Fragment Binding Pose Predictions Using Unbiased Simulations and Markov-State-Models

Stephanie Maria Linker[1,2], Aniket Magarkar[1], Jürgen Köfinger[2], Gerhard Hummer[2,3], Daniel Seeliger[1]

[1]Department of Medicinal Chemistry, Boehringer Ingelheim Pharma, Birkendorfer Str. 65, 88397 Biberach an der Riß, Germany
[2]Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue Straße 3, 60438 Frankfurt am Main, Germany.
[3]Institute for Biophysics, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany.

Author contributions

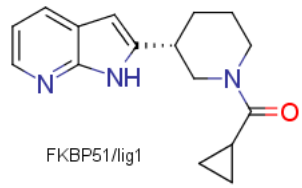
D.S. conceived the project. S.M.L., D.S. and G.H. designed experiments. S.M.L. executed the simulations and analyzed the data. S.M.L and A.M. developed MSM_bind. D.S., G.H. and J.K. supervised the work. S.M.L and D.S. wrote the manuscript.

Keywords:

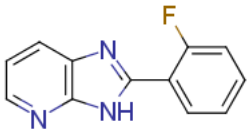
Fragment-based drug discovery, Markov-state models, binding pose prediction, molecular dynamics

Supplementary Information

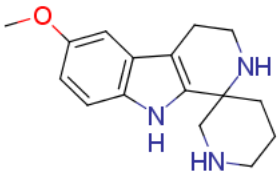
Fragment Structures



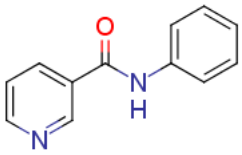
FKBP51/lig1



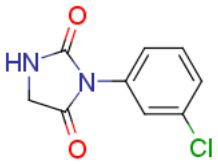
FKBP51/lig2



NE/lig1



NE/lig2



NE/lig3

In-Cluster Variability

Ligands sample different bound conformations in the binding pocket that can be exploited for lead optimization. Our protocol identified several low rmsd poses that might correspond to different binding modes. To explore potential binding pose alternatives, we analyzed the conformational variability within and between clusters. The rmsd distribution of top ranked cluster members is depicted in the violin plot in Figure S1 A and D. It contains information about the probability density as well as the median and the extrema. Most clusters have a uniform probability density, however some cluster contained outliers. Visual inspections of most representative cluster members identified two different binding modes for FKBP51/lig1 and three binding modes for NE/lig3 (Figure S1 B and E, respectively). The assignment of the top ranked clusters to these binding modes is encoded in the asterixis's color in Figure S1 A and D. The two different binding modes for FKBP51/lig1 only differ in the orientation of the cyclopropane moiety. Similarly, NE/lig3 has a binding pattern shared

by all clusters and one more diverse region. Alternative binding poses can lead to new fragment growing/linking ideas and contain valuable information about essential fragment substructures.

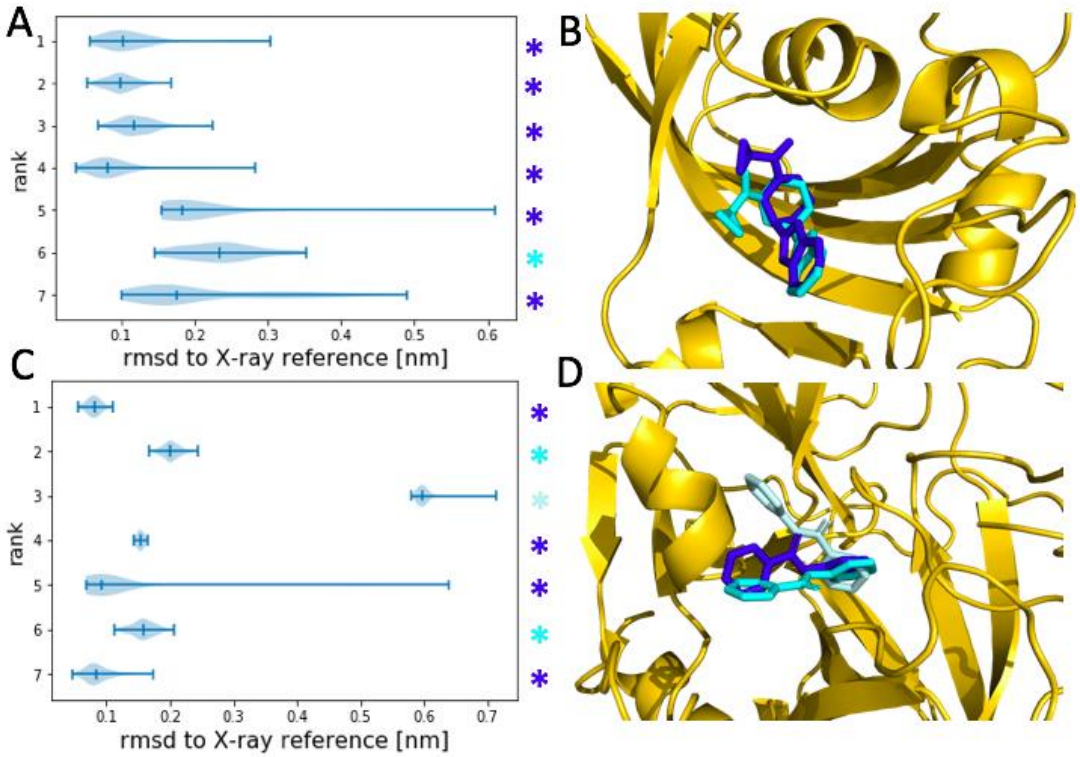


Figure S 1: Identification of alternative binding poses. A and C Confirmation variability of the first ranked clusters for FKBP51/lig1 and NE/lig3, respectively. The violin plot depicts the cluster member's rmsd to the X-ray reference. Whiskers indicate the extrema and the median is indicated with a vertical line. The colors of the asterixis correspond to the binding pose membership as depicted in B and D. B and D Alternative binding poses within the binding site.

Supplement methods

Model Preparation

The Xray crystal structures of the FK1 domain of FKBP51 (Xray resolution 1.86 Å) and neutrophil elastase (Xray resolution 1.12 Å) were obtained from the in-house database of Boehringer Ingelheim. The corresponding structure files are added to the supplement. The protonation state of all residues was determined according to the physiological pH (7.3) and their local environment. Subsequently, we removed the crystallographic waters and the bound fragment from the structure. Hydrogens of the protein were modeled as virtual sites[1]. Their natural mass was added to their binding partner, in order to preserve the total mass. The mass of hydrogens of the fragment was set to 4 u, therefore the hydrogen vibration was sufficiently slow for the chosen timestep of 4 fs. Ligand parameterization was performed as described below. Afterwards, the parametrized fragment was added to the simulation box with a distance between 1.5 Å and 3.5 Å to the protein. In order to separate the protein from its periodic boundary image, a 1.2 Å layer of water molecules was added. The system was neutralized with sodium and chloride ions to an ion concentration of 0.15 M. This results in a periodic, neutral box.

Fragment Parametrization

Ligands were parametrized according to the generalized amber force field (GAFF) procedure and converted to gromacs topologies using acpype[2]. Hydrogen masses of the ligand were set to 4 u to allow for a timestep of 4 fs.

Molecular Dynamics Simulations

Molecular dynamics simulations were performed using the GROMACS version 4.6.7[3]. The protein was described using the AMBER99SB*-ILDN forcefield[4]. The water molecules were treated using the SPCE water model[5]. After steepest decent energy minimization with 500 steps, the system was equilibrated in the NPT ensemble for 50 ps with position restrains on

the protein. The position restraints consisted of a constant force of 1000 kJ $mol^{-1}nm^{-2}$, which was applied to the backbone atoms to constrain their positions. Temperature was scaled to 300 K via stochastic velocity rescaling and pressure was set to 1 bar with the isotropic Parrinello- Rahman coupling scheme and a compressibility of 4.6 10^{-5} bar^{-1} [6]. The long range electrostatic interactions were treated within the Particle-Mesh Ewald approach with a nonbonded cutoff of 1 nm[7]. The SETTLE algorithm was used to constrain bonds and angles for of water molecules[8]. P-LINCS was used for all other bonds[9]. Hydrogens of the protein were modeled as virtual interaction sites[1].

Markov State Models

Snapshots of the simulations were generated every nanosecond. The trajectories were corrected for out of the box jumping and the protein was rmsd fitted to its starting structure. With the assumption that no major rearrangements in the short timescale of the simulation protein occur, this ensures that the same x,y,z coordinates of the ligand corresponds to the same relative position to the protein in every frame. In order to reduce the amount of data, water molecules and ions were not used for the Markov state model analysis.

To generate the Markov models PyEMMA, an open source Python library for MSMs, was used[10]. PyEMMA is able to read GROMACS molecular dynamics data formats. It provides functions for dimension reduction such as principal component analysis (PCA) and time-lagged independent component analysis (TICA). Additionally, it is capable of k-means and uniform time clustering. For other clustering methods the python scikit library was used[11]. PyEMMA contains estimators for MSMs, hidden Markov models, and some other models. It is capable of model validation and error calculation methods.

Docking

For both FKBP51 and NE an in-house Xray structure that has been crystallized with a ligand not subject to this publication has been used. Structures were loaded into MOE, all water molecules and organic molecules removed and the structure was energy minimized using standard MOE protocols[12]. Subsequently, pdbqt files for receptors and ligands were

generated using AutodockTools[13] through the PyMOL/Autodock plugin[14]. For the global docking runs, binding sites were defined such that the entire protein was contained. For the local dockings, the box center was defined at the center of the Xray ligand and a cubic box of 20Å was defined. Docking runs were then executed using vina[15].

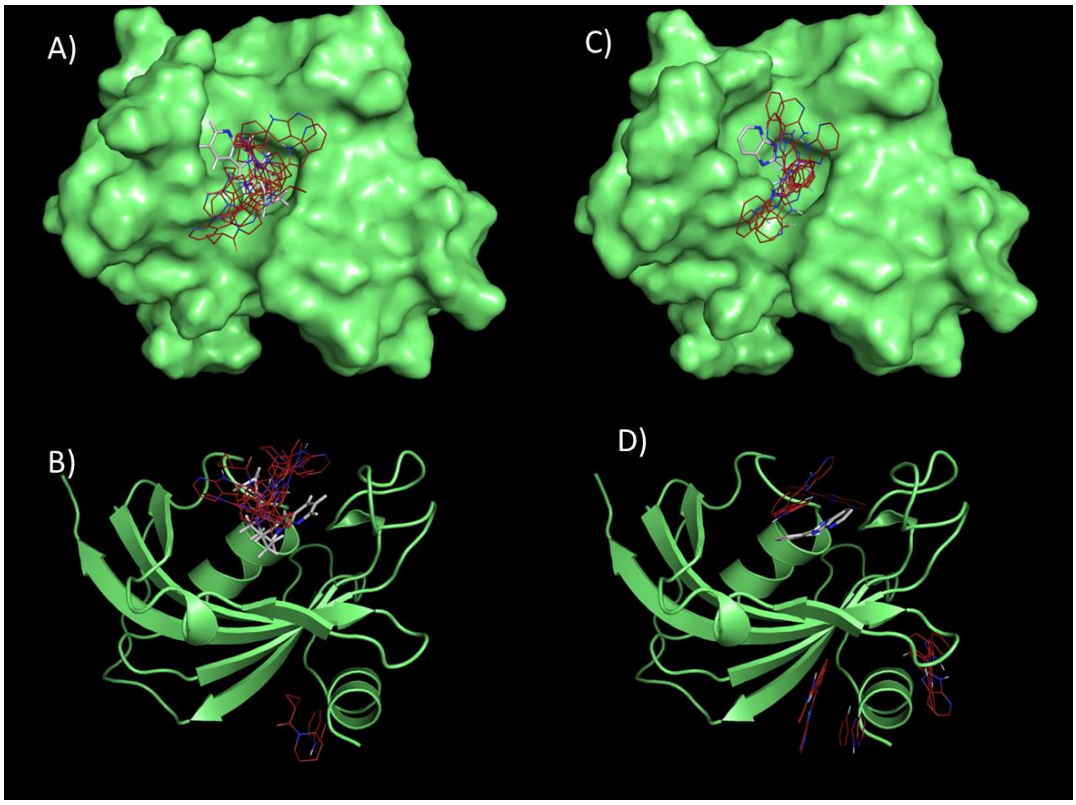


Figure S 2: Docking solutions for FKBP51. A) Local docking of ligand 1. B) Global docking of ligand 1. C) local docking of ligand 2. D) Global docking of ligand 2.

Table S1: Docking solutions FKBP51

Pose #	FKBP51/lig1 (local) RMSD[nm]	FKBP51/lig1 (global) RMSD[nm]	FKBP51/lig2 (local) RMSD[nm]	FKBP51/lig2 (global) RMSD[nm]
1	0.55	0.55	0.41	**0.41**
2	0.52	0.67	0.57	1.75
3	0.68	0.47	0.66	1.74
4	0.64	**0.35**	0.68	0.56
5	0.65	0.60	0.68	1.94
6	**0.35**	2.04	0.54	1.83
7	0.60	0.58	**0.39**	1.62
8	0.58	0.65	0.62	1.49
9	0.61	0.38	0.64	1.68

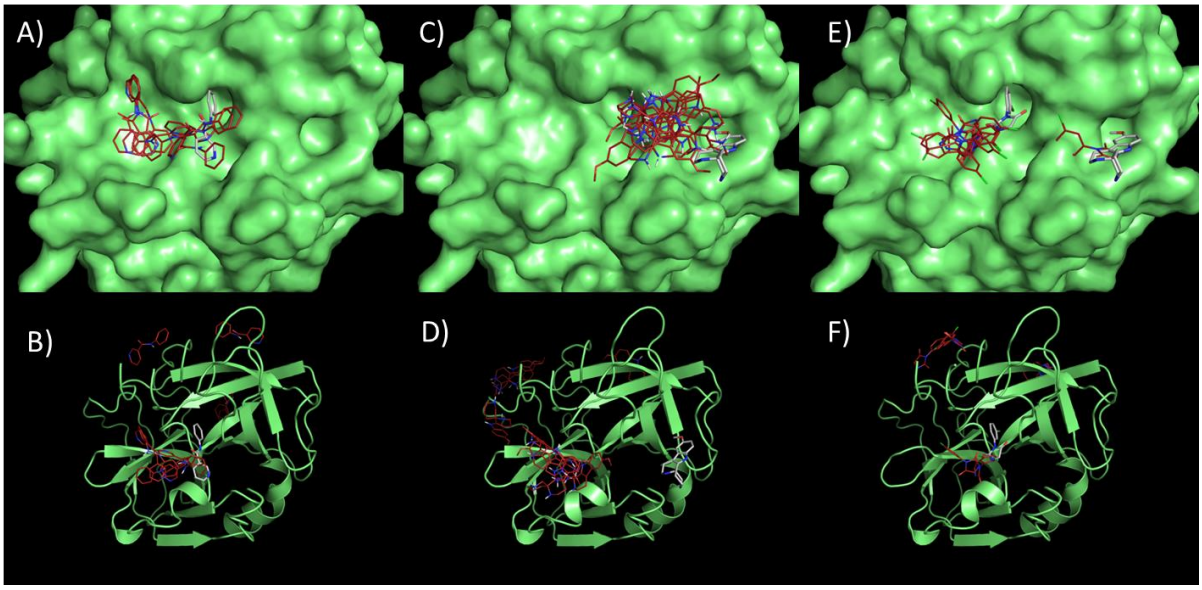


Figure S 3: Docking solutions for NE. A) Local docking of ligand 1. B) Global docking of ligand1. C) Local docking of ligand2. D) Global docking of ligand 2. E) Local docking of ligand 3. F) Global docking of ligand 3.

Table S2: Docking solutions NE

Pose #	NE/lig1 (local) RMSD[nm]	NE/lig1 (global) RMSD[nm]	NE/lig2 (local) RMSD[nm]	NE/lig2 (global) RMSD[nm]	NE/lig3 (local) RMSD[nm]	NE/lig3 (global) RMSD[nm]
1	0.78	2.06	0.83	**1.56**	0.68	2.29
2	0.90	2.51	0.79	2.07	0.79	2.32
3	0.81	0.90	0.81	3.92	0.69	1.85
4	0.86	0.78	0.74	3.34	0.78	2.32
5	0.57	**0.60**	1.09	2.07	1.10	1.96
6	**0.46**	2.22	0.73	2.00	0.78	2.34
7	0.49	0.77	**0.65**	1.93	**0.56**	0.67
8	0.65	2.49	0.91	4.01	0.64	1.86
9	0.74	0.56	0.81	2.93	0.65	**0.56**

Supplement Figures

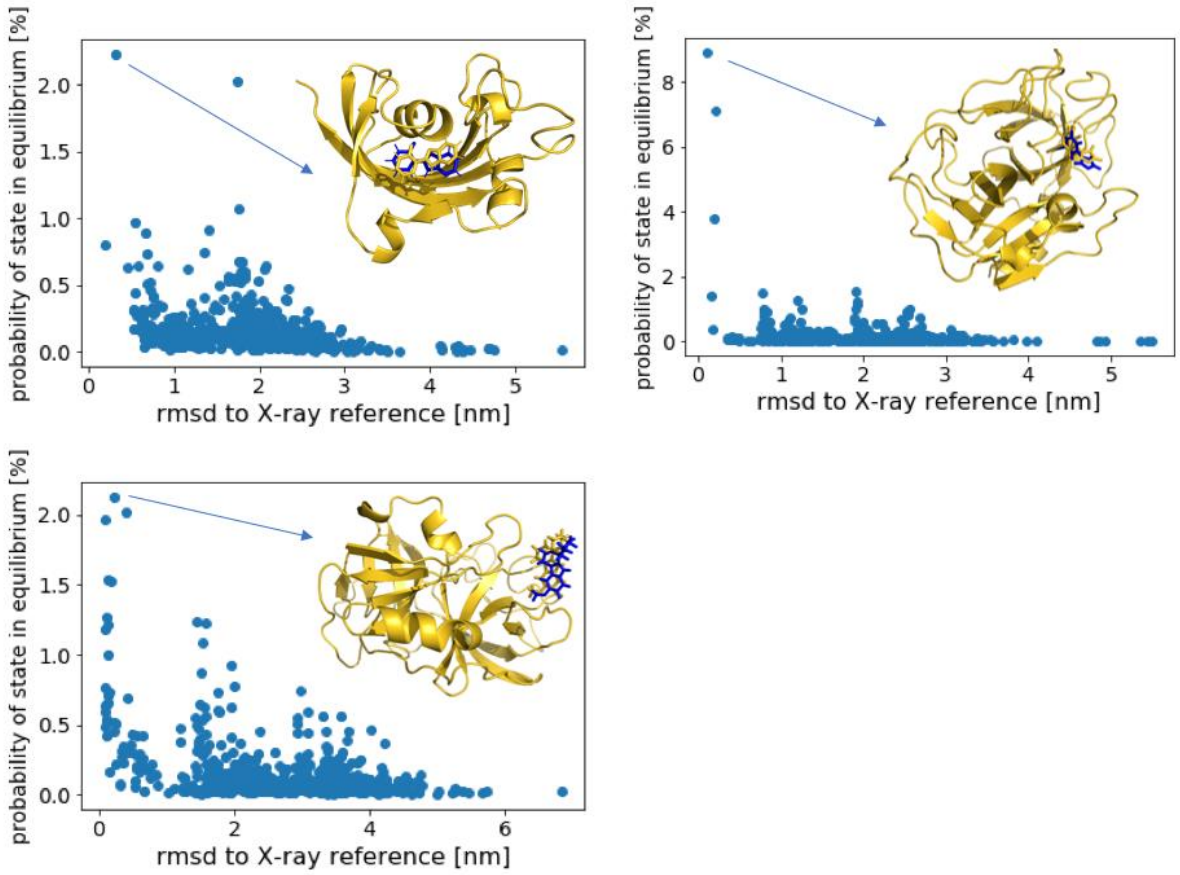


Figure S 4: Predicted equilibrium probabilities of all Markov-state models clusters plotted against their rmsd to the Xray reference. The highest ranked state is visualized in the inlay. The Xray reference is colored in yellow and the cluster in blue. A: FKBP51 with ligand 2. B: NE with ligand 3 C: NE with ligand 1

Analysis of Binding Events

In table S1 the total number of correct binding events and the total simulation time spent in the correct pose are listed. For system FKBP51/lig2, there were only two correct binding events observed with a total binding time of 66 ns. Apparently, this amount is not sufficient to predict this pose as the correct binding pose. The same applies to system NE/lig1, where the correct pose was ranked highest but not with high confidence only three binding events where observed in a total of 50 µs of simulation data.

Table S 3: Binding events and accumulated binding time

System	# of binding events	Accumulated binding time [ns]
FKBP51/lig1	7	656
FKBP51/lig2	2	66
NE/lig1	3	329
NE/lig2	20	1915
NE/lig3	18	1264

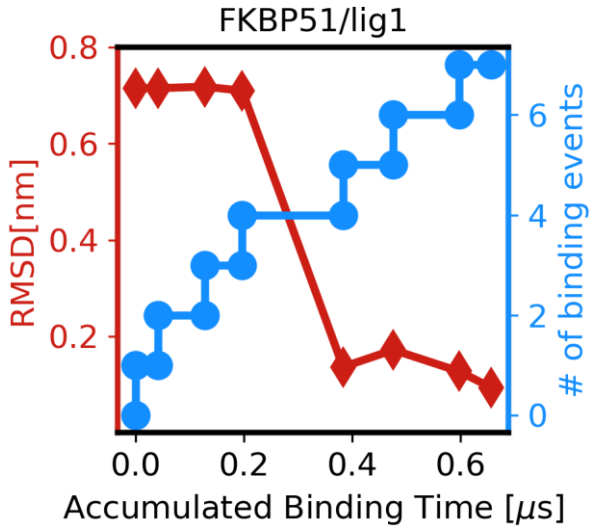


Figure S 5: rmsd of the top-ranked pose as a function of the number of binding events and accumulated binding time within the simulation data. Only few true binding events are needed to rank the correct binding pose first.

Obviously the number of true binding events, where the fragment binds into its correct position and stays there for the entire simulation, substantially influences the overall model quality. To analyze the influence of the sampling in more detail, we calculated the

dependence of the prediction on the quantity of binding events. To this end, all simulations of the FKBP51/lig1 system containing at least one frame with a rmsd < 0.3 nm to the bound state were removed from the dataset and the analysis protocol was run on the remaining simulations. Subsequently, the binding simulations were added one after another and the performance increase was observed. Figure S4 shows the rmsd of the top-ranked cluster as a function of the accumulated binding time in the simulations used to build the Markov-state model. As expected, the prediction of both protein/fragment combinations without any binding simulation does not result in a correct pose prediction. When adding simulations containing true binding event, the top-ranked cluster does not change until the simulation with the fourth binding event is added, resulting in the correct pose prediction. This data shows that for a binding pose prediction with reasonable confidence only few binding events are needed. The entire ensemble of 50 μs contains only 7 true binding events and only for about 0.65 μs of the simulation time (1.3% of the simulation data) the fragment is found in its correct binding pose. Yet the correct pose is predicted with high confidence.

Decoy detection

We have simulated 2 fragments as shown in Figure S6, which were identified as non-binders in the biochemical assays for target NE. We would like to stress that non-binder 2 has strong similarities to NE ligand 1. After repeating the same protocol with simulations and MSM bind analysis, the fragment showed some binding events, however the accumulated binding time was more than one magnitude shorter than for binders (data shown in table S4). Therefore, the fragments explore the binding pocket, but the protein-fragment interactions are not favourable enough to keep the fragment bound.

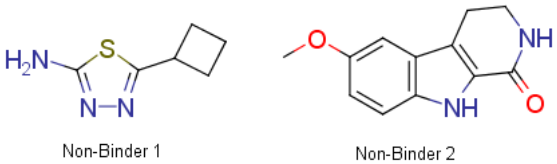


Figure S 6: Structures of non-binders

Table S 4: Binding events and accumulated binding time for non-binders

System	# of binding events in 30 µs simulations	Accumulated binding time (ns)
Non-binder 1	6	53
Non-binder 2	9	79

References:

1. Bjelkmar, P.; Larsson, P.; Cuendet, M. A.; Hess, B.; Lindahl, E., Implementation of the CHARMM Force Field in GROMACS: Analysis of Protein Stability Effects from Correction Maps, Virtual Interaction Sites, and Water Models. *Journal of Chemical Theory and Computation* **2010,** *6* (2), 459-466.
2. (a) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of computational chemistry* **2004,** *25* (9), 1157-1174; (b) da Silva, A. W. S.; Vranken, W. F., ACPYPE-Antechamber python parser interface. *BMC research notes* **2012,** *5* (1), 367.
3. (a) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E., GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation* **2008,** *4* (3), 435-447; (b) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; Van Der Spoel, D., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013,** *29* (7), 845-854.
4. (a) Sorin, E. J.; Pande, V. S., Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophysical journal* **2005,** *88* (4), 2472-2493; (b) Best, R. B.; Hummer, G., Optimized molecular dynamics force fields applied to the helix− coil transition of polypeptides. *The journal of physical chemistry B* **2009,** *113* (26), 9004-9015; (c) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* **2010,** *78* (8), 1950-1958.
5. (a) Hermans, J.; Berendsen, H. J.; Van Gunsteren, W. F.; Postma, J. P., A consistent empirical potential for water–protein interactions. *Biopolymers: Original Research on Biomolecules* **1984,** *23* (8), 1513-1518; (b) Berendsen, H. J.; Postma, J. P.; van Gunsteren, W. F.; Hermans, J., Interaction models for water in relation to protein hydration. In *Intermolecular forces*, Springer: 1981; pp 331-342.
6. Bussi, G.; Donadio, D.; Parrinello, M., Canonical sampling through velocity rescaling. *The Journal of chemical physics* **2007,** *126* (1), 014101.
7. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G., A smooth particle mesh Ewald method. *The Journal of chemical physics* **1995,** *103* (19), 8577-8593.
8. Miyamoto, S.; Kollman, P. A., Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of computational chemistry* **1992,** *13* (8), 952-962.
9. Hess, B., P-LINCS: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation* **2008,** *4* (1), 116-122.
10. Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Perez-Hernandez, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J. H.; Noe, F., PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J Chem Theory Comput* **2015,** *11* (11), 5525-42.
11. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-learn: Machine learning in Python. *Journal of machine learning research* **2011,** *12* (Oct), 2825-2830.
12. Chemical Computing Group ULC, S. S. W., Suite #910, Montreal, QC, Canada, H3A 2R7, Molecular Operating Environment (MOE), 2013.08. **2018**.

13. Huey, R.; Morris, G. M., Using AutoDock 4 with AutoDocktools: a tutorial. *The Scripps Research Institute, USA* **2008**, 54-56.
14. Seeliger, D.; de Groot, B. L., Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *Journal of computer-aided molecular design* **2010,** *24* (5), 417-422.
15. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **2010,** *31* (2), 455-461.