

MODELING (UN)BINDING KINETICS OF BIOLOGICALLY RELEVANT SYSTEMS
USING RESAMPLING OF ENSEMBLES BY VARIATION OPTIMIZATION

By

Thomas Dixon

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computational Mathematics Science and Engineering – Doctor of Philosophy
Biochemistry and Molecular Biology – Dual Major

2021

ABSTRACT

MODELING (UN)BINDING KINETICS OF BIOLOGICALLY RELEVANT SYSTEMS USING RESAMPLING OF ENSEMBLES BY VARIATION OPTIMIZATION

By

Thomas Dixon

Conventional drug design optimizes binding affinity when designing molecules to maximize efficacy. However, recent studies show that taking kinetics into account when designing drugs is necessary in some systems where the drug efficacy does not correlate with binding affinity, instead correlating with residence time (RT). To maximize the RT, knowledge of the kinetic pathway is required, but not currently feasible to determine experimentally due to the instability of the transition state. Molecular dynamics (MD) allows us to simulate these pathways with atomic resolution. However, the rare events of interest often occur at timescales as long as milliseconds to hours, and most MD trajectories are computationally limited to the microsecond timescale. In this thesis we use a variant of the Weighted Ensemble (WE) enhanced sampling algorithm, Resampling of Ensembles by Variation Optimization (REVO), to overcome the limitations of MD. This approach is more computationally efficient than conventional MD and does not alter the system’s Hamiltonian nor does it affect the force field parameters used in simulation. We use REVO simulations to produce full binding and unbinding trajectories of biologically relevant systems such as the unbinding of a radioligand bound to Translocator Protein (18kDa) (TSPO), a potential drug target in the treatment of neurodegenerative diseases. We validate these pathways by predicting kinetic rate constants and binding free energies and comparing these results to experiment. Finally, we developed new distance metrics that use experimental data to help guide simulations to a desired conformation. We tested these new distance metrics using Hydrogen deuterium exchange (HDX) data to form the ternary complex between a ligase-proteolysis-targeting chimera (PROTAC) dimer and a target protein.

Copyright by
THOMAS DIXON
2021

To my supportive fiancée and family

ACKNOWLEDGMENTS

Firstly I would like to thank my advisor Dr Alex Dickson for all of his guidance and support. He has been a kind and understanding advisor who has pushed me to be a better scientist and researcher. Secondly, I'd like to thank the members of the Dickson Lab. Working with you has been an amazing experience and I have learned so much from you all. Next I would like to thank my committee for their guidance and insight to help guide my research. I would also like to thank the Department of Computational Mathematics, Science and Engineering as well as the Biochemistry and Molecular Biology Department for hosting my studies as I pursued this degree. Thank you to my undergraduate advisors: Dr. Michelle Ammerman and Dr. Johnathan Wenzel for allowing me to experience the research process for the first time in your labs, the mentorship and guidance you gave me as I prepared for graduate school. I would finally like to thank my friends and family for all of their love, support, and for giving me a place to decompress over the last few years. I could not have done any of this without you.

TABLE OF CONTENTS

LIST OF TABLES	ix
LIST OF FIGURES	x
KEY TO ABBREVIATIONS	xx
CHAPTER 1 INTRODUCTION	1
1.1 Importance of Kinetics in Drug Design	1
1.2 Computational Methods to Determine Kinetics	3
1.2.1 Molecular Dynamics	4
1.2.2 Enhanced Sampling	6
1.2.2.1 Parallel Tempering Methods	7
1.2.2.2 Altered Potential Energy Methods	8
1.2.2.3 Trajectory Parallelization Enhanced Sampling Meth- ods	9
1.2.3 Weighted Ensemble	10
1.2.4 Resampling of Ensembles by Variation Optimization	11
1.2.5 Rate Calculations by Ensemble Splitting	12
1.2.6 Markov State Models	13
1.3 Outline of Work	15
CHAPTER 2 PREDICTING LIGAND BINDING AFFINITY FOR THE SAMPL6 CHALLENGE FROM ON- AND OFF-RATES USING WEIGHTED ENSEMBLES OF TRAJECTORIES .	18
2.1 Introduction	18
2.2 Methods	20
2.2.1 Host-guest systems	20
2.2.2 Dynamics Setup	22
2.2.3 Reweighting of Ensembles by Variance Optimization	24
2.2.4 Calculating rates by ensemble splitting	26
2.2.5 REVO simulation details	28
2.2.5.1 Note about CB8-G3-0 and CB8-G3-4	29
2.2.6 Visualization of trajectory trees	29
2.2.7 Clustering and visualization of conformation space networks	30
2.3 Results	32
2.3.1 Warped walkers	32
2.3.2 Kinetics and free energies	32
2.3.3 Trajectory trees reveal correlation between exit points	36
2.3.4 Conformation space networks reveal connection between starting poses	38
2.4 Discussion	40

CHAPTER 3 ON CALCULATING FREE ENERGY DIFFERENCES USING ENSEMBLES OF TRANSITION PATHS	42
3.1 Introduction	42
3.2 Methods	44
3.2.1 Host-guest systems	44
3.2.2 Molecular dynamics	45
3.2.3 Reweighting of Ensembles by Variance Optimization	46
3.2.4 Calculating rates by ensemble splitting	48
3.2.5 Calculating electrostatic interaction energies	50
3.3 Results	50
3.3.1 Derivation of correction terms	50
3.3.2 Extended trajectory ensembles with lower friction coefficients	54
3.3.3 Free energy estimates, correction terms and comparison with previous benchmarks	58
3.4 Discussion and Conclusion	61
CHAPTER 4 MEMBRANE-MEDIATED LIGAND UNBINDING OF THE PK-11195 LIGAND FROM TSPO	65
4.1 Introduction	65
4.2 Materials and Methods	67
4.2.1 Protein Preparation	67
4.2.2 Docking	68
4.2.3 Molecular Dynamics	68
4.2.4 REVO Resampling	69
4.2.5 Boundary Conditions	72
4.2.6 Clustering and Network Layout	72
4.2.7 Quantifying Unbinding Pathways	73
4.2.8 Calculating Non-bonded Energies	73
4.2.9 Calculating Off-Rates and Mean First Passage Times using Hill Relation	74
4.2.10 Calculating Mean First Passage Times using Markov State Models	75
4.2.11 Selecting Poses for Straightforward MD Simulations	76
4.3 Results	77
4.3.1 PK-11195 Unbinding Pathway	77
4.3.2 PK-11195 Rates and Residence Times	85
4.3.3 PK-11195 Transition State	87
4.4 Discussion and Conclusion	98
CHAPTER 5 ATOMIC-RESOLUTION PREDICTION OF DEGRADER- MEDIATED TERNARY COMPLEX STRUCTURES BY COMBINING MOLECULAR SIMULATIONS WITH HY- DROGEN DEUTERIUM EXCHANGE	101
5.1 Introduction	101
5.2 Methods	106

5.2.1	Experimental Methods	106
5.2.1.1	Cloning, expression and purification of SMARCA2 and VHL/EloB/C	106
5.2.1.2	Hydrogen Deuterium Exchange Mass Spectrometry	108
5.2.1.3	Structural Determination of SMARCA2:ACBI1:VHL Complex	110
5.2.2	Computational Methods	111
5.2.2.1	Unbound System Preparation	111
5.2.2.2	Molecular Dynamics	112
5.2.2.3	Generating Bound Ensemble	112
5.2.2.4	REVO-epsilon Weighted Ensemble method	113
5.2.2.5	Distance Metrics	114
5.2.2.6	Ternary complex docking protocol	115
5.2.2.7	HREMD simulation	116
5.2.2.8	Conformational free energy landscape determination	117
5.2.2.9	Calculating Interface RMSD	120
5.3	Results	122
5.3.1	Degraders with different efficiency induce similar ternary complex structures in X-ray crystallography.	122
5.3.2	Hydrogen Deuterium Exchange Reveals Extended Protein- Protein Interfaces	125
5.3.3	Efficient simulation of ternary complex formation using REVO Weighted Ensemble simulations	130
5.3.4	HDX improves prediction of ternary complex using docking	136
5.3.5	Conformational sampling of ternary complexes	137
5.4	Discussion	144
CHAPTER 6 SUMMARY OUTLOOK AND IMPACT		147
BIBLIOGRAPHY		151

LIST OF TABLES

Table 2.1: Pose-averaged rates and affinities	33
Table 3.1: Binding and unbinding rates as a function of friction coefficient (γ). The uncertainties shown use the standard error of the mean calculated from 5 and 10 independent REVO runs for binding and unbinding, respectively. The quantities from Chapter 2 were obtained with 5 REVO runs that used different initial conformations, each of which were 2000 cycles in length.	56
Table 3.2: Raw (ΔG^0) and corrected (ΔG_{corr}) free energy values using simulation data from three different friction coefficients. Values are in kcal/mol and uncertainties are calculated using propagation of the standard error of the mean.	59
Table 5.1: Binding affinity (K_d), efficiencies (IC50, DC50), and cooperativity (α) of PROTAC 1, PROTAC 2, and ACBI1 degraders. Ternary IC50 and binary (SMARCA2) DC50 values are reported; the cooperativity is the ratio of binary over ternary IC50. Table adapted from Farnaby et al. [208].	105
Table 5.2: Details of Hamiltonian Replica Exchange Molecular Dynamics (HREMD) simulations. Protein complexes, number of atoms in a simulation box, number of replicas used and the aggregate length of the simulations are listed.	119
Table 5.3: Details of HREMD simulations. Effective temperatures and average exchange probabilities of neighboring replicas are listed.	121
Table 5.4: Crystallographic table for protein crystal structure 7S4E SMARCA2-iso2:ACBI1: von Hippel-Lindeu protein (VHL).	123
Table 5.5: A summary of the performance of REVO simulations run with different distance metrics. Each REVO simulation ran with 48 walkers. The number of binding events (N_{binding}) counts the barrier crossings into the bound state, defined using an interface root mean square deviation (I-RMSD) < 2.0 Å. The number of simulations with binding events (Sims. w/ binding) shows the probability of binding success. The total simulation time (Sim. time) aggregates the length of all trajectories in each REVO simulation.	132
Table 5.6: Comparison of k_{on} rates between simulation and experiment for the ACBI1 PROTAC 1, and PROTAC 2 systems. The experimental rate for PROTAC 2 has not been determined yet.	136

LIST OF FIGURES

Figure 1.1:	Amount of drug concentration in the blood stream over time. Panel A shows this relationship on a linear scale. The effective minimum concentration is shown in blue and the minimal toxic concentration is shown in red. Panel B shows the log of drug concentration vs time after the maximum concentration has been reached. From the semilog plot, we can determine the elimination rate from the slope.	2
Figure 1.2:	(Left) The equation to calculate potential energy of a molecular system. r is the bond length. θ is the bond angle, ϕ is the dihedral angle and r_{ij} is the atomic distance between atoms i and j . k_r , k_θ , and k_ϕ are force constants. r_{eq} , θ_{eq} , and ϕ_{eq} are equilibrium positions. The n is multiplicity, γ is a phase shift to describe a periodic dihedral term. The e_{ij} is the Lennard-Jones well depth and r_m is the distance the potential reaches its minimum. q_i and q_j are charges for atoms i and j and ϵ_0 is the dielectric constant. (Center) Each summation calculates a separate type of energy. (Right) A pictorial representation of each type of energy. The parameters are defined by the molecular dynamic force field. This figure is modified from Ref [42].	5
Figure 2.1:	Structure of the ligands used in this study. (Top) Quinine, referred to herein as cucurbit[8]uril (CB8)-G3. (Middle) 5-hexenoic acid (deprotonated form), referred to herein as octa acid (OA)-G3. (Bottom) 4-methyl pentanoic acid (deprotonated form), referred to here ohhh in as OA-G6.	21
Figure 2.2:	Starting poses for CB8-G3. Side and top views are shown. Coloring for pose indices is consistent with Figures 2.7, 2.8 and 2.9.	22
Figure 2.3:	Starting poses for OA-G3. Side and top views are shown. Coloring for pose indices is consistent with Figures 2.7, 2.8 and 2.9.	23
Figure 2.4:	Starting poses for OA-G6. Side and top views are shown. Coloring for pose indices is consistent with Figures 2.7, 2.8 and 2.9.	23
Figure 2.5:	The REVO algorithm. Each cycle begins by running an ensemble of walkers forward in time using unbiased dynamics. The distances between the walkers are used to calculate a variance (Eq. 2.2). In the resampling loop (blue), coupled cloning and merging operations are proposed, and they are accepted only if they result in a higher V . If the proposed V is lower, the resampling loop is terminated and dynamics are continued for the next cycle.	25

- Figure 2.6: Ensemble splitting. An equilibrium host-guest binding system is split into two non-equilibrium ensembles for the calculation of on and off-rates. This is done by defining “bound” and “unbound” basins (left and right of each ensemble). The “unbinding” ensemble (top) is the set of trajectories that have most recently visited the bound basin. The “binding” ensemble (bottom) is the set of trajectories that most recently visited the unbound basin. The on and off-rates are directly computed using the time averaged trajectory flux ($\bar{\phi}_b$ or $\bar{\phi}_u$) between the ensembles. 27
- Figure 2.7: Weights of warped walkers. Weights of warping events for the unbinding (top row) and rebinding (bottom row) simulations. In both cases the points are colored according to the index of the corresponding starting pose (0, blue; 1, red; 2, yellow; 3, green; 4, brown). 33
- Figure 2.8: Spatial distribution of warped walkers. Structures of warping events for the unbinding simulations viewed from the front and back. Guest ligands are colored according to the index of the corresponding starting pose (0, blue; 1, red; 2, yellow; 3, green; 4, brown). 34
- Figure 2.9: Predicted kinetics and free energies. The calculated free energies (top), off-rates (middle), and on-rates (bottom) are shown as a function of simulation time for each starting pose in each host-guest system. The curves are colored according to the index of the starting pose as in Figures 2.7 and 2.8. The calculated binding free energies are compared with experimental measurements (horizontal red line) [123], and the computational reference (dashed black line) for each system. 35
- Figure 2.10: Trajectory trees show all cloning and merging events in a simulation. The trajectory tree for the first 1329 cycles of the OA-G3-0 unbinding simulation is shown. Each horizontal row in this tree represents a cycle, and the placement of all 48 nodes in the row is determined by minimizing an energy function (see “Visualization of trajectory trees” in Methods). solvent accessible surface area (SASA) is used to color the nodes, with blue and dark green indicating bound structures, and yellow to orange indicating unbound. 37
- Figure 2.11: Conformation space networks for the unbinding simulations. Each node in a conformation space network (CSN) represents a cluster of host-guest structures. Edges in the networks connect clusters that are seen to inter-convert in the REVO simulations. The size of each node is proportional to the number of times it was observed in the unbinding simulations. Nodes are colored according to the solvent accessible surface area of the guest molecule, as shown in the color-bars on the right. The clusters corresponding to the starting poses are labeled in each network. 39

Figure 3.1:	(A) The initial pose for the OA-G6 system (side view: left, top view: right). Note that some atoms from the host are removed in the side view for clarity. The carboxyl oxygens are shown in sphere representation. (B) The chemical structure of the G6 ligand in the deprotonated form.	45
Figure 3.2:	Splitting an equilibrium ensemble into two history-dependent ensembles using basins. The bound and unbound basins are shown in grey and light orange, respectively. The unbinding ensemble (B, top) contains all trajectories that last visited the bound basin, which are shown in black. The binding ensemble (B, bottom, also referred to as the “rebinding” ensemble) contains all trajectories that last visited the unbound basin, shown in red. Simulations in a given ensemble are terminated once they reach the destination basin and thus switch ensembles. The trajectory flux between ensembles is denoted by $\phi_{u \rightarrow b}$ and $\phi_{b \rightarrow u}$. The quantity π_b refers to the probability of the entire top ensemble, and the quantity f_b denotes the probability of the bound basin within the unbinding ensemble.	48
Figure 3.3:	(A) Average temperatures observed in short simulations for different friction coefficients (γ). (B) Probability distributions of observed temperatures from ensembles of longer simulations with different γ .	55
Figure 3.4:	Predicted on- (top) and off-rates (bottom) as a function of simulation time. Each panel is labeled according to the friction coefficient used for that set of simulations. The independent simulations are shown in shades of orange (k_{on}) and blue (k_{off}), and the averages are depicted by bold black lines.	56
Figure 3.5:	Binding (top) and unbinding (bottom) fluxes for $\gamma = 0.001 \text{ ps}^{-1}$. Fluxes are shown for each simulation individually. Parameters are the same as those used for higher γ values in the main text. Average fluxes over the simulations are shown as thick black lines.	57
Figure 3.6:	Weights of warped walkers in unbinding (top) and binding (bottom) REVO simulations for $\gamma = 0.01, 0.1$ and 1.0 ps^{-1} . Each simulation is shown in a different color.	58
Figure 3.7:	Weights of warped walkers in unbinding (top) and binding (bottom) REVO simulations for $\gamma = 0.001 \text{ ps}^{-1}$. Each simulation is shown in a different color. Parameters are the same as those used for higher γ values in the main text.	59

Figure 3.8:	Free energies as a function of friction coefficient. The dark blue line shows the uncorrected free energies calculated at three different γ values. The light blue line shows the corrected values, which are shifted upwards by 2.72 kcal/mol. The thin red line shows the value reported in Chapter 2, which employed a friction coefficient of 1.0 ps^{-1} and used a smaller dataset than is reported here. The black horizontal line shows the value of a computational reference computed using alchemical perturbation, reported in Ref. [148]. The dashed grey line shows the experimental measurement, reported in Ref. [153].	61
Figure 4.1:	TSPO-PK-11195 system. (A) Front view of the TSPO dimer in the membrane with PK-11195 bound. (B) All six starting poses are shown from the side view, along the inter-dimer axis. To compare poses, two moieties of PK-11195 are colored in black (o-chlorophenyl) and magenta (1-methylpropyl), with the rest of the molecule colored according to atom name. TM-2 is shown as transparent for clarity.	69
Figure 4.2:	Protein-ligand interaction plots for the six starting conformations. The red suns indicate that the residue has a hydrophobic contact with PK-11195. The green dashed lines show hydrogen bonds.	70
Figure 4.3:	The energy of non-bonded interactions between PK-11195 and TSPO as a function of minimum distance between PK-11195 and TSPO.	75
Figure 4.4:	Combined CSN of all REVO simulations from each starting pose. Each node in the network represents a cluster of ligand poses and is sized according to the cluster weight. Nodes are connected by edges if the ligand poses are observed to interconvert in the REVO trajectory segments. Nodes are colored according to the lipid accessible surface area (LASA). Starting poses are marked in bold and transition state poses shown in Fig. 4.5D are marked in italics.	78

- Figure 4.5: Analysis of membrane-mediated exit paths. (A) The coordinate Q_{ij} is defined as the x - y distance between the center of mass of PK-11195, shown as sticks and colored by atom type, and the line that connects the centers of mass of helix i and helix j . LP1 is not shown here for clarity. (B) The expectation values of the interaction energy between PK-11195 and TSPO (blue) and between PK-11195 and the membrane (black) are shown as a function of Q . In each case the solid line shows Q_{12} and the dashed line shows Q_{25} . The shaded region indicates the standard error over the ensemble of measurements at each Q value. (C) Probability curves projected onto Q_{12} for simulations initialized in Pose D1 (blue) and D2 (orange). Q_{12} values of the starting structures are marked with (*). (D) Poses from transition pathways with $Q \approx 0$. These poses are also labeled in the CSN of Fig. 4.4. Phe46 is shown in purple and Trp50 is shown in orange. (E) A set of poses along the Q_{12} pathway colored from bound (red) to unbound (blue). Top view is shown on the left and a front view is shown on the right. (F) The minimum PK-11195-TSPO distance and the Q_{12} value is shown for each pose in panel (E). (G) The z center of mass (COM) position as a function of Q_{12} . The red lines indicate the upper and lower bounds of the membrane as defined by the maximum and minimum z coordinate of the lipid membrane. 79
- Figure 4.6: CSN networks indicating the clusters that were observed from each initial pose. Red nodes indicate the simulations observed a TSPO-PK-11195 conformation that was clustered into that node. 80
- Figure 4.7: Expectation value for E_{int} as a function of Q_{12} . The lines are colored by residue. Only residues who have a minimum interaction energy below -3.5 kcal/mol are shown. The standard error is shown in the lighter shaded regions. 82
- Figure 4.8: Expectation value for E_{int} as a function of Q_{25} . The lines are colored by residue. Only residues who have a minimum interaction energy below -3.5 kcal/mol are shown. The standard error is shown in the lighter shaded regions. 82
- Figure 4.9: The residues with the strongest non-bonded interactions with PK-11195 on the Q_{12} pathway. This summarizes the curves in Fig. 4.7, plotting the minimum E_{int} against the Q_{12} value for which this minimum value is observed. The colors indicate the region of TSPO, blue for residues on TM-1 and black for residues on TM-2. Only residues with a non-bonded energy below -3.5 kcal/mol are shown. 83

Figure 4.10: The residues with the strongest non-bonded interactions with PK-11195 on the Q_{12} pathway. This summarizes the curves in Fig. 4.8, plotting the minimum E_{int} against the Q_{25} value for which this minimum value is observed. The colors indicate the region of TSPO, red indicates residues on the LP1 loop, black for residues on TM-2 and orange for residues on TM-5. Only residues with a non-bonded energy below -3.5 kcal/mol are shown.	83
Figure 4.11: Residues moving along with the ligand during dissociation. Expectation values of Q_{12} for individual residues are shown as a function of the Q_{12} of PK-11195.	84
Figure 4.12: Residues moving along with the ligand during dissociation. Expectation values of Q_{25} for individual residues are shown as a function of the Q_{25} of PK-11195.	84
Figure 4.13: The average dihedral angles for the Markov state model (MSM) states for four different rotatable bonds on the PK-11195 ligand.	88
Figure 4.14: The standard deviation of the dihedral angles for the MSM states for four different rotatable bonds on the PK-11195 ligand.	89
Figure 4.15: The range of the dihedral angles for the MSM states for four different rotatable bonds on the PK-11195 ligand.	90
Figure 4.16: (A) mean first passage time (MFPT) estimates using unbinding fluxes observed over the course of REVO simulations. The light shaded area shows the standard error across the three simulations conducted for each pose. (B) A bar graph of the final MFPTs comparing the Hill Relation (green), MSM simulations before (grey), and after (black) the addition of new straight forward MD simulations. Pose-specific MFPTs were computed from MSMs that were built using only trajectories generated from that starting pose. Simulations starting from pose R never entered the unbound basin and thus MFPTs could not be determined by either method. The experimental MFPT of 34 min is shown as a dashed blue line in each panel.	91
Figure 4.17: Combined conformation space network of all REVO simulations from each starting pose with the addition of frames from straightforward MD simulations, colored by (A) LASA and (B) committor probability. Starting poses are marked in bold in panel A.	92

Figure 4.18: An MSM network including both straight forward and REVO trajectories colored by pose specific committor probability values calculated using trajectories beginning in pose 4RYI. States that were not visited by these simulations are colored grey.	93
Figure 4.19: An MSM network including both straight forward and REVO trajectories colored by pose specific committor probability values calculated using trajectories beginning in pose D1. States that were not visited by these simulations are colored grey.	94
Figure 4.20: An MSM network including both straight forward and REVO trajectories colored by pose specific committor probability values calculated using trajectories beginning in pose D2. States that were not visited by these simulations are colored grey.	95
Figure 4.21: An MSM network including both straight forward and REVO trajectories colored by pose specific committor probability values calculated using trajectories beginning in pose D3. States that were not visited by these simulations are colored grey.	96
Figure 4.22: An MSM network including both straight forward and REVO trajectories colored by pose specific committor probability values calculated using trajectories beginning in pose D4. States that were not visited by these simulations are colored grey.	97
Figure 5.1: Potential energy of all replicas from HREMD simulation of Sys7 . Left to right: rank0 to rank19. A good overall between adjacent replicas suggests a sufficient number of replicas were employed and also confirmed no phase transition took place during the HREMD simulation.	117
Figure 5.2: Effective temperature trajectories of replicas 0 (red), 5 (blue), 10 (green) and 19 (grey) from HREMD simulation of Sys7	118

Figure 5.3:	Ternary complex of SMARCA2 and VCB induced by ACBI1 shows structural similarities with PROTAC 1 and PROTAC 2: a Overall perspective of SMARCA2 Isoform 2 (green) and VHL/ElonginC/ElonginB (grey) induced by degrader molecule ACBI1 (bright orange). b ACBI1-induced interface contacts between SMARCA2 and VCB. The proteins are shown in space-filling, the colors are as in a , annotated residues are among those that make the highest number of contacts (see c). c A contact map for the interface of the crystal structure. The circle size reflects the number of atoms (including hydrogen atoms) participating in interactions. d Superposition of 6HAY (purple), 6HAX (salmon), 7S4E (green) by aligning VHL (grey) shows varied conformations of the warheads of the three degraders PROTAC 1, PROTAC 2, or ACBI1 (up to 1.7 Å) resulting in alterations of SMARCA2 within the ternary complex.	124
Figure 5.4:	Peptic coverage map of proteolyzed proteins SMARCA2, VHL, Elongin C and Elongin B.	127
Figure 5.5:	Relative uptake heat map of HDX exchange data of all PROTAC molecules 1, 2 and ACBI1 bound to binary and Ternary State SMARCA2 isoform 2 bromo domain.	128
Figure 5.6:	Relative uptake heat map of HDX exchange data of all PROTAC molecules 1, 2 and ACBI1 bound to binary and Ternary State VHL.	129
Figure 5.7:	Relative uptake heat map of HDX exchange data of all PROTAC molecules 1, 2 and ACBI1 bound to binary and Ternary State Elongin C.	130
Figure 5.8:	ACBI-induced ternary complex formation of SMARCA2 isoform 2:VCB leads to protection of specific sites:a-d, SMARCA2 isoform 2(a), VHL(b), Elongin C(c), and Elongin B(d) monitored for hydrogen-deuterium exchange over time. The difference plots of each protein in the binary and ternary states are generated by subtracting the deuterium exchange of like peptides of the APO or binary from the binary or ternary states (defined as Binary Δ APO and Ternary Δ Binary), respectively. Regions that exchange significantly less than the comparative state are depicted in blue (negative), whereas regions that exchange significantly more appear in red (positive). The resultant difference plots of the binary (e), or ternary complex (f) were mapped onto the structure 7S4E. The experiments were repeated on 2 separate days.	131

Figure 5.9: Comparing the w-RMSD, number of target-ligase contacts, and triple distance metrics (Linear combination of w-RMSD, target-ligase contacts and number of target-PROTAC contacts). **(a)** The minimum I-RMSD over time during the simulation for the triple distance metric. Each green line indicates one replica and the black line is the average between all runs. The blue line is a straightforward MD simulation run on Folding@home. **(b)** The minimum I-RMSD for each distance metric. **(c)** A scatter plot of the free energy vs the I-RMSD after clustering the 6HAX simulations. The circles are colored by w-RMSD. **(d)** The predicted binding rates for PROTAC 1 system (purple) and the ACBI1 system (green). The black line is the experimental on-rate determined via Surface Plasmon Resonance (SPR). 134

Figure 5.10: Illustration of the representative prediction produced by REVO simulation and its comparison to the co-crystallized structure (Protein Data Bank (PDB) ID: 6HAX) **(a)** predicted ternary structure with I-RMSD=1.1 Å; **(b)** detail of the binding interface; **(c)** contact maps for the interfaces of co-crystallized and predicted structures. The circle size reflects the number of atoms (including hydrogens) participating in interactions; **(d)** structurally aligned prediction (green) and co-crystallized structure (pink) with a detailed PROTAC 2 comparison shown. 135

Figure 5.11: Comparing the bound ensembles determined by docking and REVO simulations with and without information from HDX for the PDB ID 6HAX ternary complex. The REVO bound ensemble is defined as structures below a warhead RMSD of 2 Å and more than 30 contacts between the target and ligase interface. The docking bound basin is defined as the 100 top structures as determined by Rosetta-scoring. **(a)** Probability density function distributions of I-RMSD values for the bound ensembles. **(b)** The percent of structures in the predicted bound ensembles below specific I-RMSD thresholds (2 Å, 2.5 Å, and 3 Å). 138

Figure 5.12: Most populated structures of SMARCA2 bound to VHL with different degrader molecules, identified by dimension reduction and clustering of HREMD simulation data. **(a-d)** Colors of VHL and SMARCA2 represent HDX protection in the presence of the degrader molecules relative to the situation in the absence of the degrader. The second ranked structures of **c** PROTAC 2 and **d** isoform 1 SMARCA2 are displayed that support HDX data, whereas the top three structures are included in Figure 5.13. Elongin B and Elongin C are also included in panel **d**. **e** The top structures of ternary complexes are compared after aligning VHL to illustrate conformational differences among top structures of ternary complexes. 139

- Figure 5.13: Cluster centroids from the three highest populated structures of SMARCA2-iso2 bound to VHL via (a) ACBI1, (b) PROTAC 1, and (c) PROTAC 2, along with their populations. Less populated structures are omitted. 140
- Figure 5.14: Free energy landscapes determined from Principle Component Analysis (PCA) projections of SMARCA2-iso2 bound to VHL via (a) ACBI1, (b) PROTAC 1, and (c) PROTAC 2. Red points indicate k -means centroids. 140
- Figure 5.15: **a** Conformational free energy landscape as a function of the first two Time-structure independent components analysis (TICA) features of SMARCA2-PROTAC2-VHL ternary complex inferred from a MSM. The ensemble of bound states from REVO simulations is shown as blue points; the crystal structure (PDB ID 6HAX) is shown as a red X. In this projection, states II and V are close to state I. **b** Network diagram of the coarse-grained MSM calculated using a lag time of 50 ns, with the stationary probabilities associated with each state indicated. **c** mean first passage time (MFPT) from one state in the MSM to another. Numbers indicate predicted MFPTs in μ s. **d-e** Comparison of the crystal structure (gray) with the lowest free energy state (cyan) and a metastable state (orange) predicted by the MSM. Arrows indicate a change of orientation relative to **d**. 142
- Figure 5.16: Contact maps from the (a) co-crystallized structure 6HAX; (b) global minimum state and (c) metastable state identified by our MSM. 143

KEY TO ABBREVIATIONS

- RT** residence time. ii, 1, 3, 16, 18, 34, 40, 43, 65, 67, 75, 85, 86, 99, 100, 148, 150
- MD** Molecular dynamics. ii, xv, 3, 4, 5, 6, 7, 8, 10, 12, 13, 14, 15, 19, 43, 46, 50, 64, 66, 68, 69, 71, 76, 86, 91, 92, 100, 103, 105, 112, 113, 116, 120, 132, 141, 144, 149
- WE** Weighted Ensemble. ii, 10, 11, 12, 13, 24, 46, 103, 104, 130, 148, 149
- REVO** Resampling of Ensembles by Variation Optimization. ii, ix, x, xi, xii, xiii, xv, xvi, xviii, xix, 11, 12, 16, 17, 20, 24, 25, 26, 27, 28, 29, 32, 33, 34, 35, 36, 38, 39, 46, 47, 48, 49, 54, 56, 58, 59, 67, 69, 71, 72, 75, 77, 78, 91, 92, 93, 94, 95, 96, 97, 99, 101, 104, 105, 113, 114, 117, 120, 130, 132, 133, 134, 135, 136, 137, 138, 141, 142, 145, 146, 147, 149, 150
- TSPO** Translocator Protein (18kDa). ii, xiii, xiv, xv, 16, 65, 66, 67, 68, 69, 71, 72, 74, 75, 77, 79, 80, 81, 83, 84, 87, 98, 99, 100, 147, 150
- HDX** Hydrogen deuterium exchange. ii, xvii, xviii, 17, 101, 103, 104, 105, 108, 109, 110, 114, 115, 126, 128, 129, 130, 132, 133, 134, 137, 138, 139, 141, 144, 146, 149
- PROTAC** proteolysis-targeting chimera. ii, ix, xvii, xviii, xix, 17, 101, 102, 104, 105, 109, 111, 112, 113, 114, 115, 119, 122, 124, 125, 128, 129, 130, 132, 134, 135, 136, 139, 140, 141, 142, 144, 145, 148, 149, 150
- HREMD** Hamiltonian Replica Exchange Molecular Dynamics. ix, xvi, xviii, 101, 106, 116, 117, 118, 119, 121, 137, 139, 141, 145
- VHL** von Hippel-Lindeu protein. ix, xvii, xviii, xix, 102, 104, 105, 106, 107, 108, 110, 111, 112, 113, 115, 117, 119, 120, 122, 123, 124, 126, 127, 128, 129, 131, 132, 139, 140, 141, 142, 144, 145, 146, 150
- I-RMSD** interface root mean square deviation. ix, xviii, 101, 120, 132, 133, 134, 135, 136, 137, 138, 144, 145, 149
- CB8** cucurbit[8]uril. x, 20, 21, 22, 27, 29, 31, 32, 33, 34, 38
- OA** octa acid. x, xi, xii, 21, 22, 23, 31, 32, 33, 36, 37, 38, 41, 44, 45, 48, 54, 55, 62
- SASA** solvent accessible surface area. xi, 36, 37, 38, 41
- CSN** conformation space network. xi, xiii, xiv, 16, 31, 38, 39, 67, 72, 75, 77, 78, 79, 85, 86
- LASA** lipid accessible surface area. xiii, xv, 76, 77, 78, 82, 86, 92

COM center of mass. xiv, 68, 73, 79

MSM Markov state model. xv, xvi, xix, 13, 14, 15, 16, 67, 85, 86, 88, 89, 90, 91, 93, 94, 95, 96, 97, 106, 120, 141, 142, 143, 145, 147

MFPT mean first passage time. xv, xix, 29, 33, 34, 75, 76, 85, 86, 91, 100, 142

SPR Surface Plasmon Resonance. xviii, 3, 108, 134

PDB Protein Data Bank. xviii, xix, 66, 67, 68, 98, 104, 105, 111, 112, 116, 122, 132, 135, 138, 142, 144

PCA Principle Component Analysis. xix, 14, 117, 118, 119, 137, 140

TICA Time-structure independent components analysis. xix, 14, 120, 141, 142, 145

GPU graphics processing unit. 6

CV collective variable. 8, 9, 11, 103, 130, 145, 149

WHAM Weighted Histogram Analysis Method. 8

RMSD root mean square deviation. 26, 27, 47, 48, 71, 103, 120, 145, 149

NAM Northrup-Allison-McCammon. 41

FEP free energy perturbation. 62

NMR nuclear magnetic resonance. 65, 66, 98, 102, 125

RAMD random accelerated molecular dynamics. 66

CGENFF CHARMM Generalized Force Field. 67

VDAC voltage dependent ion channel. 99

POI protein of interest. 101, 102, 104, 149

TPD targeted protein degradation. 102

CHAPTER 1

INTRODUCTION

1.1 Importance of Kinetics in Drug Design

Historically, ligand efficacy has been predicted by measuring the thermodynamics of ligand binding. Examples include measurements of the half maximal inhibitory concentration (IC_{50}), equilibrium dissociation, and the change in binding free energy (ΔG_{bind}) [1, 2]. Although the binding affinity has been successfully used to guide drug design, [3, 4, 5] these experiments are conducted under closed equilibrium conditions. Living organisms do not match these conditions as the body is in constant flux to maintain homeostasis. For example the concentration of the ligand takes time to distribute throughout the body after a dose is taken and does not necessarily ever reach equilibrium, as the body is also working to metabolize and eliminate the ligand from the body. To address the above, Copeland has suggested the key parameter to maximize the ligand efficacy should not be binding affinity, but rather the residence time (RT), or the average time it takes for the ligand to unbind from the target [2]. This motivates taking kinetics into consideration when developing new drugs.

After a drug is administered, it is absorbed and distributed throughout the body. The body then metabolizes and excretes the drug. When the first two processes are dominant, the concentration of drug in the blood plasma increases and when metabolizing and excreting the drug becomes dominant, the concentration decays (Figure 1.1 A). In order to observe a therapeutic effect, the drug concentration must be above a minimum threshold, called the minimum effective concentration. The amount of drug introduced into the body is bound from above by the minimal toxic concentration, or the concentration where an organism starts observing harmful effects by the drug. The range between these two thresholds makes up the therapeutic window. The goal of drug design is to widen the therapeutic window and

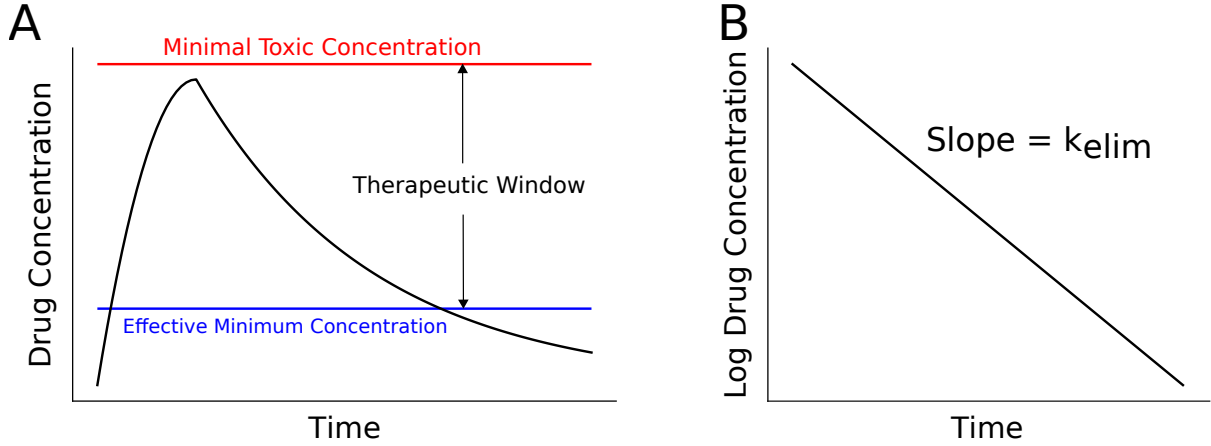


Figure 1.1: Amount of drug concentration in the blood stream over time. Panel A shows this relationship on a linear scale. The effective minimum concentration is shown in blue and the minimal toxic concentration is shown in red. Panel B shows the log of drug concentration vs time after the maximum concentration has been reached. From the semilog plot, we can determine the elimination rate from the slope.

maximize the amount of time that a given dose of a drug remains effective.

One method to maximize the time within the therapeutic window is to minimize the rate of decay, i.e. increasing its half-life [6]. If a drug has a fast half-life then doses will either need to be at higher concentrations, making it more likely that toxic effects are observed, or be taken more frequently, which requires more effort on the part of the person taking the medication. We can determine the half-life of our drug ($\tau_{1/2}$) by calculating the rate of elimination (k_{elim}), which is done by taking the log of the drug concentration in the blood after the maximum concentration has been reached and calculating the slope of the semilog plot [7] (Figure 1.1 B). The half-life is then calculated by the following relation:

$$\tau_{1/2} = \frac{\log(2)}{k_{elim}}. \quad (1.1)$$

However, there is often an observed time lag between the plasma concentration and the observed therapeutic effect [8]. To take this delay into account it was hypothesized that the drug took time to move from the blood into the tissue of interest and bind to the target of interest [9, 10]. Several previous studies have linked the change in drug concentration

and therapeutic effects with the (un)binding rate between the drug and the target protein [11, 12, 13]. Pharmacokinetic-pharmacodynamic models have been developed that integrate drug binding kinetics with thermodynamic information to predict drug activity [14]. These models significantly improve the prediction in cases of long drug RTs as previous models assumed the existence of a rapid equilibrium between drug and target and under predicted drug activity in this scenario.

The optimization of ligand kinetic rates, also known as kinetics-orientated drug design, has recently been gaining traction when designing new drugs [15, 16, 17]. This is due to identifying systems where binding affinity does not correlate with drug efficacy, rather the RT does.[18, 19, 20]. By developing drugs with longer RTs, we can ensure the therapeutic effects of the drug are observed for longer periods of time.

1.2 Computational Methods to Determine Kinetics

We have established the importance of kinetics and RTs in drug design. There are several methods to experimentally determine the kinetics from radioligand binding [21, 22], Surface Plasmon Resonance (SPR) [23, 24] and florescence assays [25, 26, 27]. However, developing drugs to optimize the ligand kinetics experimentally is difficult because we can not understand the mechanism with which the (un)binding takes place. In particular, identifying the transition state – the maximum free energy state along the unbinding pathway – is critical to determine in order to develop drugs with longer RTs. Unfortunately, characterizing these states experimentally is not feasible because the system is at this state for only an instant during the unbinding event. Therefore, we turn to computational methods to help us investigate the kinetic pathways. In particular in this thesis we will use an algorithm which can give us atomic resolution, molecular dynamics (MD), to observe these pathways.

1.2.1 Molecular Dynamics

MD is a computational algorithm that simulates molecular interactions and motion at the atomic level using classical mechanics [28]. First developed to study phase transitions using hard spheres [29] and radiation damage [30], modern computing allows simulations to study more complex phenomena such as molecular diffusion through biological membranes [31, 32], DNA supercoiling [33], and stability of intermolecular conformations [16, 34, 35, 28]. The MD algorithm works by first setting initial conditions of the system (atomic positions and velocities). Then the net force on each atom (\mathbf{F}) is calculated by:

$$\mathbf{F}(t) = -\nabla U(\mathbf{r}(t)), \quad (1.2)$$

where U is the potential energy for a set of atomic positions (\mathbf{r}), at time t . For MD simulations the force is in $\frac{\text{kcal}}{\text{mol}}$, potential energy is in the units of kcal/mol, the positions are in Å, and time is in fs.

The potential energy function and its associated parameters are determined by the force field used in the simulation. This potential energy function includes terms for bonds, angles, dihedrals between covalently bonded atoms, and non-bonded energies (electrostatics and Lennard-Jones) for non-covalently bonded atom pairs (Figure 1.2). Some common force fields used in biomolecular MD simulations are AMBER[36, 37, 38], CHARMM [39, 40], and GROMOS [41].

After the force has been calculated, we then solve Newton’s equations to determine the change in atomic positions

$$\frac{d\mathbf{r}(t)}{dt} = \mathbf{v}(t), \quad (1.3)$$

and change in atomic velocities (\mathbf{v})

$$\frac{d^2\mathbf{r}(t)}{dt^2} = \frac{d\mathbf{v}(t)}{dt} = \frac{\mathbf{F}(t)}{m}, \quad (1.4)$$

over the time step. The mass, m , has units of grams/mol, and velocities are in Å/fs. Once the changes are known, we update the positions and velocities for the atoms. We repeat this



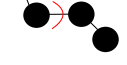
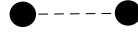
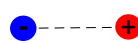
$U(\mathbf{r}) = \sum_{bonds} k_r (r - r_{eq}) +$	Bond	
$\sum_{angles} k_\theta (\theta - \theta_{eq}) +$	Angle	
$\sum_{dihedrals} k_\phi (1 + \cos [n\phi - \gamma]) +$	Dihedral	
$\sum_{LJ} \epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right) +$	Van der Waals	
$\sum_{elec} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}}$	Electrostatics	

Figure 1.2: (Left) The equation to calculate potential energy of a molecular system. r is the bond length. θ is the bond angle, ϕ is the dihedral angle and r_{ij} is the atomic distance between atoms i and j . k_r , k_θ , and k_ϕ are force constants. r_{eq} , θ_{eq} , and ϕ_{eq} are equilibrium positions. The n is multiplicity, γ is a phase shift to describe a periodic dihedral term. The ϵ_{ij} is the Lennard-Jones well depth and r_m is the distance the potential reaches its minimum. q_i and q_j are charges for atoms i and j and ϵ_0 is the dielectric constant. (Center) Each summation calculates a separate type of energy. (Right) A pictorial representation of each type of energy. The parameters are defined by the molecular dynamic force field. This figure is modified from Ref [42].

process as many times as necessary to observe the phenomenon of interest with statistical significance.

The above formulation for MD simulations was designed to keep the energy of the system constant. However, many biological experiments are not conducted under constant energy, rather at constant temperature. To keep MD simulations at constant temperature, a thermostat is required such as a Langevin thermostat using Langevin dynamics [43]. This algorithm uses the above formulation for MD as explained above, but adds two additional terms:

$$\mathbf{F}(t) = -\nabla U(\mathbf{r}(t)) - \gamma m \mathbf{v}(t) + \sqrt{2m\gamma k_b T} \mathbf{R}(t), \quad (1.5)$$

where γ is the friction coefficient in units of ps^{-1} and is generally set to 1, k_b is the Boltzmann constant in units of $\frac{kcal}{molK}$, T is the absolute temperature in K, and $\mathbf{R}(t)$ is a Gaussian random process centered at 0 and has the following properties:

$$\langle \mathbf{R}(t) \rangle = 0, \quad (1.6)$$

$$\langle \mathbf{R}(t)\mathbf{R}(t') \rangle = \delta(t - t'), \quad (1.7)$$

where δ is the Dirac delta function. The second term is designed to take energy out of the system and damp the force [44]. The third term represents random thermal fluctuations from small particles that can add energy back into the system [44]. The sum of these two terms are what maintains the temperature of the simulation [45].

MD is an excellent tool to study the biological pathways of binding and unbinding, create statistical ensembles to describe these events and make hypotheses that can be tested experimentally. However, it is difficult to observe long timescale phenomena in simulation, due to the discrepancy between the simulation time steps and the natural timescales with which these events take place. MD time steps are constrained to 1-2 fs in order to capture the fastest molecular motion (oscillation along covalent bonds). Typical ligand (un)binding events take place in the millisecond to minute timescales. This means that it would take on the order of 10^{15} time steps to simulate an event that is 1-2 seconds in real time.

Advancements in computing power and computer architecture have led to improvements in MD simulation speed. Running simulations on supercomputers using graphics processing unit (GPU)s have been able to simulate on the order of tens to hundreds of ns per day [16, 34, 35, 46, 47]. The creation of specialized supercomputers, the Anton series, specifically designed to run MD simulations can simulate on the order of μs per day [48, 49, 50]. In another approach, the Pande lab used idle computing power from personal computers of volunteers on the Folding@home network to simulate a comparable time scale [51]. However, these improvements are still not enough to be able to reach the timescales to simulate biological phenomena of interest.

1.2.2 Enhanced Sampling

MD simulations are able to model (un)binding events, but computational limitations prevent these simulations from reaching the natural timescales to simulate rare events. Many enhanced sampling algorithms have been developed to overcome this limitation and simulate

long timescale events. These can be put into three broad categories: parallel tempering, modified potential energy, and trajectory parallelization enhanced sampling methods. All these methods are capable of simulating long time events.

1.2.2.1 Parallel Tempering Methods

Parallel tempering[52] (also known as temperature replica exchange), aims to improve the breadth of MD sampling by running several copies of the system ("replicas") at different temperatures. After running for a given number of time steps, the energy is calculated for each replica. Temperature swaps between systems are suggested the energy difference between the swap is calculated by:

$$\Delta = (\beta_i - \beta_j)(E_i - E_j), \tag{1.8}$$

where E_i is the energy of the simulation at the conformation associated with temperature i (T_i) and E_j is the energy of the simulation at the conformation associated with temperature j (T_j). β_i is defined as: $\frac{1}{k_b T_i}$, and β_j is defined as $\frac{1}{k_b T_j}$ respectively. The energy is in units of kcal/mol and β has units of mol/kcal. It is worth noting that the relation $T_i < T_j$ is always true when calculating the energy difference, making the first term always positive. If the temperature swap lowers the overall energy ($\Delta < 0$) then the temperature swap occurs. However, if Δ is positive, then we can perform the temperature swap with a probability of $e^{-\Delta}$.

The idea behind parallel tempering is that it is easy to get trapped in local minima in the free energy landscape at lower temperatures. However, by having simulations at elevated temperatures, the simulation can more easily cross large energy barriers, then cool the simulations down to explore this new region at the desired temperature.

In this category of enhanced sampling algorithms, the mechanism for (un)binding does not need to be known before the simulation begins, nor is a reaction coordinate needed. However, by allowing the simulations to heat and cool in such an unnatural way, it is difficult

to generate the kinetics because when the trajectories swap energies, they are no longer continuous[53]. Additionally, there can be poor mixing of hot and cold trajectories resulting in an inaccurate landscape[54]. Finally, there needs to be an adequate number of energy levels to make sure the probability distributions between the trajectories sufficiently overlap to ensure a reasonable acceptance probability for trajectory swaps[55]. This can require several different energy levels to ensure sufficient overlap prohibitively increasing the computational cost.

1.2.2.2 Altered Potential Energy Methods

The altered potential energy category of enhanced sampling is a broad category, including but not limited to metadynamics[56, 57, 58], umbrella sampling[59] and temperature accelerated MD[60, 61], but all the algorithms bias the system along a small set of collective variable (CV) to help cross energy barriers more easily. These algorithms bias the potential energy by the following equation:

$$U_{sim} = U + U_{bias}, \tag{1.9}$$

where U is the unbiased potential energy and U_{bias} is the potential energy that is biasing the simulations. However, the method-specific means of biasing the system is quite varied. For example, in umbrella sampling the CV is divided into a series of independent windows and a harmonic potential is added to the existing Hamiltonian; this helps flatten the energy barriers in the window. MD simulations are then performed in each of the windows. Analysis is performed using the Weighted Histogram Analysis Method (WHAM) [62, 63], which works by combining data from all windows to obtain the original, unbiased free energy profile.

In contrast, in the metadynamics algorithm only one simulation is run over the entire landscape. Each time the simulation visits a location along the CV, a Gaussian, centered at that CV value is added to the Hamiltonian. As the simulation progresses, these Gaussians accumulate and allow the simulation to gradually fill up the free energy basin the simulation is stuck in, allowing it to cross energy barriers. A common analogy for this process is filling

up the free energy landscape with sand. As basins fill with sand, it becomes easier for the simulation to cross over a previously large energy barrier.

Similar to the parallel tempering algorithms, the altered potential methods make exploring the free energy landscape easier. However, there are a couple of drawbacks in using these methods. First is the need for a CV, which requires the prior knowledge about potentially critical variables that lead to the biological phenomenon of interest[64]. These variables are not trivial to determine and can lead to unphysical results if critical variables are not included. Additionally, by adding the bias to the potential energy function, the bias needs to be removed to calculate observables[56]. While there are ways to remove the bias and approximate transition rates, this process assumes that the deposited bias did not affect the dynamics at the transition state [65]. Finally, these simulations assume the system is in equilibrium, which is not always justifiable in biomolecular systems.

1.2.2.3 Trajectory Parallelization Enhanced Sampling Methods

The two categories described above enhance sampling of the free energy landscape by altering the Hamiltonian of the system being simulated. In the case of parallel tempering, the temperature swaps make determining realistic pathways impossible due to the unrealistic energy jumps between simulations. Additionally the altered potential energy methods require prior knowledge to determine good collective variables to sample along. Here we discuss a group of algorithms that do not change the Hamiltonian and thus do not bias the system, produce continuous trajectories, and give us the ability to compute path dependent observables such as kinetic rates.

Algorithms that fall into this category, such as milestoning[66, 67, 68], forward flux sampling[69, 70] and weighted ensemble[71, 72, 73], use sampling over many parallel trajectories to gain a full understanding of the energy landscape. Both the milestoning and forward flux sampling algorithms construct a number of hypersurfaces that are arranged between two basins of interest. By simulating the flux from one hypersurface to another they

get the local kinetics between the surfaces and then combine these local rates to determine the global rate of transition between the basins. However to form these surfaces, prior information is needed about the pathway to know how to construct them. In this thesis work, we have used the weighted ensemble algorithm [71, 73], described in detail in the next section, to simulate the unbinding and binding pathways because it has variants (developed in our research group)[34, 35, 74] that do not require any prior information about how to get from one basin to another.

1.2.3 Weighted Ensemble

The previous section described different algorithms to enhance the breadth of sampling from MD simulations. Here we go into detail about the enhanced sampling algorithm we use for all simulations in this thesis: weighted ensemble (WE). WE is a path sampling algorithm that attempts to focus computational resources on the low-probability states that are relevant to an observable or process of interest. An example of these states is the set of "transition states", conformations corresponding to the highest point on the free energy surface separating two basins, along an unbinding pathway. The original WE approach was outlined in 1996 by Huber and Kim [71] to simulate Brownian motion between a product and reactant basin. However, WE has been used to simulate a variety of processes from protein folding [75], to ligand (un)binding [76, 77, 78, 16, 35, 34], large scale conformational changes[79, 80, 81], ion permeation[72] and protein-protein binding[82, 83].

The WE algorithm has two distinct steps: dynamics and resampling, the combination of these two steps is called a cycle. In the dynamics step, we have a set of simulations (called walkers) that each have an associated statistical weight (w), which sum to 1. The walkers undergo MD independently for a certain amount of simulation time (τ). After each walker has completed the dynamics step, we perform resampling. In resampling we can perform two operations on the walkers called merging and cloning. Merging involves taking two walkers and choosing one to remain and one to kill. The survival probability of each walker

is proportional to its weight. The weight of the surviving walker then increases by the weight of the one that was killed. Cloning involves making exact replicas of a single walker. The weight of the original walker is divided evenly among its clones.

In the original WE algorithm, resampling is performed using bins, which are subdivided regions defined on a progress coordinate. Walkers are then cloned and merged in order to achieve a constant number of walkers in each bin (M). If there are more than M walkers in a bin, then we perform merging between these walkers in order to bring the number of walkers back down. If a bin has fewer than M walkers, then clones are made in the bin until we have the necessary number of walkers. This method increases the computational cost of the simulation when new bins are found. Another issue of using the WE algorithm when applied to simulating biologically relevant events, is that the dimensionality of these processes is typically high, and the number of bins increases exponentially with the number of dimensions[84]. This means that we would need to spend a lot of computational resources to simulate these events.

1.2.4 Resampling of Ensembles by Variation Optimization

WE is able to simulate (un)binding pathways and calculate path dependent observables. However, the algorithm divides the landscape into bins which guide the resampling. Determining a low dimensional set of CVs that sufficiently describes the pathway of interest is not trivial, especially when multiple pathways exist. Additionally, binning in high dimensional spaces becomes exponentially more computationally expensive as more dimensions are required to describe the pathway. Here we discuss a binless weighted ensemble algorithm developed by our research group: Resampling of Ensembles by Variation Optimization (REVO) [35]. In the REVO algorithm, merging and cloning is guided by an objective function called the trajectory variation (V). The variation is defined as:

$$V = \sum_i V_i = \sum_i \sum_j \left(\frac{d_{ij}}{d_0} \right)^\alpha \phi_i \phi_j, \quad (1.10)$$

where V_i is the trajectory variation contribution from walker i , d_{ij} is the distance between walkers i and j , d_0 is a characteristic distance used to normalize the variation when comparing between different distance metrics and keep the distance term unitless, and ϕ_i is the novelty term that describes the importance of individual walkers. REVO balances the exploration (distance) term with the exploitation (novelty) term using α . In this thesis the novelty is defined in terms of walker weights and can be mathematically described as:

$$\phi_i = \log(w_i) - \log\left(\frac{p_{min}}{100}\right), \quad (1.11)$$

where w_i is the weight of walker i and p_{min} is the minimum weight a walker is allowed to be. p_{min} is generally set to $1 * 10^{-12}$.

V is initially calculated and then walkers are proposed to be merged and cloned. The walker that is proposed to be cloned (walker i) is the one that has the highest walker trajectory variation, V_i , and whose cloned weight would be greater than p_{min} . A pair of walkers (walkers j and k) are selected to clone based on minimizing the expected trajectory variation loss (V_{loss}) which is defined as:

$$V_{loss} = \frac{V_k w_j + V_j w_k}{w_j + w_k}. \quad (1.12)$$

For two walkers to be eligible for merging, they need to be within a distance cutoff of each other, called the "merge distance", and the sum of their weights needs to be lower than the maximum allowed weight (p_{max}), set to 0.1 for this thesis. Once these walkers are selected, V is recalculated as though the merging and cloning operations have been performed. If V increases, the merging and cloning operations are performed. This resampling process repeats until V has been maximized. Once V is maximized, a new cycle begins and more MD is performed.

1.2.5 Rate Calculations by Ensemble Splitting

WE is capable of directly calculating observables such as on and off-rates using a technique called ensemble splitting[85, 86, 87, 88, 89]. Using this technique an equilibrium ensemble

is split into two non-equilibrium ensembles and an unbound and bound basin are defined. Starting trajectories from one basin, the rate between the two basins is computed as the flux of trajectories between the two basins.

The on-rate is defined as:

$$k_{on} = \frac{\sum_i w_i^B}{CT}, \quad (1.13)$$

where w_i^B is the weight of a walker that transitioned from the unbound basin into the bound basin, C is the concentration of the ligand that is binding in units M of and T is the aggregate simulation time in μs . The sum is over all walkers that transitioned.

Similarly the off-rate can be calculated by:

$$k_{off} = \frac{\sum_i w_i^U}{T}, \quad (1.14)$$

where w_i^U is the weight of a walker that transitioned from the bound to unbound basin.

To determine the fluxes into each basin, two sets of simulations are run: a *binding simulation* where trajectories are initialized in the unbound state and are terminated in the bound state; and an *unbinding simulation* where trajectories are initialized in the bound state and terminated in the unbound state. To achieve this, once a walker has crossed the termination boundary, it is reinitialized in the initial state by resetting its atomic conformation and velocities. The walker’s weight however, is not affected.

1.2.6 Markov State Models

Although WE does a good job in sampling low-probability states and transitions, the rate of convergence to the equilibrium probability distribution is slow and therefore, it can take a lot of computational time to get adequate sampling of the space. An additional issue is the vast amounts of data the simulation produces, which makes performing a quantitative analysis expensive in terms of computer memory and disk space. Here we describe the formulation of a Markov State Model (MSM), which uses the results from MD simulations to help build a statistically robust and efficient model that can help solve the above issues.

A MSM is a model that describes long timescale dynamics among a set of m macrostates [90]. It is based on the Markovian assumption: that the model is a memoryless description of an ergodic process[90, 91]. This means that the history of a trajectory does not affect its dynamics, the evolution of a trajectory only depends on its current position. An ergodic process implies that given infinite time any macrostate can be reached from any other state. The evolution of a trajectory from one state to another is entirely described by a transition matrix \mathbf{T} .

To generate these macrostates, the results of the MD simulation are clustered together. The issue with clustering the atomic positions directly is that the dimensionality of molecular systems is large: it is equal to $3N$, where N is the number of atoms, typically on the order of 10^4 to 10^5 . This makes the clustering a computationally expensive operation and can overshadow important long timescale dynamics with many sources of noise. Therefore it is common to reduce the dimensions of the system by dimension reduction algorithms such as Principal Component Analysis (PCA) [92, 93] and Time-structure Independent Components Analysis (TICA)[91, 92] in the construction of MSM. Additionally, the dimensions can be reduced to a set of features that describe the pathway of interest[94]. Example features can be atomic distances of a subset of atom pairs or dihedral angles. In this thesis, the MSMs did not use TICA for dimension reduction but we did use sets of distances between key atoms to featurize the space in Chapters 2 and 4, and we used PCA in Chapter 5. After the states have been defined, \mathbf{T} can be generated by counting the transitions between states that occur for a specific lag time (τ).

From an initial probability distribution on the MSM, which we call \mathbf{p}^0 , we can evolve the system by applying \mathbf{T} as follows:

$$\mathbf{p}^0 \mathbf{T} = \mathbf{p}^\tau, \tag{1.15}$$

where \mathbf{p}^τ is the probability distribution after time τ . Additionally we can evolve the simulation by multiple time steps by:

$$\mathbf{p}^0 \mathbf{T}^n = \mathbf{p}^{n\tau} \tag{1.16}$$

where n is the number of lag times for which we are evolving the system. If the system is Markovian and ergodic, then eventually the system no longer evolves when applying \mathbf{T} (i.e. when $\mathbf{p}^{n\tau} \approx \mathbf{p}^{(n+1)\tau}$) and the system reaches its stationary distribution. We define this distribution as π . Mathematically π is an eigenvector of the transition matrix with a corresponding eigenvalue of 1, which is the maximum eigenvalue. If the eigenvalues and eigenvectors of \mathbf{T} are sorted from the largest to the smallest eigenvalues, excluding the eigenvalue equal to 1, the sorted eigenvectors correspond to dynamical motions in the MSM and the larger the eigenvalue, the slower the motion.

From the MSM, we can determine the kinetics of going from one state to another by determining the relaxation time of the model. The errors for determining kinetics from MSMs come from discretization and spectral errors [90]. To minimize these errors one needs to ensure that: i) features being used to cluster the trajectory data need to capture the slow dynamical motions, ii) there are enough macrostates in the model to approximate the dynamics, and iii) to use a large enough lag time. The third requirement is challenging as longer lag times reduce the discretization error[90], but require longer MD simulations to construct the MSM which might not be feasible. To determine an appropriate lag time it is common to satisfy the Chapman-Kolmogorov test [95]. This test compares the relaxation time with respect to the lag time, and is satisfied when the relaxation time is essentially constant with respect to changes in the lag time.

1.3 Outline of Work

The overarching goal of this thesis is to apply molecular dynamics trajectories, guided by weighted ensemble algorithms, to investigate (un)binding pathways for biologically relevant systems. The specific goals are:

- Characterize pathways of host-guest (un)binding reactions.
- Estimate physical quantities such as binding and unbinding rates and binding free energies.

- Develop methodologies to apply experimental results to guide simulations.

We first test the REVO algorithm using small host-guest systems in Chapter 2. The host-guest systems is comprised of two small molecules that form a complex, the larger of two is defined as the host and the smaller is the guest. These systems are used as validation for our methodologies as they are simpler than biological systems. We determine the unbinding and binding pathways from several different initial host-guest conformations. We introduce the use of conformational space networks(CSNs) to visualize these pathways in a graph visualization to help understand the complex pathways which the guest can transition between the bound and unbound basins. From these simulations we calculate the rates and from these the $\Delta G_{binding}$ energies and compare the results to those previously reported.

Chapter 3 will expand on the work in Chapter 2. While the previous work used a correct relationship to describe a macroscopic state in equilibrium, we introduce correction terms to take into account a finite box volume, electrostatic interactions between the host and guest, and the volume of the unbound basin. Additionally we investigate the effect the Langevin dynamics friction coefficient has on the kinetics and $\Delta G_{binding}$ with and without the correction terms.

In Chapter 4, we increase the complexity to a membrane bound protein-ligand system (Translocator protein (18 kDa)-PK-11195) (TSPO). This system has a significantly longer unbinding RT which pushes the limits of the REVO algorithm. Additionally there is an interest in developing specific ligands that target TSPO for possible treatments of neurodegenerative diseases. In this chapter we simulate the unbinding process for PK-11195 from 5 different starting poses and quantify two distinct unbinding pathways for PK-11195 dissociation. We determine key residues that have strong interactions with PK-11195 along the unbinding pathways. We computed unbinding rates similarly to Chapter 2 for each pose however, we also constructed an MSM to verify these results. Finally we used committor probabilities alongside the MSM to determine the transition state for unbinding.

In Chapter 5, we investigate the binding process for a ternary complex comprised of a ligase, a proteolysis-targeting chimera (PROTAC), and a target protein. This is a more complicated process than previously explored in this thesis as we are trying to form a ternary complex involving two proteins and a small linker molecule. Since REVO is naturally designed to maximize the exploration of the landscape, we alter the resampling algorithm by only cloning walkers that have sufficiently progressed toward the target state. Additionally, we develop new distance metrics that take hydrogen deuterium exchange (HDX) experimental data to help drive the simulation to the bound state.

In Chapter 6 we take a high level look at the goals of this thesis and discuss the progress and describe next steps for improvement.

CHAPTER 2

PREDICTING LIGAND BINDING AFFINITY FOR THE SAMPL6 CHALLENGE FROM ON- AND OFF-RATES USING WEIGHTED ENSEMBLES OF TRAJECTORIES

This work was published in Journal of Computer-Aided Molecular Design volume 13, pages 1001-1012 in 2018. The work is presented here as published except that the supplemental figures are worked into the text.

2.1 Introduction

Binding affinity has long been seen as the crucial parameter for drug discovery, as it determines the proportion of drug that is bound to a receptor in solution. A wide variety of methods have emerged to predict both absolute and relative binding affinities, each with its own domain of applicability, and tradeoff between efficiency and accuracy [96, 97]. The SAMPL challenge is playing an important role to compare tools that predict affinities using blind predictions [98]. Importantly, errors can arise from both the physical model used to describe the system (e.g. forcefield, thermostat, dynamics engine), and from the sampling methodology used. The SAMPLing challenge, described in this issue, thus serves an important role in comparing the accuracy of computational methods that all employ the same physical model [99].

While the binding affinity is all that is needed to describe the action of a ligand at equilibrium, the on (k_{on}) and off-rates (k_{off}) are necessary to model drug action in general [2]. For instance, in many systems it has been observed that drug residence time (RT) ($RT = 1/k_{\text{off}}$) is the critical factor governing efficacy in living cells [18, 20, 19]. This is due to the number of factors that drive the system out of equilibrium, such as drug metabolism and elimination, the turnover of target protein, and the periodic nature of drug administration. Although $K_D = k_{\text{off}}/k_{\text{on}}$, and lower K_D can be correlated to lower k_{off} , this relationship is governed by the free energy along the ligand binding pathway, particularly the ligand binding

transition state, which is the highest point in free energy between the bound and unbound states [100]. Though the binding rate has an upper bound of $10^9 \text{ M}^{-1} \text{ s}^{-1}$, which corresponds to the “diffusion limit”, binding rates of ligands to the same target have been shown to vary over 4 orders of magnitude, which disrupts the correlation between K_D and k_{off} [101].

Prediction of k_{off} and k_{on} is challenging, as they are not state functions: they depend fundamentally on the transition path ensemble between the bound and unbound states. Computational sampling of these transition paths is in general a great challenge for molecular dynamics (MD) due to the long timescales of ligand binding and release, although in recent years, a variety of enhanced sampling methods have rose up to meet this challenge [102]. The trypsin-benzamidine system has served as a common benchmark application for enhanced sampling methods such as Adaptive Multilevel Splitting [103], SEEKR [104], adaptive [105] and traditional [106, 107] Markov state modeling, funnel metadynamics [108], as well as the WExplore method developed by our group [109]. Recently these efforts have been expanded to more challenging systems such as the unbinding of inhibitors from c-Src kinase [110] and p38 MAP kinase [56] using metadynamics, and the unbinding of the TPPU ligand from the target soluble epoxide hydrolase with WExplore [16]. The diversity of computational approaches to handle long timescale ligand binding and release events is a promising sign for the field, but comparison of methodologies is complicated – even for applications to the same system – due to differences in forcefields, boundary conditions, and integrators.

As a step toward the robust comparison of different computational methods for simulation of binding pathways, we participated in the SAMPLing challenge for the prediction of binding affinities. The SAMPLing challenge required participants to compute free energies as a function of simulation time, to compare the convergence properties and relative computational cost of different free energy calculation methods. Instead of computing free energies through alchemical perturbation, here we achieve this by explicitly simulating the binding and release processes, determining the absolute rates k_{on} and k_{off} , and computing

the binding affinity as the ratio $k_{\text{off}}/k_{\text{on}}$. We calculate the binding free energy as follows:

$$\Delta G = kT \ln \left(\frac{k_{\text{off}}}{C_0 k_{\text{on}}} \right) \quad (2.1)$$

where $kT = 0.597$ kcal/mol corresponding to a temperature of 300 K and C_0 is the reference concentration of 1 mol/L. As we broadly sample unbinding pathways from multiple starting points, we can also synthesize these results and examine how these poses are connected in the binding network.

We efficiently determine unbinding and binding rates using a further developed variant of the WExplore sampling method [84]. This is the first application of this new method, which we call Reweighting of Ensembles by Variance Optimization (REVO). This new method is also based in the weighted ensemble framework [71], where trajectories are merged and cloned, but it is the first to completely eschew the idea of dividing a space into a set of sampling regions (the possibility has previously been recognized however [73]). REVO instead directs merging and cloning operations by maximizing a measure of variance that describes the instantaneous spread of the ensemble of trajectories, which is described in the Methods section below. We visualize our REVO simulations using a branching tree network diagram, whose layout uses an energy function that takes into account the distances between the trajectories. This allows for the easy visualization of the correlation of exit point ensembles within a weighted ensemble simulation. We compare our binding affinities to computational reference values, and observe that the affinities from REVO are systematically tighter than the reference. We conclude the manuscript with a discussion of possible sources of error.

2.2 Methods

2.2.1 Host-guest systems

The host-guest systems were selected from the main SAMPL6 challenge. One system is a cucurbit[8]uril (CB8) host [111, 112], using quinine as a guest ligand (Figure 2.1). The host is a ring-shaped structure, with 8-fold rotational symmetry about the vertical axis,

and two-fold symmetry about the horizontal axis. There are thus 16 symmetry-equivalent atom mappings for this system. The second and third systems both use a Gibb deep cavity cavitand, referred to as OA, as a host [113]. Here there is only 4-fold symmetry about the vertical axis. Binding and release of two ligands is examined: 5-hexenoic acid and 4-methyl pentanoic acid, referred to as OA-G3 and OA-G6, respectively. Both of these ligands carry an explicit negative charge.

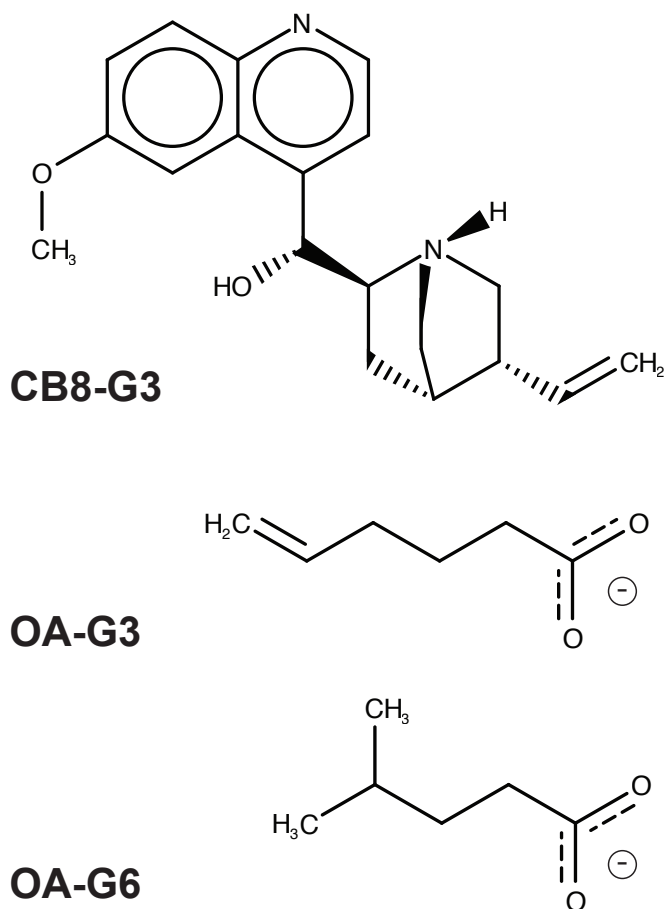


Figure 2.1: Structure of the ligands used in this study. (Top) Quinine, referred to herein as CB8-G3. (Middle) 5-hexenoic acid (deprotonated form), referred to herein as OA-G3. (Bottom) 4-methyl pentanoic acid (deprotonated form), referred to here ohhh in as OA-G6.

2.2.2 Dynamics Setup

The fifteen initial configurations (five for each host-guest system) were used as prepared by the organizers of the SAMPLing challenge without modification (Figure 2.2-2.4). The two OA systems had a cubic box with a box length of 45 Å solvated with 2586 water molecules, and contained 12 sodium ions and 3 chloride ions to neutralize the system. The CB8 system had a cubic box with a box length of 42.5 Å solvated with 2149 water molecules, and contained 6 sodium ions and 6 chloride ions to neutralize the system. OpenMM v7.1.1 [114] was used to run dynamics on the CUDA v8.0 platform. We use a Langevin integrator, with a thermostat at 300 K, a friction coefficient of 1.0 ps^{-1} , a Monte Carlo barostat to keep pressure constant at 1 atm, and a time step of 2 fs. The non-bonded forces had a cutoff of 1 nm, and were calculated using partial mesh Ewald. The simulation temperature differs slightly from that used to calculate the reference free energies (298.15 K), although we expect the resulting differences in free energy will be negligible.

CB8-G3

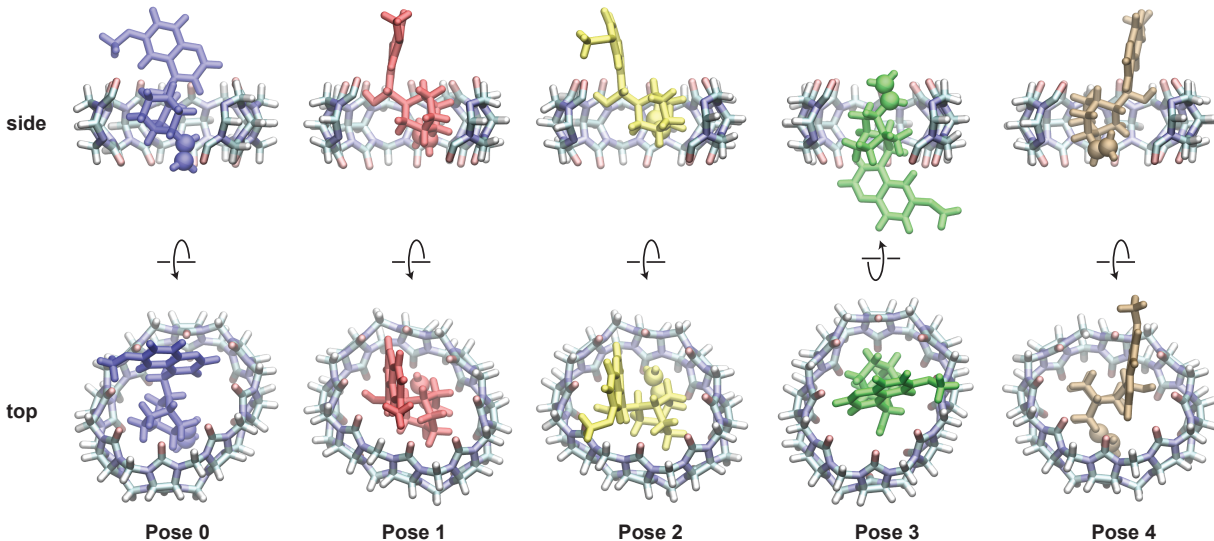


Figure 2.2: Starting poses for CB8-G3. Side and top views are shown. Coloring for pose indices is consistent with Figures 2.7, 2.8 and 2.9.

OA-G3

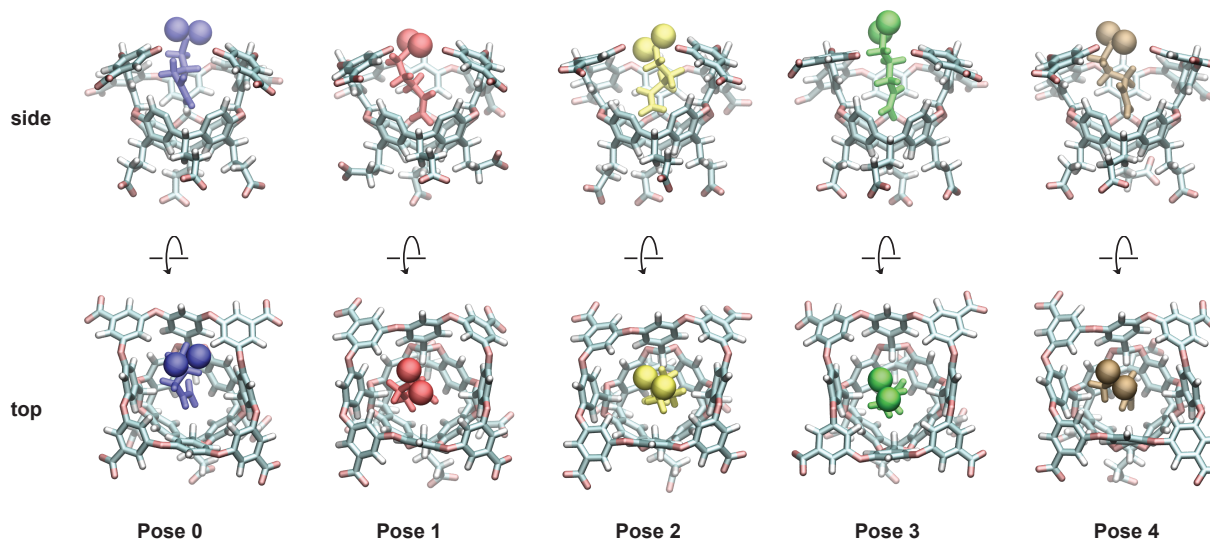


Figure 2.3: Starting poses for OA-G3. Side and top views are shown. Coloring for pose indices is consistent with Figures 2.7, 2.8 and 2.9.

OA-G6

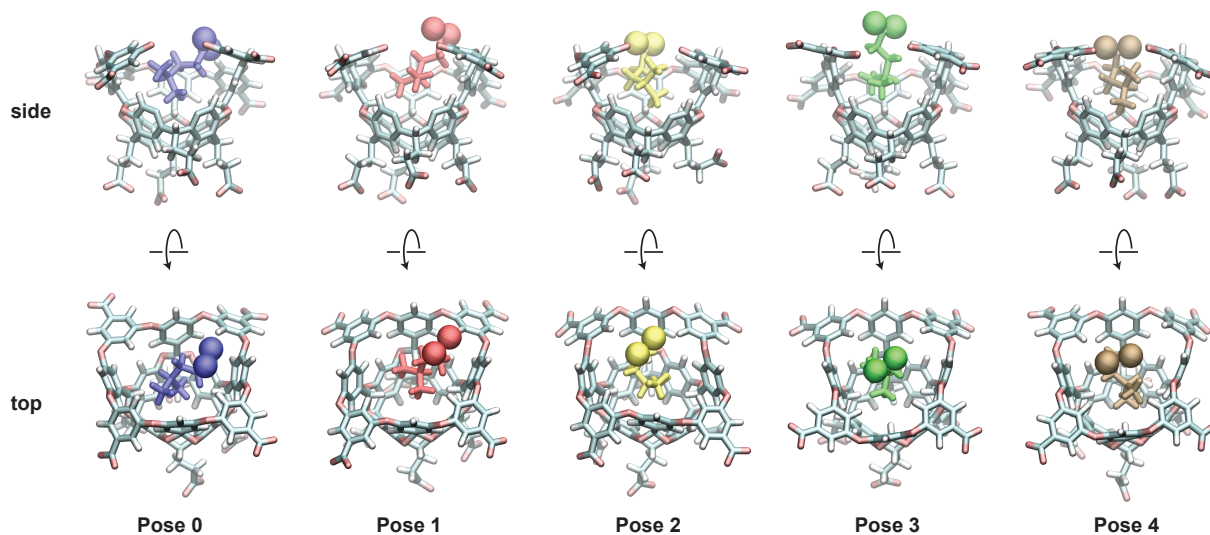


Figure 2.4: Starting poses for OA-G6. Side and top views are shown. Coloring for pose indices is consistent with Figures 2.7, 2.8 and 2.9.

2.2.3 Reweighting of Ensembles by Variance Optimization

To encourage the sampling of rare events, we developed a method based on the weighted ensemble (WE) framework [71] that we call “Reweighting of Ensembles by Variance Optimization”, or REVO. WE methods use an ensemble of trajectories (called “walkers”) that are each assigned a statistical weight, and enhance sampling through the introduction of cloning and merging steps. Initially the weights of all the walkers are equal, and are defined as $1/N_{\text{walk}}$, where N_{walk} is the total number of walkers. When walkers are *cloned*, their weight is divided among the progeny. The cloned trajectories are identical replicas of the original, with the same atomic positions and velocities. This is typically done in under-sampled regions of space, in order to boost the probability of observing rare events in the simulation. Walkers are also *merged* together, and their summed weight is given to the resulting merged walker. In practice, merging walkers A and B is accomplished by choosing a survivor (walker A is chosen with probability $\frac{w_A}{w_A+w_B}$), and discarding the other walker. Merging is typically done in over-sampled regions, with walkers that can be seen as “redundant”. The trajectory weights are only changed due to merging and cloning operations.

Previous applications of the weighted ensemble methods proceed by constructing a set of sampling regions, determining their occupancies, and using cloning and merging operations to make the occupancies as even as possible. In general, the free energy landscapes of interest are inherently high-dimensional, which makes it difficult to construct an appropriate set of regions. For this reason we were motivated to discard the notion of “regions” entirely, and direct cloning and merging operations instead by the optimization of a variance measure, V :

$$V = \sum_i V_i = \sum_i \sum_j \left(\frac{d_{ij}}{d_0} \right)^\alpha \phi_i \phi_j, \quad (2.2)$$

where the double sum is over all pairs of walkers, d_{ij} is some measure of distance between walkers i and j , d_0 is the characteristic distance, the exponent α is a parameter set here to 4, and ϕ_a is a weighting function for walker a :

$$\phi_a = \log \left(\frac{100w_a}{p_{\min}} \right), \quad (2.3)$$

where w_a is the weight of trajectory a , and p_{\min} is the lowest probability attainable by a walker, set here to be 10^{-12} . The weighting function ϕ was designed to be largest for high w_a , and to smoothly decay to a low value as w_a approaches p_{\min} .

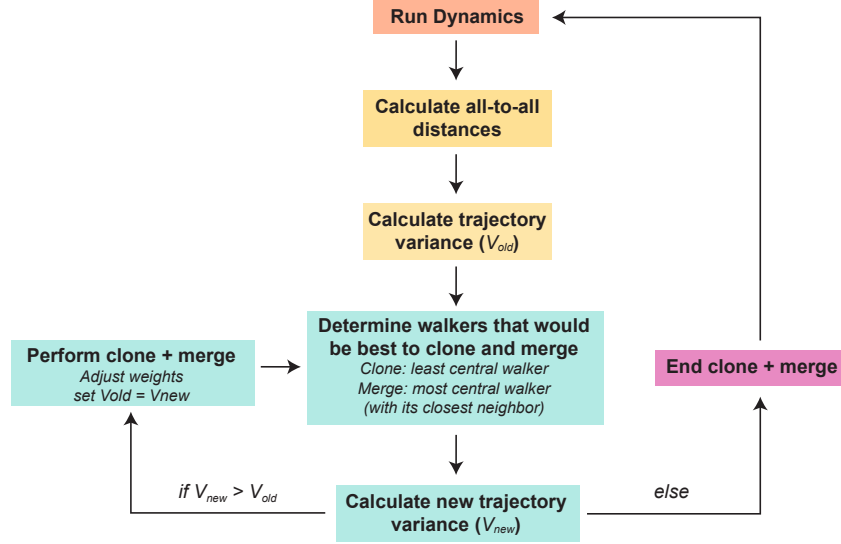


Figure 2.5: The REVO algorithm. Each cycle begins by running an ensemble of walkers forward in time using unbiased dynamics. The distances between the walkers are used to calculate a variance (Eq. 2.2). In the resampling loop (blue), coupled cloning and merging operations are proposed, and they are accepted only if they result in a higher V . If the proposed V is lower, the resampling loop is terminated and dynamics are continued for the next cycle.

The structure of the REVO “resampling” algorithm proceeds as follows (see also Fig. 2.5). Eq. 2.2 is used to compute the variance function, and the walker with the highest (“H”) and lowest (“L”) contributions to the variance are identified (e.g. with the highest and lowest V_i values). The closest walker to “L” is identified, called “C”. A coupled cloning and merging event is proposed, where “C” and “L” would be merged and “H” would be cloned. Eq. 2.2 is again used to recompute the variance, and this coupled cloning and merging move is only accepted if V increases. Further moves are proposed after recomputing “H”, “L” and “C”, and the process continues until V decreases, and the move is rejected. This way, the algorithm automatically determines the optimal number of cloning and merging events. In fact, if the

system is already in an optimal configuration, no cloning and merging operations will take place, and REVO will skip to the next dynamics step.

As in previous WExplore applications [16], a minimum and maximum walker weight was enforced (p_{\min} and p_{\max} , respectively). Note that only walkers which will not violate the walker probability boundaries (p_{\min} and p_{\max}) are eligible to be chosen as walkers “H”, “L” and “C”. In these simulations, $p_{\min} = 10^{-12}$ and $p_{\max} = 10^{-1}$, following previous work[16].

This process is general to any dynamics engine, and to any form of the distance function d_{ij} . Here we use two different distance functions to describe the unbinding and rebinding processes. For unbinding, d_{ij}^U is defined as the root mean square deviation (root mean square deviation (RMSD)), in Å, of the guest ligand between structures i and j , after aligning to the host. As mentioned in Section 2.2.1, there are multiple symmetry-equivalent mappings of the host atoms. We thus compute this distance after alignment of j to each symmetry-equivalent mapping of host i , and use the smallest such value as d_{ij}^U . For rebinding, d_{ij}^R is computed using the RMSD of both i and j to the reference starting structure:

$$d_{ij}^R = |1/d_{i0}^U - 1/d_{j0}^U|, \quad (2.4)$$

where d_{a0}^U is the distance from walker a to the reference structure. The difference between the inverse of these two quantities is used to highlight differences between small values of this quantity (e.g. between RMSD = 1.5 Å and RMSD = 2.0 Å).

2.2.4 Calculating rates by ensemble splitting

REVO, like other weighted ensemble methods, can calculate kinetic quantities on the fly, through a technique we call “ensemble splitting” [85, 115] (also referred to as “tilting” [87], or “coloring” [88, 116]). An equilibrium ensemble is split into two non-equilibrium ensembles by defining two basins, in this case the “bound” basin and the “unbound” basin (Fig. 2.6). The unbinding ensemble is defined as the set of trajectories that have most recently visited the bound basin, and the rebinding ensemble is the set of trajectories that have most recently

visited the unbound basin. The unbound basin is the set of structures where the closest host-guest interatomic distance exceeds 10 Å, as in previous work[16]. The bound basin is defined as the set of structures with guest RMSD < 1.0 Å, computed after aligning to the host. Note that a sweep over symmetry-equivalent atom mappings of the host was again conducted, so a binding event can be registered by binding to either the top or bottom of the CB8 host, for example.

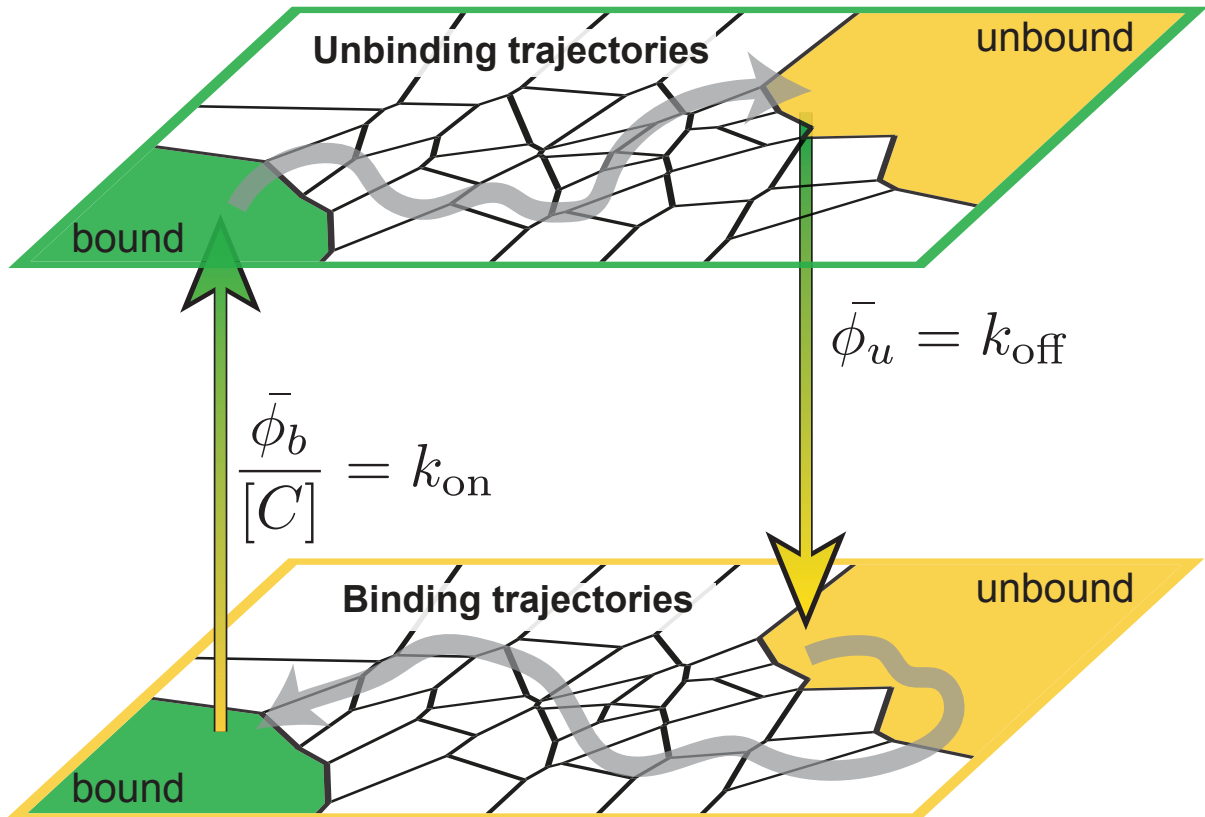


Figure 2.6: Ensemble splitting. An equilibrium host-guest binding system is split into two non-equilibrium ensembles for the calculation of on and off-rates. This is done by defining “bound” and “unbound” basins (left and right of each ensemble). The “unbinding” ensemble (top) is the set of trajectories that have most recently visited the bound basin. The “binding” ensemble (bottom) is the set of trajectories that most recently visited the unbound basin. The on and off-rates are directly computed using the time averaged trajectory flux ($\bar{\phi}_b$ or $\bar{\phi}_u$) between the ensembles.

In this work, REVO simulations are conducted explicitly either in the unbinding en-

semble, or the rebinding ensemble. After each dynamics step, any walker that has exited its ensemble (by entering the opposite basin) is identified. Its weight is recorded, and its structure is “warped” back to the starting structure. The structure recorded before warping is known as an exit point. When a walker “warps”, the atomic coordinates and velocities of the trajectory return to the starting structure. The weight of a walker does not change as a result of warping. In the unbinding ensemble, the starting structure is the initial bound pose. In the rebinding ensemble, the starting structures are exit points that were generated by the unbinding simulations.

As shown in Figure 2.6, the rates are simply calculated using the flux of trajectories (sometimes referred to as the Hill relation [117, 88]) that leave the ensemble:

$$k_{\text{off}} = \bar{\phi}_u = \frac{\sum_i w_i^U}{T}, \quad (2.5)$$

$$k_{\text{on}} = \frac{\bar{\phi}_b}{C} = \frac{\sum_i w_i^R}{CT}, \quad (2.6)$$

where $\bar{\phi}_u$ and $\bar{\phi}_b$ are the unbinding and binding flux, T is the elapsed time, the sums are over the set of exit points observed before time T , and C is the concentration of the ligand, computed as $1/V$ where V is the box volume.

2.2.5 REVO simulation details

Unbinding REVO simulations were run for 2000 cycles, with 48 walkers run for $\Delta t = 20$ ps each cycle. The exit points registered after 1000 cycles were used to initialize the rebinding REVO simulations. In some cases, fewer than 48 exit points were obtained at this point, and the walkers were randomly cloned in order to create a full set of 48 walkers. The rebinding REVO simulations were run for 200 cycles, with $\Delta t = 200$ ps per cycle. Five simulations were run for each ligand, one from each starting pose (see Figures 2.2-2.4 for starting poses). In aggregate, we ran $1.92 \mu\text{s}$ for each of the unbinding and rebinding simulations, $3.84 \mu\text{s}$ for each starting pose, or $57.6 \mu\text{s}$ over the entire set of results presented here.

2.2.5.1 Note about CB8-G3-0 and CB8-G3-4

After the conclusion of the SAMPLing challenge we found an error in the weight normalization procedure that was used to initialize the weights of the rebinding walkers when fewer than 48 exit points were observed. This affected only two simulations: CB8-G3-0 and CB8-G3-4, where only 5 and 7 exit points were observed, respectively, in the first 1000 cycles of the unbinding simulation. Due to an error, the initial weights in these rebinding simulations summed to a value greater than 1, and while this could be accounted for in the rate calculations, it was compounded by the fact that no walker in these simulations had a weight value less than $p_{\max} = 0.1$, and thus no cloning/merging moves could occur.

Surprisingly, this did not affect the calculation of the binding rate. Although the number of binding events observed in CB8-G3-0 and CB8-G3-4 (32 and 25, respectively), was much lower than the number observed in CB8-G3-1, CB8-G3-2 and CB8-G3-3 (289, 427 and 190), the total amount of weight that exited was comparable (0.62, 0.43, 0.66, 0.14, 0.50, for starting poses 0 through 4). This goes to show the downhill nature of binding in host-guest systems, as confirmed by the almost diffusion-limited k_{on} (see Table 3.1. The calculated mean first passage time (MFPT) of binding for the CB8-G3 system was 91 ns, which is well within the aggregate sampling time of each rebinding simulation (1.92 μs), again indicating why a group of straight-forward trajectories was able to produce over two dozen binding events each.

2.2.6 Visualization of trajectory trees

To visualize the correlation between exit points, we visualize REVO cloning events in a tree graph, where each node represents a walker at a given time point and the edges indicate how walkers are connected through time as can be seen in Figure 2.10. Each level (y -position) on the tree represents walkers at the same time step. The initial horizontal placement (x -position) of each node is a direct result of its parent’s position in the previous time step. If no cloning events occurred for that walker, then the node is placed directly above its parent.

If the parent was cloned, then the walkers are spread out in a fan pattern. Once the nodes are initially placed, their x -positions are minimized with a steepest descent algorithm using the following energy function:

$$E = \sum_i \left[b(x_i^t - x_i^{t-1})^2 + cw_i(x_i^t)^2 + \sum_j E_{ij}^{rep} e^{\frac{(x_i^t - x_j^t)^2}{r_o}} \right], \quad (2.7)$$

where x_i^t and x_j^t are the positions on the tree of walkers i and j at time t , x_i^{t-1} is the position of the parent at the previous time step, and w_i is the walker weight obtained from the simulation. The variables b , c , r_0 are parameters set here to 0.01, 5 and 1000 respectively. The first term causes the nodes to stay close to their parent’s position, allowing trajectories to be visually tracked through the tree more easily. The second term encourages the higher weight trajectories to stay close to $x = 0$. The third term is a pairwise repulsion term, which gives the nodes a “radius” of r_0 , and is scaled by a repulsion energy (E_{ij}^{rep}) that takes into account the molecular distance between the walkers in the simulation (d_{ij}):

$$E_{ij}^{rep} = a * \max(0, d_{ij} - d_0), \quad (2.8)$$

where a and d_0 are parameters set here to 2.5 and 2.0. d_{ij} can refer to either d_{ij}^U or d_{ij}^R when making trees of unbinding and rebinding simulations respectively. However, only trees generated from unbinding simulations are shown here. The parameters for Eq. 2.7 were selected to keep the branches generated by cloning events in the same region on the tree, as well as to keep larger weighted walkers towards the center. It is important to note that this energy minimization only affects the x -position of each node. The y -position is determined by the time step and is not used in the steepest descent algorithm. The graphs were made using NetworkX 2.2 library [118] and visualized using Gephi 0.9.2 [119].

2.2.7 Clustering and visualization of conformation space networks

All of the trajectory frames for the five starting poses of each system were clustered together using the MSMBuilder 3.8.0 library [120]. The datasets were first featurized using a vector of

host-guest distances for each system. These vectors contain 7056, 3128, and 3496 distances for the CB8-G3, OA-G3 and OA-G6 systems, respectively. A k-centers clustering algorithm was used to generate 1000 clusters using the featurized space and assign each frame of the trajectories to a cluster. The clustering was done using the Canberra distance:

$$D(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^n \frac{|p_i - q_i|}{|p_i| + |q_i|} \quad (2.9)$$

where \mathbf{p} and \mathbf{q} are host-guest distance vectors from different frames and n is the total number of distance pairs. A count matrix describing the cluster-cluster transitions was calculated. This corresponds to a Markov state model with a lag-time equal to the cycle length $\Delta t = 20$ ps.

We then construct Conformation Space Networks (CSNs) from the count matrices, which are graphical models of the transition matrix, with a node representing each row, and edges representing non-zero off-diagonal elements using CSNAnalysis [121]. Gephi 0.9.2 was used to visualize the CSN. The size of each node is proportional to the statistical population of the cluster. The smallest node was 20 times smaller than the largest node. The topology of the network was determined using a force minimization algorithm, Force Atlas, included in Gephi [122]. This algorithm includes repulsive forces for nodes that are not connected and attractive forces proportional to the weight of the edges. The directed edge weights were values between 0.1 and 100 as determined by $w_{ij} = 100p_{ij}$ where p_{ij} is the transition probability of cluster i transitioning to cluster j . Undirected edge weights were then determined as the average between the two directed edge weights. Force Atlas was applied twice, first without adjusting for node sizes which enabled the nodes to overlap, and then a second minimization adjusted for node size which prevented overlap. For visualization, all edge weights were given a uniform value. A CSN of each system is shown in Figure 2.11.

2.3 Results

2.3.1 Warped walkers

For each host-guest system we run both unbinding and rebinding REVO simulations originating from five different starting poses (Figure S1-S3), making 30 simulations total. All of these REVO simulations generated a substantial number of warping events. In general these are distributed across a wide range of weight values (Figure 2.7). For all systems it is observed that rebinding can occur with very high probability walkers ($p > 0.1$), but that unbinding occurs with much lower probability. Indeed it is the probability of the unbinding warped walkers that largely governs differences in K_D and k_{off} between the systems. The minimum weight that is achievable by a walker, p_{min} , was set to 10^{-12} in all cases. As shown in Figure 2.7, this could be increased substantially (e.g. to 10^{-3}) in the rebinding case to avoid the propagation of low-weight trajectories that will not meaningfully contribute to the binding flux.

The warping points for the unbinding simulation are shown in Figure 2.8, again using color to indicate the starting pose. Although they exhibit some strong correlation within a REVO run, together they comprise a broad distribution. For CB8-G3, both upward and downward exit pathways are sampled with roughly equal frequency, whereas for the Octa-Acid systems, the exit points are clustered towards the top of the cavitand.

2.3.2 Kinetics and free energies

The binding and unbinding rates are calculated using the sum of the weights of the warped walkers, divided by the elapsed time (see “Calculating rates by ensemble splitting” in Methods). The binding rate is calculated by dividing the binding trajectory flux by the concentration of the guest in mol/L, calculated as $C = \frac{1}{N_A V}$, where V is the box volume. The concentration ranged from 0.021 M for OA-G6 to 0.025 M for CB8-G3 and OA-G3, resulting from unit cells with side-length ranging from 4.1 nm to 4.3 nm. Running estimates of k_{on}

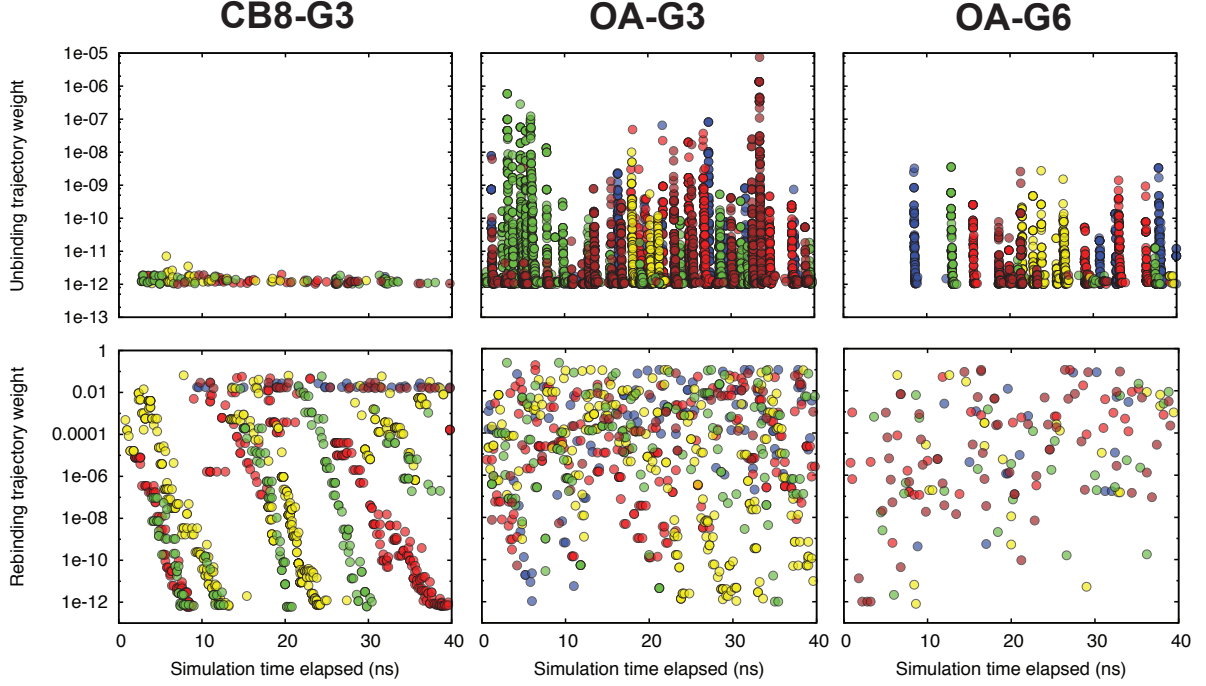


Figure 2.7: Weights of warped walkers. Weights of warping events for the unbinding (top row) and rebinding (bottom row) simulations. In both cases the points are colored according to the index of the corresponding starting pose (0, blue; 1, red; 2, yellow; 3, green; 4, brown).

Table 2.1: Pose-averaged rates and affinities

	$k_{\text{off}} \text{ (s}^{-1}\text{)}$	MFPT _{off} (s)	$k_{\text{on}} \text{ (s}^{-1} \text{ M}^{-1}\text{)}$	MFPT _{on} (ns)	$\Delta G \text{ (calc.)}$	$\Delta G \text{ (ref.) [99]}$	$\Delta G \text{ (exp.) [123]}$
CB8-G3	0.0012 ± 0.0003	860 ± 230	$4.7 \pm 0.8 \times 10^8$	92 ± 16	-16.0 ± 0.2	-10.90 ± 0.16	-6.45 ± 0.06
OA-G3	160 ± 110	0.0064 ± 0.0044	$1.2 \pm 0.2 \times 10^9$	36 ± 6	-9.5 ± 0.4	-6.70 ± 0.02	-5.18 ± 0.02
OA-G6	0.48 ± 0.11	2.1 ± 0.5	$2.8 \pm 1.0 \times 10^8$	150 ± 50	-12.1 ± 0.3	-7.18 ± 0.05	-4.97 ± 0.02

and k_{off} are shown individually for each REVO simulation in Figure 2.9, along with their average, which is calculated by averaging the trajectory flux over the set of five simulations. Large, upward jumps are observed in the rate curves whenever an exit point is recorded that has a higher weight than was previously observed.

The final average rate values, as well as the corresponding mean first passage times, are given in Table 2.1. The uncertainties of k_{on} and k_{off} (δ_{on} and δ_{off}) are determined using the standard error across the five trajectories. The uncertainties in the MFPT of binding and unbinding are calculated as $\delta_{\text{on}}/Ck_{\text{on}}^2$ and $\delta_{\text{off}}/k_{\text{off}}^2$, respectively. Finally, the uncertainty in

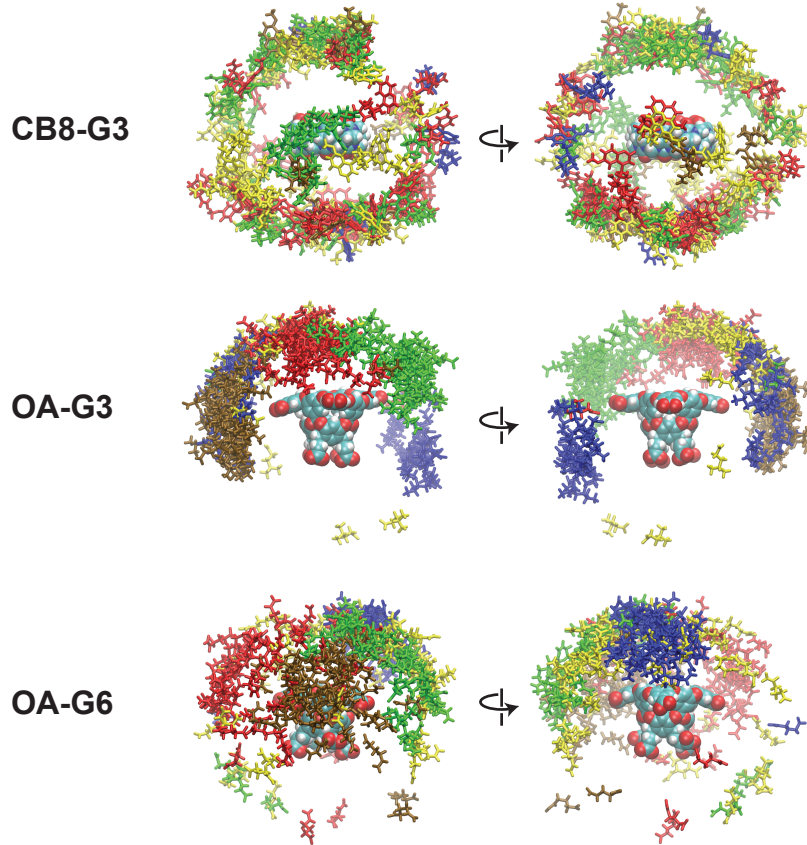


Figure 2.8: Spatial distribution of warped walkers. Structures of warping events for the unbinding simulations viewed from the front and back. Guest ligands are colored according to the index of the corresponding starting pose (0, blue; 1, red; 2, yellow; 3, green; 4, brown).

ΔG is as follows:

$$\delta_{\Delta G} = kT \sqrt{\left(\frac{\delta_{\text{off}}}{k_{\text{off}}}\right)^2 + \left(\frac{\delta_{\text{on}}}{k_{\text{on}}}\right)^2}. \quad (2.10)$$

The MFPT of unbinding demonstrate the power and scope of the REVO method: we estimate that the CB8-G3 system has an average ligand RT of 860 seconds, and we obtain multiple ligand release events for each of the five starting poses. In total, we used $9.6 \mu\text{s}$ of sampling in the CB8-G3 unbinding ensemble, resulting in an acceleration factor of $\approx 9 \times 10^7$.

With k_{off} and k_{on} in hand, the binding affinity is calculated using Eq. 2.1. This binding affinity is compared to both the experimentally measured binding affinity [123], and

a computational reference computed using alchemical free energy calculations with YANK (see [99] for more details). As shown in Figure 2.9, the host-guest affinity calculated by the rate ratio in REVO is systematically too tight when compared both the experimental and reference values. This is possibly due to finite box size effects, which is discussed further in the Discussion and Conclusions section.

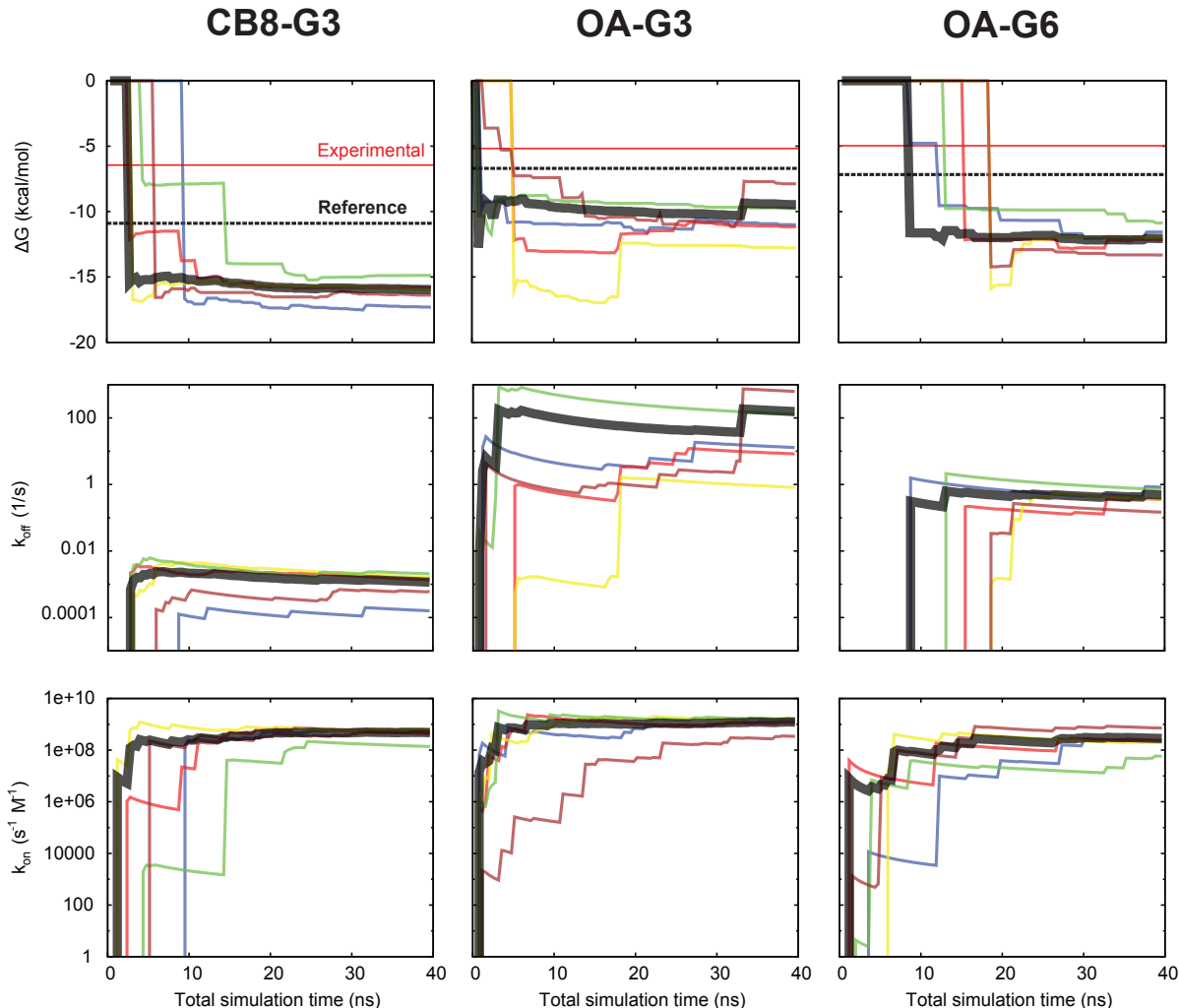


Figure 2.9: Predicted kinetics and free energies. The calculated free energies (top), off-rates (middle), and on-rates (bottom) are shown as a function of simulation time for each starting pose in each host-guest system. The curves are colored according to the index of the starting pose as in Figures 2.7 and 2.8. The calculated binding free energies are compared with experimental measurements (horizontal red line) [123], and the computational reference (dashed black line) for each system.

Moderate variation in k_{on} and k_{off} is observed across the sets of simulations for each host-guest system, which contributes to some uncertainty in the predicted rates and affinities. However, the average standard deviation in the log10 final rates ($\log 10(k)$) is 0.28 for on-rates and 0.56 for off-rates, both well under an order of magnitude. This compares very favorably with recent studies using WExplore [102, 16], where rates from individual simulations varied over several orders of magnitude.

2.3.3 Trajectory trees reveal correlation between exit points

Rates are derived from exit points, and while points from different starting poses are guaranteed to be independent, it is unclear how correlated the observations are within a given REVO simulation. We can use a tree network to observe the entire set of merging and cloning events that occur during a simulation, and to determine how closely related walkers are to one another. Additionally, one can visualize the state of the walkers through coloring the tree based on physical properties observed during the simulation, such as the solvent accessible surface area (SASA) of the guest molecule, which can help evaluate how close the guest is to unbinding from or rebinding to the host. Using this coloring, and how closely related walkers are to one another, we can visualize the correlation between a set of unbinding or rebinding events.

Figure 2.11 shows a trajectory tree for the OA-G3-0 unbinding simulation. From the tree it is clear that the majority of sampling time is spent sampling the bound state (dark green structures). However, the top inset shows that this sampling is still very active, with outliers being detected and cloned nearly every cycle, although the vast majority of these clones are merged one or two cycles later, which implies that the outlying property corresponded to a fast degree of freedom. The middle inset shows a breakout event that led to a series of exit points. The vertical “branches” show individual trajectories. Termination of a branch with high SASA (orange) correspond to exit points.

The OA-G3-0 simulation generated 966 exit points, 534 of which can be seen in Figure

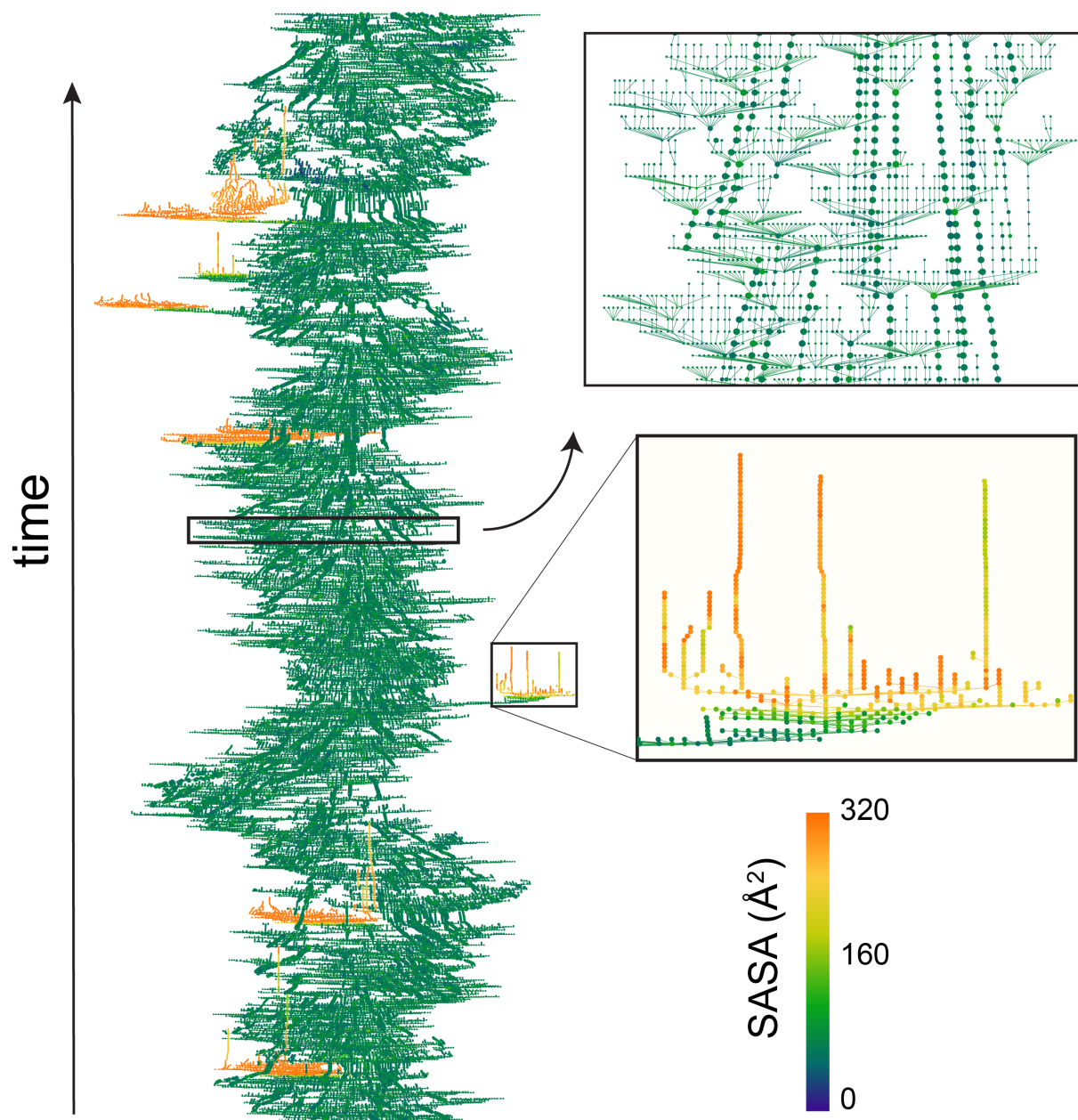


Figure 2.10: Trajectory trees show all cloning and merging events in a simulation. The trajectory tree for the first 1329 cycles of the OA-G3-0 unbinding simulation is shown. Each horizontal row in this tree represents a cycle, and the placement of all 48 nodes in the row is determined by minimizing an energy function (see “Visualization of trajectory trees” in Methods). SASA is used to color the nodes, with blue and dark green indicating bound structures, and yellow to orange indicating unbound.

2.10, which captures only the first 1329 cycles. From the tree it can be seen that many of these exit points are correlated, as they were recently cloned from common ancestors. Using the tree analysis one can observe that there are likely at least seven distinct groups of exit points that can be treated as independent observations of unbinding pathways.

In the bottom inset we see a trajectory that demonstrates transient rebinding behavior. That is, the SASA goes high (≈ 320 , orange), to medium (≈ 160 , light green), back to high again. This behavior results from a transient, loose association with the exterior of the host molecule.

2.3.4 Conformation space networks reveal connection between starting poses

Here we obtain combined estimates of k_{on} , k_{off} and K_D by averaging the transition flux from simulations with different starting poses, and in the case of the rebinding simulations, different boundary conditions. This is only appropriate if the five starting poses are all part of the same basin of attraction, and can interconvert on timescales much faster than the unbinding process. If two poses form distinct basins of attraction, then we cannot expect that the poses will have the same k_{off} , k_{on} , or K_D . To examine the connectivity of starting poses, we use the REVO trajectory segments to construct a Markov state model. We then visualize CSNs to examine how the starting poses are connected, whether they are in the same basin of attraction, and whether they share the same (un)binding pathways.

Figure 2.11 shows CSNs for the unbinding simulations of all three host-guest systems. For both OA systems a large, densely-connected ensemble of bound states is observed. As the entire set of host-guest distances was used to featurize our dataset, this heterogeneity arises from motions of the flexible chemical groups on the bottom and around the rim of the OA host molecule. Starting structure 3 in CB8-G3 is bound in the opposite orientation from the others (see Figure 2.2), although the host-molecule is symmetric to inversion about the horizontal plane. While this did not affect the kinetic measurements (which took into account symmetry-equivalent atom mappings of the host), in the CSN it forms a distinct

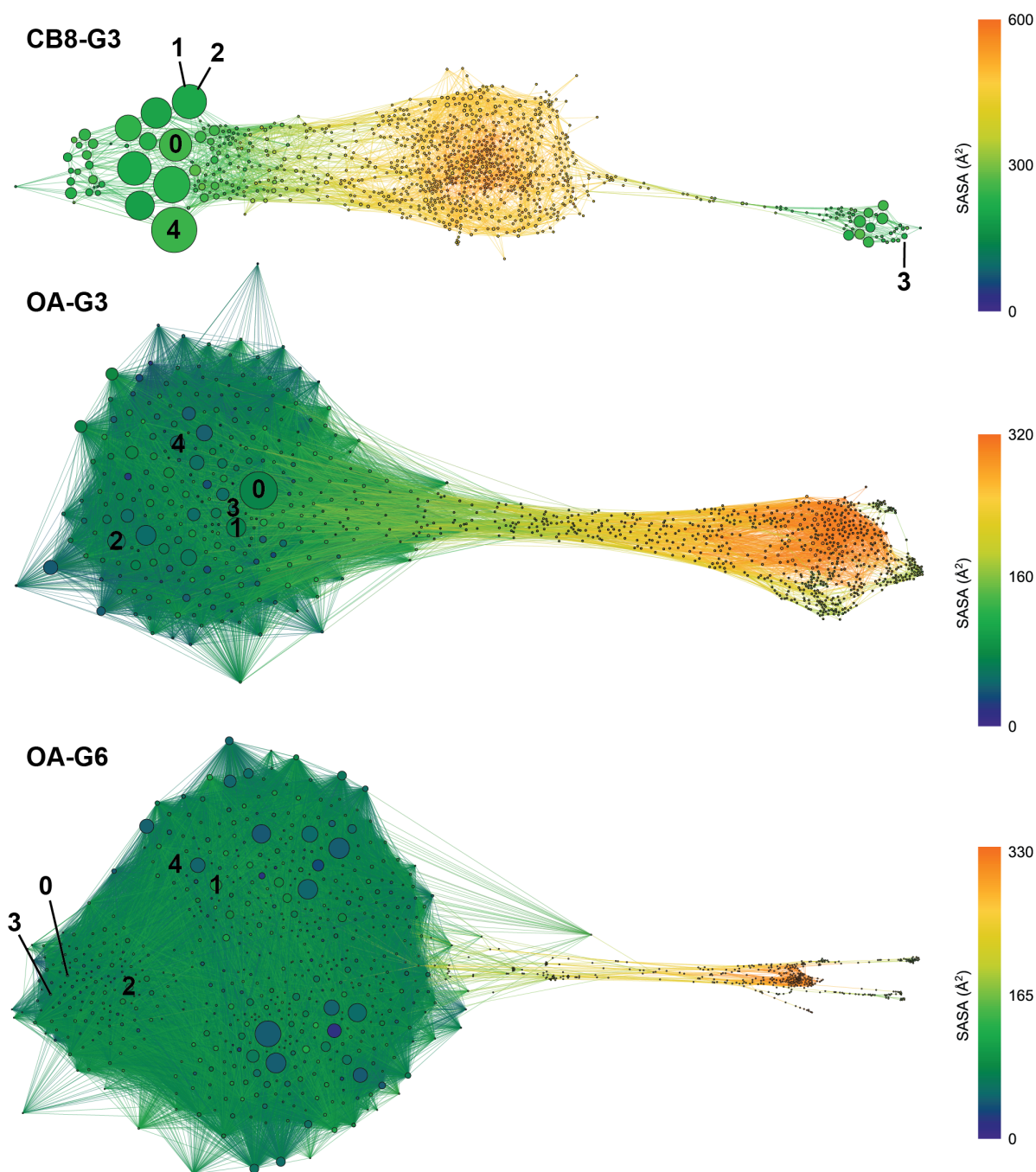


Figure 2.11: Conformation space networks for the unbinding simulations. Each node in a CSN represents a cluster of host-guest structures. Edges in the networks connect clusters that are seen to interconvert in the REVO simulations. The size of each node is proportional to the number of times it was observed in the unbinding simulations. Nodes are colored according to the solvent accessible surface area of the guest molecule, as shown in the color-bars on the right. The clusters corresponding to the starting poses are labeled in each network.

basin from the other starting poses. This allows us to observe that the ligand cannot flip between these two structures inside the host, and instead converts between the two poses only through the quasi-bound and unbound states (yellow and orange). Although here we conclude that all structures are part of the same (or symmetry-equivalent) fast-interconverting bound ensemble, this type of analysis is useful to reveal the interconversion of binding poses, and whether we should expect them to have the same calculated RTs.

2.4 Discussion

Although we obtained much information about the binding and release processes of these host-guest systems, our predicted ΔG values were systematically lower than those of a reference calculation employing the same forcefield (average 4.2 kcal/mol). These reference ΔG values were themselves systematically lower than the experimentally calculated dissociation constants (average 2.7 kcal/mol), likely arising from inaccuracies in the forcefield. The nature of the SAMPLing challenge gives us a unique opportunity to isolate these different sources of error. Below we discuss different possible sources of error in light of the analyses presented above.

In weighted ensemble simulations that calculate kinetic quantities, convergence is often the first question. Here we devoted the same amount of sampling time to the binding and unbinding processes (1.92 μ s per system per starting pose). This is more than sufficient to capture the binding process, which has a mean first passage time ranging from 36 to 160 ns. The unbinding process was much more challenging, and it is possible that longer simulations would have captured higher weight walkers exiting from the bound state. This would increase our k_{off} estimates, and K_D as well. Significantly extending the unbinding simulations and monitoring their exit rates could provide additional insight.

We also have concerns related to the size of the simulation box. This was chosen to be appropriate for standard alchemical free energy perturbations, and not for simulations of full unbinding and binding pathways. A more accurate determination of the binding rate

could be obtained with the Northrup-Allison-McCammon (NAM) method, which combines the rate of first hitting points with a committor probability to determine the binding rate [82]. Diffusion at long distances is typically efficiently simulated using Brownian dynamics. This approach has been used successfully to determine binding rates with both the weighted ensemble method [124, 83], and the SEEKR method (Simulation Enabled Estimation of Kinetic Rates) [104].

An important point is that although the reference calculations were performed with the same forcefield, the rates can sensitively depend on aspects of the forcefield that are not relevant to alchemical measurements of the affinity. As an example, in OA-G3 unbinding trajectory trees we observe long “tendrils” of unbound trajectories that are stuck in intermediate SASA values, where the guest ligand is bound to the outer surface of the host. The strength of these interactions can significantly affect our calculations of k_{off} , although they will not affect the alchemical K_D calculations.

In general, to successfully predict k_{on} and k_{off} will require optimizing the ligand forcefield terms that govern interactions that occur along binding pathways. By analogy, it is known that protein forcefields that are only trained on *folded* protein structures have difficulties representing unfolded and intrinsically disordered structures. As a community we must take care not to over-emphasize the ligand bound state in forcefield development. An extension of the SAMPL challenge to include the prediction of kinetic quantities would thus be tremendously valuable to the development of both sampling methodologies and forcefields.

CHAPTER 3

ON CALCULATING FREE ENERGY DIFFERENCES USING ENSEMBLES OF TRANSITION PATHS

This work was done in collaboration with Robert Hall, a postbaccalaureate student working in our research group. Robert ran simulations and performed analysis. I acted as a mentor and co-advisor to help Robert analyze the simulations, make meaningful conclusions and prepare the manuscript.

This work was published in *Frontiers in Molecular Biosciences* volume 6 page 106 in 2020. The work is presented here as published except that the supplemental figures are worked into the text.

3.1 Introduction

In recent years there is a growing appreciation for the utility of binding kinetics in the prediction of drug efficacy [20, 125, 126, 127, 128, 2, 129, 130, 131, 132]. Pharmacokinetic and pharmacodynamic models of drug activity in the body are inherently out of equilibrium: a drug is administered, it is absorbed, distributed to different tissues, metabolized and eliminated from the body. As such, kinetic constants of binding and release – beyond just the equilibrium constants of binding – are required to model drug action when the timescales of binding and release cannot be separated from the other competing processes [133]. The relationship between molecular structure and the kinetics of binding (also called “structure-kinetic relationships” or SKR) is complicated, as small changes to structure can change kinetic constants by orders of magnitude [128]. It is important to note that changes in kinetics are not always tied to changes in affinity [134], and that to accurately predict changes in kinetics, models of the ligand-binding transition state are needed to estimate transition-state stabilization or destabilization [135].

Computational methods that reveal structures of transition states and calculate binding

(k_{on}) and unbinding (k_{off}) rate constants for real compounds are in their infancy, but are quickly developing [102]. It is a tremendous challenge to obtain reliable values for these quantities, as 1) they depend on the entire (un)binding pathway, not just its endpoints, and 2) the timescales of ligand binding and release often exceed the capabilities of molecular dynamics (MD) simulations by orders of magnitude. Specialized computing platforms have been applied to generate continuous binding pathways [136], although the unbinding process is typically beyond the reach of MD simulation for compounds beyond millimolar drug fragments [130, 137]. Recent studies have used enhanced sampling methods in MD to simulate ligand (un)binding pathways and determine mechanisms and rate constants k_{on} and k_{off} [56, 110, 138, 101, 16, 139, 140, 141]. Some of these rate constants have shown surprisingly good agreement with experiment – given the extraordinarily long timescales involved – however these have the confounding uncertainty of force field accuracy [142, 143], there is a possibility for fortuitous cancellation of error. Unfortunately, the computational cost required to predict these quantities is typically massive [143], especially for large protein systems and ligands with extremely long residence times (RTs), precluding the study of these events under a series of different simulation conditions (e.g. forcefields, water models, polarizability).

In the field of biomolecular modeling, blind challenges – where a series of objectives are released by the organizers, and participants entries are directly judged by their agreement with experiment – have been useful catalysts for the development of predictive algorithms [144, 145, 146, 147]. Although no blind challenge currently exists for the prediction of k_{on} and k_{off} , we recently participated in the SAMPL6 SAMPLing challenge, which required participants to compute free energies as a function of simulation time and to compare the computational cost of different free energy calculation methods [148, 99, 74]. This challenge allows sampling methods to be assessed independently of force field accuracy, as all entries used the same initial coordinates, force field parameters and partial charges. Importantly, the challenge makes use of very small model systems (host-guest) that require considerably

less computational resources to simulate, which allowed us to efficiently simulate binding and release for a number of systems, determine k_{on} and k_{off} , and predict values for the binding free energy (ΔG) that would then be compared to experimental observables, as well as results from alchemical free energy perturbation methods [149, 150].

The binding free energy was determined as a function of rate constants:

$$\Delta G = -k_B T \ln \left(\frac{C^0 k_{\text{on}}}{k_{\text{off}}} \right), \quad (3.1)$$

where C^0 is a reference concentration of 1 mol/L. In this chapter, we revisit this equation in detail and explicitly examine the assumptions made when the rate constants used in Eq. 3.1 are computed through typical simulations with finite box-size and periodic boundary conditions. In Section 3.3.1 we derive three correction terms that can be easily computed and facilitate a better connection with both experiment and alchemical computational free energy calculations. To examine questions of convergence, we reproduce our binding and unbinding simulations with larger numbers of replicas and longer simulation times. We also explore the effects of the Langevin integrator on the prediction of unbinding and binding rates; in particular, how altering the friction coefficient (γ), defined in the Langevin integrator, impacts the binding and release processes. Although γ does not appear in the internal energy function, and hence cannot affect thermodynamic properties such as the binding free energy, we examine whether lower friction coefficients can accelerate the convergence of unbinding simulations.

3.2 Methods

3.2.1 Host-guest systems

The host-guest system utilized in this study is referred to as OA-G6 (Fig. 3.1), where the host, octa acid (OA) is a Gibb deep cavity cavitand, referred to as an octa acid (OA [113]. OA forms a basket-like structure with 4-fold symmetry, functionalized with four benzoic-acid substituents on the top rim of the basket and four more on the bottom. The guest ligand we

study here is 4-methyl pentanoic acid (referred to as “G6”). This ligand harbors a negative charge at the carboxyl end of the alkyl chain.

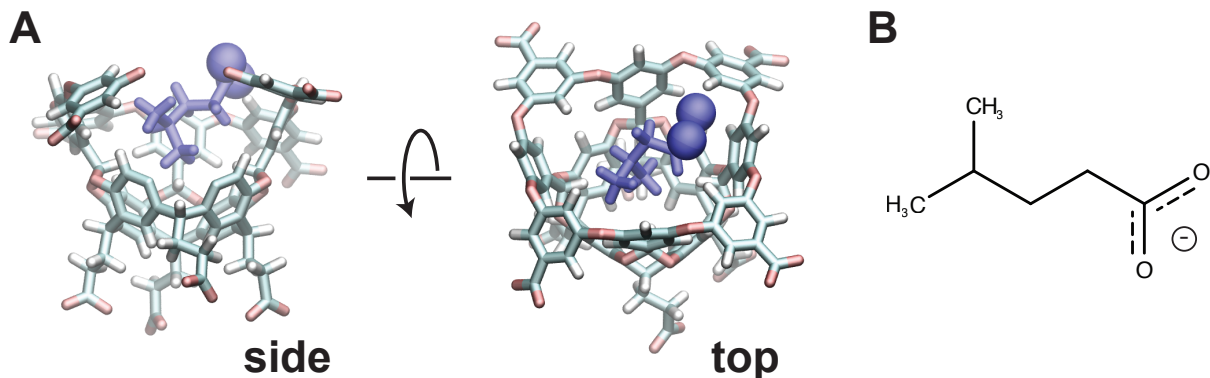


Figure 3.1: (A) The initial pose for the OA-G6 system (side view: left, top view: right). Note that some atoms from the host are removed in the side view for clarity. The carboxyl oxygens are shown in sphere representation. (B) The chemical structure of the G6 ligand in the deprotonated form.

3.2.2 Molecular dynamics

The OA-G6 configuration was obtained from the organizers of the SAMPLing challenge [148]. The system was solvated in a (roughly) cubic box with box length 4.28, 4.33 and 4.33 nm in the x , y and z dimensions, respectively. The system provided had a total of 7976 atoms: 2586 water molecules to solvate the system, 12 sodium and 3 chloride ions to neutralize the system, and the remaining atoms belonging to either the host or the guest. OpenMM v7.2.1 [114] was used to run dynamics with the CUDA v9.0.176 platform. A Monte Carlo barostat is used to maintain a constant pressure of 1 atm. A time step of 2 fs was used across all simulations.

We utilize the Langevin integrator, which uses a drag term and a noise term to account for the friction of solvent molecules and high velocity collisions that perturb the system. Langevin dynamics allows for the temperature to be controlled and can be used as a thermostat; we run all dynamics here at 300 K. Our host-guest system follows the Langevin

equation, shown below:

$$\mathbf{F}(t) = -\nabla U(\mathbf{r}(t)) - \gamma m \mathbf{v}(t) + \sqrt{2m\gamma k_B T} \mathbf{R}(t), \quad (3.2)$$

where $U(\vec{r}(t))$ is the particle interaction potential, $\vec{R}(t)$ is a random Gaussian noise term, T is the temperature, k_B is the Boltzmann constant, and γ is the friction coefficient. The friction term plays two different roles here, both modulating the second “drag” term, and the Gaussian noise. As γ approaches zero, the noise gets weaker and the dynamics becomes more deterministic. Here we run binding and unbinding simulations with γ values of 1.0, 0.1 and 0.01 ps⁻¹.

3.2.3 Reweighting of Ensembles by Variance Optimization

To generate an ensemble of ligand unbinding events, we need to employ enhanced sampling as the timescale of ligand unbinding events in this system is prohibitively long: we found in previous studies a mean first passage time of 2.1 s (Chapter 2), which is six orders of magnitude longer than the reach of conventional MD sampling. In this work, we implement the Resampling Ensembles by Variation Optimization (REVO) method, based on weighted ensemble (WE) framework, to encourage the sampling of rare unbinding/rebinding events. WE accelerates the sampling of rare events using an ensemble of trajectories that are each assigned a statistical weight [71]. The ensemble is integrated forward in time in a parallel fashion, and periodically “resampled” by cloning certain trajectories and merging others. When a trajectory is cloned, its weight is divided amongst the clones, but the multiple copies of the trajectory go on to evolve independently. By repeatedly cloning trajectories that are in undersampled regions of space we can obtain statistics on very long-timescale events using only short-timescale simulations.

The REVO was designed to efficiently perform cloning and merging operations on small ensembles of trajectories that are evolving in high-dimensional spaces [35]. This is valuable in situations where it is difficult to define one or two progress variables that capture the long-

timescale events of interest. In REVO, coupled cloning and merging operations are proposed (e.g. clone trajectory i , and merge trajectories j and k) and are accepted or rejected based on an objective function called the "trajectory variation":

$$V = \sum_i V_i = \sum_i \sum_j (d_{ij}/d_0)^\alpha \phi_i \phi_j, \quad (3.3)$$

where d_{ij} is the distance between trajectories i and j , α and d_0 are parameters, and ϕ_x is a function that measures the importance, or "novelty" of a trajectory x , which in our work here is strictly a function of the weight of the trajectory: $\phi_i = \log w_i - C$, where w_i is the weight of trajectory i and C is a constant. Trajectories with the highest V_i values in Eq. 3.3 are chosen for cloning, and those with the lowest V_i are chosen for merging. More information about the algorithm can be found in previous work [35].

We run separate simulations for the binding and unbinding processes. In our unbinding simulations, the ligands start in the bound state and are terminated as they unbind. In the rebinding simulations, the ligands start in the unbound state and are terminated as they bind. The distance function (d_{ij}) we use in Eq. 3.3 is different for these two simulation types. For the unbinding simulations, we superimpose the hosts from trajectories i and j , and then compute the root mean square deviation (RMSD) between the guest molecules, without any further alignment [77, 109]. As there is 4-fold symmetry in this system, we perform the alignment four times (once for each symmetrically-equivalent mapping) and use the smallest such distance as d_{ij} . For the rebinding simulations, we calculate the distance to the native state for each trajectory ($d_{\text{native}}(\mathbf{X}_i)$), which again takes into account the four symmetry mappings, using the lowest such distance. The distance between trajectories i and j is then calculated as $d_{ij} = |1/d_{\text{native}}(\mathbf{X}_i) - 1/d_{\text{native}}(\mathbf{X}_j)|$, where the inverse is used to prioritize differences between small values of d_{native} .

3.2.4 Calculating rates by ensemble splitting

A major advantage of the REVO method, much like other weighted ensemble methods, is that it can calculate kinetic parameters in real time as the simulation progresses. This is achieved by running separate simulations for the binding and unbinding processes, and in each case, measuring the trajectory flux into the opposite basin [85, 86, 87, 88, 89]. The unbound basin is defined as the set of structures where the closest host-guest interatomic distance is > 1 nm, following previous work [77, 109, 16]. The bound basin is defined as the set of structures where the guest RMSD (compared to the native structure) is < 0.1 nm after aligning to the host. Again, this RMSD measurement takes into account the four symmetry-equivalent mappings of OA.

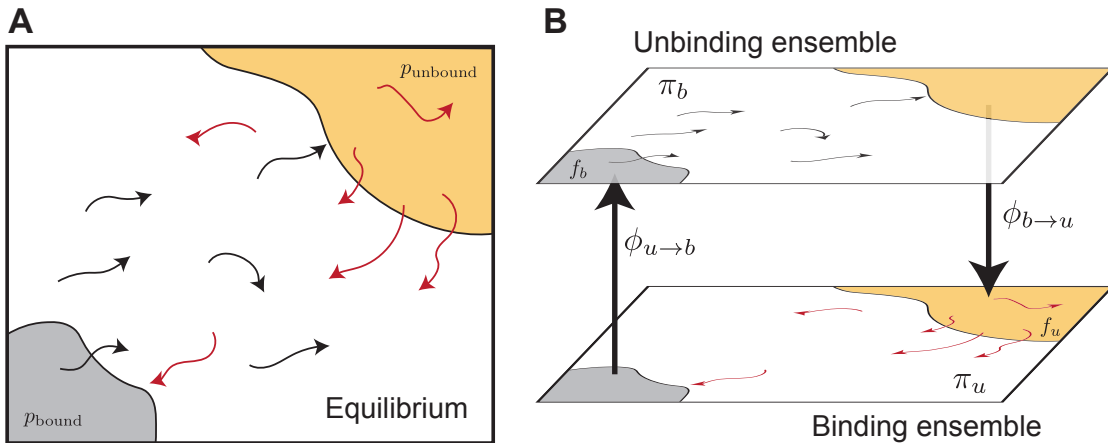


Figure 3.2: Splitting an equilibrium ensemble into two history-dependent ensembles using basins. The bound and unbound basins are shown in grey and light orange, respectively. The unbinding ensemble (B, top) contains all trajectories that last visited the bound basin, which are shown in black. The binding ensemble (B, bottom, also referred to as the “rebinding” ensemble) contains all trajectories that last visited the unbound basin, shown in red. Simulations in a given ensemble are terminated once they reach the destination basin and thus switch ensembles. The trajectory flux between ensembles is denoted by $\phi_{u \rightarrow b}$ and $\phi_{b \rightarrow u}$. The quantity π_b refers to the probability of the entire top ensemble, and the quantity f_b denotes the probability of the bound basin within the unbinding ensemble.

In our studies, the binding and rebinding REVO simulations are conducted separately. However, the methodology of obtaining on and off-rates is essentially the same. After each

dynamics step, if a walker has entered the opposite basin, as described above, its weight is recorded and its structure is “warped” back to the starting structure at the beginning of the simulation. The atomic coordinates are set to the starting structure and the velocities are reinitialized; however, the weight of the trajectory remains the same. Before the warping event to the starting structure, the structure of the walker is recorded and is referred to as an “exit point”. In our unbinding simulations, the initial starting structure is the initial bound pose provided. In our rebinding simulations, the initial starting structure is chosen from a set of exit points generated from the unbinding simulations. Therefore, the unbinding analyses were performed prior to initialization and the subsequent running of our rebinding simulations.

The off and on-rates are calculated by using the flux of trajectories into either the unbound or bound state respectively, and mathematically calculated by:

$$k_{\text{off}}(t) = \frac{\sum_i w_i}{T}, \quad (3.4)$$

$$k_{\text{on}}(t) = \frac{\sum_i w_i}{CT}, \quad (3.5)$$

where the sum is over the set of “warped” trajectories, T is the elapsed simulation time, and C is the concentration of ligand, computed as $1/V$ where V is the box volume. The box volume was approximately 80.2 nm^3 , corresponding to a concentration of ligand of 0.0207 M .

There are a few key differences between the REVO simulations discussed here and in Chapter 2. For both the unbinding and rebinding simulations in this study, the total simulation time is 2.25 times longer compared to our previous study, as our current unbinding and rebinding simulations were run for 4500 and 450 cycles, respectively. Additionally, ten independent unbinding simulations were run for each of the four friction coefficients, whereas our previous study only ran five independent simulations for each starting pose. However, only five independent rebinding simulations were run for each of the coefficients, as we observe much less variation in the k_{on} estimates. Finally, 48 walkers were used in both studies

and the time per cycle is consistent, where the unbinding simulations are 20 ps/cycle and the rebinding simulations are 200 ps/cycle.

3.2.5 Calculating electrostatic interaction energies

The electrostatic energy between the host and guest molecules for use in the second correction term was calculated as: $E_{\text{int}} = \frac{1}{4\pi\epsilon_w} \frac{Q_i Q_j}{r_{ij}}$ where Q_a is the partial charge of atom a used in the force field during simulation. r_{ij} is the interatomic distance between atoms i and j calculated by using the minimum image convention. $\epsilon_w = 6.88 \times 10^{-10}$ F/m is the permittivity of water at 300 K calculated by linear interpolation of the water dielectric constant at 298.15 and 303.15 K [151].

3.3 Results

3.3.1 Derivation of correction terms

The binding free energy can be calculated using the rate constants k_{on} and k_{off} as $\Delta G = G_{\text{bound}} - G_{\text{unbound}} = -kT \ln(K_{eq} C_0) = -kT \ln\left(\frac{C_0 k_{\text{on}}}{k_{\text{off}}}\right)$, where K_{eq} is the binding equilibrium constant, C_0 is the reference concentration of 1 mol/L, k is Boltzmann’s constant and T is the temperature in Kelvin. While this relationship is correct in the macroscopic limit, it fails to account for the box size and the volume of the unbound state in finite simulation environments with periodic boundary conditions. Here we derive a more accurate expression for the binding free energy that accounts for the finite box size in a typical MD simulation.

Our starting point is an expression for K_{eq} , which is valid for a dilute solution in thermodynamic equilibrium. We use the notation of Woo and Roux (see Eq. 4 from Ref. [152]):

$$K_{eq} = \frac{\int_{\text{bound}} d\mathbf{l} \int d\mathbf{X} e^{-\beta U}}{\int_{\text{bulk}} d\mathbf{l} \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X} e^{-\beta U}}, \quad (3.6)$$

where U is the internal energy of the system, $\beta = 1/kT$ is the inverse temperature, \mathbf{r}_1 is the center of mass of the ligand (referred to as a “guest” molecule) and \mathbf{r}_1^* is an arbitrary position of the guest in the bulk. Note that $d\mathbf{l}$ integrates over the guest positions, and $d\mathbf{X}$ integrates

over everything else: the host and the solvent degrees of freedom. Note also that K_{eq} has units of volume, as the delta function constraining the center of mass in the denominator removes three spatial degrees of freedom.

Here we examine the calculation of free energies using rates determined from split ensemble calculations (Fig. 3.2, see Section 3.2.4 for more details). We denote the probability of these two ensembles as π_b and π_u , where $\pi_b + \pi_u = 1$, and:

$$\frac{\pi_b}{\pi_u} = \frac{\phi_{u \rightarrow b}}{\phi_{b \rightarrow u}}, \quad (3.7)$$

where $\phi_{a \rightarrow b}$ is the time-averaged flux from the a ensemble to the b ensemble (i.e. across the dotted lines in Fig. 3.2). The equilibrium probability of a position \mathbf{X} can be obtained by combining estimates from both ensembles:

$$p(\mathbf{X}) = p_u(\mathbf{X})\pi_u + p_b(\mathbf{X})\pi_b, \quad (3.8)$$

where $p_a(\mathbf{X})$ is the probability of conformation \mathbf{X} in ensemble a , which is normalized such that $\int p_a(\mathbf{X})d\mathbf{X} = 1$.

Let us define the bound state as the domain of the integral in the numerator of Eq. 3.6, and the unbound state as a set of structures considered unbound in simulation (not the same as the bulk state in Eq. 3.6). These states are shown as shaded regions in Fig. 3.2. The ratio of the probabilities of these two states, at equilibrium, is given by:

$$\frac{p_{\text{bound}}}{p_{\text{unbound}}} = \frac{\int_{\text{bound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U}}{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U}}, \quad (3.9)$$

which can also be calculated in our ensemble splitting simulations:

$$\frac{p_{\text{bound}}}{p_{\text{unbound}}} = \frac{\pi_b \int_{\text{bound}} p_b(\mathbf{X}) d\mathbf{X}}{\pi_u \int_{\text{unbound}} p_u(\mathbf{X}) d\mathbf{X}} = \frac{\pi_b f_b}{\pi_u f_u}, \quad (3.10)$$

where f_a is the probability of the basin state within ensemble a .

Expanding Eq. 3.6 we have:

$$\begin{aligned} K_{eq} &= \frac{\int_{\text{bound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U}}{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U}} \frac{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U}}{\int_{\text{bulk}} d\mathbf{1} \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X} e^{-\beta U}} \\ &= \frac{\pi_b f_b}{\pi_u f_u} \frac{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U}}{\int_{\text{bulk}} d\mathbf{1} \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X} e^{-\beta U}}. \end{aligned} \quad (3.11)$$

The unbound state in simulation is far enough that the host and guest do not interact directly through van der Waals interactions, although if both molecules carry an explicit charge – as in the example considered here – there could still be significant host-guest electrostatic interactions. To account for these, we introduce another intermediate state with an altered energy function (U^*) which is the same as U except that it does not include electrostatic interactions between the host and the guest:

$$K_{eq} = \frac{\pi_b f_b}{\pi_u f_u} \frac{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U}}{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U^*}} \frac{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U^*}}{\int_{\text{bulk}} d\mathbf{1} \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X} e^{-\beta U}} \quad (3.12)$$

$$= \frac{\pi_b f_b}{\pi_u f_u} \langle e^{\beta E_{\text{int}}} \rangle_{\text{unb}}^{-1} \frac{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U^*}}{\int_{\text{bulk}} d\mathbf{1} \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X} e^{-\beta U}}. \quad (3.13)$$

where $E_{\text{int}} = U - U^*$ and the subscript “unb” indicates an ensemble average over structures in the unbound state obtained with the normal energy function U . Note the final step used the relation:

$$\frac{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U^*}}{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U}} = \frac{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{\beta E_{\text{int}}} e^{-\beta U}}{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U}} = \langle e^{\beta E_{\text{int}}} \rangle_{\text{unb}}. \quad (3.14)$$

We can now reasonably assume that the guest in the unbound state is non-interacting with the host. This allows us to write $e^{-\beta U}$ as $e^{-\beta U_G} e^{-\beta U_{HS}}$, where U_G are the terms in the energy function that depend only on the coordinates of the guest, and U_{HS} are terms that only depend on the host and the solvent. We can then pull the integral $\int d\mathbf{X} e^{-\beta U_{HS}}$ out of the numerator and denominator of the last term of Eq. 3.13:

$$\frac{\int_{\text{unbound}} d\mathbf{1} \int d\mathbf{X} e^{-\beta U^*}}{\int_{\text{bulk}} d\mathbf{1} \delta(\mathbf{r}_1 - \mathbf{r}_1^*) \int d\mathbf{X} e^{-\beta U}} = \frac{\int_{\text{unbound}} d\mathbf{1} e^{-\beta U_G}}{\int_{\text{bulk}} d\mathbf{1} \delta(\mathbf{r}_1 - \mathbf{r}_1^*) e^{-\beta U_G}}. \quad (3.15)$$

The bottom integral has the center of mass of the ligand fixed and is only over internal and rotational degrees of freedom of the ligand. This can also be separated and removed from the numerator, which simplifies the ratio to be the volume of the unbound state, defined as:

$$V_{\text{unbound}} = \frac{\int_{\text{unbound}} d\mathbf{1} e^{-\beta U_G}}{\int_{\text{guest}} d\mathbf{G}_1 e^{-\beta U_G}} = \int_{\text{box}} d\mathbf{R} \phi_u(\mathbf{R}), \quad (3.16)$$

where we use \mathbf{G}_1 to denote the internal and rotational degrees of freedom of the guest that remain after specification of \mathbf{r}_1 . The quantity $\phi_u(\mathbf{R})$ is the fraction of conformers with center

of mass \mathbf{R} that satisfy the unbound boundary conditions: here, that the guest atoms are all farther than a cutoff distance of 1 nm away from the host. This integral can be calculated by Monte Carlo, where a center of mass position and orientation of the ligand is randomly generated, and the number of successful unbound conformers is recorded:

$$V_{\text{unbound}} = V_{\text{box}} \frac{N_{\text{unbound}}}{N_{\text{trials}}}. \quad (3.17)$$

Note that for large boxes $V_{\text{unbound}} \approx V_{\text{box}}$.

Putting this all together we have:

$$K_{eq} = \frac{\pi_b f_b}{\pi_u f_u} \langle e^{\beta E_{\text{int}}} \rangle_{\text{unb}}^{-1} V_{\text{unbound}}, \quad (3.18)$$

which differs from the straightforward interpretation used in Chapter 2:

$$K_{eq}^0 = \frac{\pi_b}{\pi_u [L]} = \frac{\pi_b}{\pi_u} V_{\text{box}}. \quad (3.19)$$

Using $\Delta G = -kT \ln(K_{eq} C_0)$, we have:

$$\Delta G = \Delta G^0 - kT \ln \left(\frac{f_b}{f_u} \right) + kT \ln \langle e^{\beta E_{\text{int}}} \rangle_{\text{unb}} - kT \ln \left(\frac{V_{\text{unbound}}}{V_{\text{box}}} \right), \quad (3.20)$$

which explicitly shows ΔG as the sum of $\Delta G^0 = -kT \ln(K_{eq}^0 C_0)$ and the three newly derived correction terms. The first term will go to zero in the limit that the basin states are chosen to represent the vast majority of the probability in both the binding and unbinding ensembles. In other words, this term goes to zero when both f_b and f_u approach one. The second term is likely to only be non-negligible in the case of explicitly charged host and guest molecules and regardless would go to zero as the definition of the unbound state is moved to farther and farther distances. The third term would also go to zero for large simulation boxes, but in practice this is often not feasible due to computational constraints. Consequently, $V_{\text{unbound}}/V_{\text{box}}$ could be much less than one, introducing a correction in the positive direction. Below we calculate these three correction terms and apply them to free energy calculations.

3.3.2 Extended trajectory ensembles with lower friction coefficients

In previous work, we used a Langevin integrator with a value of $\gamma = 1 \text{ ps}^{-1}$ for the friction coefficient. As the simulations already have explicit solvent, this adds extra friction into the system that is not physical. Here we investigate whether reducing γ to values less than one will significantly affect our rate calculations. We thus run a set of trajectory ensembles at multiple values of γ and extend each ensemble to be longer than those discussed in Chapter 3 to more fully examine questions of convergence.

As γ governs the coupling to the Langevin thermostat, we determine the minimum value of γ where our target temperature (300 K) is maintained. We first ran a series of short simulations (one 10 ns trajectory for each γ) and find that temperature control is completely lost for friction coefficients less than $\gamma = 0.001$ (Figure 3.3A). We then ran longer simulations for $\gamma = 1, 0.1, 0.01$ and 0.001 , examining not only the mean temperature, but the probability of significant temperature fluctuations, which could spur anomalous results in our ligand dissociation simulations. Figure 3.3B shows the probability distribution of observed temperatures over an ensemble of 240 trajectories run for 90 ns each. For $\gamma = 0.01, 0.1$ and 1 ps^{-1} , the temperature distribution is normally distributed around the mean (300 K) as seen by the parabolic curves on a log scale. Temperature control is not fully maintained for $\gamma = 0.001 \text{ ps}^{-1}$, as shown by a rightward shift and slight widening of the parabolic distribution. We thus restrict our analysis to three values of the friction coefficient: $\gamma = 0.01, 0.1$ and 1 ps^{-1} .

We run both unbinding and rebinding REVO simulations for the OA-G6 system. For unbinding, we ran 10 simulations for each of the three friction coefficients; for rebinding, we ran 5 simulations for each coefficient, yielding a total of 30 simulations for unbinding and 15 simulations for rebinding. A set of binding and unbinding simulations were also run for $\gamma = 0.001$ – despite the impaired temperature control – which are reported in Figure 3.5. The estimates for the unbinding and binding fluxes are depicted in Figure 3.4, where each curve represents an individual REVO simulation. The averages, illustrated with a bolded line, are calculated by averaging the trajectory flux over the entire set of simulations for

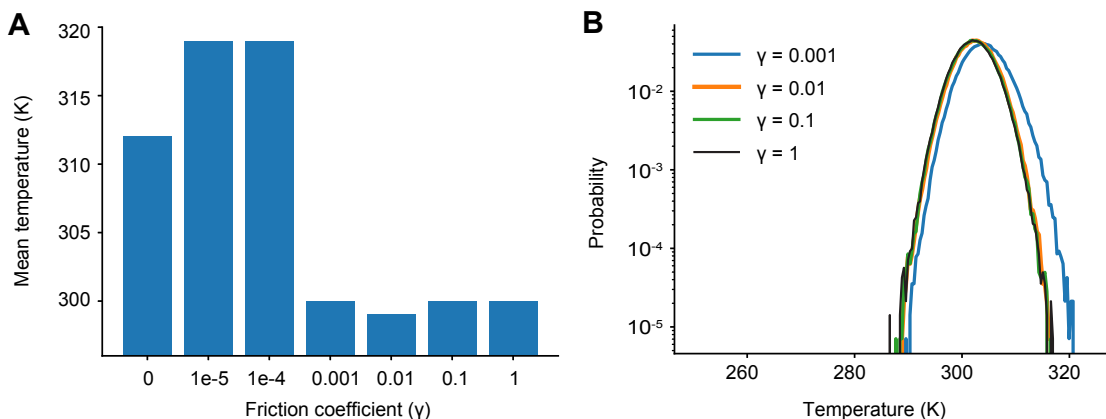


Figure 3.3: (A) Average temperatures observed in short simulations for different friction coefficients (γ). (B) Probability distributions of observed temperatures from ensembles of longer simulations with different γ .

that value of γ . The upward jumps on these plots indicate that an exit point was recorded that has a higher weight than was previously observed. A set of binding and unbinding simulations were also run for $\gamma = 0.001$ – despite the impaired temperature control – the rates of which are depicted in Figure 3.5

By reducing γ to values less than 1, we observed no change in the binding rates, and small changes to the unbinding rates which are on the border of significance. With regard to unbinding rates, the two largest friction coefficients yielded the smallest error and similar k_{off} values, where $\gamma = 1$ yielded an average off-rate of 16.4 s^{-1} and $\gamma = 0.1$ yielded an off-rate of 11.5 s^{-1} . The off-rate increased by 10-fold for $\gamma = 0.01$, although this is mostly driven by exit points observed in a single simulation. In our previous OA-G6 results using $\gamma = 1$, we calculated an unbinding rate of 0.48 s^{-1} which slightly differs from the value calculated in this study using $\gamma = 1$ (Table 3.1). Unbinding rates for $\gamma = 0.001 \text{ ps}^{-1}$ were approximately 1000-fold higher, although these are known to be affected by a higher average temperature (SI). Taking a closer look at the binding rates, we saw no discernible difference across the friction coefficients. The binding rate was approximately $10^9 \text{ s}^{-1} \text{ M}^{-1}$, for all friction coefficients, which was about 5-fold larger when compared to our previous study using $\gamma = 1$. For both

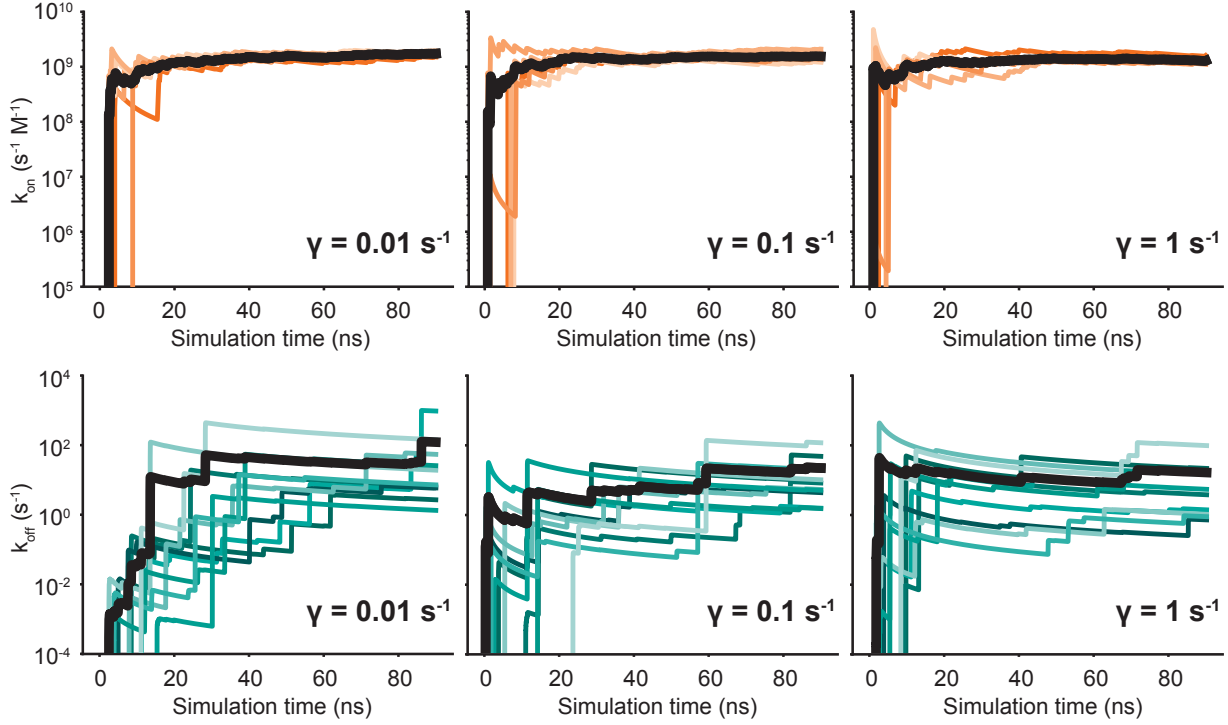


Figure 3.4: Predicted on- (top) and off-rates (bottom) as a function of simulation time. Each panel is labeled according to the friction coefficient used for that set of simulations. The independent simulations are shown in shades of orange (k_{on}) and blue (k_{off}), and the averages are depicted by bold black lines.

binding and unbinding rates we have more confidence in the results obtained here, as they are based on more extensive simulation data.

Table 3.1: Binding and unbinding rates as a function of friction coefficient (γ). The uncertainties shown use the standard error of the mean calculated from 5 and 10 independent REVO runs for binding and unbinding, respectively. The quantities from Chapter 2 were obtained with 5 REVO runs that used different initial conformations, each of which were 2000 cycles in length.

	$k_{\text{on}} (10^8 M^{-1} s^{-1})$	$k_{\text{off}} (s^{-1})$
$\gamma = 0.01$	17 ± 1	122 ± 94
$\gamma = 0.1$	16 ± 2	22 ± 12
$\gamma = 1$	13 ± 1	16.4 ± 9.4
Chapter 5 ($\gamma = 1$)	2.8 ± 1.0	0.48 ± 0.11

For both the unbinding and rebinding simulations, across all friction coefficients, we

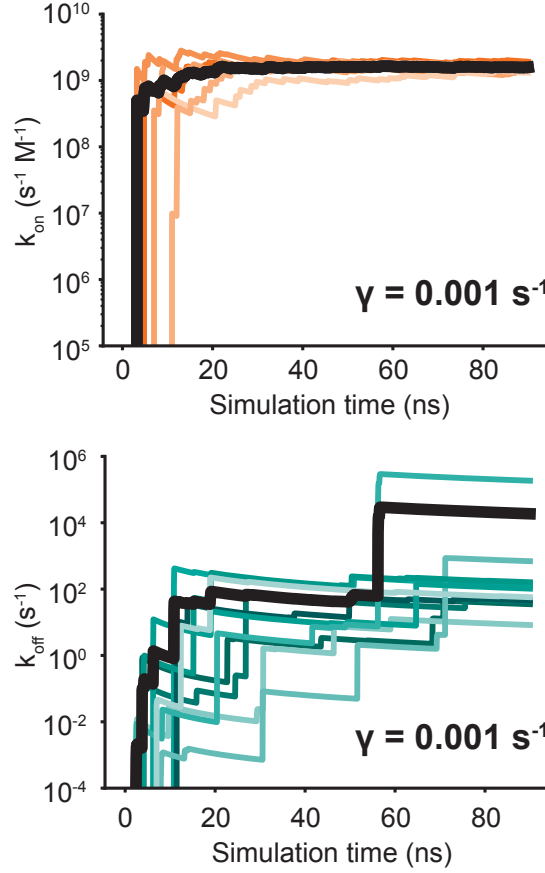


Figure 3.5: Binding (top) and unbinding (bottom) fluxes for $\gamma = 0.001 \text{ ps}^{-1}$. Fluxes are shown for each simulation individually. Parameters are the same as those used for higher γ values in the main text. Average fluxes over the simulations are shown as thick black lines.

observed at least 1000 warping events (Figure 3.7). As expected, we observe that rebinding occurs with a much higher probability when compared to unbinding, by several orders of magnitude. The unbinding walker weights are limited at the low end by the minimum walker probability (p_{min}), which is set to 10^{-12} . The rebinding walker weights are limited at the high end by the maximum walker probability (p_{max}), which is set to 10^{-1} , respectively. Figure 3.7 shows that the 10-fold larger unbinding rate for $\gamma = 0.01$ was largely due to a single unbinding point in a single simulation, which underscores the sensitivity and uncertainty of rate calculations using trajectory fluxes. Figure 3.6 shows unbinding fluxes for $\gamma = 0.001$, which is known to have elevated temperatures. There we see a large number of high-weight

unbinding events in two different simulations, leading to the 1000-fold increase in k_{off} .

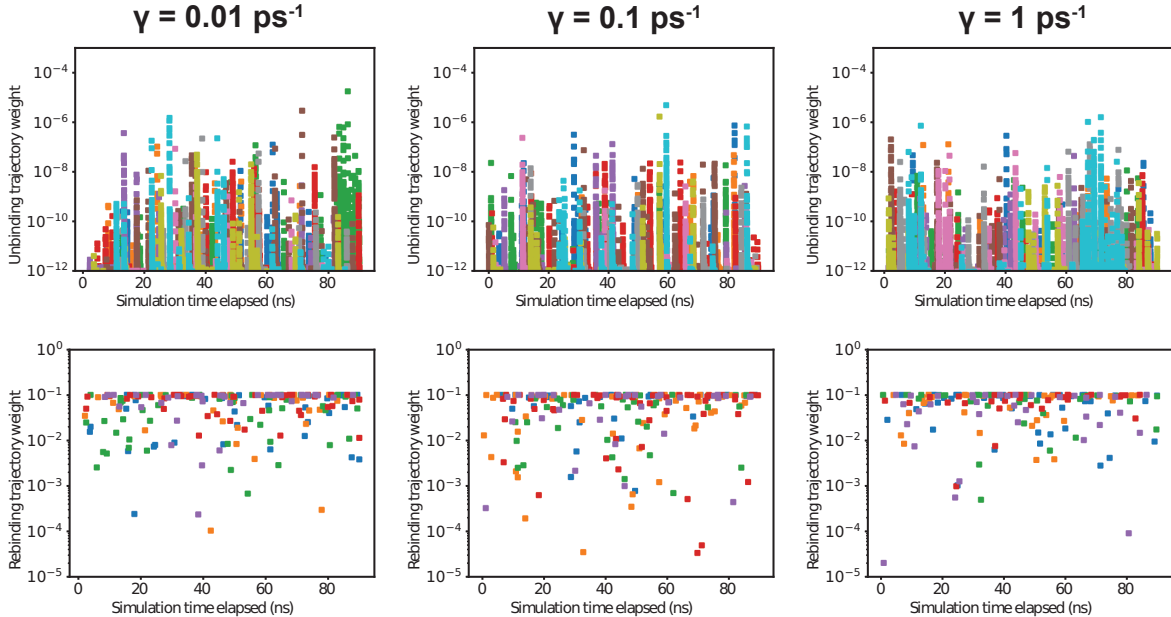


Figure 3.6: Weights of warped walkers in unbinding (top) and binding (bottom) REVO simulations for $\gamma = 0.01, 0.1$ and 1.0 ps^{-1} . Each simulation is shown in a different color.

3.3.3 Free energy estimates, correction terms and comparison with previous benchmarks

As the friction coefficient unevenly affected the rates of binding and unbinding, there was a net effect on the binding free energies. As shown in Figure 3.8 and Table 3.2, the binding free energy increases as the friction coefficient is lowered, independent of the free energy correction terms derived in Section 3.3.1. Table 3.2 shows the free energies computed using the averaged fluxes across all simulations at each γ value. For all friction coefficients, the calculated free energy was always higher than that from our previous study (-12.1 kcal/mol ; red line), even for $\gamma = 1$, signifying that extending the simulation time aided in predicting experimentally determined binding free energies.

The correction terms are calculated using data obtained from the simulations, but they are mostly functions of geometric properties of the simulation box and boundary condi-

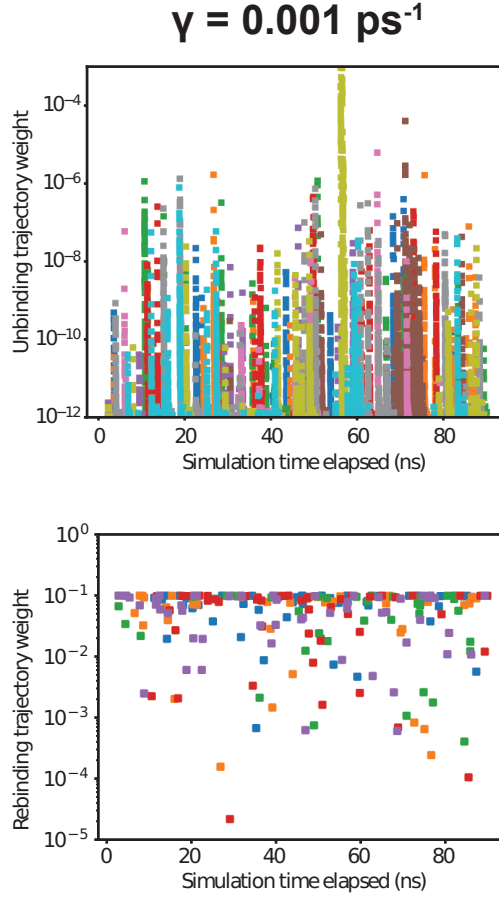


Figure 3.7: Weights of warped walkers in unbinding (top) and binding (bottom) REVO simulations for $\gamma = 0.001 \text{ ps}^{-1}$. Each simulation is shown in a different color. Parameters are the same as those used for higher γ values in the main text.

Table 3.2: Raw (ΔG^0) and corrected (ΔG_{corr}) free energy values using simulation data from three different friction coefficients. Values are in kcal/mol and uncertainties are calculated using propagation of the standard error of the mean.

	ΔG^0 (kcal/mol)	ΔG_{corr} (kcal/mol)
$\gamma = 0.01$	-9.83 ± 0.46	-7.11 ± 0.47
$\gamma = 0.1$	-10.78 ± 0.32	-8.06 ± 0.33
$\gamma = 1$	-10.85 ± 0.34	-8.13 ± 0.36
Chapter 2 ($\gamma = 1$)	-12.1 ± 1.0	-9.38 ± 1.0
Comp. ref. [148]	-	-7.0 ± 0.1
Exp. [153]	-	-4.97 ± 0.02

tions, and are not expected to change as a function of γ . The first term, $-kT \ln f_b/f_u$, was calculated to be 0.74 ± 0.10 kcal/mol, with f_b and f_u taking on values of 0.157 and 0.54

respectively. As described in Section 3.3.1, f_b is the probability of the being in the bound basin given that you are in the unbinding ensemble, which is calculated using the sum of the weights of trajectories in the bound basin, divided by the total sum of the weights of the trajectories considered. The f_b value in particular was lower than expected, indicating that our definition of the bound state might be too restrictive, even though we did account for all symmetry-equivalent conformations in our calculation of f_b .

The second term, $+kT \ln \langle e^{\beta E_{\text{int}}} \rangle_{\text{unb}}$, was calculated to be 1.64 ± 0.002 kcal/mol. This was calculated by determining the electrostatic interaction energies (see Section 3.2.5) for the set of unbound states observed in the rebinding simulations. The expectation value in the correction term again accounted for trajectory weights and was computed using 71428 interaction energy measurements that were selected from the unbound ensemble. The uncertainty was computed as the standard error of the mean of this set of energies. To calculate the third correction term, $-kT \ln \left(\frac{V_{\text{unbound}}}{V_{\text{box}}} \right)$, we directly estimated $V_{\text{unbound}}/V_{\text{box}}$ using the Monte Carlo procedure described in Section 3.3.1. The ratio was computed as 0.56 ± 0.0037 using five batches of 10000 trials each, where the uncertainty is the standard error of the mean across the sets of trials.

Together these three terms sum to 2.72 kcal/mol, which is a significant correction to the binding free energies computed here. Over half of this comes from the residual electrostatic interaction energy between the host and the guest. Note that both the host and the guest have negative charges, and the residual interaction between the two molecules is repulsive. Turning this interaction off *releases* 1.64 kcal/mol of energy, which lowers the free energy gap between the bound and unbound states. The corrected and uncorrected free energies are shown as a function of γ in Figure 3.8. For $\gamma \geq 0.01$ the calculated free energies are almost equal to within standard error and the correction terms significantly reduce the error with respect to the computational reference value [99, 148].

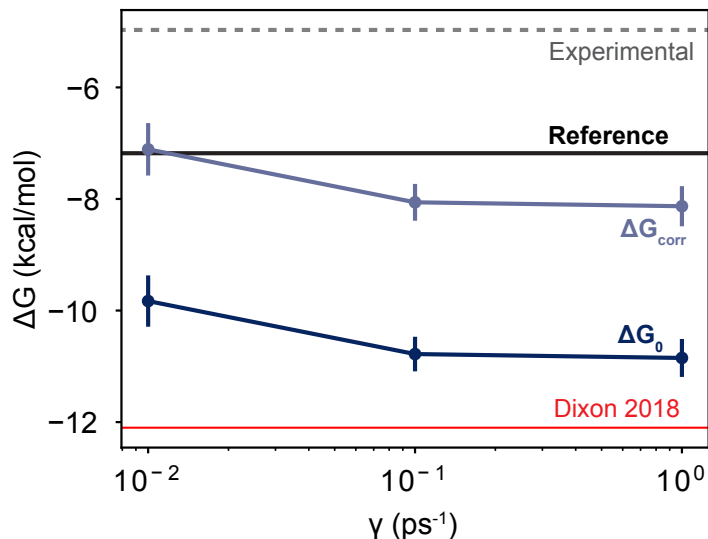


Figure 3.8: Free energies as a function of friction coefficient. The dark blue line shows the uncorrected free energies calculated at three different γ values. The light blue line shows the corrected values, which are shifted upwards by 2.72 kcal/mol. The thin red line shows the value reported in Chapter 2, which employed a friction coefficient of 1.0 ps⁻¹ and used a smaller dataset than is reported here. The black horizontal line shows the value of a computational reference computed using alchemical perturbation, reported in Ref. [148]. The dashed grey line shows the experimental measurement, reported in Ref. [153].

3.4 Discussion and Conclusion

In this study, we sought to better connect the calculation of binding and unbinding rates with the calculation of binding free energies. The rate calculations measured the microscopic fluxes of trajectories from one basin to another. These fluxes can be visualized in an extended history-dependent conformation space, where trajectories change their “color” based on which basin (“bound” or “unbound”) they have most recently visited [85, 86, 87, 88, 89]. The ratio of these rates gives a ratio of two populations: the trajectories that have most recently visited the “bound” basin and the trajectories that have most recently visited the “unbound” basin. The first correction term adjusts this ratio to instead only account for the probability contained within the basins themselves and is particular to rates that are calculated using this history-dependent formalism. The third term can be seen as a volume correction term,

which is used to accurately account for the volume in the unbound state. This is done in other approaches where restraints are used, such as umbrella sampling [154, 155, 156]. In our case the unbound state cannot be easily approximated by a geometric object, such as the volume of a spherical shell.

The second term accounts for residual interactions in the unbound ensemble. This could be used by other approaches that directly determine free energy differences between bound and unbound conformations, such as umbrella sampling. The conventional approach is to define a simulation box that is large enough such that the interactions between the host and guest are negligible in the unbound state. However, this can significantly increase the cost of the simulation. It is worth noting that umbrella sampling results for this system (OA-G6) obtained by Song et al. [156], -8.50 kcal/mol, were also below both the computational benchmark and the experimental value. Their unbound state was defined as a 20 Å distance between an atom in the guest and a dummy atom in the center of the host, which is roughly comparable to our unbound basin of 10 Å of clearance between the host and the guest. Assuming a similar value for the electrostatic correction term, it would have brought their prediction to -6.86 kcal/mol, which is in line with the computational benchmarks [148].

The electrostatic term can also be viewed as a sort of “decoupling” between the host and the guest, and it is warranted to discuss similarities and differences with similar procedures in alchemical free energy methods. They are similar in that we are computing a free energy between two Hamiltonians, one in which an interaction is turned off. We could thus use similar techniques for computing these free energy differences, such as thermodynamic integration [157, 158], BAR [159], MBAR [150, 158], or MM/PBSA [160], although here we effectively use a simple free energy perturbation (FEP) expression [161, 162]. The approaches are different in that we are only considering ensembles of structures where the interactions being turned off are relatively weak. We are assuming here – as is always the case with FEP – that the conformational ensembles of both the host and the guest are highly overlapping between the two Hamiltonians, which considerably simplifies the problem. We also note that

although we employ electrostatic decoupling to compute free energies, our simulations still reveal important information about the (un)binding kinetics and mechanism.

We also examined the role that the Langevin integrator plays in the prediction of kinetic and thermodynamic quantities. In particular, we adjusted the friction coefficient (γ), defined in the Langevin integrator, while maintaining the stability of temperature at 300 K. We did not expect that altering the friction coefficient would have an impact on the calculation of equilibrium quantities. As γ does not appear in the Hamiltonian of the system, it should not affect the probability of a given microstate $P(\mathbf{X})$, which is given by the Canonical probability density $\exp(-\beta U(\mathbf{X}))$. While we did expect it to affect rates, we expected that these effects would offset: that if unbinding was accelerated 10-fold, we would observe the binding process to be sped up by the same factor. However, we observe that the on-rate was very stable as a function of γ , while the off-rate changed slightly. One explanation is that unbinding is much more rare event when compared to rebinding, and estimates of k_{off} were not converged. Lower friction coefficients could be accelerating sampling of these events and making it easier to observe higher probability walkers unbind in our simulations.

Convergence is of utmost priority in weighted ensemble simulations that calculate kinetic quantities. In our previous study, we hypothesized that it was possible that extending the time of the unbinding simulations could capture more high weight walkers exiting from the bound state. Indeed, we observe a higher unbinding flux in this study across all friction coefficients. In Figure 3.4, we observe large upward jumps, for all γ values, even after 40 ns of simulation time per walker, which was sampling limit in our previous study. These upward jumps, as previously described, signify that an exit point was recorded that has a higher weight than previously observed. This highlights the challenges involved in accurate determination of rate fluxes for rare events. It is worth noting that by using our correction terms to account for small unbound volumes and persistent but small electrostatic interactions in the unbound state, we can keep box sizes small, allowing for better convergence of rate fluxes at fixed computational cost.

Of course the binding free energy alone is still an important quantity for drug design [163]. If one is only interested in the absolute binding free energy, calculating it through the ratio of rates is needlessly complicated; free energy is a state function and thus only depends on the endpoints of the binding pathway. The prediction of k_{off} and k_{on} themselves is challenging, since they are not state functions: they depend on the transition path ensemble between the bound and unbound state. Sampling of these physical pathways is a large challenge for MD, largely due to the long timescales of the binding and release processes. Ensuring that the ratio of rates is consistent with binding free energy calculations - as done here - provides an additional, powerful consistency check. In particular, comparing to well-converged computational benchmarks is more useful than experimental quantities, as we avoid an additional layer of uncertainty associated with the force field used to describe the system.

CHAPTER 4

MEMBRANE-MEDIATED LIGAND UNBINDING OF THE PK-11195 LIGAND FROM TSPO

This work was published in the Biophysical Journal volume 120 pages 158-167 in 2021. The work is presented here as published except that the supplemental figures are worked into the text.

4.1 Introduction

The binding affinity of a ligand to its protein target has long been viewed as the key parameter determining its efficacy. However, recent studies have shown that in some protein-ligand systems residence time (RT) correlates more strongly with efficacy than binding affinity [2]. But unlike the binding affinity, RT is not a state function; it depends on the height of the free energy barrier separating the bound and unbound states. In order to rationally design ligands for longer RTs we need to understand the (un)binding mechanism and what molecular interactions occur along the ligand (un)binding pathway.

Previous studies have shown that the Translocator protein (18 kDa) (TSPO) is one such protein where RT is important for predicting efficacy [18]. TSPO is a well-conserved membrane protein, being present in all kingdoms including prokaryotes as well as in the outer mitochondrial membrane of eukaryotes [164]. TSPO has five transmembrane α -helices (TM1-5) along with a small helical region in a 20-residue loop (hereafter denoted as the LP1 region) connecting TM-1 and TM-2 on the cytosolic side (Fig. 4.1A). While in the membrane, TSPO is largely found in a dimeric state [165]. To date, four different structures have been solved for TSPO, for both bacterial [165, 166] and mammalian [167, 168] organisms the former by X-Ray crystallography, the latter by nuclear magnetic resonance (nuclear magnetic resonance (NMR)).

While the structure of TSPO have been solved, its function remains unknown. In humans,

TSPO is highly expressed in steroidogenic tissues, consistent with the hypothesis that it is involved in the regulation of cholesterol transport across the mitochondrial membrane. Indeed, TSPO has been shown to have a high binding affinity for cholesterol [169]. There are other studies linking it to apoptosis [170, 171] and cellular stress regulation in TSPO knockout mice [172, 173], although evidence for this is mixed [174, 175]. Increased TSPO expression has also been observed in cases of neurodegenerative diseases such as Alzheimer’s and Parkinson’s diseases [176]. Relatedly, due to its high expression in areas of inflammation TSPO serves as a biomarker for neurodegenerative disease and brain trauma, and radiolabeled ligands such as [H3]-PK-11195, are commonly used in positron emission tomography (PET) scans [177]. PK-11195 is an isoquinoline carboxamide with no known therapeutic effect [175] and a RT of 34 min in the human TSPO sequence [18, 178].

Molecular dynamics (MD) simulations have been previously performed using a bound TSPO-PK-11195 complex. Researchers recently determined the unbinding pathway of PK-11195 from a rat TSPO model generated from the Protein Data Bank (PDB) 2MGY structure [139]. To generate unbinding paths they used a combination of random accelerated molecular dynamics (RAMD) [179] and steered MD [180] and determined that PK-11195 unbinds into the cytosol through the largely disordered LP1 region (Fig. 4.1A). Unfortunately, this starting structure, determined by NMR, was significantly destabilized by the detergent used in the purification [181, 182]. Also, the methods used to determine the unbinding pathway RAMD have the potential to impart bias on the predicted (un)binding path. Another group performed an induced-fit docking of PK-11195 using Glide [183] with a homology model to resemble the mammalian (mouse) TSPO structure using the PDB 4UC1 *Rhodobacter sphaeroides* structure. They simulated the TSPO-PK-11195 complex for 700 ns and did not observe significant ligand displacement, which is expected due to the extremely long RT of the TSPO-PK-11195 complex.

Here we study the unbinding mechanism for the TSPO-PK-11195 complex, using PDB 4UC1 as the TSPO starting structure [165] and using a weighted ensemble algorithm: Re-

sampling Ensembles by Variation Optimization (REVO) to generate continuous unbinding pathways without perturbing the underlying dynamics [35]. REVO has been previously applied to study ligand unbinding on a series of host-guest systems (Chapter 2 and the trypsin-benzamidine system [35]. In the next section we discuss the methodology used for the simulations: the REVO resampling algorithm, the clustering algorithm used to make the Markov State Model (MSM) and the conformational space networks (CSN)s representation, and rate calculations. In the Results and Discussion we analyze pathways found for dissociation of PK-11195 from TSPO, residues which bound strongly to PK-11195 along the observed pathways, and we compare RTs between different starting poses. We then summarize our findings and discuss how they relate to existing research.

4.2 Materials and Methods

4.2.1 Protein Preparation

The initial TSPO dimer structure is comprised of chains A and B from PDB 4UC1[165]. This x-ray crystal structure comes from the *Rhodobacter sphaeroides* with an A139T mutation to resemble human TSPO. CHARMM-GUI membrane generator[184] was used to place the TSPO complex into a membrane comprised of 174 phospholipids consisting of 53.4% phosphatidylcholine, 28.2% phosphatidylethanolamine, and 18.4% phosphatidylinositol lipids. 10268 TIP3 water molecules were inserted up to a cutoff of 10 Å from the complex and 121 potassium ions and 27 chloride ions were added to reach a salt concentration of 150 mM and to neutralize the system. The system was placed into a rectangular box with dimensions 96.4 Å x 96.4 Å x 91.8 Å. The protein was simulated using the CHARMM36 forcefield [185] and parameters for the PK-11195 ligand were obtained with CHARMM Generalized Force Field (CGENFF) [39, 40].

4.2.2 Docking

Six different PK-11195 poses were used in the simulations. Docking was carried out with Extra Precision (XP) by using Schrödinger Glide[186]. The center of mass (COM) of PK-11195 was placed at the COM of the bound Protoporphyrin IX in the chain A monomer of TSPO protein from PDB 4UC1 without any constraints. The XP docking yielded four poses (D1-D4) and the XP Gscores for the resultant poses can be seen in Fig 4.1B. A homology model of PK-11195-bound TSPO (Pose R) was generated by Xia *et al.* as a Rosetta comparative model of the mouse TSPO structure constructed using TSPO structures from *Mus musculus* (PDB 2MGY[167]), *R. sphaeroides* (PDB 4UC1[165]), and *Bacillus cereus* (PDB 4RYI [166]); more details found in Ref. [187]. The TSPO monomer bound to PK-11195 from this model was then aligned to chain A of the 4UC1 structure using PyMol 1.7.2.1 [188], and the ligand coordinates from the D1 pose were changed to reflect the new pose. The 4RYI pose was generated by X-Ray crystallography and the coordinates of the PK-11195 ligand were added to the 4UC1 structure in the same way as pose R. The system’s energy was minimized using a series of constraints with scripts provided by CHARMM-GUI for all poses. The molecular structure for each pose is shown in Fig. 4.1B and pose view diagrams are shown in Fig. 4.2.

4.2.3 Molecular Dynamics

All MD simulations were performed using OpenMM[114] v7.1.1. The time step for every simulation was 2 fs. To enforce constant temperature and pressure, a Langevin heat bath was used with a set temperature of 300K and a friction coefficient of 1 ps^{-1} was coupled to a Monte Carlo barostat set to 1 atm and volume moves were attempted every 50 time steps. The non-bonded forces were computed using the CutoffPeriodic function in OpenMM with a cutoff of 10 Å. The atomic positions and velocities are saved every 15,000 time steps, or every 30 ps of simulation time, which is the resampling period (τ) used here.

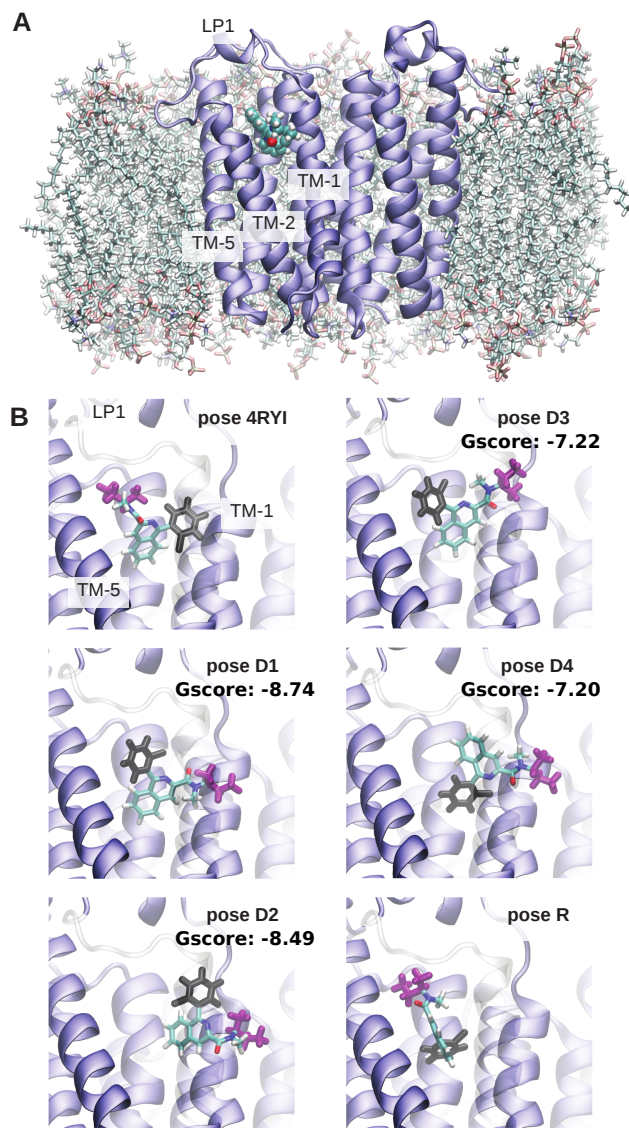


Figure 4.1: TSPO-PK-11195 system. (A) Front view of the TSPO dimer in the membrane with PK-11195 bound. (B) All six starting poses are shown from the side view, along the inter-dimer axis. To compare poses, two moieties of PK-11195 are colored in black (o-chlorophenyl) and magenta (1-methylpropyl), with the rest of the molecule colored according to atom name. TM-2 is shown as transparent for clarity.

4.2.4 REVO Resampling

To observe long timescale unbinding of PK-11195, we used a variant of the weighted ensemble algorithm: REVO[35]. In this algorithm, we perform unbiased MD simulation on 48 separate trajectories in a parallel fashion. Each of these trajectories (called "walkers") has

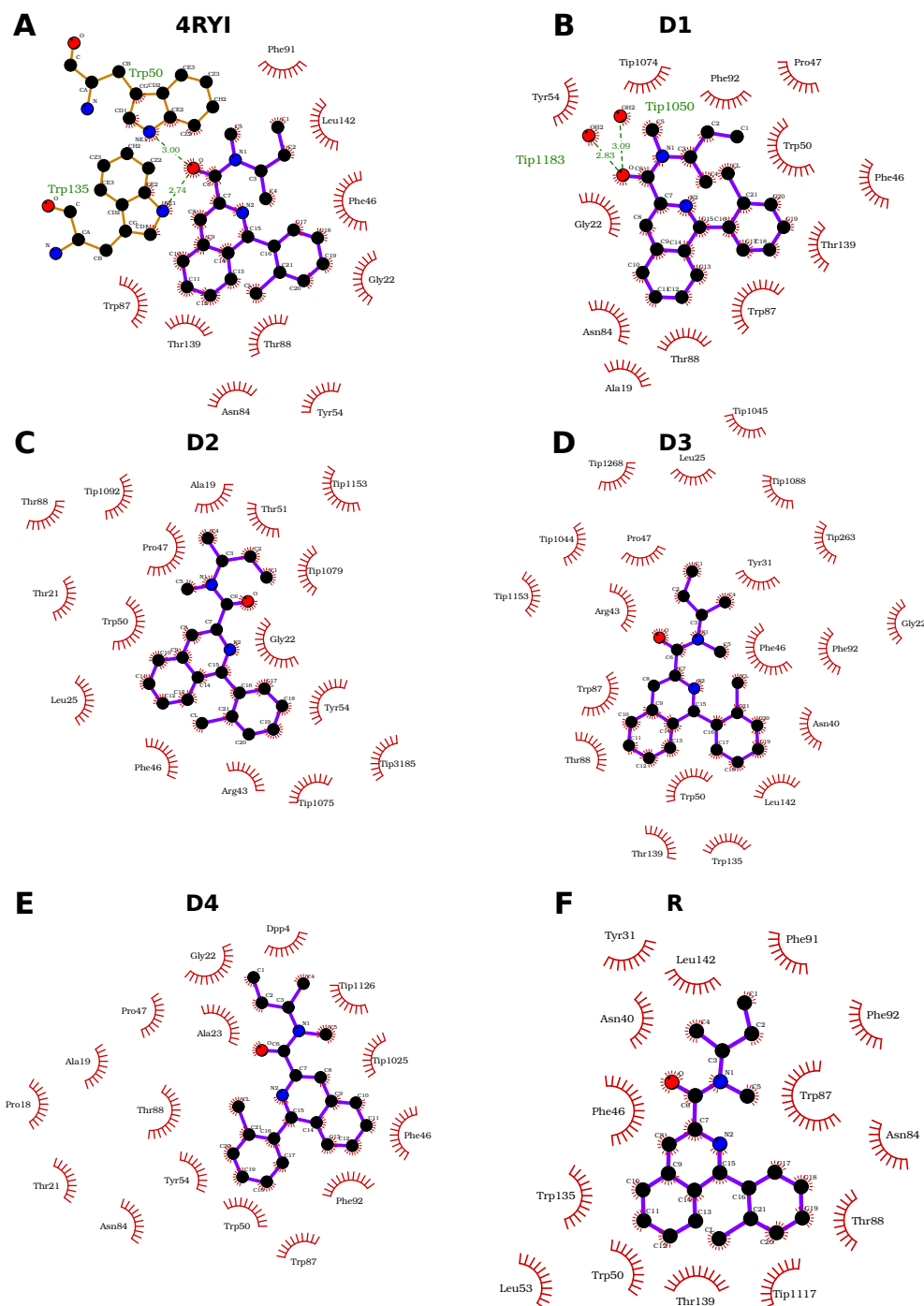


Figure 4.2: Protein-ligand interaction plots for the six starting conformations. The red suns indicate that the residue has a hydrophobic contact with PK-11195. The green dashed lines show hydrogen bonds.

a statistical weight (w) that governs the probability with which it contributes to statistical observables. With periodicity τ , a resampling procedure is performed, where similar walkers are merged together and unique walkers are cloned, as defined by a distance metric. During cloning, weights are split, and during merging, weights are added, to ensure conservation of probability.

Below we briefly describe the REVO method, focusing on the details of its application in this work. More information on the algorithm can be found in previous work [35]. In REVO, merging and cloning is done to maximize a variation function:

$$V = \sum_i V_i = \sum_i \sum_j \left(\frac{d_{ij}}{d_0} \right)^\alpha \phi_i \phi_j, \quad (4.1)$$

where d_{ij} is the distance between walker i and walker j determined using a distance metric of choice. For these simulations the distance metric used was the root mean square deviation (RMSD) of the PK-11195 atoms between each walker, following alignment to a selection of binding site atoms in TSPO. The exponent α is used to modulate the influence of the distances in the variation calculation and was set to 4 for all simulations. $d_0 = 0.148$ nm is a characteristic distance used to make V dimensionless and to normalize the variance for comparison between different distance metrics. ϕ is a novelty and here is defined as:

$$\phi_i = \log(w_i) - \log\left(\frac{p_{\min}}{100}\right). \quad (4.2)$$

The minimum weight, p_{\min} , allowed during the simulation was 10^{-12} . The walker that is selected for cloning is the one that has the highest V_i and the resultant weight of the clones is larger than p_{\min} . The two walkers selected for merging are at most 2 Å away, have a combined weight lower than the maximum allowed weight $p_{\max} = 0.1$, and is the walker pair j, k that minimizes the variation loss (V_{loss}) defined as:

$$V_{loss} = \frac{V_j w_k + V_k w_j}{w_j + w_k}. \quad (4.3)$$

Once the walkers (i, j, k) are selected, the new variation is calculated: if it increases, then these operations are performed and another (i, j, k) is proposed; if it decreases then resampling for that cycle is terminated and a new cycle of MDis performed. Three simulations

were run for each docked pose using 48 walkers and 1200 cycles, for 1.728 μ s of simulation time per simulation. In total each pose was simulated for 5.184 μ s.

4.2.5 Boundary Conditions

The overall goal of the simulations was to determine the pathways along which PK-11195 can transition from the initial starting poses to an unbound state. During the simulations, we defined PK-11195 as being unbound when the minimum distance between the ligand and TSPO was at least 10 Å. When the ligand crossed this boundary, the weight is recorded and the walker was "warped" back to the initial conformation. The structure recorded before warping is known as an exit point. When the walker warps back, the atomic positions and velocities are reset to their initial values before the simulation began. The walker weight does not change as a result of warping.

4.2.6 Clustering and Network Layout

The trajectory frames of all 18 REVO runs were clustered together using the MSMBuilder 3.8.0 python library. The frames were featurized using a vector of atomic distances between TSPO and PK-11195 atoms initially within 8 Å of each other from the 4RYI starting pose for a total of 7527 distances. A k-centers clustering algorithm was used to generate 2000 clusters using the featurized space and each frame was assigned to a cluster. The clustering was done using the Canberra distance metric. A count matrix describing the cluster-cluster transitions was calculated for a lag time of 30 ps.

We then construct a CSNs from the count matrix, which is a graphical representation of the transition matrix. Each node, representing each row of the transition, and the edges, representing non-zero off diagonal elements of the transition matrix, were determined using the CSNAnalysis package [121]. Gephi 0.9.2[119] was used to visualize the CSN. The size of each node is proportional to the statistical population of the cluster. For visualization, the smallest node was set to be 20 times smaller than the largest node. The layout of the network

was determined using a force minimization algorithm, Force Atlas included in Gephi. The algorithm repulses nodes that are not connected and attracts nodes that are connected via an edge. The strength of the attractive force is proportional to the weight of the edges. The directed edge weights were values between 0.1 and 100 as determined by $w_{ij} = 100p_{ij}$, where p_{ij} is the transition probability of cluster i transitioning to cluster j . Unidirectional edge weights were then determined using the average between the two directed edge weights. Force Atlas was applied twice. The first minimization was done without adjusting for node sizes, allowing the nodes to overlap. The second minimization adjusted for the node size and prevented overlap. For visualization, all edges are shown with a uniform line weight.

4.2.7 Quantifying Unbinding Pathways

Upon analysis of the simulation results, the only unbinding pathways observed in our simulations were PK-11195 dissociating through pairs of transmembrane helices. We therefore introduce the coordinate Q_{ij} which measures the minimum x - y distance from the COM of PK-11195 to the line formed by the COMs of helices i and j to measure the dissociation progress of PK-11195 into the membrane. Negative values indicate the COM of the ligand is closer to the center of the helical bundle, and positive values indicate the COM is closer to being fully dispersed in the membrane. All six poses had trajectories where PK-11195 traveled between transmembrane helices 1 and 2 and only pose R had trajectories where PK-11195 went between transmembrane helices 2 and 5. For pose R analysis we separate the conformations according to which value (Q_{12} or Q_{25}) is largest. Projections onto a given Q value will only use conformations for which that Q value is the largest.

4.2.8 Calculating Non-bonded Energies

We calculated the non-bonded interaction energies (E_{int}) by:

$$E_{int} = V_{LJ} + V_{ES}, \tag{4.4}$$

where V_{LJ} is the Lennard-Jones potential energy and V_{ES} is the potential energy from electrostatic interactions. The Lennard-Jones interactions were determined using a 12-6 potential given as:

$$V_{LJ} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right], \quad (4.5)$$

where r is the atomic distance between atoms, σ is the inter-atomic distance at which the potential is 0, and ϵ is the depth of the potential well. To calculate σ and ϵ we used the Lorentz-Berthelot combining rule. There was a hard cutoff distance of 10 Å when calculating the Lennard-Jones potential. The electrostatic energy was calculated using:

$$V_{ES} = \frac{1}{4\pi\epsilon_0} \frac{Q_i Q_j}{r_{ij}}, \quad (4.6)$$

where Q_a is the charge of atom a , r_{ij} is the interatomic distance between atoms i and j , $\epsilon_0 = 8.854 * 10^{-12} \frac{F}{m}$ is the permittivity of free space in farads per meter. The specific σ , ϵ , and Q , for each atom type was provided by CHARMM-36 parameter files obtained through CHARMM-GUI. Two sets of non-bonded energies were calculated: between PK-11195 and TSPO, and between PK-11195 and lipids in the membrane.

4.2.9 Calculating Off-Rates and Mean First Passage Times using Hill Relation

The rates are calculated using the flux of trajectories into the unbound basin, also known as the Hill relation[117, 85, 88], defined as

$$k_{off} = \frac{\sum_i w_i}{T}, \quad (4.7)$$

where w_i is the weight of the walker entering the unbound basin, and T is the total simulation time. During the simulations the unbound basin was defined by the 10 Å boundary condition. However, although many walkers had dissociated into the membrane, no walkers made it to the boundary. Therefore, to obtain estimates of unbinding rates, after the simulations were completed the unbound basin was redefined using a minimum distance of 5 Å as we found negligible interaction energy between PK-11195 and TSPO at this distance (Fig. 4.3). In

our simulations we observed a total of 2285 instances of trajectory crossings into the 5 Å unbound basin. This is broken down by starting pose as follows: 4RYI (47), D1 (4), D2 (1804), D3 (278), D4 (152) and R (0). In our analysis, once a walker entered the unbound basin, we ignored all future trajectories associated with that walker. This was done to prevent double-counting of unbinding transitions. The mean first passage time (MFPT), synonymous with the RT, was calculated as

$$\text{MFPT} = \frac{1}{k_{\text{off}}}. \quad (4.8)$$

The uncertainty of off-rates and MFPT for each pose is the standard error across each set of simulations.

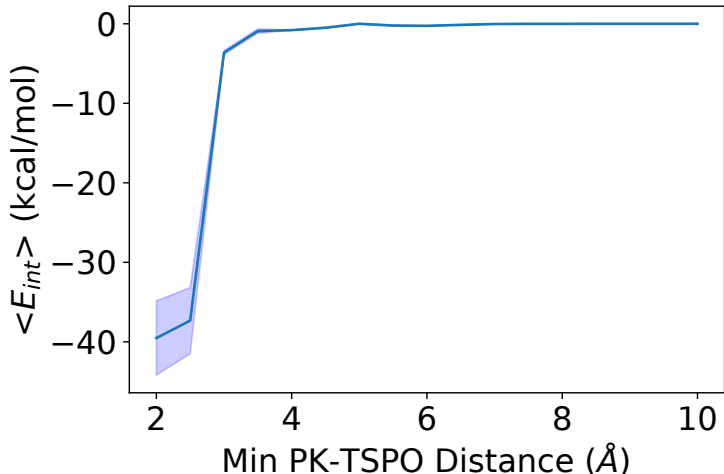


Figure 4.3: The energy of non-bonded interactions between PK-11195 and TSPO as a function of minimum distance between PK-11195 and TSPO.

4.2.10 Calculating Mean First Passage Times using Markov State Models

We create transition matrices, $T(\tau)$, for various lag times (τ) using the cluster identities from the CSN and tracking walkers through merging and cloning operations in the REVO resampler. We alter these matrices to include a probability sink for states that are unbound, defined as when PK-11195 is at least 5 Å away from the TSPO dimer. We run a Markov

chain simulation for a given starting pose and lag time by initializing a probability vector, P , where all of the probability starts at the state of a given starting pose. To progress the simulation we use the following: $P_k = P_0 T(\tau)^k$ where P_0 is the initial probability vector, and P_k is the probability vector after k time steps. We continue the simulations until all the probability accumulates in the unbound basin. We then calculate the MFPT using the following formula:

$$\text{MFPT} = \sum_k (p_k - p_{k-1}) \left(\frac{t_k + t_{k-1}}{2} \right) \quad (4.9)$$

where p_k is the probability of being unbound at time step k and t_k is the time associated with time step k . We repeat this for all initial poses and lag times to determine MFPT as a function of lag time.

4.2.11 Selecting Poses for Straightforward MD Simulations

To strengthen the accuracy of our Markov state model, we run straightforward simulations at weak points in the network. To determine these weak points, we randomly multiplied the elements of a row on the transition matrix with numbers drawn from a Gaussian distribution with a mean (μ) at 1 with a standard deviation (σ) of 0.2 and we re-normalized the row after perturbation. We rerun the Markov chain simulations to calculate the MFPT. To get a sense of how consistently the cluster alters the MFPT, we randomly perturb the transition matrix 10 times independently. Weak points in the network are determined by the clusters whose perturbations affect the MFPT the most, using the following formula: $\delta_{\text{MFPT}}/\overline{\text{MFPT}}$, where δ_{MFPT} and $\overline{\text{MFPT}}$ are the standard deviation and average of the perturbed MFPT values, respectively. For two poses, this ratio was greater than 0.2; we identified these clusters as weak points and reran straightforward MD simulations from the highest weighted structure in that cluster. From each weak point we launched 144 independent straightforward MD simulations for a length of 500 cycles (15 ns). In addition we launched trajectories from high lipid accessible surface area (LASA) clusters in the central unbound region and each

of the high-LASA states originating from pose R. In total we ran 10.8 μ s of supplemental trajectories to bolster our Markov state model.

4.3 Results

4.3.1 PK-11195 Unbinding Pathway

We comprehensively studied the TSPO-PK-11195 interaction landscape using a set of REVO simulations initialized at six different starting poses (Fig. 4.1B), simulating 5.184 μ s per pose. After the simulations were completed, all frames were clustered together into a CSN shown in Fig. 4.4, where each node represents a PK-11195 pose and the edges reveal which poses interconvert in our simulations within a 30 ps lag time. All of the starting poses form a connected network, though pose R is only connected via two low probability edges to the pose 4RYI ensemble (Fig. 4.6). The 4RYI pose is similarly connected to pose D4, but is also connected to the other docked poses via the high LASA clusters. It is worth noting that both pose 4RYI and pose R were the only poses that were not designed for this specific protein structure and were instead inserted from other protein structures after alignment. Consistent with this fact, both of these regions in the CSN do not show accumulation of probability into one or more high-probability states. Instead we observe a broader distribution among many low probability states indicating a lack of a local funneling in the energy landscape. Interestingly, all of the docked poses (D1-D4) show at least one high-probability state, although this is not necessarily at the initial docked pose itself, indicating that some relaxation is required from the docked poses to reach the true local minima.

Contrary to what was observed in previous work [139], PK-11195 did not dissociate into the solvent via the LP1 region: it instead dissociates into the membrane. The CSN shows that all the poses, besides pose R, connect directly to the unbound states, shown in yellow and orange, where PK-11195 is fully dissolved into the lipid membrane. In all of these pathways, PK-11195 exits between TM-1 and TM-2. The pose R trajectories show two different pathways that have a moderate LASA – one between TM-1 and TM-2 and another

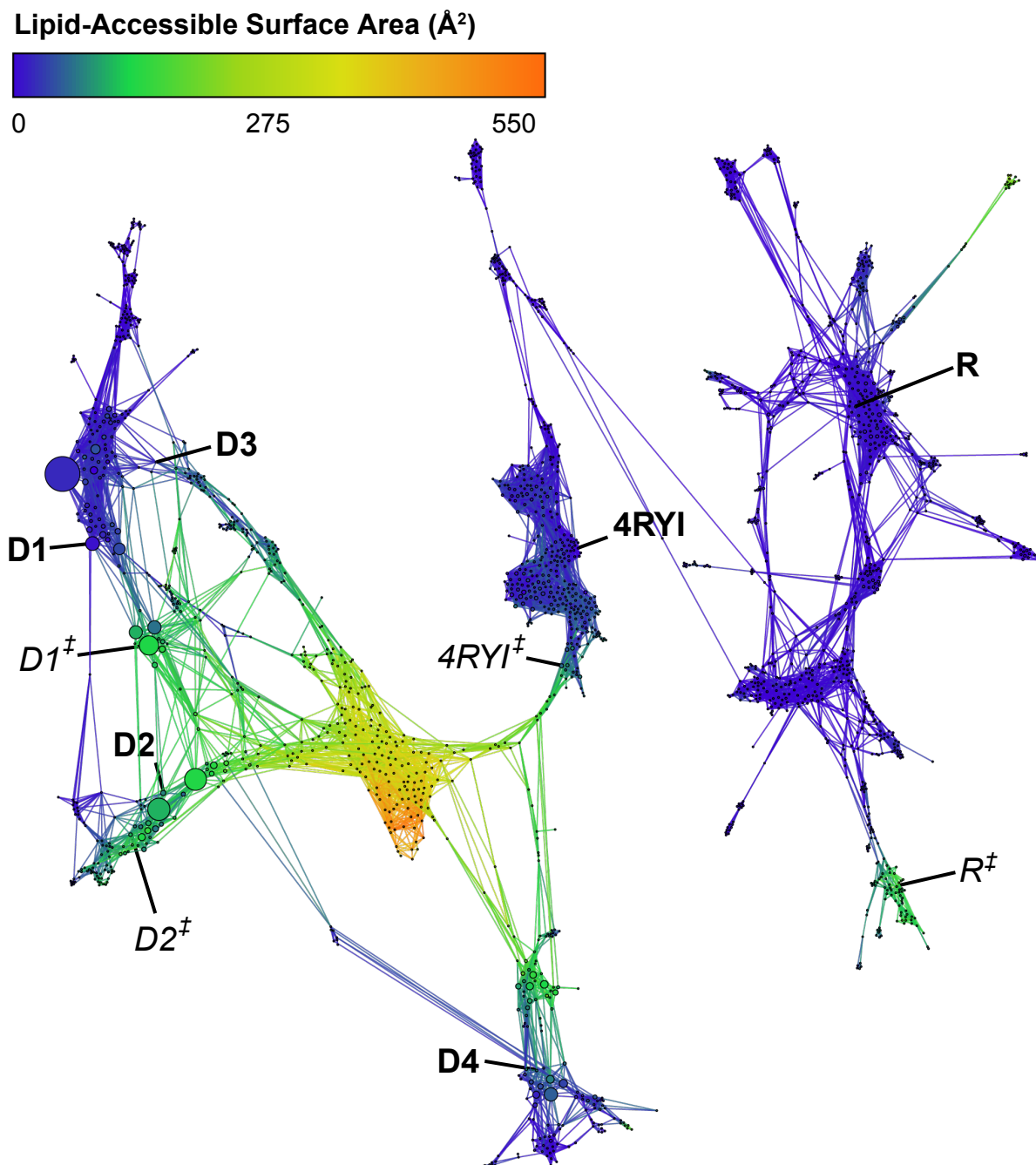


Figure 4.4: Combined CSN of all REVO simulations from each starting pose. Each node in the network represents a cluster of ligand poses and is sized according to the cluster weight. Nodes are connected by edges if the ligand poses are observed to interconvert in the REVO trajectory segments. Nodes are colored according to the LASA. Starting poses are marked in bold and transition state poses shown in Fig. 4.5D are marked in italics.

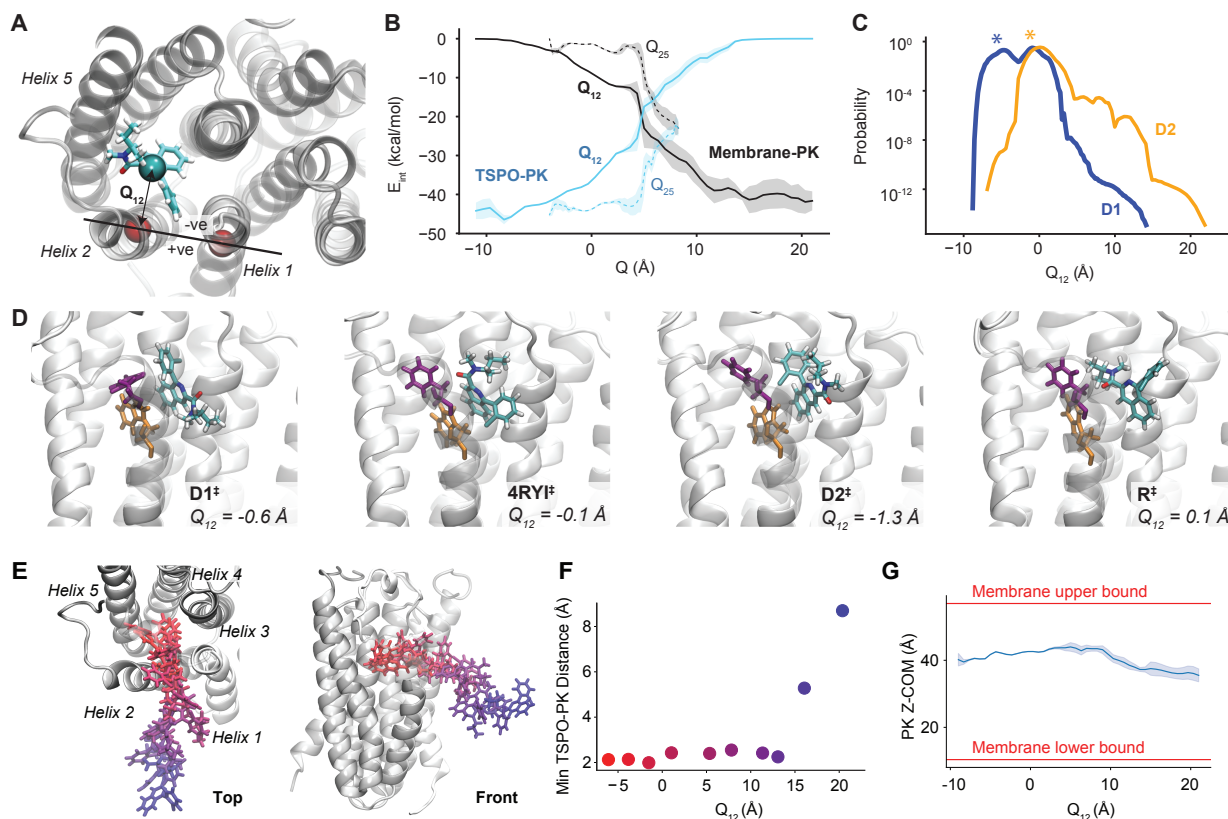


Figure 4.5: Analysis of membrane-mediated exit paths. (A) The coordinate Q_{ij} is defined as the x - y distance between the center of mass of PK-11195, shown as sticks and colored by atom type, and the line that connects the centers of mass of helix i and helix j . LP1 is not shown here for clarity. (B) The expectation values of the interaction energy between PK-11195 and TSPO (blue) and between PK-11195 and the membrane (black) are shown as a function of Q . In each case the solid line shows Q_{12} and the dashed line shows Q_{25} . The shaded region indicates the standard error over the ensemble of measurements at each Q value. (C) Probability curves projected onto Q_{12} for simulations initialized in Pose D1 (blue) and D2 (orange). Q_{12} values of the starting structures are marked with (*). (D) Poses from transition pathways with $Q \approx 0$. These poses are also labeled in the CSN of Fig. 4.4. Phe46 is shown in purple and Trp50 is shown in orange. (E) A set of poses along the Q_{12} pathway colored from bound (red) to unbound (blue). Top view is shown on the left and a front view is shown on the right. (F) The minimum PK-11195-TSPO distance and the Q_{12} value is shown for each pose in panel (E). (G) The z COMposition as a function of Q_{12} . The red lines indicate the upper and lower bounds of the membrane as defined by the maximum and minimum z coordinate of the lipid membrane.

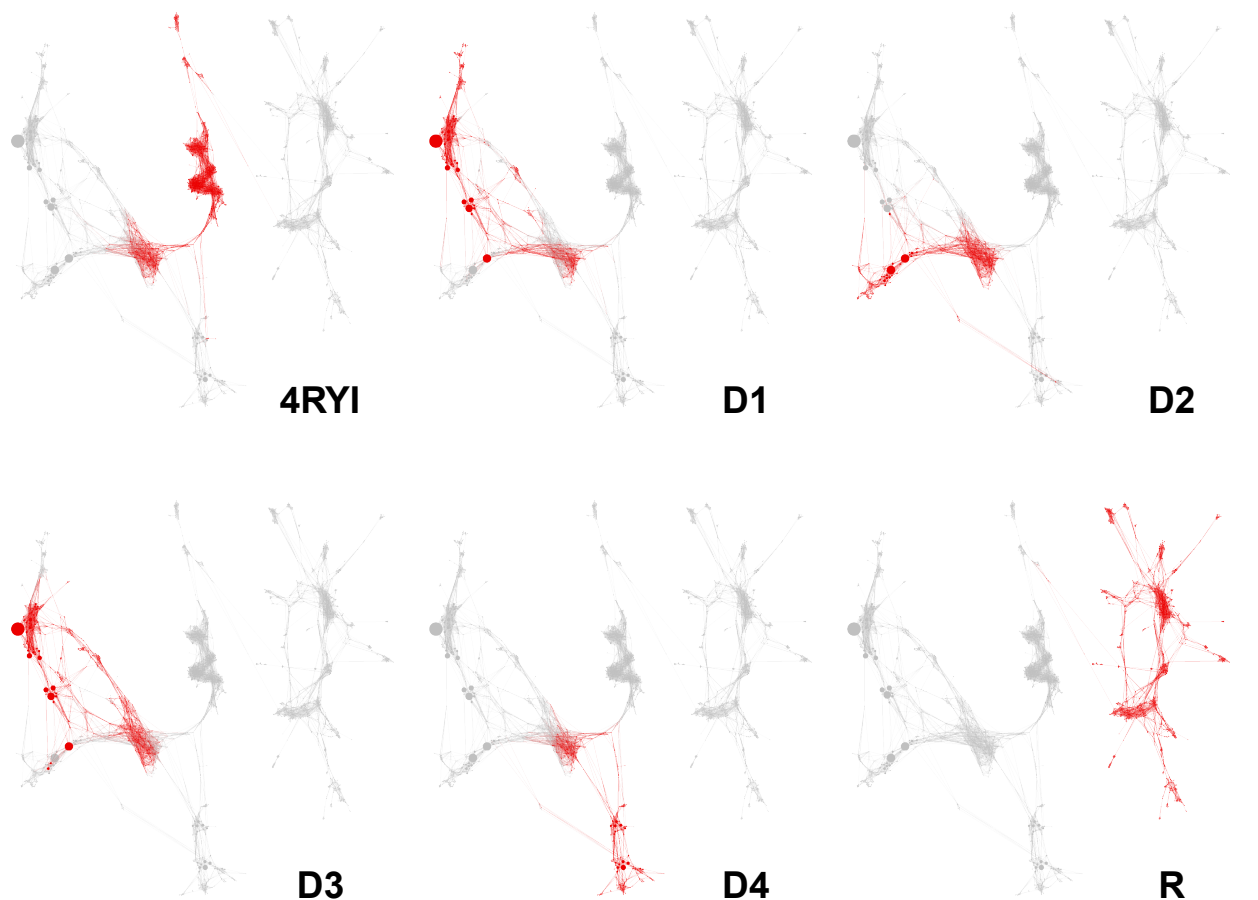


Figure 4.6: CSN networks indicating the clusters that were observed from each initial pose. Red nodes indicate the simulations observed a TSPO-PK-11195 conformation that was clustered into that node.

between TM-2 and TM-5 – where PK-11195 forms direct interactions with membrane lipids.

We introduce the coordinate Q_{ij} , which measures the minimum x - y distance from the center of mass of the ligand to the line connecting the centers of mass of helix i and helix j (Fig. 4.5A). Negative Q values indicate the ligand is within the helical bundle and positive values indicate the ligand is outside the bundle. This provides a basis to compare between different pathways and a means of obtaining general information about membrane-mediated ligand unbinding pathways. Fig. 4.5B compares the TSPO-PK-11195 interaction energy (E_{int}) with membrane-PK-11195 interaction energy. In the Q_{12} pathway (solid lines), PK-11195 interacts more closely with the lipid membrane than TSPO after about 5 Å. For

the Q_{25} pathway (dashed lines) this crossover occurs at 7.5 Å. The difference is due to differences in the orientation of PK-11195 along the two pathways. Fig. 4.5D shows the transition states labelled in Fig. 4.4 where the Q values are approximately equal to zero along each dissociation pathway. We see that each structure is still heavily informed by its starting pose, with very different PK-11195 orientations. Fig. 4.5C shows probability distributions projected onto Q_{12} for starting poses D1 and D2. This shows that although D1 started further backward on the unbinding pathway, the simulations discovered another high-probability basin around $Q_{12} = 0$, which can also be seen by the high-probability states around D1[‡]. A representative Q_{12} dissociation pathway is shown and analyzed in Fig. 4.5E and 4.5F. Note that while the Q_{12} value increases steadily along the pathway, the minimum distance between TSPO and PK-11195 (used to define the unbound state) rises rapidly only as PK-11195 reaches a Q_{12} of about 15 Å. Additionally, we track the PK-11195 center of mass as a function of Q_{12} (Fig. 4.5G). Once it gets fully dissociated into the membrane, PK-11195 does not travel closer towards the solvent in either direction. Rather it interacts strongly with the hydrophobic tails and remains at approximately the membrane midpoint over the course of our simulations.

We also measure interaction energies between PK-11195 and individual residues for all residues on TM1, TM2, TM5 and LP1 (Fig. 4.7-4.10). Early in both the Q_{12} and Q_{25} pathways, PK-11195 strongly interacts with aromatic residues Phe46 and Trp50 forming π - π interactions. These aromatic residues with long side chains follow PK-11195 along the unbinding pathway, which is observed by plotting the Q value of individual residues as a function of Q -PK-11195 (Fig. 4.11 and Fig. 4.12). Interestingly, this phenomenon occurs for smaller amino acid side chains as well: Gly22 and Pro47 both change Q value significantly over the Q_{12} pathway, indicating significant distention of the helices during unbinding.

Finally, we investigated the similarity of the PK-11195 conformations within each cluster with respect to the dihedral angles along four different rotatable bonds (see Fig. 4.13-4.15). The standard deviation for all the angles is generally low (below 85 degrees), however there

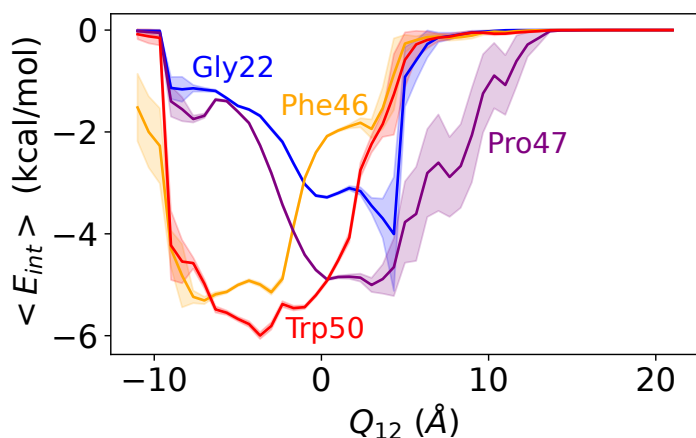


Figure 4.7: Expectation value for E_{int} as a function of Q_{12} . The lines are colored by residue. Only residues who have a minimum interaction energy below -3.5 kcal/mol are shown. The standard error is shown in the lighter shaded regions.

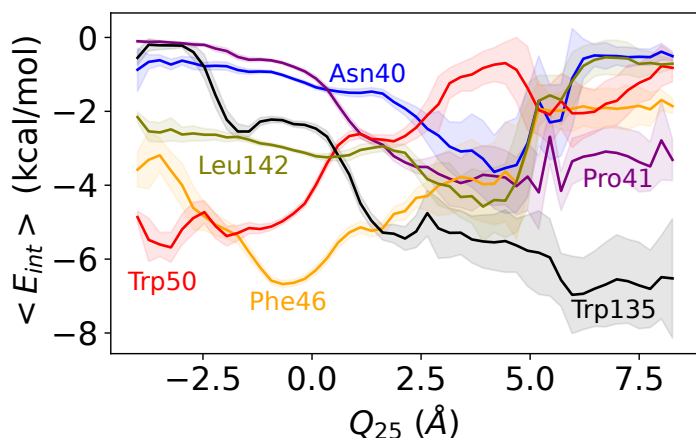


Figure 4.8: Expectation value for E_{int} as a function of Q_{25} . The lines are colored by residue. Only residues who have a minimum interaction energy below -3.5 kcal/mol are shown. The standard error is shown in the lighter shaded regions.

are clusters in high LASA regions on the network that have a higher standard deviation. This indicates that PK-11195 has more degrees of freedom when it comes to rotation when it is within the membrane. However, when looking at the overall angle range for the network clusters, there are several clusters with a high overall range, indicating that different ligand conformations are occasionally being clustered together. In particular, rotatable bond 1

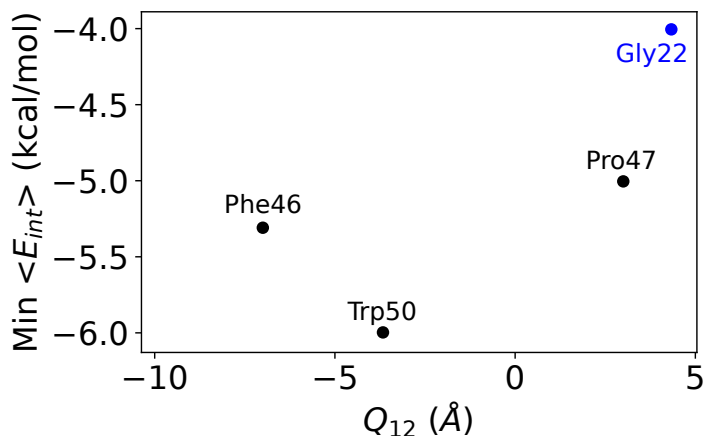


Figure 4.9: The residues with the strongest non-bonded interactions with PK-11195 on the Q_{12} pathway. This summarizes the curves in Fig. 4.7, plotting the minimum E_{int} against the Q_{12} value for which this minimum value is observed. The colors indicate the region of TSPO, blue for residues on TM-1 and black for residues on TM-2. Only residues with a non-bonded energy below -3.5 kcal/mol are shown.

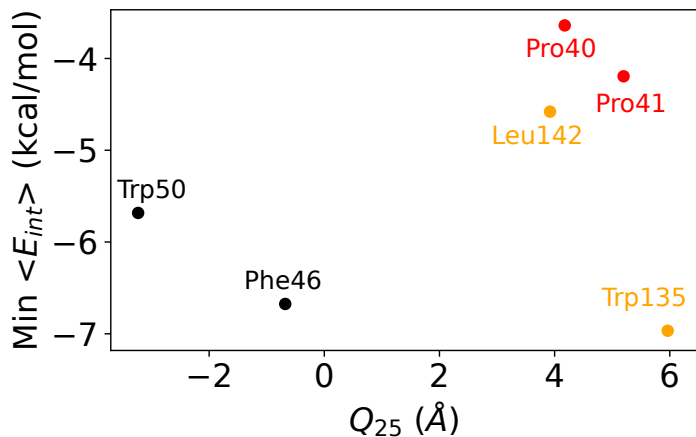


Figure 4.10: The residues with the strongest non-bonded interactions with PK-11195 on the Q_{12} pathway. This summarizes the curves in Fig. 4.8, plotting the minimum E_{int} against the Q_{25} value for which this minimum value is observed. The colors indicate the region of TSPO, red indicates residues on the LP1 loop, black for residues on TM-2 and orange for residues on TM-5. Only residues with a non-bonded energy below -3.5 kcal/mol are shown.

has a range of 90 degrees or higher for most states, and rotatable bond 2 has a range over 150 degrees in the D1-D3, D4, and R basins as well as in the states where PK-11195 has

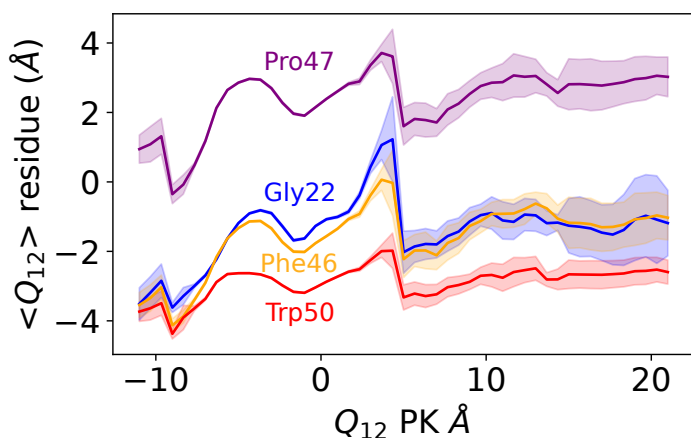


Figure 4.11: Residues moving along with the ligand during dissociation. Expectation values of Q_{12} for individual residues are shown as a function of the Q_{12} of PK-11195.

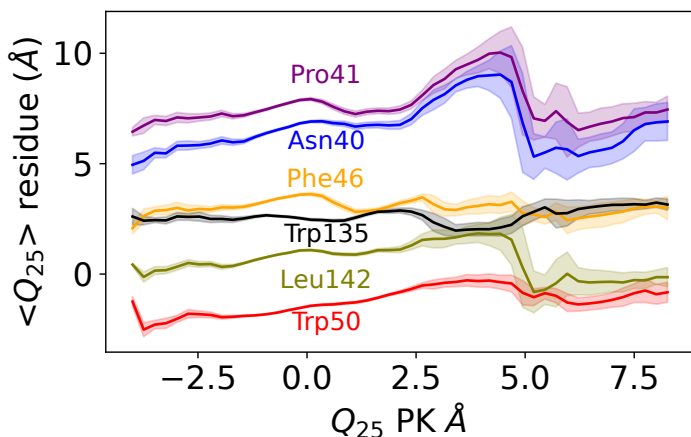


Figure 4.12: Residues moving along with the ligand during dissociation. Expectation values of Q_{25} for individual residues are shown as a function of the Q_{25} of PK-11195.

dissociated into the membrane. Therefore, it is likely that the distance metric, defined as a set of atomic distances from PK-11195 to the TSPO binding site, is good at distinguishing the PK-11195 location but not necessarily good at defining the internal coordinates of PK-11195. It is thus possible that the clustering procedure introduced some unphysical connections and the network should be seen as representing an upper bound of the connectivity between the bound states.

4.3.2 PK-11195 Rates and Residence Times

We directly estimate the unbinding rates (k_{off}) by summing the weights of the unbinding trajectories and we calculate the MFPT by inverting the unbinding rate for each starting pose (Fig. 4.16A). Pose D2 had a high unbinding flux and a predicted MFPT of less than 0.02 s, indicating a clear lack of stability with respect to the other poses. Poses D3 and D4 had predicted MFPTs of 2.6 and 4.1 minutes, respectively, still lower than the experimental measurements; these estimates are likely to continue to decrease with further simulation time. Poses 4RYI and D1 had MFPT estimates near or above the experimental MFPT (28 and 260 min, respectively). No unbinding events were observed for Pose R, implying an even longer MFPT than 260 min.

One of the issues with performing simulations via weighted ensemble is ensuring the simulations converge. A lack of convergence introduces additional uncertainty into k_{off} and MFPT calculations. To address this issue, we launch a set of Markov chain simulations using the transition matrix that constructed the CSN. Due to the unphysical connections between various clusters, we constructed pose-specific networks by only including states that were visited by trajectories that were generated from a given starting pose. We again find that the D2 pose has a low MFPT, though 2 orders of magnitude less than that calculated by the Hill relation. Calculating the MFPT using the MSM showed that all poses besides D2 were on the same order of magnitude as the experimental RT, and were within an order of magnitude of that determined by the Hill Relation. Since there were no trajectories starting from pose R that entered the unbound basin, a MFPT could not be computed for this starting pose without additional simulations.

The accuracy of the MFPT calculations however, assumes that the transition matrix determined from the simulations has converged. To test for convergence, we run additional straightforward simulations at the bottlenecks of the network and rerun the MFPT calculations by combining the old and new trajectory data. Two such bottlenecks were identified: the connections between pose R and pose 4RYI as well as between poses D2 and D4. In

order to better sample the unbound state, we also ran straightforward MD simulations from high-LASA poses in the Q_{12} and Q_{25} pathways that were seen by pose R as well as the most probable state in the unbound basin. We then reclustered and remade the CSN network to include the new frames (Fig. 4.17). Several connections were formed between pose 4RYI and pose R, which also gained connections to the other poses after reclustering. Additionally, the most probable region in the network was once again the D1-D3 basin as determined by the steady state probabilities of each state.

With the addition of the straightforward simulations, we recalculated the pose-specific MFPTs from each starting pose. The new simulations did not show pose R progress enough along the unbinding pathway to enter the unbound basin, and therefore we again could not compute a RT for this pose. The MFPT for pose D2 increased by five orders of magnitude in the new D2 MSM, but it is still the pose with the fastest unbinding pathway. This is likely a result of reclustering after the addition of the new trajectories. Accordingly, when we recalculate a new MSM that uses the new clusters but excludes the new trajectories from the transition matrix, we find only an additional slight increase of the D2 MFPT from 0.13 to 0.16 minutes.

Poses 4RYI, D1, and D4 all had MFPTs on the same order of magnitude as the original MSM simulations and D3 had an MFPT that was lower by a factor of ~ 10 . The lack of change between the MFPT calculations for slow unbinding events indicates that the original MSM had converged enough to produce a reliable estimation for those poses. In terms of stability, D2 consistently has the fastest unbinding events and is consistently the most unstable pose we simulated. Poses 4RYI, D1, D3, and D4 all have similar levels of stability, as can be seen by their similar RTs. Due to the lack of unbinding events for pose R we can not measure how stable it is in comparison to the other starting poses, but we can say that starting in the pose R basin is more stable than the other poses we simulated.

4.3.3 PK-11195 Transition State

Our final goal was to determine the location of the transition state along the unbinding pathway. Fig 4.17B, shows the committor probability of each state in the final network. The vast majority of the states have a near zero committor to the unbound state. Only once PK-11195 is dissociated into the membrane does the committor probability begin to significantly increase. We built an ensemble of transition states using the centroid structures for the two nodes with committor probabilities between 0.4 and 0.6. In this way, we estimate that the transition state – where the committor equals 0.5 – occurs when PK-11195 has begun dissociating into the membrane and has reached a Q_{12} of ~ 10 Å. For these states we find that the non-bonded interaction energy between TSPO and PK-11195 is roughly -5 kcal/mol (compared to -45 kcal/mol in the bound state), whereas the interaction energy between PK-11195 and the lipid membrane has increased to -40 kcal/mol at this Q_{12} (Fig. 4.5B).

To ensure that this result is not affected by any unphysical connections between bound poses, we also calculated the committor probability for each state in the pose-specific networks (Figs. 4.18-4.22). We determined pose-specific transition states for each of the initial poses that had unbinding events (e.g. all except pose R) and found that they were all located in the membrane after PK-11195 had dissociated from TSPO. This confirms the results from the committor probability analysis of the full network. Further, these transition states all demonstrated a mix of direct PK-11195-TSPO interactions and PK-11195-lipid interactions. Together these results suggest that the membrane presents a physical barrier that acts to trap PK-11195 near TSPO and forms the rate-limiting step of PK-11195 dissociation into the membrane.

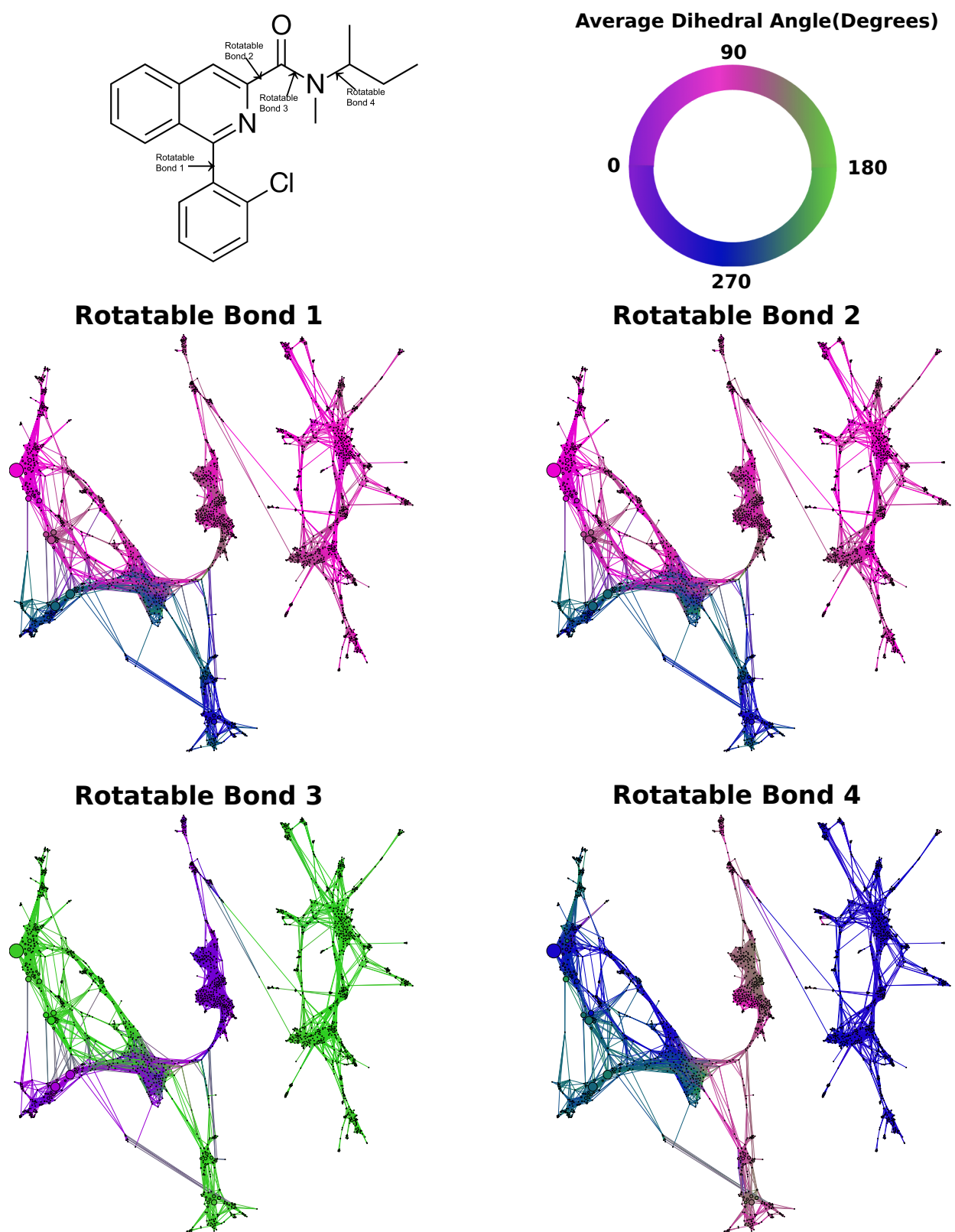


Figure 4.13: The average dihedral angles for the MSM states for four different rotatable bonds on the PK-11195 ligand.

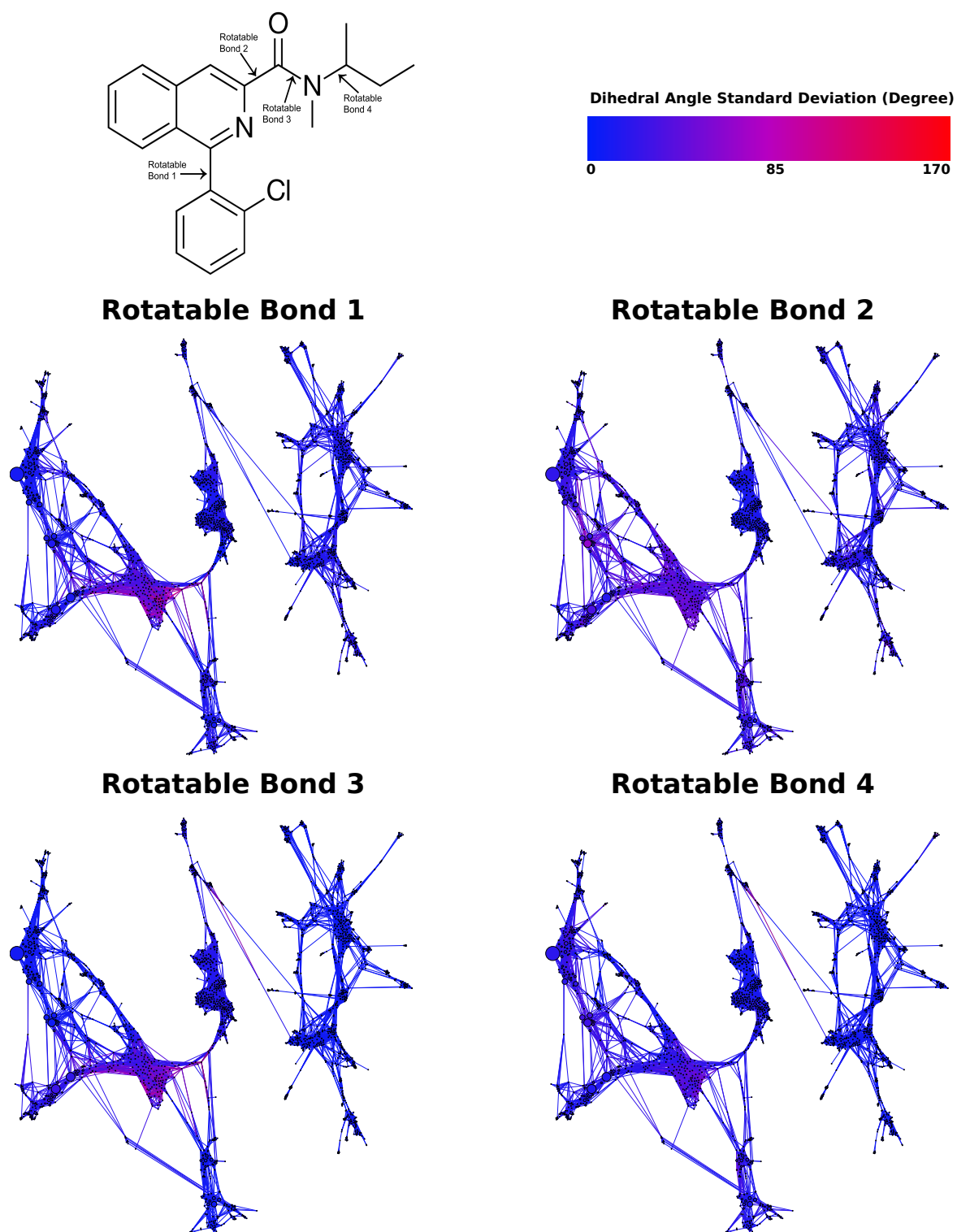


Figure 4.14: The standard deviation of the dihedral angles for the MSM states for four different rotatable bonds on the PK-11195 ligand.

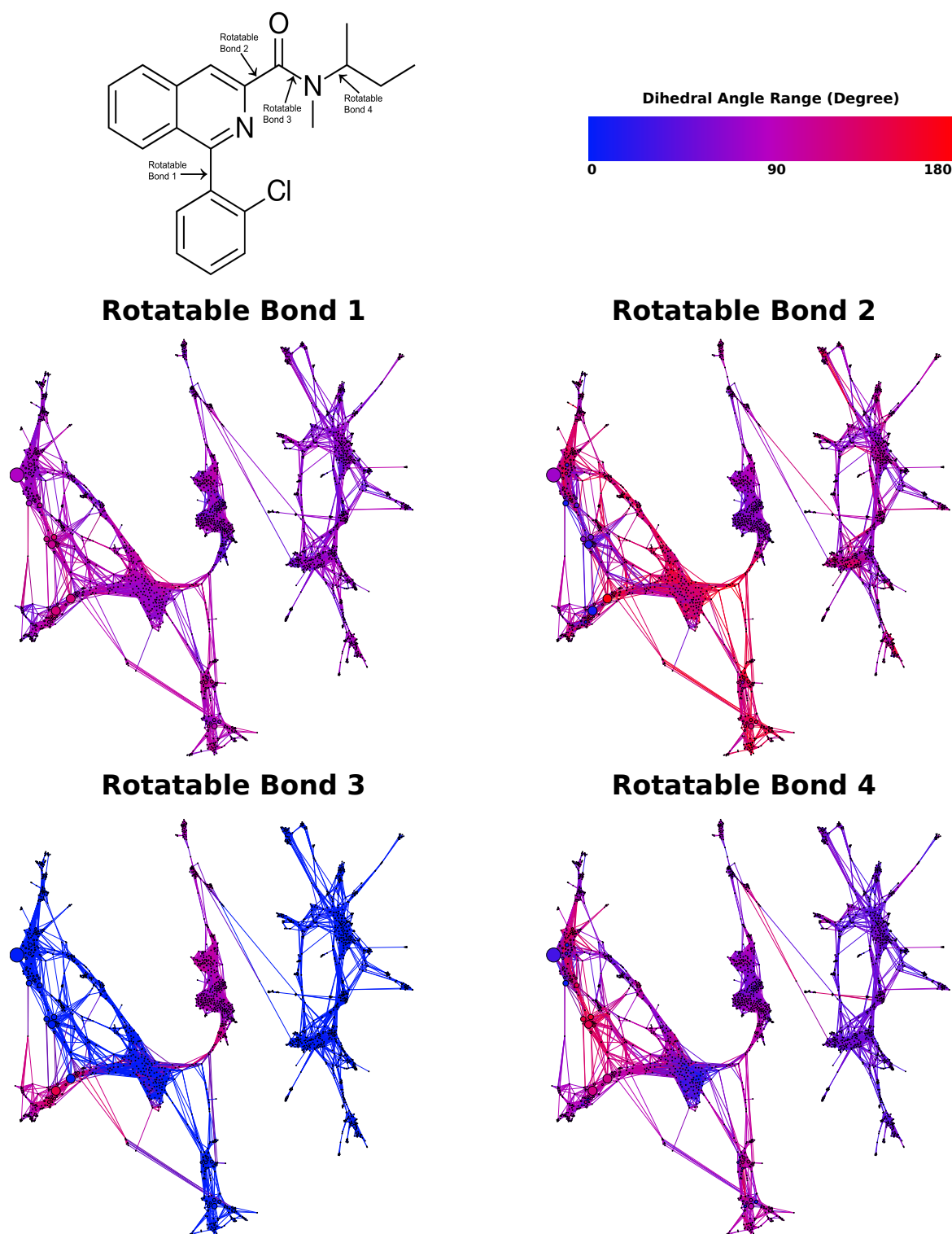


Figure 4.15: The range of the dihedral angles for the MSM states for four different rotatable bonds on the PK-11195 ligand.

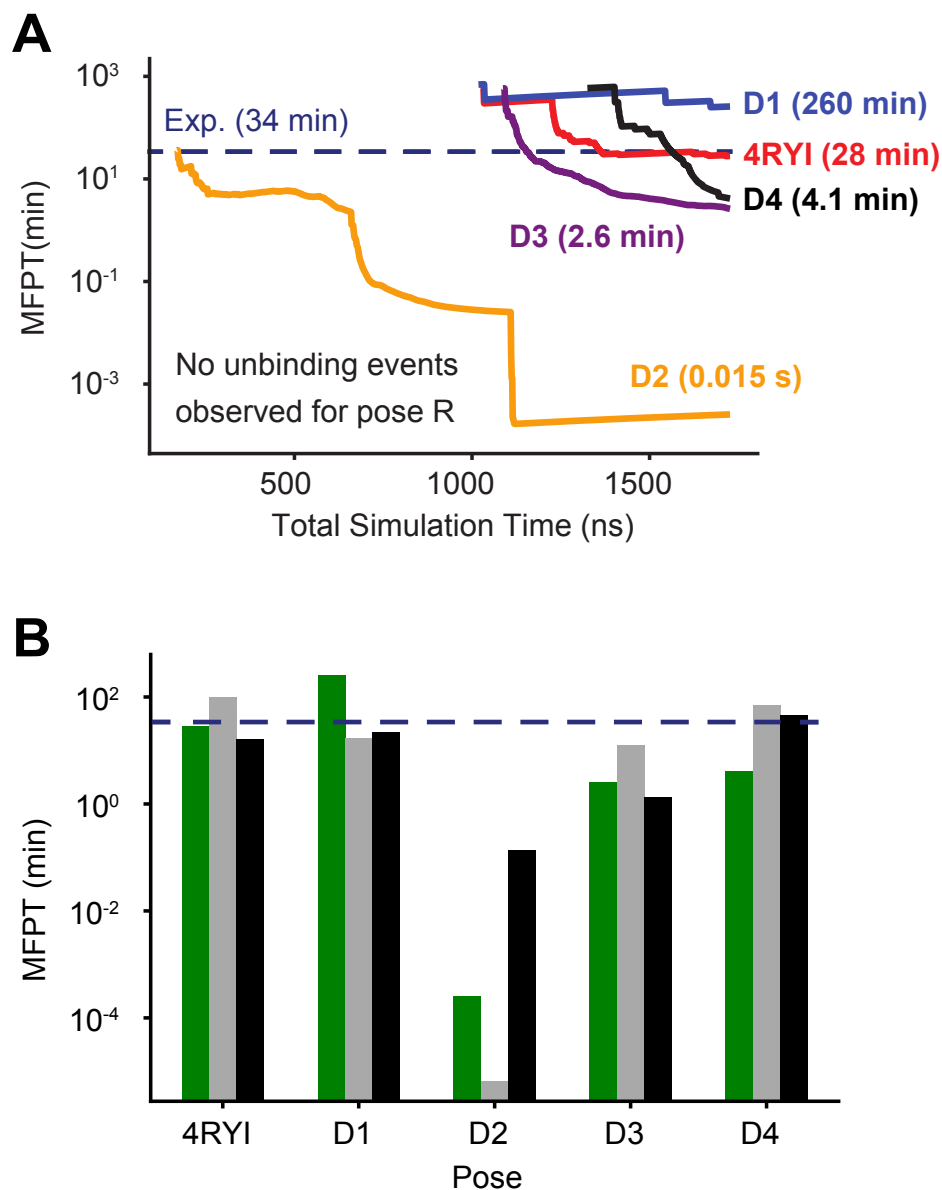


Figure 4.16: (A) MFPT estimates using unbinding fluxes observed over the course of REVO simulations. The light shaded area shows the standard error across the three simulations conducted for each pose. (B) A bar graph of the final MFPTs comparing the Hill Relation (green), MSM simulations before (grey), and after (black) the addition of new straight forward MD simulations. Pose-specific MFPTs were computed from MSMs that were built using only trajectories generated from that starting pose. Simulations starting from pose R never entered the unbound basin and thus MFPTs could not be determined by either method. The experimental MFPT of 34 min is shown as a dashed blue line in each panel.

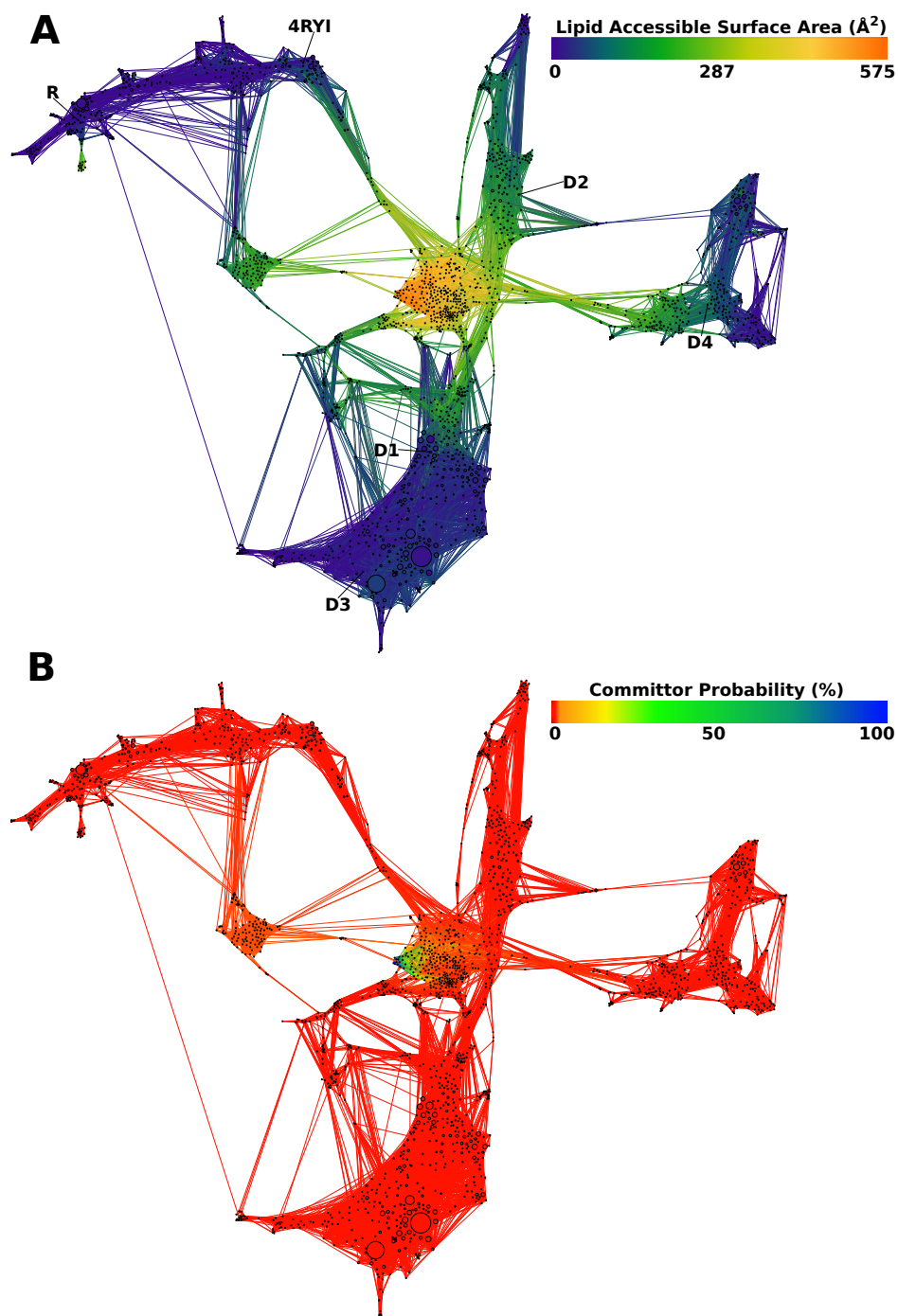


Figure 4.17: Combined conformation space network of all REVO simulations from each starting pose with the addition of frames from straightforward MD simulations, colored by (A) LASA and (B) committor probability. Starting poses are marked in bold in panel A.

4RYI

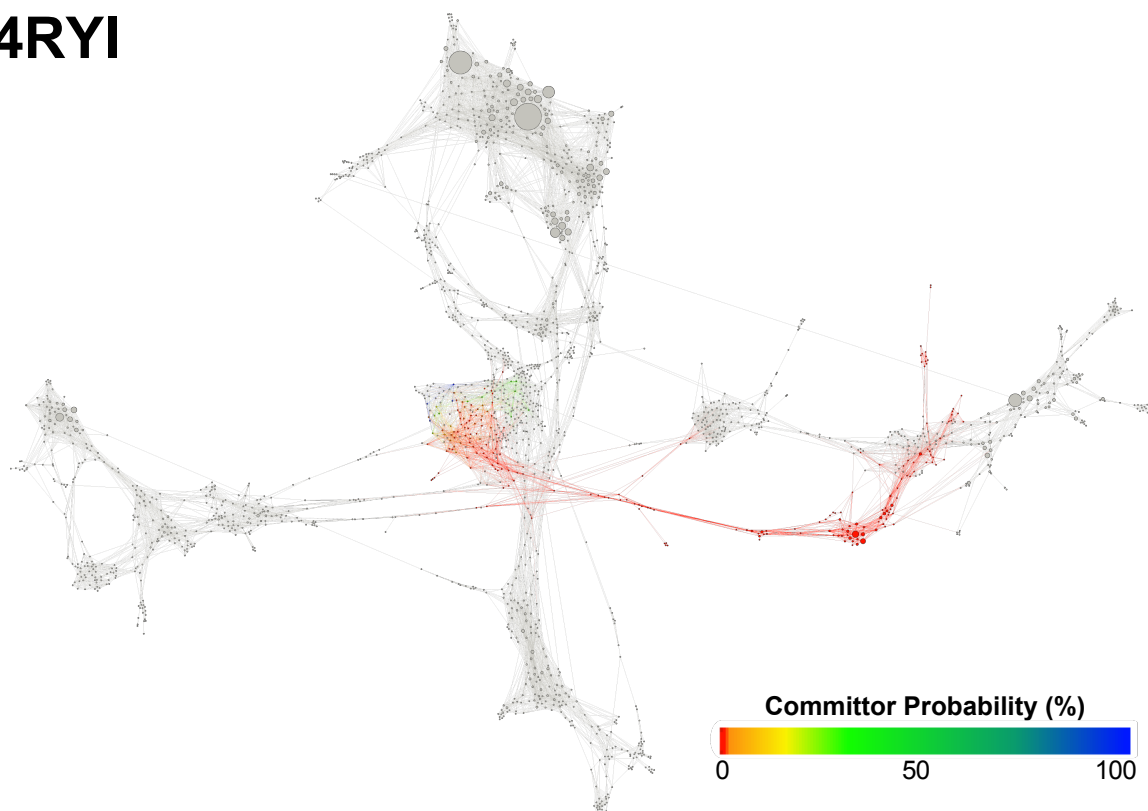


Figure 4.18: An MSM network including both straight forward and REVO trajectories colored by pose specific committor probability values calculated using trajectories beginning in pose 4RYI. States that were not visited by these simulations are colored grey.

D1

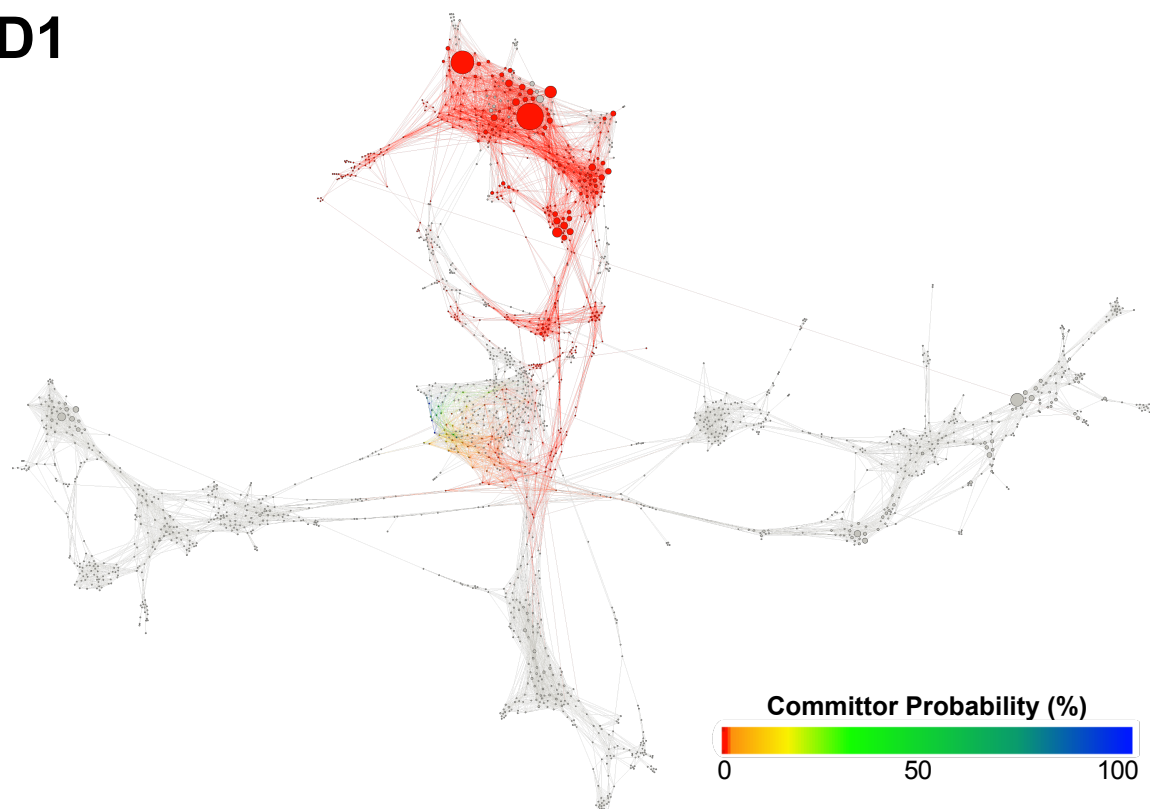


Figure 4.19: An MSM network including both straight forward and REVO trajectories colored by pose specific committor probability values calculated using trajectories beginning in pose D1. States that were not visited by these simulations are colored grey.

D2

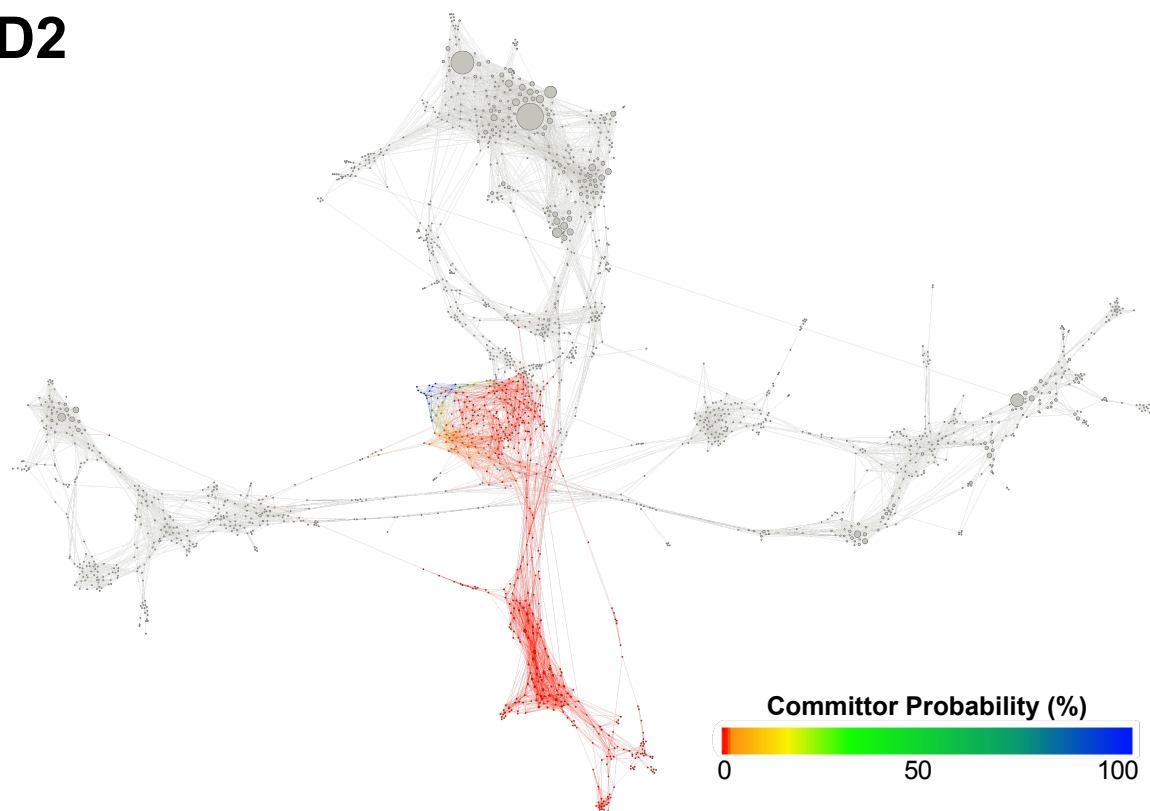


Figure 4.20: An MSM network including both straight forward and REVO trajectories colored by pose specific committor probability values calculated using trajectories beginning in pose D2. States that were not visited by these simulations are colored grey.

D3

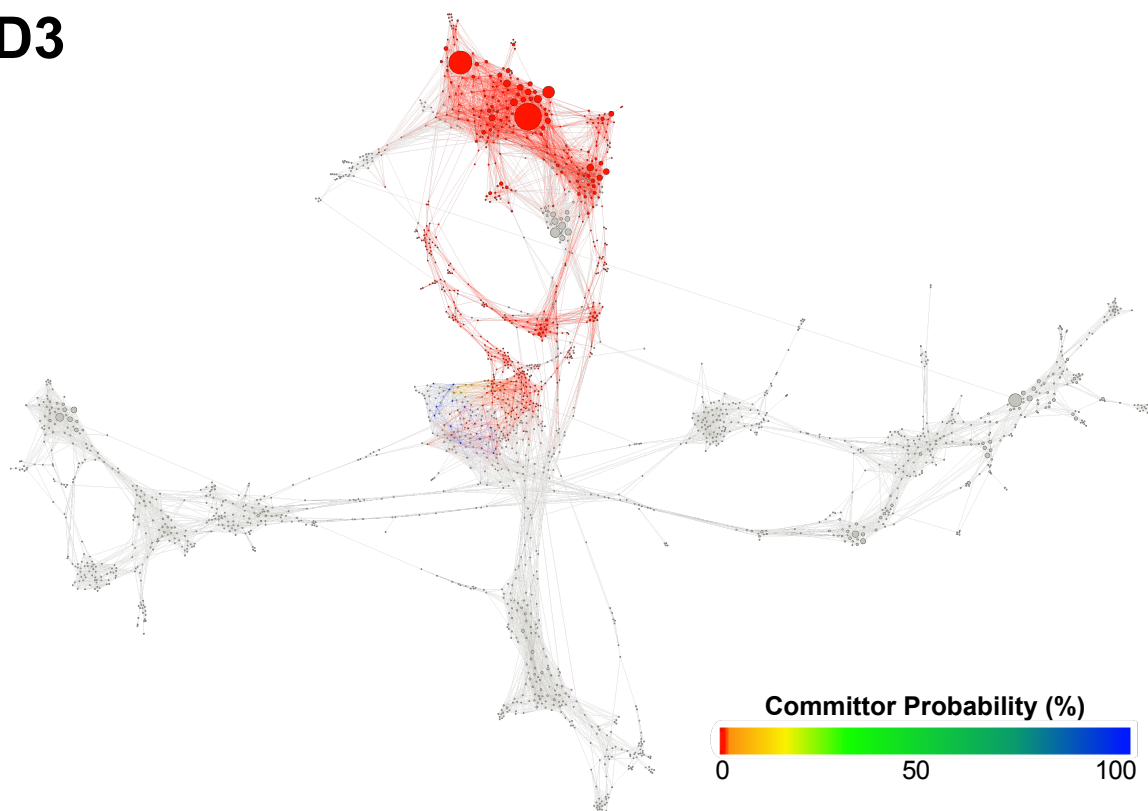


Figure 4.21: An MSM network including both straight forward and REVO trajectories colored by pose specific committor probability values calculated using trajectories beginning in pose D3. States that were not visited by these simulations are colored grey.

D4

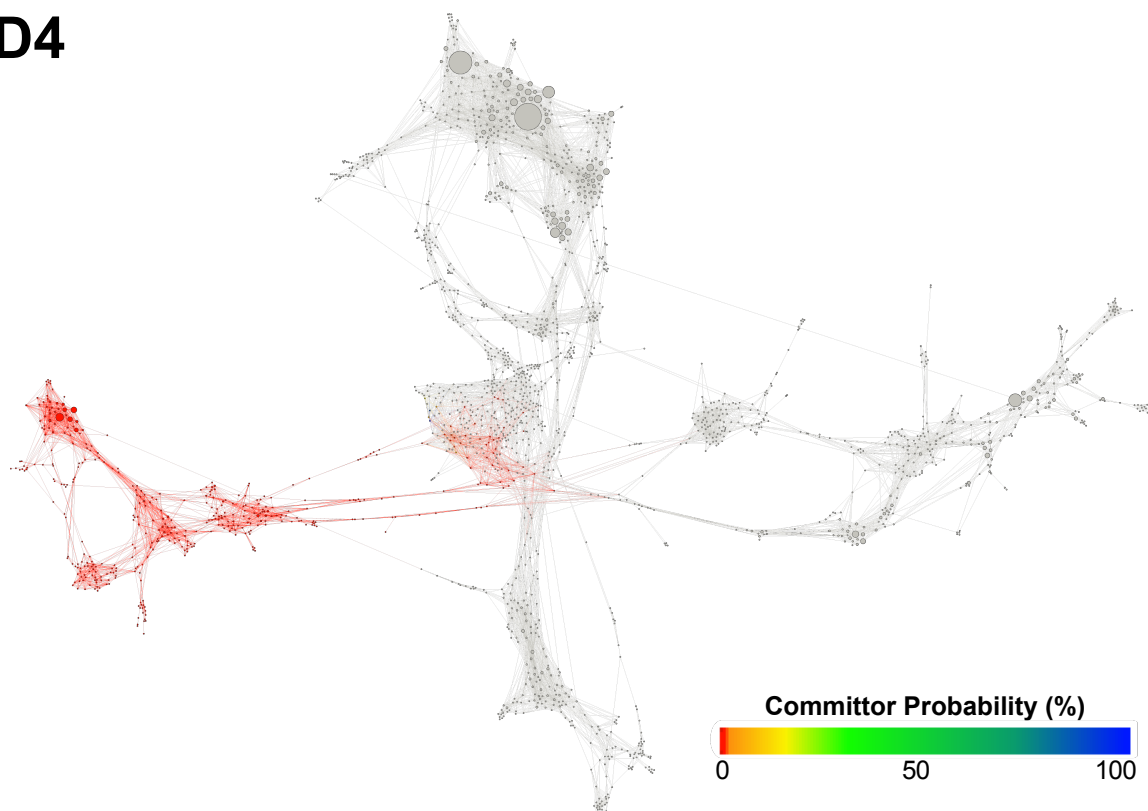


Figure 4.22: An MSM network including both straight forward and REVO trajectories colored by pose specific committor probability values calculated using trajectories beginning in pose D4. States that were not visited by these simulations are colored grey.

4.4 Discussion and Conclusion

The results of our simulation show that from all six initial PK-11195 poses, using the *R. sphaeroides* TSPO structure, the ligand dissociates into the membrane through the trans-membrane helices. We found a pathway between TM1 and TM2 and a lower probability pathway between TM2 and TM5. These pathways identify residues with which PK-11195 has high interaction energy. Among them are aromatic residues: Phe46 and Trp50 which form π - π interactions with the ligand. The interactions with the Trp50 rings are also found in different bound states. We note that the Trp50 residue happens to be highly conserved across organisms of several species and kingdoms. These stabilizing interactions could lower the barrier to entry for other TSPO ligands such as protoporphyrin-IX and heme, which are also largely aromatic.

Previous results [139] using a different starting pose and TSPO structure showed PK-11195 dissociating into the cytosol through the LP1 loop region. The TSPO structure used in the previous study was built from a homology model based on the mouse NMR TSPO structure and used the rat sequence, whereas our structure was determined from X-Ray crystallography from *R. sphaeroides* TSPO. As mentioned in the Introduction, this NMR structure was destabilized by the detergent used in the purification [181, 182], which likely affected the homology model structure as well. This, in addition to the differences in sequence, results in several key structural differences between the mouse (PDB 2MGY [167]) and *R. sphaeroides* (PDB 4UC1 [165]) structures. TM1 in the mouse structure is significantly longer and the top portion of the helix is at a drastically different angle than the helix in the structure we used in our simulations. While the LP1 region is present in both structures, the *R. sphaeroides* sequence has a small α -helix which in the mouse structure is incorporated into TM1. Finally, the LP1 region in *R. sphaeroides* has several stabilizing interactions [165] between non-bonded residues such as between Trp30-Met97, Asp32-Arg43 and Trp39-Gly141 that are not present in the mouse structure. This stabilization limits the freedom of motion of the LP1 loop, sterically hindering PK-11195 from leaving via the

LP1 pathway. In addition to TSPO structural differences, previous results were obtained using a 2:1 POPC:cholesterol lipid bilayer, while our results used an approximately 2.9:1.6:1 mixture of POPC:POPE:POPI lipids. Cholesterol is known to bind to TSPO, although known binding sites are not close to the TM1-TM2 pathway found here. Differences in lipid composition could also affect membrane fluidity, which could impact the relative probabilities of the LP1 and TM1-TM2 pathways. It will be an important goal of future work to parse the relative impact of these differences (protein sequence, protein structure and membrane composition) in determining ligand dissociation pathways.

There is interest in designing new TSPO ligands with longer RTs [18, 178]. The ligand binding transition state is the rate-limiting step of ligand binding and release, which can also be identified in simulations by a committor probability of 0.5 between the bound and unbound basins. Here we find that the ligand binding transition state occurs when the ligand has only minimal direct contact with TSPO, with a Q_{12} of ~ 10 Å. In addition to details of the bound state, this implies that TSPO ligand RT is primarily affected by properties related to membrane permittivity and diffusivity, such as hydrophobicity. These results lead to the hypothesis that the membrane composition could have a direct impact on ligand binding kinetics of PK-11195.

This work also raises questions about membrane insertion and removal along ligand binding paths. Additional REVO simulations with only PK-11195 and the lipid membrane could reveal the membrane diffusion coefficient of PK-11195 as well as rate constants for insertion and removal to form holistic models of membrane-mediated binding that stretch from solvent to binding site. A larger question is how the presence of other proteins known to interact with TSPO, such as voltage dependent ion channel (VDAC) [170] and cytochrome P450s [189] affect the unbinding/binding and insertion/removal pathways. Cholesterol could also affect the binding pathways of PK-11195, either by binding to TSPO and affecting a conformational change, or through membrane fluidity, which could affect the (un)binding rate of PK-11195 as it interacts with the membrane [190].

Although it is exciting that our predicted RTs come so close to experimental quantities, some caution should be exercised in making this comparison. First, it has been previously shown that some simulations using traditional MD force fields do not produce reliable estimations for RT [58]. However, we note that this result was mainly due to a lack of polarizability in the force field and errors in parameters that overestimate the electrostatic interactions. In our system, PK-11195 is uncharged and we do not expect these errors from the force field to dramatically influence the RT. Another thing to note is the experimental MFPT reported by Costa [18, 178] was determined using human TSPO while our simulations used the structure from *R. sphaeroides* containing a A139T mutation. While the mutation was designed to mimic the human TSPO structure [165], the human and *R. sphaeroides* sequences have low homology (30%) which could potentially result in different transition paths, transition states and unbinding rates. Furthermore, these results emphasize that we should take care to ensure consistency of the “unbound” state from simulation and experiment. In radioligand displacement assays, any ligand pose that is not sterically blocking entry of the radiolabelled competitor ligand would be considered “unbound” [191]. However, in surface plasmon resonance, a ligand would still be considered bound until it dissociated from the detergent that is bound to the chip along with TSPO. Our simulations show how differences in the definition of the unbound state can lead to significant differences in RT, and could help rationalize differences between experimental RTs obtained with different methods.

CHAPTER 5

ATOMIC-RESOLUTION PREDICTION OF DEGRADER-MEDIATED TERNARY COMPLEX STRUCTURES BY COMBINING MOLECULAR SIMULATIONS WITH HYDROGEN DEUTERIUM EXCHANGE

The research conducted in this chapter was done in collaboration with Roivant Discovery. I prepared, simulated and analyzed the REVO simulations and compared my results with other methods. Roivant Discovery performed the crystallography, hydrogen deuterium exchange, docking, and HREMD experiments and analyzed the HREMD experiments to construct the free energy landscape.

This work is not yet published but is available on BioRxiv [192]. This chapter is an excerpt from that work, presenting only what pertains to the REVO simulations. The crystallography section denotes how the reference structure for the ACBI1 PROTAC was determined and REVO attempts to replicate the warhead conformation from this structure. The HDX experiments were used to determine what residues are protected from hydrogen-deuterium exchange when the ternary complex was formed and we created a distance metric used in REVO to maximize contacts between these residues. The I-RMSD of the bound states from docking and REVO simulations were compared. The most probable states from the REVO simulations were projected onto the free energy landscape constructed via the HREMD simulations.

5.1 Introduction

Heterobifunctional degrader molecules are a class of ligands that induce proximity between a target protein of interest (POI) and a E3 ubiquitin ligase, which can ultimately lead to ubiquitination of the POI and its subsequent proteosomal degradation through a complex machinery of proteins[193]. These degrader molecules provide the opportunity of a novel therapeutic modality, single molecules induce catalytic turnover of the POI, and potentially offer an avenue for modulation of targets traditionally labeled as undruggable by classical

therapeutic strategies [194, 195, 196]. The subset of degrader molecules classified as heterobifunctionals, also known as proteolysis-targeting chimera (PROTAC) molecules, consist of two separate moieties joined by a “linker”; the “warhead” binds to the and the “ligand” binds to an E3 ligase such as Cereblon [197, 198, 199], cIAP [200], Keap1[201], and the von Hippel-Lindau disease tumor suppressor (VHL) [202, 203, 204]. In each case it is the ability of the warhead-linker-ligand degrader molecule to induce a ternary complex that is critical for bridging the interaction between the POI and an E3 ligase (which can be the native or non-native degradation partner for the POI).

The formation of the POI-degrader-E3 ternary complex is central to the targeted protein degradation (TPD) process, but how the formation of the ternary structure impacts protein degradation is still poorly understood, especially given the dynamic nature of the non-native induced proximity complex[23]. X-ray crystallography, the primary experimental technique for determining 3-dimensional structures of the ternary complex [205], provides a high resolution structure of a single conformational state, but a growing body of evidence suggests that the dynamic nature of the ternary structure is integral to the binding cooperativity (the term used to describe degree to which the binding affinity of ternary complexes are thermodynamically different than the binary counterparts) and degradation efficiency. A study targeting the degradation of Burton Tyrosine Kinase by Cereblon found that optimal protein removal was achieved through a molecule that induced a non-cooperative ternary complex, demonstrating a disconnect between binding affinity and degradation efficiency[206]. Similarly, Burton Tyrosine Kinase was also found to non-cooperatively interact with cIAP but still led to high degradation efficiency[194]. Interestingly, NMR and crystallography revealed a structural ensemble being sampled by this ternary complex, suggesting specific conformations could be responsible for efficient downstream ubiquitination[207]. In contrast, studies with SMARCA2 and VHL found that more cooperative molecules led to higher degradation efficiency[208]. Furthermore, analysis of the ternary structures revealed a high degree of similarity despite the fact that the heterobifunctional molecules displayed different degrees

of degradation efficiency[208], raising questions about relationship between static structural representations of the ternary complex and degradation efficiency. These findings and others[209, 210, 211] suggest that degradation efficiency is more complex than can be understood through the thermodynamics of binding or static structural analysis. As such, determining the dynamic ensemble of the ternary complex can reveal mechanistic insights to facilitate the design of more effective degrader molecules, especially to understand the relationship between linkers and degradation. [205, 208, 212, 213, 214].

Previous work to computationally predict ternary structures mostly consists of protein-protein docking protocols, perhaps followed by refinement of the initial structures with molecular dynamics (MD) simulations to assess the stability of the predicted models [208, 214, 215, 216, 217, 218]. However, these docking protocols fail to predict high-resolution structures (sub-2.0 Å) with high fidelity. That is, while protein-protein docking protocols have shown some promise in generating structural models of ternary complexes with reasonable resolution (often characterized as sub-10 Å root mean square deviation (RMSD) to an x-ray structure), the best structures typically fall somewhere within a long list of possible poses (often in the hundreds or thousands), demonstrating the challenge associated with the selection of high-accuracy ternary structure models.

Here, we present an integrated workflow that combines solution-state biophysical techniques with advanced MD simulations to produce atomic resolution structural ensembles of the ternary complex. Hydrogen-deuterium exchange (HDX) protection data is used as a collective variable (CV) in the MD simulation, enhancing both the speed and accuracy of the computational predictions. Furthermore, HDX data is also used as constraints for protein-protein docking when higher throughput and lower resolution models are sought, such as when screening many degrader molecules.

We use the weighted ensemble (WE) approach to perform MD simulations at biologically relevant timescales (from microseconds to milliseconds) across multiple graphics processing units (from dozens to thousands of simultaneous GPUs). This approach allows for the speed

and throughput needed to sample the conformational free energy landscape at a sufficient level to generate robust, high accuracy predictions of the ternary complex structural ensemble. WE utilizes an adaptive sampling procedure where an ensemble of unbiased trajectories are iteratively simulated and analysed so that computational resources can be optimally reallocated to regions of interest (e.g. unexplored regions of conformational phase space or regions of interest based on data from HDX experiments). Trajectories in sparsely populated regions (i.e. limited data for statistical thermodynamic calculations) are cloned in order to enhance sampling and high-probability regions with sufficient data for computing statistical thermodynamic quantities are merged so computational resources can be reallocated to the sparsely populated regions [71]. The resampling is done such that the probability of the whole simulation ensemble of “walkers” is tracked in a statistically rigorous manner [44, 76, 75, 219]. We modified a bin-less algorithm called Resampling Ensembles by Variation Optimization (REVO) [35] to more efficiently sample ternary complex formation, which is implemented in the open source software package wepy [34]. The work presented here relies on knowledge of the binding pose of the warhead to the POI and that of the ligand to the E3 ligase, which are typically known from prior experiments or can be generated with computational tools like docking or shape-based alignment. Experimental HDX data is used to determine the level at which residues are shielded from solvent upon ternary complex formation, as compared to the binary complexes (POI plus degrader or E3 ligase plus degrader).

Our ultimate goal is to understand the structural and dynamic basis for differences in degradation among a set of degrader molecules. Here, we focus on three different degrader molecules of the BAF ATPase subunit SMARCA2 isoform 2 that recruit the E3 ligase VHL. The binding affinities and cooperativity of ternary complex formation and the degradation efficiency for these three degrader molecules are summarized in Table 5.1.

Ternary complex crystal structures of PROTAC 1 (Protein Data Bank (PDB) ID: 6HAY) and PROTAC 2 (PDB ID: 6HAX) show slight variations in the interactions and orientation of the proteins in the ternary structure. In addition, we obtain the crystal structure of the highly

Table 5.1: Binding affinity (K_d), efficiencies (IC50, DC50), and cooperativity (α) of PROTAC 1, PROTAC 2, and ACBI1 degraders. Ternary IC50 and binary (SMARCA2) DC50 values are reported; the cooperativity is the ratio of binary over ternary IC50. Table adapted from Farnaby et al. [208].

	$K_d^{\text{VHL}}(nM)$	$K_d^{\text{SMARCA2}}(nM)$	IC50 (nM)	DC50 (nM)	α
PROTAC 1	98 ± 26	4500 ± 480	205 ± 15	300	12
PROTAC 2	100 ± 10	770 ± 51	45 ± 9	N/A	18
ACBI1	250 ± 64	1800 ± 980	26 ± 3	6/3.3	30

cooperative and more efficient degrader ACBI1 (PDB ID 7S4E from the work presented here) (Section 5.3.1). A static analysis of these crystal structures does not explain the difference in cooperativity and degradation of these heterobifunctional degrader molecules. To explain the different degradation profiles of these molecules, we carry out MD simulations and solution experiments, which reveal insights beyond what is defined by the crystal structure alone (Section 5.3.2).

Our results show that by including experimental solution-phase HDX data into the REVO simulations (REVO+HDX) we obtain improved throughput and accuracy of the ternary structure predictions. Starting from unbound SMARCA2 and VHL structures, REVO+HDX is able to produce structural models of ternary complexes with Interface-RMSD below 2 Å from the experimental x-ray crystal structures (Section 5.3.3). Additionally, REVO+HDX generates an ensemble of bound conformations spanning a free energy basin within 3 kcal/mol from the crystal structure (Section 5.3.5). These dynamic models describe an ensemble of energetically viable structures that could be used to study multiple aspects of the targeted protein degradation process, including binding kinetics, affinity, selectivity, cooperativity, ubiquitination, and degradation. We make prospective ternary structure predictions of the SMARCA2 isoform 1, ACBI1 and VHL:Elongin C:Elongin B, where SMARCA2 isoform 1 has a 17 amino acid extension compared to isoform 2. Our prediction reconciles the HDX data showing interaction of the isoform 1 extension with a beta-strand from VHL.

We also introduce methodology to determine the conformational free energy landscapes of these ternary complexes, which is the foundation for quantifying the populations of different

conformational states. Starting from the crystal structures, we first sample conformations using a HREMD simulation similar to solute tempering. From these simulations, we choose structures as seeds to run 10,000 simulations on Folding@Home, totalling approximately 6 ms of accumulated simulation time. We build a Markov State Model (MSM) [220] that identify the most probable structures along with their conformational free energies and kinetics of interconversion, all of which can be used to guide the design of novel degrader molecules.

5.2 Methods

5.2.1 Experimental Methods

5.2.1.1 Cloning, expression and purification of SMARCA2 and VHL/EloB/C

The SMARCA2 gene from *Homo sapiens* was custom-synthesized at Genscript with N-terminal GST tag (Ciulli 2019 Nature ChemBio) and thrombin protease cleavage site. The synthetic gene comprising the SMARCA2 (UniProt accession number P51531-1; residues 1373-1511) was cloned into pET28 vector to create plasmid pL-477. The second construct of SMARCA2 with deletion 1400-1417 (UniProt accession number P51531-2) was created as pL-478. For biotinylated SMARCA2, AVI-tag was gene synthesized at C-terminus of pL-478 to create pL-479. The VHL gene from *Homo sapiens* was custom-synthesized with N-terminal His6 tag [208] and thrombin protease cleavage site. The synthetic gene comprising the VHL (UniProt accession number P40337; residues 54-213) was cloned into pET28 vector to create plasmid pL-476. ElonginB and ElonginC gene from *Homo sapiens* was custom-synthesized with AVI-tag at C-terminus of EloB [213]. The synthetic genes comprising the EloB (UniProt accession number Q15370; residues 1-104) and EloC (UniProt accession number Q15369; residues 17-112) were cloned into pCDFDuet vector to create plasmid pL-474. For protein structural study, AVI-tag was deleted in pL-474 to create pL-524.

For SMARCA2 protein expression, the plasmid was transformed into BL21(DE3) and plated on Luria-Bertani (LB) medium containing 50 $\mu\text{g}/\text{ml}$ kanamycin at 37 °C overnight.

A single colony of BL21(DE3)/pL-477 or BL21(DE3)/pL-478 was inoculated into a 100-ml culture of LB containing 50 μ g/ml kanamycin and grown overnight at 37 °C. The overnight culture was diluted to OD₆₀₀=0.1 in 2 x 1-liter of Terrific Broth medium containing 50 μ g/ml kanamycin and grown at 37 °C with aeration to mid-logarithmic phase (OD₆₀₀ = 1). The culture was incubated on ice for 30 minutes and transferred to 16 °C. IPTG was then added to a final concentration in each culture of 0.3 mM. After overnight induction at 16 °C, the cells were harvested by centrifugation at 5,000 xg for 15 min at 4 °C. The frozen cell paste from 2 L of cell culture was suspended in 50 ml of Buffer A consisting of 50 mM HEPES (pH 7.5), 0.5 M NaCl, 5 mM DTT, 5% (v/v) glycerol, supplemented with 1 protease inhibitor cocktail tablet (Roche Molecular Biochemical) per 50 ml buffer. Cells were disrupted by Avestin C3 at 20,000 psi twice at 4 °C, and the crude extract was centrifuged at 39,000 xg (JA-17 rotor, Beckman-Coulter) for 30 min at 4 °C. Two ml Glutathione Sepharose 4 B (Cytiva) was added into the supernatant and mixed at 4 °C for 1 hour, washed with Buffer A and eluted with 20 mM reduced glutathione (Sigma). The protein concentration was measured by Bradford assay, and GST-tag was cleaved by thrombin (1:100) at 4 °C overnight during dialysis against 1 L of Buffer B (20 mM HEPES, pH 7.5, 150 mM NaCl, 1mM DTT). The sample was concentrated to 3 ml and applied at a flow rate of 1.0 ml/min to a 120-ml Superdex 75 (HR 16/60) (Cytiva) pre-equilibrated with Buffer B. The fractions containing SMARCA2 were pooled and concentrated by Amicon® Ultracel-3K (Millipore). The protein concentration was determined by OD₂₈₀ and characterized by SDS-PAGE analysis and analytical LC-MS. The protein was stored at -80 °C.

For VHL/EloB/C protein expression, the plasmids were co-transformed into BL21(DE3) and plated on Luria-Bertani (LB) medium containing 50 μ g/ml kanamycin and 50 μ g/ml streptomycin at 37 °C overnight. A single colony of BL21(DE3)/pL-476/474 or BL21(DE3)/pL-476/524 was inoculated into a 100-ml culture of LB containing 50 μ g/ml kanamycin and 50 μ g/ml streptomycin and grown overnight at 37 °C. The overnight culture was diluted to OD₆₀₀=0.1 in 6 x 1-liter of Terrific Broth medium containing 50 μ g/ml kanamycin and 50

$\mu\text{g/ml}$ streptomycin and grown at 37 °C with aeration to mid-logarithmic phase ($\text{OD}_{600} = 1$). The culture was incubated on ice for 30 minutes and transferred to 18 °C. IPTG was then added to a final concentration of 0.3 mM in each culture. After overnight induction at 18 °C, the cells were harvested by centrifugation at 5,000 g for 15 min at 4 °C. The frozen cell paste from 6 L of cell culture was suspended in 150 ml of Buffer C consisting of 50 mM HEPES (pH 7.5), 0.5 M NaCl, 10 mM imidazole, 1 mM TCEP, 5% (v/v) glycerol, supplemented with 1 protease inhibitor cocktail tablet (Roche Molecular Biochemical) per 50 ml buffer. Cells were disrupted by Avestin C3 at 20,000 psi twice at 4 °C, and the crude extract was centrifuged at 17000 g (JA-17 rotor, Beckman-Coulter) for 30 min at 4 °C. Ten ml Ni Sepharose 6 FastFlow (Cytiva) was added into the supernatant and mixed at 4 °C for 1 hour, washed with Buffer C containing 25 mM imidazole and eluted with 300 mM imidazole. The protein concentration was measured by Bradford assay. For protein crystallization, His-tag was cleaved by thrombin (1:100) at 4 °C overnight during dialysis against 1 L of Buffer D (20 mM HEPES, pH 7.5, 150 mM NaCl, 1 mM DTT). The sample was concentrated to 3ml and applied at a flow rate of 1.0 ml/min to a 120-ml Superdex 75 (HR 16/60) (Cytiva) pre-equilibrated with Buffer D. The fractions containing VHL/EloB/C were pooled and concentrated by Amicon® Ultracel-10K (Millipore). The protein concentration was determined by OD280 and characterized by SDS-PAGE analysis and analytical LC-MS. The protein was stored at -80 °C. For the Surface plasmon resonance (SPR) assay, 10 mg VHL/EloB/C protein complex was incubated with BirA (1:20), 1 mM ATP and 0.5 mM Biotin and 10mM MgCl_2 at 4 °C overnight, removed free ATP and Biotin by 120-ml Superdex 75 (HR 16/60) with the same procedure as above, and confirmed the biotinylation by LC/MS.

5.2.1.2 Hydrogen Deuterium Exchange Mass Spectrometry

Our HDX analyses were performed as reported previously with minor modifications [221, 222, 223]. HDX experiments were performed using a protein stock at the initial concentra-

tion of 200 μ M of SMARCA2, VCB in the APO, binary (200 μ M PROTAC ACBI1) and ternary (200 μ M PROTAC ACBI1) states in 50 mM HEPES, pH 7.4, 150 mM NaCl, 1 mM TCEP, 2% DMSO in H₂O. The protein samples were injected into the nanoACQUITY system equipped with HDX technology for UPLC separation (Waters Corp. [224]) to generate mapping experiments used to assess sequence coverage. Generated maps were used for all subsequent exchange experiments. HDX was performed by diluting the initial 200 μ M protein stock 13-fold with D₂O (Cambridge Isotopes) containing buffer (10 mM phosphate, pH 7.4, 150 mM NaCl) and incubated at 10 °C for various time points (0.5, 5, 30 min). At the designated time point, an aliquot from the exchanging experiment was sampled and diluted 1:13 into D₂O quenching buffer containing (100 mM phosphate, pH 2.1, 50 mM NaCl, 3M GuHCl) at 1 °C. The process was repeated at all time points, including for non-deuterated samples in H₂O-containing buffers. Quenched samples were injected into a 5- μ m BEH 2.1 X 30-mm Enzymate-immobilized pepsin column (Waters Corp.) at 100 μ l/min in 0.1% formic acid at 10 °C and then incubated for 4.5 min for on-column digestion. Peptides were collected at 0 °C on a C18 VanGuard trap column (1.7 μ m X 30 mm) (Waters Corp.) for desalting with 0.1% formic acid in H₂O and then subsequently separated with an in-line 1.8 μ MHss T3 C18 2.1 X 30-mm nanoACQUITY UPLC column (Waters Corp.) for a 10-min gradient ranging from 0.1% formic acid to acetonitrile (7 min, 5–35%; 1 min, 35–85%; 2 min hold 85% acetonitrile) at 40 μ l/min at 0 °C. Fragments were mass-analyzed using the Synapt G2Si ESL-Q-ToF mass spectrometer (Waters Corp.). Between injections, a pepsin-wash step was performed to minimize peptide carryover. Mass and collision-induced dissociation in data-independent acquisition mode (MSE) and ProteinLynx Global Server (PLGS) version 3.0 software (Waters Corp.) were used to identify the peptides in the non-deuterated mapping experiments and analyzed in the same fashion as HDX experiments. Mapping experiments generated from PLGS were imported into the DynamX version 3.0 (Waters Corp.) with quality thresholds of MS1 signal intensity of 5000, maximum sequence length of 25 amino acids, minimum products 2.0, minimum products per amino acid of 0.3, minimum PLGS

score of 6.0. Automated results were inspected manually to ensure the corresponding m/z and isotopic distributions at various charge states were assigned to the corresponding peptides in all proteins (SMARCA2, VHL, ElonC, ElonB). DynamX was utilized to generate the relative deuterium incorporation plots and HDX heat map for each peptide. The relative deuterium uptake of common peptides was determined by subtracting the weighted-average mass of the centroid of the non-deuterated control samples from the deuterated samples at each time point. All experiments were made under the same experimental conditions negating the need for back-exchange calculations but therefore are reported as relative [225]. All HDX experiments were performed twice, on 2 separate days, and a 98 and 95% confidence limit of uncertainty was applied to calculate the mean relative deuterium uptake of each data set. Mean relative deuterium uptake thresholds were calculated as described previously [221, 222, 223]. Differences in deuterium uptake that exceeded the error of the datasets were considered significant.

5.2.1.3 Structural Determination of SMARCA2:ACBI1:VHL Complex

Purified SMARCA2 and VCB in 50 mM HEPES, pH 7.5, 150 mM NaCl, 1 mM DTT were incubated in a 1:1:1 molar ratio with ACBI1 for 1 hour at room temperature. Incubated complex was subsequently injected on to a Superdex 10/300 GL increase (Cytiva) pre-incubated with 50 mM HEPES, pH 7.5, 150 mM NaCl, 1 mM DTT, 2% DMSO at a rate of 0.5 mL/min to separate any noncomplexed partners from the properly formed ternary complex. Eluted fractions corresponding to the full ternary complex were gathered and spun concentrated to 14.5 mg/mL using an Amicon Ultrafree 10K NMWL Membrane Concentrator (Millipore). Crystals were grown 1-3 μ L hanging drops by varying the ratio of protein to mother liquor from 0.5-2:0.5-2 respectively. Crystals were obtained in buffer consisting of 0.1 M HEPES, pH 7.85, 13% PEG 3350, 0.2 M sodium formate incubated at 4 °C. Crystals grew within the first 24 hours but remained at 4 °C for 5 days until they were harvested, cryo protected in an equivalent buffer containing 20% glycerol and snap frozen in LN2. Diffraction data was col-

lected at NSLS2 beamline FMX ($\lambda=0.97932 \text{ \AA}$) using an Eiger X 9M detector. Crystals were found to be in the P 21 21 21 space group with unit cell dimensions of $a=80.14$, $b=116.57$, $c=122.23 \text{ \AA}$, where $\alpha=\beta=\gamma=90^\circ$. Crystal contained two copies of the SMARCA2:ACBI1:VCB (VHL, ElonC, ElonB) complex within the asymmetric unit cell. The structure was solved by performing molecular replacement with CCP4i2 [226] PHASER using PDB ID 6HAX as the replacement model. MR was followed by iterative rounds of modeling (COOT [227]) and refinement (REFMAC5 [228, 229, 230, 231, 232, 233, 234, 235, 236]) by standard methods also within the CCP4i2 suite. Structures were refined to R_{work}/R_{Free} of 23.7%/27.5%.

5.2.2 Computational Methods

5.2.2.1 Unbound System Preparation

In this chapter, we will be using weighted ensemble to predict the Probable global transcription activator SNF2L2 (SMARCA2) and the VHL-PROTAC ternary complex for three different PROTACS: ACBI1, PROTAC1, and PROTAC2. The ternary complexes for PROTAC1 and PROTAC2 can be found in PDB 6HAY and PDB 6HAX respectively and ternary complex for ACBI1 was solved in this chapter.

The simulation box of the unbound complex was solvated with explicit waters and counter ions were added to neutralize the net charge of the system. The ACBI1 system has 24,093 water molecules, 9 chlorine ions. The PROTAC 1 system has 21191 water molecules and 10 chlorine ions. The PROTAC 2 simulations has 31,567 water atoms and 9 chlorine ions. We used the Amber ff14SB force field for the proteins and TIP3 water model. All systems were placed in rectangular boxes, with dimensions: $123 \text{ \AA} \times 76 \text{ \AA} \times 98 \text{ \AA}$ for the ACBI1 system $131 \text{ \AA} \times 84 \text{ \AA} \times 84 \text{ \AA}$ for the PROTAC 1 system and $144 \text{ \AA} \times 89 \text{ \AA} \times 91 \text{ \AA}$ for the PROTAC 2 system. We used Amber ff14SB force fields for the protein and a TIP3 water model. The PROTAC molecular parameters were generated using in-house FFGEN/FFEngine tool. The PROTACs began each simulation bound to the VHL protein with the goal to bind the

VHL-PROTAC complex to SMARCA2.

5.2.2.2 Molecular Dynamics

All MD simulations were performed using OpenMM[114] v7.5.1. The time step for every simulation was 2 fs. To enforce constant temperature and pressure, a Langevin heat bath was used with a set temperature of 300K and a friction coefficient of 1 ps^{-1} was coupled to a Monte Carlo barostat set to 1 atm and volume moves were attempted every 50 time steps. The non-bonded forces were computed using the CutoffPeriodic function in OpenMM with a cutoff of 10 Å. The atomic positions and velocities are saved every 10,000 time steps, or every 20 ps of simulation time, which is the resampling period (τ) used here. The degrader-VHL complex was constrained to maintain the complex during the simulation by using a OpenMM custom centroid force defined as:

$$\text{Centroid Force} = k * (\text{dist} - \text{edist})^2, \quad (5.1)$$

where the dist is the distance between the center of mass of PROTAC and the center of mass of VHL and the edist is the distance between the center of mass of PROTAC and center of mass of VHL of the crystal structure, and k is a constant set to $2 \text{ kcal/mol} * \text{Å}^2$.

5.2.2.3 Generating Bound Ensemble

The bound ternary complex was made using the same procedure as described in Section 5.2.2.1 The PROTAC1 system used PDB:HAY for its starting conformation, the PROTAC2 system used PDB:HAX as its initial conformation and the ACBI1 system used PDB:7S4E. Straight forward MD was performed on the bound structures for $1\mu\text{s}$ each using the same parameters as described in Section 5.2.2.2 without the constraining force between VHL and the PROTAC. We then cluster the simulations into 25 cluster representatives using a vector describing if two residues are within 4.5 Å of each other. These 25 cluster representatives are the bound ensemble and will be used as reference structures for future calculations.

5.2.2.4 REVO-epsilon Weighted Ensemble method

To observe binding of the VHL-PROTAC complex and SMARCA2 we apply a variant of the weighted ensemble algorithm REVO. Each cycle of the REVO algorithm is comprised of two parts: semi-independent MD trajectories performed in parallel and resampling. Each of the MD trajectories (called "walkers") has a statistical weight (w) that contributes to statistical observables. All simulations ran with 48 walkers. After a trajectory time of τ we perform resampling. In resampling similar walkers are merged together and unique walkers are cloned, as defined by a distance metric. During cloning, the weight is evenly divided between the resultant clones and, when walkers are merged, the weights are combined to ensure the conservation of probability.

We will describe the application of the REVO algorithm as it pertains to this study, but a more detailed explanation can be found in previous works[35, 34] and in Chapters 2 and 4. The goal of the REVO resampling algorithm is to maximize the variation function defined as:

$$V = \sum_i V_i = \sum_i \sum_j \left(\frac{d_{ij}}{d_0} \right)^\alpha \phi_i \phi_j, \quad (5.2)$$

where V_i is the walker variation, d_{ij} is the distance between walkers i and j determined using a specific distance metric, d_0 is the characteristic distance used to make the distance term dimensionless, set to 0.148 for all simulations, the α is used to determine how influential the distances are to the walker variation and was set to 6 for all the simulations. The novelty terms ϕ_i and ϕ_j are defined as: $\phi_i = \log(w_i) - \log\left(\frac{p_{min}}{100}\right)$. The minimum weight, p_{min} , allowed during the simulation was 10^{-50} . The walker with the highest variance, V_i and when the weights of the resultant clones would be larger than p_{min} , and is within distance ϵ of the walker with the maximal progress towards binding of the ternary complex was proposed to be cloned. The two walkers selected for merging were within a distance of 2 Å and have a combined weight larger than the maximal weight allowed, p_{max} , which was set to 0.1 for all

REVO simulations. The merge pair also needed to minimize:

$$\frac{V_j w_i - V_i w_j}{w_i + w_j}, \quad (5.3)$$

If the proposed merging and cloning operations increase the total variance of the simulation, the operations are performed and we repeat this process until the variation can no longer be increased. After resampling is complete, we begin a new cycle.

5.2.2.5 Distance Metrics

Three different distance metrics were used while simulating the PROTAC 2 system: Using the warhead RMSD to the crystal structure, maximizing the contact strength between protected residues identified by HDX data, and a linear combination of the warhead RMSD, contact strength between HDX-protected residues, and the contact strength between SMARCA2 and the degrader. The simulations for the other systems used the last distance metric exclusively. To compute the warhead RMSD distance metric, we aligned to the binding site atoms on SMARCA2, defined as atoms that were within 8 Å of the warhead in the crystal structure. Then the RMSD was calculated between the warhead in each frame and the crystal structure. The distance between a set of walkers i and j is defined as: $d = |\frac{1}{RMSD_i} - \frac{1}{RMSD_j}|$. The contact strength is defined by determining the distances between residues. We calculate the minimum distance between the residues and use the following to determine the contact strength:

$$strength = \frac{1}{1 + e^{-k(r-r_0)}}, \quad (5.4)$$

where k is the steepness of the curve, r is the minimum distance between any 2 residues and r_0 is the distance we want a contact strength of 0.5. We used 10 for k and 5 Å for r_0 . The total contact strength was the sum of all residue-residue contact strengths. The distance between walkers i and j was calculated by: $d = |cs_i - cs_j|$ where cs is the contact strength of a given walker.

5.2.2.6 Ternary complex docking protocol

For the purpose of quick filtering through a large number of degrader designs, we take advantage of the conventional restriction of molecular flexibility used in molecular docking methods. Following [237, 217] (Methods 4 and 4b) and [216], we assume that high fidelity structures of :warhead (i.e., SMARCA2 isoform 2:PROTAC binding moiety) and E3:ligand (i.e., VHL:PROTAC binding moiety) are known and available to be used in protein-protein docking. This docking of two proteins with bound PROTAC moieties is performed in the absence of the linker. The conformations of linker are sampled independently with an in-house developed protocol that uses implementation of fast quantum mechanical methods, CREST [238, 239, 240]. Differently from the docking protocols described in [237, 217, 216], we make use of the distance restraints derived either from the end-to-end distances of the sampled conformations of linker, or from the HDX data. Thus, before running the protein-protein docking, we generate an ensemble of conformers for linkers and calculate the values of mean (x_0) and standard deviation (sd) for the end-to-end distance. This information is then used to set the distance restraints in the RosettaDock software [241, 242]:

$$f_1(x) = \left(\frac{x - x_0}{sd}\right)^2, \quad (5.5)$$

where x is the distance between a pair of atoms in a candidate docking pose (the pair of atoms is specified as the attachment points of the linker to warhead and ligand).

When information about the protected residues is available from HDX experiments, we used them to set up a set of additional distance restraints:

$$f_{2,i}(x) = \frac{1}{1 + \exp(-m \cdot (x - x_0))} - 0.5, \quad (5.6)$$

where i is the index of a protected residue, x_0 is the center of the sigmoid function and m is its slope. As above, x_0 value was set to be the mean end-to-end distance calculated over the ensemble of linker conformers. The value of m was set to be 2.0 in all the performed docking

experiments. The type of RosettaDock-restraint is *SiteConstraint*, with specification of C α atom for each protected residue and the chain-ID of partnering protein (i.e., x in Eq.(5.6) is the distance of C α atom from the partnering protein). Thus, the total restraint-term used in docking takes the form:

$$f_{\text{restr}}(x) = w \cdot (f_1(x) + \sum_i f_{2,i}(x)), \quad (5.7)$$

where $w = 10$ is the weight of this additional score function term.

RosettaDock implements a Monte Carlo-based multi-scale docking algorithm that samples both rigid-body orientation and side-chain conformations. The distance-based scoring terms, Eq. (5.7), bias sampling towards those docking poses that are compatible with specified restraints. This allows to limit the number of output docking structures, as only those ones that pass the Metropolis criterion with the additional term of Eq. (5.7) will be considered.

Once the docking poses are generated with RosettaDock, all the pre-generated conformations of the linker are structurally aligned onto each of the docking predictions [216]. Only those structures that satisfy the RMS-threshold value of $\leq 0.3 \text{ \AA}$ are saved as PDB files. All the docking predictions are re-ranked by the values of Rosetta Interface score (I_{sc}). The produced ternary structures are examined for clashes, minimized and submitted for further investigations with MD methods.

5.2.2.7 HREMD simulation

The details of HREMD [243, 244] are shown in Figures 5.1 and 5.2, and Tables 5.2 and 5.3. For all HREMD simulations, we chose the effective temperatures, $T_0 = 300 \text{ K}$ and $T_{max} = 425 \text{ K}$ such that the Hamiltonian scaling parameter, $\lambda_0 = 1.00$ and $\lambda_{min} = 0.71$ for the lowest and the highest rank replicas respectively. The effective temperatures of intermediate replicas are listed in Table 5.3. We estimated the number of replicas (n) in such a way that the average exchange probabilities (p) between neighboring replicas were in the range of 0.3 to

0.4. We used $n=20$ and $n=24$ for SMARCA2:degrader:VHL and SMARCA2:degrader:VCB respectively. Each simulation was run for $0.5 \mu\text{s}/\text{replica}$, and a snapshot of a complex was saved every 5 ps (total 100,001 frames per replica). Finally, we performed all the analyses on only the loREVOt rank replica that ran with original/unscaled Hamiltonian.

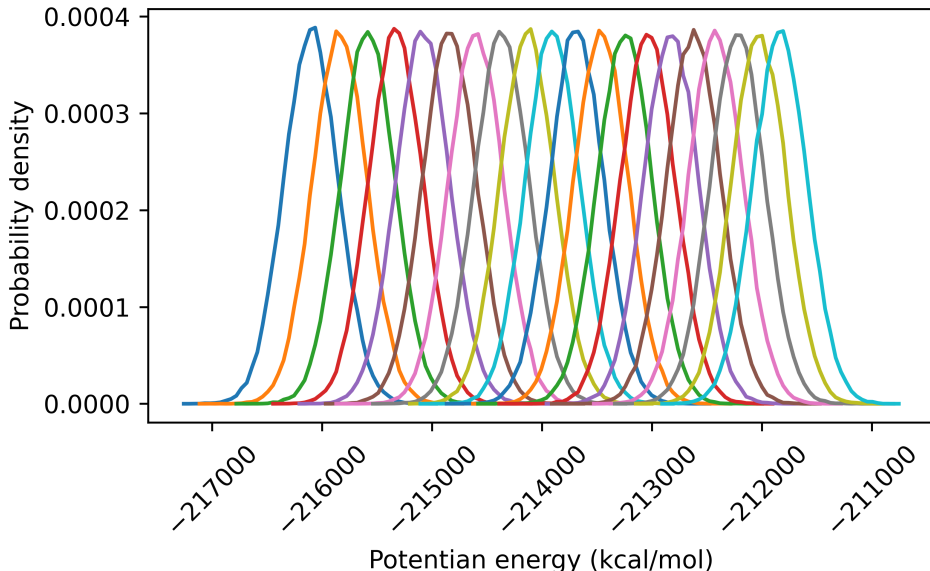


Figure 5.1: Potential energy of all replicas from HREMD simulation of **Sys7**. Left to right: rank0 to rank19. A good overlap between adjacent replicas suggests a sufficient number of replicas were employed and also confirmed no phase transition took place during the HREMD simulation.

We assessed the efficiency of sampling by observing (i) the values of p (Tables 5.2 and 5.3), (ii) a good overlap of histograms of potential energy between adjacent replicas (Figure 5.1), and (iii) a mixing of exchange of coordinates across all the replicas (Figure 5.2).

5.2.2.8 Conformational free energy landscape determination

In order to quantify to the conformational free energy landscape, we performed dimension reduction of our simulation trajectories using Principal Component Analysis (PCA). First, the simulation trajectories were featurized by calculating interfacial residue contact distances. Pairs of residues were identified as part of the interface if they passed within 5 \AA of each other

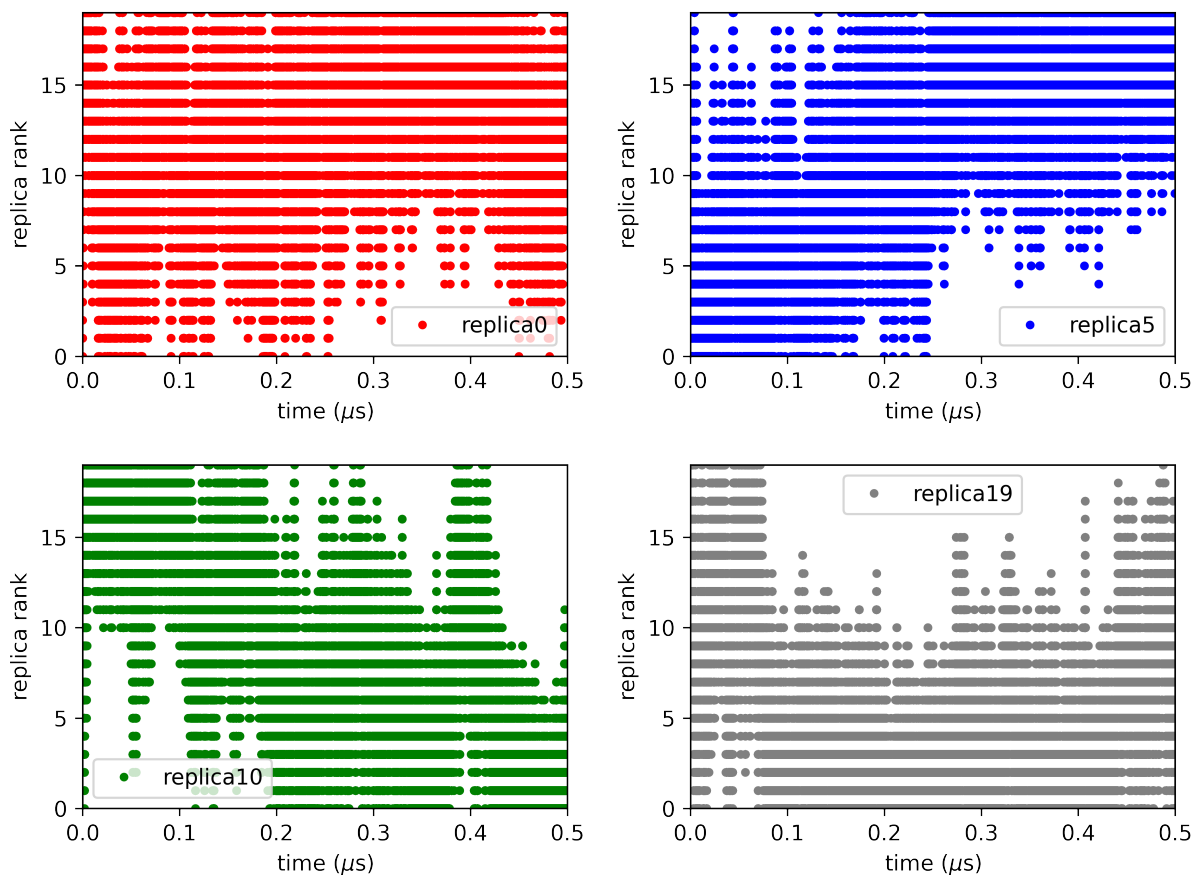


Figure 5.2: Effective temperature trajectories of replicas 0 (red), 5 (blue), 10 (green) and 19 (grey) from HREMD simulation of Sys7

during the simulation trajectory, where the distance between two residues was defined as the distance between their closest heavy atoms. PCA was then used to identify the features that contributed most to the variance by diagonalizing the covariance matrix; for each simulated system, the number of features used in our analysis was chosen as that which explained at least 95% of the variance.

After projecting the simulation data onto the resultant feature space, snapshots were clustered using the k -means algorithm. The number of clusters k was chosen using the “elbow-method”, i.e. by visually identifying the point at which the marginal effect of an additional cluster was significantly reduced. In cases where no “elbow” could be unambiguously

Table 5.2: Details of HREMD simulations. Protein complexes, number of atoms in a simulation box, number of replicas used and the aggregate length of the simulations are listed.

ID	Complex	# of atoms	# of replicas	Aggregate length (μ s)
Sys1	SMARCA2-iso1:ACBI1:VHL	116,254	20	10
Sys2	SMARCA2-iso1:ACBI1:VCB	220,573	24	12
Sys3	SMARCA2-iso2:ACBI1:VHL	117,256	20	10
Sys4	SMARCA2-iso2:ACBI1:VCB	234,724	24	12
Sys5	SMARCA2-iso2:PROTAC 1:VHL	137,347	20	10
Sys6	SMARCA2-iso1:PROTAC 2:VHL	69,696	20	10
Sys7	SMARCA2-iso2:PROTAC 2:VHL	68,820	20	10
Sys8	SMARCA2-iso2:PROTAC 2:VCB	119,082	24	12

identified, k was chosen to be the number of local maxima of the probability distribution in the PCA feature space. Interestingly, the centroids determined by k -means approximately coincided with such local maxima, consistent with the interpretation of the centroids as local minima in the free energy landscape.

To prepare the Folding@home simulations, HREMD data was featurized with interface distances and its dimensionality reduced with PCA as described above. The trajectory was then clustered into 98 k-means states, whose cluster centers were selected as 'seeds' for Folding@home massively parallel simulations. The simulation systems and parameters were kept the same as for HREMD and loaded into OpenMM where they were energy minimized and equilibrated for 5 ns in the NPT ensemble ($T = 310$ K, $p = 1$ atm) using the openmmtools Langevin BAOAB integrator with 2 fs time step. 100 trajectories with random starting velocities were then initialized on Folding@home for each of the seeds. The final dataset consists of 9800 trajectories, 5.7 milliseconds of aggregate simulation time, and 650 ns median trajectory length. This dataset is made publicly available at:

<https://console.cloud.google.com/storage/browser/paperdata>.

For computational efficiency, the data was strided to 5 ns/frame, featurized with closest heavy atom interface distances (as described above), and projected into Time-structure Inde-

pendent Components Analysis (TICA) space at lag time 5 ns using commute mapping. The dimensionality of the dataset was reduced to 339 dimensions, keeping the number of TICA necessary to explain 95% of kinetic variance. The resulting TICA space was discretized into 1000 microstates using k-means. The MSM was then estimated from the resulting discretized trajectories at lag time 50 ns using a minimum number of counts for ergodic trimming (i.e. the 'mincount_connectivity' argument in PyEMMA) of 4, as the default setting resulted in a trapped state whose connectivity between simulation sub-ensembles starting from two different seeds was observed only due to clustering noise. The validity of the MSM was confirmed by plotting the populations from raw MDcounts vs. equilibrium populations from the MSM, which is a useful test, especially when multiple seeds are used and the issue of connectivity is paramount. A hidden Markov model (HMM) was then computed using 5 macrostates to coarse-grain the transition matrix.

5.2.2.9 Calculating Interface RMSD

The quality of the structures produced by the REVO simulations were judged based on the similarity between the interface of the simulated SMARCA2-VHL complex to the bound ensemble. The SMARCA2-VHL interface was defined as residues between the two proteins with a maximum distance of 10 Å. The backbone of these residues are superimposed onto the reference structure. The interface root mean square deviation (I-RMSD) is then defined as the RMSD of C_α atoms between the simulated structure and the reference structures. This calculation is calculated for every structure in the bound ensemble and the minimum value is reported.

Table 5.3: Details of HREMD simulations. Effective temperatures and average exchange probabilities of neighboring replicas are listed.

ID	Eff. temperature, T_i (K)	Avg. exchange prob. (p)
Sys1	300, 306, 311, 317, 323, 329, 335, 341, 347, 354, 360, 367, 374, 381, 388, 395, 402, 410, 417, 425	0.30, 0.30, 0.30, 0.29, 0.29, 0.29, 0.29, 0.30, 0.31, 0.29, 0.29, 0.31, 0.32, 0.31, 0.29, 0.31, 0.29, 0.31, 0.33
Sys2	300, 305, 309, 314, 319, 324, 329, 334, 339, 344, 349, 354, 360, 365, 371, 377, 382, 388, 394, 400, 406, 412, 419, 425	0.27, 0.29, 0.29, 0.32, 0.31, 0.29, 0.32, 0.32, 0.29, 0.39, 0.36, 0.29, 0.30, 0.30, 0.30, 0.30, 0.31, 0.31, 0.29, 0.31, 0.31, 0.31, 0.34
Sys3	300, 306, 311, 317, 323, 329, 335, 341, 347, 354, 360, 367, 374, 381, 388, 395, 402, 410, 417, 425	0.35, 0.31, 0.31, 0.29, 0.30, 0.32, 0.30, 0.32, 0.33, 0.34, 0.33, 0.35, 0.35, 0.35, 0.32, 0.34, 0.33 0.35, 0.39
Sys4	300, 305, 309, 314, 319, 324, 329, 334, 339, 344, 349, 354, 360, 365, 371, 377, 382, 388, 394, 400, 406, 412, 419, 425	0.37, 0.38, 0.30, 0.30, 0.39, 0.39, 0.30, 0.34, 0.30, 0.30, 0.31, 0.28, 0.37, 0.39, 0.39, 0.36, 0.38, 0.31, 0.31, 0.31, 0.41, 0.32, 0.33
Sys5	300, 306, 311, 317, 323, 329, 335, 341, 347, 354, 360, 367, 374, 381, 388, 395, 402, 410, 417, 425	0.31, 0.32, 0.33, 0.34, 0.31 0.35, 0.35, 0.30, 0.36, 0.32, 0.36, 0.36, 0.32, 0.31, 0.31, 0.35, 0.35, 0.33, 0.38
Sys6	300, 306, 311, 317, 323, 329, 335, 341, 347, 354, 360, 367, 374, 381, 388, 395, 402, 410, 417, 425	0.28, 0.27, 0.30, 0.28, 0.28 0.28, 0.29, 0.29, 0.30, 0.29, 0.31, 0.31, 0.31, 0.29, 0.30, 0.29, 0.32, 0.32, 0.30
Sys7	300, 306, 311, 317, 323, 329, 335, 341, 347, 354, 360, 367, 374, 381, 388, 395, 402, 410, 417, 425	0.29, 0.30, 0.30, 0.29, 0.31, 0.30, 0.29, 0.29, 0.32, 0.30, 0.34, 0.31, 0.30, 0.32, 0.31, 0.34, 0.33, 0.34, 0.32
Sys8	300, 305, 309, 314, 319, 324, 329, 334, 339, 344, 349, 354, 360, 365, 371, 377, 382, 388, 394, 400, 406, 412, 419, 425	0.26, 0.28, 0.30, 0.28, 0.34, 0.27, 0.34, 0.32, 0.25, 0.34, 0.32, 0.33, 0.31, 0.30, 0.31, 0.27, 0.23, 0.25, 0.32, 0.31, 0.31, 0.29, 0.27

5.3 Results

5.3.1 Degraders with different efficiency induce similar ternary complex structures in X-ray crystallography.

The ternary complexes of SMARCA2 isoform 2 and the VHL/ElonginC/ElonginB (VCB) induced by different heterobifunctional degraders have been studied extensively [245, 208]. In particular, PROTAC 1, PROTAC 2, and ACBI1 are three prominent degrader molecules that induce a ternary SMARCA2 isoform 2:VCB complex and have quite different degradation efficiencies (see Table 5.1). Whereas crystal structures of the ternary complexes induced by PROTAC 1 (PDB ID: 6HAY) and PROTAC 2 (PDB ID: 6HAX) exist, none has been reported to date for ACBI1, the most potent degrader among them. Thus, to study the effect of different degraders on the ternary complex, we determined, as a first step, the structure of SMARCA2 isoform 2:VHL liganded by ACBI1 via X-ray crystallography. The structure was obtained using similar conditions as reported before (see Methods) [208] and solved by molecular replacement to 2.25 Å in the highest resolution shell (Table 5.4), using 6HAX as the search model (Figure 5.3a).

The degrader molecule bridges the induced interface, forming contacts with both proteins. Importantly, the ligand induces “cooperative contacts” between several amino acids of the two proteins, such as VCB:ARG69 and SMARCA2 isoform 2:PHE1463 (Figure 5.3 b,c). SMARCA2 isoform 2:ASN1464 makes critical bivalent contacts to the aminopyridazine group of ACBI1, positioning the terminal phenol group for pi stacking interactions with residues PHE1409 and TYR1421 (Figure 5.3b,c).

On the VHL side of the interface, the interactions between TYR98 and ACBI1 are consistent with those between the same residue and PROTAC 1 or PROTAC 2 (Figure 5.3b,c) [208].

The three degraders PROTAC 1, PROTAC 2, and ACBI1 bind to VHL in a near-identical fashion as their superposition reveals, upon aligning the VHL protein (see Figure 5.3d).

Table 5.4: Crystallographic table for protein crystal structure 7S4E
SMARCA2-iso2:ACBI1:VHL.

Smarca 2 ^{BD} : ACBi1 : VCB	
Data collection	
Space Group	P 21 21 21
Cell Dimension	
a, b, c, (Å)	80.14, 116.57, 122.32
α, β, γ (°)	90, 90, 90
Resolution (Å)	2.25 (2.31-2.25)*
R _{merge}	0.15
$\langle I/\sigma \rangle$	1.27*
Completeness (%)	99.9(37.89-2.25)
Redundancy	7.4
Refinement	
Resolution (Å)	2.25
No. Reflections	52206
R _{work} /R _{merge}	21.9/25.9
No. Atoms	7356
Protein	7070
Ligand/Ions	196
Water	90
B factors	
Mean B value	61.19
Ligand/Ions	57.66
Water	53.56
R.M.S deviations	
Bond length (Å)	0.009
Bond angles (Å)	1.519

* Denotes values obtained for the highest-resolution shell.

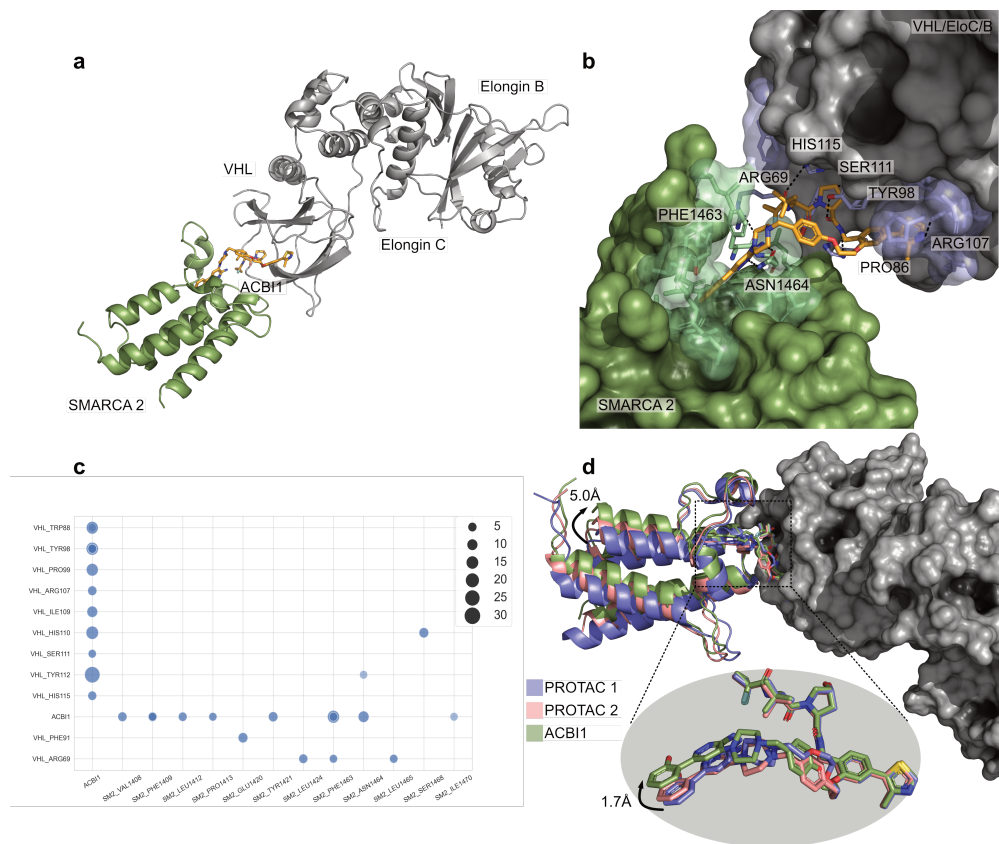


Figure 5.3: Ternary complex of SMARCA2 and VCB induced by ACBI1 shows structural similarities with PROTAC 1 and PROTAC 2: **a** Overall perspective of SMARCA2 Isoform 2 (green) and VHL/ElonginC/ElonginB (grey) induced by degrader molecule ACBI1 (bright orange). **b** ACBI1-induced interface contacts between SMARCA2 and VCB. The proteins are shown in space-filling, the colors are as in **a**, annotated residues are among those that make the highest number of contacts (see **c**). **c** A contact map for the interface of the crystal structure. The circle size reflects the number of atoms (including hydrogen atoms) participating in interactions. **d** Superposition of 6HAY (purple), 6HAX (salmon), 7S4E (green) by aligning VHL (grey) shows varied conformations of the warheads of the three degraders PROTAC 1, PROTAC 2, or ACBI1 (up to 1.7 Å) resulting in alterations of SMARCA2 within the ternary complex.

Nonetheless, the minor differences in the linker compositions, e.g. the ACBI1 linker has one additional ether group compared to the PROTAC 2 linker, yields a slight 1.7 Å twist of ACBI1 compared to the other two degraders, resulting in a subtle 5 Å “swing” of the protein (Figure 5.3d).

Our results show that, despite the differences in the linker compositions, the protein-protein interface induced by ACBI1 is structurally similar to that induced by PROTACs 1 or 2 [208]. Notwithstanding, the markedly different degradation efficiencies between these degraders[208] suggest that the (dynamic) ensemble of ternary complex structures may be fairly different among them. Consistent with other studies [207, 194], this implies that “crystallographic snapshots” are not suitable to provide a holistic view of the ensemble of all possible ternary complex structures in solution, but merely represent a subset of the relevant conformations favored by crystallization [246]. Consequently, such X-ray structures cannot fully capture the dynamic nature of the degrader-induced ternary complexes, which ultimately determines their activities and degradation efficiencies [207, 194].

5.3.2 Hydrogen Deuterium Exchange Reveals Extended Protein-Protein Interfaces

In order to assess the impact of different degrader molecules on the dynamic nature of the SMARCA2 isoform 2:VCB interactions, we performed hydrogen-deuterium exchange of the respective APO, binary and ternary (complex) species, thus characterizing the protein-protein interface in solution. This approach is a promising alternative to previous attempts at characterizing degrader ternary complexes that employed multiple crystal structures [207], nuclear magnetic resonance (NMR) [194]. Based on previously established protocols [221, 247, 223], and with the knowledge of binding constants for each of the three degraders, complex formation was determined and each protein was also found to have a 100 % sequence coverage (see 5.4) and stable deuterium exchange (see 5.5 and 5.6). To ascertain the degree of protection from solvent in the binary or ternary complex, the residue-specific uptake of

the APO or binary species was subtracted from that of the corresponding residues in the binary or ternary state (referred to as Binary Δ APO and Ternary Δ Binary), respectively. The results are summarized in difference plots that highlight the statistically significant (95% or 98% confidence interval) protection of distinct sites (see Figure 5.8a-d for the SMARCA2 isoform 2:VCB complex induced by ACBI). Importantly, protection during HDX arises due to changes in the environment around the observed residues, which could be a result of direct occlusion of solvent or conformational changes [225].

Figure 5.8a reveals that large regions of SMARCA2 isoform 2 become protected upon ternary complex formation (see Ternary Δ Binary difference plot). These stretches of protected residues, e.g. amino acids 1409-1422 and 1456-1470, overlap with the ligand binding site based on the ternary complex structure published in this work (7S4E) and those published previously (6HAY, 6HAX), which confirms the similarity of the ternary complex interface among the three degrader molecules discussed above. Additionally, there are also stretches of protected amino acids, 1394-1407, that are too distant from the established binding interface to result from complex formation (Figure 5.8a and f). Interestingly, the Binary Δ APO difference plot suggests that the SMARCA2 isoform 2-ligand binary complex was unstable under our experimental conditions (the ligand concentration is close to the dissociation constant), as there is no substantial difference between the HDX of SMARCA2 isoform 2 in presence and in absence of the ligand (Figure 5.8a and e).

On the VCB side of the interface, large regions of VHL, which resides at the direct interface of the protein complex, are protected in the presence of the ligand as indicated by the Binary Δ APO difference plot (Figure 5.8b and e). The most protected residues in the binary state are centered around amino acids 87-116, which include all 9 residues in the ligand binding site of VHL. However, there is a large amount of protection across the entire protein indicating the presence of a large allosteric network across the protein [248]. In the presence of SMARCA2 isoform 2 (see Figure 5.8b, Ternary Δ Binary difference plot), much of the allosteric network due to ligand binding can be subtracted away leaving only the most

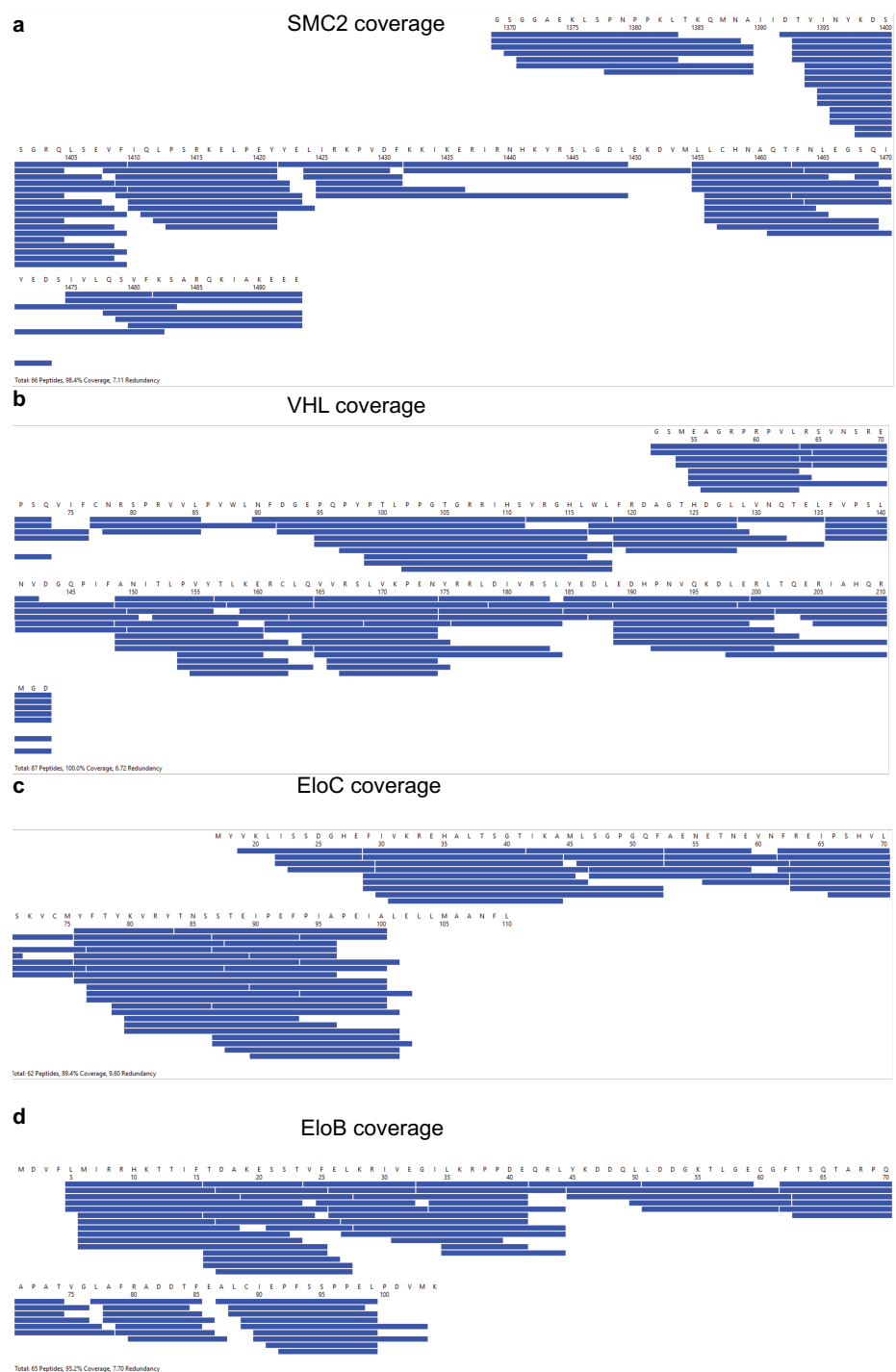


Figure 5.4: Peptic coverage map of proteolyzed proteins SMARCA2, VHL, Elongin C and Elongin B.

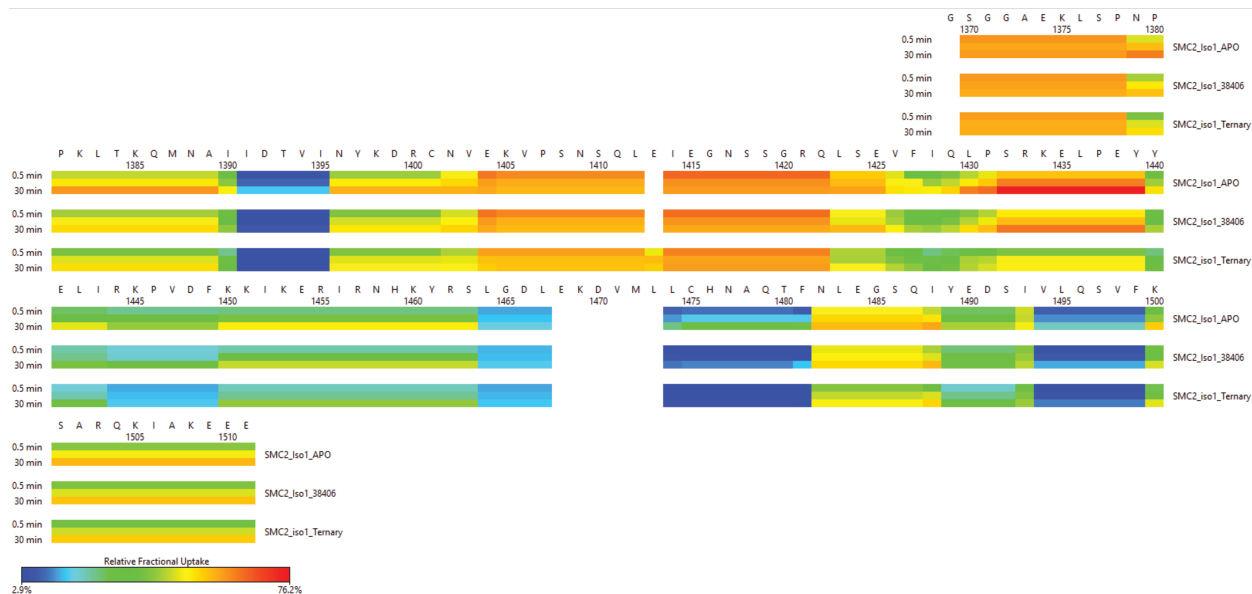


Figure 5.5: Relative uptake heat map of HDX exchange data of all PROTAC molecules 1, 2 and ACBI1 bound to binary and Ternary State SMARCA2 isoform 2 bromo domain.

significantly protected residues due to ternary complex formation (Figure 5.8b and f). In particular, residues 60-72, which house the critical interaction of ARG69 show significant protection due to ternary complex formation (Figure 5.8b and f).

Additionally, we observe continued protection of residues 166-176 and residues 187-201 on VHL (Figure 5.8b and f) as well as some regions on Elongin B and C that show protection upon ternary complex formation (see Figure 5.8c and d). Although these sites are distal from the binding interface, they spatially align with one another when mapped onto the structure (Figure 5.8c) potentially uncovering a critical network of allosteric changes induced by ACBI1.

These changes that are not observed in the snapshot provided by crystallography help explain how this molecule vastly improves biological function, i.e., perhaps by changing the orientation of the ligase to target lysine residues on SMARCA2 isoform 2, as suggested by our simulations of the SMARCA2 isoform 2:VCB complex (see below).

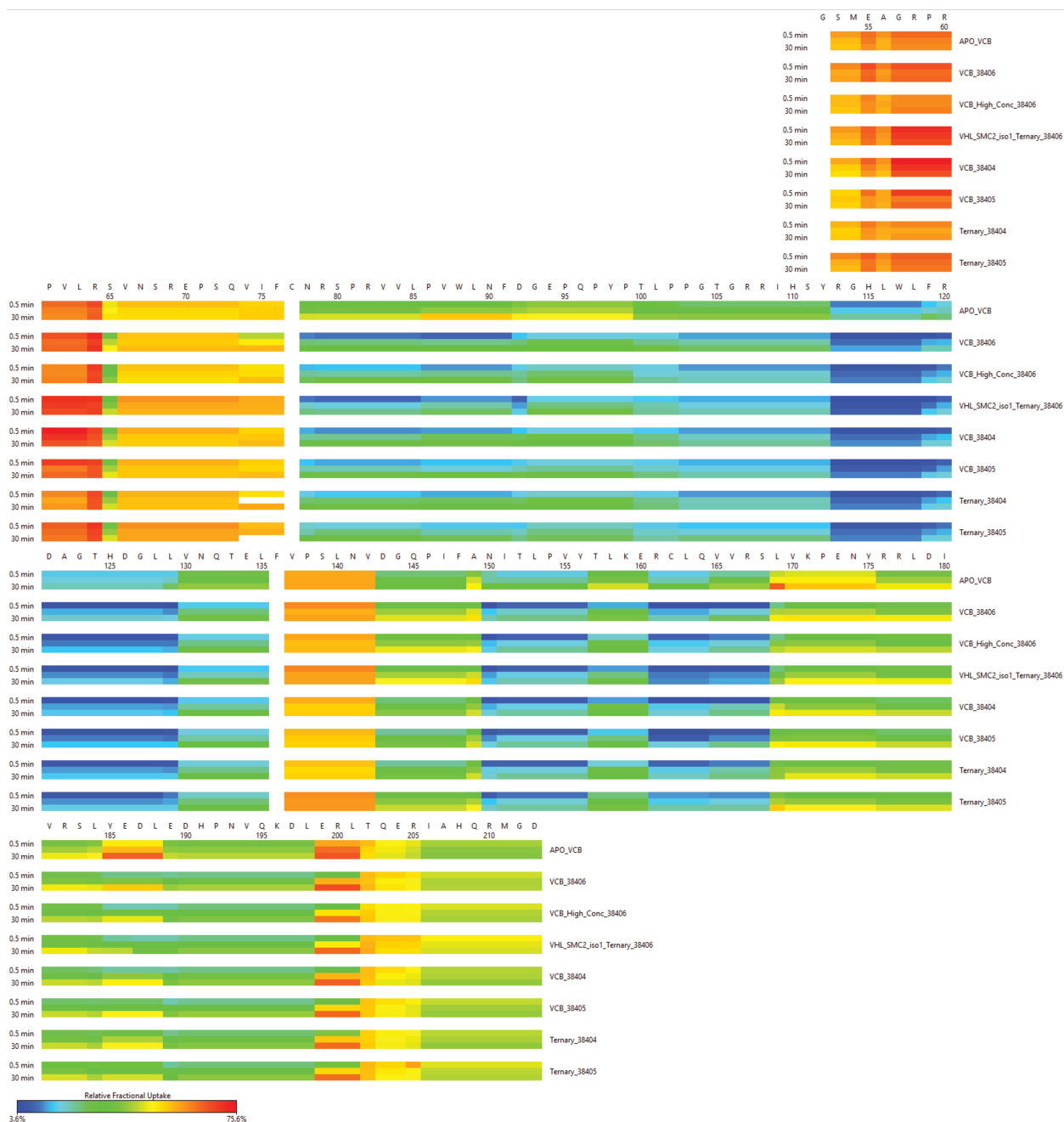


Figure 5.6: Relative uptake heat map of HDX exchange data of all PROTAC molecules 1, 2 and ACBI1 bound to binary and Ternary State VHL.

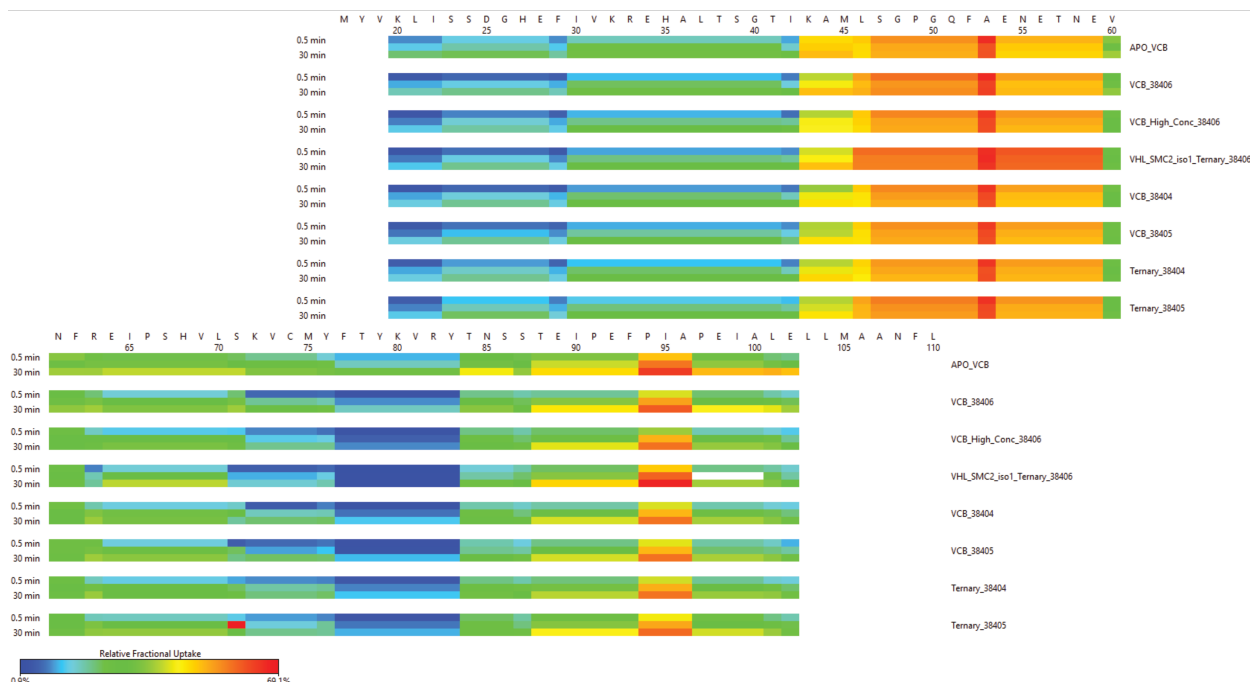


Figure 5.7: Relative uptake heat map of HDX exchange data of all PROTAC molecules 1, 2 and ACBI1 bound to binary and Ternary State Elongin C.

5.3.3 Efficient simulation of ternary complex formation using REVO Weighted Ensemble simulations

We simulate the formation of a degrader ternary complex with the weighted ensemble path sampling approach. This method has been employed before for tasks such as protein-protein binding [249]. It is noteworthy, however, that the pre-defined CVs in the current simulations are *not* informed by structural data about the ternary complex interface from X-ray crystallography experiments. In particular, we employ the WE variant: REVO[35], which optimizes an objective function called the trajectory variation (see Methods). Here we compare the performance of the REVO algorithm with three different distance metrics: 1) differences in the warhead RMSD, (RMSD between the target binding domain of the PROTAC between the reference structure and a simulation frame, hereby called w-RMSD); 2) differences in target-ligase contacts; 3) a weighted combination of metrics 1, 2, and the differences in contacts between the target and the PROTAC for the PROTAC 2 system, hereby called the

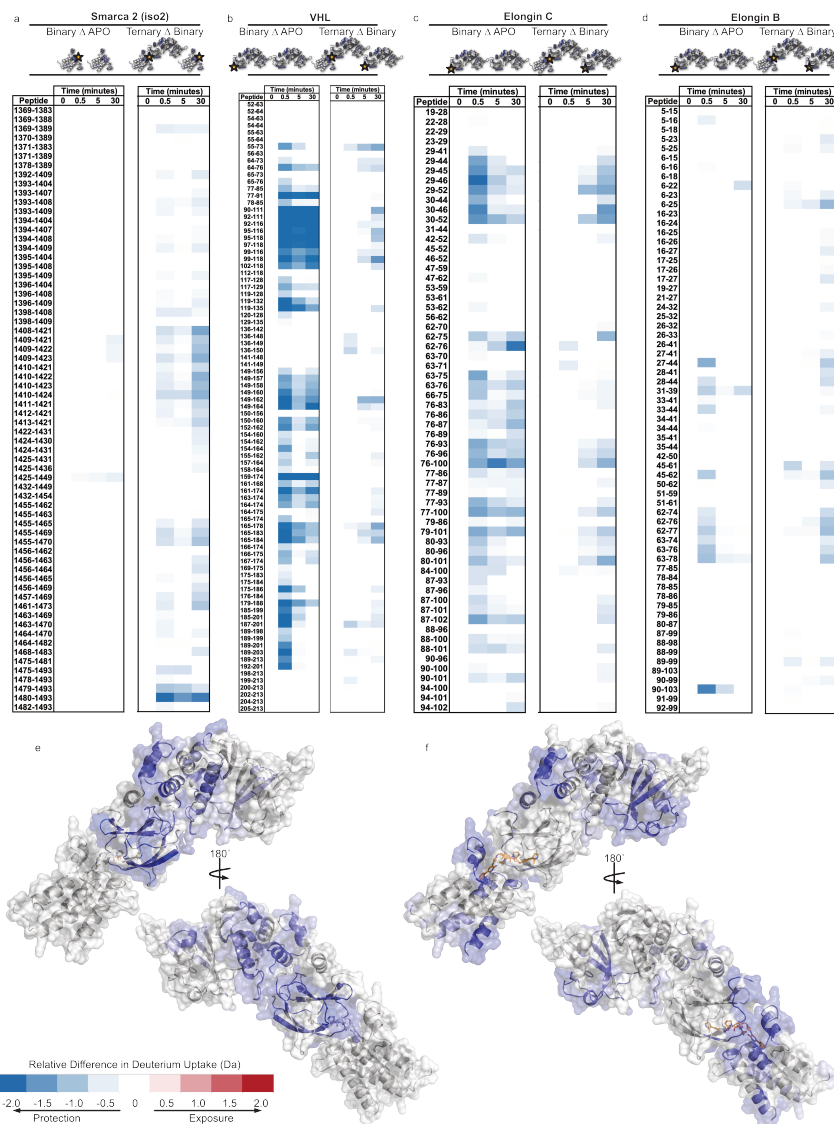


Figure 5.8: ACBI-induced ternary complex formation of SMARCA2 isoform 2:VCB leads to protection of specific sites:a-d, SMARCA2 isoform 2(a), VHL(b), Elongin C(c), and Elongin B(d) monitored for hydrogen-deuterium exchange over time. The difference plots of each protein in the binary and ternary states are generated by subtracting the deuterium exchange of like peptides of the APO or binary from the binary or ternary states (defined as Binary Δ APO and Ternary Δ Binary), respectively. Regions that exchange significantly less than the comparative state are depicted in blue (negative), whereas regions that exchange significantly more appear in red (positive). The resultant difference plots of the binary (e), or ternary complex (f) were mapped onto the structure 7S4E. The experiments were repeated on 2 separate days.

triple distance metric. The residues selected to compute the target-ligase contacts used in distance metrics 2 and 3 were selected based on residues that showed increased protection from hydrogen-deuterium exchange based on the HDX experiments. Between three and seven REVO simulations are run for each distance metric on the SMARCA2-VHL-PROTAC 2 system, and a summary of their performance is given in Table 5.5.

Table 5.5: A summary of the performance of REVO simulations run with different distance metrics. Each REVO simulation ran with 48 walkers. The number of binding events (N_{binding}) counts the barrier crossings into the bound state, defined using an I-RMSD $< 2.0 \text{ \AA}$. The number of simulations with binding events (Sims. w/ binding) shows the probability of binding success. The total simulation time (Sim. time) aggregates the length of all trajectories in each REVO simulation.

Distance Metric	N_{binding}	Sims. w/ binding	Sim. time (μs)
W-RMSD	5327	6/7	13.44
Target-Ligase Contacts	5876	2/7	13.44
Triple	3278	6/7	12.5

Encouragingly, we observe a large number of binding events with all three distance metrics examined here. The triple distance metric found the least number of binding events at 3278, whereas the target-ligase contacts distance metric sees the most binding events (5876 binding events).

We find that the low I-RMSDs are achieved early in the REVO simulations, with the minimum I-RMSD reaching $\sim 2.5 \text{ \AA}$ and stabilizing after about 500 ns of aggregate simulation time, whereas the vanilla MD simulation we ran plateaued at 10 \AA after $1.35 \mu\text{s}$ of simulation (Figure 5.9a,b). When comparing between the distance metrics both w-RMSD and the triple distance metrics were able to find sub- 2 \AA structures within an aggregate simulation time of 200 ns, whereas it took the target-ligase contacts distance metric 800 ns to find structures of the same quality (Figure 5.9b). This indicates that using the degrader orientation to the binding site may be required to quickly and consistently generate low I-RMSD structures.

Figure 5.10 shows an example of a structural prediction obtained for SMARCA2 isoform 2-PROTAC 2-VHL (PDB ID 6HAX [208]). The contact maps presented in Figure 5.10c have been obtained by the Arpeggio software [250] applied to the ternary interfaces. Each

point on the contact maps reflects the degree of interaction. As can be seen from both the aligned prediction and co-crystallized structure (Figure 5.10d) and from the contact map (panel (Figure 5.10c), the accuracy of prediction is very high.

We performed clustering on all structures produced by the 6HAX simulations by a k-means clustering algorithm into 500 macrostates using the $C\alpha - C\alpha$ distances on residues determined from the HDX experiments. Low I-RMSD states all have low values of w-RMSD (as expected) (Figure 5.9c). High free energy states have a large range of both I-RMSD values. However, the low free energy states are coincidentally below 10 Å. Five of the lowest free energy states have I-RMSD values below 3 Å.

Determining the I-RMSD of the ensemble is only possible when we have a crystal structure for the ternary complex. When trying to filter many possible degraders, it is not always feasible to solve these structures for every compound. Therefore, we need to rely on other physical quantities for predicting acceptable structures in such cases. We developed two definitions of the bound ensemble for the REVO simulations: 1) Using w-RMSD below 2 Å and 2) Using w-RMSD below 2 Å and more than 30 residue-residue contacts between the target and ligase residues that showed increased protection from hydrogen-deuterium exchange obtained from the HDX experiments. The first definition was used on the simulations where we used the warhead distance metric and the second definition was used for the target-ligase contacts and triple distance metrics. Using these definitions of the bound ensemble to filter our simulations, REVO simulations with and without HDX were able to sample low I-RMSD regions (below 2 Å), and the probability distributions had peaks below this I-RMSD threshold (Figure 5.11a). However, using the HDX data limited the bound ensemble to a maximum I-RMSD of about 4 Å whereas not including the HDX data allowed a broad bound distribution that had a maximum at 8 Å. Using this definition for the bound ensemble, 43% of structures that met this criteria had I-RMSD values below 2 Å, whereas the bound ensemble generated from REVO simulations not using HDX only selected conformations with an I-RMSD below 2 Å at 38% (Figure 5.11b). This definition of the bound

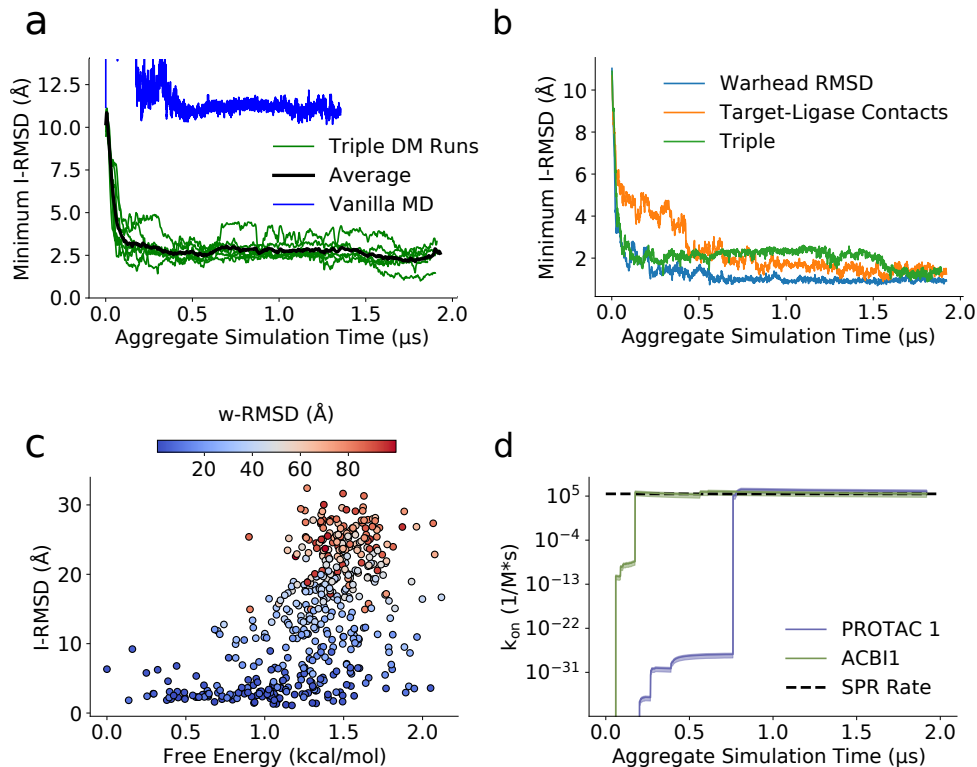


Figure 5.9: Comparing the w-RMSD, number of target-ligase contacts, and triple distance metrics (Linear combination of w-RMSD, target-ligase contacts and number of target-PROTAC contacts). (a) The minimum I-RMSD over time during the simulation for the triple distance metric. Each green line indicates one replica and the black line is the average between all runs. The blue line is a straightforward MD simulation run on Folding@home. (b) The minimum I-RMSD for each distance metric. (c) A scatter plot of the free energy vs the I-RMSD after clustering the 6HAX simulations. The circles are colored by w-RMSD. (d) The predicted binding rates for PROTAC 1 system (purple) and the ACBI1 system (green). The black line is the experimental on-rate determined via SPR.

ensemble had 90% of structures below 3 Å I-RMSD for REVO+HDX simulations. However, only 51% of structures in the REVO bound ensemble were below 3 Å. It is worth noting that both the target-ligase contacts and triple distance metric both use HDX data to help guide the simulations. However, the target-ligase contacts metric did not produce low w-RMSD structures, the lowest being just below 6 Å and thus did not contribute to the bound ensemble via this definition. Using this definition for the bound ensemble, we find that adding the HDX data during REVO, we have a higher likelihood of finding structures at lower I-RMSD compared to when we run simulations without HDX.

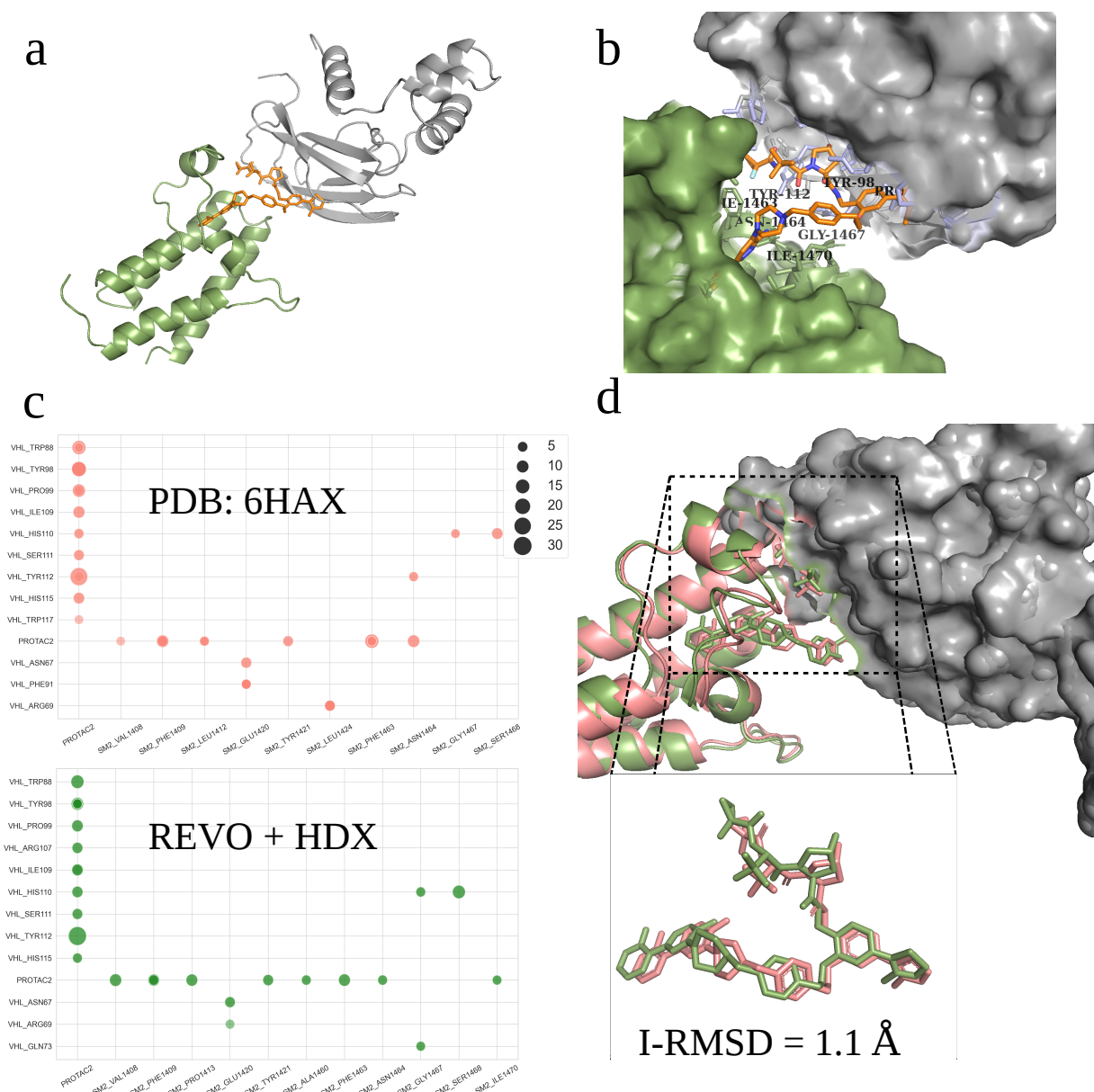


Figure 5.10: Illustration of the representative prediction produced by REVO simulation and its comparison to the co-crystallized structure (PDB ID: 6HAX) (a) predicted ternary structure with I-RMSD=1.1 Å; (b) detail of the binding interface; (c) contact maps for the interfaces of co-crystallized and predicted structures. The circle size reflects the number of atoms (including hydrogens) participating in interactions; (d) structurally aligned prediction (green) and co-crystallized structure (pink) with a detailed PROTAC 2 comparison shown.

All the above analysis was done using the PROTAC 2 system. We also performed three 1.96 μs simulations for the PROTAC 1 and ACBI1 systems using the triple distance metric, totaling 5.88 μs for each system. All the simulations for these two systems were able to produce structures of quality I-RMSD, the lowest being 0.69 Å for PROTAC 1 and 0.47 Å for ACBI1.

We next predict the on-rates on the three different PROTACs (Table 5.6) using the flux into the bound state as defined when the state reaches an I-RMSD below 2 Å. For PROTAC 1 and ACBI1, our predicted rates are on the same order of magnitude as experiment (Figure 5.9d). For PROTAC 2, we were unable to experimentally determine the on-rate so we simply report the rate predicted by simulation. However, for all three rates there were large errors. This is due to the weighted ensemble algorithm being slow to converge. To obtain better statistics, more simulation time is needed.

Table 5.6: Comparison of k_{on} rates between simulation and experiment for the ACBI1 PROTAC 1, and PROTAC 2 systems. The experimental rate for PROTAC 2 has not been determined yet.

PROTAC	Predicted Rate ($M^{-1}s^{-1}$)	Experimental Rate ($M^{-1}s^{-1}$)
ACBI1	$3 * 10^5 \pm 2 * 10^5$	$2.4 * 10^5$
PROTAC 1	$10 * 10^5 \pm 8 * 10^5$	$2.9 * 10^5$
PROTAC 2	$2.2 * 10^2 \pm 1.7 * 10^2$	N/A

5.3.4 HDX improves prediction of ternary complex using docking

Molecular docking is a very popular method for high-throughput predictions of binding poses, that follows a protocol of sampling, searching, and scoring these predictions. Considering the computational cost of the REVO+HDX method described above, docking is a viable alternative to the simulation approach in obtaining different conformations of the flexible degrader ternary complexes in a less resource-intensive and more timely fashion.

To demonstrate the usefulness of HDX data for more accurate structural predictions, we show that incorporating experimentally retrieved distance restraints into the docking protocol significantly improve its sensitivity (see Figure 5.11). Importantly, unlike Zhang et

al. [251], who derived restraints from chemical cross-linking experiments, or Eron [252], who revised the post-sampling scoring, our approach imposes distance restraints based on the statistics of the linker length in a degrader molecule – at the sampling stage (see Methods).

Figure 5.11a shows the improvement in docking predictions when augmented by HDX data as the distribution of I-RMSDs (with respect to the crystal structure) for the top-100 predictions is shifted toward smaller values upon incorporating the experimentally derived restraints (green compared to orange profile). In particular, when focusing on subsets of highly accurate structure predictions, i.e., I-RMSD $< 2 \text{ \AA}$, 2.5 \AA , or 3 \AA (see Figure 5.11b), for which, as described above, the improvement of REVO upon adding information from HDX was measurable, the performance of the docking protocol is significantly improved. Although REVO+HDX consistently outperforms the HDX-enhanced docking routine, it is striking how strongly the incorporation of HDX data can boost the accuracy of this docking protocol. Therefore, considering the significantly less computational cost of this approach (75 CPU hours for 3 independent replicas) compared to the REVO method (300 A40 GPU hours), docking, in combination with HDX, is a useful tool for the quick filtering of a large number of degrader designs.

5.3.5 Conformational sampling of ternary complexes

Our HDX data suggest that the protein complexes studied here are dynamic and sample several distinct bound conformations. We use HREMD simulations to identify these structures and quantify the free energy landscape of these complexes. First, we perform PCA of the interface distances observed in our HREMD simulations in order to reduce the dimensionality of the simulation data. The probability distribution of the highest-variance features allow us to measure a more easily interpretable free energy landscape from our simulation data than would be possible otherwise. We find that the landscape of each protein complex contained several local minima differing by only a few kcal/mol.

Using k -means clustering in the PCA feature space, we then identify distinct clusters

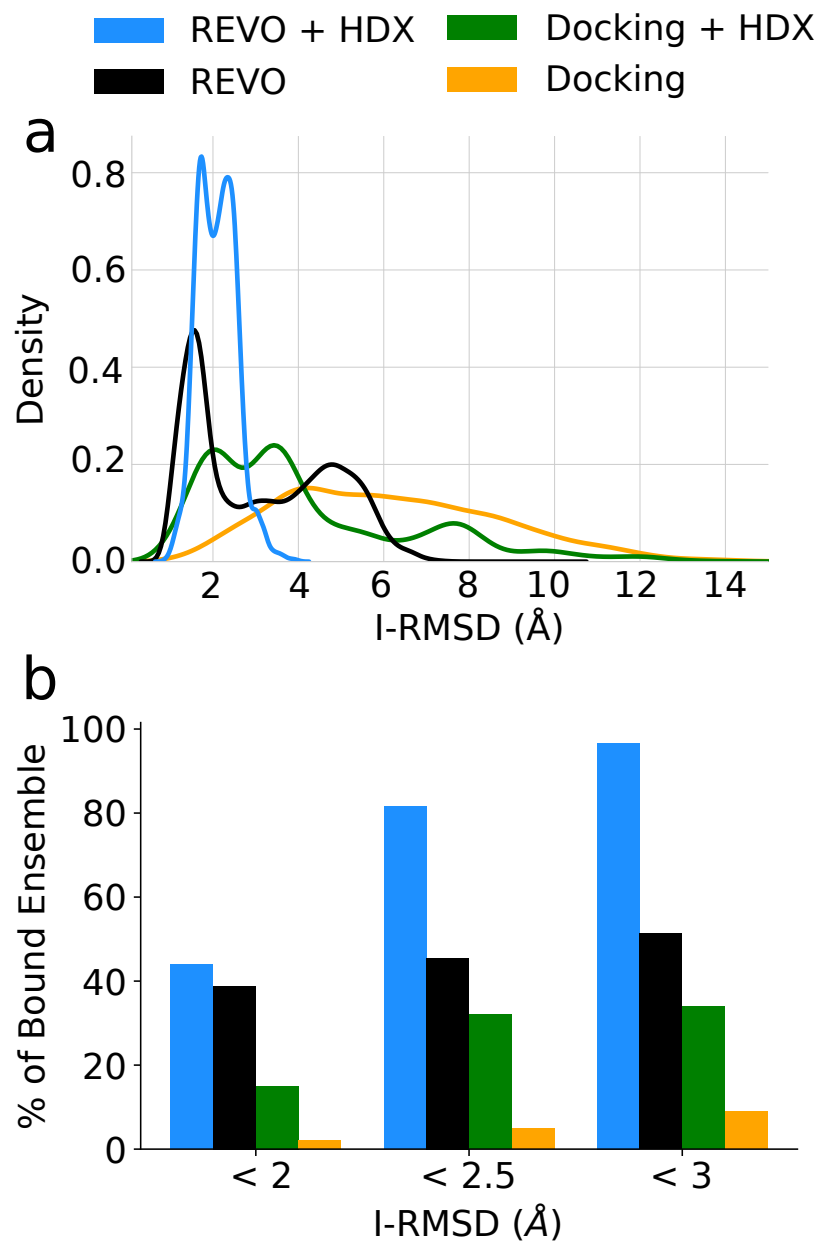


Figure 5.11: Comparing the bound ensembles determined by docking and REVO simulations with and without information from HDX for the PDB ID 6HAX ternary complex. The REVO bound ensemble is defined as structures below a warhead RMSD of 2 Å and more than 30 contacts between the target and ligase interface. The docking bound basin is defined as the 100 top structures as determined by Rosetta-scoring. **(a)** Probability density function distributions of I-RMSD values for the bound ensembles. **(b)** The percent of structures in the predicted bound ensembles below specific I-RMSD thresholds (2 Å, 2.5 Å, and 3 Å).

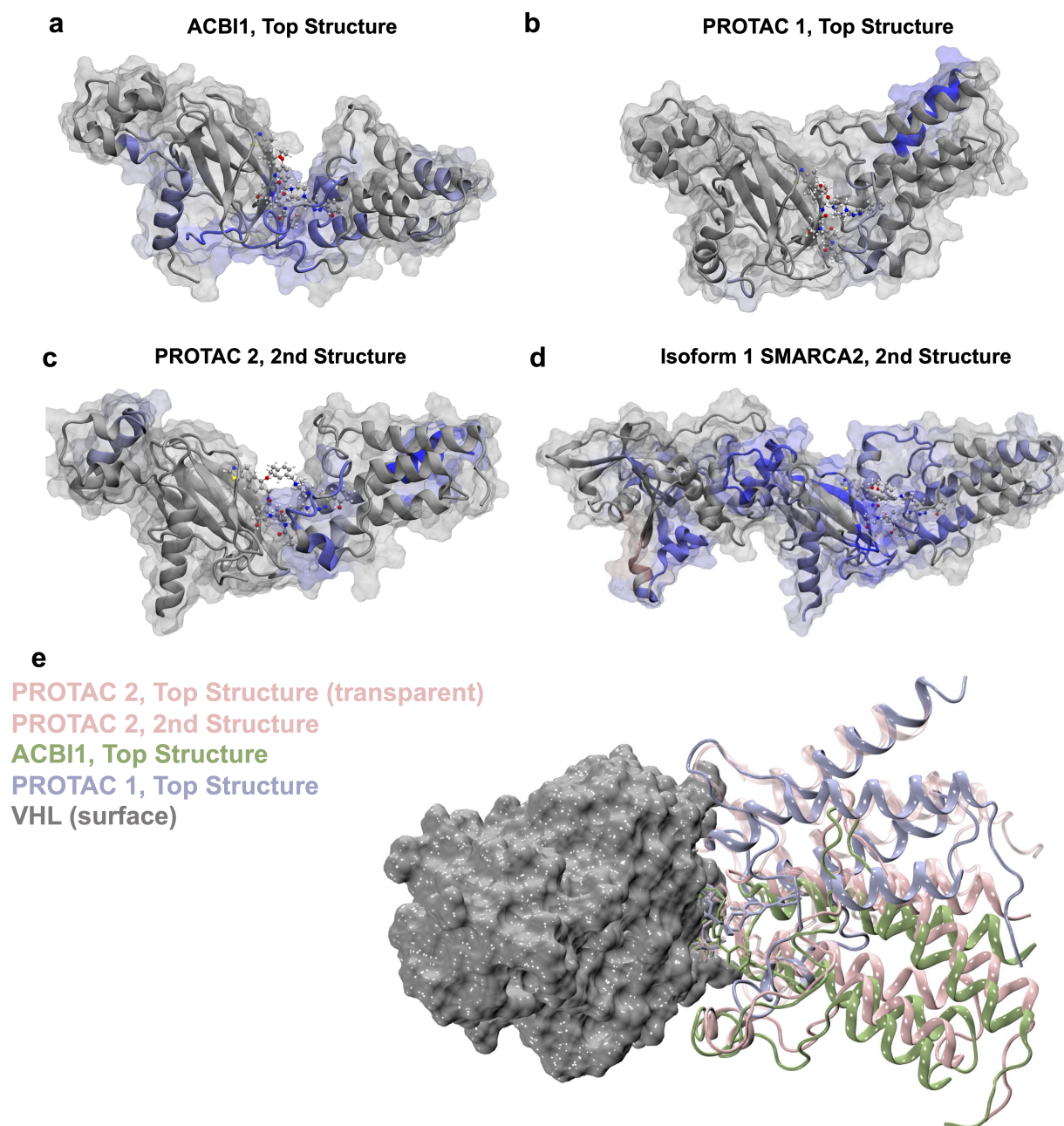


Figure 5.12: Most populated structures of SMARCA2 bound to VHL with different degrader molecules, identified by dimension reduction and clustering of HREMD simulation data. **(a-d)** Colors of VHL and SMARCA2 represent HDX protection in the presence of the degrader molecules relative to the situation in the absence of the degrader. The second ranked structures of **c** PROTAC 2 and **d** isoform 1 SMARCA2 are displayed that support HDX data, whereas the top three structures are included in Figure 5.13. Elongin B and Elongin C are also included in panel **d**. **e** The top structures of ternary complexes are compared after aligning VHL to illustrate conformational differences among top structures of ternary complexes.

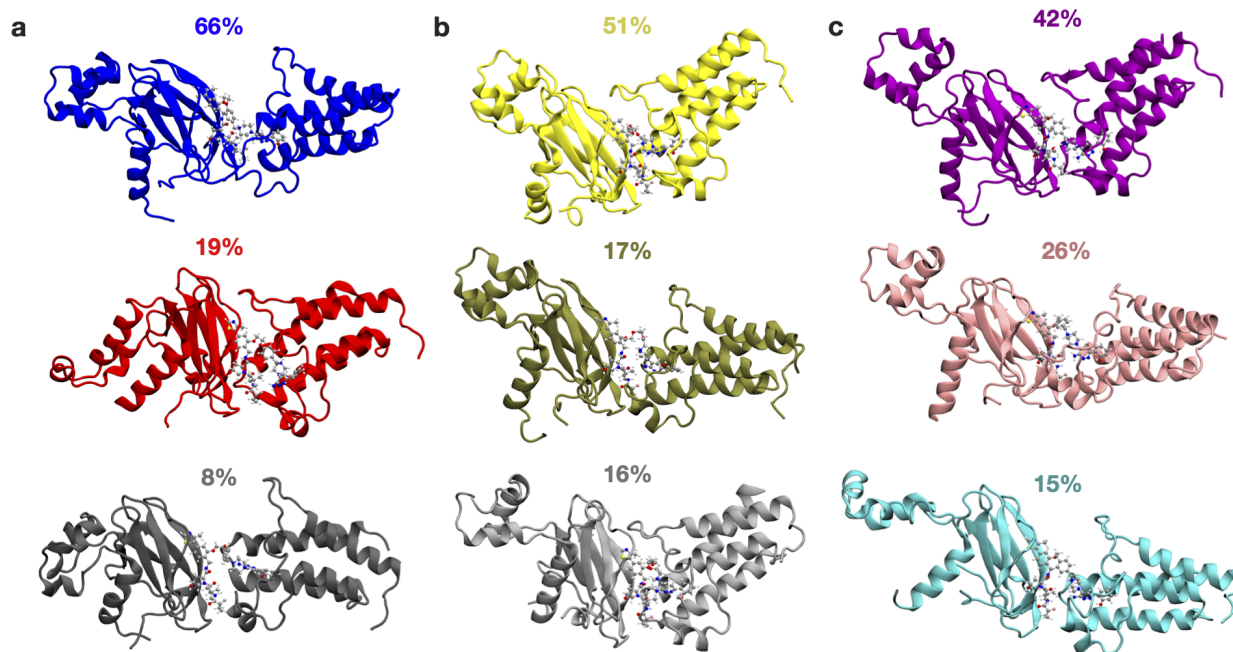


Figure 5.13: Cluster centroids from the three highest populated structures of SMARCA2-iso2 bound to VHL via (a) ACBI1, (b) PROTAC 1, and (c) PROTAC 2, along with their populations. Less populated structures are omitted.

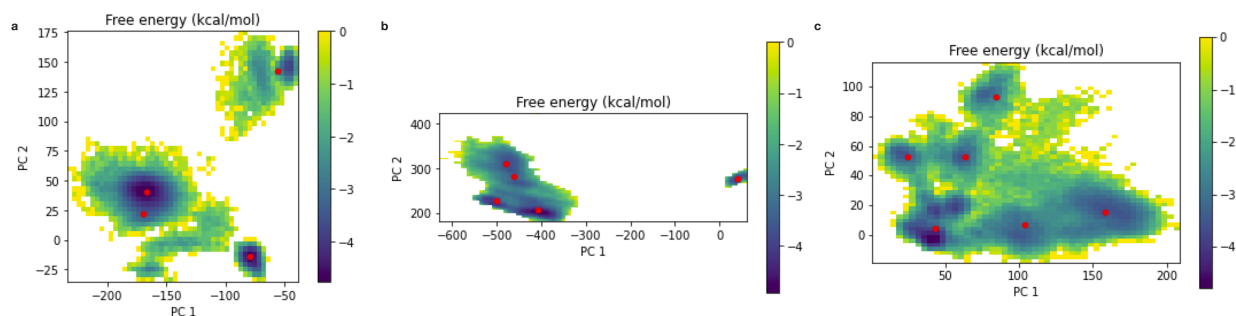


Figure 5.14: Free energy landscapes determined from PCA projections of SMARCA2-iso2 bound to VHL via (a) ACBI1, (b) PROTAC 1, and (c) PROTAC 2. Red points indicate k -means centroids.

of conformations. Cluster centers roughly correspond to local minima in the free energy landscape, see Figure 5.14. The clusters identified by k -means are consistent with our HDX protection data. Figure 5.12 shows that interface residues that were found to be protected in HDX experiments are observed to interact in either the most populated or second most populated cluster identified by k -means. Notably, this analysis shows that in the second most populated structure of Iso1-ACBI1-VCB, the helix formed by the 17 residue extension of isoform 1-SMARCA2 interacts with a beta sheet of VHL, Figure 5.12d, in accordance with HDX experiments that found this beta sheet to be protected in presence of Iso1, but not in the presence of Iso2. Similarly, highly populated structures of Iso2-ACBI1-VHL and Iso2-PROTAC2-VHL show contact between residues that were observed to be protected in HDX experiments with these PROTACs, but not with PROTAC 1, while the most populated structure of PROTAC 1 does not show these contacts.

We selected 98 representative structures from HREMD data to use as initial configurations for Folding@home (F@H) simulations of Iso2-PROTAC2-VHL. Each initial condition was cloned 100 times and run for ~ 650 ns, for a total of ~ 6 ms of simulation data. These independent MD trajectories provide the basis for fitting a MSM[253], which provides a full thermodynamic and kinetic description of the system and allows for the prediction of experimental observables of interest [220]. First, we used time-lagged independent component analysis (tICA) [254] to determine a projection of the underlying MD data. The distance between points in the TICA feature space corresponds roughly to kinetic distance.

The MSM predicts a stationary probability distribution on TICA space that is in general different from the empirical distribution of our simulation data. Interestingly, the MSM predicts that the the crystal structure is 1.5 kcal/mol higher in free energy than the global free energy minimum, while the bound structures obtained from our REVO simulations are $\sim 1.5 - 3.5$ kcal/mol above the global minimum, Figure 5.15a-c. The model also predicts a metastable state with free energy 2.2 kcal/mol (Figure 5.15e).

This model is coarse-grained to obtain a five-state MSM, of which the following three

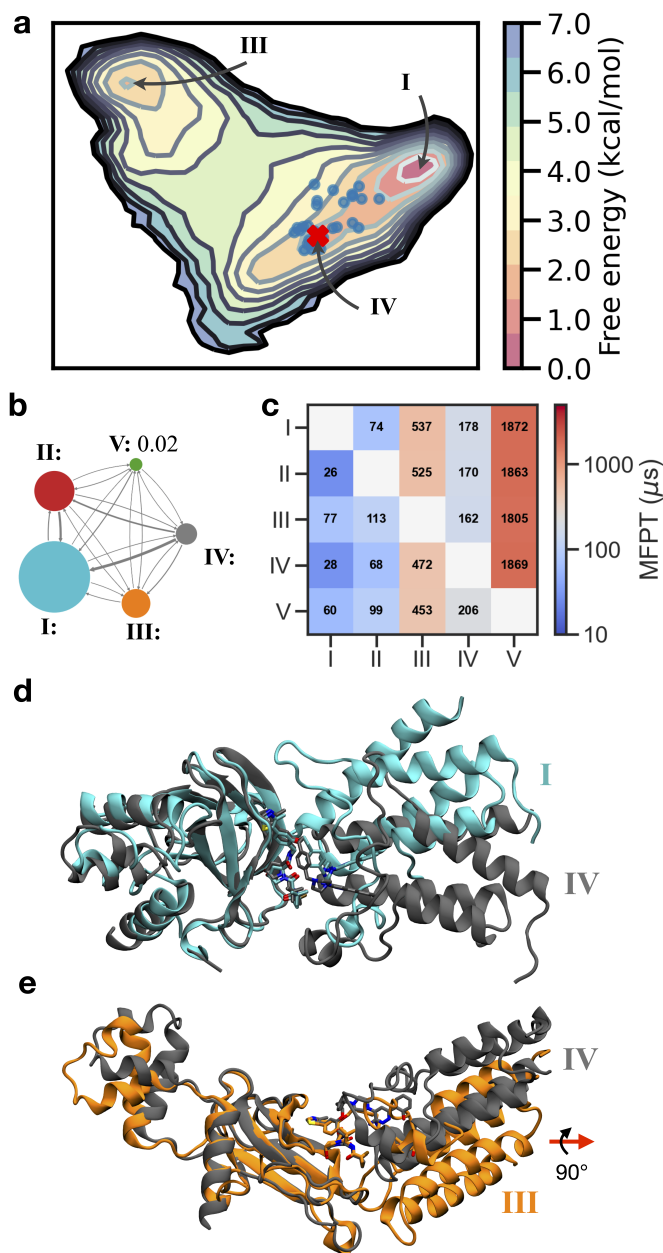


Figure 5.15: **a** Conformational free energy landscape as a function of the first two TICA features of SMARCA2-PROTAC2-VHL ternary complex inferred from a MSM. The ensemble of bound states from REVO simulations is shown as blue points; the crystal structure (PDB ID 6HAX) is shown as a red X. In this projection, states II and V are close to state I. **b** Network diagram of the coarse-grained MSM calculated using a lag time of 50 ns, with the stationary probabilities associated with each state indicated. **c** mean first passage time (MFPT) from one state in the MSM to another. Numbers indicate predicted MFPTs in μs . **d-e** Comparison of the crystal structure (gray) with the lowest free energy state (cyan) and a metastable state (orange) predicted by the MSM. Arrows indicate a change of orientation relative to **d**.

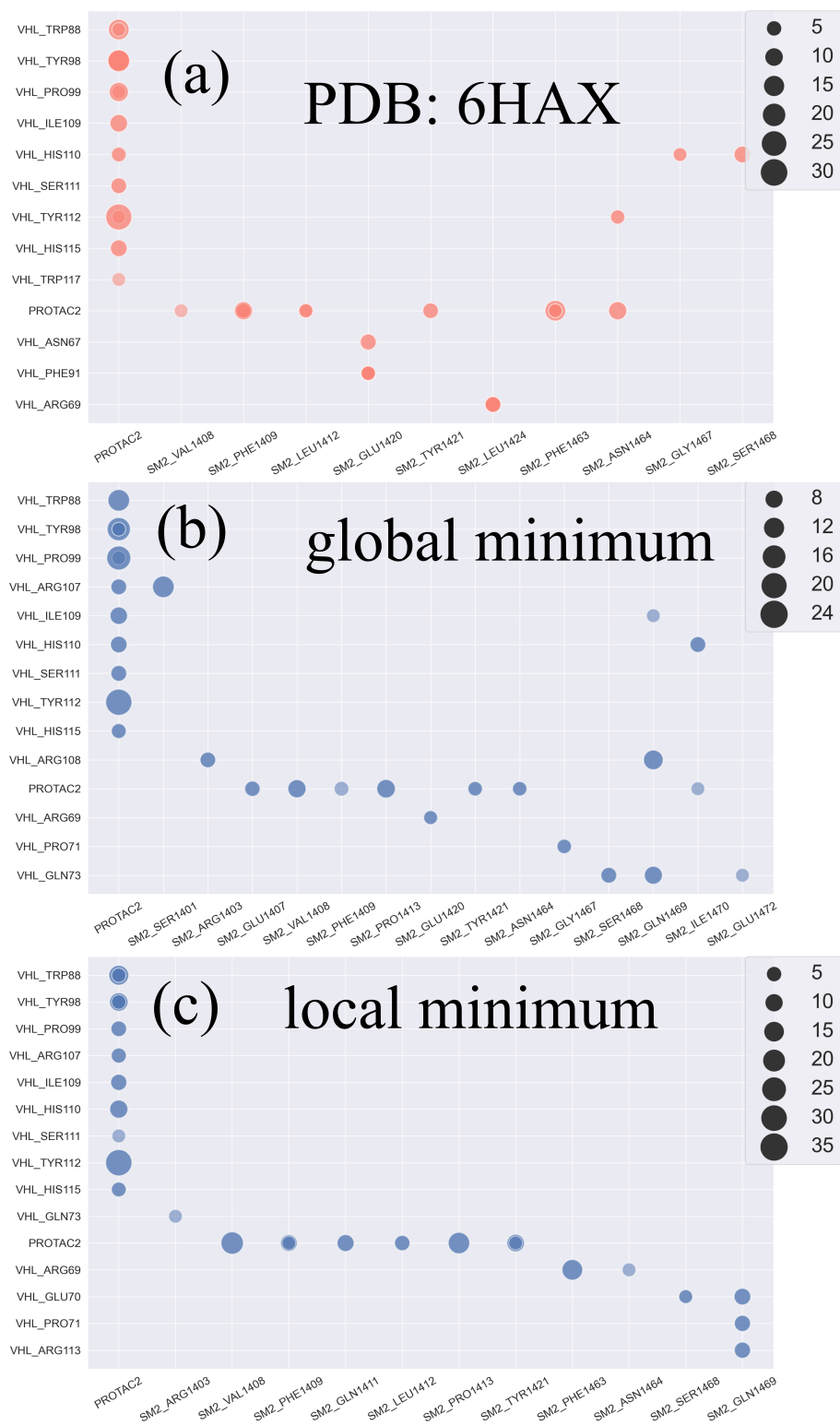


Figure 5.16: Contact maps from the (a) co-crystallized structure *6HAX*; (b) global minimum state and (c) metastable state identified by our MSM.

states are of particular interest: the global minimum state (or state I) with a stationary probability of 0.63, the metastable state III with 0.10 probability, and state IV, to which the experimental crystal structure can be assigned and which has a stationary probability of 0.05. The global minimum state differs from the crystal structure 6HAX by an I-RMSD of 3.6 Å, while the metastable state has an I-RMSD of 4.8 Å relative to the crystal structure. The global minimum state is stabilized by a large number of protein-protein contacts (Figure 5.16). Contacts between VHL and PROTAC 2 are largely unchanged between the metastable and global minimum states, likely due to the tight interaction between VHL and the PROTAC. On the other hand, the metastable state lacks contacts between PROTAC 2 and ARG29, ASN90, and ILE96 of SMARCA2. The area of the binding interface was substantially increased in both the metastable and global minimum states relative to the crystal structure: the global minimum state had a buried surface area of 2962 Å², compared to 2800 Å² for the metastable state and 2369 Å² for the crystal structure.

5.4 Discussion

Ternary complex formation is a critical step in the process of targeted protein degradation. However, studying the dynamic nature of ternary degrader complexes has posed challenges to the field due to the size of the system, degrees of flexibility, timescales for biological motions, and limited solution-phase experimental data. Here, we studied the structure and dynamics of three different degrader molecules of SMARCA2 that have similar thermodynamic binding profiles and crystal structure conformations but different degradation efficiencies. We solved the crystal structure of the ternary complex bridged by ACBI1 (PDB ID: 7S4E), which revealed a potential structural and dynamic aspect of the degradation profiles, which led to a series of solution-phase experimental studies and MD simulations in an effort to explain and predict the degradation profiles. HDX coupled with extensive MD simulations enabled the generation of atomic resolution representations of the dynamic ensembles. We highlighted the conformational landscape of SMARCA2:PROTAC 2:VHL, for which we carried

out 6 ms of accumulated simulation time using the Folding@home distributed supercomputer cluster. We propose an enhanced-seeding method that includes HREMD simulations of the ternary complex, extraction of a modest number of seed structures (100 in the work presented here), and independent simulations starting from each seed using Folding@home. We applied TICA for dimensionality reduction and built a MSM to assess the conformational free energy landscape and transitions between low-energy states. We found that the experimentally determined crystal structure is close to, but not coinciding with, the global minimum in the free energy landscape, although it is within 2 kcal/mol from the global minimum. By coarse-graining the MSM, we were able to identify five low-energy structures that contain different protein-protein interfaces between SMARCA2 and VHL. A less computationally costly solution was also explored, where we found that only 0.5 μ s of lowest rank/unbiased replica of HREMD (aggregate 12 μ s with 24 replicas) gave us a qualitatively similar conformational landscape with the same global minimum, albeit a lower resolution free energy surface. Thus, we proceeded to run HREMD on other systems of interest (PROTAC 1, ACBI1, and isoform 1 of SMARCA2; see Table 5.2 for number of replicas used and aggregate length for each system). Simulation analyses of the most probable structures for each of these ternary complexes show that ACBI1 and PROTAC 1 have different orientations of SMARCA2 relative to VHL, with PROTAC 2 sampling both orientations. We propose that this sampling procedure can be replicated for other target and E3 ligase complexes, since no target-specific information was used for the simulations in this work other than the starting x-ray structure coordinates.

We then explored the prediction of the ternary complex itself, using the binary structures of the known binding mode of the ligase ligand binding to VHL and warhead binding to SMARCA2, which is typically known in advance of designing heterobifunctional degrader molecules. We developed a protocol using REVO that is able to produce a 3-dimensional structural ensemble of high accuracy structures using the RMSD to the bound pose of the warhead as CV: we find structures less than 2 Å I-RMSD within 2-3 kcal/mol from the most

probable conformation of the ternary complex. The addition of experimental solution-phase HDX protection factor data, which identifies residues that are most likely to be precluded from solvent interactions in the context of the ternary structure, further increased the quality of the predictions. Similarly, we discovered that HDX data can improve docking results, although, overall docking has lower quality than the REVO simulations, likely due to the minimal sampling of internal protein degrees of freedom and lack of explicit solvent. We made publicly available all relevant simulation data and the source code for running REVO+HDX.

To further validate the REVO+HDX procedure, we performed a prospective prediction of the ternary complex of isoform 1 SMARCA2, with a 17 amino acid extension, the structure of which was previously unknown. Our ternary complex predictions suggest ways that the SMARCA2 extension is interacting with a beta-strand from VHL, explaining the observed HDX protection pattern. The REVO+HDX method provides an opportunity to visualize, at an atomistic level, the molecular interactions that guide ternary structure formation for complexes with previously known crystal structures. The ultimate goal is to uncover the solution-phase structural ensembles adopted by the target ligase pair, which appear to extend beyond what is observed using conventional structure determination methods. This information may provide a better representation of the factors that influence ternary complex formation, ultimately leading to downstream ubiquitination. Moreover, knowledge of critical atomic interactions provides a basis for alternative strategies to improve degrader designs, such as modifying linkers to induce specific conformational ensembles of the ternary complex that are associated with higher degradation. REVO+HDX is also able to recapitulate experimental k_{on} for ternary and binary complex formation.

CHAPTER 6

SUMMARY OUTLOOK AND IMPACT

In section 1.3 we outline the goals of this thesis. Now we go back to these overall goals and discuss the progress, what improvements can be made and the impact on the field in general.

The first goal of the thesis is to model and characterize the binding and unbinding events for systems of interest, which was successful for all systems studied in this thesis. It is worth noting however that the REVO algorithm was modified for each system. This is not surprising as the complexity of the simulation goal increases the parameters and algorithm might need to be optimized to be able to successfully simulate the desired phenomena.

In Chapters 2 and 4 we characterized these pathways with detailed Markov State Models (MSM)s constructed from high dimensional data. For the TSPO system, we are able to model multiple pathways between the bound and unbound states. We also were able to identify key residues the ligand interacted with along these unbinding pathways.

While we were able to simulate (un)binding pathways, it is unlikely that we were able to sample all likely pathways. To obtain a more complete picture of the (un)binding landscape for a given force field we can:

- Run longer simulations.
- Run more replicate simulations.
- Optimize the REVO parameters.
- Use distance metrics that better represent the system.
- Run simulations with more walkers.
- Develop or utilize more efficient algorithms.

Knowledge of the (un)binding pathway can help guide ligand design with more optimal kinetic properties. In particular, the results from Chapter 4 suggest a new pathway for PK-11195 unbinding through the membrane, where only dissociation into the solvent was previously published. The pathway determined in this thesis indicates that altering the lipophobicity of PK-11195 based compounds would affect the unbinding pathway and could significantly alter the residence time (RT). Additionally, an Roivant Discovery is using the results presented in Chapter 5 to design new proteolysis-targeting chimera (PROTAC) molecules.

The second goal was to accurately predict observables, such as rate constants, RT and ΔG_{bind} that can be compared to experiments. It is difficult to determine if the rates were accurate for the systems present in the SAMPL6 challenge as, to our knowledge, these quantities have not been published. However, we do know the ΔG_{bind} . Using the conventional definition for ΔG_{bind} we were off between 2.80 (OA-G4) and 5.10 (CB8-G3) kcal/mol from experiment. After applying the correction factors and running additional simulations on the OG-G6 system we were able to reduce the error to a maximum of 1.13 kcal/mol. Simulating the binding and unbinding is not the most computationally efficient method to predict free energy, but our simulations provide additional insight to the (un)binding mechanisms. The rates calculated for TSPO are hard to compare against experiment because the only published TSPO-PK-11195 unbinding rates are for the human protein, and we used a bacterial strain from *Rhodobacter sphaeroides* in our simulations. However, multiple starting poses showed unbinding rates within an order of magnitude over different rate calculation methods.

While we were able to predict kinetic rates, there are several sources of error. The first is that the use of the WE algorithm leads to high variation between runs as this method requires long simulations to be able to produce adequate sampling to confidently predict rates. Additionally there are errors associated with the force fields, which dictates the atomic motion in these simulations which in turn affects the path dependent rate prediction. Finally there is a potential inconsistency between the boundary conditions used in our work with

those used in experiment to decide if a ligand is (un)bound. For example in radioligand displacement assays, any ligand pose that is not sterically blocking entry of the radiolabelled competitor ligand would be considered “unbound” [191]. This does not indicate that the ligand has stopped interacting with the POI, merely that it is not blocking the radioligand from binding. Therefore, caution should be used when comparing the kinetic rates between different methods.

The final goal was to be able to use experimental data to help guide the REVO simulations to simulate a rare event, in particular binding of protein-PROTAC dimer to the target protein. When performing binding simulations, it is not always feasible to experimentally determine the bound structure as a reference beforehand, especially if there is a large set of ligand candidates. Furthermore, a particular crystal structure is not necessarily indicative of the conformations a given complex will take *in solution*. We need other forms of experimental data to help determine if the simulation has successfully discovered the bound basin. In Chapter 5, we were able to use protected residues identified by hydrogen deuterium exchange (HDX) experiments to predict which residues would be in contact in the ternary complex, and developed distance metrics to guide the simulations to maximize those contacts and compared this metric to a ligand root mean square deviation (RMSD) based distance metric. Both distance metrics were able to simulate the ternary complex formation. However, the simulations guided by HDX data had a higher probability of predicting bound structures with lower I-RMSDs after filtering using features from docking and HDX experiments.

Implementing experimental results into simulations is not just limited to distance metrics constructed in this thesis. Experimental data could also be helpful in designing a binning scheme in WE simulations. In the case of HDX data, the number of contacts between a protein and ligand could be used in ligand (un)binding simulations. Additionally, experimental data can be used to determine the CV being simulated along in potential energy biasing simulations such as metadynamics.

This thesis set out to use MD simulations to help model potential (un)binding pathways of

biologically relevant systems and test the validity of these pathways by predicting quantities such as kinetics that can be compared to experiment, which we did successfully. We first began with a test system to verify the REVO algorithm, and we were able to determine these pathways and use the kinetics to accurately calculate the binding free energy. We then moved to simulate the unbinding of PK-11195 from TSPO, an event that takes about 30 minutes experimentally and were able to model it dissociating into the membrane and predicted RTs that were on the same timescale as experiment. We finally simulated the formation of a ternary complex between a VHL-PROTAC dimer and SMARCA2 and evaluated how well different distance metrics were able to reach the bound state.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] Robert A. Copeland, David L. Pompliano, and Thomas D. Meek. Drug–target residence time and its implications for lead optimization. *Nature Reviews Drug Discovery*, 5:730–739, 2006.
- [2] Robert .A Copeland. The drug–target residence time model: a 10-year retrospective. *Nature Reviews Drug Discovery*, 15:87–95, 2015.
- [3] Visvaldas Kairys, Lina Baranauskiene, Migle Kazlauskiene, Daumantas Matulis, and Egidijus Kazlauskas. Binding affinity in drug design: experimental and computational techniques. *Expert Opinion on Drug Discovery*, 14(8):755–768, 2019.
- [4] T. Tanaji Talele, A. Santosh Khedkar, and C. Alan Rigby. Successful applications of computer aided drug discovery: Moving drugs from concept to the clinic. *Current Topics in Medicinal Chemistry*, 10:127–141, 2010.
- [5] Vijay Kumar Bhardwaj and Rituraj Purohit. Targeting the protein-protein interface pocket of aurora-a-tpx2 complex: rational drug design and validation. *Journal of Biomolecular Structure and Dynamics*, 39(11):3882–3891, 2021.
- [6] Ian M Hastings, William M Watkins, and Nicholas J White. The evolution of drug-resistant malaria: the role of drug elimination half-life. *Philosophical Transactions B*, 357:505–519, 2002.
- [7] K. Sandy Pang. A review of metabolite kinetics. *Journal of Pharmacokinetics and Biopharmaceutics*, 13:633–662, 1985.
- [8] Georges Vauquelin and Steven J. Charlton. Long-lasting target binding and rebinding as mechanisms to prolong in vivo drug action. *British Journal of Pharmacology*, 161:488–503, 2010.
- [9] L. B Sheiner, D. R. Stanski, S. Vozeh, D. Miller, and J. Ham. Simultaneous modeling of pharmacokinetics and pharmacodynamics: application to d-tubocurarine. *Clinical Pharmacology and Therapeutics*, 25:358–371, 1979.
- [10] Nicholas H.G. Holford and Lewis B. Sheiner. Kinetics of pharmacologic response. *Pharmacology & Therapeutics*, 16(2):143–166, 1982.
- [11] Angela Äbelö, Magdalena Andersson, Ann Aurell Holmberg, and Mats O. Karlsson. Application of a combined effect compartment and binding model for gastric acid inhibition of ar-ho47108: A potassium competitive acid blocker, and its active metabolite ar-ho47116 in the dog. *European Journal of Pharmaceutical Sciences*, 29(2):91–101,

2006.

- [12] Ashraf Yassen, Erik Olofsen, Albert Dahan, and Meindert Danhof. Pharmacokinetic-pharmacodynamic modeling of the antinociceptive effect of buprenorphine and fentanyl in rats: Role of receptor equilibration kinetics. *Journal of Pharmacology and Experimental Therapeutics*, 313(3):1136–1149, 2005.
- [13] H.-Y. Yun, M.-H. Yun, W. Kang, and K.-I. Kwon. Pharmacokinetics and pharmacodynamics of benidipine using a slow receptor-binding model. *Journal of Clinical Pharmacy and Therapeutics*, 30(6):541–547, 2005.
- [14] Peter J. Tongue. Drug–target kinetics in drug discovery. *ACS Chemical Neuroscience*, 9(1):29–39, 2018.
- [15] Yu Zhou, Yan Fu, Wanchao Yin, Jian Li, Wei Wang, Fang Bai, Shengtao Xu, Qi Gong, Tao Peng, Yu Hong, Dong Zhang, Dan Zhang, Qiufeng Liu, Yechun Xu, H. Eric Xu, Haiyan Zhang, Hualiang Jiang, and Hong Liu. Kinetics-driven drug design strategy for next-generation acetylcholinesterase inhibitors to clinical candidate. *Journal of Medicinal Chemistry*, 64(4):1844–1855, 2021.
- [16] Samuel D Lotz and Alex Dickson. Unbiased molecular dynamics of 11 min timescale drug unbinding reveals transition state stabilizing interactions. *Journal of the American Chemical Society*, 140(2):618–628, 2018.
- [17] Kruti B. Patel, Olga Kononova, Shuowei Cai, Valeri Barsegov, Virinder S. Parmar, Raj Kumar, and Bal Ram Singh. Botulinum neurotoxin inhibitor binding dynamics and kinetics relevant for drug design. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1865(9):129933, 2021.
- [18] Barbra Costa, Elenora Da Pozzo, Chiara Giacomelli, Elisabetta Barresi, Sabrina Taliani, Federico Da Settimo, and Claudia Martini. Tspo ligand residence time: a new parameter to predict compound neurosteroidogenic efficacy. *Scientific Reports*, 6, 2016.
- [19] Dong Guo, Thea Mulder-Kriger, Adriaan P. IJzerman, and Laura H Heitman. Functional efficacy of adenosine a_2a receptor agonists is positively correlated to their receptor residence time. *British Journal of Pharmacology*, 166:1846–1859, 2012.
- [20] Hao Lu and Peter J Tonge. Drug-target residence time: critical information for lead optimization. *Current Opinion in Chemical Biology*, 14(4):467–474, aug 2010.
- [21] Sai Kiran Sharma, Serge K. Lyashchenko, Hijin A. Park, Nagavarakishore Pillarsetty, Yorann Roux, Jiong Wu, Sophie Poty, Kathryn M. Tully, John T. Poirier, and Jason S. Lewis. A rapid bead-based radioligand binding assay for the determination of target-binding fraction and quality control of radiopharmaceuticals. *Nuclear Medicine and Biology*, 71:32–38, 2019.

- [22] Anni Allikalt and Ago Rinken. Budded baculovirus particles as a source of membrane proteins for radioligand binding assay: The case of dopamine d1 receptor. *Journal of Pharmacological and Toxicological Methods*, 86:81–86, 2017.
- [23] Michael J. Roy, Sandra Winkler, Scott J. Hughes, Claire Whitworth, Michael Galant, William Farnaby, Klaus Rumpel, and Alessio Ciulli. SPR-Measured Dissociation Kinetics of PROTAC Ternary Complexes Influence Target Degradation Rate. *ACS Chemical Biology*, 14(3):361–368, 2019.
- [24] Bojun Xiong, Guilin Jin, Ying Xu, Wenbing You, Yufei Luo, Menghan Fang, Bing Chen, Huihui Huang, Jian Yang, Xu Lin, and Changxi Yu. Identification of koumine as a translocator protein 18 kda positive allosteric modulator for the treatment of inflammatory and neuropathic pain. *Frontiers in Pharmacology*, 12:1536, 2021.
- [25] David A. Sykes, Steven J. Charlton, and Tahsin F. Kellici. *Single Step Determination of Unlabeled Compound Kinetics Using a Competition Association Binding Method Employing Time-Resolved FRET*, pages 177–194. Springer New York, New York, NY, 2018.
- [26] David A. Sykes, Leire Borrega-Roman, Clare R. Harwood, Bradley Hoare, Jack M. Lochray, Thais Gazzi, Stephen J. Briddon, Marc Nazaré, Uwe Grether, Stephen J. Hill, Steven J. Charlton, and Dmitry B. Veprintsev. *Kinetic Profiling of Ligands and Fragments Binding to GPCRs by TR-FRET*, pages 1–32. Springer International Publishing, Cham, 2021.
- [27] Kunal Khanna, Shankar Mandal, Aaron T. Blanchard, Muneesh Tewari, Alexander Johnson-Buck, and Nils G. Walter. Rapid kinetic fingerprinting of single nucleic acid molecules by a fret-based dynamic nanosensor. *Biosensors and Bioelectronics*, 190:113433, 2021.
- [28] Jacob D. Durrant and J. Andrew McCammon. Molecular dynamics simulations and drug discovery. *BMC Biology*, 9:1–9, 2011.
- [29] B.J. Alder and T.E Wainwright. Phase transition for a hard sphere system. *Journal of Chemical Physics*, 27:1208–1209, 1957.
- [30] J. B. Gibson, A. N. Goland, M. Milgram, and G. H. Vineyard. Dynamics of radiation damage. *Physical Review*, 120:1229–1253, 1960.
- [31] Alper T. Celebi, Seyed Hossein Jamali, André Bardow, Thijs J. H. Vlugt, and Othonas A. Moulton. Finite-size effects of diffusion coefficients computed from molecular dynamics: a review of what we have learned so far. *Molecular Simulation*, 47(10-11):831–845, 2021.
- [32] Richard M. Venable, Andreas Krämer, and Richard W. Pastor. Molecular dynamics

- p>simulations of membrane permeability.
- Chemical Reviews*
- , 119(9):5954–5997, 2019.
- [33] Arman Fathizadeh, Helmut Schiessel, and Mohammad Reza Ejtehadi. Molecular dynamics simulation of supercoiled dna rings. *Macromolecules*, 48(1):164–172, 2015.
 - [34] Samuel D. Lotz and Alex Dickson. Wepy: A flexible software framework for simulating rare events with weighted ensemble resampling. *ACS Omega*, 5(49):31608–31623, 2020.
 - [35] Nazanin Donyapour, Nicole M. Roussey, and Alex Dickson. REVO: Resampling of ensembles by variation optimization. *Journal of Chemical Physics*, 150(24), 2019.
 - [36] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.
 - [37] Chuan Tian, Koushik Kasavajhala, Kellon A. A. Belfon, Lauren Raguette, He Huang, Angela N. Migués, John Bickel, Yuzhang Wang, Jorge Pincay, Qin Wu, and Carlos Simmerling. ff19sb: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *Journal of Chemical Theory and Computation*, 16(1):528–552, 2020.
 - [38] Zhe Huai, Zhaoxi Shen, and Zhaoxi Sun. Binding thermodynamics and interaction patterns of inhibitor-major urinary protein-i binding from extensive free-energy calculations: Benchmarking amber force fields. *Journal of Chemical Information and Modeling*, 61(1):284–297, 2021.
 - [39] K. Vanommeslaeghe and A. D. MacKerell. Automation of the charmm general force field (cgenff) i: Bond perception and atom typing. *Journal of Chemical Information and Modeling*, 52(12):3144–3154, 2012.
 - [40] K. Vanommeslaeghe, E. Prabhu Raman, and A. D. MacKerell. Automation of the charmm general force field (cgenff) ii: Assignment of bonded parameters and partial atomic charges. *Journal of Chemical Information and Modeling*, 52(12):3155–3168, 2012.
 - [41] Chris Oostenbrink, Alessandra Villa, Alan E. Mark, and Wilfred F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6. *Journal of Computational Chemistry*, 25:1656–1676, 2004.
 - [42] Chia-en A. Chang, Yu-ming M. Huang, Leonard J. Mueller, and Wanli You. Investigation of structural dynamics of enzymes and protonation states of substrates using computational tools. *Catalysts*, 6(6), 2016.

- [43] T. Schneider and E. Stoll. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Phys. Rev. B*, 17:1302–1322, 1978.
- [44] Bin W. Zhang, David Jasnow, and Daniel M. Zuckerman. The “weighted ensemble” path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *The Journal of Chemical Physics*, 132:054107, 2010.
- [45] José Ruiz-Franco, Lorenzo Rovigatti, and Emanuela Zaccarelli. On the effect of the thermostat in non-equilibrium molecular dynamics simulations. *The European Physical Journal E*, 41(80):1302–1322, 2018.
- [46] Michael Gecht, Marc Siggel, Max Linke, Gerhard Hummer, and Jürgen Köfinger. Md-benchmark: A toolkit to optimize the performance of molecular dynamics simulations. *The Journal of Chemical Physics*, 153:144105, 2020.
- [47] Ada Sedova, John D. Eblen, Reuben Budiardja, Arnold Tharrington, and Jeremy C. Smith. High-performance molecular dynamics simulation for biological and materials sciences: Challenges of performance portability. In *2018 IEEE/ACM International Workshop on Performance, Portability and Productivity in HPC (P3HPC)*, pages 1–13, 2018.
- [48] David E. Shaw, Martin M. Deneroff, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Ierardi, István Kolossváry, John L. Klepeis, Timothy Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, and Stanley C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, 2008.
- [49] David E. Shaw, Ron O. Dror, John K. Salmon, J. P. Grossman, Kenneth M. Mackenzie, Joseph A. Bank, Cliff Young, Martin M. Deneroff, Brannon Batson, Kevin J. Bowers, Edmond Chow, Michael P. Eastwood, Douglas J. Ierardi, John L. Klepeis, Jeffrey S. Kuskin, Richard H. Larson, Kresten Lindorff-Larsen, Paul Maragakis, Mark A. Moraes, Stefano Piana, Yibing Shan, and Brian Towles. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, SC ’09, New York, NY, USA, 2009. Association for Computing Machinery.
- [50] David E. Shaw, J.P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li-Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk-Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L.

- Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the bar for performance and programmability in a special purpose molecular dynamics supercomputer. In *SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 41–53, 2014.
- [51] Vincent A. Voelz, Gregory R. Bowman, Kyle Beauchamp, and Vijay S. Pande. Molecular simulation of ab initio protein folding for a millisecond folder ntl9. *Journal of the American Chemical Society*, 132(5):1526–1528, 2010. PMID: 20070076.
- [52] David J. Earl and Michael W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7:3910–3916, 2005.
- [53] Lukas S. Stelzl and Gerhard Hummer. Kinetics from replica exchange molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 13(8):3927–3935, 2017.
- [54] Mark J. Abraham and Jill E. Gready. Ensuring mixing efficiency of replica-exchange molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 4(7):1119–1128, 2008.
- [55] Angel E. Garcia, Henry Herce, and Dietmar Paschek. Chapter 5 simulations of temperature and pressure unfolding of peptides and proteins with replica exchange molecular dynamics. In David C. Spellmeyer, editor, *Annual Reports in Computational Chemistry*, volume 2 of *Annual Reports in Computational Chemistry*, pages 83–95. Elsevier, 2006.
- [56] Rodrigo Casasnovas, Vittorio Limongelli, Pratyush Tiwary, Paolo Carloni, and Michele Parrinello. Unbinding kinetics of a p38 map kinase type ii inhibitor from metadynamics simulations. *Journal of the American Chemical Society*, 139(13):4780–4788, 2017.
- [57] Riccardo Capelli, Anna Bochicchio, Giovanni Maria Piccini, Rodrigo Casasnovas, Paolo Carloni, and Michele Parrinello. Chasing the full free energy landscape of neuroreceptor/ligand unbinding by metadynamics simulations. *Journal of Chemical Theory and Computation*, 15(5):3354–3361, 2019.
- [58] Riccardo Capelli, Wenping Lyu, Viacheslav Bolnykh, Simone Meloni, Jógvan Magnus Haugaard Olsen, Ursula Rothlisberger, Michele Parrinello, and Paolo Carloni. Accuracy of molecular simulation-based predictions of koff values: A metadynamics study. *The Journal of Physical Chemistry Letters*, 11(15):6373–6381, 2020.
- [59] Rilei Yu, Nargis Tabassum, and Tao Jiang. Investigation of α -conotoxin unbinding using umbrella sampling. *Bioorganic & Medicinal Chemistry Letters*, 26(4):1296–1300,

2016.

- [60] Cameron F. Abrams and Eric Vanden-Eijnden. Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. *Proceedings of the National Academy of Sciences*, 107(11):4961–4966, 2010.
- [61] Gabriel Stoltz and Eric Vanden-Eijnden. Longtime convergence of the temperature-accelerated molecular dynamics method. *Nonlinearity*, 31(8):3748–3769, 2018.
- [62] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.
- [63] Marc Souaille and Benoit Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer Physics Communications*, 135(1):40–57, 2001.
- [64] Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nature Review Physics*, 2:200–212, 2020.
- [65] Pratyush Tiwary and Michele Parrinello. From metadynamics to dynamics. *Phys. Rev. Lett.*, 111:230602, 2013.
- [66] Anton K. Faradjian and Ron Elber. Computing time scales from reaction coordinates by milestoning. *The Journal of Chemical Physics*, 120(23):10880–10889, 2004.
- [67] Ron Elber. Long-timescale simulation methods. *Current Opinion in Structural Biology*, 15(2):151–156, 2005. Theory and simulation/Macromolecular assemblages.
- [68] Surl-Hee Ahn, Benjamin R. Jagger, and Rommie E. Amaro. Ranking of ligand binding kinetics using a weighted ensemble approach and comparison with a multiscale milestoning approach. *Journal of Chemical Information and Modeling*, 60(11):5340–5352, 2020.
- [69] Camilo Velez-Vega, Ernesto E. Borrero, and Fernando A. Escobedo. Kinetics and reaction coordinate for the isomerization of alanine dipeptide by a forward flux sampling protocol. *The Journal of Chemical Physics*, 130(22):225101, 2009.
- [70] David Richard and Thomas Speck. Crystallization of hard spheres revisited. i. extracting kinetics and free energy landscape from forward flux sampling. *The Journal of Chemical Physics*, 148(12):124110, 2018.
- [71] Gary A. Huber and Sangtae Kim. Weighted-ensemble brownian dynamics simulations for protein association reactions. *Biophysical Journal*, 70:97–110, 1996.

- [72] Daniel M. Zuckerman and Lillian T. Chong. Weighted ensemble simulation: Review of methodology, applications, and software. *ANN REV BIOPHYS*, 46:43–57, 2017.
- [73] Daniel M. Zuckerman and Lillian T. Chong. Weighted ensemble simulation: Review of methodology, applications, and software. *Annual Review of Biophysics*, 46(1):43–57, 2017.
- [74] Tom Dixon, Samuel D. Lotz, and Alex Dickson. Predicting ligand binding affinity using on- and off-rates for the SAMPL6 SAMPLing challenge. *Journal of Computer-Aided Molecular Design*, 32(10):1001–1012, 2018.
- [75] Badi’ Abdul-Wahid, Haoyun Feng, Dinesh Rajan, Ronan Costaouec, Eric Darve, Douglas Thain, and Jesús A. Izaguirre. AWE-WQ: Fast-Forwarding Molecular Dynamics Using the Accelerated Weighted Ensemble. *Journal of Chemical Information and Modeling*, 54(10):3033–3043, 2014.
- [76] Matthew C. Zwier, Joshua L. Adelman, Joseph W. Kaus, Adam J. Pratt, Kim F. Wong, Nicholas B. Rego, Ernesto Suarez, Steveb Lettieri, David W. Wang, Michael Grabe, Daniel M. Zuckerman, and Lillian T. Chong. Westpa: An interoperable, highly scalable software package for weighted ensemble simulation and analysis. *Journal of Chemical Theory and Computation*, 11(2):800–809, 2015.
- [77] Alex Dickson and Samuel D. Lotz. Ligand Release Pathways Obtained with WExplore: Residence Times and Mechanisms. *Journal of Physical Chemistry B*, 120(24):5377–5385, 2016.
- [78] Matthew C. Zwier, Adam J. Pratt, Joshua L. Adelman, Joseph W. Kaus, Daniel M. Zuckerman, and Lillian T. Chong. Efficient atomistic simulation of pathways and calculation of rate constants for a protein–peptide binding process: Application to the mdm2 protein and an intrinsically disordered p53 peptide. *The Journal of Physical Chemistry Letters*, 7(17):3440–3445, 2016.
- [79] Bin W. Zhang, David Jasnow, and Daniel M. Zuckerman. Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin. *Proceedings of the National Academy of Sciences*, 104(46):18043–18048, 2007.
- [80] Hiroshi Fujisaki, Kei Moritsugu, Ayori Mitsutake, and Hiromichi Suetani. Conformational change of a biomolecule studied by the weighted ensemble method: Use of the diffusion map method to extract reaction coordinates. *The Journal of Chemical Physics*, 149(13):134112, 2018.
- [81] Hiroshi Fujisaki, Yasuhiro Matsunaga, and Kei Moritsugu. Weighted ensemble simulations for conformational changes of proteins. *AIP Conference Proceedings*, 2343(1):020016, 2021.

- [82] Scott H. Northrup, Stuart A. Allison, and J. Andrew McCammon. Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *Journal of Chemical Physics*, 80:1517, 1984.
- [83] Ali S. Saglam and Lillian T. Chong. Highly efficient computation of the basal k_{on} using direct simulation of protein–protein association with flexible molecular models. *The Journal of Physical Chemistry B*, 120(1):117–122, 2016.
- [84] Alex Dickson and Charles L. Brooks. Wexplore: Hierarchical exploration of high-dimensional spaces using the weighted ensemble algorithm. *The Journal of Physical Chemistry B*, 118(13):3532–3542, 2014.
- [85] Alex Dickson, Aryeh Warmflash, and Aaron R. Dinner. Separating forward and backward pathways in nonequilibrium umbrella sampling. *Journal of Chemical Physics*, 131:154104, 2009.
- [86] Alex Dickson, Mark Maienschein-Cline, Allison Tovo-Dwyer, Jeff R. Hammond, and Aaron R. Dinner. Flow-dependent unfolding and refolding of an rna by nonequilibrium umbrella sampling. *Journal of Chemical Theory and Computation*, 7:2710–2720, 2011.
- [87] Eric Vanden-Eijnden and Maddalena Venturoli. Exact rate calculations by trajectory parallelization and tilting. *Journal of Chemical Physics*, 131:044120, 2009.
- [88] Ernesto Suárez, Steven Lettieri, Matthew C. Zwier, Carsen A. Stringer, Sundar Raman Subramanian, Lillian T. Chong, and Daniel M. Zuckerman. Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. *Journal of Chemical Theory and Computation*, 10(7):2658–2667, 2014.
- [89] Ronan Costaeuec, Haoyun Feng, Jesús Izaguirre, and Eric Darve. Analysis of the accelerated weighted ensemble methodology. *Discrete and Continuous Dynamical Systems*, pages 171–181, 2013.
- [90] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics*, 134(17):174105, 2011.
- [91] C. R. Schwantes, R. T. McGibbon, and V. S. Pande. Perspective: Markov models for long-timescale biomolecular dynamics. *The Journal of Chemical Physics*, 141(9):090901, 2014.
- [92] Robert T. McGibbon, Christian R. Schwantes, and Vijay S. Pande. Statistical model selection for markov models of biomolecular dynamics. *The Journal of Physical Chemistry B*, 118(24):6475–6481, 2014.

- [93] Yuguang Mu, Phuong H. Nguyen, and Gerhard Stock. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins: Structure, Function, and Bioinformatics*, 58(1):45–52, 2005.
- [94] Ushnish Sengupta, Martín Carballo-Pacheco, and Birgit Strodel. Automated markov state models for molecular dynamics simulations of aggregation and self-assembly. *The Journal of Chemical Physics*, 150(11):115101, 2019.
- [95] Adam Kells, Alessia Annibale, and Edina Rosta. Limiting relaxation times from markov state models. *The Journal of Chemical Physics*, 149:072324, 2018.
- [96] Anita de Ruiter and Chris Oostenbrink. Free energy calculations of protein–ligand interactions. *Current Opinion in Chemical Biology*, 15:547–552, 2011.
- [97] Vytautas Gapsys, Servaas Michielssens, Jan Henning Peters, Bert L. de Groot, and Hadas Leonov. Calculation of binding free energies. In *Molecular Modeling of Proteins (Methods and Protocols)*, pages 173–209. Humana Press, New York, NY, 2014.
- [98] Matthew T. Geballe, A. Geoffrey Skillman, Anthony Nicholls, J. Peter Guthrie, and Peter J. Taylor. The sampl2 blind prediction challenge: introduction and overview. *Journal of Computer-Aided Molecular Design*, 24:259–279, 2010.
- [99] Andrea Rizzi, Steven Murkli, John McNeill, Wei Yao, Mathew Sullivan, Michael K. Gilson, Michael W. Chiu, Lyle Isaacs, Bruce C. Gibb, David L. Mobley, and John D. Chodera. Overview of the sampl6 host–guest binding affinity prediction challenge. *Journal of Computer-Aided Molecular Design*, 32:937–963, 2018.
- [100] Albert C. Pan, David W. Borhani, Ron O. Dror, and David E. Shaw. Molecular determinants of drug–receptor binding kinetics. *Drug Discovery Today*, 18:667–673, 2013.
- [101] Daria B. Kokh, Marta Amaral, Joerg Bomke, Ulrich Grädler, Djordje Musil, Hans-Peter Buchstaller, Matthias K. Dreyer, Matthias Frech, Maryse Lowinski, Francois Vallee, Marc Bianciotto, Alexey Rak, and Rebecca C. Wade. Estimation of drug-target residence times by τ -random acceleration molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 14:3859–3869, 2018.
- [102] Alex Dickson, Pratyush Tiwary, and Harish Vashisth. Kinetics of ligand binding through advanced computational approaches: A review. *Current Topics in Medicinal Chemistry*, 17:2626–2641, 2017.
- [103] Ivan Teo, Christopher G. Mayne, Klaus Schulten, and Tony Lelièvre. Adaptive multilevel splitting method for molecular dynamics calculation of benzamidine-trypsin dissociation time. *Journal of Chemical Theory and Computation*, 12:2983–2989, 2016.

- [104] Lane W. Votapka, Benjamin R. Jagger, Alexandra L. Heyneman, and Rommie E. Amaro. Seekr: Simulation enabled estimation of kinetic rates, a computational tool to estimate molecular kinetics and its application to trypsin–benzamidine binding. *The Journal of Physical Chemistry B*, 121(15):3597–3606, 2017.
- [105] S. Doerr and G. De Fabritiis. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *Journal of Chemical Theory and Computation*, 10(5):2064–2069, 2014.
- [106] Ignasi Buch, Toni Giorgino, and Gianni De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25):10184–10189, 2011.
- [107] Nuria Plattner and Frank Noé. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and markov models. *Nature Communications*, 6:7653, 2015.
- [108] Vittorio Limongelli, Massimiliano Bonomi, and Michele Parrinello. Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National Academy of Sciences of the United States of America*, 16:6358–6363, 2013.
- [109] Alex Dickson and Samuel D. Lotz. Multiple ligand unbinding pathways and ligand-induced destabilization revealed by wexplore. *Biophysical Journal*, 112:620–629, 2017.
- [110] Pratyush Tiwary, Jagannath Mondal, and B. J. Berne. How and when does an anti-cancer drug leave its binding site? *Science Advances*, 3(5):e1700014, 2017.
- [111] Hari S. Muddana, C. Daniel Varnado, Christopher W. Bielawski, Adam R. Urbach, Lyle Issacs, Matthew T. Geballe, and Michael K. Gilson. Blind prediction of host–guest binding affinities: a new sampling challenge. *Journal of Computer-Aided Molecular Design*, 26:475–487, 2012.
- [112] Frank Biedermann and Oren A. Scherman. Cucurbit[8]uril mediated donor–acceptor ternary complexes: A model system for studying charge-transfer interactions. *The Journal of Physical Chemistry B*, 116(9):2842–2849, 2012.
- [113] Haiying Gan, Christopher J. Benjamin, and Bruce C. Gibb. Nonmonotonic assembly of a deep-cavity cavitand. *Journal of the American Chemical Society*, 133(13):4770–4773, 2011.
- [114] Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrin, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS*

- Computational Biology*, 13(7):1–17, 2017.
- [115] Alex Dickson, Mark Maienschein-Cline, Allison Tovo-Dwyer, Jeff R. Hammond, and Aaron R. Dinner. Flow-dependent unfolding and refolding of an rna by nonequilibrium umbrella sampling. *Journal of Chemical Theory and Computation*, 7(9):2710–2720, 2011.
 - [116] Ronan Costaouec, Haoyun Feng, Jesús Izaguirre, and Eric Darve. Analysis of the accelerated weighted ensemble methodology. *Discrete & Continuous Dynamical Systems*, 2013:171–181, 2013.
 - [117] T. Hill. *Free energy transduction and biochemical cycle kinetics*. Academic Press, 1989.
 - [118] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science conference (SciPy 2008)*, pages 11–15, United States, 2009.
 - [119] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 3:361–362, 2009.
 - [120] Matthew P. Harrigan, Mohammad M. Sultan, Carlos X. Hernández, Brooke E. Husic, Peter Eastman, Christian R. Schwantes, Kyle A. Beauchamp, Robert T. McGibbon, and Vijay S. Pande. Msmbuilder: Statistical models for biomolecular dynamics. *Biophysical Journal*, 112:10–15, 2017.
 - [121] Alex Dickson. Csnalysis. <https://github.com/ADicksonLab/CSNAnalysis>, 2018.
 - [122] Ken Cherven. *Network Graph Analysis and Visualization with Gephi*. Packt Publishing, 2013.
 - [123] Steven Murkli, John N. McNeill, and Lyle Isaacs. Cucurbit[8]uril•guest complexes: blinded dataset for the sampl6 challenge. *Supramolecular Chemistry*, 31:150–158, 2019.
 - [124] Atipat Rojnuckarin, Dennis R. Livesay, and Shankar Subramaniam. Bimolecular reaction simulation using weighted ensemble brownian dynamics and the university of houston brownian dynamics program. *Biophysical Journal*, 79:686–693, 2000.
 - [125] Mary J Carroll, Randall V Mauldin, Anna V Gromova, Scott F Singleton, J Edward, and Andrew L Lee. Evidence for dynamics in proteins as a mechanism for ligand dissociation. *Nature Chemical Biology*, 8(3):246–252, 2012.
 - [126] Georges Vauquelin, Sophie Bostoen, Patrick Vanderheyden, and Philip Seeman. *Clozapine, atypical antipsychotics, and the benefits of fast-off D2 dopamine receptor antagonism*, volume 385. Springer, 2012.

- [127] Jianfeng Pei, Ning Yin, Xiaomin Ma, and Luhua Lai. Systems biology brings new dimensions for structure-based drug design. *Journal of the American Chemical Society*, 136(33):11556–11565, 2014.
- [128] Pelin Ayaz, Dorothee Andres, Dennis A. Kwiatkowski, Carl Christian Kolbe, Philip Lienau, Gerhard Siemeister, Ulrich Lücking, and Christian M. Stegmann. Conformational Adaption May Explain the Slow Dissociation Kinetics of Roniciclib (BAY 1000394), a Type i CDK Inhibitor with Kinetic Selectivity for CDK2 and CDK9. *ACS Chemical Biology*, 11(6):1710–1719, 2016.
- [129] Barbara Costa, Eleonora Da Pozzo, Chiara Giacomelli, Elisabetta Barresi, Sabrina Taliani, Federico Da Settimo, and Claudia Martini. TSPO ligand residence time: a new parameter to predict compound neurosteroidogenic efficacy. *Scientific Reports*, 6(August 2015):18164, 2016.
- [130] Dong Guo, Laura H. Heitman, and Adriaan P. Ijzerman. The Added Value of Assessing Ligand-Receptor Binding Kinetics in Drug Discovery. *ACS Medicinal Chemistry Letters*, 7(9):819–821, 2016.
- [131] Peter J. Tonge. Drug-Target Kinetics in Drug Discovery. *ACS Chemical Neuroscience*, page acschemneuro.7b00185, 2017.
- [132] Kin Sing Stephen Lee, Jun Yang, Jun Niu, Connie J. Ng, Karen M. Wagner, Hua Dong, Sean D. Kodani, Debin Wan, Christophe Morisseau, and Bruce D. Hammock. Drug-Target Residence Time Affects in Vivo Target Occupancy through Multiple Pathways. *ACS Central Science*, 5(9):1614–1624, 2019.
- [133] M. Bernetti, A. Cavalli, and L. Mollica. Protein-ligand (un)binding kinetics as a new paradigm for drug discovery at the crossroad between experiments and modelling. *MedChemComm*, 2017.
- [134] Dong Guo, Lizi Xia, Jacobus P. D. van Veldhoven, Marc Hazeu, Tamara Mocking, Johannes Brussee, Adriaan P. IJzerman, and Laura H. Heitman. Binding Kinetics of ZM241385 Derivatives at the Human Adenosine A_{2A} Receptor. *ChemMedChem*, 9(4):752–761, 2014.
- [135] Lauren A. Spagnuolo, Sandra Eltschkner, Weixuan Yu, Fereidoon Daryaei, Shabnam Davoodi, Susan E. Knudson, Eleanor K H Allen, Jonathan Merino, Annica Pschibul, Ben Moree, Neil Thivalapill, James J. Truglio, Joshua Salafsky, Richard A. Slayden, Caroline Kisker, and Peter J. Tonge. Evaluating the Contribution of Transition-State Destabilization to Changes in the Residence Time of Triazole-Based InhA Inhibitors. *Journal of the American Chemical Society*, 139(9):3417–3429, 2017.
- [136] R O Dror, A C Pan, D H Arlow, D W Borhani, P Maragakis, Y Shan, H Xu, and D E Shaw. Pathway and mechanism of drug binding to G-protein-coupled receptors.

- Proceedings of the National Academy of Sciences of the United States of America*, 108(32):13118–13123, 2011.
- [137] Albert C. Pan, Huafeng Xu, Timothy Palpant, and David E. Shaw. Quantitative characterization of the binding and unbinding of millimolar drug fragments with molecular dynamics simulations. *Journal of Chemical Theory and Computation*, page acs.jctc.7b00172, 2017.
 - [138] Alex Dickson. Mapping the Ligand Binding Landscape. *Biophysical Journal*, 115(9):1707–1719, 2018.
 - [139] Agostino Bruno, Elisabetta Barresi, Nicola Simola, Eleonora Da Pozzo, Barbara Costa, Ettore Novellino, Federico Da Settimo, Claudia Martini, Sabrina Taliani, and Sandro Cosconati. Unbinding of Translocator Protein 18 kDa (TSPO) Ligands: From in Vitro Residence Time to in Vivo Efficacy via in Silico Simulations. *ACS Chemical Neuroscience*, 10(8):3805–3814, 2019.
 - [140] Indrajit Deb and Aaron T. Frank. Accelerating Rare Dissociative Processes in Biomolecules Using Selectively Scaled MD Simulations. *Journal of Chemical Theory and Computation*, 15(11):5817–5828, 2019.
 - [141] Steven E. Kirberger, Peter D. Ycas, Jorden A. Johnson, Chen Chen, Michael F. Ciccone, Rinette W.L. Woo, Andrew K. Urick, Huda Zahid, Ke Shi, Hideki Aihara, Sean D. McAllister, Mohammed Kashani-Sabet, Junwei Shi, Alex Dickson, Camila O. Dos Santos, and William C.K. Pomerantz. Selectivity, ligand deconstruction, and cellular activity analysis of a BPTF bromodomain inhibitor. *Organic and Biomolecular Chemistry*, 17(7):2020–2027, 2019.
 - [142] Jian Yin, Niel M. Henriksen, David R. Slochower, Michael R. Shirts, Michael W. Chiu, David L. Mobley, and Michael K. Gilson. Overview of the sampl5 host–guest challenge: Are we doing better? *Journal of Computer-Aided Molecular Design*, 1:1–19, 2017.
 - [143] Carlo Camilloni and Fabio Pietrucci. Advanced simulation techniques for the thermodynamic and kinetic characterization of biological systems. *Advances in Physics: X*, 2018.
 - [144] Marc F. Lensink, Sameer Velankar, and Shoshana J. Wodak. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins: Structure, Function and Bioinformatics*, 85(3):359–377, 2017.
 - [145] Tristan I. Croll, Massimo D. Sammito, Andriy Kryshchuk, and Randy J. Read. Evaluation of template-based modeling in CASP13. *Proteins: Structure, Function and Bioinformatics*, 87(12):1113–1127, 2019.
 - [146] Conor D Parks, Zied Gaieb, Michael Chiu, Huanwang Yang, Chenghua Shao,

- W. Patrick Walters, Johanna M Jansen, G McGaughey, Richard A Lewis, Scott D Bembenek, Michael K Ameriks, Tara Mirzadegan, Stephen K. Burley, Rommie E. Amaro, and Michael K. Gilson. D3R grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of Computer-Aided Molecular Design*, 34:99–119, 2020.
- [147] Synapse. IDG-DREAM Drug-Kinase Binding Prediction Challenge.
- [148] Andrea Rizzi, Travis Jensen, David R Slochower, Matteo Aldeghi, Vytautas Gapys, Dimitris Ntekoumes, Stefano Bosisio, Michail Papadourakis, Niel M Henriksen, L De Groot, Zoe Cournia, Alex Dickson, Julien Michel, Michael K Gilson, R Michael, David L Mobley, and John D Chodera. The SAMPL6 SAMPLing challenge : Assessing the reliability and efficiency of binding free energy calculations. *Journal of Computer-Aided Molecular Design*, pages 1–33, 2020.
- [149] Michael K. Gilson, James A. Given, Bruce L. Bush, and J. Andrew McCammon. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophysical Journal*, 72(3):1047–1069, 1997.
- [150] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *Journal of Chemical Physics*, 129(12):124105, 2008.
- [151] Donald Archer and Peiming Wang. The Dielectric Constant of Water and Debye-Hückel Limiting Law Slopes. *Journal of Physical and Chemical Reference Data*, 19(2):371–411, 1990.
- [152] Hyung June Woo and Benoît Roux. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19):6825–6830, 2005.
- [153] Matthew R. Sullivan, Wei Yao, and Bruce C. Gibb. The thermodynamics of guest complexation to octa-acid and tetra-endo-methyl octa-acid: reference data for the sixth statistical assessment of modeling of proteins and ligands (SAMPL6). *Supramolecular Chemistry*, 31(3):184–189, 2019.
- [154] J M Torrie and J P Valleau. Non-physical sampling distributions in Monte-Carlo free-energy estimation- umbrella sampling. *Journal of Computational Physics*, 23:187–199, 1977.
- [155] Naohiro Nishikawa, Kyungreem Han, Xiongwu Wu, Florentina Tofoleanu, and Bernard R. Brooks. Comparison of the umbrella sampling and the double decoupling method in binding free energy predictions for SAMPL6 octa-acid host–guest challenges. *Journal of Computer-Aided Molecular Design*, 32(10):1075–1086, 2018.
- [156] Lin Frank Song, Nupur Bansal, Zheng Zheng, and Kenneth M. Merz. Detailed potential

- of mean force studies on host–guest systems from the SAMPL6 challenge. *Journal of Computer-Aided Molecular Design*, 32(10):1013–1026, 2018.
- [157] John G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3:300–313, 1935.
- [158] Agastya P. Bhati, Shunzhou Wan, and Peter V. Coveney. Ensemble-based replica exchange alchemical free energy methods: The effect of protein mutations on inhibitor binding. *Journal of Chemical Theory and Computation*, 15:1265–1277, 2019.
- [159] Margarita Gutiérrez, Gabriel A. Vallejos, Magdalena P. Cortés, and Carlos Bustos. Bennett acceptance ratio method to calculate the binding free energy of bace1 inhibitors: Theoretical model and design of new ligands of the enzyme. *Chemical Biology & Drug Design*, 93:1117–1128, 2019.
- [160] Eko Aditya Rifai, Marc van Dijk, Nico P. E. Vermeulen, Arry Yanuar, and Daan P. Geerke. A comparative linear interaction energy and mm/pbsa study on sirt1–ligand binding free energy calculation. *Journal of Chemical Information and Modeling*, 59:4018–4033, 2019.
- [161] Robert W. Zwanzig. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *Journal of Chemical Physics*, 22:1420–1426, 1954.
- [162] William L. Jorgensen and Laura L. Thomas. Perspective on free-energy perturbation calculations for chemical equilibria. *Journal of Chemical Theory and Computation*, 2008.
- [163] Nadine Homeyer, Friederike Stoll, Alexander Hillisch, and Holger Gohlke. Binding free energy calculations for lead optimization: Assessment of their accuracy in an industrial drug design context. *Journal of Chemical Theory and Computation*, 10(8):3331–3344, 2014.
- [164] Frederick Bonsack and Sangeetha Sukumari-Ramesh. Tspo: An evolutionarily conserved protein with elusive functions. *International Journal of Molecular Sciences*, 19(1694), 2018.
- [165] F. Li, J. Liu, Y. Zheng, R. M. Garavito, and S. Ferguson-Miller. Crystal structures of translocator protein (tspo) and mutant mimic of a human polymorphism. *Science*, 347:555–558, 2015.
- [166] Y. Guo, R. C. Kalathur, Q. Liu, R. Bruni, C. Ginter, E. Kloppmann, B. Rost, and W. A. Hendrickson. Structure and activity of tryptophan-rich tspo proteins. *Science*, 347:551–555, 2015.
- [167] M. Jaremko, Ł. Jaremko, K. Giller, S. Becker, and M. Zweckstetter. Structure of

- the mitochondrial translocator protein in complex with a diagnostic ligand. *Science*, 343:1363–1366, 2014.
- [168] M. Jaremko, Ł. Jaremko, K. Giller, S. Becker, and M. Zweckstetter. Structural integrity of the a147t polymorph of mammalian tspo. *ChemBioChem*, 16(10):1483–1489, 2015.
 - [169] H. Li and V. Papadopoulos. Peripheral-type benzodiazepine receptor function in cholesterol transport. identification of a putative cholesterol recognition/interaction amino acid sequence and consensus pattern. *Endocrinology*, 139(12):4991–4997, 1998.
 - [170] Alana M. Scarf, Lars M. Ittner, and Michael Kassiou. The translocator protein (18 kda): Central nervous system disease and drug design. *Journal of Medical Chemistry*, 52:581–592, 2009.
 - [171] L. Veenman, V. Papadopoulos, and M. Gavish. Channel-like functions of the 18-kda translocator protein (tspo): regulation of apoptosis and steroidogenesis as part of the host-defense response. *Current Pharmaceutical Design*, 13(23):2385–2405, 2007.
 - [172] H. Batoko, V. Veljanovski, and P. Jurkiewicz. Enigmatic translocator protein (tspo) and cellular stress regulation. *Trends Biochem Sci.*, 40:497–503, 2015.
 - [173] Jemma Gatliff, Daniel A. East, Aarti Singh, Maria Soledad Alvarez, Michele Frison, Ivana Matic, Caterina Ferraina, Natalie Sampson, Federico Turkheimer, and Michelangelo Campanella. A role for tspo in mitochondrial ca^{2+} homeostasis and redox stress signaling. *Cell Death & Disease*, 8:e2896, 2017.
 - [174] Lan N. Tu, K. Morohaku, P. R. Manna, S. H. Pelton, W. R. Butler, D. M. Stocco, and V. Selvaraj. Peripheral benzodiazepine receptor/translocator protein global knock-out mice are viable with no effects on steroid hormone biosynthesis. *Journal of Biological Chemistry*, 289:27444–27454, 2014.
 - [175] Lan N. Tu, Amy H. Zhao, Douglas M. Stocco, and Vimal Selvaraj. Pk11195 effect on steroidogenesis is not mediated through the translocator protein (tspo). *Endocrinology*, 156:1033–1039, 2015.
 - [176] Rainer Rupprecht, Vassilios Papadopoulos, Gerhard Rammes, Thomas C. Baghai, Jinjiang Fan, Nagaraju Akula, Ghislaine Groyer, David Adams, and Michael Schumacher. Translocator protein (18 kda) (tspo) as a therapeutic target for neurological and psychiatric disorders. *Nature Reviews Drug Discovery*, 9:971–988, 2010.
 - [177] Mara Perrone, Byung Seook Moon, Hyun Soo Park, Valentino Laquintana, Jae Ho Jung, Annalisa Cutrignelli, Angela Lopedota, Massimo Franco, Sang Eun Kim, Byung Chul Lee, and Nunzio Denora. A novel pet imaging probe for the detection and monitoring of translocator protein 18 kda expression in pathological disorders. *Scientific Reports*, 6(20422), 2016.

- [178] Barbra Costa, Chiara Giacomelli, Eleonora Da Pozzo, Sabrinia Taliani, Federico Da Settimo, and Claudia Martini. The anxiolytic etifoxine binds to tspo ro5-4864 binding site with long residence time showing a high neurosteroidogenic activity. *ACS Chemical Neuroscience*, 8:1448–1454, 2017.
- [179] Donald Hamelberg, John Mongan, and J McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics*, 120:11919–11929, 2004.
- [180] B. Isralewitz, M. Gao, and K. Schulten. Steered molecular dynamics and mechanical functions of proteins. *Current Opinion in Structural Biology*, 11(2):224 – 230, 2001.
- [181] Fei Li, Jian Liu, Nan Liu, Leslie A. Kuhn, R. Michael Garavito, and Shelagh Ferguson-Miller. Translocator protein 18 kda (tspo): An old protein with new functions? *Biochemistry*, 55:2821–2831, 2016.
- [182] Christophe Chipot, François Dehez, Jason R. Schnell, Nicole Zitzmann, Eva Pebay-Peyroula, Laurent J. Catoire, Bruno Miroux, Edmund R. S. Kunji, Gianlugi Veglia, Timothy A. Cross, and Paul Schanda. Perturbations of native membrane protein structure in alkyl phosphocholine detergents: A critical assessment of nmr and biophysical studies. *Chemical Reviews*, 118:3559–3607, 2018.
- [183] Juan Zeng, Riccardo Guareschi, Mangesh Damre, Ruyin Cao, Achim Kless, Bernard Neumaier, Andreas Bauer, Alejandro Giorgetti, Paolo Carloni, and Giulia Rossetti. Structural prediction of the dimeric form of the mammalian translocator membrane protein tspo: A key target for brain diagnostics. *International Journal of Molecular Sciences*, 19(2588), 2018.
- [184] Emilia L. Wu, Xi Cheng, Sunhwan Jo, Huan Rui, Kevin C. Song, Eder M. Dávila-Contreras, Yifei Qi, Jumin Lee, Viviana Monje-Galvan, Richard M. Venable, Jeffery B. Klauda, and Wonpil Im. Charmm-gui membrane builder toward realistic biological membrane simulations. *Journal of Computational Chemistry*, 35:1997–2004, 2014.
- [185] Jing Huang and Alexander D MacKerell Jr. Charmm36 all-atom additive protein force field: validation based on comparison to nmr data. *Journal of Computational Chemistry*, 34:2135–2145, 2013.
- [186] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medical Chemistry*, 47(7):1739–1749, 2004.
- [187] Yan Xia, Kaitlyn Ledwitch, Georg Kuenze, Amanda Duran, Jun Li, Charles R. Sanders, Charles Manning, and Jems Meiler. A unified structural model of the mam-

- malian translocator protein (tspo). *Journal of Biomolecular NMR*, 73:347–364, 2019.
- [188] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.
- [189] Andrew Midzak, Nagaraju Akula, Laurent Lecanu, and Vassilos Papadopoulos. Novel androstenediol interacts with the mitochondrial translocator protein and controls steroidogenesis. *Journal of Biological Chemistry*, 286:9875–9887, 2011.
- [190] Garima Jaipuria, Andrei Leonov, Karin Giller, Suresh Kumar Vasa, Łukasz Jaremko, Mariusz Jaremko, Rasmus Linser, Stefan Becker, and Markus Zweckstetter. Cholesterol-mediated allosteric regulation of the mitochondrial translocator protein structure. *Nature Communications*, 8:14893, 2017.
- [191] H J Motulsky and L C Mahan. The kinetics of competitive radioligand binding predicted by the law of mass action. *Molecular Pharmacology*, 25(1):1–9, 1984.
- [192] Tom Dixon, Derek MacPherson, Barmak Mostofian, Taras Dauzhenka, Samuel Lotz, Dwight McGee, Sharon Shechter, Utsab R. Shrestha, Rafal Wiewiora, Zachary A. McDargh, Fen Pei, Rajat Pal, João V. Ribeiro, Tanner Wilkerson, Vipin Sachdeva, Ning Gao, Shourya Jain, Samuel Sparks, Yunxing Li, Alexander Vinitzky, Asghar M. Razavi, István Kolossváry, Jason Imbriglio, Artem Evdokimov, Louise Bergeron, Alex Dickson, Huafeng Xu, Woody Sherman, and Jesus A. Izaguirre. Atomic-resolution prediction of degrader-mediated ternary complex structures by combining molecular simulations with hydrogen deuterium exchange. *bioRxiv*, 2021.
- [193] Tao Wu, Hojong Yoon, Yuan Xiong, Sarah E. Dixon-Clarke, Radosław P. Nowak, and Eric S. Fischer. Targeted protein degradation as a powerful research tool in basic biology and drug target discovery. *Nature Structural & Molecular Biology*, 27(7):605–614, 07 2020.
- [194] James Schiemer, Reto Horst, Yilin Meng, Justin I. Montgomery, Yingrong Xu, Xidong Feng, Kris Borzilleri, Daniel P. Uccello, Carolyn Leverett, Stephen Brown, Ye Che, Matthew F. Brown, Matthew M. Hayward, Adam M. Gilbert, Mark C. Noe, and Matthew F. Calabrese. Snapshots and ensembles of btk and ciap1 protein degrader ternary complexes. *Nature Chemical Biology*, 17:152–160, 2021.
- [195] Matthieu Schapira, Matthew F. Calabrese, Alex N. Bullock, and Craig M. Crews. Targeted protein degradation: expanding the toolbox. *Nature Reviews Drug Discovery*, 18(12):949–963, 2019.
- [196] Kevin G. Coleman and Craig M. Crews. Proteolysis–Targeting Chimeras: Harnessing the Ubiquitin–Proteasome System to Induce Degradation of Specific Target Proteins. *Annual Review of Cancer Biology*, 2(1):1–18, 2017.

- [197] Mary E. Matyskiela, Weihong Zhang, Hon-Wah Man, George Muller, Godrej Khambatta, Frans Baculi, Matthew Hickman, Laurie LeBrun, Barbra Pagarigan, Gilles Carmel, Chin-Chun Lu, Gang Lu, Mariko Riley, Yoshitaka Satoh, Peter Schafer, Thomas O. Daniel, James Carmichael, Brian E. Cathers, and Philip P. Chamberlain. A Cereblon Modulator (CC-220) with Improved Degradation of Ikaros and Aiolos. *Journal of Medicinal Chemistry*, 61(2):535–542, 2018.
- [198] Philip P Chamberlain, Antonia Lopez-Girona, Karen Miller, Gilles Carmel, Barbra Pagarigan, Barbara Chie-Leon, Emily Rychak, Laura G Corral, Yan J Ren, Maria Wang, Mariko Riley, Silvia L Delker, Takumi Ito, Hideki Ando, Tomoyuki Mori, Yoshinori Hirano, Hiroshi Handa, Toshio Hakoshima, Thomas O Daniel, and Brian E Cathers. Structure of the human Cereblon–DDB1–lenalidomide complex reveals basis for responsiveness to thalidomide analogs. *Nature Structural & Molecular Biology*, 21(9):803–809, 2014.
- [199] B. A. Kochert, R. E. Iacob, T. E. Wales, A. Makriyannis, and J. R. Engen. Hydrogen-deuterium exchange mass spectrometry to study protein complexes. *Methods in Molecular Biology*, 1764:153–171, 2018.
- [200] Nobumichi Ohoka, Keiichiro Okuhira, Masahiro Ito, Katsunori Nagai, Norihito Shibata, Takayuki Hattori, Osamu Ujikawa, Kenichiro Shimokawa, Osamu Sano, Ryokichi Koyama, Hisashi Fujita, Mika Teratani, Hirokazu Matsumoto, Yasuhiro Imaeda, Hiroshi Nara, Nobuo Cho, and Mikihiro Naito. In Vivo Knockdown of Pathogenic Proteins via Specific and Nongenetic Inhibitor of Apoptosis Protein (IAP)-dependent Protein Erasers (SNIPERs)*. *Journal of Biological Chemistry*, 292(11):4556–4570, 2017.
- [201] Jieli Wei, Fanye Meng, Kwang-Su Park, Hyerin Yim, Julia Velez, Prashasti Kumar, Li Wang, Ling Xie, He Chen, Yudao Shen, Emily Teichman, Dongxu Li, Gang Greg Wang, Xian Chen, H. Ümmit Kaniskan, and Jian Jin. Harnessing the E3 Ligase KEAP1 for Targeted Protein Degradation. *Journal of the American Chemical Society*, 143(37):15073–15083, 2021.
- [202] A Rodriguez-Gonzalez, K Cyrus, M Salcius, K Kim, C M Crews, R J Deshaies, and K M Sakamoto. Targeting steroid hormone receptors for ubiquitination and degradation in breast and prostate cancer. *Oncogene*, 27(57):7201–7211, 2008.
- [203] Wai-Ching Hon, Michael I Wilson, Karl Harlos, Timothy DW Claridge, Christopher J Schofield, Christopher W Pugh, Patrick H Maxwell, Peter J Ratcliffe, David I Stuart, and E Yvonne Jones. Structural basis for the recognition of hydroxyproline in hif-1 α by pvh1. *Nature*, 417(6892):975–978, 2002.
- [204] Kathleen M. Sakamoto, Kyung B. Kim, Akiko Kumagai, Frank Mercurio, Craig M. Crews, and Raymond J. Deshaies. Protacs: Chimeric molecules that target proteins to the Skp1–Cullin–F box complex for ubiquitination and degradation. *Proceedings of*

- the National Academy of Sciences*, 98(15):8554–8559, 2001.
- [205] Scott J Hughes and Alessio Ciulli. Molecular recognition of ternary complexes: a new dimension in the structure-guided design of chemical degraders. *Essays in Biochemistry*, 61(5):505–516, 2017.
 - [206] Adelajda Zorba, Chuong Nguyen, Yingrong Xu, Jeremy Starr, Kris Borzilleri, James Smith, Hongyao Zhu, Kathleen A. Farley, WeiDong Ding, James Schiemer, Xidong Feng, Jeanne S. Chang, Daniel P. Uccello, Jennifer A. Young, Carmen N. Garcia-Irrizary, Lara Czabaniuk, Brandon Schuff, Robert Oliver, Justin Montgomery, Matthew M. Hayward, Jotham Coe, Jinshan Chen, Mark Niosi, Suman Luthra, Jaymin C. Shah, Ayman El-Kattan, Xiayang Qiu, Graham M. West, Mark C. Noe, Veerabahu Shanmugasundaram, Adam M. Gilbert, Matthew F. Brown, and Matthew F. Calabrese. Delineating the role of cooperativity in the design of potent PROTACs for BTK. *Proceedings of the National Academy of Sciences*, 115(31):201803662, 2018.
 - [207] R. P. Nowak, S. L. DeAngelo, D. Buckley, Z. He, K. A. Donovan, J. An, N. Safaee, M. P. Jedrychowski, C. M. Ponthier, M. Ishoe, T. Zhang, J. D. Mancias, N. S. Gray, and E. S. Bradner, J. E. Fischer. Plasticity in binding confers selectivity in ligand-induced protein degradation. *Nature Chemical Biology*, 14(7):706–714, 2018.
 - [208] William Farnaby, Manfred Koegl, Michael J Roy, Claire Whitworth, Emelyne Diers, Nicole Trainor, David Zollman, Steffen Steurer, Jale Karolyi-Oezguer, Carina Riedmueller, et al. Baf complex vulnerabilities in cancer demonstrated via structure-based protac design. *Nature chemical biology*, 15(7):672–680, 2019.
 - [209] Hai-Tsang Huang, Dennis Dobrovolsky, Joshiawa Paulk, Guang Yang, Ellen L Weisberg, Zainab M Doctor, Dennis L Buckley, Joong-Heui Cho, Eunhwa Ko, Jaebong Jang, et al. A chemoproteomic approach to query the degradable kinome using a multi-kinase degrader. *Cell chemical biology*, 25(1):88–99, 2018.
 - [210] Daniel P Bondeson, Blake E Smith, George M Burslem, Alexandru D Buhimschi, John Hines, Saul Jaime-Figueroa, Jing Wang, Brian D Hamman, Alexey Ishchenko, and Craig M Crews. Lessons in protac design from selective degradation with a promiscuous warhead. *Cell chemical biology*, 25(1):78–87, 2018.
 - [211] Carl C Ward, Jordan I Kleinman, Scott M Brittain, Patrick S Lee, Clive Yik Sham Chung, Kenneth Kim, Yana Petri, Jason R Thomas, John A Tallarico, Jeffrey M McKenna, et al. Covalent ligand screening uncovers a rnf4 e3 ligase recruiter for targeted protein degradation applications. *ACS chemical biology*, 14(11):2430–2440, 2019.
 - [212] Michael Zengerle, Kwok-Ho Chan, and Alessio Ciulli. Selective Small Molecule Induced Degradation of the BET Bromodomain Protein BRD4. *ACS Chemical Biology*,

- 10(8):1770–1777, 2015.
- [213] Morgan S Gadd, Andrea Testa, Xavier Lucas, Kwok-Ho Chan, Wenzhang Chen, Douglas J Lamont, Michael Zengerle, and Alessio Ciulli. Structural basis of PROTAC cooperative recognition for selective protein degradation. *Nature Chemical Biology*, 13(5):514–521, 2017.
 - [214] Andrea Testa, Scott J. Hughes, Xavier Lucas, Jane E. Wright, and Alessio Ciulli. Structure-Based Design of a Macrocyclic PROTAC. *Angewandte Chemie International Edition*, 59(4):1727–1734, 2020.
 - [215] Daniel Zaidman, Jaime Prilusky, and Nir London. PROsettaC: Rosetta Based Modeling of PROTAC Mediated Ternary Complexes. *Journal of Chemical Information and Modeling*, 60(10):4894–4903, 2020.
 - [216] Nan Bai, Palani Kirubakaran, and John Karanicolas. Rationalizing PROTAC-mediated ternary complex formation using Rosetta. *Journal of Chemical Information and Modeling*, 61(3):1368–1382, 2021.
 - [217] Michael L Drummond, Andrew Henry, Huifang Li, and Christopher I Williams. Improved Accuracy for Modeling PROTAC-Mediated Ternary Complex Formation and Targeted Protein Degradation via New In Silico Methodologies. *Journal of Chemical Information and Modeling*, 60(10):5234–5254, 2020.
 - [218] Muhammed Shaheer, Ravi Singh, and M Elizabeth Sobhia. Protein degradation: a novel computational approach to design protein degrader probes for main protease of SARS-CoV-2. *Journal of Biomolecular Structure and Dynamics*, pages 1–13, 2021.
 - [219] Haoyun Feng, Ronan Costaeuec, Eric Darve, and Jesús A. Izaguirre. A comparison of weighted ensemble and Markov state model methodologies. *The Journal of Chemical Physics*, 142(21):214113, 2015.
 - [220] Brooke E Husic and Vijay S Pande. Markov state models: From an art to a science. *Journal of the American Chemical Society*, 140(7):2386–2396, 2018.
 - [221] Kevin B. Dagbay, Nicolas Bolik-Coulon, Sergey N. Savinov, and Jeanne A. Hardy. Caspase-6 Undergoes a Distinct Helix-Strand Interconversion upon Substrate Binding*. *Journal of Biological Chemistry*, 292(12):4885–4897, 2017.
 - [222] Kevin B. Dagbay and Jeanne A. Hardy. Multiple proteolytic events in caspase-6 self-activation impact conformations of discrete structural regions. *Proceedings of the National Academy of Sciences of the United States of America*, 114(38):E7977–E7986, 2017.
 - [223] Derek J. MacPherson, Caitlyn L. Mills, Mary Jo Ondrechen, and Jeanne A. Hardy.

- Tri-arginine exosite patch of caspase-6 recruits substrates for hydrolysis. *Journal of Biological Chemistry*, 294(1):71–88, 2019.
- [224] Thomas E. Wales, Keith E. Fadgen, Geoff C. Gerhardt, and John R. Engen. High-Speed and High-Resolution UPLC Separation at Zero Degrees Celsius. *Analytical and Bioanalytical Chemistry*, 80(17):6815–6820, 2008.
- [225] Thomas E. Wales and John R. Engen. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrometry Reviews*, 25(1):158–170, 2006.
- [226] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, and K. S. Wilson. Overview of the ccp4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography*, 67(4):235–242, 2011.
- [227] P. Emsley and K. Cowtan. Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography*, 60(12):2126–2132, 2004.
- [228] G. N. Murshudov, A. A. Vagin, and E. J. Dodson. Application of maximum likelihood refinement. *Refinement of Protein structures, Proceedings of Daresbury Study Weekend*, 1996.
- [229] G. N. Murshudov, A. A. Vagin, and E. J. Dodson. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D: Biological Crystallography*, 53:240–255, 1997.
- [230] N. J. Pannu, G. N. Murshudov, E. J. Dodson, and R. A. Read. Incorporation of prior phase information strengthen maximum-likelihood structure refinement. *Acta Crystallographica Section D: Biological Crystallography*, 54:1285–1294, 1998.
- [231] G. N. Murshudov, A. Lebedev, A. A. Vagin, K. S. Wilson, and E. J. Dodson. Efficient anisotropic refinement of macromolecular structures using fft. *Acta Crystallographica Section D: Biological Crystallography*, 55:247–255, 1999.
- [232] M. Winn, M. Isupov, and G. N. Murshudov. Use of tls parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallographica Section D: Biological Crystallography*, 57:122–133, 2001.
- [233] R. Steiner, A. Lebedev, and G. N. Murshudov. Fisher’s information matrix in maximum likelihood molecular refinement. *Acta Crystallographica Section D: Biological Crystallography*, 59:2114–2124, 2003.
- [234] M. Winn, G. N. Murshudov, and M. Z. Papiz. Macromolecular tls refinement in remlcp at moderate resolutions. *Methods in Enzymology*, 374:300–321, 2003.

- [235] P. Skubak, G. N. Murshudov, and N. S. Pannu. Direct incorporation of experimental phase information in model refinement. *Acta Crystallographica Section D: Biological Crystallography*, 60:2196–2201, 2004.
- [236] A. A. Vagin, R. S. Steiner, A. A. Lebedev, L. Potterton, S. McNicholas, F. Long, and G. N. Murshudov. Refmac5 dictionary: organisation of prior chemical knowledge and guidelines for its use. *Acta Crystallographica Section D: Biological Crystallography*, 60:2284–2295, 2004.
- [237] Michael L. Drummond, Andrew Henry, Huifang Li, and Christopher I Williams. Improved Accuracy for Modeling PROTAC-Mediated Ternary Complex Formation and Targeted Protein Degradation via New In Silico Methodologies. *Journal of Chemical Information and Modeling*, 60(10):5234–5254, 2020.
- [238] Philipp Pracht, Fabian Bohle, and Stefan Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.*, 22:7169–7192, 2020.
- [239] Stefan Grimme. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *Journal of Chemical Theory and Computation*, 15(5):2847–2862, 2019.
- [240] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, 2019.
- [241] Jeffrey J. Gray, Stewart Moughon, Chu Wang, Ora Schueler-Furman, Brian Kuhlman, Carol A. Rohl, and David Baker. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, 331(1):281–299, 2003.
- [242] Nicholas A Marze, Shourya S Roy Burman, William Sheffler, and Jeffrey J Gray. Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics*, 34(20):3461–3469, 2018.
- [243] Giovanni Bussi. Hamiltonian replica exchange in gromacs: a flexible implementation. *Molecular Physics*, 112(3-4):379–384, 2014.
- [244] Lingle Wang, Richard A. Friesner, and B. J. Berne. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (rest2). *J The Journal of Physical Chemistry B*, 115(30):9431–9438, 08 2011.
- [245] Xingui Liu, Xuan Zhang, Dongwen Lv, Yaxia Yuan, Guangrong Zheng, and Daohong Zhou. Assays and technologies for developing proteolysis targeting chimera degraders.

Future Medicinal Chemistry, 12(12):1155–1179, 2020.

- [246] M. C. Deller, L. Kong, and B. Rupp. Protein stability: A crystallographer’s perspective. *Acta Crystallographica Section F Structural Biology Communications*, 72(2):72–95, 2016.
- [247] K. B. Dagbay and J. A. Hardy. Multiple proteolytic events in caspase-6 self-activation impact conformations of discrete structural regions. *Proceedings of the National Academy of Sciences of the United States of America*, 114(38):E7977–E7986, 2017.
- [248] E. S. Gallagher and J. W. Hudgens. Mapping protein-ligand interactions with proteolytic fragmentation, hydrogen/deuterium exchange-mass spectrometry. *Methods in Enzymology*, 566, 2016.
- [249] Ali S. Saglam and Lillian T. Chong. Protein–protein binding pathways and calculations of rate constants using fully-continuous, explicit-solvent simulations. *Chemical Science*, 10(8):2360–2372, 2018.
- [250] Harry C Jubb, Alicia P Higuero, Bernardo Ochoa-Montano, Will R Pitt, David B Ascher, and Tom L Blundell. Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures. *Journal of Molecular Biology*, 429(3):365–371, 2017.
- [251] Mengru Mira Zhang, Brett R. Beno, Richard Y.-C. Huang, Jagat Adhikari, Ekaterina G. Deyanova, Jing Li, Guodong Chen, and Michael L. Gross. An integrated approach for determining a protein–protein binding interface in solution and an evaluation of hydrogen–deuterium exchange kinetics for adjudicating candidate docking models. *Analytical Chemistry*, 91(24):15709–15717, 2019.
- [252] S. Eron. Finding a way out of the labyrinth: degrader-induced ternary complex modeling. Finding a way out of the labyrinth: degrader-induced ternary complex modeling. July 14, 2021 *The Protein Society 35th Anniversary Symposium*, July 7-9, 12-14, 2021, 2021.
- [253] Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation*, 11:5525–5542, 2015.
- [254] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72:3634–3637, 1994.