Results section and SI, the gain in precision afforded by replica-exchange revealed that binding free energies calculated with the method described previously³² were dependent on the volume of the GCMC region, and had statistically resolvable systematic differences with free energies calculated by double decoupling. This prompted a re-examination of the equation used to calculate binding free energies with GCI. While the original GCI equation—for the free energy to transfer water from ideal gas to the system of interest at a fixed volume—is correct, the standard state corrections were missing for the binding free energy. Derived in the Supplementary Information (SI), the equation to calculate relative standard state binding free energies between an initial number of waters N_i and final number N_f is given by

$$\beta\Delta G_{\text{bind}}^o(N_i \rightarrow N_f) = N_f B_f - N_i B_i - (N_f - N_i) \left[\beta\mu'_{\text{sol}} + \ln \left(\frac{V_{\text{sys}}}{V^o} \right) \right] - \int_{B_i}^{B_f} N(B) dB, \quad (5)$$

where B_k is the Adams parameter for which there are an average of N_k waters, V_{sys} is the volume of the GCMC region, V^o is the standard state volume of bulk water, μ'_{sol} is the excess chemical potential of the water in bulk water. For practical purposes, μ'_{sol} is the hydration free energy of the simulated water model.

In contrast to the previous binding energy equation, this has a term for the volume of the GCMC region, and no longer depends on the factorial of N_i and N_f . As shown in the SI, these changes come from removing the implicit assumption in the previous derivation that both the free energies of the simulated system and reservoir had the same contribution from the ideal gas.

In addition to the theoretical derivation, the effect of the volume term of the GCI equation on the calculated binding free energies has been empirically tested using a very well converged, idealized protein-ligand system based on Scytalone Dehydratase. The binding free energies of water molecules should be independent of the size of the GCMC sub-volume and these results demonstrate this. These results are available in the SI, and confirm that Equation 5, is the correct form of the GCI equation.

The new GCI equation is also consistent with the following equilibrium condition for water binding

$$\mu_{\text{sys}} = \mu_{\text{sol}}, \quad (6)$$

where μ_{sys} is the chemical potential for water in the simulated system, and μ_{sol} is the chemical potential for water in bulk solvent. This relation is also derived in the SI. Owing to the previous implicit assumption about the ideal gas contribution to the free energies, this was previously erroneously stated as the equality of *excess* chemical potentials.³² Equation 6 makes it straightforward to set the B value for a GCMC simulation where water molecules in the system of interest are in equilibrium with bulk water:

$$B_{\text{equil}} = \beta\mu'_{\text{sol}} + \ln\left(\frac{V_{\text{sys}}}{V_o}\right), \quad (7)$$

Given that the hydration free energy and density of bulk water only needs to be calculated once per water model, one can set the Adams parameter, or chemical potential, prior to running a GCMC simulation, where equilibrium with bulk water is desired.

Methods

Simulations were performed to (1) assess whether replica-exchange of B values in GCMC titration simulations (RE-GCMC) improves the consistency and accuracy of repeated GCMC simulations and free energies calculated with GCI, (2) determine the degree of precision with which the new GCMC methodology agrees with double-decoupling in biomolecular systems. The protein system bovine pancreatic trypsin inhibitor (BPTI) was used to address these aims, chosen due to its prior use as a model system.⁴³

BPTI was used to test the ability of RE-GCMC to improve the sampling of GCMC titration simulations and to compare free energies calculated with GCI and replica-exchange to estimates from double decoupling. All simulations focused on a small cavity in BPTI that binds to three water molecules. In our previous investigations,³² this system was found to be the worst sampled of all ten systems studied, and thus will most clearly highlight any improvements over the previous methodology. The simulated structure was taken from

protein data bank (PDB) entry 5PTI, and was prepared as previously described.³² The protein backbone and side chain atoms were sampled over the angles and dihedrals. Backbone moves within ProtoMS are performed by rigid-body translations and rotations, centered on the intersection of the $C\alpha$ -N, and C=O bond vectors. The structure was solvated inside a half harmonically restrained, spherical droplet of water, with a radius chosen such that the edge of the droplet was approximately 15 Å away from the surface of the protein.

All simulations were performed with the freely available Monte Carlo simulation package ProtoMS, version 3.3.⁴⁴ The TIP4P water model⁴⁵ and the AMBER14SB force-field⁴⁶ were used. Non-bonded interactions were residue based switching cut-off at 10 Å.

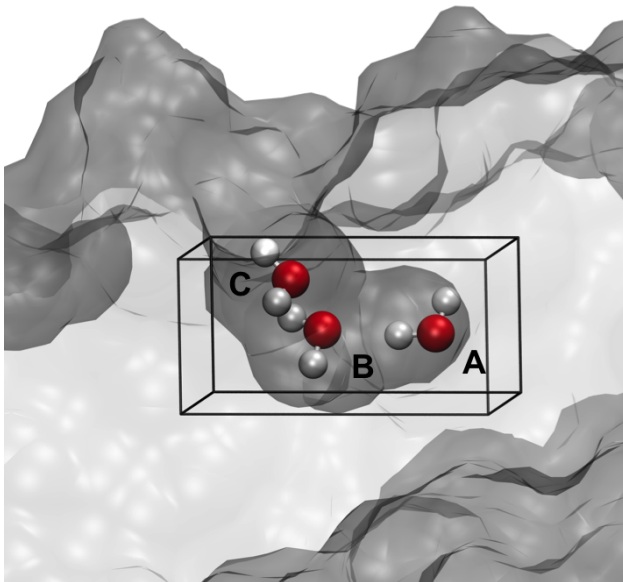


Figure 1: Cavity in the BPTI protein, with waters A, B and C bound. The surface of the protein is shown with a transparent grey surface. The GCMC box covering the water sites is delineated with black lines.

GCMC

In ProtoMS, GCMC insertion and deletion moves are only made inside a user-defined, cuboidal subvolume of the system. Once inserted, water molecules can translate and rotate. Insertion, deletion and translation/rotation moves for GCMC are attempted with

equal probability. To keep track of the number of waters inside the GCMC subvolume, a hard-wall constraint is applied to inserted water molecules to prevent them from drifting away, and bulk solvent water molecules are not allowed to enter from outside the subvolume.

Water molecules were inserted and deleted within the box over the three-water cavity shown in Figure 1. The box had dimensions $5.0 \times 4.0 \times 8.0 \text{ \AA}^3$, and an origin of (29.0, 5.0, -2.0) \AA in the reference frame of the PDB structure. Titration simulations were run with integer B -values between -31.0 and 0.0, inclusive. For equilibration, 1 million (M) GCMC-only moves were performed with the rest of the system fixed, followed by 1M moves of GCMC and protein and solvent configuration sampling, where each repeat was equilibrated independently. The production simulations comprised 10 repeats of 100M moves. When fully sampling the whole system, protein, bulk solvent, and grand canonical insertion, deletion and translation/rotation sampling were trialled with the ratios 461:39:167:167:167, respectively.

To assess the impact of replica-exchange on the consistency of the GCMC results, simulations were performed without replica-exchange, and with replica-exchange with B -value exchanges attempted with randomly selected nearest neighbours every 100,000, 200,000, 500,000 and 1M MC moves.

Double Decoupling

For each water location found with GCMC, decoupling simulations were performed to determine the binding free energy of each water. Decoupling was performed over 16 alchemical λ states, where the LJ and Coulombic terms were scaled simultaneously. Moves were split between protein, bulk water and decoupled water at a ratio of 402:98:1 respectively. The water molecules were decoupled sequentially, from weakest to strongest bound. Where the free energies of multiple waters are similar, calculations were repeated with a different order of decouplings. 500,000 equilibration and 40M production moves were performed for each water at each λ value. Each simulation was repeated four times. Soft-cores (soft66 in ProtoMS package)^{47–49} were used for decoupling calculation with $\delta=0.2$ and $\delta_c=2.0$ used for

the decoupled water molecule. The free energy to decouple the water from the system was determined using MBAR.⁵⁰

A harmonic restraint with a force constant of 2 kcal mol⁻¹ Å⁻² was used on the oxygen of the water being decoupled at all λ values. A gas phase correction of

$$\Delta G_{rest}^{gas} = k_B T \ln \left(\frac{V_{sim}}{V_o} \right) \quad (8)$$

where

$$V_{sim} = \left(\frac{2\pi k_B T}{k} \right)^{\frac{3}{2}} \quad (9)$$

was applied to account for the removal of the restraint from the decoupled system.⁵¹ This is analogous to the volume term introduced in the GCI equation, Equation 5. Prompted by the higher precision obtained in RE-GCMC and unlike our previous study,³² the free energy penalty of applying the harmonic restraint in the bound simulation was calculated using Bennett’s Acceptance Ratio method from 40,000 Monte Carlo simulations steps with six equally spaced λ values of the restraint—from 0 kcal mol⁻¹ Å⁻² to 2 kcal mol⁻¹ Å⁻². No symmetry correction was applied to water molecules. More details on the location of the restraints and the resulting free energy corrections are available in the SI.

Results

Replica-exchange with BPTI

GCMC calculations were performed on BPTI with and without replica-exchange in B . The titration results from 10 repeats with no replica-exchange and the most frequent replica-exchange (attempted every 100,000 MC steps) are shown in Figure 2. The attempted exchanges between B replicas was greater than 89% for all RE frequencies. More details of this are available in the SI.

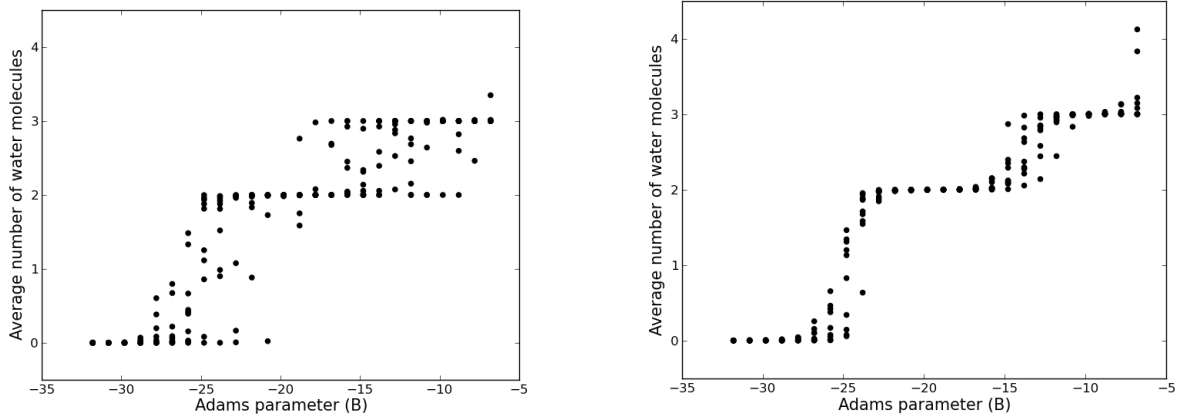


Figure 2: GCMC titration data for BPTI system, without replica-exchange (left) in B and replica-exchange every 100,000 steps (right). Each point corresponds to the average number of water molecules at a given B value. The first 200,000 MC steps have been excluded as equilibration.

Figure 2 shows that replica-exchange has improved the repeatability of the titration data. The Kolmogorov-Smirnov⁵² test was used to test the statistical significance of this apparent reduction in uncertainty. All values of N from simulations with and without replica-exchange were median centered to focus only on the spread of the N s and not the relationship with B . The distributions of median centered N with and without replica-exchange were found to be significantly different with a p-value of 0.0135. The relationship between B and N should be monotonically increasing owing to the relationship between B value and average water occupancy (see equations 1 and 2). The monotonicity of each individual titration repeat, therefore, provides an additional quantifiable test of sampling performance. The Kendall rank correlation coefficient, denoted τ , is an appropriate measure of the monotonicity as it does not require a linear relationship between B and N . A perfect monotonically increasing relationship between B and N is indicated by $\tau = 1$. The mean τ with one standard error for the non-replica-exchange titration data is 0.86 ± 0.01 , compared to 0.98 ± 0.00 for RE-GCMC, averaged over all replica-exchange frequencies. Hence, the titration data from RE-GCMC is more physically sound than without replica-exchange.

Using equation 5, the titration data for all the BPTI simulations were used to calculate binding free energies for each replica-exchange frequency. As the purpose is to evaluate

the extent to which the uncertainty in the GCI free energy decreases with replica-exchange, Figure 3 shows the boxplots of the median-centered binding free energy for one, two and three waters, whose locations can be seen in Figure 1. The boxplots were generated by bootstrap sampling the titration data and calculating the GCI binding free energy of each sample. A bootstrap sample consisted of one randomly sampled N value from the set of 10 repeats for each of the 32 B values and the titration curve was estimated as previously described.³² Both the interquartile ranges and the range of the data (illustrated by the ‘whiskers’) are larger for the protocols with no replica-exchange. The statistical significance of this reduction in spread with RE was assessed using the bootstrap samples of the free energy to bind three waters. The variance of the binding free energy was calculated in batches of 10 bootstrap samples for all replica-exchange data and separately for free energies calculated without replica-exchange. Using these samples, we estimate that the probability that the variance with replica-exchange is equal to or greater than the variance without replica-exchange to be 2.6%. Thus, the reduction in the variance of calculated free energies is statistically significant.

This result is consistent both with the increased monotonicity and the reduced titration variance found for RE-GCMC. Therefore, RE-GCMC improves the repeatability of individual titration simulations. There is no clear replica-exchange frequency that is significantly better than the rest (see Figure S.5).

Figure 4 compares the binding free energies calculated for the three-water BPTI network using GCI and double decoupling. The standard deviation over the repeats is used to illustrate the relative uncertainty of each method. The distribution of the error of original GCI binding free energy equation³² without replica-exchange overlaps with the distribution of error with the new standard state formulation (equation 5) and double decoupling. With the benefit of the higher precision granted by replica-exchange, the standard state binding free energy formulation of GCI is more clearly in agreement with the standard state binding free energies as calculated via double decoupling. Indeed, the standard deviation of

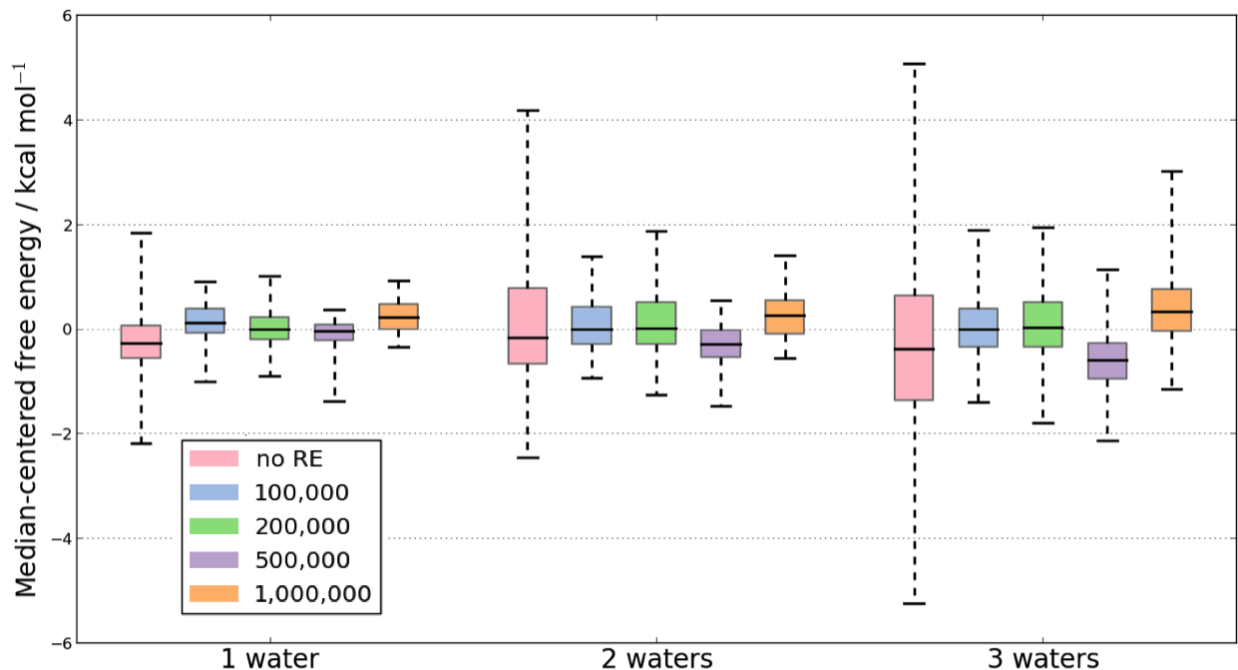


Figure 3: Boxplot of the median-centered free energies for each protocol, where errors have been calculated over 1000 bootstrapping steps of 10 repeats. In each case it is the free energy difference between an empty GCMC region, to a one, two and three water network, respectively. Replica-exchange with GCI produces free energies that have a consistently tighter distribution than GCI free energies calculated without replica-exchange.

GCI is comparable to, if not smaller than double decoupling. In addition to this, the GCI method is easier to use than double decoupling, as all of the water binding free energies are determined in a single simulation. GCMC does not require the a priori knowledge of both the number of waters and the location of those waters that is needed for double decoupling simulations. Double decoupling methods also require restraints or constraints within the simulations. Both restraints and constraints are needed to reduce the sampling of the water molecule as it is decoupled from the system, but constraints are also able to prevent other nearby water molecules occupying the site. GCMC is a significantly easier method to implement, and with the addition of RE, is able to provide results with comparable variance.

The detailed breakdown of the BPTI thermodynamic cycle is shown in Figure 5. The

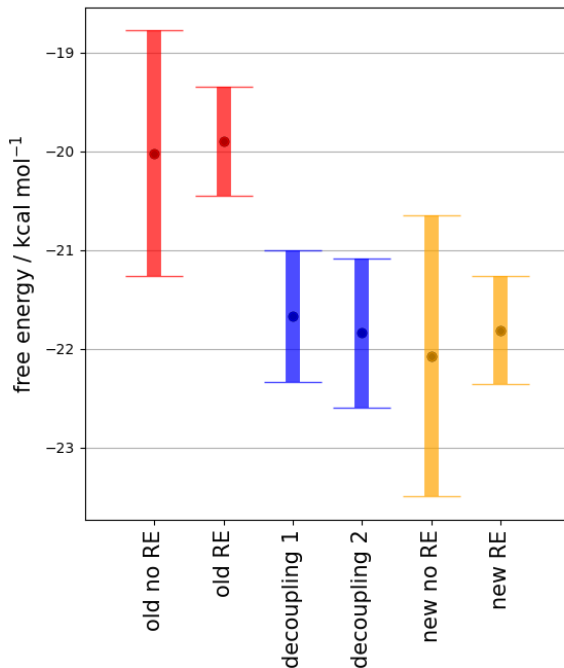


Figure 4: The binding free energies of the three water network in BPTI calculated using different methods. To highlight the intrinsic uncertainty of each method, the colored bars indicate the standard deviation, as opposed to the standard error, over all repeats. Blue results are from sequential decoupling of the water network, with *decoupling 1* waters are decoupled in the order C-B-A, and *decoupling 2* in the order C-A-B. Red results show the binding free energies calculated with the original formulation of GCI binding free energy equation for both RE (frequency: 100,000) and non-RE simulations, and the orange results are from the same simulations determined using Equation 5.

binding free energies of the water molecules in the BPTI system were determined by decoupling calculations. As Figure 2 shows, two water molecules appear to couple into the BPTI system simultaneously (waters A and B), followed by a third molecule at a higher chemical potential (water C). As waters A and B couple in simultaneously, the decoupling of these water molecules has been attempted in both orders (i.e. A then B, and B then A) with double decoupling. The results are shown in figure 5, and the closures of the thermodynamic cycles (shown in red) are consistent to within error. The free energy of the three water network is -21.81 (0.50), -21.67 (0.68) and -21.84 (0.74) kcal mol⁻¹ as calculated by GCI with RE-GCMC and the two double decoupling pathways. As GCI couples the first

Discussion and Conclusions

This work has used replica-exchange of the chemical potential (equivalently, Adams parameter) in grand canonical Monte Carlo (GCMC) to calculate the absolute binding free energies of small networks of water molecules to protein cavities. The use of replica-exchange was found to significantly reduce the variance of the water occupancies at each applied chemical potential, which resulted in a significant reduction in variances of the absolute binding free energies as calculated with grand canonical integration (GCI). The decreased variance of the GCMC simulations facilitated the precise comparison of water binding free energies calculated using double decoupling and the standard state GCI binding free energy equation developed herein. The free energies calculated from both methods were found to be in agreement.

In this study, the absolute binding free energies calculated from GCMC (using GCI) have lower statistical uncertainties than the free energies calculated via alchemical decoupling (estimated using MBAR). This difference becomes more pronounced the more water molecules are decoupled in spite of the fact that, in our hands, free energy calculations with GCMC require fewer CPU hours than alchemical decoupling in BPTI and simpler applications of constraints/restraints. For binding free energy calculations with greater numbers of water molecules, we expect the gap between accuracy and performance to widen between GCMC and alchemical methods.

The extent to which the precision and speed of the GCMC binding free energy methods carries over from water molecules to small, drug-like molecules remains to be seen. The use of GCMC with unbiased insertion moves, as used here for water, becomes increasingly inefficient for larger molecules due to higher probabilities of steric clashes. Nevertheless, GCMC has already been successfully applied to small fragments, both with unbiased insertion moves⁵³ and using cavity bias.⁵⁴ The techniques developed here, therefore, may hold promise for the rigorous calculation of absolute binding free energies of small fragments via GCMC. Larger molecules will no doubt require the application of configurational biasing techniques^{33,35,55}

although these do come with the additional computational effort required to compute the bias. The use of replica-exchange for absolute free energy calculations with GCI comes with a negligible increase in the computational burden, as the exchange of information across CPU nodes is orders of magnitude faster than the run-time of the molecular simulation.

A curious feature of GCMC simulations is the requirement that the volume must remain constant. This is so that the simulations correctly sample the grand canonical (μVT) ensemble. As protein-ligand binding affinities are typically measured under constant pressure conditions, a natural question is to what extent GCMC-like techniques can be used in constant-pressure simulations, such as those in the isothermal-isochoric (NpT) ensemble. Despite the fact that the NpT -type ensembles require the number of particles to remain constant, there are a number of Markov chain Monte Carlo techniques that have insertion and deletion moves, in addition to volume moves, as part of their sampling repertoire. For instance, the Gibbs-ensemble Monte Carlo method^{56,57} attempts to replicate the physically impossible μpT ensemble by allowing two simulated compartments to exchange volume and molecules or particles. Although the volume and number of particles in either compartment can fluctuate, the combined volume and the total number of particles must remain constant. Other techniques employ the semi-grand canonical ensemble (see, for example⁵⁸ or⁵⁹), in which the composition of the system can change (i.e. allowing different fragments to interconvert) under the constraint that the total number of particles remains constant. The semi-grand canonical ensemble also requires the *difference* in the chemical potentials of the interconverting molecules to remain constant so that, when there are only two interconverting molecular species, the semi-grand canonical ensemble can be described as a $NpT\Delta\mu$ ensemble. In controlling the ratio of two species, the $\Delta\mu$ parameter operates—in practice—very much like the μ parameter in GCMC, so that semi-grand canonical simulations are amenable to enhanced sampling via the replica-exchange method described herein.

Absolute binding free energy calculations with GCMC can calculate both the location and binding free energy of networks of waters as well as the relative stability of the individual

waters. The methods developed here reduce both the bias and variance of the predictions made with GCI compared to its original implementation. Coupled with its ease of use compared to double decoupling, we hope GCI will be a powerful tool in structure-based design, particularly in cases where one seeks to rigorously quantify the effect of structural waters on ligand binding affinities.

Acknowledgement

The authors thank the EPSRC and NIH for funding. GAR is supported by the Memorial Sloan Kettering Cancer Center, NIH grant P30 CA008748 and EPSRC grant EP/J010189/1. HBM is supported by the EPSRC funded CDT in Theory and Modelling in Chemical Sciences, under grant EP/L015722/1. CCA is supported by an EPSRC Doctoral Training Centre grant EP/G03690X/1. JWE holds a Royal Society Wolfson Research Merit Award. We acknowledge iSolution at the University of Southampton for computing time on the Iridis4 cluster.

Supporting Information Available

The Supplementary Information contains the derivation of the replica-exchange criterion (equation 4), the derivation of standard state grand canonical integration binding free energy expression (equation 5), and the derivation of the equilibrium condition in equation 6. Further investigation and analysis of the new standard state grand canonical integration equation was carried out on scytalone dehydratase test systems with the methods and results fully described. The data show how the standard state binding free energy equation does not depend on the volume of the GCMC region as expected, and that the calculated free energies agree with double decoupling calculations to a high degree of precision. Additional information is also provided on the methods, analysis, and results of the bovine pancreatic trypsin inhibitor simulations. This information includes the acceptance rates for GCMC insertion and deletion moves, acceptance rates for replica-exchange moves, the degree to which

replica-exchange reduces the variance of estimated water occupancies, as well as details on the GCMC subvolumes and double decoupling restraints that were used. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- (1) Gathiaka, S.; Liu, S.; Chiu, M.; Yang, H.; Stuckey, J.; Kang, Y.; Delproposto, J.; Kubish, G.; Dunbar, J.; Carlson, H.; Burley, S.; Walters,; Amaro, R.; Feher, V.; Gilson, M. D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions. *J. Comp. Aided Mol. Des.* **2016**, *30*, 651–668.
- (2) Yin, J.; Henriksen, N.; Slochower, D.; Shirts, M.; Chiu, M.; Mobley, D.; Gilson, M. Overview of the SAMPL5 host-guest challenge: Are we doing better? *J. Comp. Aided Mol. Des.* **2017**, *31*, 1–19.
- (3) Mikulskis, P.; Cioloboc, D.; Andrejić, M.; Khare, S.; Brorsson, J.; Genheden, S.; Mata, R.; Söderhjelm, P.; Ryde, U. Free-energy perturbation and quantum mechanical study of SAMPL4 octa-acid host-guest binding energies. **2014**, *28*, 375–400.
- (4) Michel, J.; Foloppe, N.; Essex, J. W. Rigorous Free Energy Calculations in Structure-Based Drug Design. *Mol. Inf.* **2010**, *29*, 570–578.
- (5) Lin, Y.-L.; Meng, Y.; Huang, L.; Roux, B. Computational Study of Gleevec and G6G Reveals Molecular Determinants of Kinase Inhibitor Selectivity. *J. Am. Chem. Soc.* **2014**, *136*, 14753–14762.
- (6) Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Predictions of Ligand Selectivity from Absolute Binding Free Energy Calculations. *J. Am. Chem. Soc.* **2017**, *139*, 946–957.
- (7) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist’s Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53*, 5061–5084.

- (8) Huggins, D. J.; Sherman, W.; Tidor, B. Rational Approaches to Improving Selectivity in Drug Design. *J. Med. Chem.* **2012**, *55*, 1424–1444.
- (9) Lam, P. Y.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N. Rational Design of Potent, Bioavailable, Nonpeptide Cyclic Ureas as HIV Protease Inhibitors. *Science* **1994**, *263*, 380–384.
- (10) Chen, J. M.; Xu, S. L.; Wawrzak, Z.; Basarab, G. S.; Jordan, D. B. Structure-Based Design of Potent Inhibitors of Scytalone Dehydratase: Displacement of a Water Molecule from the Active Site. *Biochemistry* **1998**, *37*, 17735–17744.
- (11) Biela, A.; Sielaff, F.; Terwesten, F.; Heine, A.; Steinmetzer, T.; Klebe, G. Ligand Binding Stepwise Disrupts Water Network in Thrombin: Enthalpic and Entropic Changes Reveal Classical Hydrophobic Effect. *J. Med. Chem.* **2012**,
- (12) Vollmuth, F.; Geyer, M. Interaction of Propionylated and Butyrylated Histone H3 Lysine Marks with Brd4 Bromodomains. *Angew. Chem. Int. Ed.* **2010**, *49*, 6768–6772.
- (13) Kadirvelraj, R.; Foley, B. L.; Dyekjær, J. D.; Woods, R. J. Involvement of Water in Carbohydrate-Protein Binding: Concanavalin A Revisited. *J. Am. Chem. Soc.* **2008**, *130*, 16933–16942.
- (14) Mikol, V.; Papageorgiou, C.; Borer, X. The Role of Water Molecules in the Structure-Based Design of (5-Hydroxynorvaline)-2-cyclosporin: Synthesis, Biological Activity, and Crystallographic Analysis with Cyclophilin A. *J. Med. Chem.* **1995**, *38*, 3361–3367.
- (15) Nasief, N. N.; Tan, H.; Kong, J.; Hangauer, D. Water Mediated Ligand Functional Group Cooperativity: The Contribution of a Methyl Group to Binding Affinity is Enhanced by a COO- Group Through Changes in the Structure and Thermodynamics

- of the Hydration Waters of Ligand-Thermolysin Complexes. *J. Med. Chem.* **2012**, *55*, 8283–8302.
- (16) 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.
- (17) Hamelberg, D.; McCammon, J. A. Standard Free Energy of Releasing a Localized Water Molecule from the Binding Pockets of Proteins: Double-Decoupling Method. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.
- (18) García-Sosa, A. T.; Mancera, R. L. Free Energy Calculations of Mutations Involving a Tightly Bound Water Molecule and Ligand Substitutions in a Ligand-Protein Complex. *Mol. Inf.* **2010**, *29*, 589–600.
- (19) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization. *J. Am. Chem. Soc.* **2009**, *131*, 15403–15411.
- (20) Bodnarchuk, M. S.; Viner, R.; Michel, J.; Essex, J. W. Strategies to Calculate Water Binding Free Energies in Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2014**, *54*, 1623–1633.
- (21) Davis, A. M.; Teague, S. J.; Kleywegt, G. J. Application and Limitations of X-ray Crystallographic Data in Structure-based Ligand and Drug Design. *Angew. Chem. Int. Ed.* **2003**, *42*, 2718–2736.
- (22) Carugo, O.; Bordo, D. How Many Water Molecules can be Detected by Protein Crystallography? *Acta Crystallogr. Sect. D* **1999**, *55*, 479–483.
- (23) Wlodawer, A.; Minor, W.; Dauter, Z.; Jaskolski, M. Protein Crystallography for Non-crystallographers, or how to get the Best (but not More) from Published Macromolecular Structures. *FEBS J.* **2008**, *275*, 1–21.

- (24) Nittinger, E.; Schneider, N.; Lange, G.; Rarey, M. Evidence of Water Molecules - A Statistical Evaluation of Water Molecules Based on Electron Density. *J. Chem. Inf. Model.* **2015**, *55*, 771–783, PMID: 25742501.
- (25) Yu, B.; Blaber, M.; Gronenborn, A. M.; Clore, G. M.; Caspar, D. L. D. Disordered Water Within a Hydrophobic Protein Cavity Visualized by X-ray Crystallography. *Proc. Nat. Acad. Sci.* **1999**, *96*, 103–108.
- (26) Nucci, N. V.; Pometun, M. S.; Wand, A. J. Site-resolved Measurement of Water-Protein Interactions by Solution NMR. *Nat. Struct. Mol. Biol.* **2011**, *18*, 245–249.
- (27) Ross, G. A.; Morris, G. M.; Biggin, P. C. Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites. *PLOS ONE* **2012**, *7*, 1–13.
- (28) Shaw, D. E. et al. Anton 2: Raising the Bar for Performance and Programmability in a Special-purpose Molecular Dynamics Supercomputer. Piscataway, NJ, USA, 2014; pp 41–53.
- (29) Stone, J. E.; Hardy, D. J.; Ufimtsev, I. S.; Schulten, K. GPU-accelerated Molecular Modeling Coming of Age. *J. Mol. Graph. Model.* **2010**, *29*, 116–125.
- (30) Denisov, V. P.; Halle, B.; Peters, J.; Hörlein, H. D. Residence Times of the Buried Water Molecules in Bovine Pancreatic Trypsin Inhibitor and its G36S Mutant. *Biochemistry* **1995**, *34*, 9046–9051.
- (31) Ernst, J. A.; Clubb, R. T.; Zhou, H. X.; Gronenborn, A. M.; Clore, G. M. Demonstration of Positionally Disordered Water Within a Protein Hydrophobic Cavity by NMR. *Science* **1995**, *267*, 1813–1817.
- (32) 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.

- (33) Mezei, M. A cavity-biased (T , V , μ) Monte Carlo Method for the Computer Simulation of Fluids. *Mol. Phys.* **1980**, *40*, 901–906.
- (34) Woo, H.-J.; Dinner, A. R.; Roux, B. Grand canonical Monte Carlo Simulations of Water in Protein Environments. *J. Chem. Phys.* **2004**, *121*, 6392–6400.
- (35) Shelley, J. C.; Patey, G. N. A configuration bias Monte Carlo method for water. *J. Chem. Phys.* **1995**, *102*, 7656–7663.
- (36) Earl, D. J.; Deem, M. W. Parallel tempering: Theory, Applications, and New Perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910–3916.
- (37) Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional Replica-Exchange Method for Free-Energy Calculations. *J. Chem. Phys.* **2000**, *113*, 6042–6051.
- (38) Woods, C. J.; Essex, J. W.; King, M. A. The Development of Replica-Exchange-Based Free-Energy Methods. *J. Phys. Chem. B* **2003**, *107*, 13703–13710.
- (39) Woods, C. J.; Essex, J. W.; King, M. A. Enhanced Configurational Sampling in Binding Free-Energy Calculations. *J. Phys. Chem. B* **2003**, *107*, 13711–13718.
- (40) Itoh, S. G.; Damjanović, A.; Brooks, B. R. pH Replica-Exchange Method Based on Discrete Protonation States. *Proteins* **2011**, *79*, 3420–3436.
- (41) Adams, D. J. Chemical Potential of Hard-Sphere Fluids by Monte Carlo Methods. *Mol. Phys.* **1974**, *28*, 1241–1252.
- (42) Guarnieri, F.; Mezei, M. Simulated Annealing of Chemical Potential: A General Procedure for Locating Bound Waters. Application to the Study of the Differential Hydration Propensities of the Major and Minor Grooves of DNA. *J. Am. Chem. Soc.* **1996**, *118*, 8493–8494.
- (43) Olano, L. R.; Rick, S. W. Hydration Free Energies and Entropies for Water in Protein Interiors. *J. Am. Chem. Soc.* **2004**, *126*, 7991–8000.

- (44) Bodnarchuk, M.; Bradshaw, R.; Cave-Ayland, C.; Genheden, S.; Cabedo-Martinez, A.; Michel, J.; Ross, G.; Woods, C. ProtoMS. 2017; protoms.org.
- (45) 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.
- (46) 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.
- (47) Özal, T. A.; Peter, C.; Hess, B.; van der Vegt, N. F. A. Modeling Solubilities of Additives in Polymer Microstructures: Single-Step Perturbation Method Based on a Soft-Cavity Reference State. *Macromolecules* **2008**, *41*, 5055–5061.
- (48) Steinbrecher, T.; Mobley, D. L.; Case, D. A. Nonlinear Scaling Schemes for Lennard-Jones Interactions in Free Energy Calculations. *J. Chem. Phys.* **2007**, *127*, 214108.
- (49) Steinbrecher, T.; Joung, I.; Case, D. A. Soft-core Potentials in Thermodynamic Integration: Comparing One- and Two-step Transformations. *J. Comp. Chem.* **2011**, *32*, 3253–3263.
- (50) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105+.
- (51) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.* **1997**, *72*, 1047–1069.
- (52) Smirnov, N. Table for Estimating the Goodness of Fit of Empirical Distributions. *Ann. Math. Statist.* **1948**, *19*, 279–281.

- (53) Lakkaraju, S. K.; Raman, E. P.; Yu, W.; MacKerell, A. D. Sampling of Organic Solutes in Aqueous and Heterogeneous Environments Using Oscillating Excess Chemical Potentials in Grand Canonical-like Monte Carlo-Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 2281–2290.
- (54) Clark, M.; Guarnieri, F.; Shkurko, I.; Wiseman, J. Grand Canonical Monte Carlo Simulation of Ligand-Protein Binding. *J. Chem. Inf. Model.* **2006**, *46*, 231–242.
- (55) Smit, B. Grand Canonical Monte Carlo Simulations of Chain Molecules: Adsorption Isotherms of Alkanes in Zeolites. *Mol. Phys.* **1995**, *85*, 153–172.
- (56) Panagiotopoulos, A. Z. Direct Determination of Phase Coexistence Properties of Fluids by Monte Carlo Simulation in a New Ensemble. *Mol. Phys.* **1987**, *61*, 813–826.
- (57) Panagiotopoulos, A. Z.; Quirke, N.; Stapleton, M.; Tildesley, D. J. Phase Equilibria by Simulation in the Gibbs Ensemble. *Mol. Phys.* **1988**, *63*, 527–545.
- (58) Bernardin, F. E.; Rutledge, G. C. Semi-Grand Canonical Monte Carlo (SGMC) Simulations to Interpret Experimental Data on Processed Polymer Melts and Glasses. *Macromolecules* **2007**, *40*, 4691–4702.
- (59) Moučka, F.; Lísal, M.; Škvor, J.; Jirsák, J.; Nezbeda, I.; Smith, W. R. Molecular Simulation of Aqueous Electrolyte Solubility. 2. Osmotic Ensemble Monte Carlo Methodology for Free Energy and Solubility Calculations and Application to NaCl. *J. Phys. Chem. B* **2011**, *115*, 7849–7861.

Graphical TOC Entry

