Electronic Supplementary Material (ESI) for Faraday Discussions.

Electronic Supporting Information for: A machine learning based intramolecular potential for a flexible organic molecule

Daniel J. Cole,[†] Letif Mones,[‡] and Gábor Csányi[*,‡]

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

[‡]*Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom*

E-mail: gc121@cam.ac.uk

S1 Computational Methods

The initial protein-ligand coordinates were constructed from the 1W7H[1] and 3FTY[2] PDB files. For p38 kinase, the 213 residues closest to the ligand were retained. For leukotriene A4 hydrolase, 309 protein residues were retained, and the configuration of the ligand was adjusted by inspection with reference to the QM data as explained in the main text. Monte Carlo simulations were performed using the MCPRO software package.[3] In both cases, all protein residues within 12.5 Å of the ligand were free to move, and backbone motions were controlled using the concerted rotation algorithm.[4] The protein-ligand complexes were solvated in 25 Å water caps and the JAWS algorithm[5] was run to determine initial water molecule distributions around the ligand. For the simulations of 3BPA in water, the molecule was solvated in a 25 Å water cap. Following standard MCPRO protocols, a solute move is undertaken every 5 steps in the bound simulations, and every 60 steps in the unbound.

Free energy perturbation (FEP) calculations employed 11 λ windows of simple overlap sampling[6] to perturb from the GAP to the OPLS/CM1A potentials as described in the main text. At each λ window, four replicas of the system were run in parallel with replica exchange with solute tempering (REST) enhanced sampling applied to the ligand.[7–9] Both bound and unbound Monte Carlo (MC) simulations comprised 10 million (M) configurations of equilibration, and 30 M configurations of averaging. Figure S1 shows the convergence of the relative binding free energy correction with respect to the number of MC steps. In the REST approach, high temperature replicas of the system facilitate crossing of any high energy barriers to sampling, and replica exchange ensures correct Boltzmann sampling in the room temperature replica. REST temperature scaling factors were chosen to be exponentially distributed (25, 86, 160, 250 ^oC), and exchange attempts were made every 10 000 MC steps. Each REST replica runs on a separate cpu, and each λ window requires around 60 hrs using GAP-v2, compared to around 20 hrs for OPLS. Note that ligand Monte Carlo moves are not attempted at every step, which is why the GAP simulation is only a factor of three slower using the current set up.

Representative snapshots from the Monte Carlo simulations were generated using the Bio3D software package.[10] A principal component analysis (PCA) was performed on the ensemble of computed structures, and a cluster analysis was performed in the space of the first two PCs, separating the frames into two clusters. The average structure of each group was computed, and a representative snapshot was selected that is closest (smallest RMSD) to the average structure. Figure 5 in the main text displays the representative snapshot from cluster 1 of each simulation overlaid on the relevant experimental crystal structure.

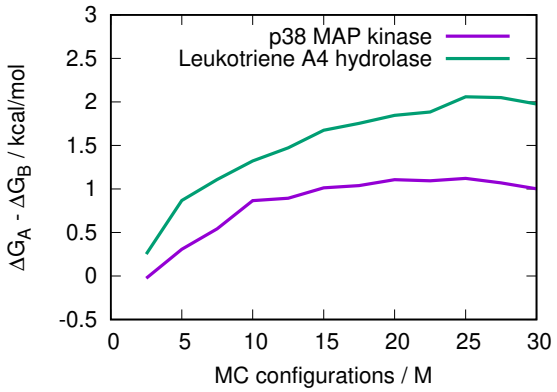

Figure S1: Convergence of the correction to the binding free energy (eq 5 of main text) with respect to the number of Monte Carlo configurations sampled during FEP.

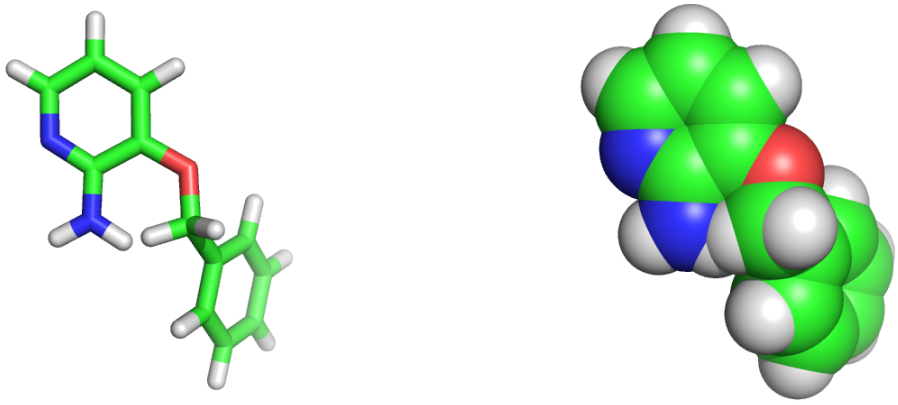

Figure S2: Stick and space-filling representations of the small molecule, 3BPA, extracted from its complex with leukotriene A4 hydrolase (PDB: 3FTY). An unphysical clash between the aminopyridine and –CH_2– linker is evident.

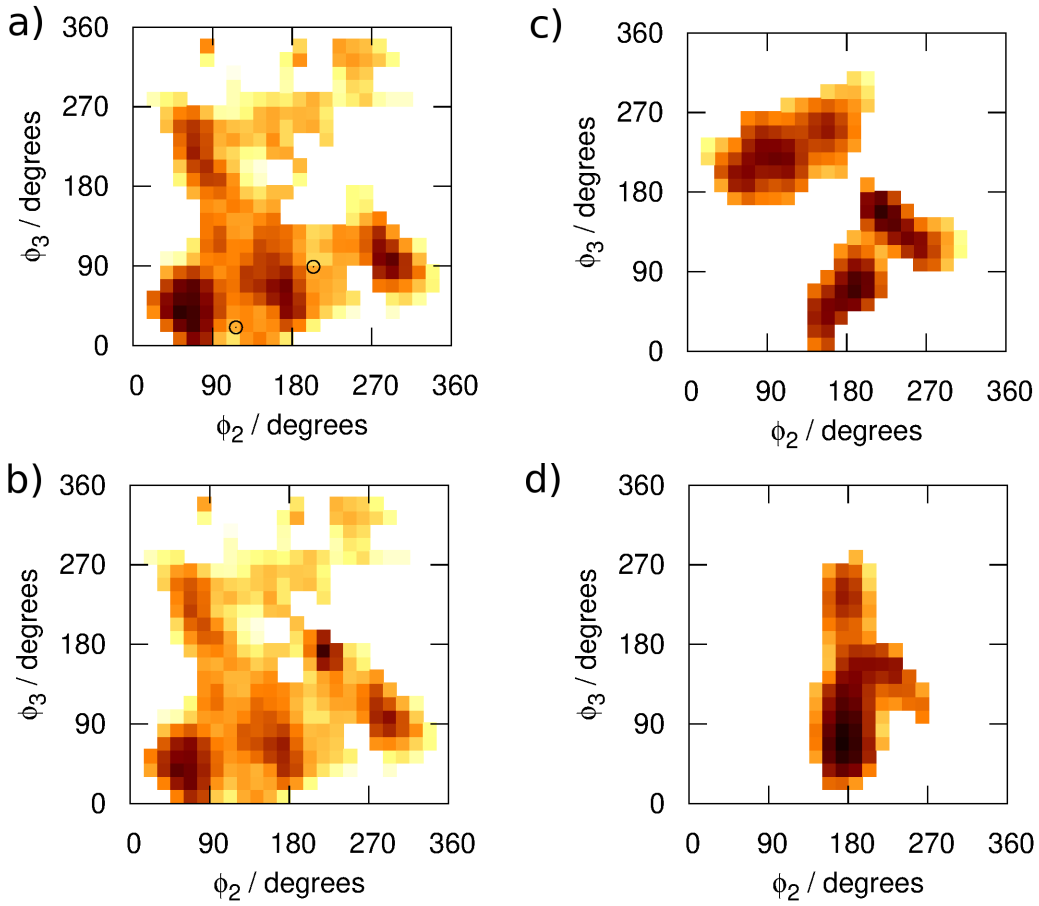

Figure S3: Distributions of the ϕ_2 and ϕ_3 dihedral angles (plotted as $\log(p_{\phi_2,\phi_3})$) sampled in (a) training set 1, (b) training set 2, (c) MC simulations with GAP-v2, and (d) MC simulations with OPLS. Similar to Figure 3 of the main text, GAP-v2 samples a large area of conformational space, which is well covered by the training data.

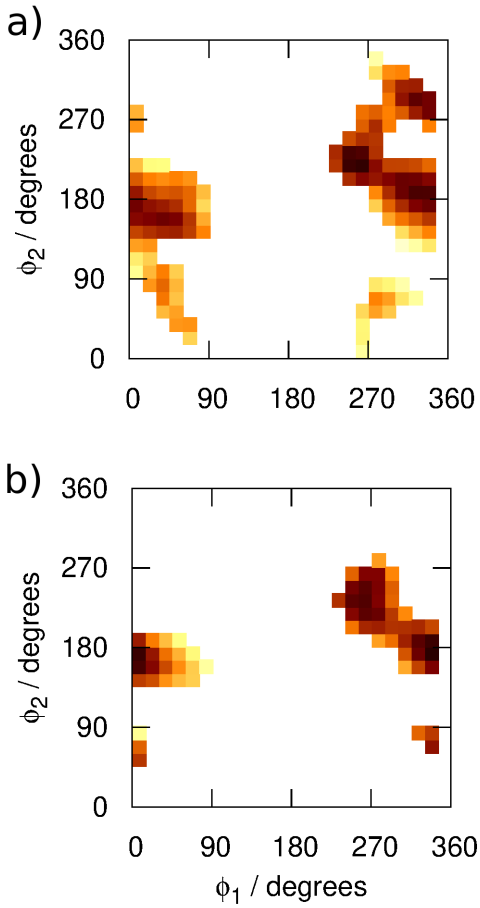

Figure S4: Distributions of the ϕ_1 and ϕ_2 dihedral angles (plotted as $\log(p_{\phi_1,\phi_2})$) sampled in (a) duplicate MC simulations with GAP-v2, and (b) duplicate MC simulations with OPLS. Distributions are similar to those shown in Figure 3 of the main text (note that in the unbound simulations $\phi_2 = 90°$ and $\phi_2 = 270°$ are equivalent by symmetry). The MM simulation of leukotriene A4 hydrolase spends longer in the bound conformation in the second run ($\phi_1 \sim 270°$, $\phi_2 \sim 270°$), but unbinds by the end of the simulation (representative snapshots are provided in the Supporting Information data).

References

(1) Hartshorn, M. J.; Murray, C. W.; Cleasby, A.; Frederickson, M.; Tickle, I. J.; Jhoti, H. *J. Med. Chem.* **2005**, *48*, 403–413.

(2) Davies, D. R.; Mamat, B.; Magnusson, O. T.; Christensen, J.; Haraldsson, M. H.; Mishra, R.; Pease, B.; Hansen, E.; Singh, J.; Zembower, D.; Kim, H.; Kiselyov, A. S.; Burgin, A. B.; Gurney, M. E.; Stewart, L. J. *J. Med. Chem.* **2009**, *52*, 4694–4715.

(3) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689–1700.

(4) Ulmschneider, J. P.; Jorgensen, W. L. *J. Chem. Phys.* **2003**, *118*, 4261–4271.

(5) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2009**, *113*, 13337–13346.

(6) Jorgensen, W. L.; Thomas, L. L. *J. Chem. Theory Comput.* **2008**, *4*, 869–876.

(7) Wang, L.; Berne, B. J.; Friesner, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 1937–1942.

(8) Cole, D. J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2014**, *10*, 565–571.

(9) Cole, D. J.; Janecek, M.; Stokes, J. E.; Rossmann, M.; Faver, J. C.; McKenzie, G. J.; Venkitaraman, A. R.; Hyvönen, M.; Spring, D. R.; Huggins, D. J.; Jorgensen, W. L. *Chem. Commun.* **2017**, *53*, 9372–9375.

(10) Grant, B. J.; Rodrigues, A. P. C.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. D. *Bioinformatics* **2006**, *22*, 2695–2696.