Electronic supplementary information for:

Computation of Protein-Ligand Binding Free Energies using Quantum Mechanical Bespoke Force Fields

Daniel J. Cole,[a] Israel Cabeza de Vaca,[b] William L. Jorgensen[b]

[a] School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom.

[b] Department of Chemistry, Yale University, New Haven, Connecticut 06520-8107, United States.

S1. FORCE FIELD DERIVATION

DFT Calculation. Ground state electron densities of the lysozyme protein (2628 atoms), and the six small molecules, were computed using the ONETEP linear-scaling density functional theory (DFT) code [1] with the PBE exchange-correlation functional. ONETEP uses a basis set of spatially truncated nonorthogonal generalized Wannier functions (NGWFs) localized on each atom [2]. Four NGWFs were used for each atom (except hydrogen, which used one). The NGWFs were expanded in a periodic sinc (psinc) basis with a plane-wave energy cutoff of 980 eV and were spatially truncated with localization radii of 10 Bohr. Core electrons were treated using OPIUM norm-conserving pseudopotentials [3]. To account for induction effects in the condensed phase, the electron density was computed using an implicit solvent model, in which the dielectric cavity is defined by an isosurface of the vacuum electron density [4]. The dielectric constant outside the cavity was set to 10 for the lysozyme protein, and 4 for the small molecules. Partitioning of the polarized electron density was performed using the DDEC atoms-in-molecule method in ONETEP, with a mixing parameter (γ) of 0.02 [5].

Non-bonded parameters. Charges and Lennard-Jones parameters were obtained directly from the partitioned atomic electron density using methods described in detail elsewhere [6]. In brief, the charges are obtained simply by integrating the atomic electron density over all space and subtracting the nuclear charge. Only atom-centered charges were used in this study. In classical force fields, short-ranged repulsion due to electron density overlap and longer ranged, attractive van der Waals (or dispersion) interactions are modelled by the Lennard-Jones potential:

$$E_{LJ}=\sum\left(\frac{A^{ij}}{r_{ij}^{12}}-\frac{C_6^{ij}}{r_{ij}^{6}}\right)$$

where r_{ij} is the interatomic distance and A and C_6 are parameters that determine the strengths of the interactions. Following Ref. [6], the C_6 parameters are determined using the Tkatchenko-Scheffler relation, by re-scaling the C_6 coefficient of an atom in vacuum by its relative atomic volume in the molecule (itself computed from the QM atomic electron density) [7]. Heteronuclear parameters are determined via a geometric combining rule:

$$C_6^{ij}=\sqrt{\left(C_6^i C_6^j\right)}$$

In this way, the C_6 parameters are derived directly from the QM electron density, rather than fitting to experiment. In Figure 1 in the main text, the summed C_6 parameters per residue are given by:

$$C_6 = \sum C_6^i$$

The A coefficients are determined from C_6 and the van der Waals radius of an atom in the molecule using methods discussed in Ref. [6].

Bonded parameters. For the lysozyme protein, bond and angle force field equilibrium values and harmonic force constants are assigned from the library developed in Ref. [8], and backbone and main side chain torsion parameters have been fit to reproduce quantum mechanical torsional scans in Ref. [9]. Remaining side chain torsional parameters and improper terms are taken from the OPLS-AA/M force field [10]. For the ligands, bond and angle parameters were derived using the modified Seminario method [8], as implemented in the QUBEKit software [11]. The QM Hessian matrix was computed using the ωB97X-D exchange-correlation functional with a 6-311++G(d,p) basis set. Due to the rigidity of the molecules studied here, all torsional parameters were taken from the OPLS force field [12].

S2. FREE ENERGY CALCULATIONS

Protein-ligand systems were prepared using the T4-lysozyme L99A protein in complex with benzene (PDB ID: 4W52) as a reference [13]. Crystal water molecules and counter ions were removed. For the starting structures of the remaining analogs, benzene was replaced using maximum overlap of the non-hydrogen atoms. The calculations of absolute binding free energies with the free energy perturbation (FEP) method were carried out using the MCPRO software (version 3.2, modified to accept protein-specific non-bonded parameters) [14] by annihilating the ligand both unbound in water and bound to the protein using the single topology method. Annihilation was performed by decoupling the electrostatic and the Lennard-Jones (LJ) terms. First the charges were turned off linearly with the λ parameter, followed by the LJ terms using the 1-1-6 soft-core potential [15].

During the LJ annihilation process, a hard-wall (HW) potential constraint was applied to keep the ligand in the binding site, thus restraining the movement to a sphere of radius 2.8 Å located at the benzene geometric center of the initial X-ray structure. This radius was chosen to allow the ligand exploration of the entire binding site during the simulation [16]. Due to the HW constraint, a correction term was included in the binding free energy estimation:

$$\Delta G_{HW} = -k_B T ln\left(\frac{V_{eff}}{V_0}\right)$$

where V_{eff} is the effective volume of ligand, that is, the volume of the HW sphere and V_0 is 1660 Å^3 (corresponding to a 1 M standard state), T is the temperature, and k_B is the Boltzmann constant [17]. The final absolute binding free energy was computed using:

$$\Delta G = \Delta G_{unbound} - \Delta G_{bound} + \Delta G_{HW}$$

The complexes were solvated in a 25 Å water sphere, with 1514 TIP4P water molecules, centered on the binding site. Water molecules in close contact with the protein/ligand complex were automatically removed using the MCPRO software. Nonbonding energy terms used a 10 Å cutoff. Side-chain and ligand move frequencies were set to 10 and 3. Backbone move frequencies were set to 7 and 11 for concerted rotations with angles (CRA) and pivot [16]. CRA parameters parameters c_1, c_2, and c_3 were assigned values of 100, 8, and 20, respectively. The unbound ligand simulations were performed in a sphere of 1500 water molecules where the ligand move frequency was 60 configurations; the remaining parameters were kept the same as in the bound simulations.

The protein and ligand energetics were described using the QUBE force field (Section S1), and water with TIP4P. Charge and LJ annihilations used 15 and 18 λ windows of simple overlap sampling respectively [18]. For *p*-xylene and *o*-xylene, for which reorientation of residue V111 is expected, each window consisted of 20 million (M) configurations of equilibration and 60M configurations of averaging for the bound simulations. For the remaining analogs, 5M/25M configurations of equilibration/averaging was used and the simulations were run in triplicate. All unbound simulations were run for 20M/60M configurations. Figure S1 shows the evolution of various components of the binding free energy with simulation length for lysozyme-benzene and lysozyme-*p*-xylene. In agreement with Ref. [16], the free energy is generally well-converged after around 25M averaging steps, though reorientation of the V111 side chain can take longer (hence the longer MC runs for *p*-xylene and *o*-xylene). Table S3 confirms that the results do not change by more than 0.3 kcal/mol even if the second half of the averaging stages of the bound simulations are discarded.

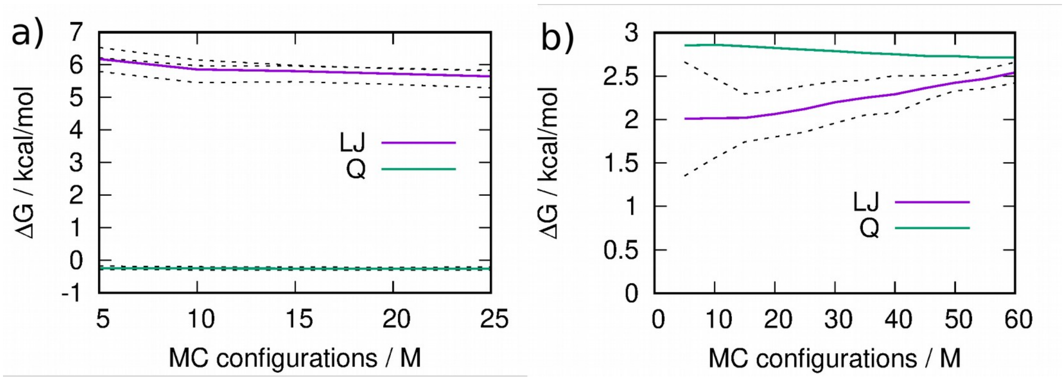

Figure S1. *Evolution of the average Lennard-Jones (LJ) and charge (Q) annihilation free energy for the bound leg of the a) lysozyme-benzene and b) lysozyme-p-xylene FEP simulations. Where multiple simulations have been run, the individual runs are shown as dashed lines.*

Hydration free energies (Table S2) were computed using:

$$\Delta G = \Delta G_{gas} - \Delta G_{unbound} + LRC$$

where ΔG_{gas} is obtained by decoupling the charge interactions in the gas phase using 15 λ windows of simple overlap sampling and 1M/5M Monte Carlo steps per window. LRC is a *post hoc* long range correction to account for van der Waals interactions neglected beyond the cut-off [19].

Enhanced Sampling. The replica exchange with solute tempering (REST) method increases conformational sampling in defined, localized regions of the system, thereby improving the consistency of the calculation by reducing the dependence of the free energy on the starting structure [20]. The REST method is implemented in MCPRO as described in Refs. [21,22]. At each λ window, four replicas were run in parallel with REST enhanced sampling applied to the ligand and the protein residue V111. For residues in the REST region, the dihedral rotation and non-bonded force field terms are effectively re-scaled to reduce potential energy barriers in high temperature replicas of the system. REST scaling factors were chosen to be exponentially distributed (25, 86, 160, 250 ^{o}C). Exchange attempts between pairs of neighboring replicas were attempted every 10 000 MC steps. Free energy changes were computed from the room temperature ensemble. The REST protocol was used alongside the 'flip' algorithm in which selected dihedral angles undergo attempted jumps that are much larger than typical MC moves [21,23]. The flip algorithm was applied to the side chain χ_1 dihedral angle in V111. The jumps were of random size in the range 60° to 300°.

Table S1. *Full comparison between experimental and computed protein-ligand binding free energies for a range of literature force fields.* [a]*Using the OPLS force field [16].* [b]*Using the AMBER/AM1BCC force field and the confine-release protocol [24].*

	ΔG^{exp}	**ΔG^{QUBE}**	**ΔG^{a}**	**ΔG^{b}**
benzene	-5.19	-5.97	-7.68	-3.95
p-xylene	-4.67	-4.41	-4.98	-3.59
o-xylene	-4.60	-4.98	-2.90	-3.23
benzofuran	-5.46	-6.95	-7.21	-3.66
indole	-4.89	-3.84	-4.35	-1.37
indene	-5.13	-4.01	-5.87	-1.63
MUE		**0.85**	**1.26**	**2.09**

Table S2. *Computed hydration free energies (kcal/mol), compared with experiment (where available).*

	ΔG^{exp} [25]	ΔG^{QUBE}
benzene	-0.86	-0.23
p-xylene	-0.80	+1.60
o-xylene	-0.90	+2.05
benzofuran	N/A	+1.01
indole	N/A	-3.63
indene	N/A	+0.68

Table S3. *Test of the convergence of computed protein-ligand binding free energies with number of MC steps. $\Delta G^{25M/60M}$ uses the full dataset (as reported in Table 1), and $\Delta G^{12.5M/30M}$ uses the first half of the simulations only. The maximum change in free energies over the second half of the simulation is less than 0.3 kcal/mol.*

	ΔG^{exp}	$\Delta G^{25M/60M}$	$\Delta G^{12.5M/30M}$
benzene	-5.19	-5.97	-6.15
p-xylene	-4.67	-4.41	-4.15
o-xylene	-4.60	-4.98	-5.03
benzofuran	-5.46	-6.95	-6.75
indole	-4.89	-3.84	-3.86
indene	-5.13	-4.01	-3.80
MUE		**0.85**	**0.93**

REFERENCES

[1] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne, *J. Chem. Phys.*, 2005, **122**, 084119.

[2] C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Diequez, M. C. Payne, *Phys. Rev. B.*, 2002, **66**, 035119.

[3] For information about the Opium pseudopotential generation project, see: http://opium.sourceforge.net/index.html

[4] J. Dziedzic, H. H. Helal, C.-K. Skylaris, A. A. Mostofi, M. C. Payne, *EPL*, 2011, **95**, 43001.

[5] L. P. Lee, N. G. Limas, D. J. Cole, M. C. Payne, C.-K. Skylaris, T. A. Manz, *J. Chem. Theory Comput.*, 2014, **10**, 5377.

[6] D. J. Cole, J. Z. Vilseck, J. Tirado-Rives, M. C. Payne and W. L. Jorgensen, *J. Chem. Theory Comput.*, 2016, **12**, 2312.

[7] A. Tkatchenko, M. Scheffler, *Phys. Rev. Lett.*, 2009, **102**, 073005.

[8] A. E. A. Allen, M. C. Payne, D. J. Cole, *J. Chem. Theory Comput.*, 2018, **14**, 274.

[9] A. E. A. Allen, M. J. Robertson, M. C. Payne, D. J. Cole, 2019, https://doi.org/10.26434/chemrxiv.7565222.v1

[10] M. J. Robertson, J. Tirado-Rives, W. L. Jorgensen, *J. Chem. Theory Comput.*, 2015, **11**, 3499.

[11] J. T. Horton, A. E. A. Allen, L. S. Dodda and D. J. Cole, 2018, https://doi.org/10.26434/chemrxiv.7247045.v1

[12] W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, *J. Am. Chem. Soc.*, 1996, **118**, 11225.

[13] M. Merski, M. Fischer, T. E. Balius, O. Eidam, B. K. Shoichet, *PNAS*, 2015, **112**, 5039.

[14] W. L. Jorgensen, J. Tirado-Rives, *J. Comput. Chem.*, 2005, **26**, 1689.

[15] T. Steinbrecher, D. L. Mobley, D. A. Case, *J. Chem. Phys.*, 2007, **127**, 214108.

[16] I. Cabeza de Vaca, Y. Qian, J. Z. Vilseck, J. Tirado-Rives and W. L. Jorgensen, *J. Chem. Theory Comput.*, 2018, **14**, 3279.

[17] J. Hermans, L. Wang, *J. Am. Chem. Soc.*, 1997, **119**, 2707.

[18] W. L. Jorgensen, L. L. Thomas, *J. Chem. Theory Comput.,* 2008, **4**, 869.

[19] W. L. Jorgensen, J. D. Madura, C. J. Swenson, *J. Am. Chem. Soc.*, 1984, **106**, 6638-6646.

[20] L. Wang, B. J. Berne and R. A. Friesner, *PNAS*, 2012, **109**, 1937–1942.

[21] D. J. Cole, J. Tirado-Rives and W. L. Jorgensen, *J. Chem. Theory Comput.*, 2014, **10**, 565–571.

[22] D. J. Cole, M. Janecek, J. E. Stokes, M. Rossmann, J. C. Faver, G. J. McKenzie, A. R. Venkitaraman, M. Hyvönen, D. R. Spring, D. J. Huggins and W. L. Jorgensen, *Chem. Commun.*, 2017, **53**, 9372–9375.

[23] L. L. Thomas, T. J. Christakis, W. L. Jorgensen, *J. Phys. Chem. B,* 2006, **10**, 21198.

[24] D. L. Mobley, A. P. Graves, J. D. Chodera, A. C. McReynolds, B. K. Shoichet and K. A. Dill, *J. Mol. Biol.*, 2007, **371**, 1118.

[25] D. L. Mobley and J. P. Guthrie, *J. Comput. Aided Mol. Des.*, 2014, **28**, 711-720.