# Supplementary Information for Water sites, networks, and free energies with grand canonical Monte Carlo

Gregory A. Ross, Michael S. Bodnarchuk, Jonathan W. Essex

## 1 Theoretical Background

In the grand canonical ensemble, a system at constant temperature and volume is coupled to a reservoir of a molecular species such that molecules are allowed to exchange between the two. In this work, we will consider situations when the reservoir consists of ideal gas, and when the reservoir consists of bulk water. The chemical potential of a system where only one species can vary, defined as

$$\mu = \left( \frac{\partial F(N,V,T)}{\partial N} \right)_{T,V}, \tag{S1}$$

controls the ratio of the number particles in the system and reservoir. For the purposes of this section, $N$ refers to the instantaneous number of molecules in the system, as opposed to the average number. In the above equation, $F$ is the Helmholtz free energy of the system coupled to the reservoir. For the Unlike the Gibbs ensemble[1], the reservoir in the grand canonical ensemble is not explicitly considered as it is completely defined by the chemical potential. When the system and reservoir are two phases, for example, liquid and gas respectively, chemical equilibrium is established when the chemical potential of the system of interest and reservoir, denoted $\mu_{\text{sys}}$ and $\mu_{\text{res}}$ respectively, are equal.

Equation S1 implies that $N$ is a continuous variable, suggesting that $F$ is valid for non-integer values of $N$. Generally, however, this is *not* the case; Equation S1 is only valid in the thermodynamic limit, that is, when $N \rightarrow \infty$. One can understand this by considering that while a non-integer number of molecules is "alchemical", the more molecules there are, the more continuous the quantity appears to be. Treating $N$ as a continuous variable is, as Kirkwood noted, "a convenient mathematical device"[2].

Free energy can be decomposed into an ideal part, denoted $F_{\text{ideal}}$, and an *excess* part, denoted $F_{\text{ex}}$, so that $F = F_{\text{ideal}} + F_{\text{ex}}$. The excess free energy contains the contribution from potential energy, such as from inter and intra molecular interactions. Similarly, one can define the excess chemical potential as

$$\mu'(N,V,T) = \mu(N,V,T) - \mu_{\text{ideal}}(N,V,T), \tag{S2}$$

where $\mu'$ is the excess chemical potential, and the dependence on $N$, volume $V$ and temperature $T$ has been made explicit. Clearly, $\mu'$ is zero for ideal systems. By utilising a finite difference approximation to equation S1 in the thermodynamic limit (see, for instance, McQuarrie[3]), one can reason that for the system of interest

$$\begin{aligned} \mu' &\approx \frac{\Delta F_{\text{ex}}}{\Delta N_{\text{sys}}} && \text{(S3)} \\ &= F_{\text{ex}}(N_{\text{sys}}+1) - F_{\text{ex}}(N_{\text{sys}}), && \text{(S4)} \end{aligned}$$

where the $N_{\text{sys}}$ is the number of molecules in the system of interest, and $\Delta N_{\text{sys}}$ is set equal to 1. Thus, the excess chemical potential approximates the free energy required to couple one additional molecule to the system from ideal gas when $N_{\text{sys}} \to \infty$. Thus, $\mu'$ is an important quantity in free energy calculations. In a pioneering paper, Widom derived an equation (not stated here) that can be utilised in the determination of $\mu'$ from molecular simulations involving particle insertions[4].

Widom's method formed the foundation of multiple grand canonical methodologies and free energy calculations. For example, Hummer and co-workers have combined Widom's insertion method and Bennett's method of overlapping histograms to calculate the binding free energies of water molecules to non-polar cavities, such as carbon nanotubes and fullerenes[5;6]. The free energy to hydrate a large site could be evaluated using Hummer's method using multiple sequential calculations.

In a seminal paper, Adams laid the foundations for many of the current implementations of GCMC by introducing an "insert" and "delete" move along with the standard translation and rotation move types in Monte Carlo simulations[7]. In Adams' formulation, the acceptance probabilities for inserting a molecule from the reservoir and deleting a particle are, respectively, given by

$$p(N_{\text{sys}} \to N_{\text{sys}} + 1) = \min\left[1, \frac{1}{N_{\text{sys}} + 1} \exp(B) \exp(-\beta \Delta E)\right], \tag{S5}$$

$$p(N_{\text{sys}} \to N_{\text{sys}} - 1) = \min\left[1, N_{\text{sys}} \exp(-B) \exp(-\beta \Delta E)\right], \tag{S6}$$

where $\beta = 1/k_B T$ denotes the inverse temperature with $k_B$ as Boltzmann's constant, $\Delta E$ is change in energy from the trial move, $N_{\text{sys}}$ is the number of water molecules in the system, and the parameter $B$ is known as the Adams parameter. Discussed in more detail below, we motivate $B$—in a manner slightly differently, but equivalently, to Adams—as

$$B = \mu\beta + \ln\left(\frac{V}{\Lambda^3}\right), \tag{S7}$$

where $\Lambda$ is the thermal wavelength of water, and $V$ denotes the volume over which GCMC moves take place. The benefit of using $B$ is that the constants $\mu$, $V$ and $\Lambda$ are absorbed into a single term. Like temperature, $B$ must be set prior to running a GCMC simulation. The parameter $B$ influences the probability that a water is inserted or deleted; fewer waters are found on average at lower values of $B$ that at higher values of $B$. Note that simulating at a constant $B$ guarantees that the simulation is run at a constant $\mu$: the basic requirement for simulating the grand canonical ensemble.

As discussed in Section 2.1, a significant drawback to implementing equations S5 and S6 is that, prior to this work, it has not been clear what $B$ (or $\mu$) should be set by the user to produce the number of water molecules that would be present if the simulated system were placed in physical contact with bulk water. Notably, Guarnieri and Mezei circumvented this issue by, instead of running one simulation at a single $B$, they ran many GCMC simulations at a range of $B$ values[8]. The analogy between $\mu$ and temperature means that varying $B$ can be thought of as performing a type of simulated annealing on the chemical potential. As $B$ is changed from high to low, the number of water molecules in the system decreases, and, as $B$ is a type of energetic potential, the water molecules are present at lower values of $B$ are more strongly bound than those that disappear at higher $B$ values. Thus, the simulated annealing method can rank order molecules by binding affinity[9], although the numeric values of the binding free energies are unknown.

Extra clarity on $B$ can be gained by considering GCMC between liquid and ideal gas phases and applying the equilibrium condition $\mu_{\text{sys}} = \mu_{\text{res}}$ to equation S7. We denote the equilibrium number of molecules of the system of interest and reservoir as $N_{\text{sys,equil}}$ and $N_{\text{res,equil}}$ respectively. The chemical potentials for the system and ideal gas reservoir are, respectively, given by

$$\mu_{\text{sys}} = \mu'_{\text{sys}} + k_B T \ln\left(\frac{N_{\text{sys,equil}}\Lambda^3}{V}\right), \tag{S8}$$

and

$$\mu_{\text{res}} = k_B T \ln\left(\frac{N_{\text{res,equil}}\Lambda^3}{V}\right). \tag{S9}$$

These relations are well known[10]; $\mu_{\text{res}}$ is that of an ideal gas, and describes the kinetic energy contribution to free energy, whereas $\mu_{\text{sys}}$ comprises of both potential (encoded in $\mu'$) and kinetic energy contributions. Combining equation S7 with equations S8 and S9 when the volumes of the system and ideal gas reservoir are equal, we arrive at

$$B = \beta\mu'_{\text{sys}} + \ln(N_{\text{sys,equil}}) \tag{S10}$$

$$\text{and}\quad B = \ln(N_{\text{res,equil}}), \tag{S11}$$

respectively. Equation S11 reveals that $B$ establishes the size of the particle reservoir at equilibrium: the larger the reservoir, the higher the probability of accepting an insertion move and rejecting a deletion move. When simulating the grand canonical ensemble, $N_{\text{sys,equil}}$ fluctuates about its mean value. Accordingly, Adams originally *defined* $B$ to equal $\beta\mu' + \ln\langle N_{\text{sys,equil}}\rangle$[7], where the angular brackets denote an ensemble average. Note that this implies there is a one-to-one relationship between a chosen $B$ and the resulting $\langle N_{\text{sys,equil}}\rangle$.

Selecting which Adams value (or equivalently, chemical potential) to run with GCMC is straightforward when studying phases of pure substances. The situation is more complicated when considering binding processes in a mixed medium. If one knows in advance that there should be $\langle N^*_{\text{sys,equil}}\rangle$ molecules in a system, then one can find the correct $B$ to use, denoted $B^*$, by performing many simulations with different $B$s and selecting the one that yields the $\langle N_{\text{sys,equil}}\rangle$ that is closest to $\langle N^*_{\text{sys,equil}}\rangle$. This has been implemented by Mezei and co-workers in a dCpG-proflavine complex[11]. Mezei and other groups have implemented different iterative simulation procedures to determine $B^*$ from a known $\langle N^*_{\text{sys,equil}}\rangle$[12;13;14].

Returning to the example of GCMC between two phases, it is straightforward to determine coupling free energies[10]. Combining equations S4, S11, and S10 one can find that

$$F_{\text{ex}}(N_{\text{sys}}+1) - F_{\text{ex}}(N_{\text{sys}}) = k_B T \ln\left(\frac{N_{\text{res,equil}}}{N_{\text{sys,equil}}}\right). \tag{S12}$$

Despite appearing valid only for phase-equilibrium and not for water-protein binding, an equivalent relationship to this was used by Clark et al. to predict the binding affinities for fragments in T4 lysozyme[15], and later by Bodnarchuk et al. to predict the binding affinities of individual water molecules[16]. Interestingly, both Clark and Bodnarchuk found that $k_B T \ln\left(\frac{N_{\text{res,equil}}}{N_{\text{sys,equil}}}\right)$ yielded inconsistent values when $B$ was high, and consistent, as well as accurate, values when $B$ was low. These particular findings are explained in the Section 2.6.

In summary, prior to this work, there were two unresolved problems with GCMC. First, the determination of what chemical potential to use to replicate physical contact with bulk water. Second, how GCMC simulations conducted at multiple $B$ parameters can be used to reliably calculate free energies.

# 2 Expanded theoretical results

## 2.1 Chemical potential and water binding

One of the primary aims of running GCMC on water in a buried cavity in protein is to efficiently produce the density that would occur at equilibrium with bulk solvent. While this density can be achieved with an appropriate choice of chemical potential, it is not known *a priori* which value one should choose. Previous GCMC treatments required that the number of bound water molecules was already known, after which a series of sequential or iterative simulations would adjust the chemical potential to reproduce this number[13;12;14]. When the occupancy of a cavity is not known, a choice for the chemical potential is that of bulk water, which is inspired by phase equilibrium for pure substances, where the condition for thermodynamic equilibrium is that the chemical potentials of each phase are equal. However, equal chemical potentials for water in a protein cavity and bulk is *not* the equilibrium condition for water-protein binding. To show why this is so, this section considers the thermodynamics of water binding to a macromolecule. These considerations naturally lead to a derivation of equation 6 of the main text, which is later re-derived in Section 2.5 within the formalism of grand canonical integration (GCI). In contrast to the other sections of the Supplementary Material, the isothermal-isobaric ensemble is used in order to exploit well known physical chemistry definitions.

Consider the reaction for water, $W$, binding to a macromolecule, $M$, to form a water-macromolecule complex, $WM$, in bulk water:

$$\nu W_{\text{sol}} + M_{\text{sol}} \rightleftharpoons WM_{\text{sol}}, \tag{S13}$$

where the subscript sol refers to the fact that the chemical species are in bulk water solvent, and the stoichiometric coefficient, $\nu$, refers to the number of water molecules that are extracted from the bulk solvent to form a single complex with the macromolecule, which is equivalent to the water occupancy. Thermodynamic equilibrium for this reaction occurs when[3]

$$\nu\mu_{W,\text{sol}} + \mu_{M,\text{sol}} = \mu_{WM,\text{sol}}, \tag{S14}$$

where $\mu_{X,sol}$ denotes the chemical potential for species $X$ in water solvent. Not only does the above condition not include the chemical potential for water in the macromolecule – which is what is required to perform GCMC – but the stoichiometry of the reaction is, in general, unknown. We proceed by establishing an equilibrium condition that can be used to deduce the appropriate chemical potential to sample water in a macromolecule using GCMC.

The chemical potential for a species $X$ in solvent can be expressed as

$$\mu_{X,\text{sol}} = \mu^{o}_{X,\text{sol}} + k_B T \ln(a_{X,\text{sol}}), \tag{S15}$$

where $\mu^{o}_{X,\text{sol}}$ is the chemical potential at the standard state and $a_{X,\text{sol}}$ is the activity of the species. The standard Gibbs binding free energy for the water-macromolecule reaction is defined as

$$\begin{aligned}\Delta G^{o}_{\text{bind}}(\nu) &= \mu^{o}_{WM,\text{sol}} - \nu\mu^{o}_{W,\text{sol}} - \mu^{o}_{M,\text{sol}} && \text{(S16)}\\ &= -k_B T \ln\left[\frac{a(\nu)_{WM,\text{sol}}}{a^{\nu}_{W,\text{sol}}\, a_{M,\text{sol}}}\right], && \text{(S17)}\end{aligned}$$

where the dependency of the standard binding free energy and activity of the water-macromolecule complex on $\nu$ has been made explicit. The definition of the standard binding free energy in terms of the activities is used in the following analysis such that standard concentrations are implicit.

Even though the stoichiometry of the binding reaction, $\nu$, is unknown, clarity can be gained by considering what stoichiometry minimizes the standard binding free energy. The condition for the minimum binding free energy can be found by investigating $\frac{\partial \Delta G^{o}_{\text{bind}}(\nu)}{\partial \nu} = 0$. As $\nu$ is a discrete variable, the partial differential will treated with the finite difference approximation

$$\frac{\Delta\Delta G^{o}_{\text{bind}}(\nu)}{\Delta\nu} = \Delta G^{o}_{\text{bind}}(\nu+1) - \Delta G^{o}_{\text{bind}}(\nu), \tag{S18}$$

where $\Delta\nu = 1$. Setting the above equal to zero, defining $\Delta G^{o}_{\mathrm{bind}}(\nu+1)$ and $\Delta G^{o}_{\mathrm{bind}}(\nu)$ using equation S17, and rearranging, one finds that

$$\begin{aligned} -k_BT\ln\left[\frac{a(\nu+1)_{WM,\mathrm{sol}}}{a(\nu)_{WM,\mathrm{sol}}}\right] &= -k_BT\ln[a_{W,\mathrm{sol}}] \\ -k_BT\ln\left[\frac{a(\nu+1)_{WM,\mathrm{sol}}}{a_{W,\mathrm{ideal}}\,a(\nu)_{WM,\mathrm{sol}}}\right] &= -k_BT\ln\left[\frac{a_{W,\mathrm{sol}}}{a_{W,\mathrm{ideal}}}\right], \end{aligned} \tag{S19}$$

where in the last line the natural logarithm of the activity of water in the ideal gas has been added to both sides of the equation. This last step is significant, as both sides of the equation can be interpreted as standard binding free energies for two different reversible reactions. The left-hand side is the standard free energy to add a water molecule from ideal gas to the water-macromolecule complex, and the right-hand side is the standard Gibbs free energy to transfer a water molecule from ideal gas to bulk water, which is explicitly considered in Section 2.4. As implied by equation S2 and the preceding paragraph, the left- and right-hand side of S19 can be interpreted as the excess chemical potential of water in the complex and the excess chemical potential of water in bulk solvent respectively. Thus, we conclude that at the minimum standard binding free energy state

$$\begin{aligned} \mu'_{W,WM}(\nu) &= \mu'_{W,\mathrm{sol}} \\ &= \mu'_{\mathrm{hyd}}, \end{aligned} \tag{S20}$$

where $\mu'_{W,WM}(\nu^*)$ is the excess chemical potential of water in the complex with optimal occupancy $\nu^*$, and $\mu'_{W,sol}$ is the excess chemical potential of water in bulk water solvent, which is nothing more than the hydration free energy of water, denoted $\mu'_{\mathrm{hyd}}$. In Adams's formulation of GCMC, one sets the Adams value $B$ and measures the average number of inserted water molecules at equilibrium. From equation S10 (equation 1 of the main text), one can determine the excess chemical potential of the GCMC region and thus the equilibrium state with bulk water.

It is important to note that equation S20 (equation 6 of the main text) implies that the chemical potentials of the GCMC region and bulk water are, in general, not equal at equilibrium. Following from the relation between the excess chemical potential and the chemical potential in equation S8, the equality of the excess chemical potentials of a water-macromolecule complex and bulk water allows the water densities in the complex to differ from that of bulk water.

## 2.2 Grand canonical integration

We seek an equation through which the Helmholtz free energy to couple multiple molecules from the ideal gas reservoir to a system of interest can be calculated with GCMC. However, instead of the Helmholtz free energy, denoted $F(N,V,T)$, the characteristic state function of the grand canonical ensemble is the grand potential, denoted $\Omega(\mu,V,T)$. In the thermodynamic limit[18], $F(N,V,T)$ and $\Omega(\mu,V,T)$ are related to each other via the Legendre transformation

$$\Omega(\mu,V,T) = F(N,V,T) - N\mu, \tag{S21}$$

where $N$ is the average number of inserted molecules for an applied $\mu$ at equilibrium; subscripts and angular brackets signifying, respectively, equilibrium and ensemble averaged variables have been omitted for notional simplicity. Throughout, we consider systems and ideal gas reservoirs at the same temperature and volume so that, henceforth, the Helmholtz free energy and grand potential, are respectively, abbreviated as $F(N)$ and $\Omega(\mu)$.

In GCMC, one can change the average number of bound water molecules from an initial number $N_i$ to a final $N_f$ by altering the applied chemical potential from $\mu_i$ to $\mu_f$ respectively. Using equation S21, the difference in Helmholtz free energy, $F(N_f) - F(N_i) = \Delta F(N_i \rightarrow N_f)$, between these states is related to $\Omega(\mu_f) - \Omega(\mu_i) = \Delta\Omega(\mu_i \rightarrow \mu_f)$ via

$$\begin{aligned}\Delta F(N_i \to N_f) &= \Delta\Omega(\mu_i \to \mu_f) + N_f\mu_f - N_i\mu_i \\ &= \Delta\Omega(B_i \to B_f) + k_BT\big(N_fB_f - N_iB_i\big) - k_BT\big(N_f - N_i\big)\ln\left(\frac{V}{\Lambda^3}\right), \end{aligned} \tag{S22}$$

where, using equation S7, $\mu$ has been expressed in terms of $B$ to make the above directly applicable to Adams' formulation of GCMC. The Helmholtz free energy required to insert molecules from the ideal gas reservoir to the system of interest is given by

$$\Delta F_{\text{trans}}(N_i \to N_f) = \Delta F_{\text{sys}}(N_i \to N_f) - \Delta F_{\text{res}}(N_i \to N_f), \tag{S23}$$

where $\Delta F_{\text{sys}}$ and $\Delta F_{\text{res}}$ refer to the system of interest and the ideal gas reservoir respectively, and can be computed using equation S22. The free energy $\Delta F_{\text{trans}}$ is our quantity of interest, so that evaluating $\Delta F_{\text{sys}}$ and $\Delta F_{\text{res}}$ will complete the derivation.

To evaluate $\Delta F_{\text{sys}}$, an expression for $\Delta\Omega_{\text{sys}}$ is required. A relation valid in both statistical mechanics and thermodynamics that has proved fruitful in a previous GCMC study[19] is

$$\left(\frac{\partial\Omega}{\partial\mu}\right)_{V,T} = -N(\mu), \tag{S24}$$

where it has been made explicit that $N$ is a function of $\mu$. Importantly, as $\mu$ – or $B$ – is user defined, $N(\mu)$ is a measurable and controllable quantity in a GCMC simulation. Thus, the change in the grand potential that occurs by altering the chemical potential from an initial $\mu_i$ to a final $\mu_f$ can be found with

$$\begin{aligned}\Delta\Omega_{\text{sys}}(\mu_i \to \mu_f) &= -\int_{\mu_i}^{\mu_f} N(\mu)\ \mathrm{d}\mu \\ &= -k_BT\int_{B_i}^{B_f} N(B)\ \mathrm{d}B, \end{aligned} \tag{S25}$$

where, using equation S7, the variable of integration has been changed to the Adams parameter. The limits of integration have also been changed to Adams values; the constants relating $\mu'$ to $B$ are not required as $B_i$ and $B_f$ – similarly to $\mu_i$ and $\mu_f$ – are *defined* as the Adams values that correspond an average of $N_i$ and $N_f$ waters, respectively. In congruence with equation S24, both $V$ and $T$ are kept constant in GCMC, so that by simulating a system at various $B$s, $\Delta\Omega_{\text{sys}}$ can be computed using numerical integration of $N(B)$. We therefore have

$$\beta\Delta F_{\text{sys}}(N_i \to N_f) = -\int_{B_i}^{B_f} N(B)\ \mathrm{d}B + \big(N_fB_f - N_iB_i\big) - \big(N_f - N_i\big)\ln\left(\frac{V}{\Lambda^3}\right). \tag{S26}$$

Next, an expression for $\Delta F_{\text{res}}$ is required. Section 2.3.1 of this supplement shows that if $\Delta F_{\text{res}}$ is found by evaluating $\Delta\Omega_{\text{res}}$ via integration (like in equation S25), one is left with an expression that is only valid for large $N$. Instead, an approach that is valid for both small and large $N$ is to consider the free energy for an ideal gas of indistinguishable particles[3],

$$\beta F_{\text{res}}(N) = -\ln\left[\frac{1}{N!}\left(\frac{V}{\Lambda^3}\right)^N\right], \tag{S27}$$

which implies that

$$\beta\Delta F_{\mathrm{res}}(N_i \to N_f) = \ln\left(\frac{N_f!}{N_i!}\right) - (N_f - N_i)\ln\left(\frac{V}{\Lambda^3}\right). \tag{S28}$$

Combining this with equations S23 and S26, we arrive at our primary theoretical result

$$\beta\Delta F_{\mathrm{trans}}(N_i \to N_f) = N_f B_f - N_i B_i + \ln\left(\frac{N_i!}{N_f!}\right) - \int_{B_i}^{B_f} N(B)\ \mathrm{d}B, \tag{S29}$$

which is equation 3 of the main text. The contribution due to translation, $\ln\left(\frac{V}{\Lambda^3}\right)$, has cancelled out when taking the difference between $\Delta F_{\mathrm{sys}}$ and $\Delta F_{\mathrm{res}}$ as both the ideal gas reservoir and the system are taken to be at equal volume. Likewise, contributions from ideal rotations (not considered in the above) also cancel out in $\Delta F_{\mathrm{trans}}$.

Equation S29 allows for the free energy to simultaneously couple multiple molecules to be efficiently computed, providing that there are GCMC data over a range of $B$ values. All that is required is that one measures $N(B)$, and numerical integration. As the application of equation S29, depends heavily on the accuracy of the measured $N(B)$, a fitted model to smooth over the GCMC titration data was introduced in equation 4 of the main text.

An important issue whether equation S29 (3 of the main text) is valid for a small number of molecules, given that a thermodynamic relationship (equation S21) was used as the starting point for its derivation. A key feature of the derivation was the use of the statistical mechanical relationship for the free energy of an ideal gas (equation S27), such that equation S29 ostensibly contains both micro- and macroscale elements. While the Results section of the main text presents numerical evidence for the consistency of equation S29 with replica exchange thermodynamic integration, the most stringent validation that equation S29 is correct for a low number of molecules is presented in Section 2.7 of this supplement, where it is analytically proven that the free energy to transfer a *single* water molecule from an ideal gas reservoir to a system of interest calculated from equation S29 is equal to that of the logistic equation (equation 2 of the main text), where the latter can be derived from statistical mechanics.

Explorations of how purely thermodynamic derivations result in a different version of equation 3 are presented in sections 2.3 and 2.3.1.

## 2.3 GCI in the thermodynamic limit

Here, we re-derive equation equation S29 (3 of the main text) from a different starting point from the previous section. The purpose is to explore to what extent the thermodynamic relations break down when dealing with a low number of molecules, and the regime of accuracy of equation 6. The route investigated here is similar to an approach by Fan et al.[20], which involves integration of the excess chemical potential.

To begin, equations S1 and S2 imply that a general definition of $\mu'$ is

$$\mu'(N) = \left(\frac{\partial F_{\mathrm{ex}}(N, V, T)}{\partial N}\right)_{T,V}, \tag{S30}$$

which implies that the free energy to insert additional molecules from the ideal gas to a system of interest, starting from an initial number, denoted $N_i$, to a final number, denoted $N_f$, is simply

$$\Delta F_{\mathrm{trans}}(N_i \to N_f) = \int_{N_i}^{N_f} \mu'_{\mathrm{sys}}(N)\ \mathrm{d}N \tag{S31}$$

To proceed, we note that $B$ can be considered a function of $N$, and $N$ can be considered as a function of $B$. The functions that map between these two variables are monotonically increasing because – as equations S5 and S6 show – the probability to insert a molecule increases with $B$. With this in mind, we can re-write equation S10 as

$$\beta\mu'_{\mathrm{sys}}(N) = B(N) - \ln(N). \tag{S32}$$

Combining this with equation S31 and integrating over $N$, we arrive at

$$\begin{aligned}\beta\Delta F_{\mathrm{trans}}\big(N_i \to N_f\big) &= \int_{N_i}^{N_f} \Big(B(N) - \ln(N)\Big)\ \mathrm{d}N \\ &= \int_{N_i}^{N_f} B(N)\ \mathrm{d}N - N_f\big(\ln(N_f) - 1\big) + N_i\big(\ln(N_i) - 1\big)\end{aligned} \tag{S33}$$

This, like equation S29 can be used to calculate the free energy to change the number of molecules in the system to and from any number within the interval between $N_i$ and $N_f$. It is thermodynamically exact, but, as shown below, differs from statistical mechanics for a low number of molecules. To express equation S33 in the form of equation S29, the variable of integration must be changed from $N$ to $B$, which can be achieved with the following identity:

$$\int\limits_{B_i}^{B_f} N(B)\ \mathrm{d}B + \int\limits_{N_i}^{N_f} B(N)\ \mathrm{d}N = B_f N_f - B_i N_i, \tag{S34}$$

which follows from a known relation for inverse function integration[21]. Inserting this into equation S33, we arrive at

$$\beta\Delta F_{\mathrm{trans}}\big(N_i \to N_f\big) = N_f B_f - N_i B_i - \underbrace{N_f\big(\ln(N_f) - 1\big)}_{\text{Stirling's approx.}} + \underbrace{N_i\big(\ln(N_i) - 1\big)}_{\text{Stirling's approx.}} - \int_{B_i}^{B_f} N(B)\ \mathrm{d}B, \tag{S35}$$

which is the same as equation S29, except for the sections encompassed by the curly brackets. In their place, equation S29 has $\ln(\frac{N_i!}{N_f!})$, and inspection of equation S35 shows that these terms are large $N$ approximations to this term. The approximation is known as Stirling's approximation, in which

$$\lim_{x\to\infty} x(\ln(x) - 1) = \ln(x!) \tag{S36}$$

This approximation was never explicitly invoked when deriving equation S35. Instead, it arose naturally out of the language of thermodynamics, in which $N$ is large and continuous. Therefore, the most significant assumption in the derivation of equation S35 was the macroscale definition of $\mu'$ in equation S30. This implies the microscale correction, denoted $R(N_i, N_f)$, to equation S31, so that

$$\Delta F_{\mathrm{trans}}(N_i \to N_f) = \int_{N_i}^{N_f} \mu'_{\mathrm{sys}}(N)\ \mathrm{d}N + k_B T R(N_i, N_f), \tag{S37}$$

where

$$R(N_i, N_f) = N_f\big(\ln(N_f) - 1\big) - N_i\big(\ln(N_i) - 1\big) + \ln\left(\frac{N_i!}{N_f!}\right) \tag{S38}$$

for identical particles. This correction, found by taking the difference of equation S35 and S29, removes the implicit application of Stirling's approximation when calculating insertion free energies and ensures equality between equation S37 and 3. When dealing with small number of particles, the correction is small, but significant. For instance, when going from 0 to 1 particle, $R(0, 1) = -1$. For appreciable number of particles, $R(N_i, N_f)$ is negligible. For example, $R(10, 11) = -0.05$. As finding the minimum of a binding free energy curve can be considered as a series of relative free energy calculations, the approximation given in equation 6 of the main text to find the optimal number of bound molecules can be used with greater and greater confidence as the optimal number of particles increases.

### 2.3.1 Particle creation in an ideal gas

The free energy to create particles in an ideal gas reservoir was computed in Section 2.2 of this supplement with equation S29 by considering the analytical solution to the free energy of an ideal gas in equation S27. In this section, we re-derive the equation from the grand canonical isotherm in equation S24 to show that the result is only valid in for a large number of molecules.

Using S25, the Adams value will be changed from $B'_i$ to $B'_f$ so the the number of molecules created in the reservoir goes from $N_i$ to $N_f$. Using the fact that $B = \ln N_{\text{res}}$ for an ideal gas (see equation S11), we find that

$$\begin{aligned}\Delta\Omega_{\text{res}} &= -k_B T \int_{B'_f}^{B'_i} \exp(B)\ \mathrm{d}B \\ &= -k_B T\big(\exp(B'_f) - \exp(B'_i)\big) \\ &= -k_B T(N_f - N_i), \end{aligned} \tag{S39}$$

which, using equation S22, implies that

$$\begin{aligned}\beta\Delta F_{\text{res}}(N_i \to N_f) &= -(N_f - N_i) + k_B T(N_f B_f - N_i B_i) + (N_f - N_i)\ln\frac{V}{\Lambda^3} \\ &= \underbrace{N_f\big[\ln(N_f) - 1\big]}_{\text{Stirling's approx.}} - \underbrace{N_i\big[\ln(N_i) - 1\big]}_{\text{Stirling's approx.}} + (N_f - N_i)\ln\frac{V}{\Lambda^3}. \end{aligned} \tag{S40}$$

Here, as with equation S35, Stirling's large $N$ approximation to $\ln(N!)$ has emerged implicitly. In this case, it is the use of Legendre transform in equation S21 that is responsible for this approximation. The Legendre transformations in thermodynamic can be shown to be valid when $N \to \infty$[18].

## 2.4 Hydration Helmholtz free energy of N waters

The central result of this study, equation S29 (equation 3 of the main text), relates a changes in the grand potential, $\Omega(\mu, V, T)$, to a change in the Helmholtz free energy, $F(N, V, T)$. We are concerned with establishing the equilibrium number of water molecules in a GCMC region when that region can exchange particles with bulk water. We require a thermodynamic expression for hydrating $N$ water molecules in an infinitely sized water bath. This is most easily derived using the Gibbs free energy, denoted $G(N, P, T)$, where $P$ is the pressure. The Gibbs hydration free energy can then be related back to the Helmholtz hydration free energy in the thermodynamic limit.

Let $G_{\text{water}}(N_b)$ denote the absolute Gibbs free energy of a water bath that contains $N_b$ solvent molecules at a constant pressure and temperature. Allowing $N_b \to \infty$ implies that $N_b\mu_{\text{water}} = G_{\text{water}}(N_b)$, due to the extensivity of the free energy[3]. Similarly, $(N_b + N)\mu_{\text{water}} = G_{\text{water}}(N_b + N)$ Thus, the free energy to increase the number of water molecules by $N$ when $N_b$ is large is

$$\begin{aligned}\Delta G_{\text{water}}(N,P,T) &= G_{\text{water}}(N_b+N,P,T) - G_{\text{water}}(N_b,P,T) \\ &= (N_b+N)\mu_{\text{water}} - N_b\mu_{\text{water}} \\ &= N\mu_{\text{water}} \end{aligned} \tag{S41}$$

Similarly, the free energy to create an additional $N$ particles in an infinitely sized ideal gas reservoir, denoted $\Delta G_{\text{res}}(N)$, has the same form. Therefore, the hydration free energy of $N$ molecules is given by

$$\begin{aligned}\Delta G_{\text{hyd}}(N,P,T) &= \Delta G_{\text{water}}(N,P,T) - \Delta G_{\text{res}}(N,P,T) \\ &= N\mu_{\text{water}} - N\mu_{\text{res}} \\ &= N(\mu_{\text{water}} - \mu_{\text{res}}) \\ &= N\mu'_{\text{hyd}}, \end{aligned} \tag{S42}$$

where $\mu'_{\text{hyd}}$ is the excess chemical potential of a single water molecules in bulk water, and the last line follows from the definition of the excess chemical potential.

Finally, as in the thermodynamic limit the Gibbs hydration free energy and the Helmholtz free energy are equivalent due to the negligible contribution of volume changes when $P$ is constant[10]. We therefore have

$$\Delta F_{\text{hyd}}(N,V,T) = N\mu'_{\text{hyd}} \tag{S43}$$

when inserting $N$ water molecules from the ideal gas phase to bulk water.

## 2.5 Equilibrium with bulk water

The section will derive the equation 6 of the main text, which occurs when the simulated GCMC region is in equilibrium with bulk water when it is allowed to exchange water molecules. Starting from equation 5 of the main text, the binding free energy of $N$ water molecules is given by

$$\begin{aligned}\Delta F_{\text{bind}}(N) &= \Delta F_{\text{trans}}(N) - \Delta F_{\text{hyd}}(N). \\ &= \Delta F_{\text{trans}}(N) - N\mu'_{\text{hyd}}, \end{aligned} \tag{S44}$$

where equation S43 has been used in the last line. Grand canonical integration (equation S29) can be used to evaluate $\Delta F_{\text{trans}}(N)$, and $\mu'_{\text{hyd}}$ is a known quantity. The number of water molecules that minimizes the binding free energy, denoted $N^*$, is the most likely occupancy of the GCMC region. This can be expressed mathematically as

$$N^* = \underset{N}{\arg\min}\left[\Delta F_{\text{trans}}(N) - N\mu'_{\text{hyd}}\right], \tag{S45}$$

as we seek the argument that minimizes $\Delta F_{\text{bind}}(N)$. If the GCMC region is assumed to be in the thermodynamic limit, we can approximate $\Delta F_{\text{trans}}(N)$ with equation S31, which was derived and discussed in Section 2.3. Strictly speaking, $N$ is limited to the integer numbers. However, in the thermodynamic limit one may treat $N$ as a continuous variable, so that the binding free energy of $N$ waters can be approximated to be

$$\Delta F_{\text{bind}}(N) \approx \int_0^N \mu'_{\text{sys}}(N^\dagger)\, \mathrm{d}N^\dagger - N\mu'_{\text{hyd}}, \tag{S46}$$

where $N^{\dagger}$ is a dummy variable. Equation S45 can be evaluated by solving $\frac{\mathrm{d}\Delta F_{\text{bind}}(N)}{\mathrm{d}N} = 0$. Doing so produces

$$\mu'_{\text{sys}}(N^*) = \mu'_{\text{hyd}}, \tag{S47}$$

as stated in equation 6 of the main text and equation S20. Finding where $\mu'(N)$ equals the hydration free energy determines $N^*$. As equation S10 shows, $\mu'(N)$ depends on $B$ and the average number of inserted molecules at equilibrium, and does not require the binding free energy to be evaluated. Inserting equation S47 into S10 produces

$$B^* = \beta\mu'_{\text{hyd}} + \ln(N^*) \tag{S48}$$

where $B^*$ is the value that gives the equilibrium number of water molecules in a GCMC simulation. As stated, both this and equation S47 are only approximately true when $N$ is small, and exact in the thermodynamics limit. Nevertheless, as the end of Section 2.3 of this supplement shows, the approach to the large $N$ limit is fast, so that equations S47 and S48 can be reliably applied to cavities that have around 10 waters bound.

## 2.6 Kinetic analysis of GCMC

In this section, we derive equation 2 of the main text. While it is straightforward to do so rigorously, kinetic arguments are used in order to explain the observation by Clark[15] and Bodnarchuk et al.[16] that equation S12 becomes more valid as $B$ decreases.

First, we model a GCMC simulation as the simple two state reaction $W_{\text{sys}} \rightleftharpoons W_{\text{res}}$, in which water molecules are exchanged to and from the system of interest and ideal gas reservoir respectively. Water molecules are inserted from the ideal gas reservoir into the system with the rate constant $k_{\text{ins}}$ and deleted from the system and moved to the reservoir with a rate constant $k_{\text{del}}$. The kinetic equations governing the change in the average number of particles are given by

$$\frac{dN_{\text{sys}}}{dt} = -k_{\text{del}}N_{\text{sys}} + k_{\text{ins}}N_{\text{res}} \tag{S49}$$

$$\frac{dN_{\text{res}}}{dt} = k_{\text{del}}N_{\text{sys}} - k_{\text{ins}}N_{\text{res}} \tag{S50}$$

Equilibrium is defined when $\frac{dN_{\text{sys}}}{dt} = \frac{dN_{\text{res}}}{dt} = 0$, at which point the number of molecules in the system and ideal gas reservoir are $N_{\text{sys,equil}}$ and $N_{\text{res,equil}}$ respectively. At equilibrium, one can show that

$$\frac{N_{\text{sys,equil}}}{N_{\text{res,equil}}} = \frac{k_{\text{del}}}{k_{\text{ins}}} := K_d, \tag{S51}$$

where $K_d$ is the dissociation constant. The free energy to transfer a molecule from the ideal gas reservoir to the system is denoted $\Delta F_{\text{trans}}$, and is related to $K_d$ via

$$\beta\Delta F_{\text{trans}} = \ln K_d. \tag{S52}$$

Identifying $\Delta F_{\text{trans}}$ with $F_{\text{ex}}(N_{\text{sys}}+1) - F_{\text{ex}}(N_{\text{sys}})$ and using equations S51 and S52, we arrive at equation S12, which was used by Clark et al.[15] and Bodnarchuk et al.[16] to predict binding free energies. Thus, the equations S49 and S50 represents the dynamical system corresponding to equilibrium relation in equation S12. However, this kinetic model allows molecules from the ideal gas reservoir to be added to the system without limit, so long as the reservoir is large enough. As a cavity can only fit a finite number of molecules, we must incorporate the effect of saturation in the kinetics to more realistically model equilibrium in GCMC. The following heuristic model is based on a simple population growth model[22]:

$$\frac{dN_{\text{sys}}}{dt} = -k_{\text{del}}N_{\text{sys}} + k_{\text{ins}}N_{\text{res}}\left(1 - \frac{N_{\text{sys}}}{N_{\text{max}}}\right) \tag{S53}$$

$$\frac{dN_{\text{res}}}{dt} = k_{\text{del}}N_{\text{sys}} - k_{\text{ins}}N_{\text{res}}\left(1 - \frac{N_{\text{sys}}}{N_{\text{max}}}\right), \tag{S54}$$

where $N_{\text{max}}$ is the maximum number of water molecules that can fit in a cavity. This model may not be valid when a cavity can contain more than one water molecule, because one would have to account for exchanges between neighbouring sites, as well as cooperative and destabilisation effects.

As the number of inserted waters reaches this maximum capacity in equations S53 and S54, the positive contribution in equation S53 is zero, implying no further growth in $N_{\text{sys}}$. Pertinently, in the low particle limit, where $N_{\text{sys}} \ll N_{\text{max}}$, equations S53 and S54 reduce to equations S49 and S50 respectively. Thus, for a system described by equations S53 and S54, we have the following:

$$\lim_{N_{\text{sys,equil}} \to 0} k_B T \ln\left(\frac{N_{\text{sys,equil}}}{N_{\text{res,equil}}}\right) = \Delta F_{\text{trans}}. \tag{S55}$$

This behaviour can be observed in tables 1 and 2 of Clark[15], and table 3 of Bodnarchuk[16], and was previously attributed to sampling inadequacy. In Section 3.3 of this supplement we also verify equation S55 for out test systems. With the above kinetic models, it is apparent that equation S12 is only valid in cases when the number of inserted molecules is low with respect to the capacity of the cavity.

Solving equations S53 and S54 at equilibrium for the case $N_{\text{max}} = 1$, we find that

$$N_{\text{sys,equil}} = \frac{N_{\text{res,equil}}}{K_d + N_{\text{res,equil}}}, \tag{S56}$$

which has the same form as the well known equation for a ligand binding to a macromolecule. For instance, if the concentration of the ligand, macromolecule, and complex are denoted $[L]$, $[M]$ and $[ML]$ respectively,

$$\frac{[ML]}{[M_0]} = \frac{[L]}{K_d + [L]}, \tag{S57}$$

where $[M_0]$ is the initial concentration of the macromolecule[3]. As $[L]$ is typically approximated by the concentration of unbound (or free) ligand, it is directly comparable to the amount of molecules in the reservoir, $N_{\text{res,equil}}$. Also, $\frac{[ML]}{[M_0]}$ varies between 0 and 1, just like $N_{\text{sys,equil}}$ when $N_{\text{max}} = 1$. The correspondence equation S56 has with equation S57 serves to validate this result and the kinetic model described equations S53 and S54.

Finally, inserting $B = \ln N_{\text{res,equil}}$ (from equation S11) and $\beta\Delta F_{\text{trans}} = \ln K_d$ we arrive at

$$N_{\text{sys,equil}}(B) = \frac{1}{1 + \exp\left(\beta\Delta F_{\text{trans}} - B\right)}, \tag{S58}$$

which is equation 2 of the main text. Therefore, with many GCMC simulations at $B$ values, one can estimate the free energy to couple a molecule to a system by fitting equation S58 to $N_{\text{sys,equil}}$ from each simulation, as the only free parameter is $\Delta F_{\text{trans}}$. This method is distinct from GCI. The benefit of fitting equation S58 to estimate $\Delta F_{\text{trans}}$ as opposed to equation S55 is that data from all GCMC simulations at various $B$ values are utilized, compared to using only the simulations that have a low $N_{\text{sys,equil}}$. This will produce estimates of a lower variance.

## 2.7 Equivalence of GCI and logistic equation

In this work, we have introduced two equations that can be used calculate free energies using GCMC. The first, we have named grand canonical integration (GCI), involves the evaluation of equation S29 (equation 3 of the main text) with GCMC titration data and can, in theory, calculate the binding free energy of an arbitrary number of molecules. The second method applies to sites that can only bind a *single* water molecule, and free energies are obtained by fitting the logistic curve given by equation S58 (equation 2 of the main text) to GCMC titration data. In this section, we show analytically that free energies described by both equations are equal.

To differentiate between them, this section will label gas-system transfer free energies calculated with GCI as $\Delta F_{\text{gci}}$, and free energies calculated by fitting the logistic equation S58 as $\Delta F_{\text{log}}$.

To calculate the free energy to transfer a single water from the ideal gas reservoir to a system of interest, equation S29 becomes

$$\beta\Delta F_{\text{gci}}(0 \to 1) = B_f - \int_{B_i}^{B_f} N(B)\ \mathrm{d}B, \tag{S59}$$

as $N_i = 0$ and $N_f = 1$. Evaluating the integral with $N(B)$ given by equation S58 we get

$$\begin{aligned} \int_{B_i}^{B_f} N(B)\ \mathrm{d}B &= \int_{B_i}^{B_f} \frac{1}{1+\exp\left(\beta\Delta F_{\text{log}} - B\right)}\ \mathrm{d}B \\ &= \int_{B_i}^{B_f} \frac{\exp(B - \beta\Delta F_{\text{log}})}{1+\exp\left(B - \beta\Delta F_{\text{log}}\right)}\ \mathrm{d}B \\ &= \int_{B_i}^{B_f} \log\Big[1+\exp\left(B-\beta\Delta F_{\text{log}}\right)\Big]\ \mathrm{d}B \\ &= \log\left[\frac{1+\exp(B_f - \beta\Delta F_{\text{log}})}{1+\exp(B_i - \beta\Delta F_{\text{log}})}\right]. \end{aligned} \tag{S60}$$

The limits of integration must be chosen such that $N(B_i) = 0$ and $N(B_f) = 1$. As $N(B)$ is the logistic function given by equation S58, we require that $B_i - \beta\Delta F_{\text{log}} \ll 0$, and $B_f - \beta\Delta F_{\text{log}} \gg 0$. When $B_i - \beta\Delta F_{\text{log}} \to -\infty$ we get

$$1 + \exp(B_i - \beta\Delta F_{\text{log}}) = 1.$$

When $B_f - \beta\Delta F_{\text{log}} \to +\infty$ we find that

$$1 + \exp(B_f - \beta\Delta F_{\text{log}}) = \exp(B_f - \beta\Delta F_{\text{log}}).$$

Putting these limits into equation S60 shows that,

$$\int_{B_i}^{B_f} N(B)\ \mathrm{d}B = B_f - \beta\Delta F_{\text{log}}. \tag{S61}$$

Together with equation S59, this relation proves that

$$\Delta F_{\text{gci}}(0 \to 1) = \Delta F_{\text{log}}, \tag{S62}$$

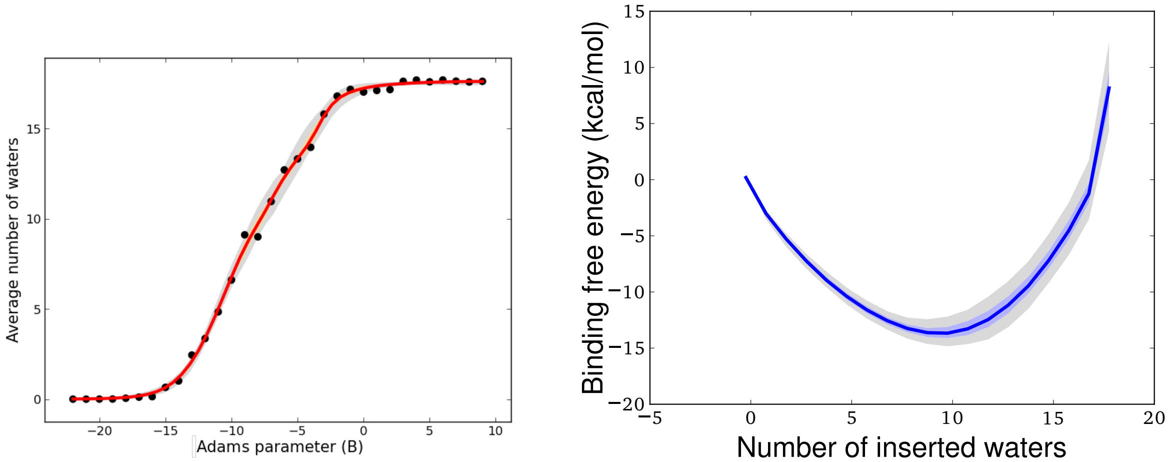
as stated in the main text. This means that free energies calculated with both methods are equal. The correspondence between equations S29 and S58 serves as a strong validation of the theoretical results presented here. This correspondence can also be derived using equation S37 as a starting point.

# 3 Methods and Results

## 3.1 Bulk water density

| Structure | Adams values simulated |
|---|---|
| 6.5×6.5×6.5 Å$^3$ | -22, -21, -20, -19, -18, -17, -16, -15, -14, -13, -12, -11, -10, -9, -8, |
| cavity in bulk water | -7, -6, -5, -4, -3, -2, -1, 0, +1, +2, +3, +4, +5, +6, +7, +8, +9 |

Table S1: The Adams values used in the GCMC simulations of a sub-volume of bulk water.

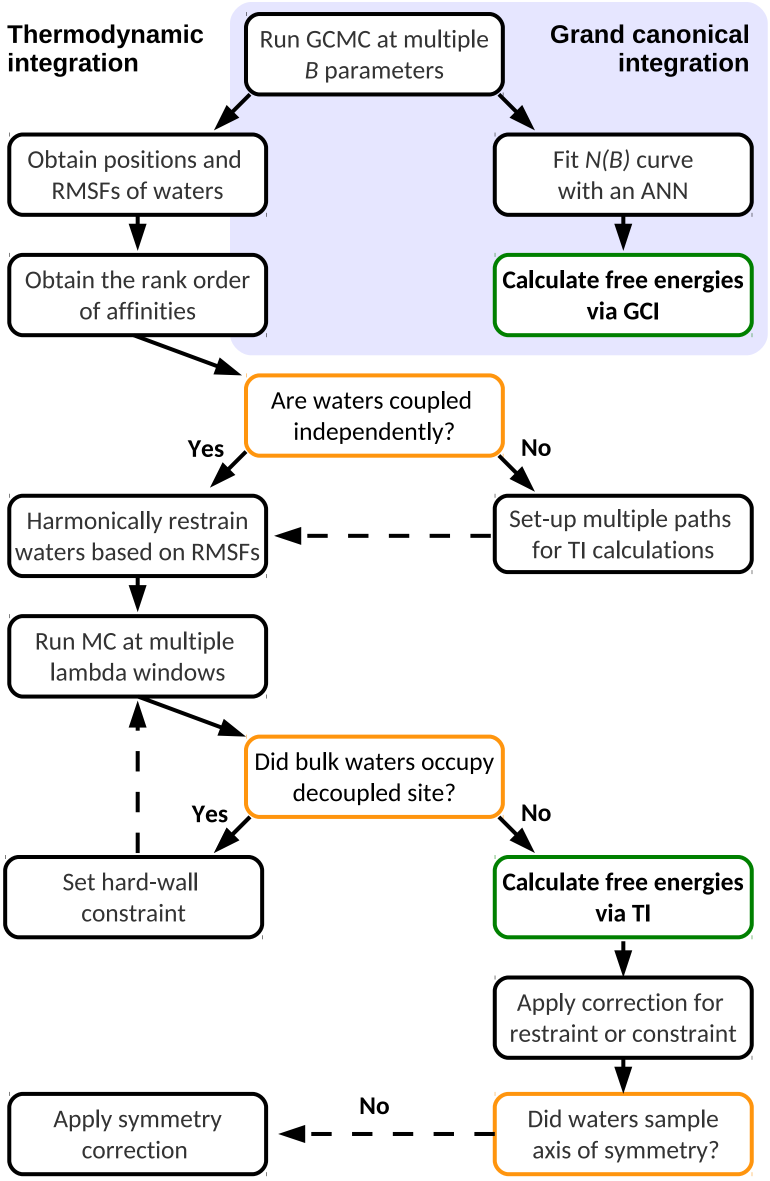


(a)

(b)

Figure S1: (a)Titration plot showing how the average occupancy of the a sub-volume within bulk water increases as the Adams value is increased. The fitted artificial neural network model is shown in red, and the 90% confidence region from 1000 bootstrap fits is shown in grey. The titration plot was fitted with equation 4 of the main text using three logistic terms ($m = 3$). At $B = -8.5$, bulk water density (nine waters for this cavity at approximately 1000 kg/m$^3$) is reproduced within the 6.5×6.5×6.5 Å$^3$ GCMC volume. (b) The binding free energy of the sub-volume as a function of water occupancy. The minimum free energy occurs at nine water molecules (to the nearest integer) which is as expected for bulk water. The minimum binding free energy equals $-13.8 \pm 1.0$ kcal/mol, and can be interpreted as the negative of the free energy required to empty the sub-volume of water in bulk water.

## 3.2 Free energy calculations with thermodynamic integration



Thermodynamic integration
Run GCMC at multiple B parameters
Grand canonical integration
Obtain positions and RMSFs of waters
Fit N(B) curve with an ANN
Obtain the rank order of affinities
Calculate free energies via GCI
Are waters coupled independently?
Yes
No
Harmonically restrain waters based on RMSFs
Set-up multiple paths for TI calculations
Run MC at multiple lambda windows
Did bulk waters occupy decoupled site?
Yes
No
Set hard-wall constraint
Calculate free energies via TI
Apply correction for restraint or constraint
No
Apply symmetry correction
Did waters sample axis of symmetry?

Figure S2: A schematic of how the transfer free energies of water molecules for the three water cavities in Chk-1 and BPTI were calculated with grand canonical integration (GCI), and sequentially with replica exchange thermodynamic integration (RETI). The locations where water molecules were decoupled from with RETI were taken from the GCMC simulations, as were the root mean squared fluctuation (RMSF) at each site. The order in which waters appeared with $B$ in the GCMC simulations indicates the relative affinity of each site[8]. This same order determined the sequence in which the waters were decoupled with RETI. Using RETI requires the careful consideration of restraints, constraints, and corrections, which become more difficult to set-up and apply when cavities can support more than one water molecule. Such considerations do not apply to GCI methodology, which is highlighted in blue. A monotonically increasing artificial neural network (ANN) was fitted to the average number of inserted water molecules at each $B$ to ensure the integration was carried out over a smooth function.

### 3.2.1 Single water cavities

| Structure | Adams values simulated |
|---|---|
| BPTI | -34, -33, -32, -31, -30, -29, -28, -27, -26, -25, -24, -23, -22.6, -22.2, -22, -21.8, -21.4, -21, -20.6, -20.2, -20, -19.8, -19.4, -19, -18.6, -18.2, -18, -17.8, -17.4, -17, -16, -15, -14, -13, -12, -11, -10, -9, -8, -7, -6, -5, -4, -3 |
| BPTI, *equilibrated around water* | -32, -32, -31, -31, -30, -30, -29, -29, -28, -28, -27, -27, -26, -26, -25, -25, -24, -24, -23, -23, -22, -22, -21, -21, -20, -20, -19, -19, -18, -18, -17, -17 |
| SD L01 | -22, -22, -21, -21, -20, -20, -19, -19, -18, -18, -17, -17, -16, -16, -15, -15, -14, -14, -13, -13, -12, -12, -11, -11, -10, -10, -9, -9, -8, -8, -7, -7 |
| SD L01, *equilibrated around water* | -23, -23, -22, -22, -21, -21, -20, -20, -19, -19, -18, -18, -17, -17, -16, -15,-16, -15, -14,-14, -13, -13, -12, -12, -11, -11, -10, -9, -10, -9, -8, -8 |
| SD L03 | -16, -15, -14, -13, -12, -11, -10, -10, -9, -9, -8, -8, -7.5, -7, -7, -6.5, -6, -6, -5.5, -5, -5, -4, -4, -3, -3, -2, -2, -1, -1, 0, 3, 4 |
| SD L03, *equilibrated around water* | -23, -23, -22, -22, -21, -21, -20, -20, -19, -18, -17, -17, -16, -16, -15, -15, -14, -14, -13, -12, -13, -12, -11, -11, -10, -10, -9, -8, -9, -8 |

Table S2: The Adams values used in the GCMC simulations of the single water cavities in the protein scytalone dehydratase (SD) with ligands L01 and L03, and bovine pancreatic trypsin inhibitor (BPTI). The starting structures for the GCMC simulations labelled *equilibrated around water* were taken from the last frame of a 45M move simulation in which the water in present in the cavity.

| System | PDB code | Oxygen coordinates | GCMC box origin | GCMC box dimensions |
|---|---|---|---|---|
| BPTI | 5PTI | (32.74, 4.03, 10.65) | (30.60, 2.45, 8.57) | (4.40, 3.52, 3.65) |
| SD | 3STD | (26.42, 13.87, 36.56) | (24.30, 11.90, 34.50) | (4.00, 4.00, 4.00) |

Table S3: The starting water oxygen coordinates used in the RETI calculations, and the box coordinates and dimensions for the GCMC simulations. Coordinates and dimensions are given in the form of (x, y, z) and are in the same coordinate frame as their respective PDB structures. For SD, the same location is studied for the ligands L1 and L3.

| System | Restraint/Constraint | Size | Symmetry correction? |
|---|---|---|---|
| BPTI | Harmonic | 45.54 kcal/mol/Å$^2$ | Yes |
| SD L1 | Hard-wall | 1.8 Å | No |
| SD L3 | Hard-wall | 1.8 Å | No |

Table S4: The final set of restraints/constraints that were applied to the single-water cavities for the RETI calculations. The size column refers either to the radius of the applied hard-wall, or spherically symmetric force constant.

### 3.2.2 Three water cavities

| Structure | Adams values simulated |
|---|---|
| BPTI | -35, -34, -33, -32, -31, -30, -29, -28.6, -28.5, -28.5, -28.3, -28, -27.7, -27.2, -27, -26.3, -26, -25.7, -25.5, -25.2, -25.1, -25, -24.3, -24, -23.7, -23.2, -23, -22.3, -22, -21.7, -21.2, -21.0, -21, -20.9, -20.3, -19.7, -19, -17.5, -16, -15, -14 |
| Chk-1 | -30, -29, -28, -27, -26, -25, -24, -23, -22, -21, -21, -20, -20, -19, -19, -18, -18, -17, -17, -16, -16, -15, -15, -14, -13, -12, -11, -10, -9, -8, -7, -6, |

Table S5: The Adams values used in the GCMC simulations of the three water cavities.

The RETI simulations were more technically demanding than the GCMC simulations. Details of the final set of simulations that were used to hydrate the cavities are shown in Tables S7 and S8. As described in the schematic in Figure S2, the initial set of restraints on the water molecules were harmonic, and were based on the root mean squared fluctuation (RMSF) of the water molecule in the GCMC simulations. The RMSF is shown in the 'Size' column of Tables S7 and S8. The harmonic force constant that was applied is as described in Hamelberg and McCammon[23]. If bulk water was found to have drifted into the decoupled site, the simulations were re-run with a hard-wall constraint, the radius of which is also shown in the 'Size' column. Each simulation was visually inspected to see whether the water molecule in question rotated about its axis of symmetry. If it did not, a symmetry correction was applied to the calculated free energy.

| System | Label | Coordinates (x,y,z) | GCMC box origin | GCMC box dimensions |
|---|---|---|---|---|
| BPTI | A | (31.62, 6.99, 1.37) | (28.03, 3.48, -2.173) | (7.02, 5.46, 8.58) |
| (PDB code: 5PTI) | B | (32.01, 7.27, 4.13) | | |
| | C | (32.48, 5.81, 0.202) | | |
| Chk-1 | X | (12.59, -3.518, 14.41) | (7.38, -6.56, 12.19) | (7.79, 8.89, 6.16) |
| (Private structure) | Y | (9.99, -2.66, 15.50) | | |
| | Z | (11.69, -0.416, 15.65) | | |

Table S6: The coordinates of the water oxygen atoms that were used in the alchemical decoupling simulations, as well as the GCMC box coordinates and dimensions. The coordinates for BPTI refer to cavity that can bind 3 water molecules, which is distinct from the cavity used to bind one water molecule. These coordinates are the mean positions taken from the GCMC simulations in which the cavity was filled, and are in the reference frame of the PDB file.

| Transition | Restraints/Constraints | Size | Symmetry correction? |
|---|---|---|---|
| Empty → X | harmonic | 9.67 kcal/mol/Å$^2$ | Yes |
| X → X & Y | harmonic | 13.41 kcal/mol/Å$^2$ | No |
| X & Y → X & Y & Z | harmonic | 6.41kcal/mol/Å$^2$ | No |

Table S7: The calculation pathways and set of restraints used to couple water molecules into the three water cavity in Chk-1. Waters are labelled as X, Y and Z, where their coordinates are shown in table S6. Whilst waters Y and Z appeared concurrently in the GCMC simulations with increasing $B$, the occupancy of Y was higher than that of Z, and so was coupled-in first in the RETI simulations.

| Transition | Restraint/Constraint | Size | Symmetry correction? |
|---|---|---|---|
| Empty → A | hard-wall | 1.8 Å | Yes |
| Empty → B | hard-wall and plug | 1.8 Å | No |
| A → A & B | hard-wall | 1.8 Å | Yes |
| B → A & B | harmonic | 31.71 kcal/mol/Å$^2$ | No |
| A & B → A & B & C | harmonic | 19.96 kcal/mol/Å$^2$ | No |

Table S8: The calculation pathways and final set of restraints/constraints used to couple water molecules into the three water cavity in BPTI. The size column refers either to the radius of the applied hard-wall, or spherically symmetric force constant. Waters are labelled as A, B and C, where their coordinates are shown in Table S6. Two routes to coupling waters A and B was computed as they appeared together with equal occupancy in the GCMC simulations.

### 3.3 Comparison between the method by Clark et al. and logistic model

The main text presents a method to calculate the coupling free energy of a molecule by fitting equation 2 (equation S58 of this supplement), a logistic equation, to GCMC titration data. This method is more general than the previously reported technique by Clark et al.[15], which is shown in equation S12. Equation S12 is shown to be a limiting case of our logistic model in equation S55. Here, the free energies computed with the logistic fit are compared to those computed with the method by Clark et al.

The basis of the method by Clark et al. is based on equations S10 and S11, with which one can show

$$\mu'_{\text{sys}} = k_B T \ln\left(\frac{N_{\text{res,equil}}}{N_{\text{sys,equil}}}\right) \tag{S63}$$

$$= k_B T B - k_B T \ln N_{\text{sys,equil}}, \tag{S64}$$

so that $\mu'_{\text{sys}}$ can be calculated from GCMC simulations as $B$ is known, and $N_{\text{sys,equil}}$ can calculated from a simulation. Clark et al. and Bodnarchuk et al.[16] then used $\mu'_{\text{sys}}$ to approximate the free energy of insertion from an ideal gas, denoted $\Delta F_{\text{trans}}$. However, as shown in Section 2.6, this only becomes valid as $N_{\text{sys,equil}} \rightarrow 0$, during which $\mu'_{\text{sys}} \rightarrow \Delta F_{\text{trans}}$, a constant. As Figures S3 and S4 show, most of the simulation data has to be discarded, and human judgement must be used to decide the point at which $\mu'_{\text{sys}}$ appears constant, introducing subjectivity into the estimate of $\Delta F_{\text{trans}}$. In contrast, least squares fitting of the logistic model uses all of the simulation data, negating the need for human intervention when estimating $\Delta F_{\text{trans}}$. Figures S3 and S4 show that GCMC data for single water molecules are well described by the logistic model. Nevertheless, Table S9 show that both methods produce free energy estimates that are in good agreement.

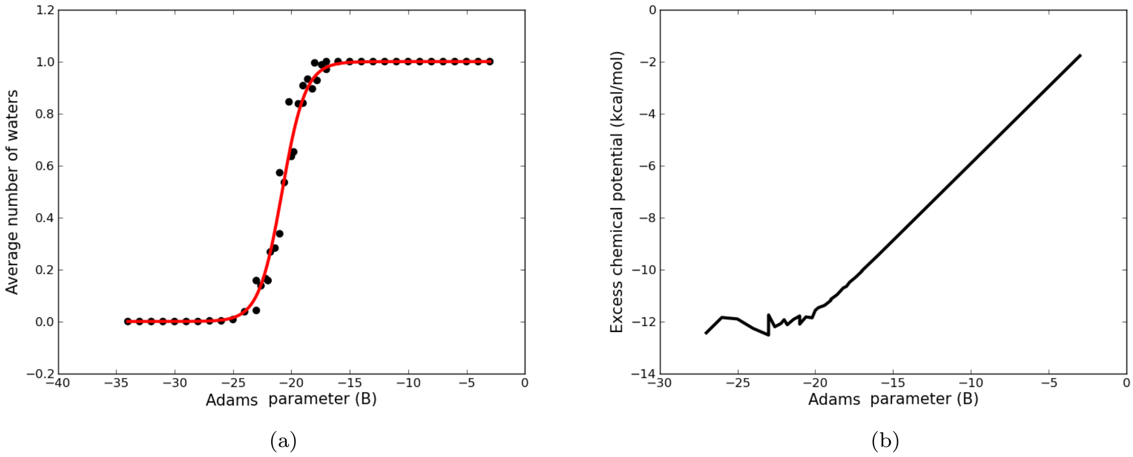
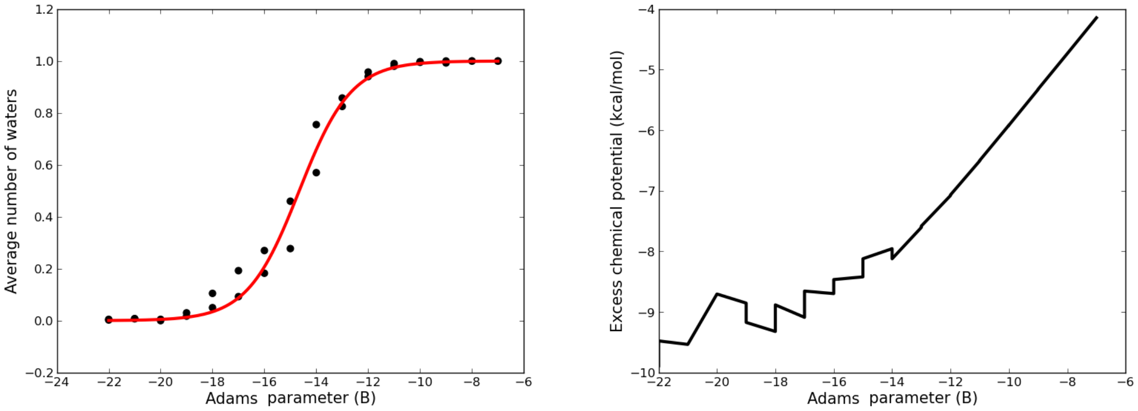


Figure S3: GCMC simulation data for the single water cavity in BPTI. (a) Titration plot showing how the average occupancy of the cavity decreases as the Adams parameter is lowered. The fitted logistic model is shown in red; its close agreement with the data indicates that GCMC for single water sites is well described by equation S58 (equation 2 of the main text). The coupling free energy is equal to the point of inflection multiplied by $k_B T$. (b) The estimated excess chemical potential as a function of the Adams parameter, used to estimate the coupling free energy using Clark et al.'s method[15]. Human judgement is used to determine where the excess chemical appears constant, at which point the average value is taken. The range used to compute the value in Table S9 was $B = -27$ to $B = -22$.



(a) (b)

Figure S4: GCMC simulation data for the single water cavity in SD L01. (a) GCMC titration plot with fitted logistic model in red. (b) The estimated excess chemical potential as a function of the Adams parameter. The noise in the excess chemical potential introduces subjectivity when determining the range to estimate the coupling free energy, which is not the case when fitting equation S58 (equation 2 of the main text). For the value quoted in Table S9, the range $B = -20$ to $B = -15$ was chosen.

| Water | Logistic Fit | Clark limit method |
|---|---|---|
| BPTI | -12.32 ± 0.07 | -12.10 ± 0.26 {9} |
| BPTI *equilibrated around water* | -11.63 ± 0.11 | -11.41 ± 0.22 {8} |
| SD L01 | -8.68 ± 0.09 | -8.76 ± 0.34 {11} |
| SD L01 *equilibrated around water* | -9.52 ± 0.06 | -9.72 ± 0.31 {13} |
| SD L03 | -3.51 ± 0.06 | -3.52 ± 0.30 {20} |
| SD L03 *equilibrated around water* | -3.49 ± 0.06 | -3.34 ± 0.24 {10} |

Table S9: Comparison of the logistic fit method (equation 2 in main text and equation S58 of this supplement) and the method proposed by Clark et al.[15], which is represented by equation S55. The error in the logistic fit has been calculated using 1000 bootstrap samples, whereas for the limit method of Clark et al., the quoted error is the standard deviation of the values that were used when averaging, with the number of data points used to the average are shown within curly brackets.

## 3.4 Artificial Neural Network

As described in the main text, a single layer artificial neural network (ANN) was used to smooth the GCMC titration data such that integration in GCI could be reliably performed. While there are numerous packages one can use to fit ANNs, two features of this project meant that it was most fruitful to code our own tool: first, the relationship between Adams parameter and average number of inserted waters should be monotonically increasing, and second, we sought to experiment with different loss, or cost, functions to compensate for the noisy titration data in the three water cavity of BPTI.

A single layer monotonic ANN fitting tool was written in python using both the NumPy[24] and SciPy[25] packages. The ANN described in equation 4 of the main text can be made to be monotonic increasing if all the free parameters (commonly referred to as weights) are greater than or equal to zero. To accomplish this, the loss function (described below) was minimized using the L-BFGS-B algorithm, which allows for input of variable constraints.

| | Free energy (kcal/mol) to couple | | |
|---|---|---|---|
| Loss function | 1 water | 2 waters | 3 waters |
| $L_{\mathrm{SE}}(r)$ | -16.3 | -31.9 | -44.0 |
| $L_{\mathrm{A}}(r)$ | -16.0 | -31.3 | -43.4 |
| $L_{\mathrm{PH}}(r, c = 1)$ | -15.4 | -31.0 | -43.3 |
| $L_{\mathrm{PH}}(r, c = 0.1)$ | -15.4 | -31.0 | -43.5 |

Table S10: Comparison of coupling free energies calculated with GCI for BPTI using different loss functions when fitting to the titration data. The squared error loss function, $L_{\mathrm{SE}}(r)$, disagrees by up to 0.9 kcal/mol with the values calculated with the other loss functions, due to the poor fit it produces (see Figure S5). The free energies corresponding to $L_{\mathrm{A}}(r)$ are approximate due to the instability of this loss function when optimizing using gradient descent.

### 3.4.1 Loss function

The loss function is a measure of the performance of a fitted function, and encompasses the intuitive notion of error. Denoting the explanatory variable, $X$, the fitted function as $f(X)$, and the target as $Y$, the residual of the $i$th pair $(x_i, y_i)$ is given by $r_i = y_i - f(x_i)$. The total loss over $n$ pairs is therefore $\sum_{i=1}^{n} L(r_i)$. A common choice of loss function is the squared error

$$L_{\mathrm{SE}}(r) = r^2, \tag{S65}$$

which has the benefit of being differentiable everywhere, and is thus suitable to gradient based optimizers. Squared loss yields a unique solution and allows for tractable analytical evaluation of gradients. However, when applied to noisy data, $L_{\mathrm{SE}}(r)$ may be dominated by outliers due to the quadratic dependence of $r$. A popular alternative which is more robust to outliers is absolute loss

$$L_{\mathrm{A}}(r) = |r|. \tag{S66}$$

However, this is not differentiable at $r = 0$, which makes optimization cumbersome and prone to error without dedicated algorithms. The pseudo Huber loss function[26], defined as

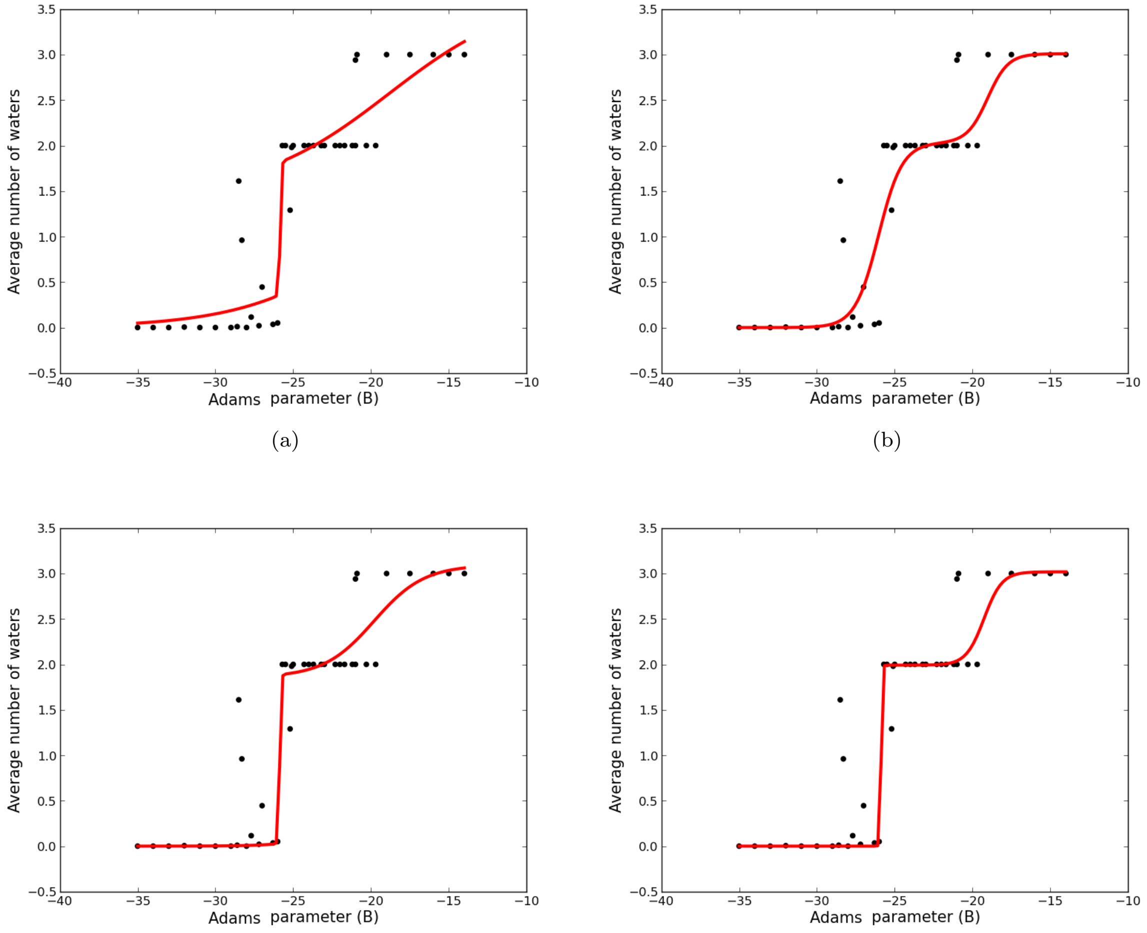
$$L_{\mathrm{PH}}(r, c) = c^2 \left( \sqrt{1 + \frac{r}{c}} - 1 \right), \tag{S67}$$

where $c$ is a free parameter, is a compromise between $L_{\mathrm{SE}}(r)$ and $L_{\mathrm{A}}(r)$. It is an approximation of the Huber loss function[27], which is quadratic when $r \leq c$ and linear otherwise. In contrast to the Huber loss function, $L_{\mathrm{PH}}(r, c)$ is differentiable everywhere, and therefore more suitable for 'out of the box' optimizers, such as the ones encountered in SciPy. The parameter $c$ determines the scale at which outliers are treated; as $c$ is lowered, the effect of outliers is reduced.

Figure S5 shows how the different loss functions changed the fitted ANN on the BPTI GCMC data, and Table S10 shows how those fits affect the calculated free energies. Free energies corresponding to the ANN that had been fitted using $L_{\mathrm{A}}$ are only approximate, as the optimization was carried out using gradient descent, which is not well suited to this loss function.

The parameter $c$ for the pseudo Huber loss function was experimented with until the fits were qualitatively improved over those produced by squared loss. At the end state of three coupled water molecules, the pseudo Huber and absolute loss functions are in good agreement, but differ by almost 1 kcal/mol with the free energy calculated with squared loss. The free energies reported in the main text are calculated using the pseudo Huber loss function with $c = 0.1$.

Figure S5 shows that the GCMC data for BPTI is noisy. This is due to poor insertion *and* deletion rates, where the latter is responsible for the significant outlier at $B = -28.5$. In the first set of simulations at $B = -28.5$, there were an average of 1.6 waters in the cavity. A repeat of the same $B$ value produced an average of 0.0 waters in the cavity, indicating that molecule trapping (slow deletion rates), rather than slow insertion may have the most significant effect. While techniques such as the cavity bias method[28] can increase insertion rates, deletion rates can also be small. As demonstrated with BPTI, GCMC titrations with ANN fitting alleviates the effects of poor sampling, as all $B$ values affect the line of best fit, and the detrimental effect of outliers can be lessened with loss functions that are robust to outliers.



(a)

(b)

(c)

(d)

Figure S5: GCMC titration data (black dots) for three site water cavity in BPTI. Artificial neural network fitted with squared error (a) mean absolute error (b) pseudo Huber, $c = 1$ (c) and pseudo Huber $c = 0.1$ (d) loss functions, the latter of which was used to calculate the free energies shown in the main text. The squared error loss function produces a qualitatively poor fit, with the line pulled by the outliers around $B \approx -28$, and the step between $B = -20$ and $B = -15$ fitted as a straight line. The other loss functions capture two steps more cleanly and are not adversely affected by the noisy values around $B \approx -28$.

## References

[1] Panagiotopoulos, A. Z. *Mol. Phys.* **1987**, *61*, 813.

[2] Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300.

[3] D. A. McQuarrie, *Statistical Mechanics*; Harper & Row: USA, 1976.

[4] Widom, B. *J. Chem. Phys.* **1963**, *39*, 2808.

[5] Vaitheeswaran, S.; Rasaiah, J. C.; Hummer, G. *J. Chem. Phys.* **2004**, *121*, 7955.

[6] Vaitheeswaran, S.; Yin, H.; Rasaiah, J. C.; Hummer, G. *Proc. Natl. Acad. Sci.* **2004**, *101*, 17002.

[7] Adams, D. J. *Mol. Phys.* **1974**, *28*, 1241.

[8] Guarnieri, F.; Mezei, M. *J. Am. Chem. Soc.* **1996**, *118*, 8493.

[9] Clark, M.; Guarnieri, F.; Shkurko, I.; Wiseman, J. *J. Chem. Inf. Model.* **2005**, *46*, 231.

[10] Naim, A. B.; Marcus, Y. *J. Chem. Phys.* **1984**, *81.*

[11] Resat, H.; Mezei, M. *Biophys. J.* **1996**, *71*, 1179.

[12] Speidel, J. A.; Banfelder, J. R.; Mezei, M. *J. Chem. Theory Comput.* **2006**, *2*, 1429.

[13] Malasics, A.; Gillespie, D.; Boda, D. *J. Chem. Phys.* **2008**, *128*, 124102.

[14] Lakkaraju, S. K.; Raman, E. P.; Yu, W.; MacKerell, A. D. *J. Chem. Theory Comput.* **2014**, *10*, 2281.

[15] Clark, M.; Meshkat, S.; Wiseman, J. S. *J. Chem. Inf. Mod.* **2009**, *49*, 934.

[16] Bodnarchuk, M. S.; Viner, R.; Michel, J.; Essex, J. W. *J. Chem. Inf. Model.* **2014**, *54*, 1623.

[17] Deng, Y.; Roux, B. *J. Chem. Phys.* **2008**, *128*, 115103.

[18] Zia, R. K. P.; Redish, E. F.; McKay, S. R. *Am. J. Phys.* **2009**, *77*, 614.

[19] Puibasset, J. *J. Phys. Chem. B* **2005**, *109*, 480.

[20] Fan, C.; Do, D. D.; Nicholson, D.; Ustinov, E. *Mol. Phys.* **2013**, *112*, 60.

[21] Key, E. *College Math. J.* **1994**, *25*, 136.

[22] Renshaw, E. *Modelling biological populations in space and time*; Cambridge University Press: UK, 1991.

[23] Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 7683.

[24] van der Walt, S.; Colbert, S. C.; Varoquaux, G. *Comput. Sci. Eng.* **2011**, *13*, 22.

[25] Jones, E.; Oliphant, T.; Peterson, P. et al. *SciPy: Open source scientific tools for Python*, 2001–, [Online; accessed 2014-12-15].

[26] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision, Second Edition*, Cambridge University Press: UK, 2004.

[27] Huber, P. J. *Ann. Math. Stat.* **1964**, *35*, 73.

[28] Mezei, M. *Mol. Phys.* **1980**, *40*, 901.