Supplementary Information for 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

Replica Exchange with grand canonical Monte Carlo

In the grand canonical ensemble, the equilibrium probability density for a system composed of N identical molecules with configuration $\mathbf{r}$ at a chemical potential μ, volume V, and inverse temperature β, is given by[1]

$$\pi(\mathbf{r}, N|\mu, V, \beta) = \frac{1}{\Xi(\mu, V, \beta)} \frac{V^N}{\Lambda^{3N} N!} \exp\left(\beta\mu N - \beta U(\mathbf{r})\right), \tag{S.1}$$

[*]To whom correspondence should be addressed
[†]Memorial Sloan Kettering Cancer Center.
[‡]University of Southampton

where $\Xi(\mu, V, \beta)$ is the grand canonical partition function (the normalization constant), Λ is the thermal wavelength of the molecule, and $U(\mathbf{r})$ is the potential energy of the system. In this work, we consider a super ensemble composed of M independent replica systems with the same volume and temperature but each with a different chemical potential, for which the probability density for a given microstate has the form

$$\pi(\mathbf{r}_1, N_1, \mu_1; ...; \mathbf{r}_M, N_M, \mu_M | V, \beta) \propto \prod_{i=1}^{M} \frac{V^{N_i}}{\Lambda^{3N_i} N_i!} \exp\left(\beta \mu_i N_i - \beta U(\mathbf{r}_i)\right). \tag{S.2}$$

We seek an unbiased Monte Carlo scheme for this expanded ensemble that allows for chemical potentials between replicas to be swapped with the aim of enhancing the sampling in GCMC titration simulations. In the Metropolis-Hastings algorithm, the probability to move from state a to b, with equilibrium probabilities π_a and π_b respectively, is accepted according to the following probability:

$$A(a \to b) = \min\left[1, \frac{\pi_b \, p(b \to a)}{\pi_a \, p(a \to b)}\right], \tag{S.3}$$

where $p(a \to b)$ is the transition probability for going from state a to state b. By attempting moves that have transition probabilities equal to the reverse transition probability (i.e. $p(a \to b) = p(b \to a)$), we need only consider the ratio of equilibrium densities (π_a/π_b) to find the acceptance probability. Attempting exchanges between uniformly selected replica pairs fulfills this requirement. If, in the super ensemble specified by equation S.2, the ith and jth chemical potentials were swapped and all others kept constant, the ratio of equilibrium densities can be shown, after minimal algebra, to be

$$\frac{\pi_b(\mathbf{r}_i, N_i, \mu_j; \mathbf{r}_j, N_j, \mu_i, ...|V, \beta)}{\pi_a(\mathbf{r}_i, \mu_i, N_i; \mathbf{r}_j, N_j, \mu_j, ...|V, \beta)} = \exp\big(\beta(\mu_i - \mu_j)(N_j - N_i)\big). \tag{S.4}$$

The simple relation on the right-hand side is the only function that needs to be evaluated in the acceptance test for a GCMC replica exchange swap. Notably, the right-hand side of the above does not depend on $U(\mathbf{r})$ or V as the proposal does not involve a change of configuration or volume for any of the replicas.

Standard state binding free energies

The aim of this section is to derive an expression for the Gibbs binding free energy for N water molecules to a subvolume of a system of interest, within which water has been sampled with GCMC. In particular, the goal is to derive an expression that can be used with grand canonical integration (GCI). The binding free energy is given by

$$\Delta G_{\text{bind}}(N_i \to N_f) = \Delta G_{\text{sys}}(N_i \to N_f) - \Delta G_{\text{sol}}(N_i \to N_f) \tag{S.5}$$

where $\Delta G_{\text{sys}}(N_i \to N_f)$ is the Gibbs free energy to change the number of water molecules from N_i to N_f in the system of interest and $\Delta G_{\text{sol}}(N_i \to N_f)$ is the Gibbs free energy to change the number of water molecules in bulk water similarly. Although an expression for the binding free energy of water that exploited GCI was previously used by Ross et al.,[2] the expression did not evaluate *standard state* binding free energies. The following analysis derives a standard state binding free energy equation for GCI.

As shown previously in S41,[2] the change in free energy of the solvent is given by

$$\Delta G_{\text{sol}}(N_i \to N_f) = (N_f - N_i)\mu_{\text{sol}} \tag{S.6}$$

where μ_{sol} is the chemical potential of bulk water. See, for instance, McQuarry,[1] for more details on the relationship between chemical potential and Gibbs free energy. A general expression for μ_{sol} is

$$\mu_{\text{sol}} = \mu'_{\text{sol}} + k_B T \ln(\rho_{\text{sol}} \Lambda^3), \tag{S.7}$$

where μ'_{sol} is the excess chemical potential, $k_B T$ is Boltzmann's constant multiplied by temperature, ρ_{sol} is the number density of bulk water, and Λ is the de Broglie thermal wavelength of a water molecule.[3]

The system contribution to the binding free energy, $G_{\text{sys}}(N_i \to N_f)$, will be evaluated using the grand canonical integration (GCI) equation. The GCI equation gives the difference in Helmholtz free energy to transfer water molecules from ideal gas to the system of interest, which is denoted $\Delta F_{\text{trans}}(N_i \to N_f)$ and is defined as

$$\Delta F_{\text{trans}}(N_i \to N_f) = \Delta F_{\text{sys}}(N_i \to N_f) - \Delta F_{\text{ideal}}(N_i \to N_f), \tag{S.8}$$

where $\Delta F_{\text{ideal}}(N_i \to N_f)$ is the Helmholtz free energy to change the number of molecules in ideal gas from N_i to N_f. The above free energies refer to changing the number of molecules in the same fixed volume, denoted V_{sys}. As is common, the approximation $\Delta G_{\text{sys}}(N_i \to N_f) \approx \Delta F_{\text{sys}}(N_i \to N_f)$ will be used. This approximation is often invoked due to the small contribution changes in pressure have on differences of Gibbs free energies under physiological conditions.[3] With this approximation and equations S.6 and S.8, we have

$$\Delta G_{\mathrm{bind}}(N_i \rightarrow N_f) = \Delta F_{\mathrm{trans}}(N_i \rightarrow N_f) + \Delta F_{\mathrm{ideal}}(N_i \rightarrow N_f) - (N_f - N_i)\mu_{\mathrm{sol}}. \tag{S.9}$$

Owing to the explicit inclusion of the ideal gas free energy, the above differs from the expression for the binding free energy considered in equation 5 (and S44 in the Supplementary Information) in reference,[2] where it was implicitly assumed that $F_{\mathrm{ideal}}(N_i \rightarrow N_f) = 0$. That assumption is not made in the following analysis.

As described previously,[2] $\Delta F_{\mathrm{trans}}(N_i \rightarrow N_f)$ can be calculated by sampling water at a range of different chemical potentials, or, equivalently, Adams values and evaluating

$$\Delta F_{\mathrm{trans}}(N_i \rightarrow N_f) = k_B T \left[N_f B_f - N_i B_i + \ln\left(\frac{N_i!}{N_f!}\right) - \int_{B_i}^{B_f} N(B) dB \right], \tag{S.10}$$

where B_k is the Adams value in which an average of N_k waters are present in V_{sys}. As the average number of water molecules changes with the applied Adams value, N appears as a function of B in the integral on the right-hand side.

The Helmholtz free energy for N ideal gas particles in a volume V has the analytical expression

$$F_{\mathrm{ideal}}(N) = -k_B T \ln\left[\frac{1}{N!}\left(\frac{V}{\Lambda^3}\right)^N\right], \tag{S.11}$$

(see, for instance[1]) such that

$$\Delta F_{\mathrm{ideal}}(N_i \rightarrow N_f) = -k_B T \left[\ln\left(\frac{N_i!}{N_f!}\right) + (N_f - N_i)\ln\left(\frac{V_{\mathrm{sys}}}{\Lambda^3}\right)\right]. \tag{S.12}$$

Using the expression for μ_{sol} (equation S.7), $\Delta F_{\mathrm{trans}}(N_i \rightarrow N_f)$ (equation S.10), and

$\Delta F_{\text{ideal}}(N_i \to N_f)$ (equation S.12), we arrive at

$$\beta\Delta G_{\text{bind}}(N_i \to N_f) = N_f B_f - N_i B_i - (N_f - N_i)\left[\beta\mu'_{\text{sol}} + \ln\left(\rho_{\text{sol}} V_{\text{sys}}\right)\right] - \int_{B_i}^{B_f} N(B)dB, \tag{S.13}$$

where $\beta = 1/k_B T$ has been included for notational simplicity. In contrast to the expression previously presented,[2] the above expression lacks the N factorial terms, and has the extra term $\ln(\rho_{\text{sol}} V_{\text{sys}})$. When the solvent is in the standard state with density ρ^o, the standard state volume of water is defined as $V^o = 1/\rho^o_{\text{sol}}$,[4] so that the standard Gibbs binding free energy is given by

$$\beta\Delta G^o_{\text{bind}}(N_i \to N_f) = N_f B_f - N_i B_i - (N_f - N_i)\left[\beta\mu'_{\text{sol}} + \ln\left(\frac{V_{\text{sys}}}{V^o}\right)\right] - \int_{B_i}^{B_f} N(B)dB. \tag{S.14}$$

Equilibrium in grand canonical Monte Carlo

This section derives the condition for thermodynamic equilibrium for water binding to the system of interest using the above framework. When discussing *thermodynamic* equilibrium, we are obliged to consider water-protein binding in the thermodynamic limit, which occurs when $N \to \infty$ and $V \to \infty$. Although this limit must be taken by necessity, it will allow for some simplifying approximations. The starting point for this derivation is the expression for the Gibbs binding free energy given in equation S.9. Thermodynamic equilibrium is established when the binding free energy for N waters is at a minimum.[5] Denoting $\Delta G_{\text{bind}}(0 \to N)$ as $\Delta G_{\text{bind}}(N)$, we seek an expression that satisfies

$$\frac{d\Delta G_{\text{bind}}(N)}{dN} = 0 \tag{S.15}$$

A useful expression shown for this purpose—disscused by Ross et al.[2]—is that in the thermodynamic limit, the Helmotz free energy to transfer water N molecules from ideal gas to the system of interest is approximately given by

$$\Delta F_{\text{trans}}(N) = \int_0^N \mu'_{\text{sys}}(N) dN, \tag{S.16}$$

where $\mu'_{\text{sys}}(N)$ is the excess potential of the system of interest and $\Delta F_{\text{trans}}(N) = \Delta F_{\text{trans}}(0 \rightarrow N)$. (Section 2.3 of the Supplementary Information in reference[2] shows that equation S.16 is indeed a large N approximation to equation S.10.) With this and equations S.9 and S.11, the binding free energy is given by

$$\Delta G_{\text{bind}}(N) = \int_0^N \mu'_{\text{sys}}(N) dN - k_B T \ln \left[\frac{1}{N!} \left(\frac{V}{\Lambda^3} \right)^N \right] - N\mu_{\text{sol}}. \tag{S.17}$$

To evaluate equation S.15, we must differentiate the above expression, which is hindered by the presence of the factorial term. As we are concerned with thermodynamic equilbrium, in which the thermodynamic limit is invoked, we can use Stirling's approximation for $\ln(N!)$:

$$\ln(N!) \approx N \ln(N) - N, \tag{S.18}$$

and whose error decreases as N increases. With the relations immediately above and equation S.7, we have

$$\frac{d\Delta G_{\text{bind}}(N)}{dN} = \mu'_{\text{sys}} - \mu'_{\text{sol}} + k_B T \ln\left(\frac{N}{V\rho_{sol}}\right) \tag{S.19}$$

$$= 0 \tag{S.20}$$

Recognising N/V as the number density of the system of interest, denoted ρ_{sys}, we arrive at

$$\mu'_{\text{sol}} + k_B T \ln(\rho_{\text{sol}}) = \mu'_{\text{sys}} + k_B T \ln(\rho_{\text{sys}}) \tag{S.21}$$

$$\implies \mu'_{\text{sol}} + k_B T \ln(\rho_{\text{sol}}\Lambda^3) = \mu'_{\text{sys}} + k_B T \ln(\rho_{\text{sys}}\Lambda^3) \tag{S.22}$$

which, by equation S.7, is equivalent to stating that the chemical potentials of water in the system of interest and bulk solvent are equal. Previously, Ross et al. derived the equality of the *excess* chemical potentials as the condition for equilibrium,[2] which was due to the omission of the standard state volume correction for binding free energies. However, determining equilibrium via the equality of excess chemical potential will likely result in only a small error, as large difference in the densities are required to significantly affect the determination of equilibrium. For example, at 300 K, $K_B T \ln(\rho_{\text{sys}}/\rho_{\text{sol}})$ contributes roughly 1 kcal/mol for every factor of 6 in the density ratio.

Scytalone Dehydratase

Scytalone dehydratase (SD) in complex with two congeneric ligands was used to quantify the level of precision that could be reached by comparing free energies calculated with GCI with RE-GCMC and double decoupling (DD) calculations in independent simulations. For consistency with a previous study,[6] the ligands are referred to as L1 and L3, Figure S.1. Both ligands have two putative bridging water sites, waters D and E, that have been previ-

ously identified[6] and are shown in Figure S.2.

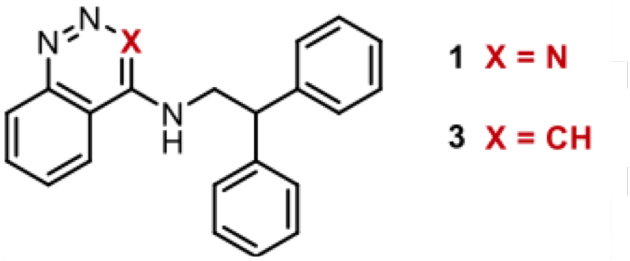


Figure S.1: Structure of L1 and L3, both of which bind to SD

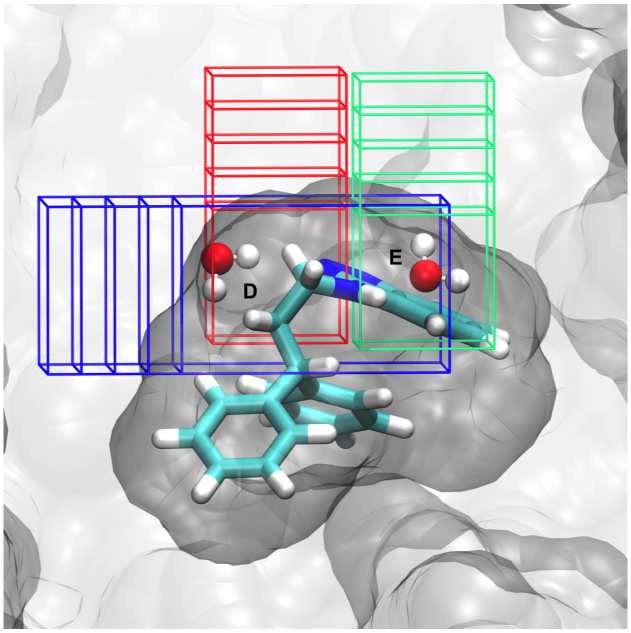


Figure S.2: Ligand L3 bound to SD, with water D and E present. The active site of SD is shown with a transparent grey surface. The incrementally increasing GCMC boxes for each calculation are shown; red - water D, green - water E and blue for the box encompassing both waters. Each box repeatedly increased in 1 Å increments. The increasing volume of the GCMC region covers protein, not accessible to water.

To compare directly the binding free energies calculated with GCI with DD free energies, GCMC simulations with a box encompassing each individual water site were performed for both ligands, and with a box covering both hydration sites simultaneously. To test for any GCMC region size dependence of the new GCI equation (equation S.10), each GCMC simulation was repeated with different box sizes. Each GCMC box was extended along one axis

in 1 Å increments over a 5 Å range into regions of high protein density (details available in Table S.1). As the larger volumes do not increase the number of accessible hydration sites, the free energies calculated with GCI should be the same for each set of GCMC boxes. Simulations were repeated four times for each GCMC box. A diagram of L3 bound to SD, with the water positions D and E as well as the various GCMC boxes used is shown in Figure S.2.

Methodology

The protein and ligand structures were generated from PDB 3STD, where a structurally similar ligand is bound to the protein. To improve the precision of the calculations as much as possible, the protein and ligand configurations of the SD complexes were not sampled and no surrounding bulk water solvent was added to the system. This is therefore a model system. Only the water molecules shown in Figure S.2 were sampled for these simulations. As the protein is not sampled, the protein structure was chosen from a fully sampled GCMC simulation with L1 where both water A and B are bound. A protein scoop of 15 Å was chosen to be consistent with other literature simulations of this protein.[7]

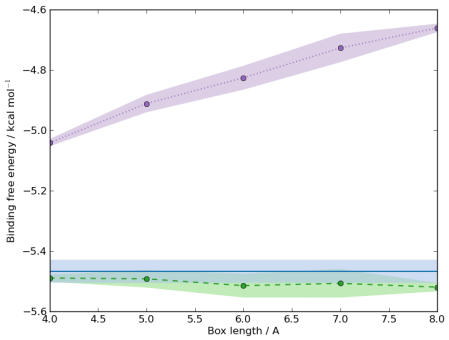
The protocols for GCMC and decoupling simulations are the same as the BPTI system, however with the reduced system sampling described above. A replica exchange frequency of every 100,000 moves was chosen. For the L3 complex, water E has a lower binding free energy than water D, so water E was included for the GCMC simulations of water D. For the L1 complex, the individual-box GCMC simulations were repeated both with and without the other water molecule.

Results

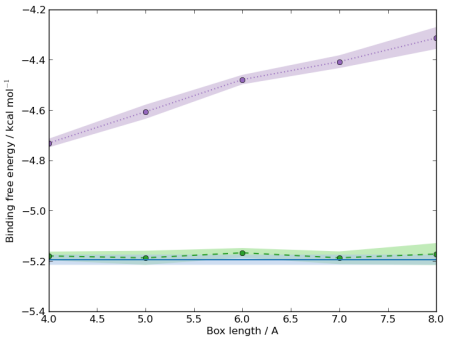
The reduced variance of the RE-GCMC simulations facilitates a precise comparison between the binding free energies calculated with GCI and free energies calculated in separate, alchemical DD simulations. The corrections required for DD method have been included, and are the same as those used fot BPTI in the main text. The new GCI equation (Equation S.14) contains two modifications to the original equation: a new volume correction term, and the removal of the factorial dependence on the number of molecules. As the latter term would have zero effect on the free energy calculated for one water molecule, free energies calculated for single waters will isolate the effect of the volume correction. Therefore, simulations with SD were also used to verify whether the new standard state binding free energy equation (Equation S.14) was not erroneously dependent of the volume of the GCMC box.

Figure S.4 shows the binding free energy for the SD water molecules D and E calculated with RE-GCMC and the standard state GCI binding free energy equation (equation S.14), the original GCI binding free energy equation,[2] and via DD. As can be seen, the corrected standard state GCI binding free energy equation produces binding free energies that are within 0.05 kcal mol^{-1} of the binding free energies computed via alchemical double decoupling. In contrast, the previous, non-standard state formulation of the GCI binding free energy produces energies that are distinct from the decoupling results, with a small but statistically significant volume dependence.

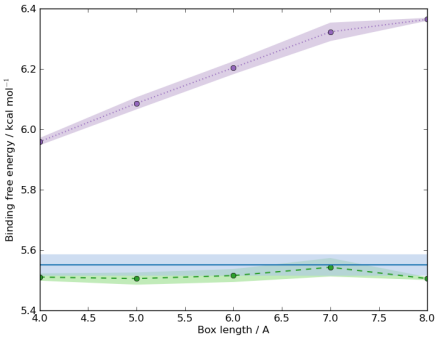
The binding free energy of each two water network has also been calculated in a single step, using a GCMC volume region covering both sites (see Figure S.2). All GCMC binding free energies in Figure S.4 are calculated with the corrected GCI equation, and the values are within error of alchemical DD irrespective of which decoupling pathway is chosen.



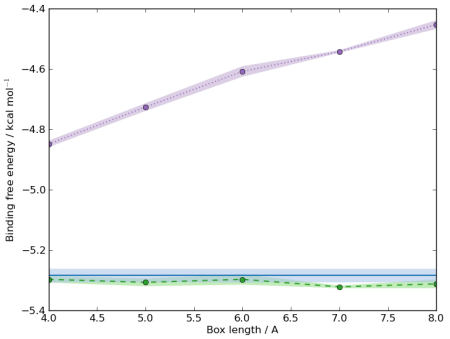
Water D, ligand 1



Water E, ligand 1



Water D, ligand 3



Water E, ligand 3

Figure S.3: Binding free energy of waters in SD. Dotted line (purple) - GCMC results without volume correction, dashed line (green) - GCMC result with volume correction, solid line (blue) - double decoupling result. For each, the shaded region shows the 95% confidence interval calculated from four repeats.

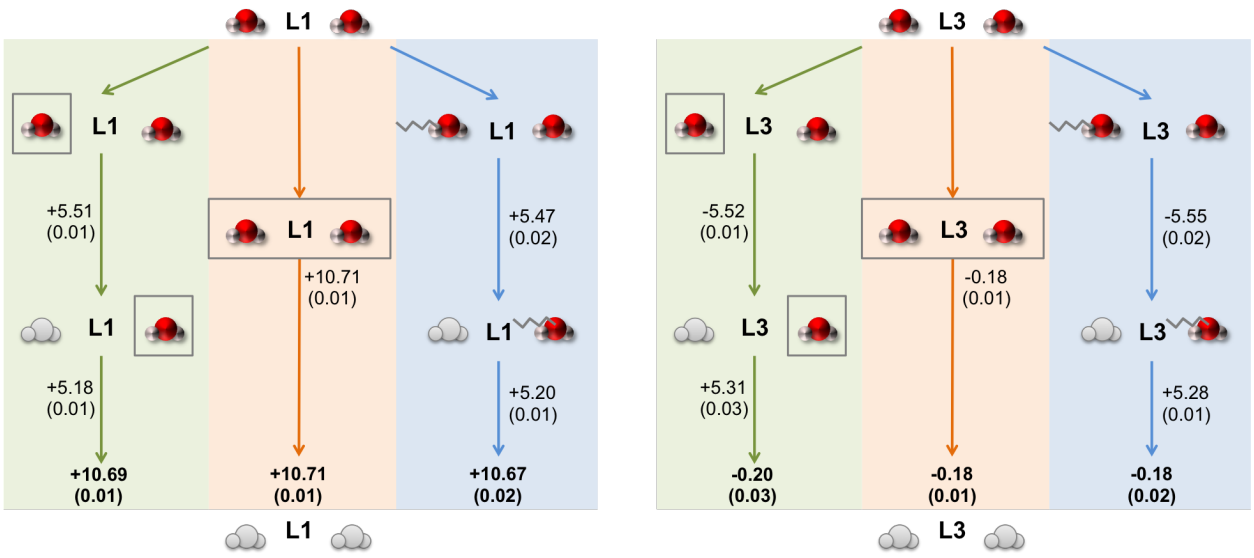


Figure S.4: Thermodynamic cycles of the Scytalone Dehydratase system with L1 and L3. Blue arrows - restraint double decoupling, green arrows - single water GCMC and orange arrows - two water GCMC. Standard errors shown in parenthesis from four repeats. Grey waters represent those removed from the system. Free energies are in kcal mol^{-1}

Bovine pancreatic trypsin inhibitor

Additional methodological details

GCMC subvolumes

Table S.1: Location and dimensions of GCMC boxes. The origin of the box is the coordinates of the lowest corner of the box and dimension shows the length in Å of the box in each dimension. Where a range is provided for the dimension, a series of GCMC calculations have been performed with each box length at 1 Å intervals along the range.

	box origin			box length		
system	x	y	z	x	y	z
one D	24.141	11.225	32.916	4	4 - 8	4
one E	27.913	11.260	28.713	4	4 - 8	4
one both	26.000	10.500	30.000	5	5	8 - 12
three D	24.141	11.225	34.000	4	4 - 8	4
three E	27.9135	11.260	28.713	4	4 -8	4
three both	26.000	10.500	30.000	5	4	8 - 12

Double decoupling restraints

For every water, a restraint of strength k = 2 kcal mol^{-1}Å $^{-2}$ was used. This corresponds to a V_{sim} of 2.54 Å^3 (Equation 9), and therefore a correction of -1.46 kcal mol^{-1} for the gas phase restraint (Equation 8). The method to calculate the bound phase restraint correction is outlined in the main text.

Table S.2: Details of the center of the restraint applied to the oxygen atom of each decoupled water molecule. Calculation of the gas phase restraint is explained above. Calculation of the bound phase restraint was performed using a short simulation, explained in the main text, and repeated four times.

	restraint origin				
system	x	y	z	ΔG^{gas}_{rest}	ΔG^{bound}_{rest}
A	31.705	7.133	1.254	-1.46	+0.33
B	32.184	7.273	4.121	-1.46	+0.10
C	32.310	5.881	0.041	-1.46	+0.64
one D	26.585	13.658	36.700	-1.46	+0.14
one E	29.913	13.260	30.713	-1.46	+0.21
three D	27.520	13.723	36.826	-1.46	+0.10
three E	30.119	13.546	30.417	-1.46	+0.06

Additional results

Replica exchange improves the consistency of titration data

To quantify the consistency of the BPTI titration data with and without replica exchange, the root-mean squared variance of the water occupancy at each B value, averaged over all B values and repeats for a given replica exchange frequency was calculated, excluding the first 200,000 MC steps of each repeat as equilibration. Lower values of this consistency measure indicate which set of repeats have more reliable titration data. Error bars were generated for this measure by sampling with replacement B values for each repeat. Figure S.5 shows that RE-GCMC produces simulations that are significantly more consistent than without replica exchange, irrespective of the frequency with which neighboring replicas were exchanged.

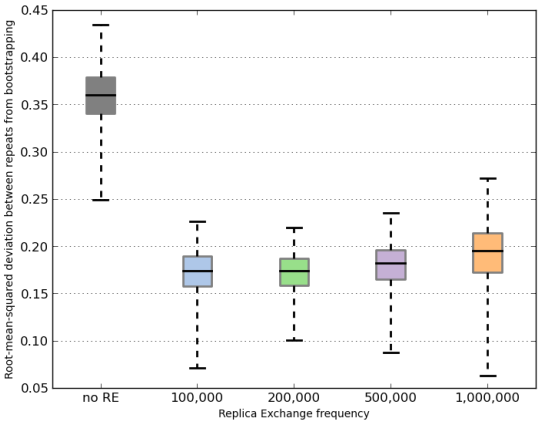


Figure S.5: The root-mean squared variance of the water occupancy at each B value, averaged over all B values and repeats for each replica exchange frequency for BPTI. The box plots show the minimum, first quartile, median, third quartile, and maximum of the data.

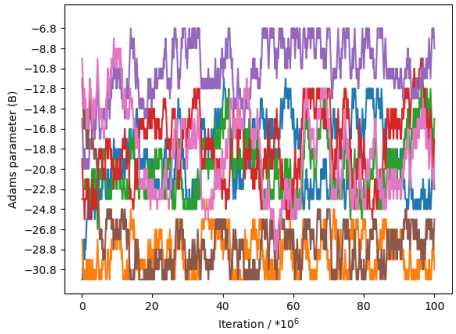
The median centered average water occupancy at each B value for the RE and all of the non-RE data was compared using the Kolmogorov-Smirnov test, which found the distribution of the values to be significantly different with a p-value of 1.3%. The distributions between the data for different RE frequencies were found to have p-values >5%, suggesting the water occupancies at each B value are drawn from the same distribution, irrespective of the RE frequency.

Acceptance rates and replica exchange sampling

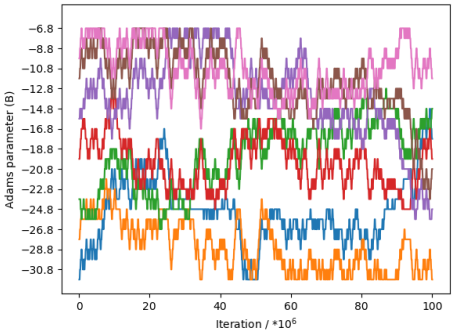
Table S.3: Percentage acceptance rates for Adams parameter (B) exchange moves for each replica exchange (RE) frequency.

RE Frequency	B exchange (%)
100,000	89.9
200,000	89.7
500,000	89.8
1,000,000	89.4

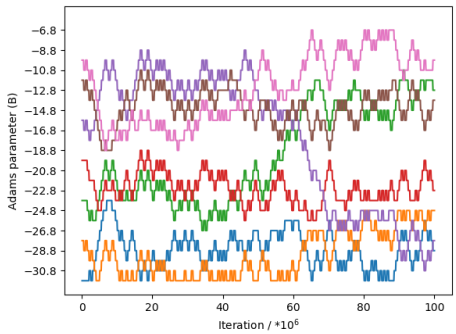
The acceptance rates of the water insertion and deletion moves are 0.002% and 0.003% respectively, on average over all B values and for all RE frequencies.



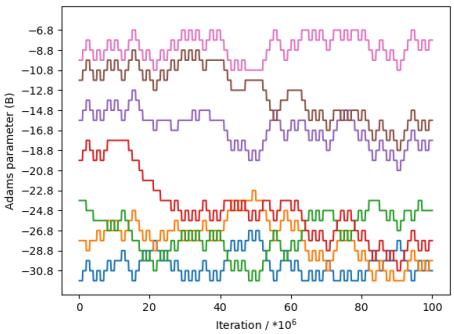
100,000



200,000



500,000



1,000,000

Figure S.6: Pathway of replicas through B space over example titration simulations at various exchange frequencies. Only seven replicas out of twenty-four, which were initially equally spaced in B before equilibration, have been shown for clarity. The water occupancy for B values below -24.8 was 0 for all repeats.

References

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

(2) 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.

(3) Ben-Naim, A.; Marcus, Y. Solvation Thermodynamics of Nonionic Solutes. *J. Chem. Phys.* **1984**, *81*, 2016–2027.

(4) 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.

(5) Atkins, P.; de Paula, J. *Atkins' Physical Chemistry*, 9th ed.; Oxford University Press, 2009.

(6) 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.

(7) 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.