Supplementary Information

December 16, 2018

1 Simulation protocol

1.1 System preparation

1.1.1 Trypsin

A crystal structure of trypsin bound to benzamidine was downloaded (PDB ID: 3PTB).[1] Coordinates were prepared by removing crystallographic ligands except for benzamidine. Prime (Schrödinger, Inc.) was used to model in missing side-chains. The crystallographic benzamidine was removed and placed at least 10 Å from the binding site. Hydrogen atoms were added, and protein chain termini capped with the neutral groups acetyl and methylamide. Titrateable residues were left in their dominant protonation state at pH 7.0.

Internal waters were added with Dowser,[2] then the prepared structure was solvated with in a $(70\ \text{Å})^3$ box of TIP3P waters, solvated with 0.15 M NaCl, and neutralized by removing sodium ions using Dabble.[3] Final system dimensions were $70 \times 70 \times 70\ \text{Å}^3$, including about 9700 waters, 16 sodium ions, and 26 chloride ions. The protein had a minimum clearance of 10.8 Å to the edge of the simulation box.

1.1.2 β2AR

Simulations of the β2 adrenergic receptor were based on the crystal structure of the carazolol-β2AR complex (PDB ID: 2RH1).[4] The T4 lysozyme fusion protein comprising intracellular loop 3 was deleted, along with co-crystallized ligands other than cholesterol. As with the trypsin system, Prime was used to model in missing-side chains, add hydrogen atoms, and add neutral capping groups to protein chain termini. Titrateable residues were left in their dominant protonation state at pH 7.0, except for Glu122$^{3.41}$ and Asp79$^{2.50}$, which were protonated. A palmitoylation was added to Cys341.[5]

The prepared protein structure was aligned on the transmembrane helices to the Orientation of Proteins in Membranes (OPM) database[6] and internal waters added with Dowser. Ten dihydroalprenolol ligands (protonated at the tertiary amine nitrogen as is predicted at pH 7.0) were placed above and below the protein on the Z axis so that they would be in water rather than lipid.

The software Dabble[3] was used to insert the system into a palmitoyl-oleoyl-phosphatidylcholine (POPC) bilayer, solvate with 0.15 M NaCl in explicit TIP3P waters, and remove sodium ions until the system is neutral. Final system dimensions were $75 \times 75 \times 115\ \text{Å}^3$, including about 105 lipids, 14000 waters, 24 sodium ions, and 38 chloride ions. The solute had a minimum clearance of 10.6 Å to the edge of the simulation box.

1.2 MD simulation force field parameters

We used the CHARMM36 parameter set for protein molecules, lipid molecules, and salt ions, and the CHARMM TIP3P model for water; protein parameters incorporated CMAP terms.[7–10] Parameters for benzamidine were in the CHARMM General Force Field,[11] and the ParamChem server[12,13] used to assign parameters from this force field to dihydroalprenolol. Full parameter sets are available upon request.

1.3 MD simulation protocol

Simulations were performed on GPUs using the CUDA version of PMEMD (Particle Mesh Ewald Molecular Dynamics) in Amber16.[14] Prepared systems were minimized, then equilibrated as follows: The system was heated using the Langevin thermostat from 0 to 100 K in the NVT ensemble over 12.5 ps with harmonic restraints of 10.0 kcal $\cdot$ mol $^{-1}\cdot$ Å^{-2} on the non-hydrogen atoms of lipid, protein, and ligand, with initial velocities sampled from the Boltzmann distribution. The system was then heated to 310 K over 125 ps in the NPT ensemble with semi-isotropic (for β2AR) or anisotropic (for trypsin) pressure coupling and a pressure of one bar. Further equilibration was performed at 310 K with harmonic restraints on the protein and ligand starting at 5.0 kcal$\cdot$ mol $^{-1}\cdot$Å^{-2} and reduced by 1.0 kcal$\cdot$mol$^{-1}\cdot$Å^{-2} in a stepwise fashion every 2 ns, for a total of 10 ns of additional restrained equilibration.

Production simulations were conducted in the NPT ensemble at 310 K and 1 bar, using a Langevin thermostat with a collision frequency of 1.0 ps^{-1} and Monte Carlo barostat with a pressure relaxation time of 1.0 ps and 0.4 ns between volume change attempts. These simulations were run for a total of t_{sim} ns: for trypsin, $t_{\text{sim}} = 10$ ns, and for β2AR, $t_{\text{sim}} = 40$ ns.

Simulations used periodic boundary conditions and a time step of 4.0 fs, with hydrogen mass repartitioning.[15] Bond lengths to hydrogen atoms were constrained using SHAKE. Non-bonded interactions were cut off at 9.0 Å, and long-range electrostatic interations were computed using the particle mesh Ewald (PME) method with an Ewald coefficient β of approimately 0.31 Å and B-spline interpolation of order 4. The FFT grid size was chosen such that the width of a grid cell as approximately 1 Å. Trajectory snapshots were saved every 200 ps.

2 Adaptive sampling protocol

2.1 Trajectory preparation

Before featurizing, trajectories were reimaged into a common periodic box with CPPTRAJ.[16] To reduce the size of these trajectories and to simplify later system building, lipids, waters, and ions were removed from these reimaged trajectories. In order to avoid biasing the model due to the restraints on the ligand during equilibration, we omitted the 10 ns equilibration and reimaged only the t_{sim} ns of production trajectory.

2.2 Featurization and dimensionality reduction

We featurize with the *multi-ligand contact featurizer*, implemented as a custom featurizer for MSMBuilder.[17] Each entry in the feature vector is the log of the minimum distance between heavy atom i on the ligand and residue j on the receptor, for a total of ij features per ligand. Treating ligand atoms separately captures ligand orientation, and using all protein residues yields ligand-protein interactions and weak capture of protein conformation. The log of the distance is used as ligands far from the protein (floating in solvent or stuck in the membrane) should be regarded as less distinct.

We did not set a distance cutoff or threshold as others have done[18] as dimensionality reduction will remove any unimportant features and we did not want to arbitrarily remove data that could be useful to the model.

Since the number of features is quite high, the dimensionality of the dataset is reduced by projecting the feature vectors into a space accounting for the slowest evolving coordinates in time using time independent component analysis (tICA).[19] The number of tICs retained is a user-configurable parameter, but for the trials described in this paper it was set to 15. As the timescales of transition we are interested in is small, we set the tICA lag parameter $\tau = 1$ ns for the β_2AR system and $\tau = 0.2$ ns (equivalent to no lag) for the trypsin-benzamidine system. These choices are fairly arbitrary. Another implementation[20] retained only 3 tICA components using an unspecified lag time, while other work did not reduce dataset dimensionality at all.[18]

2.3 Clustering

Initial geometric clustering is performed using the Mini-batch KMeans algorithm,[21] selected for its parallelism and speed on large datasets. We fit to a large number of clusters as many will be combined with each other in the next step—1000 microstate clusters for the β2AR system, and 100 for the trypsin system. The larger value for the β2AR system was selected due to the system featuring a membrane in which the ligands may partition.

In order to cluster kinetically, a microstate Markov State Model is created and Robust Perron Cluster Analysis (PCCA+)[22] is used to lump quickly transitioning microstates into macrostates. We build both models with a lag time $\tau = 5$ ns. For the β2AR system we fit to 50 macrostates, and for the less complex trypsin system, 20. When choosing the number of macrostates, it is better to have too many than too few, as lumping a meaningful state with an unimportant one could result in failure to resample the meaningful state. Too many macrostates means that some microstates that truly represent the same macrostate could be treated separately and both resampled—a much more minor type of error.

2.4 Markov State Model construction

The final Markov State Model needs to be a single connected component whose transition matrix is non-singular in order for the equilibrium populations to be defined and macrostates to be assignable.

If every frame in one simulation is assigned to the same cluster, and that cluster is not otherwise seen in the remaining simulations, the resulting transition matrix can be singular. To avoid this problem, we initialize the counts matrix with all values of 1×10^{-6}. This does not have a major effect on the model's eigenvalues or implied timescales, and allows our sampling to explore disconnected regions of protein-ligand space.

We also do not require the model to be ergodic—that is, we omit the usual trimming step where only the maximal strongly ergodic subgraph of the data is used to build the MSM. This prevents discarding of uncommon or disconnected clusters, and enables resampling scores to be defined for all clusters.

The Markov State Model used to determine which states to resample is typically quite far from converged. As sampling is by definition lacking at the beginning of the trial, the MSM contains quite a bit of error at this point, but for our purposes this is alright as we are using it as means of evaluating our discretized state space for what to resample.

Note that it is important that the desired number of macrostates exceeds the number of samplers, as otherwise resampling becomes complicated.

2.5 Scoring functions

For the counts scoring function, the assigned cluster data was counted to determine the number of frames assigned to each cluster. The population scoring function uses built-in functionality in MSMBuilder, where the equilibrium populations are accessible as an attribute of the MSM. The hub scores scoring function is also implemented in MSMBuilder as a function that takes a MSM as a parameter.

2.6 System building

Candidate simulation frames were aligned on either protein backbone (trypsin) or transmembrane helices (β2AR) to the initial prepared system using the python interface to VMD.[23]

What constitutes the "bulk solvent" as determined by manual inspection is usually a combination of several macrostates—one where the ligand is in the water, and several where it is near the lipid or protein. To automatically identify solvent, the top three macrostates with either the highest hub scores, largest equilibrium populations, or largest count (depending on the chosen scoring function) are designated as the bulk clusters.

To ensure diversity in sampling, the top N scoring clusters (corresponding to the number of sampling simulations) are always resampled. To resample these clusters, a simulation frame where a ligand assigned to the desired cluster is present is chosen at random. Water, lipid, and ions have already been removed from the trajectories, so the frame contains only the protein (in whatever conformation the ligand is associating with), and ligand(s). In the case of multiple ligands, ligands not assigned to the cluster of interest have their coordinates zeroed to mark them as available for the greedy step.

The remaining ligands are placed greedily in each of the N simulation systems with the following algorithm: while there are ligands left to place, select a cluster to sample that has not already been selected for this system with selection probability inversely proportional to either hub score, population, or count. Select a simulation frame where a ligand is assigned to the cluster. If the ligand is too close to the edge of the simulation box, is within 2.5 Å of protein or is within 10.0 Å of another placed ligand, reject it and try another cluster. If 20 rejections have happened, select a ligand from the bulk solvent cluster, repeatedly choosing simulation frames for the ligand position if it would be placed too close. If this fails 50 times, place the solvent ligand without the closeness criteria (in practice, this rarely to never happens).

Ligands placed in solvent are forbidden from being too close to the edges of the simulation box, as this can result in crashes during MD simulation. The minimum allowed distances to the protein and other ligands are user-configurable parameters.

The system with placed ligands is then inserted into a lipid membrane (if applicable), solvated with a water box of desired size, and ions added. The system is then parameterized and is ready to be simulated. This final assembly step is performed by Dabble.

3 Analysis

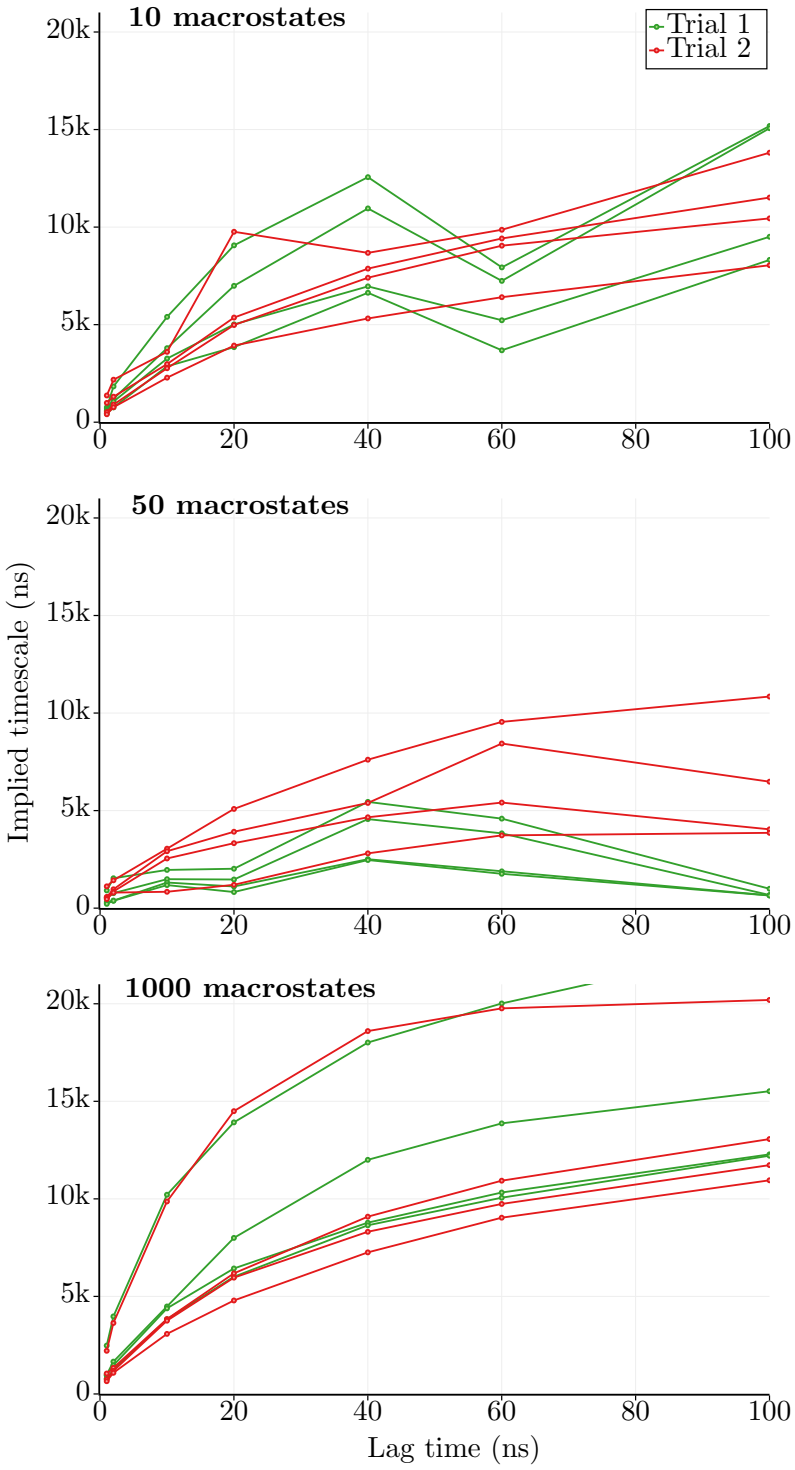
3.1 Density calculation

For the cholesterol location sampling plots, a map of ligand density was constructed as follows: the simulation box was divided into a 1 Angstrom grid, and the value of each grid square incremented for each frame of simulation where any ligand atom was in the square. The grid values were divided by the total number of frames, yielding a normalized measurement of where the ligand goes that is roughly analogous to the density maps produced by X-ray crystallography. The ligand density grid

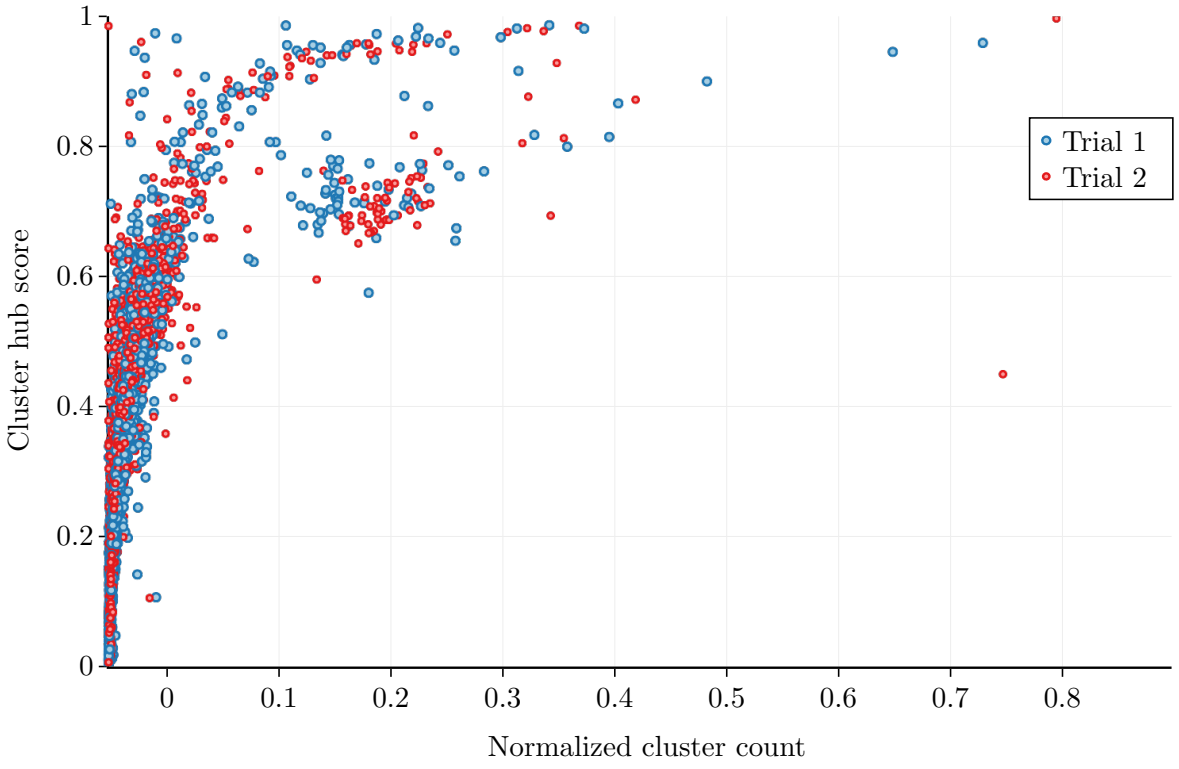
was then summed over each cholesterol position, using a box-like approximation for the locations as they are oriented nearly exactly along the Z axis of the simulation box.

For the "within 5 Å of protein" plots, the ligand density grid was summed over grid locations that are within 5 Å of protein.

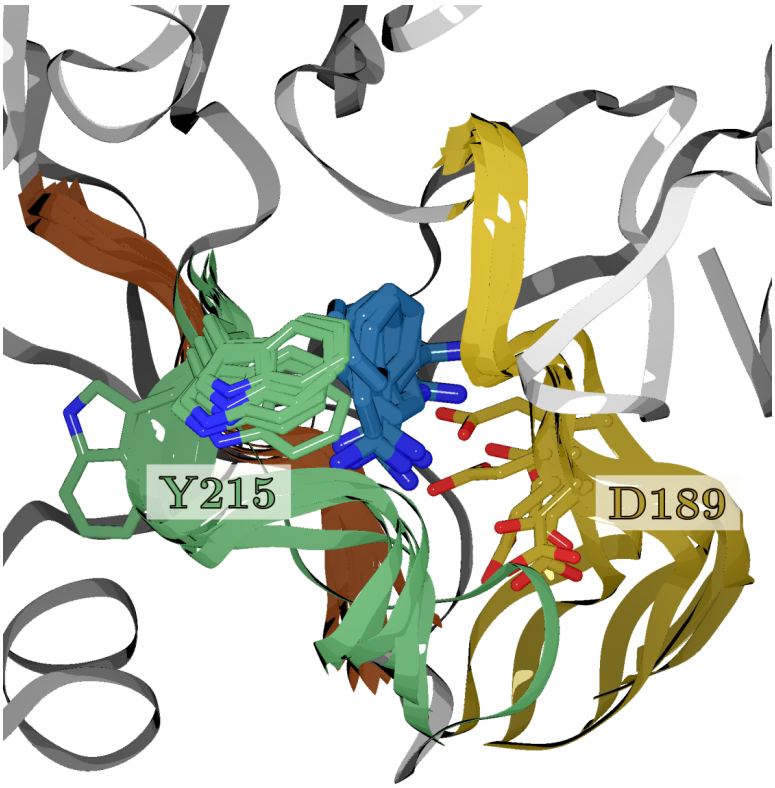
4 Supplementary Figures



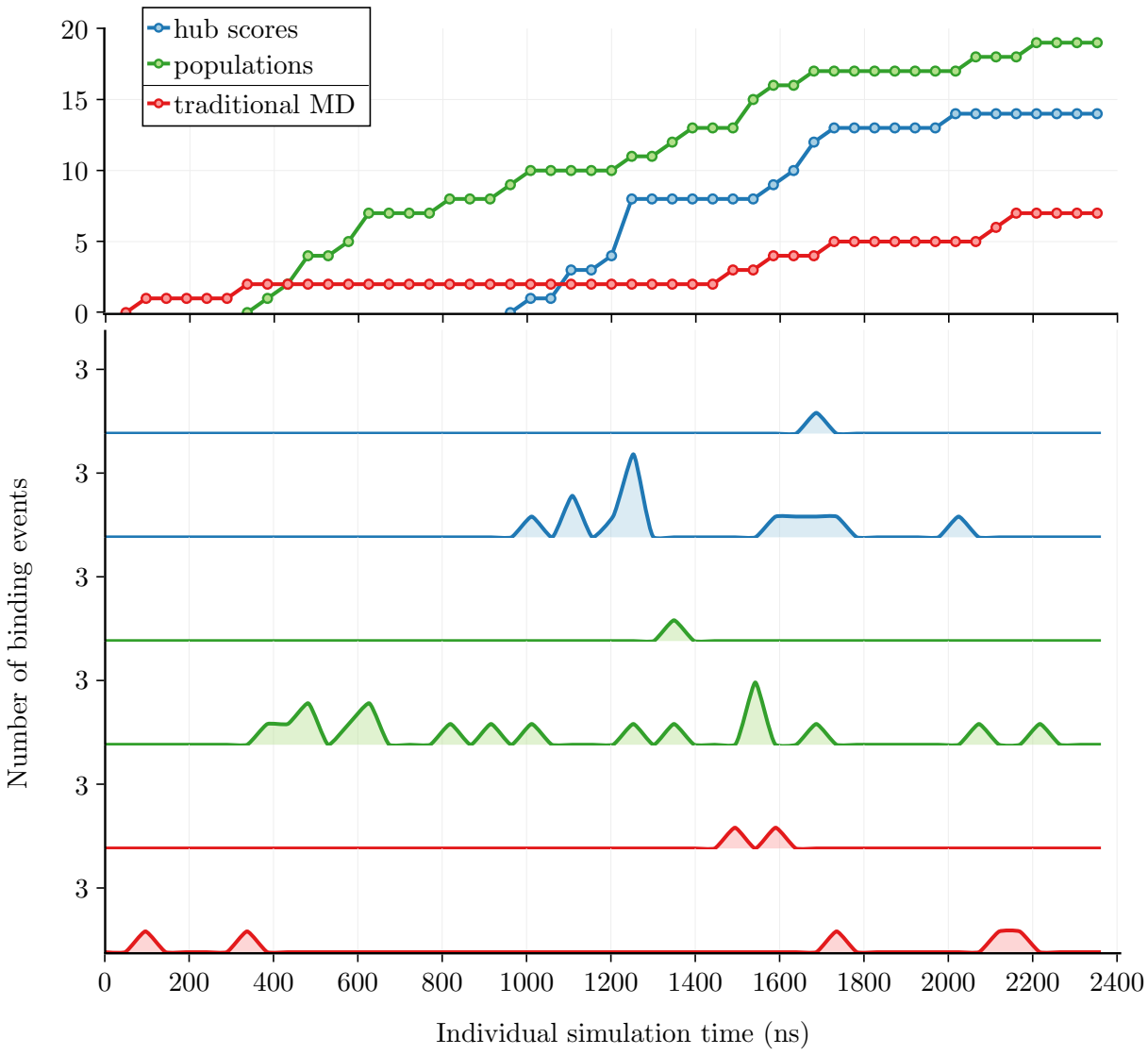
Supplementary Figure 1: Slowest four implied timescales vs. MSM lag time τ for MSMs built with 10, 50, and 1000 macrostates. To allow testing of longer lag times and avoid possible sampling biases, both traditional MD runs of the β_2AR system were used to train the models.



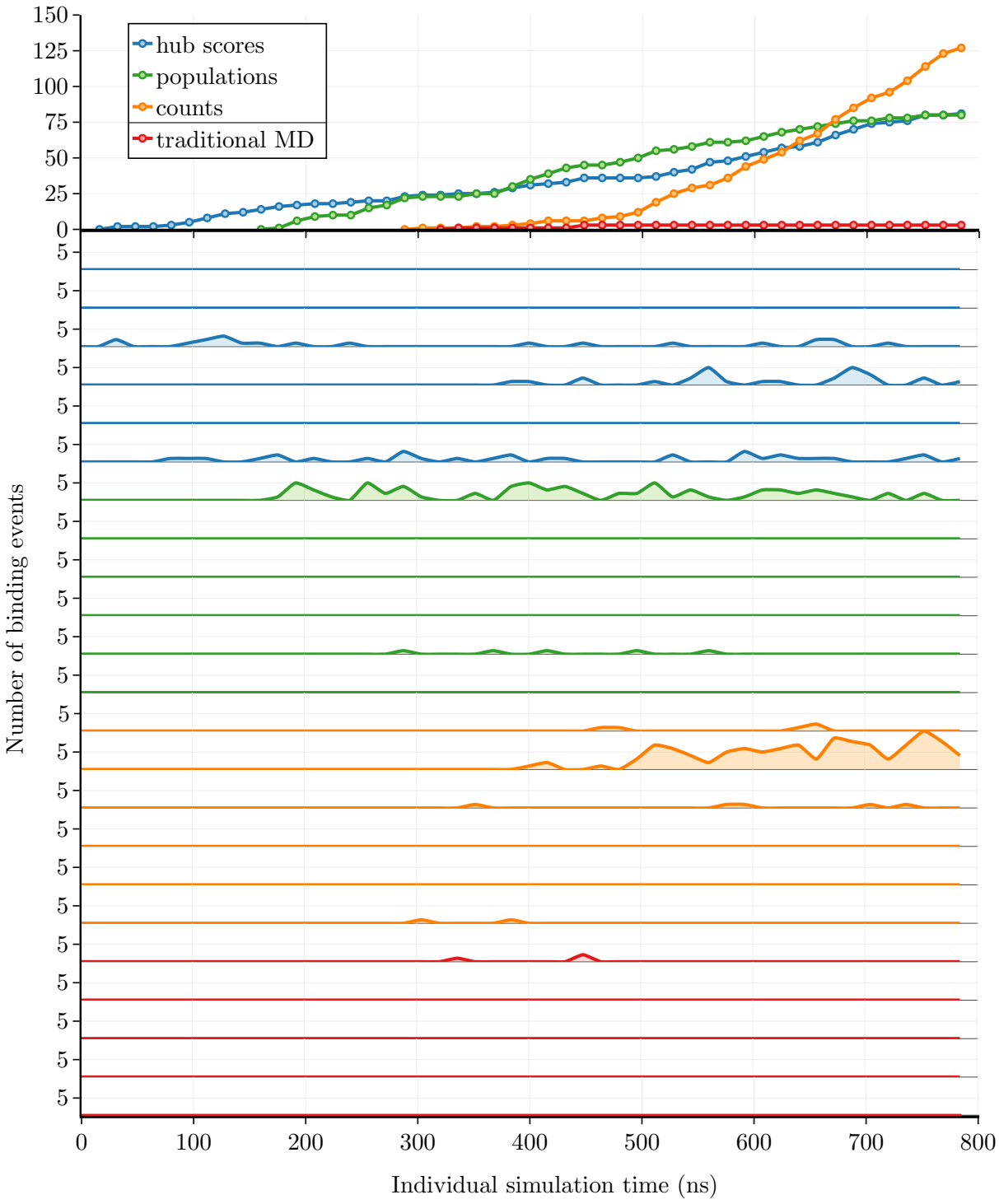
Supplementary Figure 2: Relationship between cluster count and its hub score for both runs of the β_2AR system with the hub scores criterion, for all clusters and rounds. Counts have been normalized within each round to allow for comparison.



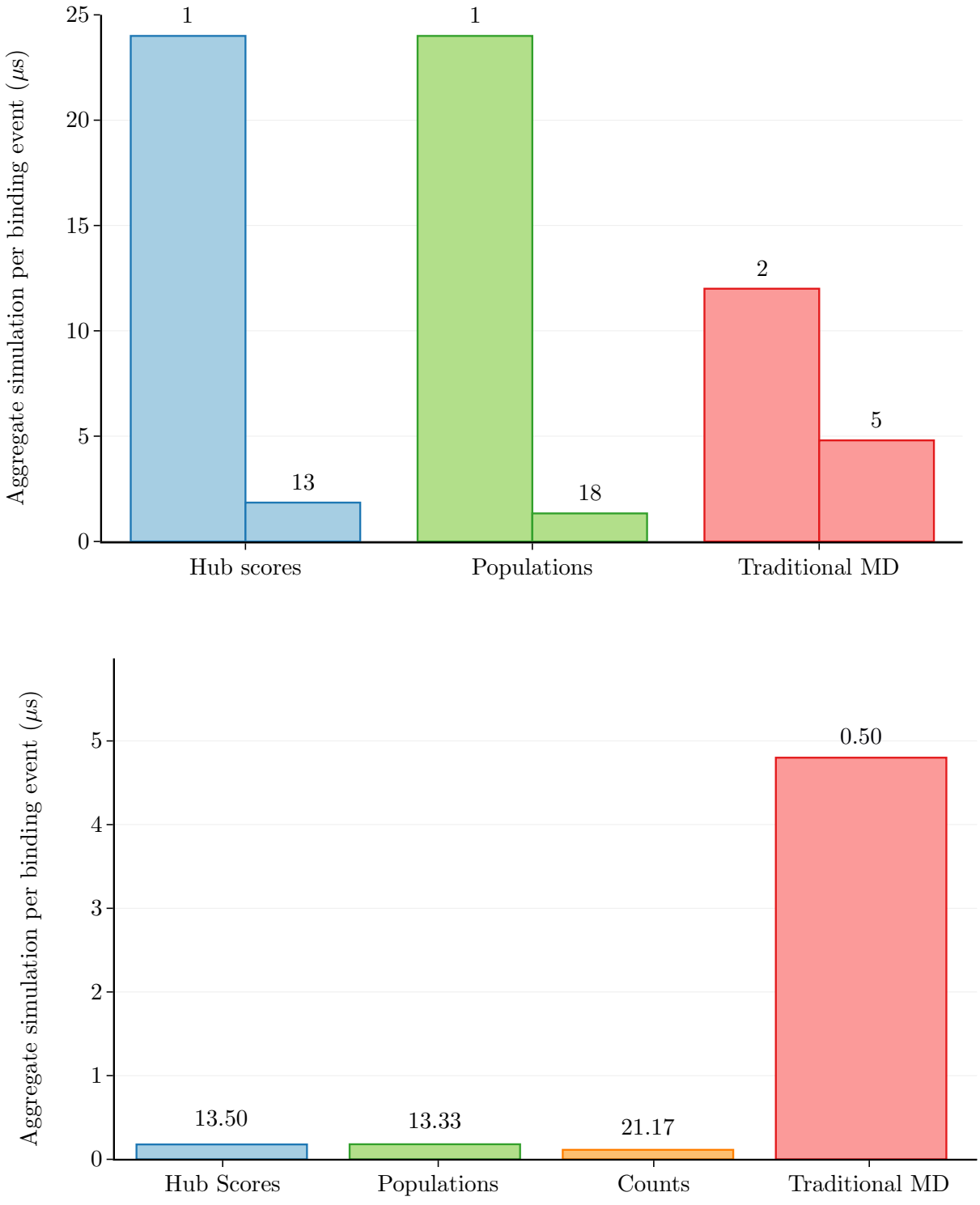
Supplementary Figure 3: Overlaid are seven independent simulation frames assigned to the bound cluster obtained after 80 rounds of adaptive sampling with the hub scores criterion on the trypsin-benzamidine system. Benzamidine is shown in blue sticks. The three loops forming the binding pocket are shown for all frames in orange, green, and yellow, with key residues Asp189 and Trp215 shown as sticks.



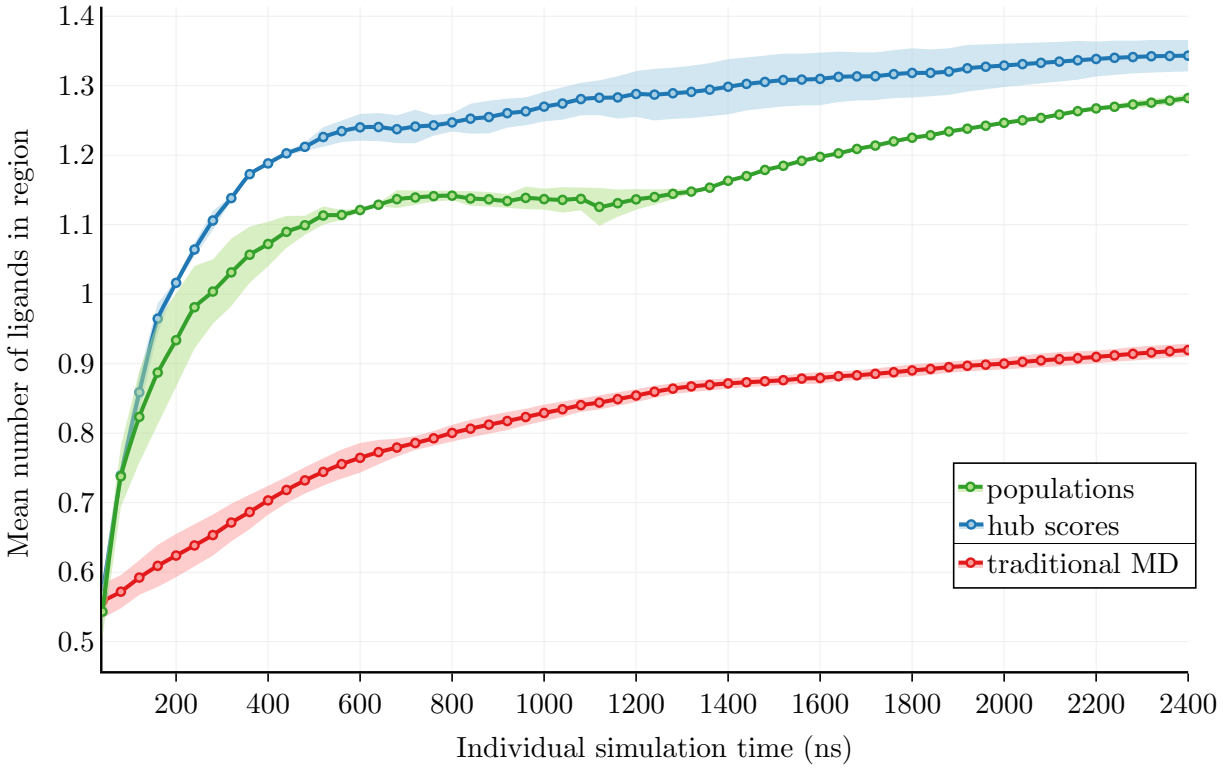
Supplementary Figure 4: At top, cumulative number of binding events observed for dihydroalprenolol to β_2AR over all trials. Below, each trace corresponds to a single trial with 20 independent simulations and an individual simulation length of 40 ns. Binding events are defined as the ligand going from RMSD over 3 Å to less than 2 Å.



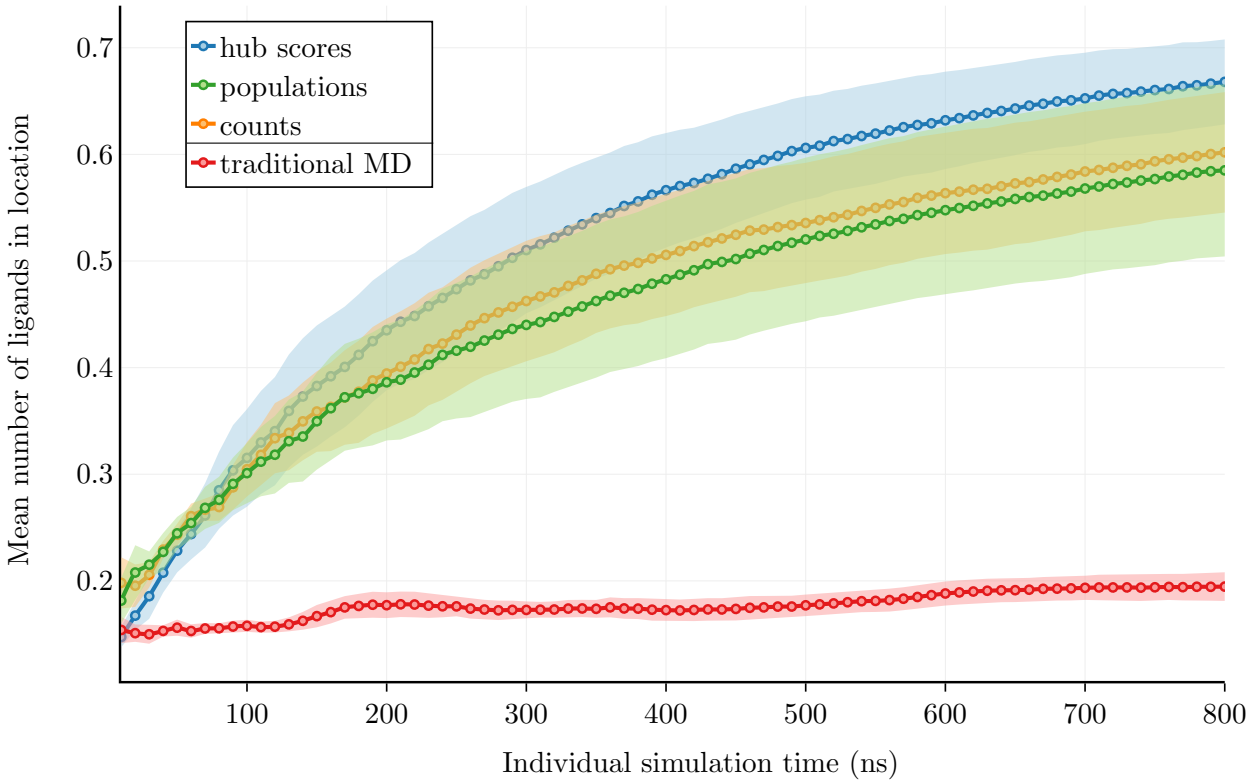
Supplementary Figure 5: At top, cumulative number of binding events observed for benzamidine binding to trypsin. Below, each trace corresponds to a single trial with 10 independent simulations and an individual simulation length of 10 ns. Binding events are defined as the ligand going from RMSD greater than 3 Å to less than 2 Å.



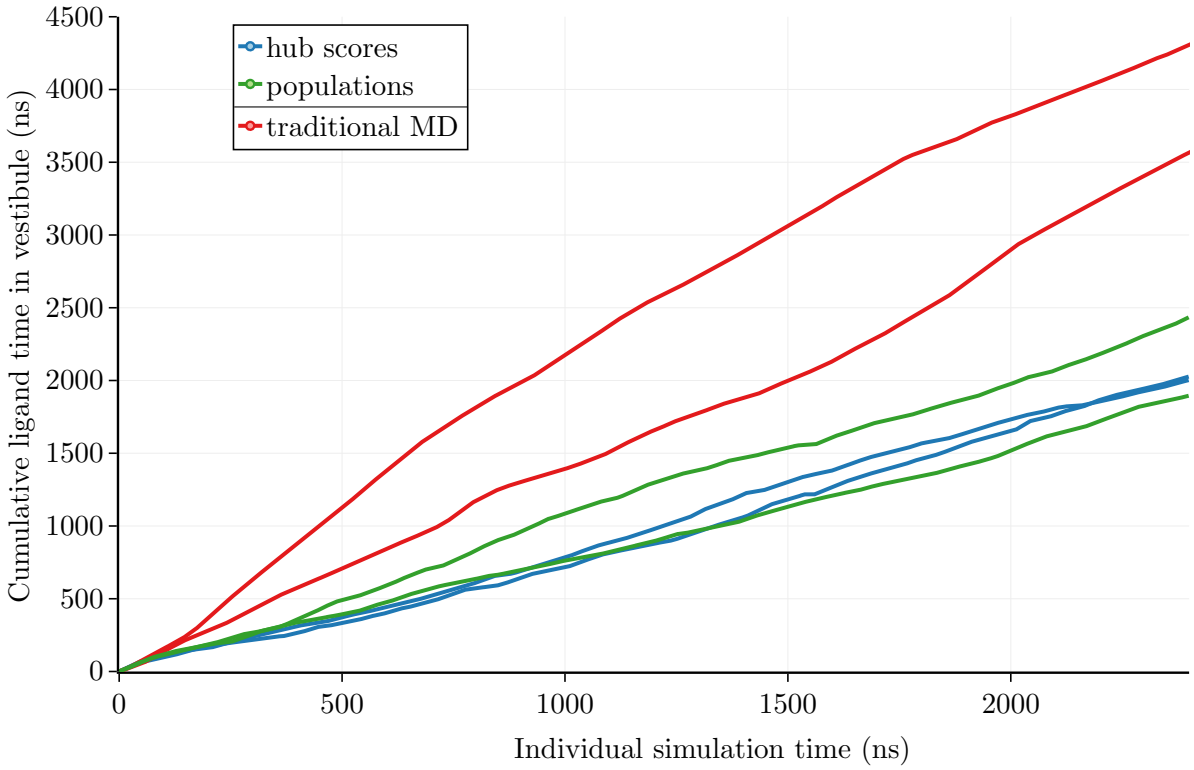
Supplementary Figure 6: Total simulation time required per binding event for all conditions for the β_2AR (top) and trypsin (bottom) systems. The total number of binding events in the run was divided by the total aggregate simulation time run. The number of binding events in the run is shown above each bar. Runs for β_2AR as shown individually, while mean values are shown for trypsin due to lack of binding events in some trials.



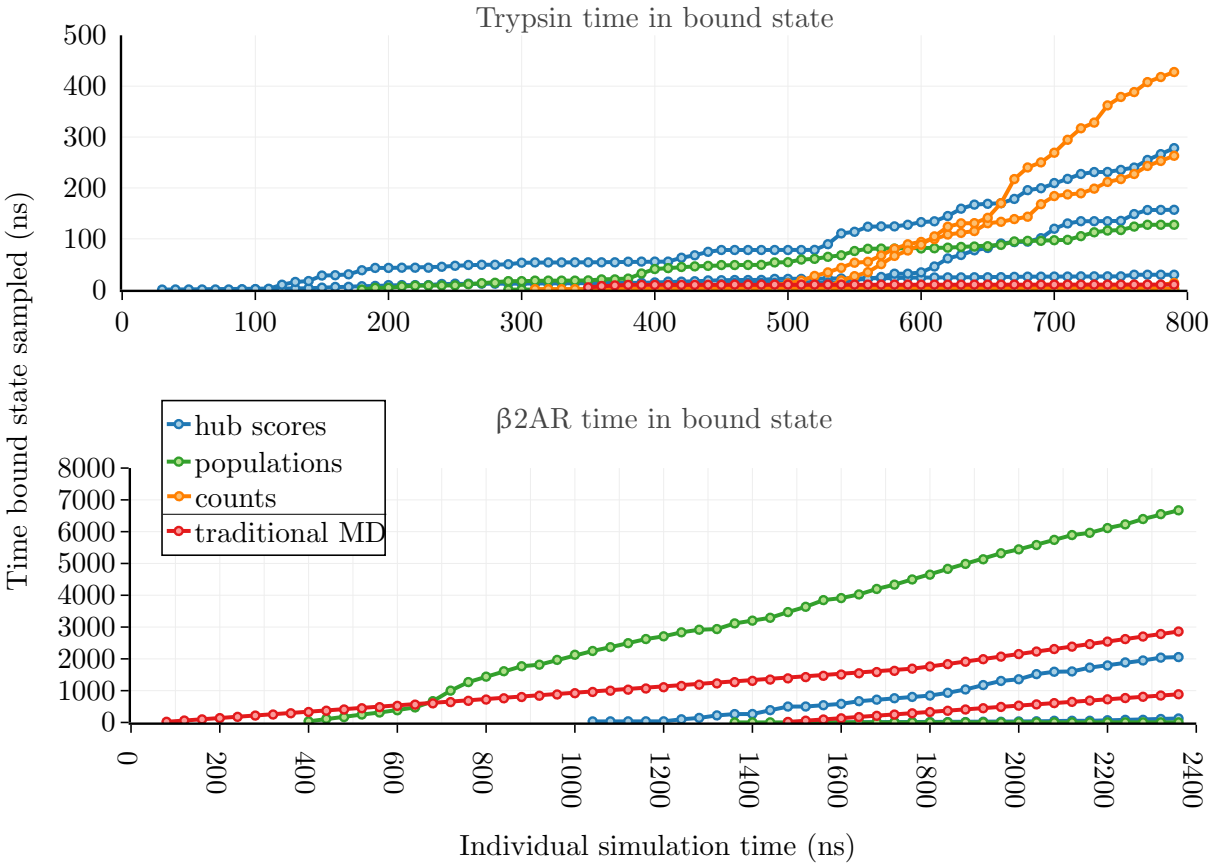
Supplementary Figure 7: Sampling over time of regions within 5 Å of β_2AR by dihydroalprenolol, in units of the mean number of ligands in the region. As there are ten dihydroalprenolol ligands in the simulation box, this value can be greater than one. Shaded areas indicate the standard error of the mean.



Supplementary Figure 8: Sampling over time of regions within 5 Å of trypsin by benzamidine, in units of the mean number of ligands in the region. This is equivalent to the probability of finding a benzamidine ligand in the region. Shaded areas indicate the standard error of the mean.



Supplementary Figure 9: Sampling of the intracellular vestibule region on β_2AR, which is the intermediate state in the binding pathway of dihydroalprenolol to this receptor, over time. For the ligand to be in the vestibule, following alignment to the initial structure, at least half of the ligand heavy atoms (9 of 18) need to be in the box defined by $-10 < x < 10$ and $-10 < y < 10$ and $12 < z < 27$.



Supplementary Figure 10: Total time spent in the crystallographic pose (as defined by RMSD to crystal structure less than 2 Å) over time for trypsin-benzamidine (top) and the β_2AR system (bottom). For clarity, only trials with binding events are plotted.

References

[1] Marquart, M.; Walter, J.; Deisenhofer, J.; Bode, W.; Huber, R. *Acta Crystallographica Section B* **1983**, *39*, 480–490.

[2] Hermans, J.; Xia, X.; Zhang, L.; Cavanaugh, D. Dowser. 2003.

[3] Betz, R. M. Dabble. 2017.

[4] Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-j.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. *Science* **2007**, *318*, 1258–1266.

[5] O'Dowd, B. F.; Hnatowich, M.; Caron, M. G.; Lefkowitz, R. J.; Bouvier, M. *Journal of Biological Chemistry* **1989**, *264*, 7564–7569.

[6] Lomize, M. A.; Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I. *Bioinformatics* **2006**, *22*, 623–625.

[7] Huang, J.; MacKerell, A. D. *Journal of Computational Chemistry* **2013**, *34*, 2135–2145.

[8] Klauda, J. B.; Venable, R. M.; Freites, J. A.; Connor, J. W. O.; Tobias, D. J.; Mondragonramirez, C.; Vorobyov, I.; Mackerell, A. D.; Pastor, R. W. *Journal of Physical Chemistry B* **2011**, *114*, 7830–7843.

[9] Best, R. B.; Mittal, J.; Feig, M.; MacKerell, A. D. *Biophysical Journal* **2012**, *103*, 1045–1051.

[10] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. *Journal of Chemical Theory and Computation* **2012**, *8*, 3257–3273.

[11] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D. J. *Journal of Computational Chemistry* **2010**, *31*, 671–690.

[12] Vanommeslaeghe, K.; MacKerell, A. D. *Journal of Chemical Information and Modeling* **2012**, *52*, 3144–3154.

[13] Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. *Journal of Chemical Information and Modeling* **2012**, *52*, 3155–3168.

[14] Case, D. et al. AMBER 2016. 2016.

[15] Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. *Journal of Chemical Theory and Computation* **2015**, *11*, 1864–1874.

[16] Roe, D. R.; Cheatham III, T. E. *Journal of Chemical Theory and Computation* **2013**, *9*, 3084–3095.

[17] Harrigan, M. P.; Sultan, M. M.; Hernández, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; McGibbon, R. T.; Pande, V. S. *Biophysical Journal* **2017**, *112*, 10–15.

[18] Doerr, S.; De Fabritiis, G. *Journal of Chemical Theory and Computation* **2014**, *10*, 2064–2069.

[19] Naritomi, Y.; Fuchigami, S. *Journal of Chemical Physics* **2011**, *134*.

[20] Doerr, S.; Harvey, M. J.; Noe, F.; De Fabritiis, G. *Journal of Chemical Theory and Computation* **2016**, *12*, 1845–1852.

[21] Sculley, D. *Proceedings of the 19th international conference on World wide web - WWW '10* **2010**, 1177.

[22] Kube, S.; Weber, M. *Journal of Chemical Physics* **2007**, *126*.

[23] Humphrey, W.; Dalke, A.; Schulten, K. *Journal of Molecular Graphics* **1996**, *14*, 33–38.