Supplementary Information for Ligand binding free energies with adaptive water networks: two-dimensional grand canonical alchemical perturbations

Hannah E. Bruce Macdonald,[†] Christopher Cave-Ayland,[†] Gregory A. Ross,[‡] and Jonathan W. Essex[*,†]

School of Chemistry, University of Southampton, Southampton, SO17 1BJ, UK, and Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, New York 10065, United States

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

Simulation protocols

Table S.1: Details of GCMC region used for each system. The GCMC region is cuboidal. B_{eq} is calculated from the GCMC volume using Equation 2

System	origin (x,y,z)	length (x,y,z) /Å	Volume /Å^3	B_{eq}
SD	24.141, 11.225, 32.916	4.000, 4.000, 4.000	64.0	-9.70
A_{2A}	-44.253, 0.565, -47.602	9.784, 6.533, 7.844	501.4	-7.65

[*]To whom correspondence should be addressed
[†]University of Southampton
[‡]MSK

Table S.2: MC move ratios for each simulation performed. A - indicates that no GCMC type moves were performed.

Simulation	solvent	protein	solute	GC insertion	GC deletion	GC sampling	n equilibration / M	n production / M
SD AP	280	218	2	-	-	-	5	40
SD DD	280	218	2	-	-	-	5	40
SD GCAP	280	218	2	167	167	167	5*	80
A_{2A} GCAP	376	118	7	167	167	167	10*	120
A_{2A} naïve	376	118	7	-	-	-	10	60

Move ratios and number of MC moves performed for each type of simulation. n MC moves performed quoted in number of million (M) steps. Classical alchemical perturbations are indicated as AP. An asterisk indicates that simulations involving GCMC moves have two equilibration steps, each of the number of steps quoted. The initial equilibration stage is purely GC moves (insertion, deletion and GC water sampling equally), while the second is with the move ratios presented).

Table S.3: B value ranges for 2D-GCAP simulations, where B_{min} and B_{max} are inclusive. Interval shows the distance between neighbouring B values and N_B is the number of B values simulated.

System	B_{min}	B_{max}	Interval	N_B
SD lig 1 + 3	-19.7	-3.7	1	19
SD lig 1 + 2	-18.7	-9.7	1	10
SD lig 2 + 3	-12.7	-3.7	1	10
A_{2A} (all pairs)	-21.15	-7.65	1.5	10

stepwise perturbation of EG leg

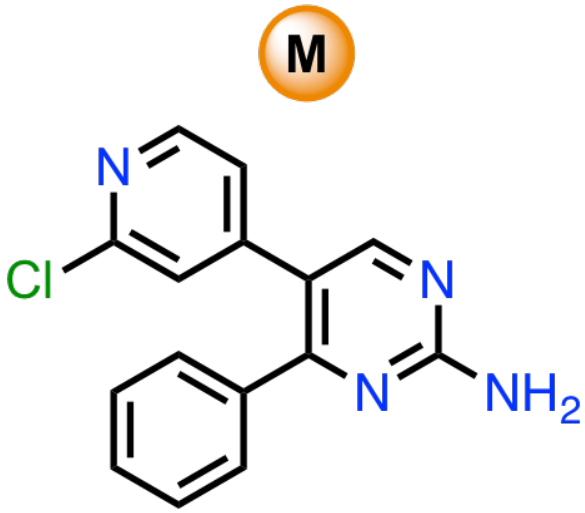

Figure S.1: Ligand M (for mutant), not included in the published dataset,[1] but used as the mid-point for the E - G leg, as this perturbation requires both the growing and shrinking of different R groups. It is more straightforward to calculate the relative free energy of both E - M and G - M and use this to calculate the E - G leg. M was calculated to have lower affinity than any of E, F or G.

hydration free energies

Table S.4: Relative free energy perturbations for ligands in the gas phase, and bulk solvent phase. ΔG_{hyd} is the relative free energy of hydration of the two ligands, calculated from ΔG_{sol} - ΔG_{gas} . ΔG_{sol} is used to calculate ΔG_{bind}. All energies are in kcal·mol^{-1}. Energies and standard errors for SD are calculated using MBAR from four repeats, and A_{2A} from three.

Perturbation	ΔG_{gas}	ΔG_{sol}	ΔG_{hyd}
2 to 1	-100.9 (0.0)	-101.1 (0.1)	-0.1 (0.1)
2 to 3	-12.5 (0.0)	-11.5 (0.1)	1.0 (0.1)
3 to 1	-90.122 (0.0)	-91.3 (0.1)	-1.2 (0.1)
F to E	-5.7 (0.0)	-4.9 (0.2)	0.8 (0.2)
F to G	-43.9 (0.0)	-43.7 (0.1)	0.2 (0.1)
E to G	-38.7 (0.1)	-39.0 (0.2)	-0.4 (0.2)

A_{2A} experimental binding affinities

The original publication of the A_{2A} ligand set considered herein[1] provides both K_i and K_D results for the set of ligands. As the free energy accuracy aimed for is typically 1 kcal mol^{-1}, we wanted to select a set of ligands where the relative free energies were within 1 kcal mol^{-1} for demonstrating the GCAP methodology. The relative experimental free energies were considered as this reduces any possible systematic differences between the two measurements.

$$\Delta G_{K_D} = -k_B T ln(K_D) \tag{S.1a}$$

$$\Delta G_{K_i} = -k_B T ln(K_i) \tag{S.1b}$$

So the difference in the relative free energy for a pair of ligands (x and y), between the two methods, can be calculated from:

$$\Delta\Delta G(x-y)_{K_D} - \Delta\Delta G(x-y)_{K_i} = k_B T ln\left(\frac{K_D(x)}{K_D(y)}\right) - k_B T ln\left(\frac{K_i(x)}{K_i(y)}\right) \tag{S.2}$$

If the absolute value of Equation S.2 is less than 1 kcal mol^{-1} then the perturbation was considered for GCAP simulations. As crystal structures are only available for ligand G and E, any ligands where the binding mode was unclear, i.e. where either ring A or ring B was asymmetrically substituted, were excluded, as the ligand may bind in either orientation. This excludes ligands B, J, K and L. Ligand E is asymmetrically substituted, but the binding mode is available from the crystallographic structure. Of the 8 remaining ligands, only 7 have published data for both K_i and K_D. This results in 42 possible pairs of ligands. Of the 42 pairs, only 8 pairs satisfied the requirement that Equation S.2 was less than $\pm$ 1 kcal mol^{-1}; EF, EG, EH, EI, FG, GH, GI and HI. Of these, the ligands E, F and G were chosen as both ligands E and G have crystallographic structures available, and the differences in

the ligand seem significant enough to displace or disrupt active site water molecules.

Naïve solvation of A_{2A}

For comparison to other available methods, the A_{2A} simulations were also performed with a naïve solvation. The naïve simulation refers to the system being set up using ProtoMS set up tools, where the system is solvated based the available pocket volume and simulated with the NVT ensemble. The set-up places three water molecules within the GCMC region, shown in Figure S.2. The water molecules will be sampled with solvent MC steps.

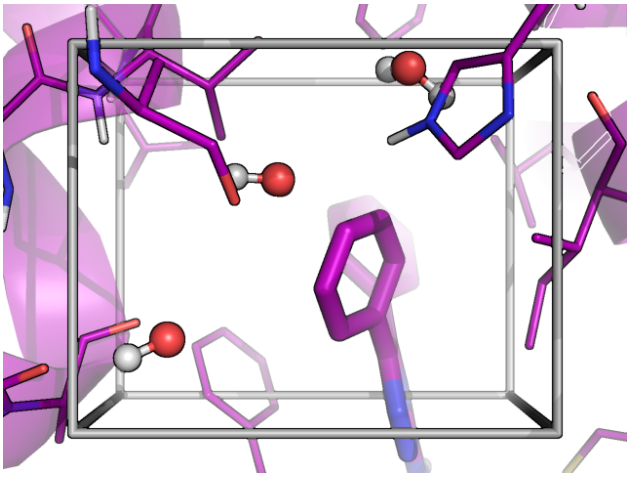

Figure S.2: The initial placement of water molecules in the naïve solvation simulations. This naïve solvation is used with all ligands, but an unsubstituted scaffold is shown for clarity. The GCMC box is not included in the naïve simulations, but is shown in light grey for ease of comparison to Figure 11.

GCAP hydration of A_{2A}

2D-GCAP results

To create the free energy surfaces, PMFs are calculated along B using GCI, and along λ using thermodynamic integration. These are combined to generate a free energy surface using least-squares fitting. For PMFs along B, free energy values for states with non-integer occupancies are determined by linear interpolation of the binding free energy curves output by GCI.

In principle it is possible to calculate free energy surfaces directly using MBAR. The free energies produced between states with differing B values will include contributions from changes in chemical potential however that are not physically meaningful in the context of the binding free energies of interest in this work. The above approach produces consistent Helmholtz free energy surfaces using GCI. MBAR free energy differences between states at the same B value are consistent with NVT free energy cycles using sequential ligand perturbations and water molecule double decoupling.

In all cases, $\lambda = 0$ corresponds to the larger ligand, and $\lambda = 1$ to the smaller.

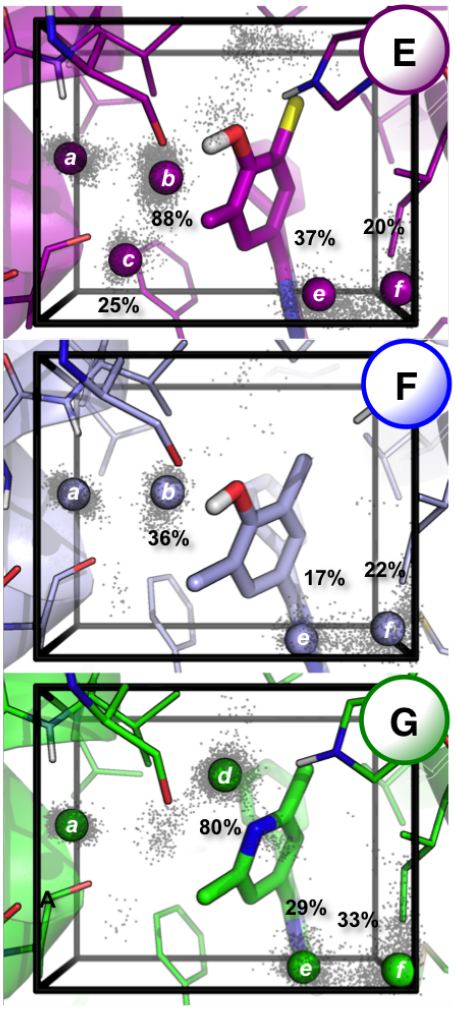

Figure S.3: GCMC cluster locations top to bottom for ligands E (purple), F (light blue) and G (green) shown as sticks. Protein is represented as cartoon, with residues shown as lines. Carbon atoms are colored per ligand, with oxygen (red), nitrogen (dark blue), chlorine (yellow) and hydrogen (white). Any non-polar hydrogen atoms are removed for clarity. GCMC cluster centers have been labelled $a - f$, with water occupancies labelled for waters that are present $< 95\%$ of the simulation. All of the GCMC water oxygens from the simulation of each ligand are shown as small grey points. Cluster centres (large spheres) have been calculated using clustering and density scripts available in ProtoMS.[2] Cluster centers e and f are consistent for all three ligands so have been excluded from the main text.

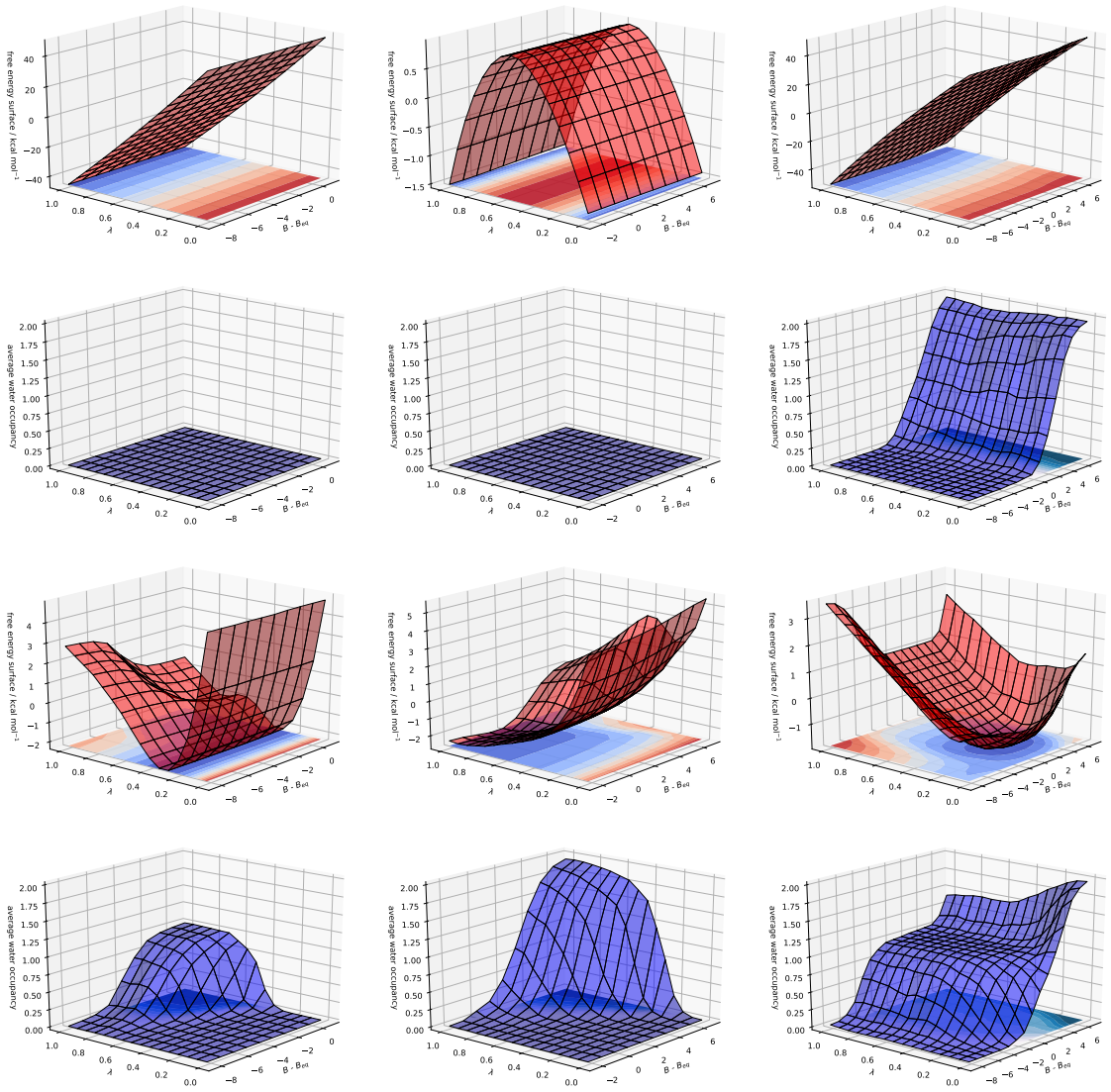

Figure S.4: 2D-GCAP results for SD. Columns left to right: Ligands 1-2, 2-3, 3-1. Rows top to bottom: electrostatic surface, electrostatic solvation, van der Waals surface, van der Waals solvation.

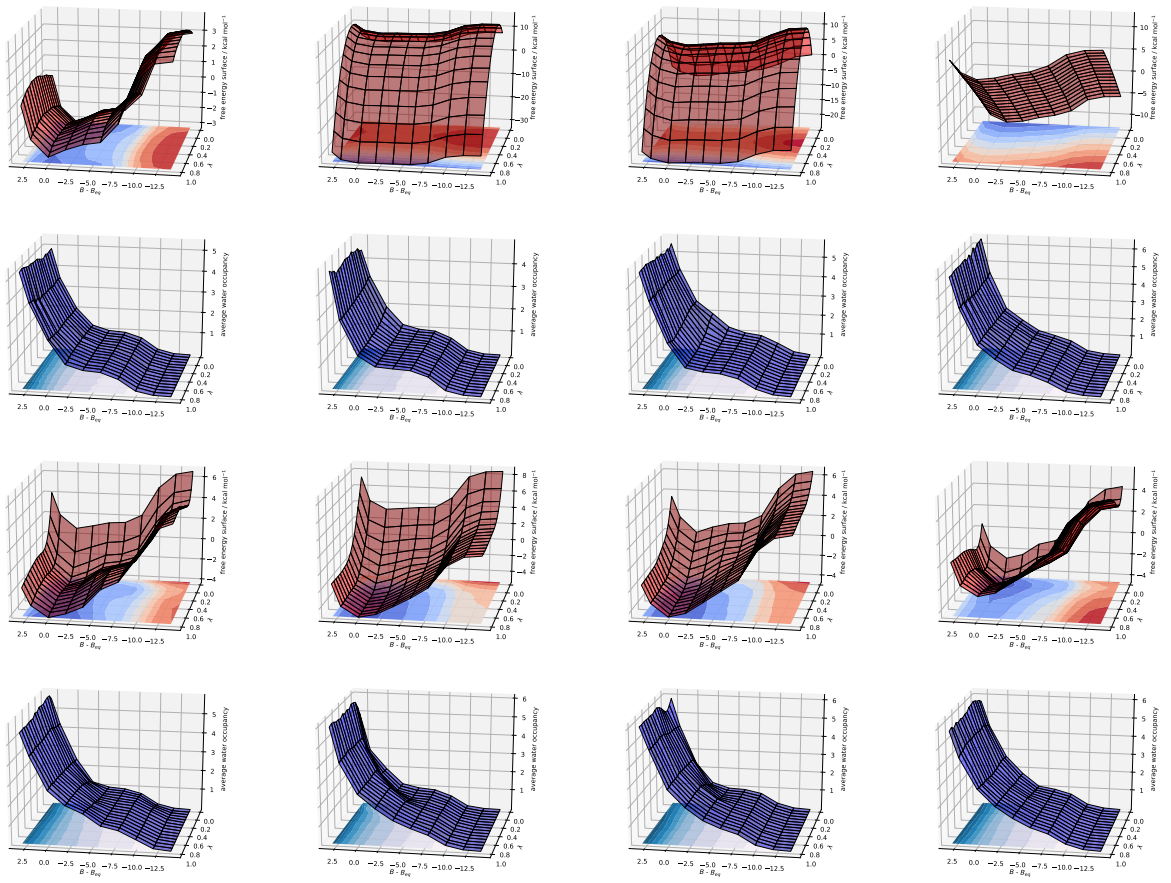

Figure S.5: 2D-GCAP results for A_{2A}. Columns left to right: Ligands F-E, F-G, E-M, G-M. Rows top to bottom: electrostatic surface, electrostatic solvation, van der Waals surface, van der Waals solvation.

A_{2A} His278 protonation

The active site histidine (His278) was ϵ protonated by Maestro set up tools. As the residue is in close proximity to the ligand and the GCMC region, the simulations were repeated also with the δ protonation state. The results of this are shown in Figure S.6. This changes the rank ordering of the ligands, with ligand E stabilised, and ligand G destabilised. The relative destabilisation of ligand G may be rationalised as it is the only complex that contains a His278-water hydrogen bond, Figure 11. As the ϵ protonated form was suggested in the set up, and has significantly better experimental agreement, this was shown in the main text. The δ protonated results show how sensitive results can be to choices made in the system set up - whether that be location of water molecules (demonstrated by the naïve results in the manuscript), and by the effect of protonation here. While it may be possible to sample the alternate rotamers through simulation, the flip requires is unlikely to be sampled during MC simulations. Simulations starting from the alternate histidine rotamers have not been considered herein. Ideally, GCAP would be performed with a constant-pH protocol that would exchange the titratable active site residues within the simulation; however this is beyond the scope of the paper.

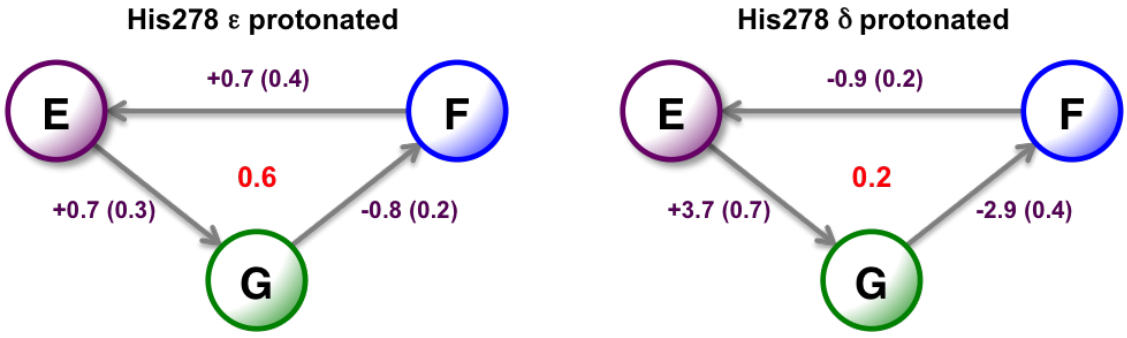

Figure S.6: Relative ligand binding free energies for A_{2A} ligands with the two protonation states of active site residue, His278

References

(1) Congreve, M.; Andrews, S. P.; Doré, A. S.; Hollenstein, K.; Hurrell, E.; Langmead, C. J.; Mason, J. S.; Ng, I. W.; Tehan, B.; Zhukov, A.; Weir, M.; Marshall, F. H. Discovery of 1,2,4-triazine derivatives as adenosine A(2A) antagonists using structure based drug design. *J. Med. Chem* **2012**, *55*, 1898–1903.

(2) ProtoMS. 1999; `www.protoms.org`.