

Incorporating Hotspot Mapping and Allostery in Structure Based Drug Design

by

Phani Ghanakota

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Medicinal Chemistry)
in the University of Michigan
2015

Doctoral Committee:

Professor Heather A. Carlson, Chair
Professor Charles L. Brooks III
Assistant Professor Barry J. Grant
Professor Henry I. Mosberg

© Phani Ghanakota, 2015

Table of Contents

List of Figures	v
List of Tables	xiii
List of Appendices	xvi
Abstract	xvii
Chapter 1. Introduction	1
1.1 Multiple Solvent Crystal Structures	1
1.2 Computational approaches for reproducing MSCS	2
1.2.1 Static protein conformation based approaches	2
1.2.2 Molecular dynamic simulation based approaches	5
1.3 Overview of thesis.....	22
Chapter 2. Comparing pharmacophore models derived from X-ray and NMR ensembles.....	24
2.1 Abstract.....	24
2.2 Introduction	25
2.3 Methods.....	26
2.3.1 Protein Preparation.....	26
2.3.2 Probe Flooding, Minimization, and Clustering.....	27
2.3.3 Creation of ligand and decoy databases.....	29
2.3.4 Evaluation of Pharmacophore Models.	30
2.4 Results and Discussion	31
2.4.1 Protein conformational sampling in crystal structures and NMR ensembles.	31
2.4.2 Comparison of crystal and NMR pharmacophore models.....	33
2.4.3 Evaluation of MPS pharmacophore models	42
2.4.4 Locating and characterizing extraneous elements in crystal pharmacophore models. ...	48
2.5 Conclusions	52
2.6 Supplementary Information.....	53
Chapter 3. Moving Beyond Active-site Detection: MixMD Applied to Allosteric Systems.....	54
3.1 Abstract.....	54
3.2 Introduction	55
3.3 Methods.....	57

3.3.1	MixMD simulation setup.....	57
3.3.2	Parametrization of acetate and methyl ammonium for use in MixMD.....	58
3.3.3	Processing MixMD results.....	59
3.4	Results and Discussion	60
3.4.1	Choice of protein targets and conformations.....	60
3.4.2	Identifying and ranking hotspots on the protein surface	61
3.4.3	Mapping active and allosteric sites with MixMD	64
3.4.4	Alternative probes for mapping charged binding sites.....	66
3.4.5	Identifying co-activator/cofactor and protein oligomerization/ packing interfaces (Evaluating MixMD maps at lower sigma values)	74
3.4.6	Limits of conformational sampling with MixMD	75
3.5	Conclusion.....	76
Chapter 4.	Free energies and entropies of binding sites identified by MixMD simulations	78
4.1	Abstract.....	78
4.2	Introduction	79
4.3	Methods.....	80
4.3.1	Simulation of 5%v/v box of MixMD probes to obtain expected occupancies (no proteins present)	80
4.3.2	Deriving free energies from MixMD simulations.....	81
4.4	Results and Discussion	83
4.4.1	The maximum free energy of a probe is dictated by system setup.....	83
4.4.2	Free energy calculations from mixed-solvent simulations	85
4.4.3	Free energies of MixMD binding sites calculated by occupancies	86
4.4.4	Ranking MixMD binding sites based on configurational entropy.....	90
4.4.5	Entropies across MixMD binding sites.....	93
4.4.6	Validating configurational entropies obtained from MixMD binding sites	94
4.4.7	Comparison of entropies across protein targets	99
4.4.8	Conclusion.....	102
4.5	Supplementary Information.....	103
Chapter 5.	Hitting an Undruggable Target: A Blinded Test of MixMD on Heat Shock Protein 27	104
5.1	Abstract.....	104
5.2	Introduction	105

5.3	Methods.....	107
5.3.1	MixMD simulation setup.....	107
5.3.2	Analyzing MixMD results	108
5.3.3	Deriving free energies from MixMD simulations.....	108
5.3.4	NMR HSQC experiments with Hsp27	109
5.4	Results and Discussion	109
5.4.1	Mapping binding sites on HSP27 using MixMD	109
5.4.2	Comparing ¹⁵ N, ¹ H HSQC spectra of organic probe molecules to MixMD probe mapping 113	
5.4.3	Experimental support from crystallography	115
5.4.4	High Throughput Screening – Derived Inhibitors.....	116
5.5	Conclusions	116
Chapter 6.	Conclusions and Future Directions	118
6.1	Significant contributions of this thesis.....	118
6.2	Future Directions	120
Appendices.....		124
References		175

List of Figures

Figure 1-1. An overlay of the all the MSCS crystal structure of elastase is shown. The co-solvent molecules cluster within the active-site. The active-site is denoted with a red circle. The list of crystal structures includes 2FOE, 2FOD, 2FOG, 2FOH, 2FOF, 2FOA, 2FOB, 2FO9, 2FOC. 2

Figure 2-1. A detailed description of the MPS method is presented using benzene probes as an example. A) The protein active site is flooded with 500 probe molecules. B) These molecules are minimized, independent of each other using the MUSIC routine in BOSS. C) The probe molecules are clustered and represented by “parent probes” which are the single benzene molecule with the best interaction energy in each cluster in B. D) Steps A-C are carried out on all structures in the ensemble, and they are overlaid using the wRMSD method. E) Clusters of “parent probes” are identified manually. F) Consensus clusters are identified when at least 50% of the conformations contain a parent probe in the same location. All probes farther than 10Å of the center of the active site are ignored. The center of each pharmacophore element is derived from the center of mass of the parents in the consensus cluster, and the radius of the element is set by the RMSD of the parents. 28

Figure 2-2. Receiver Operator Characteristic plots are shown for two cases, the first one corresponds to an exaggerated case of the more traditional use of ROC plots for continuous scores such as those obtained from docking results. The second plot illustrates the ROC plots generated by screening MPS pharmacophore models. Each discrete point on this line corresponds to the pharmacophore screening results obtained by gradually increasing the radii of the pharmacophore elements from 1× to 3× RMSD. Thus a label “5/6 2.33×” denotes the screening results from a MPS pharmacophore model whose radii have been multiplied by 2.33 and requires 5 of its 6 pharmacophore elements to be matched for a hit to be identified. 31

Figure 2-3. MPS pharmacophores are shown with 1× RMSD radii, which indicates tighter or looser position constraints. Pharmacophore models for all the protein targets are color coded to represent different interactions: Red – Donor, Blue – Acceptor, Purple – Doneptor, Green – Aromatic, and Cyan – Hydrophobic. A) The MPS pharmacophore model for Src SH2 derived from X-ray structures. B) The MPS pharmacophore model for Src SH2 derived from the NMR ensemble. C) The ligands from X-ray structures overlaid on top of the Src SH2 X-ray model. D) The ligands from X-ray structures overlaid on top of the Src SH2 NMR model. 35

Figure 2-4. Coloring and radii of the pharmacophore elements are the same as in Figure 2-3. A) The MPS pharmacophore model for GRB2 SH2 derived from X-ray structures. B) The MPS pharmacophore model for GRB2 SH2 derived from the NMR ensemble. C) The ligands from X-ray structures overlaid on top of the GRB2 SH2 MPS X-ray pharmacophore model. D) The ligands from X-ray structures overlaid on top of the GRB2 SH2 MPS NMR pharmacophore model. In Figure 2-4A and Figure 2-4B, tryptophan 121 is rendered as sticks between the pY+1 and pY+2 surfaces. 37

Figure 2-5. Coloring and radii of the pharmacophore elements are the same as in Figure 2-3. A) The MPS pharmacophore model for FKBP12 derived from X-ray structures. B) The MPS pharmacophore model for FKBP12 derived from the NMR ensemble. C) The ligands from X-ray structures overlaid on top of the FKBP12 X-ray model. D) The ligands from X-ray structures overlaid on top of the FKBP12 NMR model. . 39

Figure 2-6. Coloring and radii of the pharmacophore elements are the same as in Figure 2-3. MPS pharmacophore models derived from PPAR- γ X-ray and NMR ensembles are shown along with a representative protein conformation. The tyrosine residue 473 which is part of Helix 12 is shown in pink. In the X-ray ensemble, there is limited sampling of the tyrosine residue 473, which corresponds to the active form of the protein. In the NMR ensemble; this residue samples the inactive conformation of the protein. A) The MPS pharmacophore model for PPAR- γ derived from X-ray structures. B) The MPS pharmacophore model for PPAR- γ derived from the NMR ensemble. C) The ligands from X-ray structures overlaid on top of the PPAR- γ MPS X-ray pharmacophore model. D) The ligands from X-ray structures overlaid on top of the PPAR- γ MPS NMR pharmacophore model. In Figure 2-6A and Figure 2-6B, the location of the binding site is shown by rendering rosiglitazone as a stick model obtained from the PDB ID: 1ZGY (colored brown)..... 41

Figure 2-7. ROC plots of crystal and NMR pharmacophore models of Src SH2 are shown along with a label for the model that displays the best performance. The best performing crystal pharmacophore model was at $2.66 \times$ RMSD using seven out of ten pharmacophore elements. The best performing NMR pharmacophore model was achieved at $2.66 \times$ RMSD using all six pharmacophore elements. 45

Figure 2-8. ROC plots of crystal and NMR pharmacophore models of Grb2 SH2 are shown along with a label for the model that displays the best performance. The best performing crystal pharmacophore model was at $3.00 \times$ RMSD using eight out of nine pharmacophore elements. The best performing NMR pharmacophore model was achieved at $3.00 \times$ RMSD using all six pharmacophore elements. 45

Figure 2-9. ROC plots of crystal and NMR pharmacophore models of FKBP12 are shown along with a label for the model that displays the best performance. The best performing crystal pharmacophore model was at $2.66 \times$ RMSD using seven out of fourteen pharmacophore elements. The best performing NMR pharmacophore model was achieved at $1.00 \times$ RMSD using all four pharmacophore elements. 46

Figure 2-10. ROC plots of crystal and NMR pharmacophore models of PPAR- γ are shown along with a label for the model that displays the best performance. The best performing crystal model was at $2.33 \times$ RMSD using five out of six pharmacophore elements. The best performing NMR model was achieved at $1.33 \times$ RMSD using all three pharmacophore elements. The NMR pharmacophore model was built from an NMR ensemble that samples the inactive conformation, so this model was expected to perform poorly. 48

Figure 2-11. ROC plots are shown for the PPAR- γ crystal pharmacophore model with cutoffs of 8\AA and 7\AA from the center of the binding site. The best performing crystal pharmacophore model at a cutoff of 8\AA was $1.00 \times$ RMSD using four out of five pharmacophore elements and $1.33 \times$ RMSD using four out of four pharmacophore elements for a cutoff of 7\AA 50

Figure 2-12. ROC plots of Grb2 SH2 crystal model truncated to 8Å and 7Å are shown. The best performing model at a cutoff of 8Å was 3.00 × RMSD using eight out of eight pharmacophore elements and 2.33 × RMSD using six out of six pharmacophore elements for a cutoff of 7Å. 50

Figure 2-13. ROC plots of FKBP12 crystal pharmacophore model truncated to 9Å and 8Å are shown along with a label for the model that displays the best performance. The best performing crystal pharmacophore model at a cutoff of 9Å was 2.66 × RMSD using seven out of thirteen pharmacophore elements and 3.00 × RMSD using seven out of ten pharmacophore elements for a cutoff of 8Å. 51

Figure 2-14. A) FKBP12 crystal pharmacophore model and B) FKBP12 NMR pharmacophore model. The models are overlaid on the protein which is color coded by C_α RMSD after wRMSD alignment. The color scale ranges from Blue (0.1 Å RMSD) to Red (3.8 Å RMSD). The increased flexibility of the NMR ensemble reduces the consensus across the probes used in constructing the model, which removes several elements present in the crystal pharmacophore model. 51

Figure 2-15. ROC plots of Src SH2 crystal pharmacophore model truncated to 9Å and 8Å are shown. The best performing crystal pharmacophore model at a cutoff of 9Å was 2.66 × RMSD using six out of eight pharmacophore elements and 2.00 × RMSD using five out of six pharmacophore elements for a cutoff of 8Å. 52

Figure 3-1 (A) The final snapshot from a simulation of ~2.5%v/v mixture of acetate, methyl ammonium and water for 5ns demonstrates proper mixing was achieved. Acetate ions are colored purple, methyl ammonium ions are colored green and water molecules are colored white. (B) Adequate mixing of acetate, methyl ammonium probe molecules in MixMD was confirmed by radial distribution functions that displayed a probability of 1.0 at long range distances. 59

Figure 3-2 The MixMD maps for Abl Kinase are contoured at varying sigma values from 90σ to 20σ to show the degree of molecular surface mapped by the probe atoms. The maps are color coded to represent MixMD maps derived from different probes . Orange - acetonitrile, blue – isopropanol, and magenta – pyrimidine. At 90σ (Figure 3-2A), the allosteric site shows the highest occupied points. Maps contoured at 85σ, show a second hotspot in the active site. In contouring the MixMD maps from Figure 3-2C-Figure 3-2F at successively lower sigma values, additional hotspots appear and are numbered based on their order of appearance. Unless sites are mapped by more than one probe type when contoured at 20σ they are ignored. The active site ligand (PDB ID: 3KFA, Green) and allosteric site ligand (PDB ID: 3K5V, brown) are only shown for reference in E and F to orient the viewer towards the location of the active and allosteric sites. We emphasize that no ligands were present in the MixMD simulations. 63

Figure 3-3 The first four hotspots from the Abl Kinase MixMD maps identified the active and allosteric sites. The hotspot rankings are shown on top of the protein structure. The active site ligand (PDB ID: 3KFA, Green) and the allosteric site ligand (PDB ID: 3K5V, Brown) are shown for reference. (A) The four hotspots that map the active and allosteric site are shown contoured at 20σ with the spurious sites not shown. (B) MixMD maps of Abl Kinase contoured at 35σ (all spurious sites are shown) are shown with examples (where available) of molecules from the PDB database bound in probe mapped locations on the protein surface. The crystal structure of the full length Abl protein (PDB ID: 1OPK) was aligned to

show the Kinase and SH2 domain interface mapped by the fourth hotspot in MixMD. A tyrosine residue at the packing interface is shown in black (PDB ID: 1OPL). The allosteric and competitive ligands are shown in brown (PDB ID: 3K5V) and green (PDB ID: 3KFA) respectively..... 68

Figure 3-4 (A) The location of just the top four hotspots contoured at 20σ on the surface of the androgen receptor are shown. The active site ligand (PDB ID: 2AM9, Green) and the allosteric site ligand (PDB ID: 2PIX, Brown) are shown for reference. A part of the alpha helix obstructing the view of the active site ligand has been hidden to provide a better view of the hotspots mapping the active site. (B) The MixMD maps of Androgen receptor are shown contoured at 35σ to demonstrate that hotspots ranked lower than the top four hotspots, correspond to locations that can be easily desolvated. The different molecules are color coded as follows, Black – PDB ID: 2QPY – Nuclear Receptor Co-Activator 2, Yellow – PDB ID: 4HLW – Glycerol, Pink – PDB ID: 2QPY – Protein Packing Interface..... 69

Figure 3-5 Just the top four hotspots for Pdk 1 Kinase are numbered and shown contoured at 20σ and raw occupancy maps are shown at 35σ . The first hotspot maps the hinge region of the active site. The second hotspot is located at the top of the protein. A cosolvent bound at this site is overlaid on top from another PDK1 Kinase protein structure (PDB ID: 3RWQ, Pink). The third ranked hotspot is known to bind a peptide (PDB ID: 3QC4, Yellow) and the fourth hotspot corresponds to the allosteric site. The active site ligand (PDB ID: 3RCJ, Green) and the allosteric site ligand from (PDB ID: 4AW0, Brown) are shown for reference..... 70

Figure 3-6 The top four hotspots for Farnesyl Pyrophosphate Synthase are contoured at 20σ . The first hotspot maps the dimer interface. The second hotspot maps the allosteric site (PDB ID: 3N5J, Brown). The third and fourth ranked hotspots map two different sub sites in the active site (PDB ID: 4DEM, Green). The protein was simulated as a monomer. However the dimer is shown to illustrate that the top ranked hotspot is located at the dimer interface. A tyrosine residue from one of the monomers colored dark green is shown to overlap with the first ranked hotspot at the dimer interface..... 71

Figure 3-7 Chk1 Kinase is shown with just the top four hotspots contoured at 20σ . While the first and the fourth hotspots map the active (PDB ID: 1ZYS, Green) and allosteric site (PDB ID: 3JVS, Brown). The second and the third ranked hotspot are located in the peptide substrate binding groove. 72

Figure 3-8 Glucokinase is shown with just the top four hotspots contoured at 20σ . The first hotspot extensively maps the allosteric site (PDB ID: 3H1V, Brown). The fourth ranked hotspot maps the ATP binding site on Glucokinase (PDB ID: 3FGU, Grey). However no examples of molecules could be found that bound to the second and third ranked hotspots for Glucokinase. B) Only a few very small sites are present 35σ , and they are clearly lower ranked and less occupied. 72

Figure 3-9 (A) The location of the top 4 hotspots is shown for Protein Tyrosine Phosphatase 1B (PTP1B). The first hotspot maps the location of a cosolvent binding site (PDB ID: 3RWQ, Pink). The maps occlude the visibility of this cosolvent). The second hotspot maps the allosteric site (PDB ID: 1T49, Brown). The third and fourth ranked hotspots are located in close proximity to protein packing interfaces (Hotspot 3 is located near the protein packing interface - Cyan colored – PDB ID: 2CMC, Hotspot 4 is located near another protein packing interface – PDB ID: 4GRY– Pea colored). (B) As PTP1B has a charged active site, we were interested to see if charged probes could map these sites. A MixMD simulation of acetate and

methyl ammonium was carried out and the top four hotspots ranked in the order in which they appear are shown contoured at 20σ . Acetate hotspots are colored red and methyl ammonium hotspots are colored blue. A fragment of the Insulin Receptor is overlaid on top of the protein (PDB ID: 1G1F, Black). The top two ranked hotspots which correspond to the acetate ion overlap both these sites which are known to bind phosphorylated tyrosine residues. 73

Figure 3-10 MixMD simulations were performed using the inactive (PDB ID: 3KFA, Figure 3-10A) and active (PDB ID: 1M52, Figure 3-10B) forms of Abl Kinase. The top six ranked hotspots are shown for both conformations to illustrate the rearrangement of the hotspot rankings. The allosteric and active sites were ranked first and second respectively irrespective of the conformation of ABL Kinase used for MixMD. The activation loop is colored red to show the difference in this region between the two protein conformations. The third ranked hotspot in the inactive conformation (Figure 3-10A) is now occupied by the activation loop in the active form of ABL Kinase, this leads to a rearrangement in ranking of several sites on the protein. Two protein packing interfaces are shown colored black (PDB ID: 1OPL) and pink (PDB ID: 3QRK). The allosteric site ligand shown for reference is colored brown (PDB ID: 3K5V)..... 75

Figure 4-1. The process of obtaining observed occupancies and free energies from MixMD simulations is depicted in subfigures a-f. a) The grid points are sorted from highest to lowest occupancy, based on the counts of the probe's CoM. The size of the red circles on the grid indicates high vs low occupancies. The top-three grid points with the highest occupancies are shown for the purpose of demonstration. b) The grid point with the highest occupancy is taken to be the center of the first probe. All grid points enclosed within the volume of a probe are added to obtain its observed occupancy. c) After processing a given probe location, the grid points associated with this probe are removed from the search process. d) The observed occupancy is calculated for the second probe centered on the next grid point with the highest occupancy. e) Upon obtaining the occupancy of the probe at this second grid point, it is removed from the search process. e) This process is continued until all the grid points are exhaustively searched and assigned to a probe location..... 82

Figure 4-2. The relationship between free energy and expected occupancy is shown above using acetonitrile as an example. The free energy values for acetonitrile are calculated using equation (1) for observed occupancies ranging from 0.1 to 1 while varying the expected occupancy from 0.000046839 to 0.000071094. While the magnitude of the free energy values varies with the expected occupancy (left to right), the difference in free energies (spacing within the columns) increase or decrease same the amount for each expected occupancy. Thus, the relative rankings between different occupancies remain the same. 84

Figure 4-3. The normalized distribution profile of ΔG_{bind} for the top-10 MixMD probes is shown. Across the seven protein targets studied, binding free energies for isopropanol and pyrimidine were found to be more favorable than acetonitrile. Acetonitrile distribution is colored yellow, isopropanol distribution is colored purple, and pyrimidine distribution is colored purple. 89

Figure 4-4. The ligand efficiencies for the top-10 probes from MixMD simulations of seven protein systems are presented in the units kcal/mol-HA. Across the seven protein targets studied, ligand efficiencies for acetonitrile were more favorable than isopropanol and pyrimidine. Acetonitrile

distribution is colored yellow, isopropanol distribution is colored purple, and pyrimidine distribution is colored purple..... 90

Figure 4-5. The concept of entropy as the density of states is applied within the volume of a probe sphere. Each grid point within the volume is considered a state. The probability of each state (p_i) for each heavy atom is calculated using equation (4)..... 92

Figure 4-6. The distribution of $-\Delta S_{\text{probe}}$ for the top-50 MixMD probes ranked by free energy are presented for acetonitrile (orange), isopropanol (blue), and pyrimidine (purple). As expected, moving from the bulk into the binding site where the probes are restricted is unfavorable, thus $-\Delta S_{\text{probe}}$ values are positive..... 94

Figure 4-7. Acetonitrile HA densities are presented for the maximum, median, and minimum entropies reported in Table 4-4. The CoM that defines the binding site of the acetonitrile probe is shown as an orange sphere for reference. The normalized occupancies of all the atoms in A, B, and C are contoured at 0.005. The density of nitrogen atoms is colored blue, the density of the carbon atom in the middle of acetonitrile is colored cyan, and the density of the terminal carbon is colored brown. A) The $-\Delta S_{\text{probe}}$ is at a maximum, making this the most constrained probe in our dataset. Consequently, all atoms of the acetonitrile probe can be clearly seen in this case. B) The acetonitrile with the median $-\Delta S_{\text{probe}}$ shows some structure in the configurational sampling but also some latitude. C) Density for the case of minimum $-\Delta S_{\text{probe}}$ shows that the probe molecule at this location is freely rotating, and is close to the entropy of the bulk. As a result, the density is smeared out and overlapping..... 97

Figure 4-8. Normalized HA occupancies of isopropanol are presented for the maximum, median, and minimum entropies reported in Table 4-4. The CoM that defines the binding site of the isopropanol probe is shown as a blue sphere for reference. The density of all the atoms in A, B, and C are contoured at 0.005. The density of oxygen atoms is colored red, the density of the central carbon is colored cyan, and the two terminal carbons are colored blue and brown. A) The maximum $-\Delta S_{\text{probe}}$ example is the most constrained probe in our dataset. Consequently, all atoms of the isopropanol probe can be clearly seen in this case. B) The $-\Delta S_{\text{probe}}$ in this case is at the median of all processed sites, there is some structure in the probe molecule. Notably, the hydroxyl oxygen is sampling two hydrogen-bonding interactions. C) For the case of minimum $-\Delta S_{\text{probe}}$, the molecule at this location is freely rotating, and is close to the entropy of the bulk. As a result, the density is smeared out and can only be seen partly..... 98

Figure 4-9. Pyrimidine per probe normalized density is presented for the maximum, median, and minimum entropies reported in Table 4-4. The CoM that defines the binding site of the pyrimidine probe is shown as a purple sphere for reference. The normalized occupancies of all the atoms in A, B, and C are contoured at 0.005. The density of two nitrogen atoms are blue and green, whereas the density of carbon atoms is colored brown, purple, yellow, and orange. A) In the maximum $-\Delta S_{\text{probe}}$ case, the molecule is very constrained. Consequently, all atoms of the pyrimidine probe can be clearly seen in this case. B) The $-\Delta S_{\text{probe}}$ in this case is at the median of all processed sites, there is some structure in the probe molecule. Notably, the molecule is rotating and giving HA densities with a torus shape. It appears that the nitrogens are sampling three locations, separated by roughly 120° . C) For the minimum $-\Delta S_{\text{probe}}$ case, the probe molecule at this location is freely rotating, and is close to the entropy of the bulk. As a result, the density is smeared out and cannot be seen..... 99

Figure 5-1. The all atom MixMD maps contoured at 20σ along with the average protein structure of Hsp27 are shown above. The protein surface is white, the isopropanol density is blue, the pyrimidine density is purple, and the acetonitrile density is orange. The binding sites identified by MixMD are marked with red circles, and their symmetric relationships are shown by solid and dashed red circles. A) The front view of Hsp27 with the sites 1, 4, 5, and 6 is shown. B) The back view of Hsp27 is shown with site 2 and 3..... 111

Figure 5-2. The symmetry averaged free energies of the top 6 MixMD sites are shown for acetonitrile, isopropanol, and pyrimidine. All solvent maps are contoured at 20σ and are shown as a black mesh. A) The front face of Hsp27 along with sites 1, 4, 5, and 6 are shown. B) The back face of hsp27 with sites 2 and 3 are shown..... 112

Figure 5-3 MixMD maps of Hsp27 with acetonitrile, isopropanol, and pyrimidine are shown contoured at 20σ . These MixMD maps are color coded as orange for acetonitrile, blue for isopropanol, and purple for pyrimidine. Hsp27 residues that shift in at least two different co-solvent NMR experiments are shown colored green. Residues missing assignment and prolines which are invisible to the NMR HSQC experiment are colored grey. 114

Figure 5-4. A recently reported crystal structure of the LXL motif of the c-terminal region of Hsp27 (PDB ID: 4MJH) is shown in black stick model overlaid with the MixMD maps. The leucine residues from the LXL motif can be seen to bind site 4 and site 5 thereby validating these sites identified via MixMD. MixMD density is contoured at 20σ and is color coded to represent acetonitrile (orange), isopropanol (blue), and pyrimidine (purple)..... 116

Figure B-1 The names of the atoms within the probes A) Acetate and B) Methyl ammonium used in MixMD are presented..... 139

Figure E-1 MixMD maps of Farnesyl Pyrophosphate Synthase (FPPS) contoured at 35σ are shown with examples (where available) of molecules from the PDB database bound in probe mapped locations on the protein surface. FPPS functions as a dimer and a second copy of the dimer counterpart is shown in green with a tyrosine residue rendered as a stick model to illustrate the overlap of this residue with the MixMD maps. A protein packing interface rendered as cartoon is shown using PDB ID: 2P1C (Black). The allosteric and competitive ligands are shown for reference using the crystal structures PDB ID: 3N5J – Brown (Allosteric ligand) and PDB ID: 4DEM – Green (Competitive ligand). 160

Figure E-2 MixMD maps of Protein Tyrosine Phosphatase 1B (PTP1B) contoured at 35σ are shown with examples (where available) of molecules from the PDB database bound in probe mapped locations on the protein surface. The different protein packing interfaces and examples of cosolvent molecules known to bind PTP1B and mapped by MixMD are color coded as follows, PDB ID: 4GRY – Pea, PDB ID: 2CMC – Cyan, PDB ID: 2CMB – Black, PDB ID: 1GWZ– Purple, PDB ID: 1T49 – Brown (Allosteric ligand) and PDB ID: 2CMB – Green (Competitive ligand). 161

Figure E-3 The ranking of the top-four sites is shown for MixMD simulations starting from the active conformation of Abl Kinase (PDB ID: 1M52) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 130σ to 20σ (Shown in figures A – F)..... 162

Figure E-4 The ranking of the top-four sites is shown for MixMD simulations starting from Androgen Receptor (PDB ID: 2AM9) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 100σ to 20σ (Shown in figures A – F). 163

Figure E-5 The ranking of the top-four sites is shown for MixMD simulations starting from Pdk1 Kinase (PDB ID: 3RCJ) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 100σ to 20σ (Shown in figures A – F). 164

Figure E-6 The ranking of the top-four sites is shown for MixMD simulations starting from Farnesyl Pyrophosphate Synthase (PDB ID: 4DEM) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 80σ to 20σ (Shown in figures A – F). 165

Figure E-7 The ranking of the top-four sites is shown for MixMD simulations starting from Glucokinase (PDB ID: 3IDH) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 95σ to 20σ (Shown in figures A – F). 166

Figure E-8 The ranking of the top-four sites is shown for MixMD simulations starting from CHK1 Kinase (PDB ID: 1ZYS) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 110σ to 20σ (Shown in figures A – F). 167

Figure E-9 The ranking of the top-four sites is shown for MixMD simulations starting from Protein Tyrosine Phosphatase 1B (PDB ID: 2CMB) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 110σ to 20σ (Shown in figures A – F). 168

Figure E-10 The ranking of the top-four sites is shown for MixMD simulations starting from Protein Tyrosine Phosphatase 1B (PDB ID: 2AM9) using acetate (red) and methyl ammonium (blue) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 150σ to 20σ (Shown in figures A – F). 169

Figure F-1 A detailed description of the MixMD protocol used to predict and rank binding sites is shown. The MixMD maps are decreased gradually from high sigma values to low sigma values. This process is depicted using the following sigma cutoff values a) 55σ b) 45σ c) 35σ d) 30σ e) 25σ and f) 20σ 172

List of Tables

Table 1-1. A chronological list of cosolvent simulation techniques that have been used to identify binding sites on proteins is presented. The co-solvents and protein systems used in their studies are also shown.	5
Table 2-1. The number of high-affinity and low-affinity inhibitors/agonists used to validate MPS pharmacophore models are reported below for every protein target.	30
Table 2-2. The range of pair-wise C_{α} RMSDs for all the crystal and NMR structures reflects the greater flexibility in the NMR ensemble.	32
Table 2-3. The range of heavy-atom RMSDs in the binding site for the crystal structures and NMR ensembles provides support for the greater flexibility of the NMR ensemble.	33
Table 3-1 The protein structures used in MixMD are listed. The range of the all-atom RMSD for residues within 4Å of the allosteric site is shown from the MixMD starting conformation to protein conformations with allosteric ligands bound.	61
Table 4-1. The Expected occupancy for a grid point and the volume of a probe are presented for the MixMD probes acetonitrile, isopropanol, and pyrimidine.	80
Table 4-2. The ΔG_{bind} of the probes acetonitrile, isopropanol, and pyrimidine within the top-four MixMD sites (identified using our all atom binning method) are presented for the protein targets Ablkinase, Androgen receptor, Pdk1 kinase, Farnesyl Pyrophosphate Synthase, Chk1 kinase, Glucokinase, and Protein Tyrosine Phosphatase 1B. On rare occasions, the binding site identified by MixMD accommodated more than one probe. These sites were further divided in to subsites A and B.	87
Table 4-3. Maximum entropy at 300K (in kcal/mol) for a freely rotating and translating probe molecule is calculated. Under such conditions every grid point within the volume of a probe will be occupied with equal probability (p_{bulk}).	93
Table 4-4. The change in configurational entropy when moving a co-solvent from the bulk to the protein binding site were calculated using the top-fifty MixMD probes ranked by free energy from all seven allosteric systems. The minimum, median and maximum $-\Delta S_{\text{probe}}$ in this dataset was reported for each probe at 300K. The proteins to which these values belong along with the rank of the probe according to free energy are provided in brackets.	96
Table 4-5 The entropic penalties ($-\Delta S_{\text{probe}}$) of the MixMD binding sites are computed at 300K and are presented for the top-four MixMD binding sites.	100

Table 5-1. The free energies and ligand efficiencies of acetonitrile, isopropanol, and pyrimidine for the MixMD binding sites are presented below. For site 2, there were two acetonitrile probe molecules bound in the same site, as a result, the one with weaker binding is listed in parenthesis.	113
Table A-1 The residue and atom used for flooding the active site of proteins with probe molecules is shown for each protein.....	124
Table A-2 Crystal pharmacophore model coordinates and radius for Src SH2, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1O49.	125
Table A-3 Crystal pharmacophore model coordinates and radius for Src SH2 at a cutoff of 9Å, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1O49.	125
Table A-4 Crystal pharmacophore model coordinates and radius for Src SH2 at a cutoff of 8Å, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1O49.	125
Table A-5 NMR pharmacophore model coordinates and radius for Src SH2, the location of the pharmacophore model is relative to the NMR ensemble of FKBP12, PDB ID: 1FKR.	126
Table A-6 Crystal pharmacophore model coordinates and radius for Grb2 SH2, the location of the pharmacophore model is relative to the crystal structure of Grb2 SH2, PDB ID: 1JYR.	126
Table A-7 Crystal pharmacophore model coordinates and radius for Grb2 SH2 at a cutoff of 8Å, the location of the pharmacophore model is relative to the crystal structure of Grb2 SH2, PDB ID: 1JYR....	126
Table A-8 Crystal pharmacophore model coordinates and radius for Grb2 SH2 at a cutoff of 7Å, the location of the pharmacophore model is relative to the crystal structure of Grb2 SH2, PDB ID: 1JYR....	127
Table A-9 NMR pharmacophore model coordinates and radius for Grb2 SH2, the location of the pharmacophore model is relative to the NMR ensemble of Grb2 SH2, PDB ID: 1XON.	127
Table A-10 Crystal pharmacophore model coordinates and radius for FKBP12, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1FKB.....	127
Table A-11 Crystal pharmacophore model coordinates and radius for FKBP12 at a cutoff of 9Å, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1FKB.	128
Table A-12 Crystal pharmacophore model coordinates and radius for FKBP12 at a cutoff of 8Å, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1FKB.	128
Table A-13 NMR pharmacophore model coordinates and radius for FKBP12, the location of the pharmacophore model is relative to the NMR ensemble of FKBP12, PDB ID: 1FKR.	128
Table A-14 Crystal pharmacophore model coordinates and radius for PPAR-γ, the location of the pharmacophore model is relative to the crystal structure of PPAR-γ, PDB ID: 1ZGY.....	129
Table A-15 Crystal pharmacophore model coordinates and radius for PPAR-γ at a cutoff of 8Å, the location of the pharmacophore model is relative to the crystal structure of PPAR-γ, PDB ID: 1ZGY.	129
Table A-16 Crystal pharmacophore model coordinates and radius for PPAR-γ at a cutoff of 7Å, the location of the pharmacophore model is relative to the crystal structure of PPAR-γ, PDB ID: 1ZGY.	129

Table A-17 NMR pharmacophore model coordinates and radius for PPAR- γ , the location of the pharmacophore model is relative to the NMR ensemble of PPAR- γ , PDB ID: 2QMV.	129
Table A-18 PDB IDs and references of 22 SRC5H2 crystal structures used to create the Src SH2 X-ray MPS pharmacophore model.	129
Table A-19 PDB IDs and references of 18 GRB2 crystal structures used to create the Grb2 SH2 X-ray MPS pharmacophore model.	130
Table A-20 PDB IDs and references of 22 PPARGAMMA crystal structures used to create the PPAR- γ X-ray MPS pharmacophore model.	130
Table A-21 PDB IDs and references of 20 FKBP12 crystal structures used to create the FKBP12 X-ray MPS pharmacophore model.	130
Table A-22 PDB IDS and references of NMR ensembles used to create the NMR MPS pharmacophore models.....	130
Table A-23 ROC plot data for Src SH2 Crystal Pharmacophore model.	131
Table A-24 ROC plot data for Src SH2 Crystal pharmacophore model at a cutoff of 9Å.	131
Table A-25 ROC plot data for Src SH2 Crystal pharmacophore model at a cutoff of 8Å.	131
Table A-26 ROC plot data for Src SH2 NMR pharmacophore model.	132
Table A-27 ROC plot data for Grb2 Crystal pharmacophore model.	132
Table A-28 ROC plot data for Grb2 Crystal pharmacophore model at a cutoff of 8Å.	132
Table A-29 ROC plot data for Grb2 Crystal pharmacophore model at a cutoff of 7Å.	133
Table A-30 ROC plot data for Grb2 NMR pharmacophore model.	133
Table A-31 ROC plot data for FKBP12 Crytal Pharmacophore model.....	133
Table A-32 ROC plot data for FKBP12 Crystal Pharmacophore model at a cutoff of 9Å.	134
Table A-33 ROC plot data for FKBP12 Crystal pharmacophore model at a cutoff of 8Å.	135
Table A-34 ROC plot data for FKBP12 NMR pharmacophore model.	136
Table A-35 ROC plot data for PPAR- γ Crystal pharmacophore model.....	136
Table A-36 ROC plot data for PPAR- γ Crystal pharmacophore model at a cutoff of 8Å.....	136
Table A-37 ROC plot data for PPAR- γ crystal pharmacophore model at a cutoff of 7Å.	137
Table A-38 ROC plot data for PPAR- γ NMR pharmacophore model.....	138
Table B-1 The OPLS force field parameters used for acetate and methyl ammonium in MixMD simulations are provided in the table below.	139
Table G-1 The Hsp27 NMR chemical shift perturbation (CSP) normalized for each cosolvent are shown for acetonitrile, isopropanol, and pyrimidine. Residues that shift significantly are defined as those CSPs above the normalized average of 0.02 and are highlighted in pink.	173

List of Appendices

Appendix A	Raw data for MPS Pharmacophore models	124
Appendix B.	OPLS parameters for Acetate and Methyl Ammonium	139
Appendix C.	Python Script for Calculating MixMD Free energies	140
Appendix D.	Python Script for Calculating MixMD Entropies	149
Appendix E.	MixMD maps for Allosteric Systems	160
Appendix F.	MixMD maps for Heat Shock Protein 27	170
Appendix G.	NMR data for Heat Shock Protein 27	173

Abstract

Hotspots are defined as regions on the protein surface that disproportionately contribute to binding free energy. Mixed-solvent molecular dynamics (MixMD) is a hotspot mapping technique that relies on molecular dynamics simulations of binary solvent mixtures. Previous work in the group on MixMD has established the technique's effectiveness in capturing binding sites of small organic compounds. The MixMD approach embraces full protein flexibility while allowing for competition between probes and water. Sites preferentially mapped by probe molecules are more likely to be hotspots. First, we establish a rigorous protocol for the identification of hotspots on the binding surface. There are two important requirements: 1) the high-ranking hotspots must be mapped at very high signal-to-noise ratio and 2) the hotspots must be mapped by multiple probes. We have focused our probe molecule repertoire to include acetonitrile, isopropanol, and pyrimidine as these probes allowed us to capture a range of interaction types that include hydrophilic, hydrophobic, hydrogen-bonding and aromatic interactions. Second, we use MixMD to identify both competitive and allosteric sites on proteins. The test cases include Abl Kinase, Androgen Receptor, Chk1 Kinase, Glucokinase, Pdk1 Kinase, Protein-Tyrosine Phosphatase 1B, and Farnesyl Pyrophosphate Synthase. The success of the technique is demonstrated by the fact that the top four sites map the competitive and allosteric sites. We then present methodological developments for characterizing the free energies and entropies of binding sites identified by MixMD. Finally, the significance of these findings is strengthened by a successful prospective application of MixMD on Heat Shock Protein 27. Taken together, these studies demonstrate the powerful utility of MixMD in structure based drug design.

Chapter 1. Introduction

1.1 Multiple Solvent Crystal Structures

Some of the first studies of protein crystal structures with co-solvents were motivated by the difference in catalytic activities in the presence of organic solvents. Subsequently, the crystal structure of Subtilisin was solved in the presence of water and acetonitrile to determine if the difference in activities resulted from geometric changes in the active site (1). Similar studies were performed with γ -Chymotrypsin in Hexane (2). While there was no difference in the active site of the protein in both cases, these studies formed the basis for the seminal work by Dagmar Ringe and co-workers who proposed the use of multiple solvent crystal structures (MSCS) for locating binding sites on proteins (3). Initial validation studies for MSCS were performed on elastase using acetonitrile as the co-solvent (3). Acetonitrile in this study was found to map the active site and crystal packing interfaces in this protein. Subsequent studies with elastase extended the range of solvents to include acetone, dimethylformamide, 5-hexene-1,2-diol, isopropanol, ethanol, trifluoroethanol (4). Interestingly, when the range of solvents was extended, clusters of co-solvent molecules were found to populate the active site (Figure 1-1) and were dispersed near the crystal packing region. These results demonstrate the important concept that potential binding sites can be identified by requiring them to bind a diverse set of co-solvent molecules. Following this work, several studies have come to a similar conclusion using MSCS (5–7).

One of the limitations of MSCS is that many protein crystals are destabilized by the co-solvents. This results in a loss of resolution at best and no useful spectra at worst. In fact, the MSCS method was developed using cross-linked proteins to circumvent the issue. The limited application makes simulation methods an important component of using co-solvents on a wide variety of protein systems.

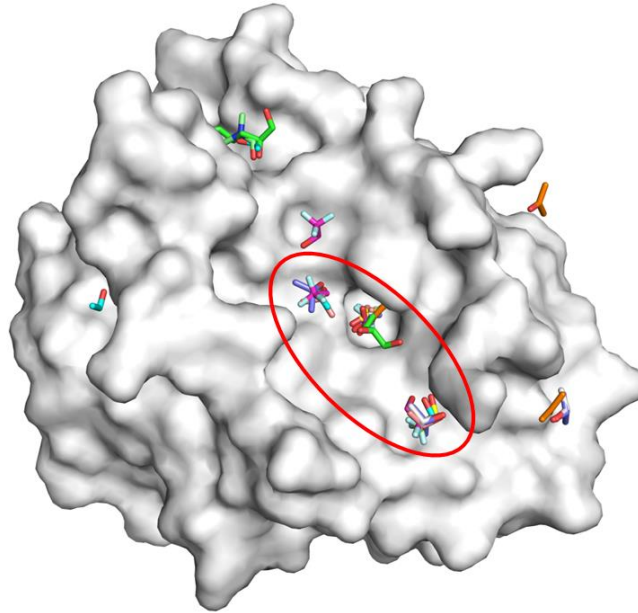


Figure 1-1. An overlay of the all the MSCS crystal structure of elastase is shown. The co-solvent molecules cluster within the active-site. The active-site is denoted with a red circle. The list of crystal structures includes 2FOE, 2FOD, 2FOG, 2FOH, 2FOF, 2FOA, 2FOB, 2FO9, 2FOC.

1.2 Computational approaches for reproducing MSCS

Several computational approaches have been developed, inspired by the MSCS technique. These can be broadly classified into two categories based on their reliance on static conformations or molecular dynamics (MD) simulations.

1.2.1 Static protein conformation based approaches

Static protein conformations are readily available from many sources including experimental (X-ray and NMR) and computational approaches (MD). Some of the first probe mapping techniques were performed using static protein conformations. All these approaches involve some form of probe minimization on the protein surface in vacuum. A few of the most important ones are described in further detail below.

1.2.1.1 Multiple Copy Simultaneous Search (MCSC)

In the MCSC method, the probes acetate, methanol, methyl ammonium, methane, and water are minimized onto the protein surface (8, 9). Several copies of the probes are dispersed in the site of interest on protein ranging from 1,000 – 5,000. Prior to minimization, probes with interaction energies greater than 5 kcal/mol are removed. The remaining probe molecules are then minimized independent of each other using the time dependent Hartree approximation (10). At regular intervals, a single copy of probe molecules converging to the same location is retained. Convergence is achieved during the minimization procedure typically after 3,000 – 6,000 steps. The remaining probes are then further sampled on a grid limited to the vicinity of the final minimized location of probes. This is achieved by fixing the center of mass of the probe at each grid point and exploring the rotational degrees of freedom. These results are then visualized as density maps. In the first application of MCSC, the sialic acid binding site of influenza virus hemagglutinin was examined. The probes molecules were found to satisfactorily map the binding sialic acid binding site. In a follow up study on HIV-1 protease, the technique was extended to include N-methyl acetamide (NMA) as an additional probe. Using the MSCS approach, favorable locations for NMA in the active site were used successfully to reconstruct the binding orientation of MVT-101, a peptide known to bind HIV-1 protease (11).

1.2.1.2 Multiple Protein Structures Based Receptor Pharmacophore models (MPS)

Our MPS method is an experimentally verified computational mapping approach for obtaining receptor based pharmacophore models (12–22). In this approach benzene, ethane, and methanol probe molecules are flooded onto the binding site of interest. These probe molecules are then minimized independent of each other using a Monte Carlo method called Multi Unit Search for Interacting Conformers implemented in the BOSS program (23). The minimized probe molecules are clustered to identify favorable interaction sites on the protein surface. These clusters are then converted into pharmacophore elements. Overlapping benzene and ethane clusters are converted to hydrophobic pharmacophore elements, benzene clusters are converted into aromatic pharmacophore elements. Donor, acceptor, and Doneptor

pharmacophore elements are obtained from methanol clusters. The size of the pharmacophore elements is reflected by the root mean square deviation of the elements in the cluster. Using the MPS method, the first receptor based pharmacophore model was derived for HIV-1 protease (13). Subsequently several optimization studies were undertaken to obtain robust pharmacophore models using structures from X-ray, NMR (17) and MD simulations (14, 15, 19). A common theme throughout MPS pharmacophore development was the positive impact of protein flexibility on MPS pharmacophore model performance. Additionally, MPS pharmacophore models were shown to exhibit species specificity for human and *Pneumocystis carinii* variants of Dihydro Folate Reductase (16). Several MPS pharmacophore models were experimentally validated by the identification of inhibitors from MPS pharmacophore screening. Small molecules that target the MDM2-p53 interaction were identified using MPS pharmacophore models (18). More recently, pharmacophore models created from an allosteric site on HIV-1 protease were experimentally verified to be active against drug resistant strains of HIV-1 protease (21, 22).

1.2.1.3 FTMAP

FTMAP developed by Sandor Vajda and co-workers is a mapping technique that samples billions of probe molecules on densely space grid. This mapping is performed using sixteen different probe molecules which include ethanol, isopropanol, isobutanol, acetone, acetaldehyde, dimethyl ether, cyclohexane, ethane, acetonitrile, urea, methylamine, phenol, benzaldehyde, benzene, acetamide, and N,N-dimethylformamide. Sampling of several copies of many different probe molecules is achieved by an energy function that is evaluated using a fast Fourier transform correlation approach. While techniques mentioned earlier include van der Waals and electrostatic interactions terms, FTMAP energy function includes cavity terms to reward hydrophobic enclosure and statistical knowledge-based pairwise potential to account for solvation effects. In a first application of FTMAP, the binding sites of the proteins elastase and renin were shown to be mapped by probe molecules. The FTMAP technique was also applied to the proteins DJ-1 and glucocerebrosidase (24). The binding sites identified by FTMAP were shown to be in agreement with subsequent MSCS solved for these proteins (24). Favorable

results were also found upon application of FTMAP to Ras GTPase (25) and Hen Egg-White Lysozyme (26).

1.2.2 Molecular dynamic simulation based approaches

MD-based approaches involve simulation of proteins with water and co-solvents, which allows one to take protein flexibility and competition with water into account while mapping the protein surface. Such techniques have the potential to present a cost effective and widely applicable alternative to MSCS. Several approaches that use MD-based methods are summarized in Table 1-1 and will be described in further detail. An important concept that is universal across these approaches is that maps of the protein surface are created using grids that display the count of co-solvents across the MD simulations. These occupancy maps identify binding sites by the grid points that most frequently contain the co-solvents (also called a high density of probes).

Table 1-1. A chronological list of cosolvent simulation techniques that have been used to identify binding sites on proteins is presented. The co-solvents and protein systems used in their studies are also shown.

Method	Co-solvents	Protein systems
Barril approach (MDmix) (27–30)	Isopropanol, ethanol, acetonitrile, methanol, acetamide	Thermolysin, p53 core domain, Porcine Pancreatic Elastase, MDM2, LFA-1/ICAM-1, PTP-1B, MAP kinase p38, Androgen receptor, Hen egg-white lysozyme, Heat shock protein 90 N-terminal domain, HIV-1 protease
MacKerell approach (SILCS) (31–40)	Benzene, propane, water (both hydrogen and oxygen patterns for water), acetonitrile,	Trypsin, α -thrombin, HIV-protease, FKBP12, Factor Xa, NadD, Ribonuclease A, cytokine IL-2, P38 MAP kinase, Dihydrofolate Reductase, Fibroblast Growth Factor Receptor 1 kinase,

	methanol, formamide, acetaldehyde, methyl-ammonium, acetate, imidazole	Adenosine deaminase, Estrogen Receptor- α Ligand-Binding Domain, AmpC β -lactamase, Androgen receptor, Peroxisome proliferator activated receptor- γ , Metabotropic glutamate receptor, β 2-Adrenergic receptor, BTB domain of BL6
Fersht Approach (41)	Isopropanol	p53-Y220C
Yang and Wang approach (42–45)	Isopropanol, phenol, trimethylamine N-oxide	Thermolysin, Bcl-xL, Mcl-1, interleukin-1 receptor
Carlson approach (MixMD) (46–48)	Acetonitrile, isopropanol, pyrimidine, N-methyl acetamide, acetate, methyl-ammonium	Hen egg-white lysozyme, Elastase, p53-core, RNase A, Thermolysin, Abl kinase, Androgen Receptor, Chk1kinase, FPPS, Glucokinase, Pdk1kinase, PTP1B
Cafilisch approach (49, 50)	Dimethylsulfoxide, methanol, ethanol	FKBP12, Bromodomains (BAZ2B and CREBBP)
GSK and Bahar approach (51)	Isopropanol, isopropylamine, acetic acid, acetamide	MDM2, PTP1B, LFA-1, Kinesin Eg5, p38 MAP kinase
Tan and Abell approach (52, 53)	Benzene, chlorobenzene	Polo-box domain (PBD) of polo-like kinase 1 (Plk1), MDM2, Interleukin-2, MCL-1, Bcl-xL
Gorfe approach (pMD) (54)	Isopropanol	K-ras

1.2.2.1 Barril Approach (MDmix)

The first cosolvent-based simulations for mapping protein surfaces were reported by Barril and co-workers (27). In this approach, simulations of isopropanol and water at concentrations of 20%v/v are run for at least 16ns. The approach was evaluated by its ability to reproduce the

locations of isopropanol molecules located in MSCS structures of Thermolysin (6), p53 core domain (55) and Porcine Pancreatic Elastase (4). The maps of the protein surfaces were broken down into separate occupancy grids for isopropanol's hydroxyl oxygen and methyl carbons. While it is reported that the densities from the isopropanol matched the location of isopropanol, there are several additional sites that are mapped on the protein surface that are not discussed, despite their likelihood of complicating prospective applications of the method. Also, the reasons for not comparing the density of the entire isopropanol probe with the location of isopropanol found in MSCS structures were not discussed. Using this approach, they calculated free energies with equation (1) where N_i and N_o are the bin counts at grid point i and the expected bin count in the absence of the protein. This is a measure of the free energy change for moving an atom from the bulk to grid point i . In their case, this atom could be the oxygen atom or the methyl groups of co-solvent isopropanol.

$$\Delta G_i = -T \ln \left(\frac{N_i}{N_o} \right) \quad (1)$$

When the "atomic" free energies were computed, they note that in some cases, the ΔG_{bind} exceeded the empirical limit of free energy of -1.5 kcal/mol on a per atom bases observed by Kuntz and Kollmann (56). Subsequent work from our group set the limit as -1.75 kcal/mol (57) and it is unclear if the values reported exceeded this limit as well. The authors propose that this behavior for isopropanol is a result of partial phase separation cause by apolar patches on the protein surface. Thus for isopropanol, the expected bin counts were rescaled so that the free energy values conformed to the limit of -1.5 kcal/mol. The maximal affinity of the probe molecules were then estimated using the principle that atoms in a drug-sized molecule are not only involved in establishing affinity, but also form a framework for allowing molecules to optimize such interactions. The authors note that probe molecules are under no such constraint and their free energies on a per atom basis could be much higher. As such, their maps could be used to establish an upper limit for the volume of drug-like molecules. In validating this concept, a comparison is made between the maximal limits established by their approach and examples of drug molecules with the most favorable free energies. Comparisons between

predicted and observed free energies were made for the protein targets MDM2, LFA—1/ICAM-1 complex, Protein Tyrosine Phosphatase 1B, MAP kinase p38, Androgen receptor.

In a follow up study, prompted by our work establishing the necessity of full protein flexibility for proper mapping of hotspots with co-solvent MD simulations(46), they examined the relationship between protein flexibility and its effect on binding free energy (29). They derived a logarithmic relationship between flexibility and its effect on ΔG_{bind} . They concluded that if the restrained protein has a preformed binding site, ΔG_{bind} would become more favorable as the entropic cost of restraining the protein had already been paid. However, if this was not the case, then clashes with the protein binding site would make ΔG_{bind} unfavorable.

More recently, they have moved to a setup where two co-solvent simulations are performed separately with 20% ethanol in water and 20% acetamide in water (30). An updated simulation protocol consists of 3 runs of 20ns while holding the heavy atoms in the protein restrained at a 0.01 kcal/mol.Å² potential. The method was validated on Heat Shock protein 90 N-terminal domain (Hsp90) and HIV-1 protease. Pharmacophore models created from ligands bound in crystal structures of these proteins were compared with the binding free energy maps calculated by equation (1). They observed that some key features in the HIV-protease pharmacophore model were not mapped and proposed to extend the technique by using other probes in the future. Furthermore, they state that using atoms within the probe molecule to define pharmacophore elements is limited by the assumption that these atoms behave independently of the probe molecules as a whole. We note that the authors no longer use the alcohol parameters from their first paper, which we suspect caused the difficulties with phase separation that they found (48).

1.2.2.2 MacKerell approach (Site-Identification by Ligand Competitive Saturation, SILCS)

SILCS is by far the technique that has made the most progress, expanding use of cosolvent simulations from only identifying binding sites to improvements such as pharmacophore modelling, free energy perturbation, and developing methods for sampling occluded pockets in proteins. The first study used a 1M benzene and propane solution on BCL-6 protein (31).

Benzene probes are used to identify aromatic interactions, propane molecules are used to identify aliphatic interactions, and water molecules are used to report upon the hydrogen-bond donating and accepting properties. Unique to the SILCS methodology, is the use of a dummy atom positioned at the geometric center of benzene and the central carbon of propane. A repulsive interaction term on these dummy atoms was necessary as the co-solvents under high concentration were found to aggregate. SILCS results were analyzed using simulation data generated from 10 runs of 5ns. Notably, a weak restraint of 0.01 kcal/mol.Å² is placed on the C α atoms during the simulations. Snapshots from the simulations are combined and visualized as density, described as “FragMaps”. Results from SILCS simulations of BCL-6 were verified by their ability to predict biologically relevant binding sites on the protein.

In a second study across a much wider set of examples (trypsin, α -thrombin, HIV protease, FKBP12, NadD, ribonuclease A), the length of SILCS simulations was increased to 20ns and the authors converted their FragMaps to Grid Free Energies (GFE) (32). These GFE were computed in a manner similar to the Barril approach, wherein equation (1) is used to report upon the free energy at each grid point. Using these GFEs, crystal ligand poses were found to score higher than decoy sets. Ligands were scored by assigning each atom in the ligand to one of aromatic, aliphatic, hydrogen-bond donor, and hydrogen-bond acceptor types. These atom types correspond to different probes used in SILCS simulations. After bringing the crystal ligands into the GFE frame of reference, the atom type of the ligand and its position within the grid were used to obtain the free energy value from the corresponding GFE grid. These values were then summed to arrive at the Ligand Grid Free Energy score for a given pose of the ligand.

In a follow up to this study, the authors assessed the use of free energy perturbation to expand the range of fragments that can be predicted to bind to proteins (34). Using benzene as an example, they first demonstrate that relative hydration energies for moving to mono-substituted benzene were correctly captured with an R² of 0.95. These benzene analogues were chosen based on experimental binding affinities that existed for ligands in α -thrombin and P38 MAP kinase. Then, a comparison was made between single-step free energy perturbations of benzene to its analogues with changes in experimental binding free energy that involved a

similar transition. It is exciting that promising results were obtained for α -thrombin, but the same could not be said for P38 MAP kinase. This highlights the inherent limitations of extrapolating FEP results from fragments to those found in drug-like ligand molecules.

SILCS simulations were also used to present an optimum solution for balancing target flexibility and possible denaturation in cosolvent-based simulations (33). Using various levels of positional restraints on Interleukin-2 (IL-2), the authors found that allowing for full protein flexibility resulted in denaturation of the protein in certain runs. In our application of MixMD, we have not seen such target denaturation, but these studies in no uncertain terms demonstrate that target denaturation should be considered a possibility when running co-solvent simulations and adequate inspection of the protein's behavior should be performed to detect them. In order to alleviate these concerns, recently we have moved from a 50% w/w concentration of co-solvent to a 5%v/v setup in MixMD. The authors in this SILCS study present two strategies to overcome unfolding problems, removing trajectories that denature or restraining the backbone of the protein while performing cosolvent simulations. It could be argued that the first option seems more appealing since restraining the protein will limit the breathing motions, thereby hampering the identification of cryptic pockets on the protein surface. In the case of IL2, this did not seem to be an issue, and these cryptic pockets were found even when using a restrained potential on the protein. It is interesting that the authors' simulation with acetonitrile at 50% w/w concentration did not map the binding site in IL2. In our application of MixMD across several allosteric systems, we also found that some binding sites were not mapped by acetonitrile.

In the SILCS Tier-II update (35), more co-solvents were introduced. In addition to benzene and propane, new probes included methanol, formamide, acetaldehyde, methyl ammonium, and acetate. All the aforementioned co-solvents were simulated in a single box of protein and water using a concentration 0.2 M for each probe. The simulations were performed for 20ns using a $0.12 \text{ kcal/mol.}\text{\AA}^2$ to prevent the unfolding of the protein in high concentration of probes, and repulsion terms were used to prevent aggregation. The densities of the co-solvents were combined in the following manner for analysis, generic polar (benzene and propane carbons),

generic neutral donor (methanol and formamide polar hydrogens), generic neutral (methanol, formamide and acetaldehyde oxygens), positive donor (methyl ammonium polar hydrogens), and negative acceptor (acetate oxygens). Using these combined atom grids, GFE values were computed. These were then contoured at various values for each grid type and compared by visual inspection of the overlap with example of ligands from crystal structures. The technique was validated using Factor Xa, P38 MAP kinase, RNase A, and HIV protease. In addition to a visual inspection, the authors developed a suite of scoring functions based around the LGFE scoring scheme that they used earlier. A Monte Carlo-based sampling of the ligands within the GFE grids gave the best correlation between the scores generated and the experimental binding affinity of the ligands. This approach worked for Factor Xa, P38 MAP kinase, and RNase A, but the values were anti-correlated for HIV protease. The authors note that this deviation of HIV protease behavior emphasized how measures of affinity obtained from GFE come from co-solvents and do not reflect the configurational entropy and strain in real ligands.

SILCS simulations have also been converted to pharmacophore models (36, 38). In their initial study, the pharmacophore models were derived from benzene, propane, and water locations from SILCS ternary simulations. The authors found that generating pharmacophore models from SILCS simulations using a GFE cutoff of -1.2 kcal/mol for aromatic/aliphatic FragMaps and -0.5 kcal/mol for water-based, hydrogen-bond donor/acceptor to be an ideal starting point. Grid points that were below the earlier mentioned GFE cutoffs were then clustered using a distance cutoff of 1 Å, 2.8 Å, and 2.6 Å for the water, aromatic, and aliphatic SILCS maps, respectively. These clusters are converted to “FragMap features” which are modelled as spheres whose center is the center of cluster. The radius of the FragMap feature is defined as the radius that encloses all the grid points that belong to this cluster. FragMap features were not allowed to have a radius greater than 2.5 Å for hydrophobic and 1.5 Å for hydrogen-bond features. The sum of the GFE within each cluster is then reported as the Feature Grid Free energy (FGFE) of the FragMap feature. In a subsequent step, the FragMap features are converted to pharmacophore elements. The most important considerations in generating pharmacophore elements was the use of overlapping FragMaps features for defining aromatic|aliphatic features and donor|acceptor pharmacophore elements. Overlapping clusters of aromatic and

aliphatic FragMap features are considered aromatic|aliphatic pharmacophore elements. Given that these simulations are conducted in a ternary system, it might be hard to establish overlapping features and this may be an area where simulations using a single co-solvent have an advantage. Using an automated approach, water locations of high density were converted to donor, acceptor, or donor|acceptor pharmacophore elements. These pharmacophore elements are then combined in different combinations and ranked using the cumulative FGFE of the elements in the pharmacophore model. That measure is called the hypothesis grid free energy (HGFE). The pharmacophore models using this approach were obtained from SILCS simulations of HIV-1 protease, Factor Xa, and dihydrofolate reductase. Pharmacophore models lowest HGFE values using 3 to 6 pharmacophores were selected for screening. These pharmacophore models were then screened against ligands and decoys from the directory of useful decoys (DUD) dataset (58). A hit was reported when all pharmacophore elements in the model matched features in the ligands using MOE (59). Furthermore, a comparison is made between results from pharmacophore screening to the docking programs Dock (60) and AutoDock (61). The authors note that the best performing SILCS pharmacophore model outperformed results from Dock and AutoDock. A comparison is also made with a receptor-based pharmacophore model technique based on hydration data (62), and the authors note the superior performance of their approach.

In a more elaborate study, the authors used SILCS Tier-II to obtain pharmacophore models (38). The primary advantage served by this approach was the use of co-solvents that allowed them to probe hydrogen-bond donating and accepting capabilities. This meant they could move away from using water to obtain such information. As the number of co-solvents expanded, the authors were able to add more pharmacophore element types to their repertoire. The additions included positive-donor and negative-acceptor pharmacophore elements. Also, excluded volumes were placed wherever grid points were not occupied by water or other co-solvents. In screening the pharmacophore models, all pharmacophore elements were used. However, certain pharmacophore elements were required for a match, which the authors describe as “key features”. In testing their pharmacophore models, these key features were selected after sorting all the pharmacophore elements based on the FGFE value. The authors

note that using 3 or 4 key features resulted in the best enrichment. When 5 or more key features were used, degradation in performance was observed. The effects of HGFE on model performance was also tested, wherein it was found that models that performed well for the most part had a low HGFE. The performance of the pharmacophore models using this approach was tested against the systems that were used in their earlier approach for pharmacophore models (HIV-1 protease, Factor Xa, and dihydrofolate reductase). Additional systems were also used to test SILCS pharmacophore models, including P38MAP kinase (P38 MAP), Fibroblast Growth Factor Receptor 1 kinase (FGFr1), Adenosine deaminase (ADA), Estrogen Receptor- α Ligand-Binding Domain (ER), and AmpC β -lactamase (AmpC). The data sets for evaluating the performance of pharmacophore models were obtained from the DUD dataset. Screening results for SILCS pharmacophore models were also compared with results from Dock, AutoDock, AutoDock Vina, Full Protein Pharmacophore, and Hydration Site Restricted Pharmacophore. In comparing across all the methods, the authors note that SILCS pharmacophore models outperformed other methods except the case of AmpC. For most of the proteins, an area under the curve of 0.7 was observed for ROC plots when SILCS pharmacophore models were screened. However, FGFr1 and P38 MAP kinase yielded values that were lower than 0.6. Interestingly, similar results were seen for these proteins with the other methods, suggesting that they were challenging targets for virtual screening.

Further advancements in the application of SILCS were made by the development of the Grand Canonical-like Monte Carlo (GCMC) approach coupled with molecular dynamics (MD) simulations (37). In this method, the excess chemical potential of water and solutes is varied to arrive at the target concentrations during the simulation process. The method in brief involves the simulated system being coupled to a reservoir of water and co-solvents. The water/co-solvent molecules from the reservoir are inserted/deleted from the reservoir into the system being simulated or translated and rotated if they are already present in the system. These moves are accepted or rejected based on Metropolis criteria, which depends on the change in energy upon the occurrence of the move, the target density, and excess chemical potential. Following several such moves (100,00 moves when used for simulating the protein), an MD simulation is performed. Finally, the excess chemical potential is changed. This change in excess

chemical potential is based on a function of the deviation of the current concentration from the target concentration of the species under consideration. The whole process described above is repeated several times till the excess chemical potential converges. This approach was validated by reproducing the hydration free energies of the co-solvent molecules used in SILCS-Tier II simulations. Following this validation, the authors investigated the use of the method to map the occluded binding site of T4 lysozyme L99A mutant (T4-L99A). Following the GCMC-MD procedure, the occluded binding site of T4-L99A was successfully mapped by SILCS simulations. Moreover, the LGFE values correlated with a R^2 of 0.72 to the experimental binding affinities for the different molecules that are known to bind within this occluded pocket.

The GCMC-MD approach was further applied to several systems with occluded ligand-binding pockets (40). These systems included androgen receptor (AR), peroxisome proliferator activated- γ (PPAR γ), metabotropic glutamate receptor (mGluR), and β 2-adrenergic receptor (β 2AR). The occluded binding sites in all the protein targets were successfully mapped during SILCS simulations. Furthermore, a SILCS pharmacophore model obtained from β 2AR was screened against a compound collection of 1.8 million from the CHEMBRIDGE and MAYBRIDGE libraries. Following an elaborate procedure of docking with AutoDock Vina (63) into the active and in-active conformations of β 2AR, molecules were identified that preferentially bound the active conformation. The hits were clustered and handpicked. Of the 16 molecules that were handpicked and tested, seven were found to be active. At this point, it is unclear if the molecules target the binding site of β 2AR, but this exciting result nevertheless points to the utility of co-solvent simulations in prospective structure based drug design.

1.2.2.3 Fersht Approach

Fersht and co-workers have used isopropanol-based co-solvent simulations to study a cancer causing mutant of the p53 protein (41). This mutant protein, named p53-Y220C, has a mutation of a tyrosine residue (present in the wild type p53) to a cysteine residue resulting in a pocket being opened. The authors investigated the use of co-solvent simulations to identify druggable binding sites on the protein surface. Interestingly, the site with the highest isopropanol density was located at the dimer interface. Two other sites were also found, one within the cavity

created by the mutation and another which the authors could not account for. The authors note that during their experimental fragment screen, they were only able to identify hits that targeted the mutation-induced cavity on p53-Y220C and could not find hits for the other two sites. The setup and execution of the isopropanol simulations was similar to the Barril approach where a concentration of 20% v/v was used. In the initial equilibration period, the protein was simulated at 600K while placing a restraint on the heavy atoms of the protein to allow for the distribution of probes. This was followed by an equilibration of 1ns followed by 19ns of production simulation under constant pressure at 300K. Binding free energies for the isopropanol molecules were also estimated using the Barril approach.

1.2.2.4 Yang and Wang approach

In their first use of co-solvent simulations, Yang and Wang compared the co-solvent locations in MSCS structures of Thermolysin. This important study was the first to compare free energies obtained from equation (1) with more rigorous statistical mechanics based approaches. For this study, the MSCS structure of Thermolysin with three different probes (isopropanol, phenol and acetone) was used. Their primary focus was on two isopropanol sites identified on Thermolysin named site 1 and site 2 that appeared at high concentrations and low (2 and 5% isopropanol), respectively. Site 2 was also mapped by phenol and acetone whereas site 1 was not. The double decoupling method (64, 65) was initially used to compute the free energies of site 1 and site 2. In applying this technique, they note that site 2 (-4.87 kcal/mol) had a more favorable free energy for binding isopropanol compared to site 1 (-3.25 kcal/mol); this observation was consistent with the identification of site 2 at a lower concentration of isopropanol. For site 1, they note the free energy changes for the different co-solvents ranged from -3.35 to -4.32 kcal/mol. These results from the double decoupling method were compared with the values obtained from performing and computing the free energies of isopropanol using the Barril approach. The binding free energy for isopropanol in site 1 and site 2 were found to be -3.91 and -5.01 kcal/mol. The authors note that the values computed using the co-solvent simulation was higher compared to the more rigorous double decoupling method, but both methods give free energies of binding within 1 kcal/mol for both sites, which is the basic limit of free energy

calculations. We consider the agreement in the methods much more interesting than differences within error of the techniques.

In a subsequent study, the authors compared simulations of the protein Bcl-xL in water and isopropanol (43). Starting from conformations obtained from one apo and three holo Bcl-xL crystal structures, 32ns simulations in water were shown to exhibit hydrophobic collapse which prevented Bcl-xL from adopting conformations that allowed it to bind to its partners. However, in the presence of isopropanol, conformations that resembled those used to bind with other partners were retained. Furthermore, the authors note that the hotspots identified on the protein surface changed based on the starting conformation used for co-solvent simulations of Bcl-xL. They suggest that such information in principle allows one to target different conformations separately. In continuation of their earlier work, co-solvent simulations using isopropanol (20% v/v), phenol (10% v/v), and 2M trimethylamine N-oxide (TMAO) were performed on Bcl-xL and Mcl-1 (44). In that study, the authors note that there were similarities and differences in the location of hotspots within both the proteins. Using this information, it was suggested that the differences in the location of hotspots within the active site between the two proteins could be exploited to obtain potent and selective drug-like molecules.

More recently, hotspots on the protein surface of the ectodomain of interleukin-1 receptor type 1 (IL-1R1) were investigated using co-solvent simulations of phenol (10% v/v) (45). The authors' primary motivation for using phenol co-solvent simulations came from the frequent observation of these groups in fragment screening libraries for targeting protein-protein interactions. Co-solvent simulations were used to investigate three druggable sites identified using Sitemap (66, 67), which were named P1, P2, and P3. As P1 and P3 could already be identified from crystal structures, they focused their attention on assessing the druggability of the P2 site using co-solvent simulations. While Sitemap identified four conformations in which the P2 site was deemed as druggable, co-solvent simulations identified only two conformations of the protein in which the P2 site exhibited high affinity for phenol co-solvent. These studies highlight the importance of including protein flexibility in assessing druggability of proteins. Based on this analysis, further efforts were focused on one of the two conformations that

adopted a novel conformation. Using *in silico* screening methodology, fragments that bound to the P2 site were identified and further simulations of these fragments revealed that when bound to the P2 site, these fragments restricted the conformations accessible to IL-1R1.

1.2.2.5 Carlson approach (MixMD)

Our approach for performing co-solvent simulations is called the mixed-solvent molecular dynamics (MixMD). MixMD involves binary solvent simulations of proteins with water and water-miscible, organic probe solvents. An emphasis on using water-miscible organics as co-solvent distinguishes our approach from other techniques. A first step in validating MixMD was evaluating its ability to reproduce the co-solvent binding location obtained from MSCS experiments. Using the acetonitrile binding site in Hen egg-white lysozyme (HEWL) as a test case, MixMD was shown to recapture the binding location of acetonitrile(46). Our first MixMD simulations used a 50% w/w concentration of acetonitrile and were run five times for 10 ns duration. The last 2ns of these simulations were used for obtaining the preferential location of acetonitrile binding on the protein surface. This work also noted the importance of protein flexibility on the accuracy of mapping the acetonitrile binding site. Using a series of MixMD simulations wherein the protein was subjected to varying levels of restraint, the acetonitrile binding site was mapped accurately without spurious minima only when full protein flexibility was allowed. When the protein was held rigid, we found that the acetonitrile binding site was mapped as strongly as other local minima. Interestingly, when the backbone of the protein was held rigid and the side chains were allowed to full explore different conformations, the local minima on the protein surface persisted, but the acetonitrile binding site was not mapped well. These studies certainly highlight the notion that the rugged landscape found when performing minimizations in vacuum are not an artifact of doing simulations in vacuum but result from using a rigid conformation.

In a follow up to the first MixMD study, we focused on extending the approach to protic solvents (47). Isopropanol was used as the co-solvent, and several proteins were used as test cases: elastase, HEWL, P53 core, RNase, and Thermolysin. MixMD results were shown to be in excellent agreement with the isopropanol binding sites found in MSCS of these proteins. During

the course of optimizing the technique, the number of runs and the simulation length were also investigated. The importance of multiple, short simulations was highlighted, and using 10 runs of 20ns was found to be optimal.

More recently, the importance of probe parameters was established by comparing co-solvent simulations using our approach and parameters for isopropanol to that of the Barril approach (48). These co-solvent simulations were performed on Thermolysin using a 50% w/w isopropanol. To our surprise, the co-solvents separated into two phases when using Barril's parameters. Our simulations based on OPLS parameters for alcohols remained evenly mixed. This result made us step back and evaluate water-cosolvent mixtures alone without proteins. We used radial distribution functions to monitor miscibility. We recommend that all co-solvent simulations included RDFs of the solvents to show proper behavior of the environment. This is just as important as monitoring protein's RMSD to show no unfolding. We investigated the use of several different organic probes for MixMD simulations. Upon testing eleven different solvents, six were found to have even mixing with TIP3P water. These co-solvents were acetonitrile, isopropanol, acetone, N-methyl acetamide, imidazole, and pyrimidine.

1.2.2.6 *Cafilisch approach*

Cafilisch and co-workers performed simulations of FK506 binding protein (FKBP) with dimethylsulfoxide (DMSO) (49). Ten simulations lasting for 70ns each using 50 molecules of DMSO (~440mM) were performed. DMSO primarily mapped the active site of FKBP in these simulations. The authors also note that using DMSO concentrations higher or lower by a factor of two did not change the obtained results. Interestingly, the binding and unbinding events of DMSO in these simulations were used to obtain the dissociation constant of DMSO for the active site. These values were in agreement with results from experiments.

In a follow up study, co-solvent simulations of the two bromodomains, zinc finger domain 2B (BAZ2B) and the binding protein of the cAMP response element binding protein (CERBBP), were performed (50). These simulations were conducted separately using the co-solvents DMSO, methanol, and ethanol. Two 0.5 μ s simulations for each co-solvent were performed using 50 co-

solvent molecules (~440mM). Co-solvent simulations were able to successfully map the acetyl-lysine binding site of CREBBP. Furthermore, the location of DMSO in these simulations was in perfect agreement with the position of DMSO found in a crystal structure of CREBBP (68). Similar mapping of the acetyl-lysine binding site by different co-solvents was also noted for BAZ2B. The authors note that there were several binding and unbinding events of the co-solvents observed in the simulation. An analysis of the kinetics of co-solvent binding revealed that unbinding events for DMSO and ethanol were slower than methanol possibly due to their larger size and hydrophobicity. Interestingly, an analysis of the water molecules within the acetyl-lysine binding revealed that while some were retained during the entire simulation, others were transiently replaced by co-solvents. Based on this information, the authors proposed the use of water molecules that do not exchange with co-solvent be used in high-throughput docking studies. Furthermore, they suggested that hydroxyl substituents could be designed into ligands when water molecules are replaced by co-solvents.

1.2.2.7 GSK and Bahar approach

Bahar and co-workers utilize co-solvent simulations to address the druggability of proteins (51). In this approach, two co-solvent simulations are conducted, one in the presence of isopropanol, and another using a mixture of acetamide, acetic acid, and isopropylamine. The ratio of probes to water was set at one probe molecule for every 20 water molecules. This corresponds to ~2.3M probe concentration in the co-solvent simulations. Several simulations of varying time length of 32 and 40ns were performed. Free energies were calculated using equation (1). However, it is important to note that the free energies were calculated based on the maximally occupied grid point in the volume of an entire probe. This is a very important distinction from other approaches where these measures are reported on a per-atom basis. These free energies calculated for volumes of the size of a probe were termed “interaction spots”. Reasonable constraints were placed on the definition of these interaction spots. They were required to not overlap with other interaction spots. Only those interaction spots with energy lower than -1 kcal/mol were considered, and the energy of an interaction spot was determined to be that of the central grid point (all other grid points within the radius of the probe were eliminated). In

co-solvent simulations using mixtures, the radius of the interaction site was the sum of the radii of all the probes used in the simulation. An interaction spot was given a charge based on the fraction of time it is occupied by a charged probe. The interaction spots were then clustered using a 6.2Å distance to identify druggable sites under the constraint that the clusters can have a charge of no more than $2e^-$. Finally, maximum achievable free energies of binding were obtained from the free energies of the interaction spots within the clusters.

These co-solvent simulations were applied to a test set of five proteins, which included murine double mutant-2 (MDM2), protein tyrosine phosphatase 1B (PTP1B), lymphocyte function-associated antigen 1 (LFA-1), vertebrate kinesin-5 (Eg5), and p38 mitogen-activated protein kinase. The authors found that the maximal free energies of binding computed using their approach are in perfect agreement with the affinities of the best known ligands for the binding sites on these proteins. Interestingly in MDM2, the occluded binding site was opened for access only in co-solvent simulations. Similar results were obtained for LFA-1 and Eg5 where rearrangement of side chains resulted in access to the allosteric site. The authors attribute the opening of partially occluded sites to the use of an annealing procedure during the equilibration protocol wherein the system was heated to 600K under a restraint placed on the heavy atoms to prevent unfolding. Furthermore, in comparing the water and co-solvent simulations, it was noted that the probe molecules prevented hydrophobic collapse of binding sites during the equilibration. In Eg5, the pocket opening happened more frequently when a mixture of polar and charged co-solvents were used instead of isopropanol. In P38 MAP kinase, the druggability of the allosteric site was better captured by a mixture of probes instead of the use of isopropanol alone. These results certainly highlight the advantages of using probe mixtures over single co-solvent simulations. The authors note that many drug molecules are either charged or zwitterionic in nature, so mixtures of probes that include charged co-solvents may be required for many druggable proteins.

1.2.2.8 *Tan and Abell approach*

Tan, Abell, and co-workers present an alternative approach for performing co-solvent simulations using a low concentration of benzene (0.2M) (52). In an application of this method

to the polo-box domain (PBD) of polo-like kinase 1, they note that a tyrosine residue lining the secondary binding site of this protein adopts a closed conformation during water simulations. However, when co-solvent simulations were performed with benzene, this residue flipped to open a cryptic pocket. Furthermore, a ligand was successfully designed to take advantage of this cryptic binding site. These studies highlight the potential of co-solvent simulations to open cryptic pockets on the protein surface that can then be targeted through SBDD.

More recently, the authors were motivated by the abundance of halogens in drug-like molecules to focus on the use of chlorobenzene as a co-solvent (53). They note that chlorobenzene aggregates when used at a concentration of 0.2M and thus decreased the concentration of the probe molecules to 0.15M. This decrease in chlorobenzene concentration necessitated an increase the simulation length from 5ns to 10ns to achieve adequate sampling. Protein targets with halogenated ligands were selected to test the approach. This set of test cases included MDM2, Mcl-1, interleukin-2, and Bcl-xL. In starting their co-solvent simulations, they chose conformations of the protein where these halogen binding sites were absent. For the most part, simulations were able to identify cryptic binding sites on the protein surface. The authors note that the only site not mapped by co-solvent simulations was in Bcl-xL, but opening that site required major rearrangement of helices.

1.2.2.9 Gorfe approach (pMD)

Gorfe and co-workers have investigated the location of hotspots on the protein surface of K-ras using pMD, an approach that uses isopropanol based co-solvent simulations (54). In their approach a simulated annealing procedure was used similar to the one reported by Bahar and GSK collaborators. Here, the system was initially restrained using a 4 kcal/mol/Å² heavy-atom restraint and heated to 650K followed by cooling to 310K. This procedure in their opinion prevented kinetic trapping of the probe molecules inside the protein. Following further equilibration wherein the restraints on the protein were gradually removed, the system was simulated for three runs for varying lengths of time ranging from 30 to 100 ns. Further analysis was then performed by combining the three runs. The results were visualized by converting each grid point to free energy values using equation (1). The maps were then subsequently

contoured at -0.5 kcal/mol for visualization. In an approach similar to the one adopted by Bahar and co-workers, maximal free energies were calculated for binding sites. The grid point with the most favorable free energy was identified, and all other points within a 5Å radius were discarded. After exhaustively processing the grid points in this manner, the retained points were clustered using a 6Å clustering distance. “Druggable sites” were defined as clusters with four or more interaction points and “subsites” were defined as clusters with two or three interaction points. Five druggable sites and three subsites were identified on K-ras. These sites were then found to capture known allosteric sites on K-ras. An additional comparison was made between pockets identified using the curvature analysis MDpocket (69) and pMD simulation maps. The authors note that some of the sites were not identified by MDpocket as they did not conform to the definition of a pocket. Thereby, the authors point to the advantage of using co-solvent maps to identify binding sites as opposed to those obtained from techniques such as MDpocket that rely on protein curvature (69). A comparison was also made between water simulations and pMD using MDpocket, wherein they found that pockets formed during pMD simulations were larger in size.

1.3 Overview of thesis

The major areas addressed in this thesis include method development and application of MPS and MixMD techniques developed in the Carlson lab.

Chapter 2 describes our application of the MPS method to several protein targets with X-ray and NMR ensembles. The MPS pharmacophore models from both ensembles are compared and contrasted. Across all protein targets, NMR pharmacophore models are shown to outperform X-ray models. Reasons for the superior performance of NMR models and the observation of several extraneous pharmacophore elements in the X-ray pharmacophore models are discussed.

Chapter 3 presents a protocol for identifying binding sites from MixMD simulations using acetonitrile, isopropanol, and pyrimidine. Using an extensive test set of allosteric protein targets, MixMD is shown to successfully recapture the competitive and allosteric sites. It is

notable that such success was achieved starting from protein conformations with no allosteric ligands bound. Chapter 4 then describes our approach for deriving free energies of binding and entropies from MixMD simulations. Drawing upon concepts from Statistical Mechanics, methods for calculating these important thermodynamic measures are put forth and validated. Careful consideration is given to the limitations of free energy calculations from co-solvent simulations. This study highlights the pitfalls of using co-solvent simulations to establish maximal affinities of ligands for binding sites that are not described in similar techniques put forth by other groups. Finally, Chapter 5 describes the first successful prospective application of a co-solvent simulation technique. Using Heat Shock Protein 27 (Hsp27) as a test case, binding sites identified by MixMD are shown to bind drug-like molecules or represent sites of biological relevance. Furthermore, a direct comparison between MixMD results and NMR chemical shift data for Hsp27 at 2% v/v concentration demonstrates a high level of agreement between theory and experiments. The method development and application of MixMD across Chapter 3 – Chapter 5 describe the successful application of MixMD for detecting binding site culminating in the first ever blinded application of a co-solvent simulation technique in a prospective manner on Hsp27.

Chapter 2. Comparing pharmacophore models derived from X-ray and NMR ensembles

2.1 Abstract

NMR and X-ray crystallography are the two most widely used methods for determining protein structures. Our previous study examining NMR vs X-Ray sources of protein conformations showed improved performance with NMR structures when used in our Multiple Protein Structures (MPS) method for receptor-based pharmacophores (17). However, that work was based on a single test case, HIV-1 protease, because of the rich data available for that system. New data for more systems are available now, which calls for further examination of the effect of different sources of protein conformations. The MPS technique was applied to Growth factor receptor bound protein 2 (Grb2), Src SH2 homology domain (Src-SH2), FK506-binding protein 1A (FKBP12), and Peroxisome proliferator-activated receptor- γ (PPAR- γ). Pharmacophore models from both crystal and NMR ensembles were able to discriminate between high-affinity, low-affinity, and decoy molecules. As we found in our original study, crystal pharmacophore models had more pharmacophore elements compared to their NMR counterparts. The crystal-based models exhibited optimum performance only when pharmacophore elements were dropped. In addition to the comparison between NMR and X-Ray pharmacophore models, we note that X-ray pharmacophore models retain performance when pharmacophore elements at the periphery are eliminated using a cutoff-based approach. These studies suggest that additional pharmacophore elements seen at the periphery in X-ray models arise as a result of decreased protein flexibility and make very little contribution to model performance.

2.2 Introduction

Experimental techniques such as crystallography and NMR provide a window into the microscopic world, allowing one to observe the many varied conformations that proteins sample in atomic detail. While proteins are recognized to be inherently flexible and sample different conformations, computational techniques have yet to fully exploit this information in structure-based drug design (SBDD). Several approaches have been designed to account for protein flexibility in drug discovery, based on experimental methods such as crystallography and NMR or computational approaches such as molecular dynamics (MD) simulations (70, 71). While MD simulations can accommodate these requirements, it is resource intensive and time consuming (14, 15).

In this study, we compare ensembles of protein conformations from crystal and NMR structures, which were readily available. Our Multiple Protein Structures (MPS) method for creating receptor-based pharmacophore models is an experimentally verified, computational technique that leverages ensembles of protein conformations. The use of many protein conformations reveals areas of the binding site that have consistent criteria for complementarity and cause the least entropic penalty (12, 13). Each conformation of the protein binding site is mapped to determine the essential pharmacophore elements required to complement the pocket. MPS then overlays all the structures of the ensemble to identify pharmacophore sites that are common to more than 50% of the structures. This consensus of pharmacophore sites describes the essential elements that a ligand must contain to bind the target.

One of our previous studies compared the performance of pharmacophore models of HIV-1 protease derived from a collection of crystal structures and an NMR ensemble, using the MPS technique (17). For that system, the pharmacophore models from the NMR ensemble encoded a more accurate representation of the essential features of the active site while maintaining selectivity for inhibitors over decoy molecules. This was a direct consequence of the greater flexibility observed in the NMR ensemble over the collection of crystal structures. HIV-1

protease is more flexible than most protein targets, and it is important to determine how universal this finding may be. In this study, we extended the MPS technique to several new protein targets. Again, we find that incorporating the greater protein flexibility of an NMR ensemble in the MPS method translates into an improvement in the quality and performance of pharmacophore models over those created with collections of crystal structures.

There are very few systems with both NMR and crystal structures available in the Protein Data Bank (72). Even fewer are of biomedical interest so that databases of known inhibitors can be generated to test method performance. Growth factor receptor bound protein 2 (Grb2), Src SH2 homology domain (Src-SH2), FK506-binding protein 1A (FKBP12), and Peroxisome proliferator-activated receptor γ (PPAR- γ) protein targets met all the required criteria. We used ligand-bound crystal structures and NMR models whenever possible in order to ensure a fair comparison. However, due to the lack of such structures for FKBP12 and PPAR- γ , apo NMR structures were used for these particular proteins. Here, we demonstrate that the lesser protein conformational sampling seen in the crystal-structure ensemble leads to the identification of non-essential pharmacophore elements. In order to further probe the location and origin of these extraneous pharmacophore elements, the crystal pharmacophore models were systematically truncated. This resulted in retention of crystal pharmacophore model performance in most of the target cases studied. In the following sections, we discuss reasons for the retention in performance upon truncation of crystal pharmacophore models and present results across all the protein targets examining the ability of NMR and crystal pharmacophore models to identify inhibitors/agonists over decoy molecules.

2.3 Methods

2.3.1 Protein Preparation.

In order to ensure a uniform setup across NMR and crystal structures, all protein-ligand complexes were stripped of hydrogen atoms which were added later using Molprobit (73). Additionally, structures were manually assigned histidine protonation state and histidine/asparagine/glutamine flips. These protein-ligand complexes were then visually

inspected and corrected where necessary for errors in hydrogen placement and ligand bond orders, followed by partial charge assignment based on MMFF94 force field (74) for the ligand and AMBER ff99 force field for the rest of the protein as implemented in Molecular Operating Environment (MOE 2010.10) (59). The hydrogen atoms were then minimized while the heavy atoms were restrained.

2.3.2 Probe Flooding, Minimization, and Clustering.

The protein structures extracted from the minimized protein-ligand complexes were then flooded separately with 500 molecules of benzene, ethane, and methanol probes with a flooding radius of 10 Å from the center of the active site using PyMOL (Figure 2-1A) (75). The atom used to define the center of the active site for flooding in each protein target is provided in the supplemental information. The probe molecules were then subjected to a low temperature, gas-phase minimization using the Multi Unit Search for Interacting Conformers (MUSIC) routine in BOSS (23). This type of minimization keeps the protein fixed and does not allow the probe molecules to interact with one another. The minimized probes in each individual protein structure were then grouped by a Jarvis-Patrick clustering method, and the group was represented by the probe with the best interaction energy (the “parent” of the cluster). (Figure 2-1C) (20). This process was repeated for all the structures in the crystal and NMR ensembles separately. Structures from each ensemble were then aligned using our weighted RMSD method (wRMSD) (Figure 2-1D) (76). The parent probes for each structure in the ensemble were then compared across the ensemble to identify consensus clusters that represent conserved interactions in more than 50% of all protein conformations/models (Figure 2-1E). Pharmacophore elements were centered at the average coordinates for the corresponding parent probes. The RMS deviation of the parent probes in each consensus cluster formed the radii of the corresponding pharmacophore elements. Using this approach, Donor, Acceptor, and Doneptor (both donor and acceptor) pharmacophore elements were derived from Methanol clusters. Aromatic pharmacophore elements were derived from Benzene clusters, and hydrophobic pharmacophore elements were created from overlapping

Benzene and Ethane clusters. Any pharmacophore element lying outside the 10 Å flooding sphere was not included in the final pharmacophore model.

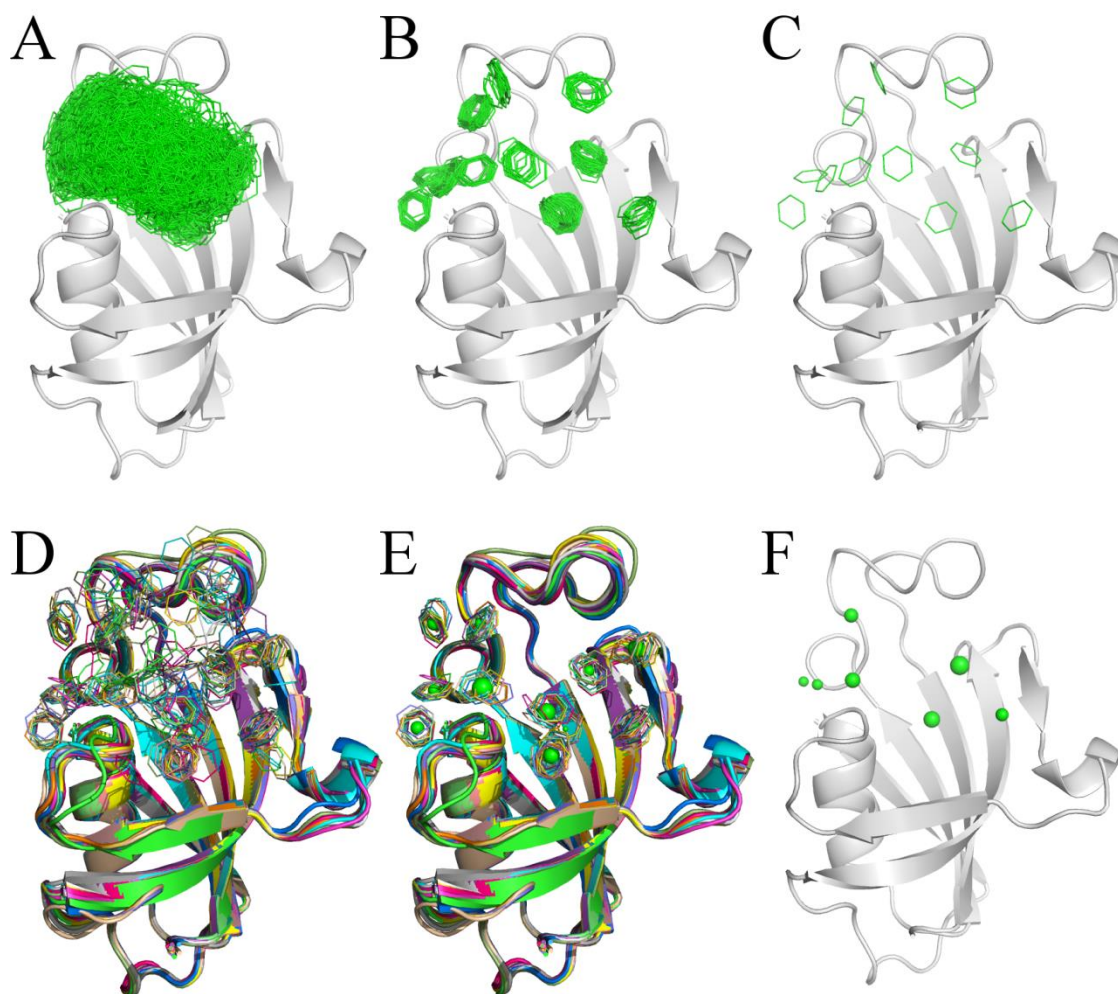


Figure 2-1. A detailed description of the MPS method is presented using benzene probes as an example. A) The protein active site is flooded with 500 probe molecules. B) These molecules are minimized, independent of each other using the MUSIC routine in BOSS. C) The probe molecules are clustered and represented by “parent probes” which are the single benzene molecule with the best interaction energy in each cluster in B. D) Steps A-C are carried out on all structures in the ensemble, and they are overlaid using the wRMSD method. E) Clusters of “parent probes” are identified manually. F) Consensus clusters are identified when at least 50% of the conformations contain a parent probe in the same location. All probes farther than 10Å of the center of the active site are ignored. The center of each pharmacophore element is

derived from the center of mass of the parents in the consensus cluster, and the radius of the element is set by the RMSD of the parents.

2.3.3 Creation of ligand and decoy databases.

Databases of inhibitors for the four protein targets were gleaned from the ChEMBL Database (77). $IC_{50} \leq 50$ nM was the cutoff between high-affinity and low-affinity inhibitors. As a sufficient number of high-affinity inhibitors with $IC_{50} \leq 50$ nM could not be found for Src SH2, the cutoff for high-affinity inhibitors was relaxed to $IC_{50} \leq 500$ nM (only for Src SH2). In order to ensure that the compounds in the inhibitor databases were structurally diverse, they were grouped by 85% Tanimoto similarity, calculated using the MACCS fingerprint (78) in MOE 2010.10 (59). The inhibitors with the highest affinity in each set of similar structures were retained. A previously reported decoy database of 2324 drug like molecules obtained from the Comprehensive Medicinal Chemistry Index (79) was used as the decoy set (16). To rule out the possibility of finding inhibitors in this decoy set that bind to protein targets under consideration, a similarity search was performed for each of the 2324 decoy molecules against the ChEMBL-derived database using a path based similarity fingerprint implemented in OEGraphSim toolkit (80). Using a Tanimoto coefficient of $\geq 85\%$ as a measure of similarity, if a ChEMBL molecule similar to the decoy molecules existed, its activity profile was searched in the ChEMBL database and any reported activity against the pfam family of the target under consideration resulted in the removal of the compound from the decoy set. This allowed us to establish a decoy set of molecules for each protein that is unlikely to contain any true positives. OMEGA (81) was then used to generate multiple conformations for the inhibitors and decoys, limiting the number of conformations to 300 while imposing a heavy-atoms RMSD constraint of 2 Å for rejection of similar conformers along with a 25 kcal/mol energy cutoff. The number of molecules in the high-affinity, low-affinity, and decoy sets used for each protein target are provided in Table 2-1.

Table 2-1. The number of high-affinity and low-affinity inhibitors/agonists used to validate MPS pharmacophore models are reported below for every protein target.

Protein	Number of High-Affinity Inhibitors/Agonists	Number of Low-Affinity Inhibitors/Agonists	Decoys
Grb2	61	97	2303
Src-SH2	16	145	2303
FKBP12	78	96	2324
PPAR-γ	54	119	2149

2.3.4 Evaluation of Pharmacophore Models.

Performance of the pharmacophore models were evaluated using Receiver Operator Characteristic (ROC) plots. For pharmacophore models, each ligand is evaluated with simple fit/no fit criteria, rather than a score. With scores, ROC plots are generated by plotting more and more relaxed scores (Figure 2-2). In our use of MPS, relaxation is accomplished by two means: 1) systematically increasing the radii of the pharmacophore elements and 2) varying the number of pharmacophore elements that need to be matched to identify a molecule as a hit. Each line on the ROC plot corresponds to results obtained from a model as the radii of its elements are systematically increased from 1 \times to 3 \times RMSD (see Figure 2-2). Each line shows hits for ligands required to match all N features of the pharmacophore model, N-1 matches, N-2 matches, etc. To illustrate this point, the label “5/6 sites, 2.66 \times RMSD” indicates hits fit five of a six-feature pharmacophore model with radii set to 2.66 times the RMSD. Using this approach, each pharmacophore model under varying degrees of relaxation was then screened against inhibitor and decoy databases using MOE 2010.10. As the database was split into high- and low-affinity data sets, this allowed us to observe if pharmacophore models exhibited selective preference for high-affinity over low-affinity inhibitors. The number of hits in the inhibitor and decoy sets were then plotted in a ROC plot. Ideally, one would expect the best results to lie in

the upper left quadrant of the ROC plot, indicating that the pharmacophore models exhibit greater selectivity for inhibitors/agonists over decoys. The distance of all the data points (distance = $\sqrt{(\%falsepositives - 0)^2 + (\%truepositives - 100)^2}$) on the ROC plot were calculated from (0,100) which represented the ideal case scenario where all inhibitors are identified and no decoys are detected by the pharmacophore model, the pharmacophore model with the least distance from (0,100) was considered to have the best performance in screening the ligand sets.

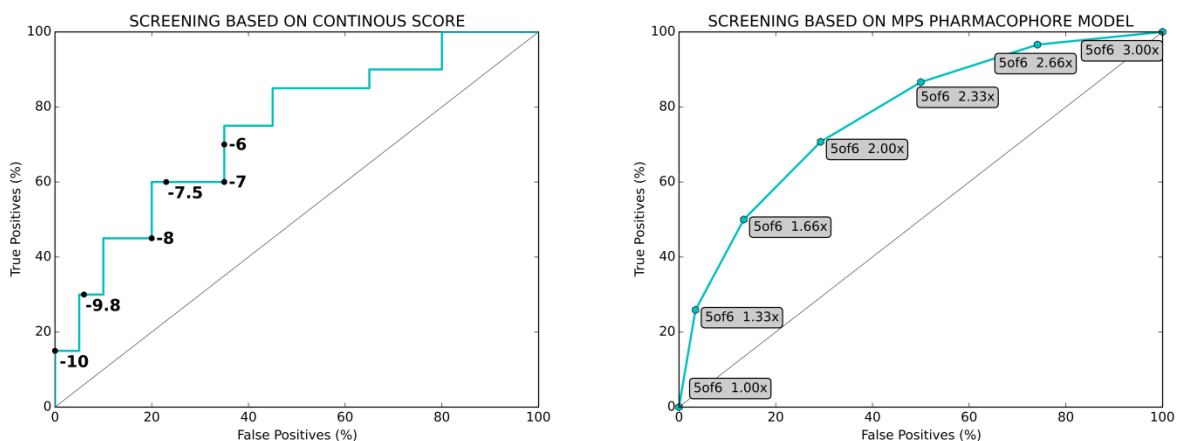


Figure 2-2. Receiver Operator Characteristic plots are shown for two cases, the first one corresponds to an exaggerated case of the more traditional use of ROC plots for continuous scores such as those obtained from docking results. The second plot illustrates the ROC plots generated by screening MPS pharmacophore models. Each discrete point on this line corresponds to the pharmacophore screening results obtained by gradually increasing the radii of the pharmacophore elements from 1× to 3× RMSD. Thus a label “5/6 2.33×” denotes the screening results from a MPS pharmacophore model whose radii have been multiplied by 2.33 and requires 5 of its 6 pharmacophore elements to be matched for a hit to be identified.

2.4 Results and Discussion

2.4.1 Protein conformational sampling in crystal structures and NMR ensembles.

As expected, crystal structures had limited conformational variation in their C_{α} backbone as compared to the NMR structures. The C_{α} RMSD of the crystal structures and NMR ensembles

for each protein was wRMSD aligned to the highest resolution crystal structure and the first model from the NMR ensemble, respectively. The results are summarized in Table 2-2 and illustrate the greater flexibility in the NMR ensembles. The greater flexibility in the NMR structures is also reflected in the heavy atom RMSD of the binding-site residues in the NMR ensemble as seen in Table 2-3. The NMR ensembles appear to be derived from high quality data, and we believe the conformational sampling exhibited in the NMR ensemble represents true sampling of these proteins in solution. While there were no unusually large conformational changes between the NMR and crystal structures of Grb2 SH2, Src SH2, and FKBP12, the differences in the conformations of PPAR- γ NMR and crystal structures represented a more important conformational change from the perspective of identifying MPS pharmacophore models. The PPAR- γ NMR structures represented an inactive conformation of the protein, while the crystal structures represented an active conformation of the protein.

Table 2-2. The range of pair-wise C_{α} RMSDs for all the crystal and NMR structures reflects the greater flexibility in the NMR ensemble.

Protein	Crystal Structures	NMR Ensemble
	C_{α}RMSD	C_{α}RMSD
Grb2	0.22 – 0.68 Å	0.74 – 2.32 Å
Src SH2	0.16 – 0.91 Å	0.82 – 1.68 Å
FKBP12	0.51 – 0.98 Å	0.95 – 1.43 Å
PPAR-γ	0.69 – 2.43 Å	1.36 – 2.8 Å

Table 2-3. The range of heavy-atom RMSDs in the binding site for the crystal structures and NMR ensembles provides support for the greater flexibility of the NMR ensemble.

Protein	Crystal Structures	NMR Ensemble
	Binding Site RMSD	Binding Site RMSD
Grb2	0.33 - 1.71 Å	1.02 - 2.08 Å
Src SH2	0.29 - 1.18 Å	1.00 - 1.79 Å
FKBP12	0.64 - 2.57 Å	1.58 - 3.95 Å
PPAR-γ	1.37 - 2.03 Å	1.49 - 3.21 Å

2.4.2 Comparison of crystal and NMR pharmacophore models.

In general, the pharmacophore models had certain characteristic traits that were common across all protein targets investigated in this study. Pharmacophore models from NMR structures had fewer pharmacophore elements which were greater in size, as compared to their crystal-pharmacophore counterparts.

2.4.2.1 Comparison of the Src SH2 pharmacophore models

Src SH2 is an important component in the auto regulation of its kinase domain; upon phosphorylation the C-terminus of the protein binds to the SH2 domain and results in the distortion of the kinase active site (82). The SH2 domain of Src binds with high affinity to phosphorylated peptides and recognizes the peptide sequence pYEEI with high affinity (83). The phosphotyrosine moiety of peptide ligands bind to the pY site, and the pY+1, pY+2, pY+3 sub-sites that determine specificity lie C terminal to this phosphotyrosine binding site (Figure 2-3A, Figure 2-3B). Pharmacophore models from both crystal and NMR structures reproduced key essential features in the different sub-sites required for binding substrates. As shown in Figure 2-3B, the NMR pharmacophore model for Src SH2 had six pharmacophore elements compared to the ten pharmacophore elements in the crystal model (Figure 2-3A). The pY site in the

crystal pharmacophore model had two extra elements, a donor and a doneptor that were absent in the NMR model. Additionally, an extra hydrophobic element in the pY+2 and an aromatic element in the pY+3 pockets were located in the crystal pharmacophore model. The elements in the NMR pharmacophore model represented a subset of those seen in the crystal pharmacophore model. Interestingly, while the exact location of the phosphotyrosine moiety was not mapped in the NMR and crystal pharmacophore models, a doneptor element in close proximity was identified whose location appeared shifted between the two pharmacophore models. The lone hydrogen-bonding element in the pY+3 pocket changed from a hydrogen-bond donor in the crystal pharmacophore model to a hydrogen-bond acceptor in the NMR pharmacophore model. This resulted from differing positions of a tyrosine residue lining the pY+3 pocket. Overlaying the crystal-structure ligands with the pharmacophore models as shown in Figure 2-3C emphasizes the observation that most elements from the NMR model overlap with the ligands in contrast to the crystal model where many elements failed to do so. This is particularly surprising because the ligands in Figure 2-3D are from the crystal structures, not the NMR model.

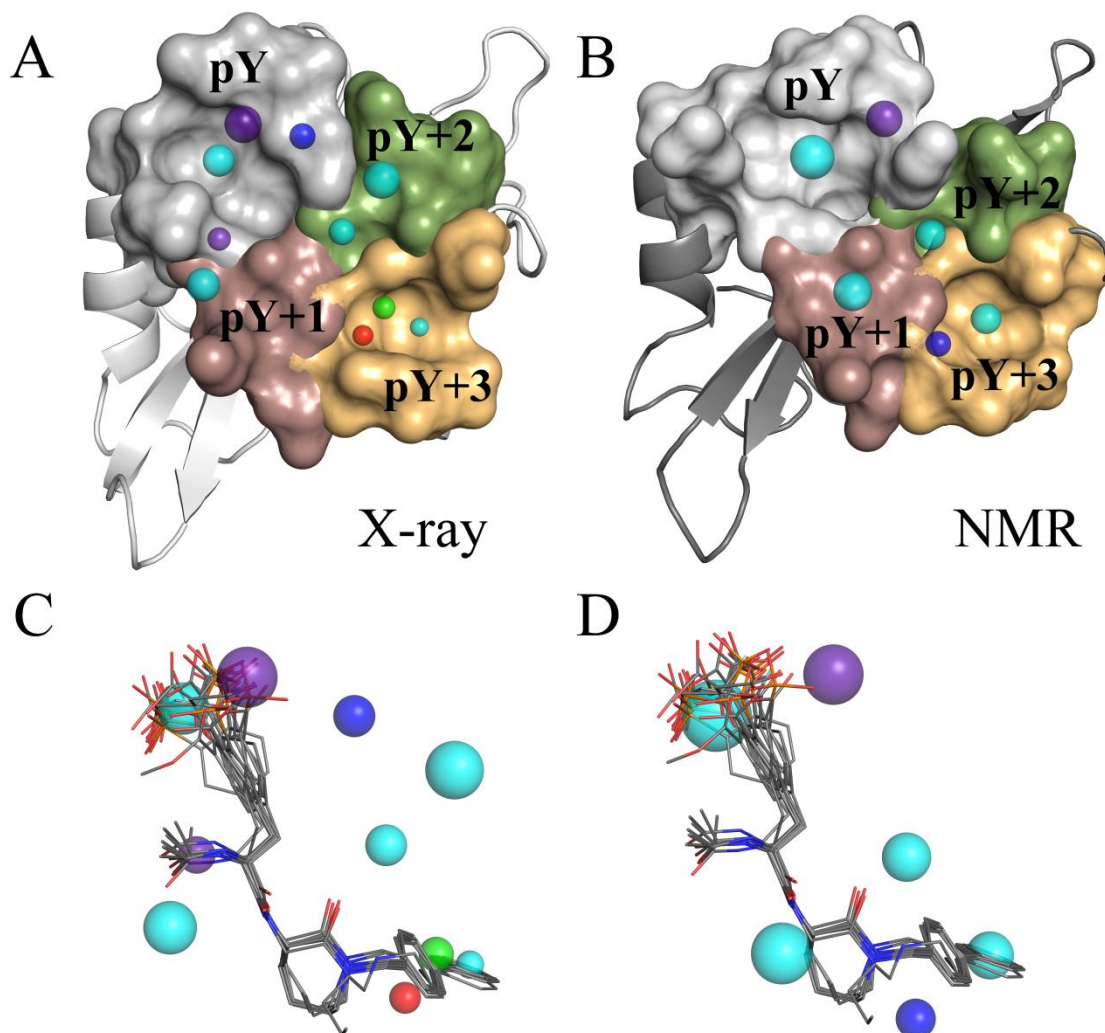


Figure 2-3. MPS pharmacophores are shown with $1\times$ RMSD radii, which indicates tighter or looser position constraints. Pharmacophore models for all the protein targets are color coded to represent different interactions: Red – Donor, Blue – Acceptor, Purple – Doneptor, Green – Aromatic, and Cyan – Hydrophobic. A) The MPS pharmacophore model for Src SH2 derived from X-ray structures. B) The MPS pharmacophore model for Src SH2 derived from the NMR ensemble. C) The ligands from X-ray structures overlaid on top of the Src SH2 X-ray model. D) The ligands from X-ray structures overlaid on top of the Src SH2 NMR model.

2.4.2.2 Comparison of the Grb2 SH2 pharmacophore models

Grb2 is an adaptor protein and consists of two SH3 domains and one SH2 domain. The SH2 domain of Grb2 binds to phosphorylated peptides of the general sequence pYXNX and adopts a

similar fold seen in other SH2 domains such as Src SH2. Sub-sites in the protein active site that accommodate ligands follow a similar nomenclature as noted above for Src SH2 and are named pY, pY+1, pY+2 (shown in Figure 2-4A and Figure 2-4B). A key difference between the Grb2 and Src SH2 domains is Trp 121 in Grb2 which is part of the specificity determining EF loop (according to the naming convention described in Eck et al. (84)) that blocks the large pY+3 pocket seen in Src SH2. As a result, phosphotyrosine peptides that bind to Grb2 SH2 domain adopt a beta turn instead of binding in an extended conformation occupying the pY+3 sub-site in Src SH2 (85). MPS pharmacophore models in this study were obtained for the SH2 domain of Grb2. Pharmacophore models reproduced key features of the active site which included the phosphotyrosine binding location in the pY subsite and essential interactions seen across all ligands in the pY+1 pocket (shown in Figure 2-4A and Figure 2-4B). A hydrophobic element that overlapped the benzene ring of the phosphotyrosine residue was seen in the pY subsite of both pharmacophore models (Figure 2-4C and Figure 2-4D). The pY subsite of the crystal pharmacophore model displayed an additional acceptor element not found in the NMR model. As seen in Figure 2-4C, this acceptor element overlaps with the carbonyl group of the amide bond linking the phosphotyrosine residue to the residue preceding it. Three pharmacophore elements in the pY+1 pocket appear in similar locations in the crystal and NMR models. The only difference between the two pharmacophore models in the pY+1 pocket was an additional doneptor element in the NMR model. The pY+2 sub site followed a similar trend and had more elements in the crystal pharmacophore model. A doneptor and hydrophobic element consistent with a key interaction is seen in both NMR and Crystal pharmacophore model (see Figure 2-4C and Figure 2-4D).

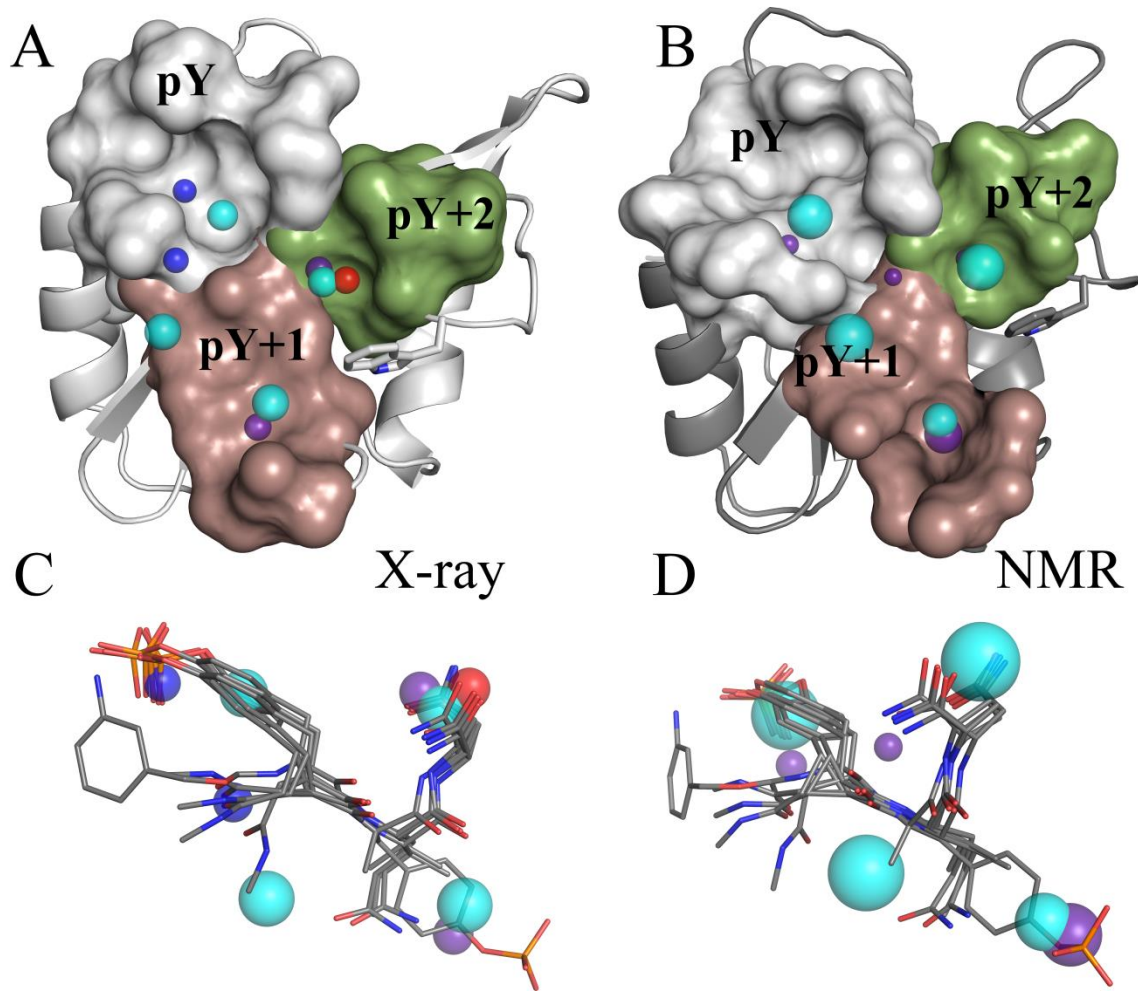


Figure 2-4. Coloring and radii of the pharmacophore elements are the same as in Figure 2-3. A) The MPS pharmacophore model for GRB2 SH2 derived from X-ray structures. B) The MPS pharmacophore model for GRB2 SH2 derived from the NMR ensemble. C) The ligands from X-ray structures overlaid on top of the GRB2 SH2 MPS X-ray pharmacophore model. D) The ligands from X-ray structures overlaid on top of the GRB2 SH2 MPS NMR pharmacophore model. In Figure 2-4A and Figure 2-4B, tryptophan 121 is rendered as sticks between the pY+1 and pY+2 surfaces.

2.4.2.3 Comparison of the FKBP12 pharmacophore models

FKBP12 is a peptidyl prolyl cis/trans-isomerase that catalyzes the isomerization of proline amide bonds in proteins and peptides, and it is known to act as an immunosuppressant in complex with FK506 or Rapamycin (86, 87). The proline ring of a substrate sits in the center of the active site, which is a hydrophobic pocket lined with aromatic and hydrophobic residues with a tryptophan residue forming the base of the pocket (88, 89). Pharmacophore models from both crystal and NMR structures identify this key hydrophobic pocket in the center of the active site as illustrated in Figure 2-5A and Figure 2-5B. While the crystal pharmacophore model identifies several elements at the periphery of the active site, the absence of most of these elements from the NMR pharmacophore model is quite apparent and can be attributed to the lack of consensus clusters in the more flexible regions of the NMR ensemble of FKBP12. An overlay of the pharmacophore models with the ligands bound to FKBP12 in the crystal structures (as seen in Figure 2-5C and Figure 2-5D) provides a more detailed understanding of the location of the elements. The only hydrogen-bonding element in the NMR pharmacophore model is a donor element that closely overlaps the carboxylic acid region of the proline residue in the ligands, making hydrogen-bonding interactions with the backbone of the protein (Figure 2-5D). It is interesting to note that some of the elements that are exclusive to the crystal pharmacophore model do not overlap with any ligands from the FKBP12 protein ligand complexes. This does not appear to be the case for the NMR model where most ligands from the crystal structures overlap with all pharmacophore elements of the NMR model.

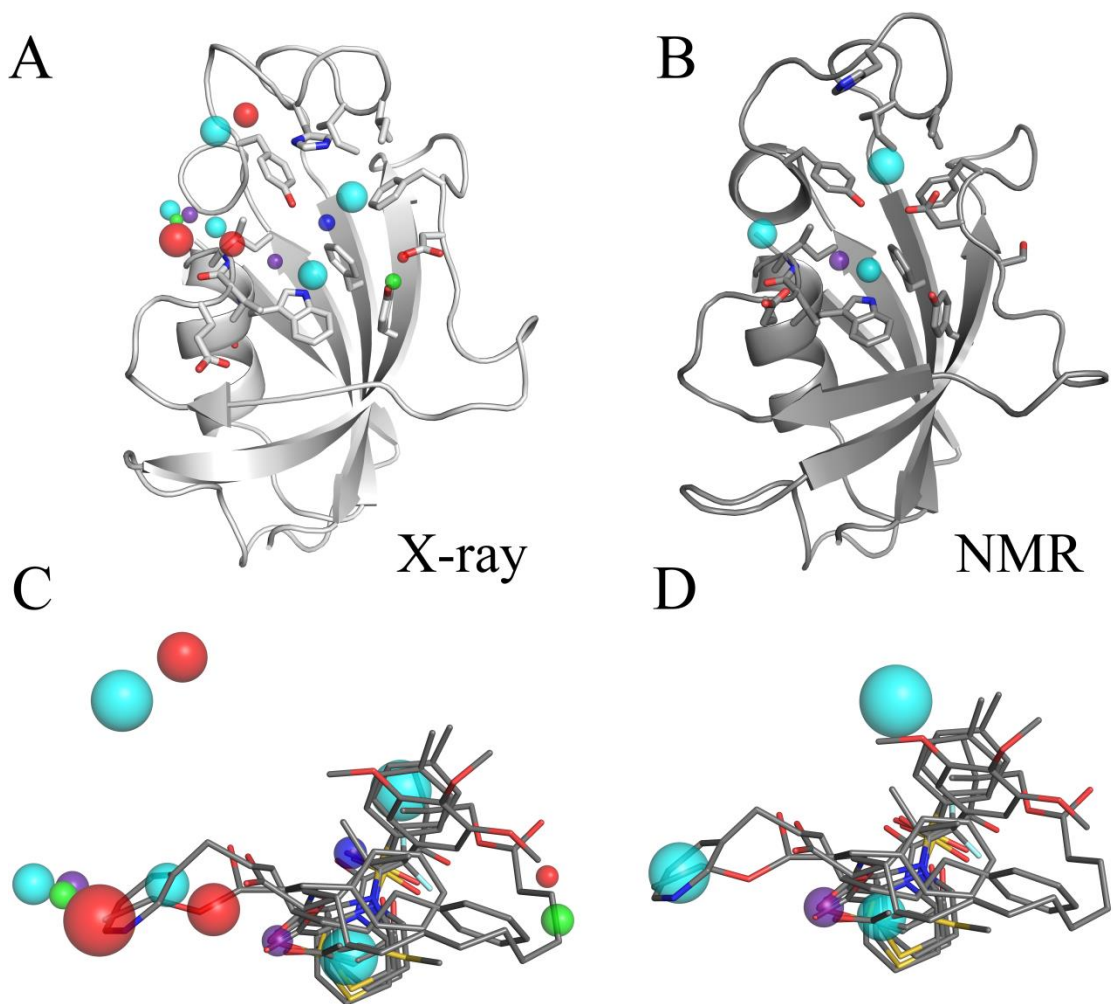


Figure 2-5. Coloring and radii of the pharmacophore elements are the same as in Figure 2-3. A) The MPS pharmacophore model for FKBP12 derived from X-ray structures. B) The MPS pharmacophore model for FKBP12 derived from the NMR ensemble. C) The ligands from X-ray structures overlaid on top of the FKBP12 X-ray model. D) The ligands from X-ray structures overlaid on top of the FKBP12 NMR model.

2.4.2.4 Comparison of the PPAR- γ pharmacophore models

PPAR- γ is a ligand-activated transcriptional factor. It primarily consists of a ligand-binding domain and a DNA-binding domain (90). PPAR- γ agonists bind to the ligand-binding domain and stabilize helix 12 located at the C-terminus, resulting in a conformational change to a closed form of helix 12. PPAR- γ agonists stabilize helix 12 through a network of hydrogen bonds

involving Tyr 473 in helix 12 and several polar residues in the vicinity (His 449, His 323, and Ser 289) typically through a carboxylic acid or thiazolidinedione moiety (91). MPS pharmacophore models in this study were obtained by flooding at the center of the active site in the ligand-binding domain. The crystal pharmacophore model displayed six elements which included a doneptor element that mapped the functional moiety of PPAR- γ agonists that stabilizes helix 12 (see Figure 2-6A). Four of the six elements in the crystal pharmacophore model were seen to overlay well with the crystal structure ligands. As a ligand-bound NMR ensemble for PPAR- γ was not available, we had to use an apo NMR ensemble for creating the pharmacophore model. The NMR pharmacophore model had three elements and the doneptor element mimicking the key functional moiety of PPAR- γ agonists was absent (see Figure 2-6). Helix 12 in the NMR ensemble (nine apo structures) sampled a wide variety of open conformations (see Figure 2-6) that did not resemble the well-ordered, hydrogen-bonding network seen in the agonist bound PPAR- γ crystal structures. Consequently, the flexibility and the absence of a doneptor element that mapped the PPAR- γ agonist functional moiety was expected. A hydrophobic element that overlapped with the agonists near helix 12 in the crystal pharmacophore model appeared shifted in the NMR model and more closely mapped the location of the Tyr 473 seen in the crystal structures. It is important to note that while the NMR pharmacophore model mapped important locations of the protein, these locations were less important for ligand binding and of more relevance for the conformational change going from the inactive form of the protein to the ligand-bound, activated form. The only element in common between the NMR and crystal pharmacophore model was a hydrophobic element located at the entrance to the active site of PPAR- γ . Interestingly, the aromatic element in the center of the active site of the crystal pharmacophore model was replaced by a donor element in the NMR pharmacophore model, presumably due to the bent nature of the helix in this region that exposes a cysteine residue (Cys 285) backbone amide in the NMR ensemble.

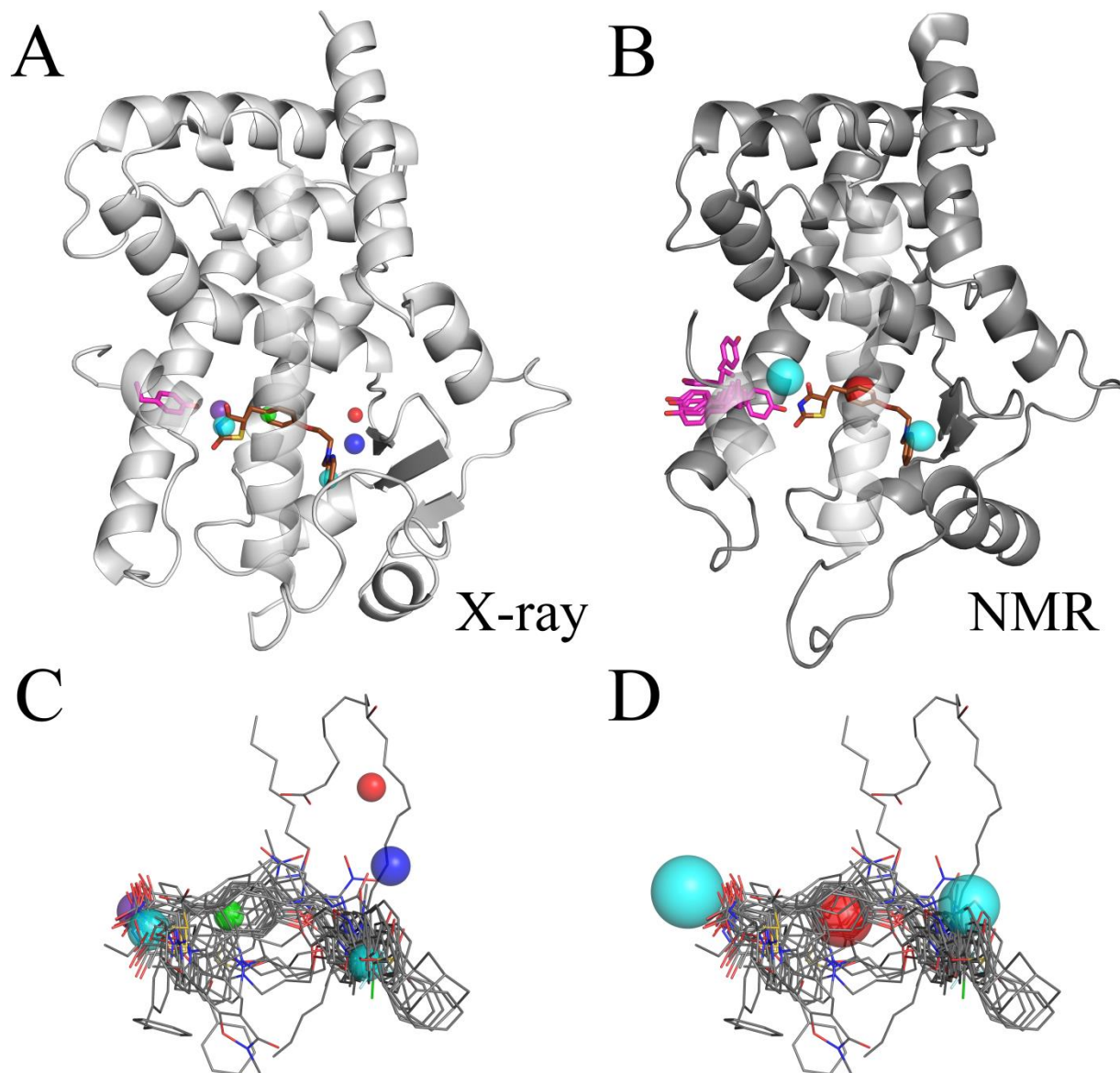


Figure 2-6. Coloring and radii of the pharmacophore elements are the same as in Figure 2-3. MPS pharmacophore models derived from PPAR- γ X-ray and NMR ensembles are shown along with a representative protein conformation. The tyrosine residue 473 which is part of Helix 12 is shown in pink. In the X-ray ensemble, there is limited sampling of the tyrosine residue 473, which corresponds to the active form of the protein. In the NMR ensemble; this residue samples the inactive conformation of the protein. A) The MPS pharmacophore model for PPAR- γ derived from X-ray structures. B) The MPS pharmacophore model for PPAR- γ derived from the NMR ensemble. C) The ligands from X-ray structures overlaid on top of the PPAR- γ MPS X-ray

pharmacophore model. D) The ligands from X-ray structures overlaid on top of the PPAR- γ MPS NMR pharmacophore model. In Figure 2-6A and Figure 2-6B, the location of the binding site is shown by rendering rosiglitazone as a stick model obtained from the PDB ID: 1ZGY (colored brown).

2.4.3 Evaluation of MPS pharmacophore models

Pharmacophore models from NMR ensembles and crystal structures were evaluated by their ability to identify inhibitors/agonists over decoy molecules. In addition, the databases were split into high- and low-affinity data sets to provide a further measure of the effectiveness of the models. Certain unifying trends were observed across all models in general. This included the ability of the pharmacophore models to identify high-affinity inhibitors over low-affinity inhibitors while maintaining selectivity over decoy molecules. Furthermore, NMR models consistently had fewer pharmacophore elements.

The increased flexibility in the NMR ensemble, coupled with the requirement of defining an element only when similar interactions are made across more than 50% of the protein conformations, decreases the likelihood of requiring elements in locations where proteins exhibit greater flexibility. Our goal with MPS is to identify the most essential interactions, so these more variable regions are inherently down played. Regions like these may complement some ligands, but they are not essential to all ligand binding. Furthermore, locations such as these may be associated with entropic costs. With the exception of PPAR- γ , NMR models for all protein targets exhibited optimum performance when all pharmacophore elements were required to identify a hit. This provided support to the argument that conformational flexibility in the NMR ensemble enables one to identify hot spots that closely map functional groups retained across diverse inhibitors and represent the most essential features of the active site. The poor performance of PPAR- γ NMR pharmacophore model can be attributed to the fact that the apo NMR ensemble samples an inactive form of the protein and the resulting pharmacophore model is devoid of the ability to identify agonists that bind to the active form of the protein. Nevertheless, hot spots mapped by the PPAR- γ NMR pharmacophore model

represent locations of protein side chains in the active form of PPAR- γ as seen in the crystal structures and hence resemble sites relevant for conformational transition from the inactive to the active form of the protein.

Conversely, crystal pharmacophore models were too limited in protein flexibility, which resulted in the identification of more non-essential elements. The crystal model's performance was poor when all the pharmacophore elements were used to identify hits. In contrast to NMR pharmacophore models, elements had to be dropped in order to obtain optimum performance for crystal models, which suggested that the extraneous elements identified in the crystal pharmacophore model represented non-essential sites, meaning that inhibitors/agonists may or may not have the chemical features. The addition of these elements hindered performance. Dropping elements from the crystal pharmacophore model allowed a more diverse set of inhibitors to be identified; however, it also resulted in a concomitant increase in the number of decoy molecules identified.

Across all protein targets, there was no trend on the number of elements required to be dropped from the crystal pharmacophore model in order to achieve optimum performance. The lack of such a trend highlights the uncertain nature of the task of optimizing crystal pharmacophore models for protein targets where sufficient data to evaluate model performance does not exist. It is clear that pharmacophore models derived from NMR ensembles do not suffer from such drawbacks since they can be screened by using all of their pharmacophore elements to identify a diverse set of inhibitors, completely obviating the need to drop elements to achieve optimum performance.

Our objective was to approach the problem in an unbiased manner excluding any prior knowledge of the binding mode of ligands and the function of the protein target while preparing MPS pharmacophore models. Hence, we flooded the protein structures with probe molecules at the center of the binding cavity using a 10Å flooding radius across all protein targets the same as one might center a "docking box" on the middle of a binding site. While

flooding in this study was performed with 500 probe molecules, we also evaluated the effect of flooding with 2000 probe molecules and found that several non-essential sites were identified in both NMR and crystal structures, limiting the performance of crystal and NMR pharmacophore models (data not shown).

ROC plots for each protein target, characteristics that deviate from general trends discussed above, and reasons for such anomalies are elaborated in further detail for each protein target below.

2.4.3.1 Performance of the pharmacophore models of the Src SH2

The optimal NMR pharmacophore model (6/6 sites, $2.66 \times$ RMSD, Figure 2-7) identified 93.7% of high-affinity inhibitors and 10.5% decoys. The best performing crystal pharmacophore model (7/10, $2.66 \times$ RMSD, Figure 2-7) identified a similar number of inhibitors at the expense of 11.9% decoys. Pharmacophore models from both NMR and crystal structures were able to distinguish high-affinity inhibitors from low-affinity inhibitors. While optimal pharmacophore models from both models identified a similar number of inhibitors, it is important to note that the NMR pharmacophore model achieved this task using all the elements in the pharmacophore model, unlike the crystal pharmacophore model where three elements had to be dropped in order to achieve a similar result.

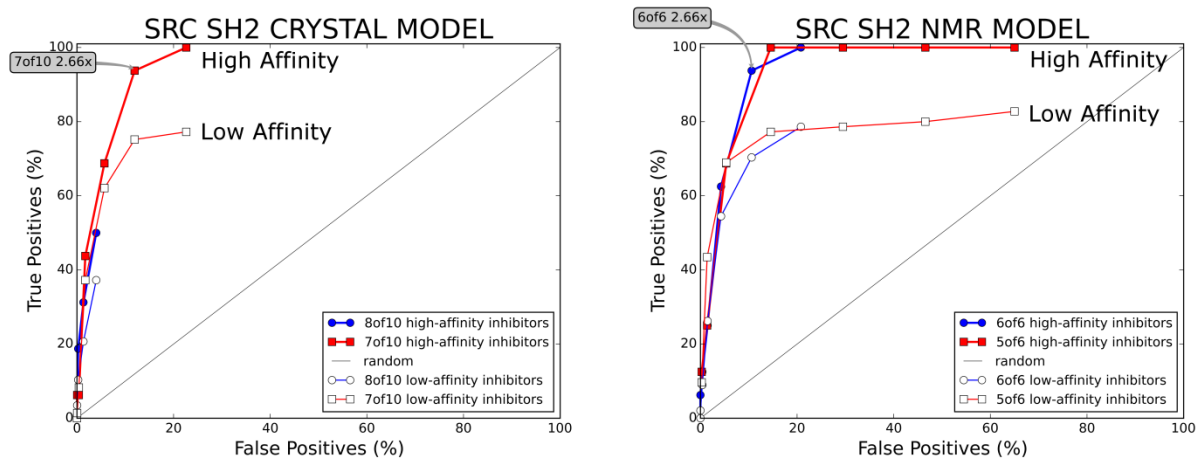


Figure 2-7. ROC plots of crystal and NMR pharmacophore models of Src SH2 are shown along with a label for the model that displays the best performance. The best performing crystal pharmacophore model was at $2.66 \times$ RMSD using seven out of ten pharmacophore elements. The best performing NMR pharmacophore model was achieved at $2.66 \times$ RMSD using all six pharmacophore elements.

2.4.3.2 Performance of the pharmacophore models for Grb2

Pharmacophore models from both crystal and NMR Grb2 models were successful in differentiating inhibitors over decoys and high-affinity over low-affinity inhibitors. The NMR pharmacophore model of Grb2 displayed optimum performance when all elements of the pharmacophore model were included (8/8 sites, $3 \times$ RMSD, Figure 2-8). This pharmacophore model identified 98.3% of high-affinity inhibitors and 5.6 % of decoys. In contrast, the crystal pharmacophore model exhibited optimum performance identifying 98.3% of high-affinity inhibitors and 7.5% of decoys when screened with a model (8/9, $3 \times$ RMSD, Figure 2-8) where one element was dropped. The crystal pharmacophore models displayed similar characteristics observed for Src SH2 where progressively dropping pharmacophore elements improved the ability of the models to identify a diverse set of inhibitors.

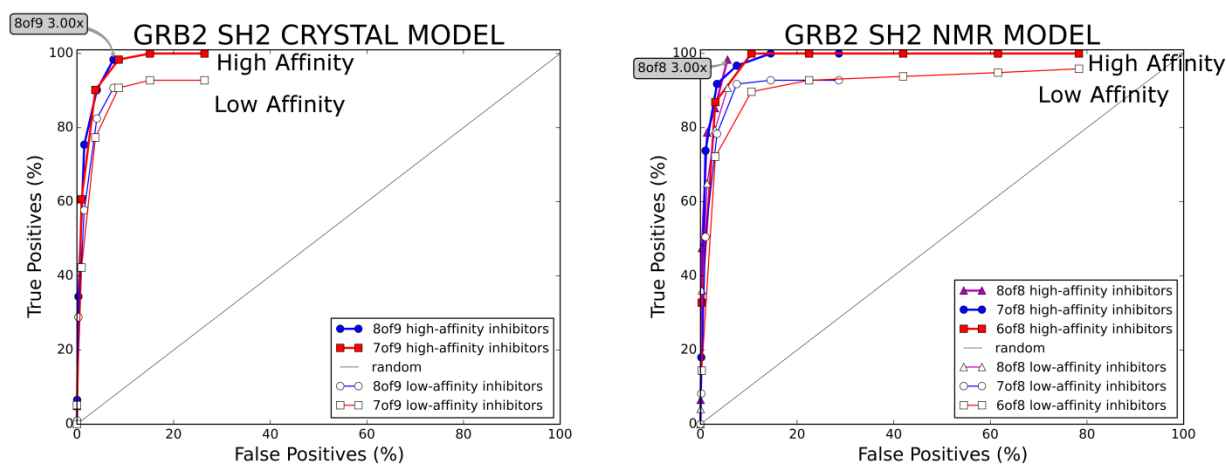


Figure 2-8. ROC plots of crystal and NMR pharmacophore models of Grb2 SH2 are shown along with a label for the model that displays the best performance. The best performing crystal pharmacophore model was at $3.00 \times$ RMSD using eight out of nine pharmacophore elements.

The best performing NMR pharmacophore model was achieved at $3.00 \times$ RMSD using all six pharmacophore elements.

2.4.3.3 Performance of the pharmacophore models of FKBP12

The FKBP12 NMR pharmacophore model was rather simple and had only four pharmacophore elements. Its optimum model (4/4, $1 \times$ RMSD, Figure 2-9) identified 73% of the high-affinity inhibitors and 10.3 % of decoys. In contrast, the crystal pharmacophore model identified 14 pharmacophore elements. Elements of the NMR model represented a subset of the crystal pharmacophore model. The inability to identify either inhibitors or decoys using the full crystal pharmacophore model points to the fact that several of these sites are not required for inhibitors to exhibit activity, and they severely limit performance of the crystal pharmacophore model. The crystal pharmacophore model showed decent performance only when half of the pharmacophore elements were dropped, and the optimum model (7/14, $2.66 \times$ RMSD, Figure 2-9) identified 80.7% of high-affinity inhibitors at the expense of 28.4% decoys.

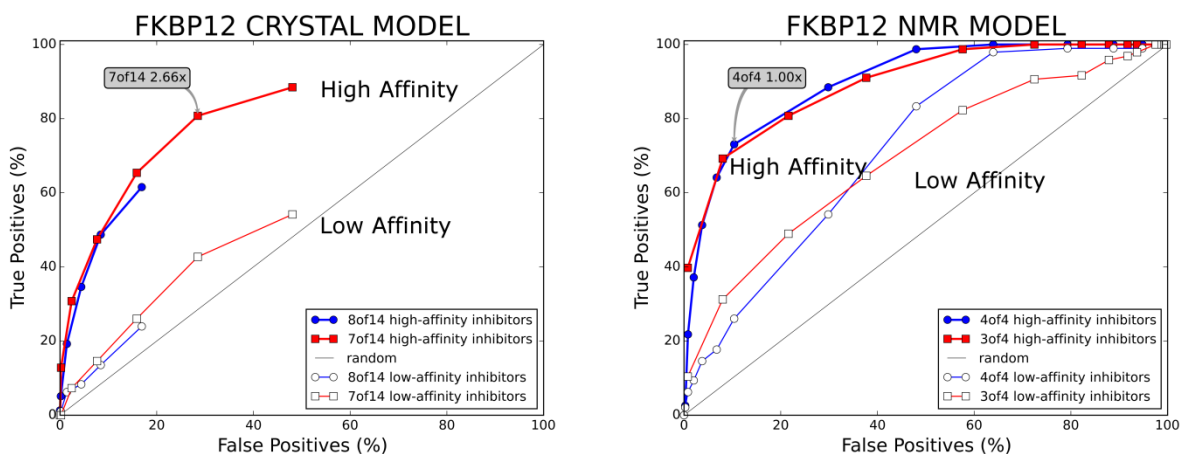


Figure 2-9. ROC plots of crystal and NMR pharmacophore models of FKBP12 are shown along with a label for the model that displays the best performance. The best performing crystal pharmacophore model was at $2.66 \times$ RMSD using seven out of fourteen pharmacophore elements. The best performing NMR pharmacophore model was achieved at $1.00 \times$ RMSD using all four pharmacophore elements.

2.4.3.4 Performance of the pharmacophore models of the PPAR- γ

PPAR- γ presented a unique case where the crystal pharmacophore model displayed better performance than the NMR pharmacophore model. These differences in performances reflect the different conformations upon which crystal and NMR pharmacophore models were built. The crystal model was derived from protein conformations of the active form where helix 12 sampled a “bound conformation” in all of the ligand-bound crystal structures. In contrast, the NMR pharmacophore model was based on apo, inactive conformations in the NMR ensemble where helix 12 was found in the open conformation.

The optimum crystal pharmacophore model (5/6, $2.33 \times$ RMSD, Figure 2-10) identified 90.7% of high-affinity agonists and 21.6 % decoys. When all six pharmacophore elements of the crystal model were used to screen for agonists, very few were identified as hits. Moreover, under such a constraint, the crystal pharmacophore displayed a selective preference for low-affinity agonists over high-affinity agonists. Similar results were found for our previous MPS models for the protein dihydrofolate reductase derived from crystal structures (16, 19).

The NMR pharmacophore model performed poorly in identifying both high-affinity and low-affinity agonists, and again the flipped specificity was seen. The optimum NMR pharmacophore model (3/3, $1.33 \times$ RMSD, Figure 2-10) identified 69.7% low-affinity inhibitors and 37.5% decoy molecules. This emphasizes the importance of pharmacophore elements absent in the NMR pharmacophore model that interact with the tyrosine residue (Tyr 473) in helix 12 and their key role in defining a high-affinity PPAR- γ agonists (91).

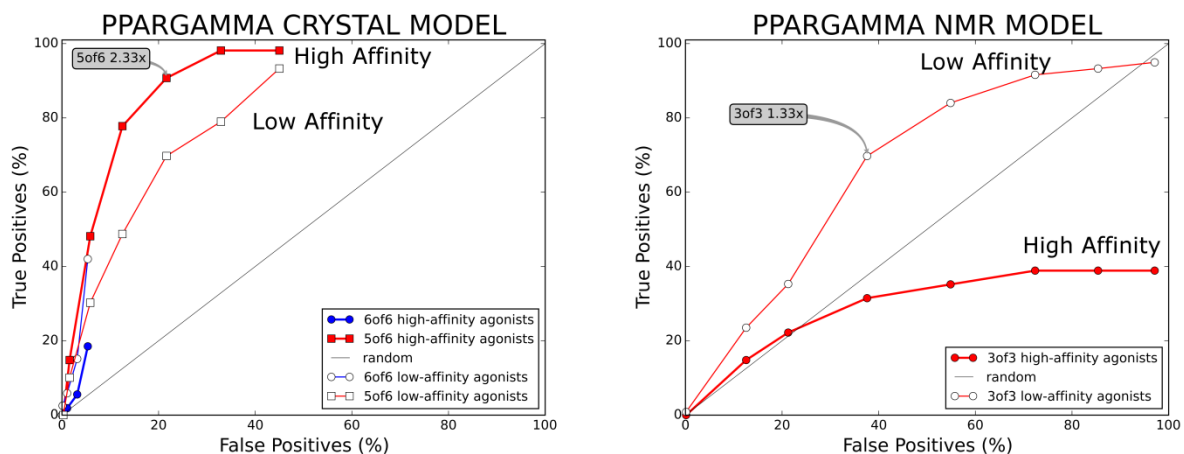


Figure 2-10. ROC plots of crystal and NMR pharmacophore models of PPAR- γ are shown along with a label for the model that displays the best performance. The best performing crystal model was at $2.33 \times$ RMSD using five out of six pharmacophore elements. The best performing NMR model was achieved at $1.33 \times$ RMSD using all three pharmacophore elements. The NMR pharmacophore model was built from an NMR ensemble that samples the inactive conformation, so this model was expected to perform poorly.

2.4.4 Locating and characterizing extraneous elements in crystal pharmacophore models.

Extraneous elements limited the performance of crystal pharmacophore models, and they required these extra elements to be dropped (at random) in order to achieve optimal performance. Given that crystal structures represent the predominant method of structure determination, it is important to develop methods that circumvent these issues. While dropping elements at random to achieve optimum performance presents one such alternative, it is unclear how to do this in a prospective case since this seems to vary by protein target. In order to investigate the location of these extraneous elements and their impact on model performance, we removed pharmacophore elements beyond a defined cutoff radius from the center of the active site. The resulting truncated model was screened against inhibitor and decoy molecules as was done earlier.

In PPAR- γ and Grb2, removing elements at the periphery of the active site resulted in pharmacophore models that exhibited optimum performance when screened using all of their pharmacophore elements. PPAR- γ (Figure 2-11) and Grb2 (Figure 2-12) exhibited such behavior when screened after using a cutoff of 7 Å and 8 Å, respectively, to truncate pharmacophore models. Interestingly, minor imperfections in the pharmacophore model such as the identification of low-affinity inhibitors over high-affinity inhibitors seen with the original PPAR- γ crystal pharmacophore model disappeared using a cutoff-based truncation. This is particularly important when the active form was only available in crystal structures.

This approach was found to be most useful for FKBP12, where removing pharmacophore elements beyond 8 or 9 Å of the center of the active site resulted in models that required less elements to be dropped in order to achieve optimal performance (see Figure 2-13). It appears that crystal packing effects cause side chains at the periphery of this active site to be artificially constrained, creating false consensus elements in the pharmacophore models. This is clearly seen Figure 2-14. Coloring the backbone based on C_{α} RMSD shows that most of the pharmacophore elements unique to the crystal pharmacophore model drop out in the NMR pharmacophore model due to higher protein flexibility.

Interestingly, when cutoffs were applied to the model of Src SH2, degradation in performance was observed (see Figure 2-15). This was due to the removal of pharmacophore elements located at the periphery in the pY+3 pocket. The pY+3 pocket is known to determine specificity and contributes significantly to the binding affinity of inhibitors. Inhibitors with larger hydrophobic residues in the pY+3 pocket are known to bind with higher affinity to Src SH2 (92).

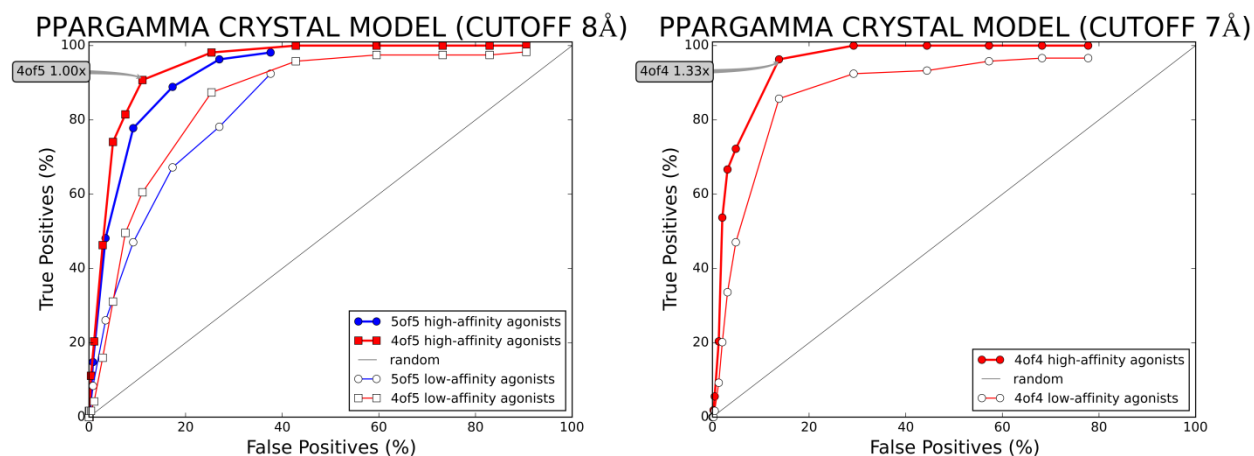


Figure 2-11. ROC plots are shown for the PPAR- γ crystal pharmacophore model with cutoffs of 8 Å and 7 Å from the center of the binding site. The best performing crystal pharmacophore model at a cutoff of 8 Å was 1.00 \times RMSD using four out of five pharmacophore elements and 1.33 \times RMSD using four out of four pharmacophore elements for a cutoff of 7 Å.

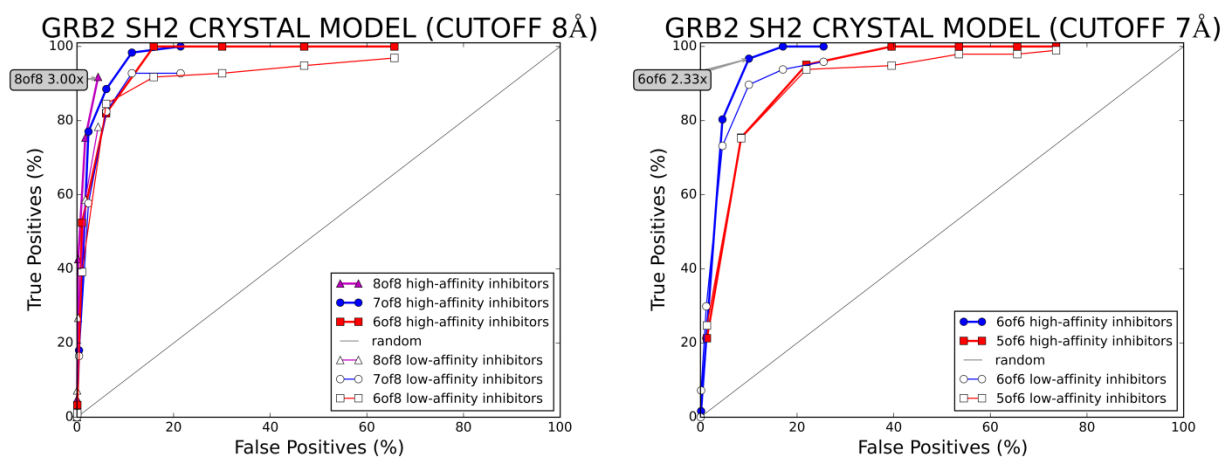


Figure 2-12. ROC plots of Grb2 SH2 crystal model truncated to 8 Å and 7 Å are shown. The best performing model at a cutoff of 8 Å was 3.00 \times RMSD using eight out of eight pharmacophore elements and 2.33 \times RMSD using six out of six pharmacophore elements for a cutoff of 7 Å.

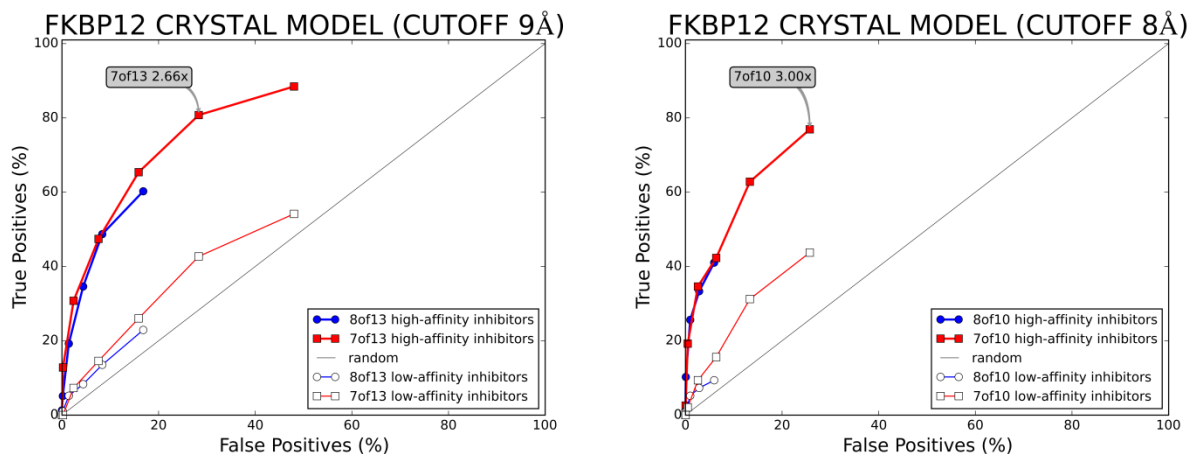


Figure 2-13. ROC plots of FKBP12 crystal pharmacophore model truncated to 9Å and 8Å are shown along with a label for the model that displays the best performance. The best performing crystal pharmacophore model at a cutoff of 9Å was 2.66 × RMSD using seven out of thirteen pharmacophore elements and 3.00 × RMSD using seven out of ten pharmacophore elements for a cutoff of 8Å.

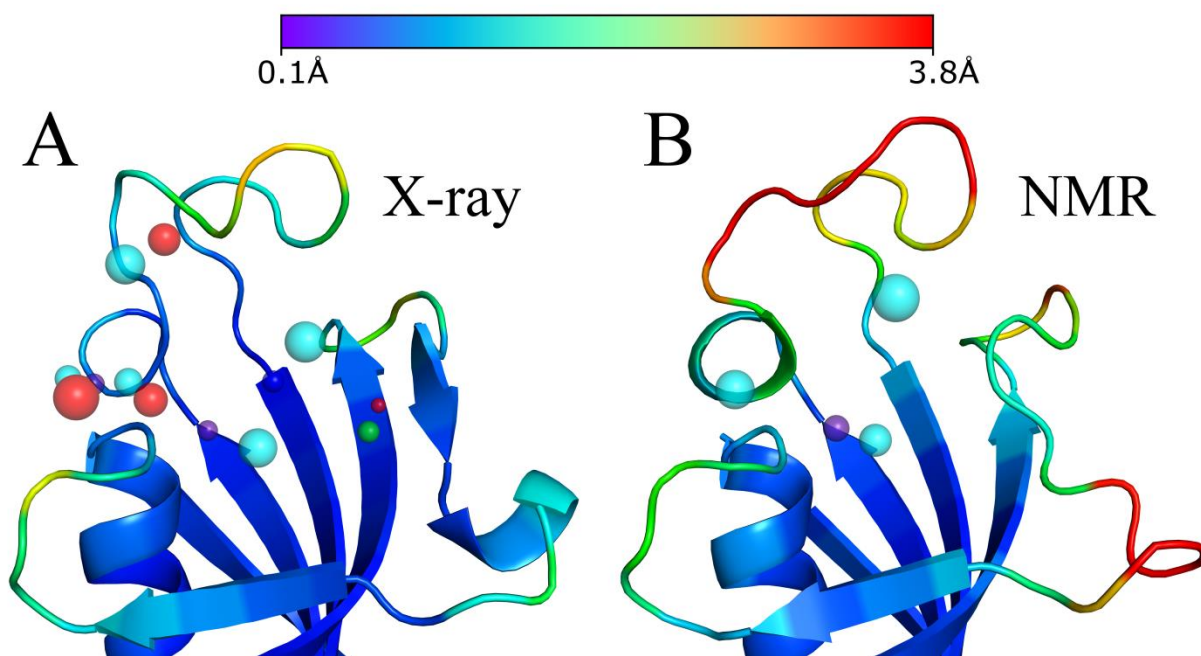


Figure 2-14. A) FKBP12 crystal pharmacophore model and B) FKBP12 NMR pharmacophore model. The models are overlaid on the protein which is color coded by C_α RMSD after wRMSD

alignment. The color scale ranges from Blue (0.1 Å RMSD) to Red (3.8 Å RMSD). The increased flexibility of the NMR ensemble reduces the consensus across the probes used in constructing the model, which removes several elements present in the crystal pharmacophore model.

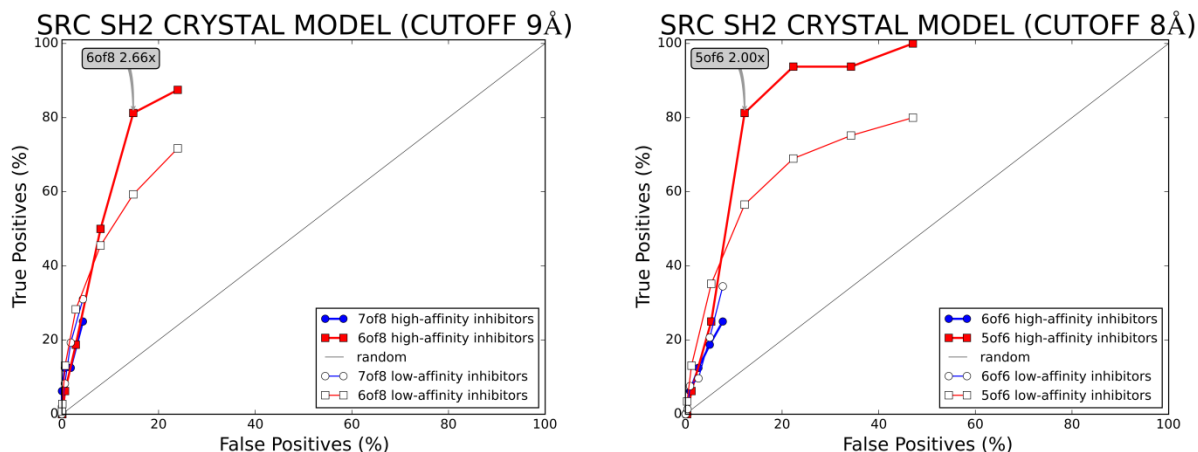


Figure 2-15. ROC plots of Src SH2 crystal pharmacophore model truncated to 9Å and 8Å are shown. The best performing crystal pharmacophore model at a cutoff of 9Å was 2.66 × RMSD using six out of eight pharmacophore elements and 2.00 × RMSD using five out of six pharmacophore elements for a cutoff of 8Å.

2.5 Conclusions

MPS pharmacophore models displayed selective preference for high-affinity inhibitors over low-affinity inhibitors. NMR pharmacophore models exhibited optimum performance when screened using all of their pharmacophore elements, an observation that lends support to the argument that greater flexibility in the NMR ensembles aids in identifying only essential pharmacophore elements. In contrast, crystal pharmacophore models identified a greater number of pharmacophore elements, some of which had to be dropped in order to improve performance. However, the number of elements to drop varied across the protein targets.

In order to understand the location and impact of these extraneous pharmacophore elements on model performance, we have truncated pharmacophore models using different cutoffs from the center of the active site. The X-ray pharmacophore models retained performance for the

most part upon truncation of pharmacophore elements at the periphery. These results confirm that the extraneous pharmacophore element are primarily located at the periphery and do not contribute any value to pharmacophore model performance. This study highlights the relationship of protein flexibility with MPS pharmacophore model performance.

2.6 Supplementary Information

Appendix A provides additional information on pharmacophore model coordinates, used in this study, ROC plot data, and a list of PDB structures used for the NMR and crystal structures used to build the pharmacophore models.

Chapter 3. Moving Beyond Active-site Detection: MixMD Applied to Allosteric Systems

3.1 Abstract

Mixed-solvent molecular dynamics (MixMD) is a hotspot-mapping technique that relies on molecular dynamics simulations of binary solvent mixtures. Previous work on MixMD has established the technique's effectiveness in capturing binding sites of small organic compounds. The MixMD approach embraces full protein flexibility while allowing competition between probes and water. Sites preferentially mapped by probe molecules are more likely to be hotspots. First, we establish a rigorous protocol for the identification of hotspots on the binding surface. There are two important requirements: 1) hotspots must be mapped at very high signal to noise ratio and 2) the hotspots must be mapped by multiple probes. We have focused our probe molecule repertoire to include acetonitrile, isopropanol, and pyrimidine as these probes allowed us to capture a range of interaction types that include hydrophilic, hydrophobic, hydrogen-bonding and aromatic interactions. Charged probes were needed for mapping one target. Second, we used MixMD to identify both competitive and allosteric sites on proteins. In order to demonstrate the robust nature and wide applicability of the technique, a combined total of 5 μ s of MixMD was applied across several protein targets known to exhibit allosteric modulation. The protein test cases were Abl Kinase, Androgen Receptor, Chk1 Kinase, Glucokinase, Pdk1 Kinase, Protein-Tyrosine Phosphatase 1B, and Farnesyl Pyrophosphate Synthase. The success of the technique is demonstrated by the fact that the top-four sites solely map the competitive and allosteric sites. While the lower-ranked sites consistently map multimerization interfaces, other biologically relevant sites, or crystal packing interfaces.

3.2 Introduction

Traditional structure based drug design (SBDD) often relies on targeting the active site as a means of inhibiting protein function. However, such an approach may prove to be challenging in some protein targets. Allosteric sites on proteins allow an opportunity to circumvent such issues. Allostery has traditionally been defined as the modulation of function as a result of an effector binding at a site distant from the orthosteric site. Our evolving understanding of allosteric modulation has moved us from the sole view of an induced fit and conformational change driven mechanism to include mechanisms dominated by population shift and conformational selection. Indeed, several studies have shown the existence of allostery in the absence of any notable change between the allosteric effector bound and unbound conformations, further strengthening the argument of a more dynamic view of allosteric mechanisms (93). Allostery is clearly important for drug design. It has a role in regulatory feedback mechanisms in controlling the activity of many enzymes. This provides an avenue for one to develop drugs to target allosteric sites for curing diseases (94). Furthermore, targeting allosteric sites can allow one to circumvent decreased effectiveness of inhibitors targeting the orthosteric/active site as a result of escape mutations. Moreover, in certain cases it has been shown that targeting allosteric sites allows one to achieve selectivity when structural similarities in the orthosteric sites across multiple subtypes of the same protein prevents one from achieving desired selectivity (95).

Many discoveries of allosteric sites have risen through serendipitous approaches involving high throughput screens (96). Experimental approaches such as tethering thiol containing small molecules to cysteine residues on the protein surface have also found success in identifying allosteric sites (97). There are several computational techniques that complement the detection of these allosteric sites. Computational methods for the detection of allosteric sites range from sequence-based analysis of evolutionarily conserved residues to decipher the allosteric network (98) to molecular dynamic simulations that attempt to detect an allosteric network through correlated motion of residues (99). These methods while promising have only been applied to a

handful of protein targets and further assessment needs to be done to evaluate their robustness.

In order to take advantage of allosteric sites, it is also essential to assess if such sites are druggable and thereby amenable to drug discovery efforts. Common experimental approaches to assess the druggability of sites on the protein surface include NMR-based fragment screening (100) and crystallography-based methods such as the multiple solvent crystal structures (MSCS) technique (101, 102). Computational probe mapping techniques, inspired by such experimental approaches, provide a cost-effective alternative and allow one to overcome practical challenges in implementing experimental methods. MixMD is one such probe-mapping technique that embraces the dynamic aspect of proteins. The MixMD method uses a molecular dynamics (MD) simulation of the protein in a binary solvent of water and a miscible, organic probe to determine the location where probes preferentially bind. Our earlier efforts in optimizing the MixMD technique have demonstrated that in order to map true hotspots, one needs to take full protein flexibility into account (103). Furthermore, we have optimized the conditions to reduce the number of spurious minima identified on the protein surface (104). Spurious sites are common in other similar methods. Probe-mapping techniques similar with MD have been put forth by several groups, the first to be reported used MD simulations with isopropanol as a single probe at a concentration of 20% v/v (105). A second probe-mapping technique termed SILCS utilized a 1M benzene, 1M propane in water as the solvent mixture to carry out MD simulations (106, 107). The third technique used either isopropanol or a mixture of small fragments (acetic acid, acetamide, isopropylamine, and isopropanol) at a concentration of 20% v/v (100). All probe-mapping techniques reported thus far rely on binning the probe locations onto a grid and identifying hotspots through some form of free energy of binding calculation. Each probe-mapping technique has its merits and drawbacks. Emphasizing on water-miscible organic probes, extending the technique in the pursuit of unknown targets, and using conditions amenable to experimental methods have influenced our methodology development and distinguish our method from similar methods.

Probe-mapping techniques such as MixMD take protein flexibility and competition of organic molecules with water into account. In principle, using drug-like fragments should facilitate the assessment of druggability of plausible binding sites on the protein surface. Competition with water in MixMD allows one to explicitly assess if unfavorable solvation effects can impede binding. In this study, we extend MixMD in pursuit of allosteric sites and show that MixMD can map both active and allosteric sites on proteins. To this end, we have identified a set of protein targets with confirmed allosteric sites. The protein targets used were monomers and included Abl Kinase, Androgen Receptor, Pdk1 Kinase, Farnesyl Pyrophosphate Synthase, Chk1 Kinase, Glucokinase and Protein Tyrosine Phosphatase 1B (PTP1B). In order to provide a robust evaluation of MixMD, it is essential to avoid pre-organization of the allosteric sites. In keeping with this philosophy, we have used proteins with competitive ligands bound, but no allosteric ligands. In subsequent sections, we present full details of our methodology and show the effectiveness of MixMD in identifying both allosteric and competitive sites on proteins.

3.3 Methods

3.3.1 MixMD simulation setup

Simulations were started from a protein conformation with no allosteric ligand bound. The protein structures were stripped of water molecules and any cofactors or active-site ligands. This was followed by the addition of hydrogen atoms using Protonate 3D in MOE (59). The asparagine and glutamine residues were flipped as necessary to achieve optimal hydrogen bonding. Histidine residue tautomerizations were corrected when required. A sufficient number of sodium or chloride ions were added to neutralize the system using the tleap suite of AmberTools (108). A layer of probe molecules was added around the protein using tleap followed by the addition of a sufficient number of TIP3P (109) water molecules as necessary to create a 5% v/v ratio of probe to water. The force field parameters for the probes acetonitrile, isopropanol, and pyrimidine were from our previous work (110). Molecular dynamic simulations were carried out in AMBER 11 (108) using the FF99SB (111) force field. The SHAKE algorithm (112) was used to restrain bonds to hydrogen atoms and a time step of 2 fs was used to integrate the equations of motion. Particle Mesh Ewald approximation as implemented for

the GPUs, PMEMDCUDA (113) was used. Non-bonded interactions were limited to a 10 Å cutoff and an Anderson Thermostat was used to maintain temperature at 300 K. Using this approach, three separate simulations with the probes acetonitrile, isopropanol, and pyrimidine were setup for each protein target. The systems were then subjected to an equilibration protocol to gradually increase the temperature and allow proper relaxation of all the atoms in the system as described previously (104). This was followed by a simulation of 20 ns. For each protein and probe, ten such simulations were carried out resulting in 200 ns of cumulative production simulation time.

3.3.2 Parametrization of acetate and methyl ammonium for use in MixMD

Force field parameters for acetate and methyl ammonium used in MixMD simulations of PTP1B were developed using the same approach as outlined in our previous work for other organic water miscible probes (110). In brief, this approach involves the use of OPLS force field parameters for nonbonded interaction terms. These nonbonded interaction terms were converted for use in AMBER using the conversion factor $r_{min} = 2^{\frac{1}{6}} * \sigma/2$, where σ is the OPLS nonbonded interaction term and r_{min} in AMBER is the distance between the atoms at which the Lennard Jones interaction term is at its lowest energy minimum. The OPLS force field parameters used for acetate and methyl ammonium are presented in the 0.

In order to validate these parameters, a mixture of ~2.5%v/v of acetate and methyl ammonium in a box of TIP3P water was subjected to 5ns of simulation under constant pressure after a series of equilibration steps. The box was ~ 85Å × 85Å × 85Å and consisted of ~64,000 atoms. The equilibration procedure consisted of 49,000 steps of conjugate gradient minimization. Then, the system was gradually heated from 10 to 300K over 20 ps. This was followed by a 2ns constant pressure equilibration. Adequate mixing of acetate and methyl ammonium was confirmed by calculating the radial distribution functions (RDF) of the Oxygen-Oxygen distance in acetate and Nitrogen-Nitrogen distance in methyl ammonium. These RDFs were computed with the radial command in ptraj by using the last 1ns of the simulation data and a bin size of 0.1 Å. The RDFs for acetate and methyl ammonium (shown in Figure 3-1B) converge to 1.0 at long range distances confirming the adequate mixing of probes and relatively uniform

distribution in the box. A snapshot of the last frame (shown in Figure 3-1A) further depicts the even mixing of acetate and methyl ammonium.

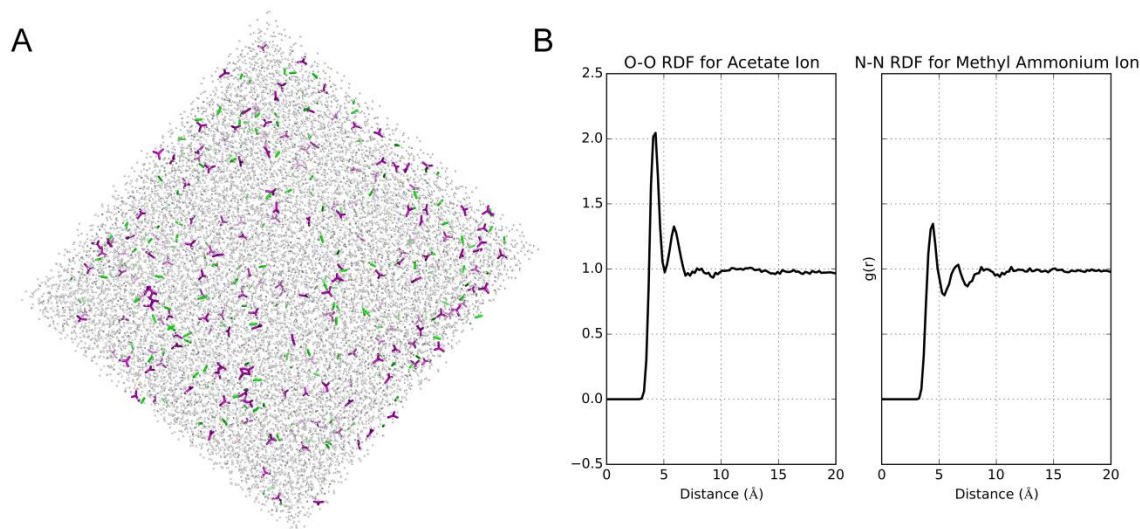


Figure 3-1 (A) The final snapshot from a simulation of $\sim 2.5\%$ v/v mixture of acetate, methyl ammonium and water for 5ns demonstrates proper mixing was achieved. Acetate ions are colored purple, methyl ammonium ions are colored green and water molecules are colored white. (B) Adequate mixing of acetate, methyl ammonium probe molecules in MixMD was confirmed by radial distribution functions that displayed a probability of 1.0 at long range distances.

3.3.3 Processing MixMD results

The location of all atoms within the probe molecules from the last five nanoseconds of the ten runs for each protein target were binned onto a grid of 0.5 \AA spacing using the Ptraj module from AmberTools (108). The raw bin counts in each of the grid points were converted to sigma values using the equation $(x - \mu)/\sigma$ where μ is the mean of the binned grid data and σ is the standard deviation of the binned grid data. This allows us to represent the location of the probes in a manner commonly implemented for electron density from X-ray crystallography. The resulting maps were contoured at various sigma values and examined in the presence of the average protein structure to identify locations of maximal occupancy. A higher sigma value

for a particular location on the grid signifies a higher residence time for a probe molecule at that particular location across all ten MixMD simulation runs. The maps in this study have been color coded as orange for acetonitrile, blue for isopropanol, and magenta for pyrimidine to represent the respective probe simulations from which they have been derived. These maps were visualized in PyMOL (75).

3.4 Results and Discussion

3.4.1 Choice of protein targets and conformations

The definition of allostery is broad and in general is used to imply anything that does not modulate a protein's activity by interacting with the competitive site. Under such a definition, there are an innumerable number of protein targets that one can choose from to test the effectiveness of techniques to identify allosteric sites. In order to avoid misinterpretation of allosteric sites, we focused our attention on those targets for which experimental data clearly supported an allosteric mechanism; moreover, we limited our choice of protein targets to those that had a verified allosteric site confirmed through crystallography. This allowed a proper and fair comparison of our MixMD mapping results for competitive and allosteric molecules in crystal structures. *In order to provide a robust analysis of the technique, we chose to start MixMD simulations from crystal structures with no allosteric ligand bound.* Complex allosteric mechanisms exist where allosteric effectors modulate the quaternary relationship of a multimeric complex of a protein. Simulating a large, multimeric complex is computationally expensive. As an example, Bacterial L-lactate dehydrogenase converts pyruvate to L-lactate, and the protein exists in a tetrameric state that has either high or no affinity for the substrate, depending on binding and unbinding of the allosteric effector fructose 1,6-bisphosphate. In order to accurately map the allosteric sites, such a system would need to be simulated as a tetramer which can be computationally expensive (114, 115). These large systems were left out from the current analysis and will be the subject of a future study. Careful curation left us with seven protein targets (Abl Kinase, Androgen Receptor, Pdk1 Kinase, Farnesyl Pyrophosphate Synthase (FPPS), Glucokinase, Chk1 Kinase, and PTP1B) for which, there were conformations

with a competitive ligand bound but no allosteric ligands. Apo structures with no ligands were not available for all the systems studied. The PDB IDs of the protein conformation used as the starting conformation for MixMD setup are given in Table 3-1.

Table 3-1 The protein structures used in MixMD are listed. The range of the all-atom RMSD for residues within 4Å of the allosteric site is shown from the MixMD starting conformation to protein conformations with allosteric ligands bound.

Protein Name	Starting conformation used for MixMD	Range of all atom RMSD of allosteric site (Between MixMD starting structure and all Allosteric ligand bound structures in the PDB)
ABL Kinase	3KFA (Chain A)	7.93 – 8.02 Å
Androgen Receptor	2AM9 (Chain A)	1.01 – 1.67 Å
PDK1 Kinase	3RCJ (Chain A)	1.07 – 2.15 Å
Farnesyl Pyrophosphate Synthase	4DEM (Chain F)	1.26 – 2.14 Å
Glucokinase	3IDH (Chain A)	1.05 – 2.82 Å
CHK1 Kinase	1ZYS (Chain A)	0.45 – 0.87 Å
Protein Tyrosine Phosphatase 1B	2CMB (Chain A)	2.02 – 2.12 Å

3.4.2 Identifying and ranking hotspots on the protein surface

Assessing the relative importance of hotspots mapped on the protein surface is essential in establishing their significance. We assessed the mapped sites based on several criteria. First and foremost, *sites mapped at a high sigma value were given greater preference since these sites represent maximally occupied sites*. Second, hotspots must be mapped by more than one probe

type, which implies “bindability” by diverse chemical functionalities. Indeed such an approach to identifying hotspots has been highlighted by Vajda and co-workers in their FT-MAP technique where sites mapped by multiple probes were identified as hotspots (116). It is important to note that our binary solvent setup is essential when we require sites to be mapped by multiple probes. This is a condition that cannot be met in ternary solvent simulations that have been reported earlier (100, 106, 107). This gives MixMD a distinct advantage.

To illustrate the identification of hotspots with MixMD, Figure 3-2 shows Abl Kinase with probe occupancies contoured at varying sigma values. At 90σ in Figure 3-2A, the only hotspot mapped is in the allosteric site. However, upon decreasing it to 85σ (Figure 3-2B), we see that a second hotspot appears in the active site of Abl Kinase. As we continue decreasing the sigma value to 75σ (Figure 3-2C), a third hotspot appears which maps the hinge region of the active site in Abl Kinase. In lowering the sigma value further, we see that a fourth site appears on the protein surface at 50σ . As the maps are contoured at lower sigma values successively from Figure 3-2D- Figure 3-2F, the sites that have already been identified increase in size and start to fuse, while less relevant sites appear. This fusion of mapped hotspots can be seen in the case of hotspot 2 and hotspot 3 that collectively map the entire competitive binding site (Figure 3-2F). Contouring hotspots at 20σ allowed us to examine the full extent to which the probes had mapped the various binding sites on the protein surface. This was a common feature across all the protein targets in this study. Throughout this study, we have found it ideal to focus on the top four hotspots. Unless otherwise stated, from here on all maps are contoured in two ways, one at 20σ to show the full extent to which the top four hotspots map the surface of the protein and another showing the raw occupancy maps at 35σ that clearly show the presence of the top-four sites before other spurious minima. For clarity, those spurious sites are not shown in the 20σ figures.

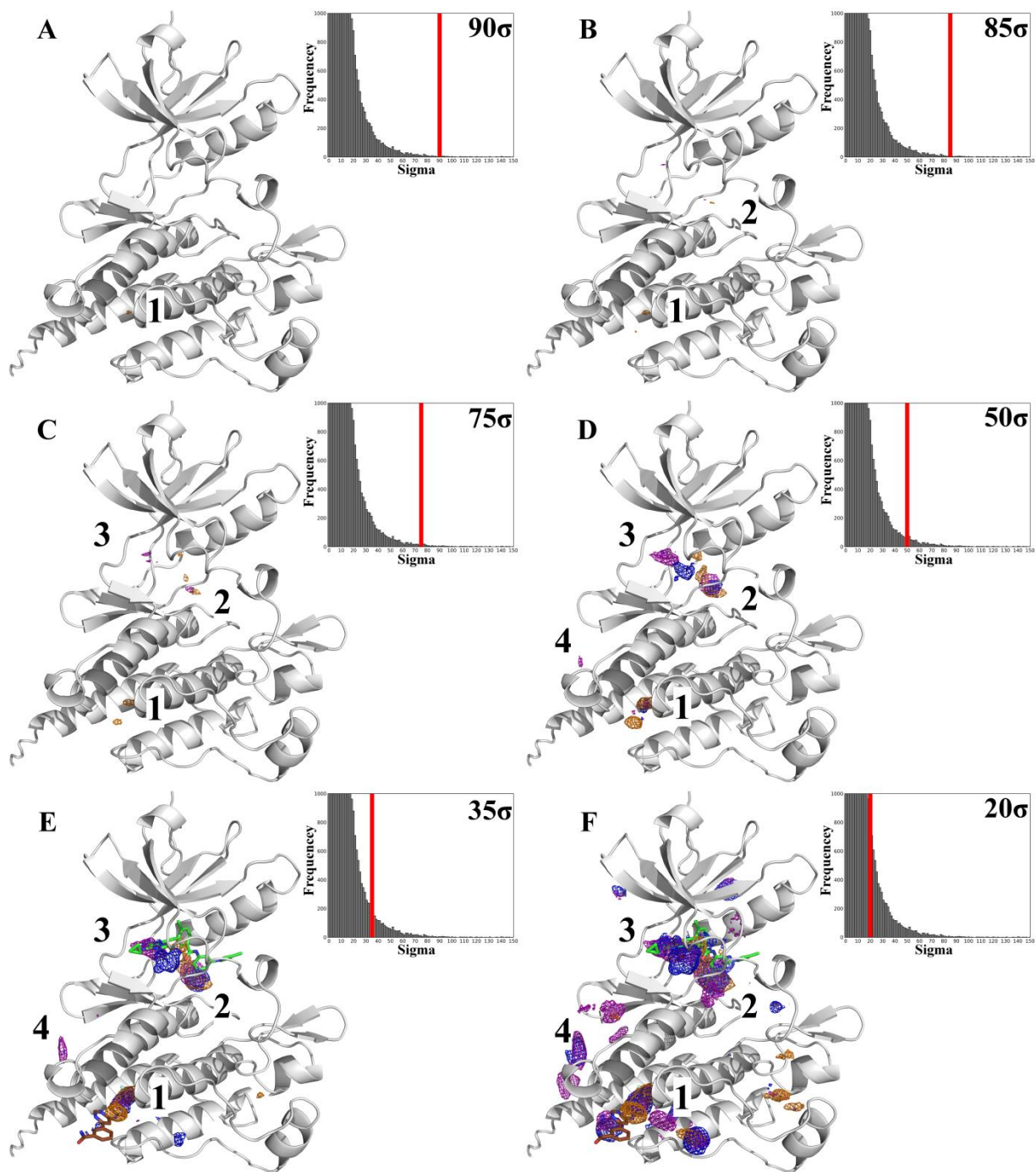


Figure 3-2 The MixMD maps for Abl Kinase are contoured at varying sigma values from 90σ to 20σ to show the degree of molecular surface mapped by the probe atoms. The maps are color coded to represent MixMD maps derived from different probes . Orange - acetonitrile, blue – isopropanol, and magenta – pyrimidine. At 90σ (Figure 3-2A), the allosteric site shows the highest occupied points. Maps contoured at 85σ, show a second hotspot in the active site. In

contouring the MixMD maps from Figure 3-2C-Figure 3-2F at successively lower sigma values, additional hotspots appear and are numbered based on their order of appearance. Unless sites are mapped by more than one probe type when contoured at 20σ they are ignored. The active site ligand (PDB ID: 3KFA, Green) and allosteric site ligand (PDB ID: 3K5V, brown) are only shown for reference in E and F to orient the viewer towards the location of the active and allosteric sites. We emphasize that no ligands were present in the MixMD simulations.

3.4.3 Mapping active and allosteric sites with MixMD

Identifying the hotspots as described in the previous section allowed us to evaluate the importance of various mapped locations on the protein surface. In analyzing the MixMD maps, we have found that the active and allosteric sites are captured in the first-four hotspots for all systems examined. As mentioned earlier in the case of Abl Kinase, all four of these hotspots correspond to the active and allosteric sites or sub-sites thereof. In Abl Kinase, the first hotspot lies in the allosteric site whereas the second and third hotspots map the entire active site (Figure 3-3A). However, it was interesting to observe the fourth hotspot at the side of the protein. Upon checking the protein data bank for molecules that may complement this hotspot, we found that this hotspot location provides the binding interface for the SH2 domain present in the full length protein. As shown in Figure 3-3B, the structure of the full length protein of Abl has a tyrosine residue from the SH2 domain occupying the location of the fourth hotspot. Clearly, this site has an important role in the functionality of Abl kinase. One can envision that targeting such a site may likely disrupt the function of the kinase and thereby achieve allosteric modulation of ABL Kinase function. It is important to stress that while some sites mapped by MixMD may have no known allosteric regulatory role, these may be leveraged in the future to yield such a response.

In the case of the androgen receptor, the first-four hotspots map the active and allosteric sites (Figure 3-4A). As observed for Abl Kinase, the individual hotspots map sub-sites of the active and allosteric site which when contoured at successively lower sigma values fuse to map the entire binding site when contoured at 20σ . It is notable that for Androgen receptor the active and the allosteric site are the only ones mapped in the first-four sites.

While such striking results were not achieved for other targets such as PDK1 Kinase, it is nonetheless important that allosteric sites and active sites were consistently captured in the first-four ranked hotspots (Figure 3-5). The active-site hinge region in PDK1 Kinase was mapped as the top hotspot whereas the allosteric site was mapped by the fourth hotspot. The second hotspot could be traced to a cosolvent binding location (shown in Figure 3-5) and the third hotspot corresponded to the binding location of the Proline ring of a peptide bound in the 3QC4 crystal structure of PDK1 Kinase. These results suggest that MixMD identifies sites that could be easily desolvated, a prerequisite for “druggable” binding sites. Similar results were seen for Chk1 Kinase where the active site near the hinge region was mapped as the top hotspot (Figure 3-7). The allosteric site was ranked as the fourth hotspot and the hotspots ranked second and third denoted the peptide substrate binding location on the protein surface.

MixMD simulations were performed on the proteins as monomers. This allowed the simulations to be completed in a reasonable amount of time. Farnesyl Pyrophosphate Synthase was interesting in this regard, since it functions as a dimer, but we simulated it as a monomer because the active and allosteric sites do not involve the second monomer (of course, a second active and allosteric site are contained in that second monomer). We assumed MixMD would identify part of the dimer interface. It is notable that the interface contains the first hotspot. A tyrosine residue from one of the monomers overlaps with the first hotspot as shown in Figure 3-6. This provides promising evidence in support of the use of MixMD as a technique to probe the location of biologically relevant binding partners of protein-protein interactions. The allosteric site was mapped by the second hotspot, and two sub-sites of the active site in Farnesyl Pyrophosphate Synthase were mapped by the third and fourth hotspots.

In two protein targets, we found that the active site was not mapped by probes. Glucokinase, which binds sugar molecules in its active site, was not mapped. Instead, the sugar-binding site was mapped by water molecules. As sugar-binding proteins are generally not considered druggable, it provides evidence in support of the argument that MixMD assists in the identification of druggable binding sites. While the active site was not mapped in Glucokinase, the allosteric site was extensively mapped by the first hotspot when contoured at 20σ (Figure

3-8). However, the second and third ranked hotspots could not be traced back to examples of molecules that could bind at these locations. The ATP binding site on Glucokinase was mapped as the fourth hotspot.

Protein Tyrosine Phosphatase 1B (PTP1B) was the other protein target where the active site was not mapped. The active site of PTP1B is charged and is known to bind phosphorylated residues. The probes used for MixMD are not charged and as a result did not map the binding site. Similar results were obtained by Bahar and GSK collaborators who had initially carried out a simulation of isopropanol and found no mapping of the active site (100). However, their simulation of a mixture of probes mapped the active site with acetate probes (100). While we were unable to map the active site with our current set of probes in MixMD, the allosteric site was however mapped as expected (Figure 3-9A) and was ranked as the second hotspot. The first, third, and fourth hotspots captured cosolvent and protein-interactions sites as shown in Figure 3-9A.

The identification of the active and allosteric sites in the first-four hotspots was a recurring theme across all the protein targets. The rest of the four hotspots in each protein mostly corresponded to cofactor or cosolvent binding locations and protein-packing interfaces, which are in principle easier to desolvate.

3.4.4 Alternative probes for mapping charged binding sites

Probes selected for MixMD simulations represent fragments derived from drug-like molecules. These probes are water soluble and easily locate desolvable sites on the protein surface. However, we were not able to map the charged binding sites in the active site of PTP1B, with our drug-like probes. This is no surprise as our probes do not complement charge. Moreover, targeting the charged binding site of PTP1B in the context of drug discovery has proved difficult as multiple iterations of medicinal chemistry efforts have been met with limited success in replacing the charged site on inhibitors (117, 118). Several reviews and druggability detection methods on the subject have expressed the view of PTP1B as an undruggable target due to the difficult and slow progress in optimizing compounds (118–120). In fact, this is one of the

primary reasons drug discovery efforts targeting phosphatases have been ignored in favor of kinases, even though kinases and phosphatases are known to work in tandem to regulate major disease-related pathways (118). The discovery of an allosteric site on PTP1B has renewed interest in alternative strategies of targeting PTP1B (121). Our difficulty in mapping the active site of PTP1B with our current set of probes is in line with this growing body of evidence. Of course, we were interested to see if these sites could be mapped by MixMD using charged probes. Bahar and co-workers had noted this in their work and used a cocktail of probes in their simulation of PTP1B to show the binding site was mapped by acetate probes (100). In keeping with our protocol for performing MixMD simulations, we identified methyl ammonium and acetate ions as suitable probes for a ternary solvent simulation. We chose to use a ternary solvent system in this particular case as it allowed us to carry out simulations in a charge-neutral condition. A MixMD simulation of the protein at approx. 2.5 %v/v concentration of each probe was carried out in a similar manner outlined in the methods section. Of the first-four sites mapped on the protein surface, the first two sites mapped two adjacent pockets in the active site. PTP1B acts as a negative regulator of insulin signaling by dephosphorylating tyrosine residues of the insulin receptor which inactivates it and reduces insulin signaling (122). A fragment of the phosphorylated insulin receptor is shown aligned to the MixMD map in Figure 3-9B which demonstrates that the two adjacent hotspots overlap with the phosphorylated tyrosine residues and have functional relevance. This illustrates that while the primary focus of our study was to identify druggable, easily desolvable hotspots on the protein surface, the MixMD technique can easily be adapted to identify other functional sites of relevance by tailoring the set of probes used for performing MixMD simulations.

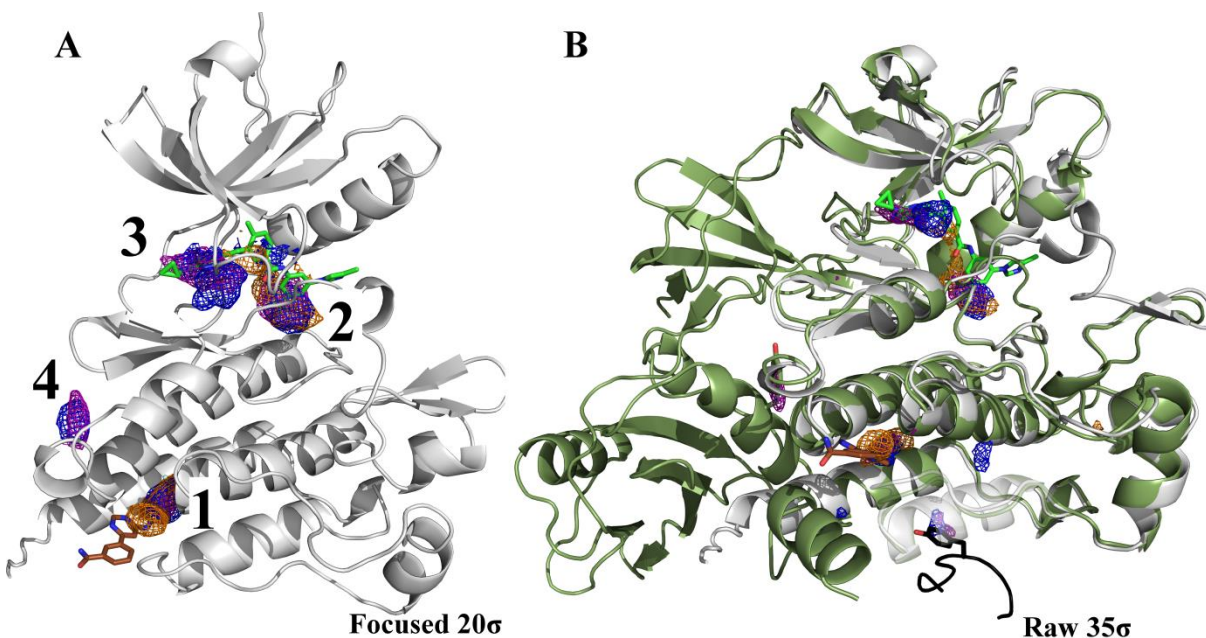


Figure 3-3 The first four hotspots from the Abl Kinase MixMD maps identified the active and allosteric sites. The hotspot rankings are shown on top of the protein structure. The active site ligand (PDB ID: 3KFA, Green) and the allosteric site ligand (PDB ID: 3K5V, Brown) are shown for reference. (A) The four hotspots that map the active and allosteric site are shown contoured at 20σ with the spurious sites not shown. (B) MixMD maps of Abl Kinase contoured at 35σ (all spurious sites are shown) are shown with examples (where available) of molecules from the PDB database bound in probe mapped locations on the protein surface. The crystal structure of the full length Abl protein (PDB ID: 1OPK) was aligned to show the Kinase and SH2 domain interface mapped by the fourth hotspot in MixMD. A tyrosine residue at the packing interface is shown in black (PDB ID: 1OPL). The allosteric and competitive ligands are shown in brown (PDB ID: 3K5V) and green (PDB ID: 3KFA) respectively.

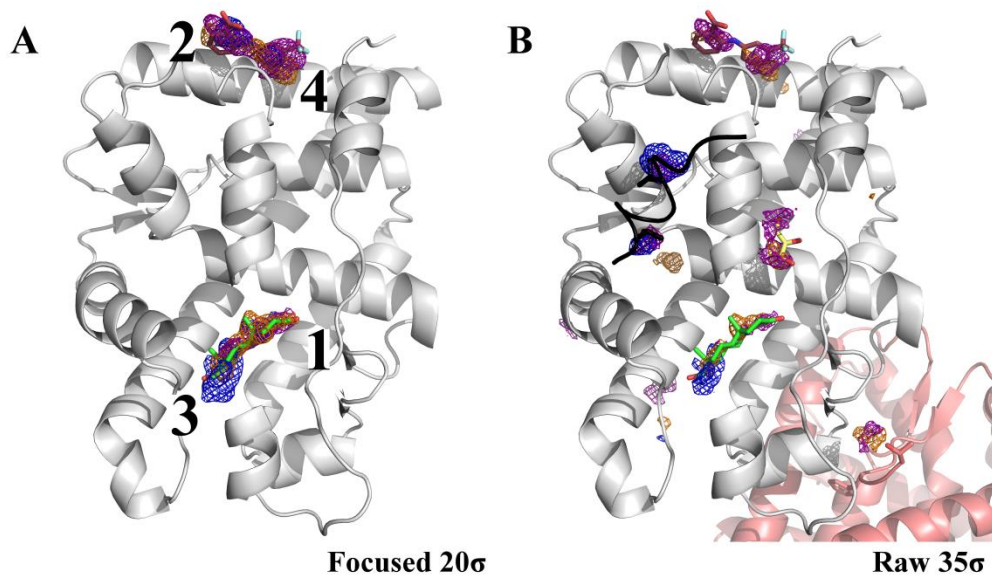


Figure 3-4 (A) The location of just the top four hotspots contoured at 20 σ on the surface of the androgen receptor are shown. The active site ligand (PDB ID: 2AM9, Green) and the allosteric site ligand (PDB ID: 2PIX, Brown) are shown for reference. A part of the alpha helix obstructing the view of the active site ligand has been hidden to provide a better view of the hotspots mapping the active site. (B) The MixMD maps of Androgen receptor are shown contoured at 35 σ to demonstrate that hotspots ranked lower than the top four hotspots, correspond to locations that can be easily desolvated. The different molecules are color coded as follows, Black – PDB ID: 2QPY – Nuclear Receptor Co-Activator 2, Yellow – PDB ID: 4HLW – Glycerol, Pink – PDB ID: 2QPY – Protein Packing Interface.

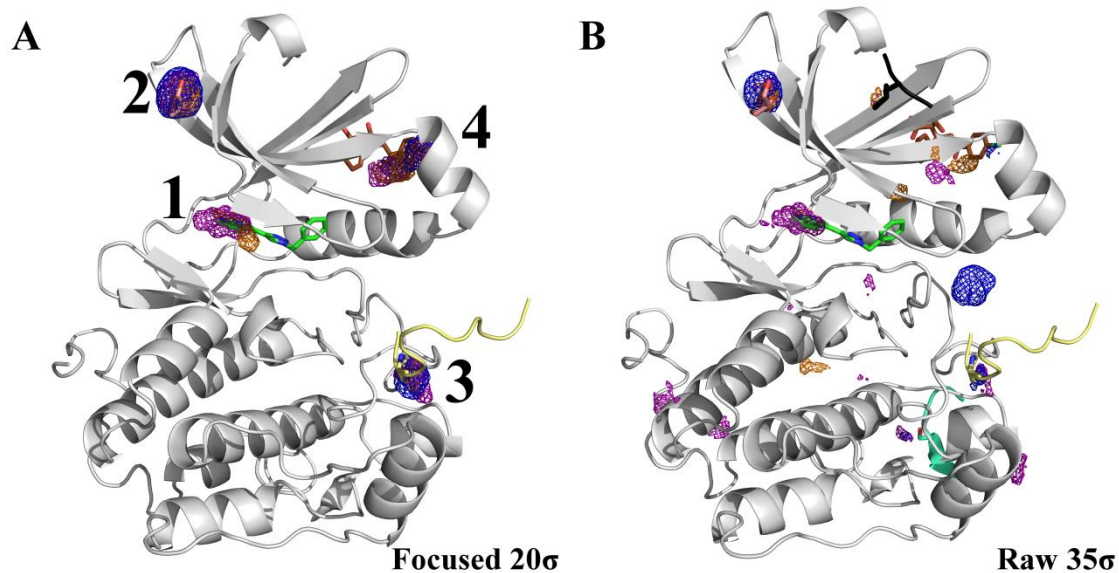
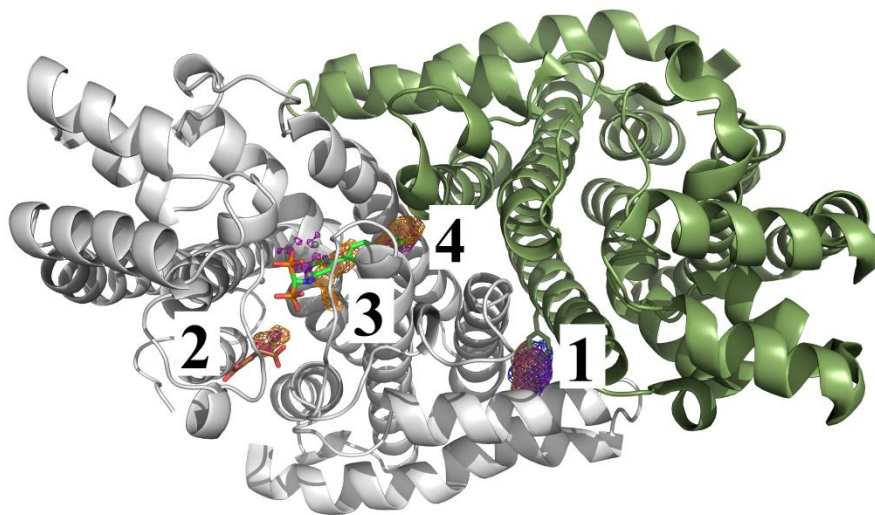


Figure 3-5 Just the top four hotspots for Pdk 1 Kinase are numbered and shown contoured at 20 σ and raw occupancy maps are shown at 35 σ . The first hotspot maps the hinge region of the active site. The second hotspot is located at the top of the protein. A cosolvent bound at this site is overlaid on top from another PDK1 Kinase protein structure (PDB ID: 3RWQ, Pink). The third ranked hotspot is known to bind a peptide (PDB ID: 3QC4, Yellow) and the fourth hotspot corresponds to the allosteric site. The active site ligand (PDB ID: 3RCJ, Green) and the allosteric site ligand from (PDB ID: 4AW0, Brown) are shown for reference.



Focused 20 σ

Figure 3-6 The top four hotspots for Farnesyl Pyrophosphate Synthase are contoured at 20 σ . The first hotspot maps the dimer interface. The second hotspot maps the allosteric site (PDB ID: 3N5J, Brown). The third and fourth ranked hotspots map two different sub sites in the active site (PDB ID: 4DEM, Green). The protein was simulated as a monomer. However the dimer is shown to illustrate that the top ranked hotspot is located at the dimer interface. A tyrosine residue from one of the monomers colored dark green is shown to overlap with the first ranked hotspot at the dimer interface.

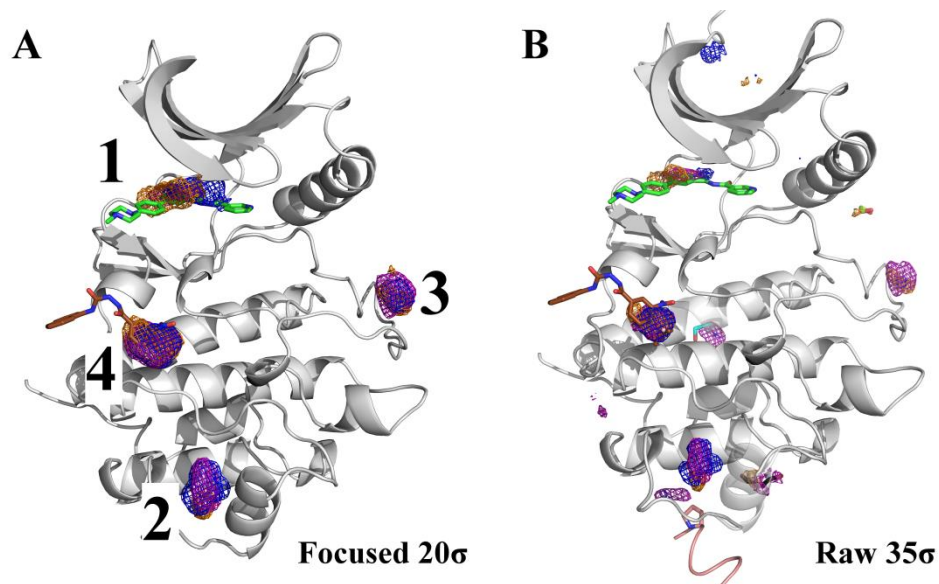


Figure 3-7 Chk1 Kinase is shown with just the top four hotspots contoured at 20 σ . While the first and the fourth hotspots map the active (PDB ID: 1ZYS, Green) and allosteric site (PDB ID: 3JVS, Brown). The second and the third ranked hotspot are located in the peptide substrate binding groove.

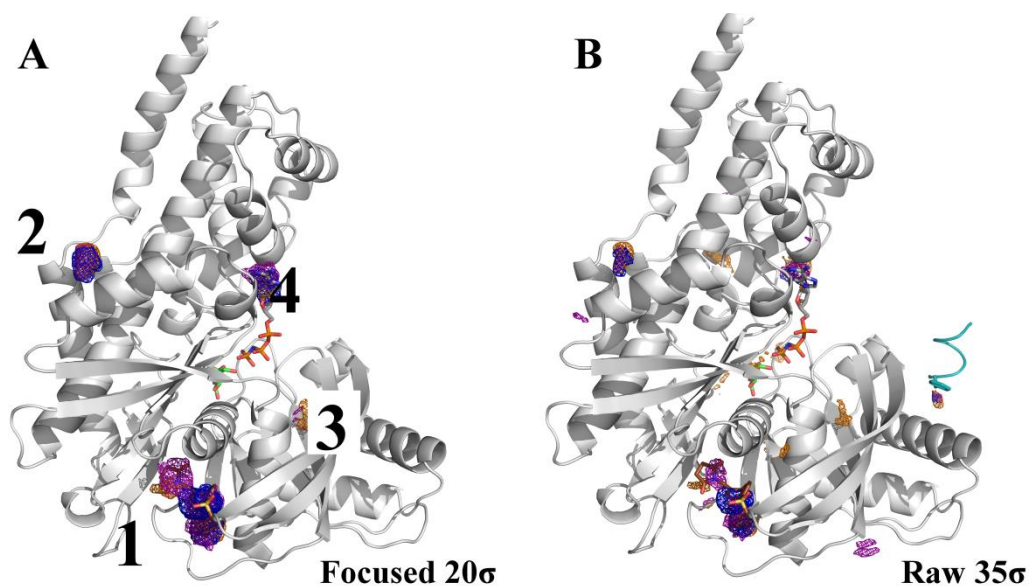


Figure 3-8 Glucokinase is shown with just the top four hotspots contoured at 20 σ . The first hotspot extensively maps the allosteric site (PDB ID: 3H1V, Brown). The fourth ranked hotspot maps the ATP binding site on Glucokinase (PDB ID: 3FGU, Grey). However no examples of molecules could be found that bound to the second and third ranked hotspots for Glucokinase.

B) Only a few very small sites are present 35σ , and they are clearly lower ranked and less occupied.

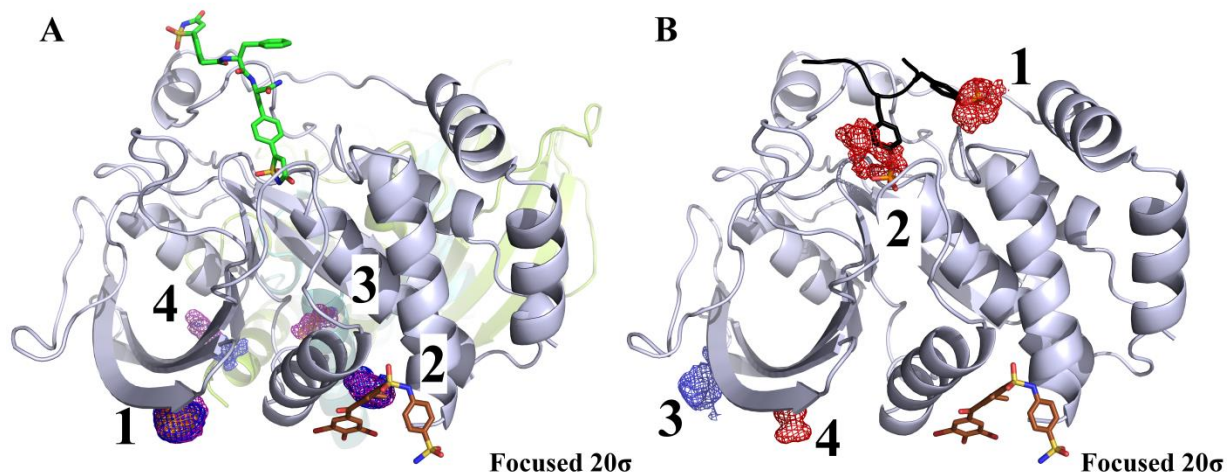


Figure 3-9 (A) The location of the top 4 hotspots is shown for Protein Tyrosine Phosphatase 1B (PTP1B). The first hotspot maps the location of a cosolvent binding site (PDB ID: 3RWQ, Pink). The maps occlude the visibility of this cosolvent). The second hotspot maps the allosteric site (PDB ID: 1T49, Brown). The third and fourth ranked hotspots are located in close proximity to protein packing interfaces (Hotspot 3 is located near the protein packing interface - Cyan colored – PDB ID: 2CMC, Hotspot 4 is located near another protein packing interface – PDB ID: 4GRY– Pea colored). (B) As PTP1B has a charged active site, we were interested to see if charged probes could map these sites. A MixMD simulation of acetate and methyl ammonium was carried out and the top four hotspots ranked in the order in which they appear are shown contoured at 20σ . Acetate hotspots are colored red and methyl ammonium hotspots are colored blue. A fragment of the Insulin Receptor is overlaid on top of the protein (PDB ID: 1G1F, Black). The top two ranked hotspots which correspond to the acetate ion overlap both these sites which are known to bind phosphorylated tyrosine residues.

3.4.5 Identifying co-activator/cofactor and protein oligomerization/ packing interfaces (Evaluating MixMD maps at lower sigma values)

The identification of the active and allosteric sites in the first-four hotspots was a recurring theme across all the protein targets. The rest of the four hotspots in each protein mostly corresponded to cofactor or cosolvent binding locations and protein-packing interfaces, which are in principle easier to desolvate.

While the active and allosteric sites were preferentially mapped at high sigma values, MixMD sites with lower sigma values could provide relevant information for SBDD. This possibly stems from the fact that we were able to identify cosolvent, cofactor, and substrate binding locations in addition to active and allosteric sites. Our interest in evaluating MixMD maps at a lower sigma value stemmed from the fact that in addition to active and allosteric sites we were able to identify cosolvent, cofactor and substrate binding locations. These results motivated us to evaluate MixMD maps at lower sigma values. For instance, when the androgen receptor MixMD maps are contoured at 35σ (Figure 3-4B), most of the sites mapped by multiple probes can be traced back to sites of biological and functional relevance. The androgen receptor is a nuclear receptor activated by nuclear receptor co-activators; one such co-activator, Nuclear Receptor Co-activator 2 (NCOA2), is shown in Figure 3-4B, including several other cosolvent and protein packing interfaces. This agreement between MixMD and the location of experimentally known sites provides additional support that MixMD properly samples the protein surface for “desolvable” sites without getting stuck in irrelevant local minima. However, our reliability in identifying binding partners from the PDB for low ranked hotspots decreased, possibly as a consequence of lower sigma value and weaker binding. Similar results were observed for other protein targets upon which MixMD was performed.

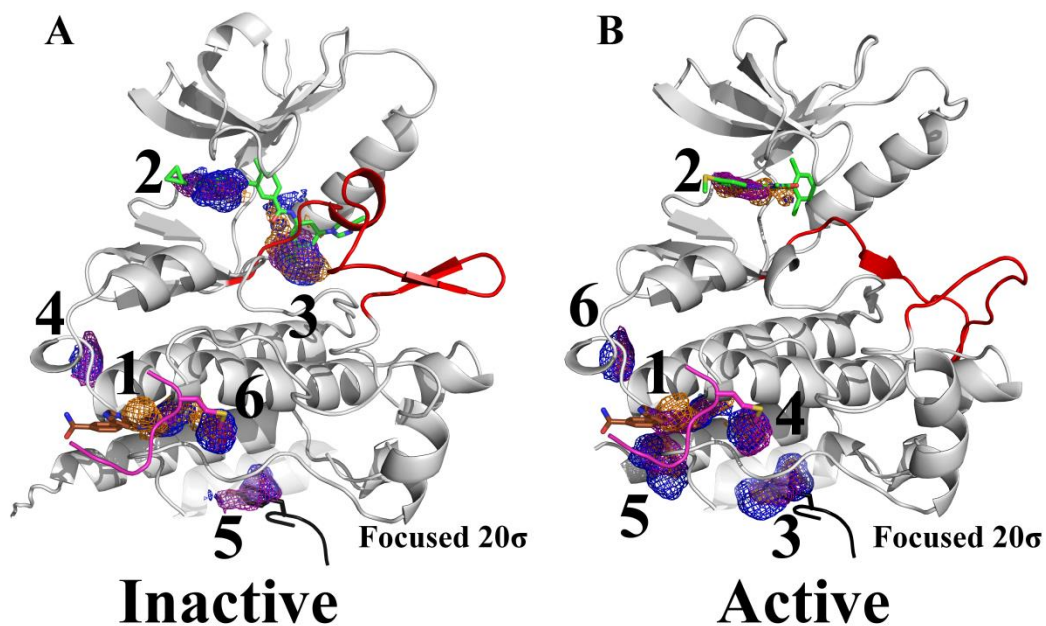


Figure 3-10 MixMD simulations were performed using the inactive (PDB ID: 3KFA, Figure 3-10A) and active (PDB ID: 1M52, Figure 3-10B) forms of Abl Kinase. The top six ranked hotspots are shown for both conformations to illustrate the rearrangement of the hotspot rankings. The allosteric and active sites were ranked first and second respectively irrespective of the conformation of ABL Kinase used for MixMD. The activation loop is colored red to show the difference in this region between the two protein conformations. The third ranked hotspot in the inactive conformation (Figure 3-10A) is now occupied by the activation loop in the active form of ABL Kinase, this leads to a rearrangement in ranking of several sites on the protein. Two protein packing interfaces are shown colored black (PDB ID: 1OPL) and pink (PDB ID: 3QRK). The allosteric site ligand shown for reference is colored brown (PDB ID: 3K5V)

3.4.6 Limits of conformational sampling with MixMD

Proteins exist in an ensemble of functionally relevant conformations and studying the effect of starting conformation on MixMD performance is imperative. In our current test set, primarily focused on proteins known to exhibit allosteric modulation, we set out to identify protein targets with multiple conformations with an established biological significance. Abl Kinase,

being a well studied system had crystal structures of both the active and inactive forms of the enzyme thereby allowing us to examine the effect of starting conformation on MixMD performance. While there was no inter-conversion between the two conformational states of Abl Kinase during the MixMD simulations, the competitive and allosteric sites were consistently mapped as the top two ranked hotspots starting from either conformation. Interestingly, the third ranked hotspot in MixMD simulations of the inactive form of Abl Kinase (Figure 3-10A) was missing in MixMD results from the active form (Figure 3-10B). Upon further examination, the reason for the absence of this hotspot became obvious. The activation loop of ABL Kinase in the active form occupies the hotspot present in the inactive form and precludes the mapping of this site by probe molecules. In addition, we have also observed rearrangement in the ranking of hotspots between the two conformations. For instance, the fourth ranked hotspot in the inactive form of Abl Kinase (Figure 3-10A) now drops to sixth rank in the active form of Abl Kinase (Figure 3-10B). In its place, a hotspot in the peptide substrate binding site takes precedence in the active form of Abl Kinase. This is in perfect agreement from the standpoint of catalytic activity, as the active form of Abl Kinase binds peptide substrates to phosphorylate them. These subtle changes in MixMD rankings starting from different conformations open up exciting prospects for the use of MixMD in understanding the functional relevance of different conformations of proteins.

3.5 Conclusion

MixMD simulations map hotspots on protein surfaces while allowing for protein flexibility and competition with water. In this study, we have successfully demonstrated the application of MixMD to several allosteric systems, identifying the active and allosteric sites within the top four sites. In addition, sites that do not correspond to active and allosteric sites represent locations of cofactor binding sites and protein multimerization/packing interfaces. While our choice of probes reflects the need to map druggable and desolvable binding sites, we have proven that one can easily extend the technique by employing a different set of probes to map charged binding sites if needed, using PTP1B as an example. We have also explored the role of protein starting conformation on MixMD, using Abl Kinase as a test case. The subtle changes in

MixMD rankings between the active and inactive conformations of Abl Kinase were found to reflect the underlying differences in the functional relevance of these protein conformations. In future studies, we intend to explore the relationship between protein starting conformation and MixMD results in further detail in order to establish the significance of these findings.

Chapter 4. Free energies and entropies of binding sites identified by MixMD simulations

4.1 Abstract

In our most recent efforts with MixMD, we were able to successfully capture the active and allosteric sites within the top-four MixMD hotspots. In this study, we describe our approach for obtaining the thermodynamic profile of the binding sites identified by MixMD. First, we establish a framework for calculating free energies from MixMD simulations. Second, we present a means to obtain a relative ranking of the binding sites by their configurational entropy. The theoretical maximum and minimum free energy and entropy values achievable under such a framework along with the limitations of the techniques are discussed. Using this approach, the free energy and relative entropy ranking of the top-four MixMD binding sites across the allosteric protein targets Ablkinase, Androgen receptor, Pdk1kinase, Farnesyl Pyrophosphate Synthase, Chk1 kinase, Glucokinase, and Protein Tyrosine Phosphatase 1B were computed and analyzed.

4.2 Introduction

Mixed-solvent simulations have gained increasing prominence with the advancements in computing power. Several such techniques have been reported in the literature (27, 32, 46, 51, 53). The ability to incorporate full protein flexibility and direct competition of organic compounds with water make these molecular dynamics methods an attractive alternative to existing approaches. For instance, docking ignores such contributions or incorporates them only to a limited extent (71). Our MixMD approach uses binary-solvent simulations of water and water-miscible, organic probes (46–48). Recently, we have applied MixMD on a test set of allosteric proteins. The application of MixMD on this test set demonstrated that one could capture the active-site and allosteric sites within the top-four sites. The success of the technique certainly suggests that MixMD holds great promise as a tool for druggability assessment. Identifying druggable binding sites is an important first step in choosing which sites on a protein surface to target. Additional information detailing each binding site would allow one to make a more informed decision on which sites to target. Thermodynamic measures such as free energy and entropy values fall in this important category. It is more straightforward to optimize enthalpy-driven binding affinity with typical SBDD scoring functions. Such considerations merit the development of techniques that can be used to obtain additional data on local thermodynamic properties. Techniques that estimate free energy of a binding site from mixed-solvent simulations have been reported by several groups (27, 29, 32, 51). Such measures have been used to predict maximal affinity of drug-like ligands, and most of the methods decompose the free energy of organic probes onto a sub-atomic grid. In this study, we demonstrate the drawbacks of making such assumptions and propose a framework for the calculation of free energies. Furthermore, efforts are made to obtain a relative ranking in terms of configurational entropies of probe molecules, using the well-established concept of entropy as a measure of the density of states (123). Such measures allow one to examine the interplay of binding site and probe structures on each other. Taken together, these studies construct and demonstrate the utility of a suite of computational techniques that one can use to fully characterize binding sites obtained from MixMD simulations.

4.3 Methods

4.3.1 Simulation of 5%v/v box of MixMD probes to obtain expected occupancies (no proteins present)

Simulations of TIP3P water (109) and 5% v/v boxes of acetonitrile, isopropanol, and pyrimidine were performed. These simulations were setup in a similar manner outlined in our earlier work on validating probe parameters (110). The 5% v/v boxes of probes and water were prepared to be $\sim 50\text{\AA} \times 50\text{\AA} \times 50\text{\AA}$ size. The boxes were simulated in AMBER12 (124) using SHAKE (112) and a time step of 1fs. Following an initial minimization, the system was gradually heated to 300K at constant volume. An initial 2ns equilibration run was followed by 20ns of constant-pressure simulation. The center of mass (CoM) of each probe's location in the last 5 ns of 10 runs were binned onto a grid of 0.5 \AA spacing, using an in-house modified version of cpptraj from AmberTools14 (125). If there were no bias by the protein, the expected occupancy per grid point is simply the number of probe molecules divided by the number of grid points. The expected occupancies for a grid point and the volume of a probe for a 5% v/v simulation are presented in Table 4-1.

Table 4-1. The Expected occupancy for a grid point and the volume of a probe are presented for the MixMD probes acetonitrile, isopropanol, and pyrimidine.

Probe	Expected Occupancy per grid point	Probe radius	Probe volume (no. of grid points)	Expected Occupancy for volume of probe
Acetonitrile	7.109e-05	2.24 \AA	47.16 \AA^3 (389)	0.002346102
Isopropanol	5.108e-05	2.54 \AA	68.74 \AA^3 (515)	0.002911845
Pyrimidine	4.683e-05	2.62 \AA	75.28 \AA^3 (619)	0.002669823

4.3.2 Deriving free energies from MixMD simulations

Free energies from MixMD simulations were derived using a process illustrated in Figure 4-1. Initially, using an in-house modified version of cpptraj module in AmberTools14, the CoM of all the probes from MixMD simulations was “binned” onto a grid of 0.5 Å spacing. MixMD simulation data from the last 5ns of 10 runs for each probe were used to perform the binning. These raw bin counts reflect the number of snapshots (amount of time) a probe molecule has spent at a particular location. The raw bin counts are then converted to occupancies by dividing the bin count at each grid point with the number of MixMD simulation snapshots that were used to obtain the initial raw bin counts.

The grid point with the highest occupancy is taken to be the center of the first probe site. The occupancy of all grid points within an enclosing sphere of the volume of the probe, centered on this grid point, are summed to determine the observed occupancy for this probe location (Figure 4-1B). In a similar manner, the next grid point with the second highest grid occupancy is taken to be the center of the second probe site. Again, the occupancy of the second site is calculated summing the grid points within the volume of the probe sphere. (Figure 4-1D). This process is iteratively repeated until all grid points are assigned to probe locations.

In order to calculate the free energies from these observed occupancies, one needs to compare them to expected occupancies in Table 4-1, using equation (1). The free energy values from equation (1) estimate the change in free energy of moving a probe molecule from the bulk into the binding-site location. A negative value for this free energy change indicates that it is more favorable for the probe molecule to be in a binding site location compared to the bulk.

$$\Delta G_{bind} = -RT \ln \left(\frac{\sum_i^{sphere} occupancy(i)}{\sum_i^{sphere} expected\ occupancy} \right) \quad (1)$$

where (i) is every grid point in the probe’s volume and the expected occupancy is constant.

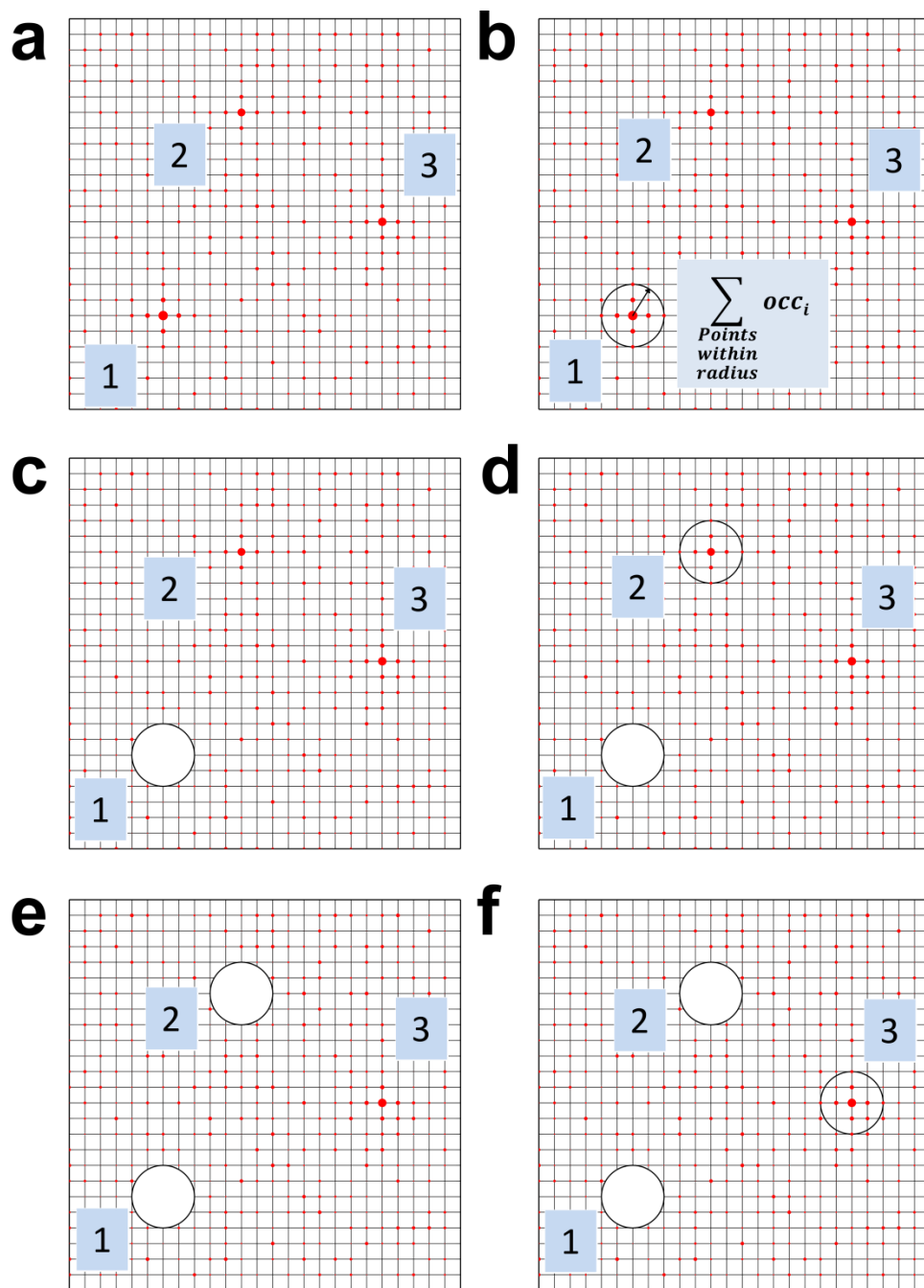


Figure 4-1. The process of obtaining observed occupancies and free energies from MixMD simulations is depicted in subfigures a-f. a) The grid points are sorted from highest to lowest occupancy, based on the counts of the probe's CoM. The size of the red circles on the grid indicates high vs low occupancies. The top-three grid points with the highest occupancies are shown for the purpose of demonstration. b) The grid point with the highest occupancy is taken to be the center of the first probe. All grid points enclosed within the volume of a probe are

added to obtain its observed occupancy. c) After processing a given probe location, the grid points associated with this probe are removed from the search process. d) The observed occupancy is calculated for the second probe centered on the next grid point with the highest occupancy. e) Upon obtaining the occupancy of the probe at this second grid point, it is removed from the search process. e) This process is continued until all the grid points are exhaustively searched and assigned to a probe location.

4.4 Results and Discussion

4.4.1 The maximum free energy of a probe is dictated by system setup

The oversimplification of obtaining free energy values using equation (1) does come with its own set of limitations which have not been highlighted in previous studies (27, 32, 51). Free energies obtained from calculations such as these are subject to the concentration of probe molecules used in the mixed-solvent simulation. This concept can be best illustrated by deriving the maximum free energy values achievable under such a framework, $\Delta G_{\text{bind}}(\text{max})$. At best, a probe molecule can occupy a given probe volume for the entire simulation, so the maximum occupancy at any particular site cannot exceed 1. Using a maximum observable occupancy of 1 and the expected occupancies for 5% v/v MixMD simulations (Table 4-1), one arrives at -2.14 kcal/mol, -2.17 kcal/mol, and -2.11 kcal/mol as the $\Delta G_{\text{bind}}(\text{max})$ for acetonitrile, isopropanol, and pyrimidine, respectively. This corresponds to $K_d(\text{max})$ of 27.7 mM, 26.3 mM, and 29.0 mM, respectively. Using a lower concentration of probe molecules within the same volume of a simulation would result in lower expected occupancies and more favorable free energies for the maximum occupancy state. Conversely, using a higher concentration of probe molecules would result in higher expected occupancies and poorer $\Delta G_{\text{bind}}(\text{max})$.

Free energy calculations using similar mixed-solvent simulations have been used by other groups to propose upper limits on the maximum achievable affinity possible for any/all drug-like molecules at a given site (27, 32, 51). Our findings call in to question, the rationale for setting an upper limit on the binding free energy for drug molecules, particularly when the

values are inherently dictated by the system setup and the concentration of probes used to perform the simulations.

A more appropriate use for such free energy measures lie in relative ranking. Even as expected occupancies increase or decrease, the relative ranking between the sites remains the same. This concept is illustrated by calculating the free energy of acetonitrile based on an observed occupancy ranging from 0.1 to 1 at increasing expected occupancy from 0.000046839 to 0.000071094 (Figure 4-2). While the magnitude of the free energy values changes, the spacing (relative ranking) between sites of different occupancy remains the same.

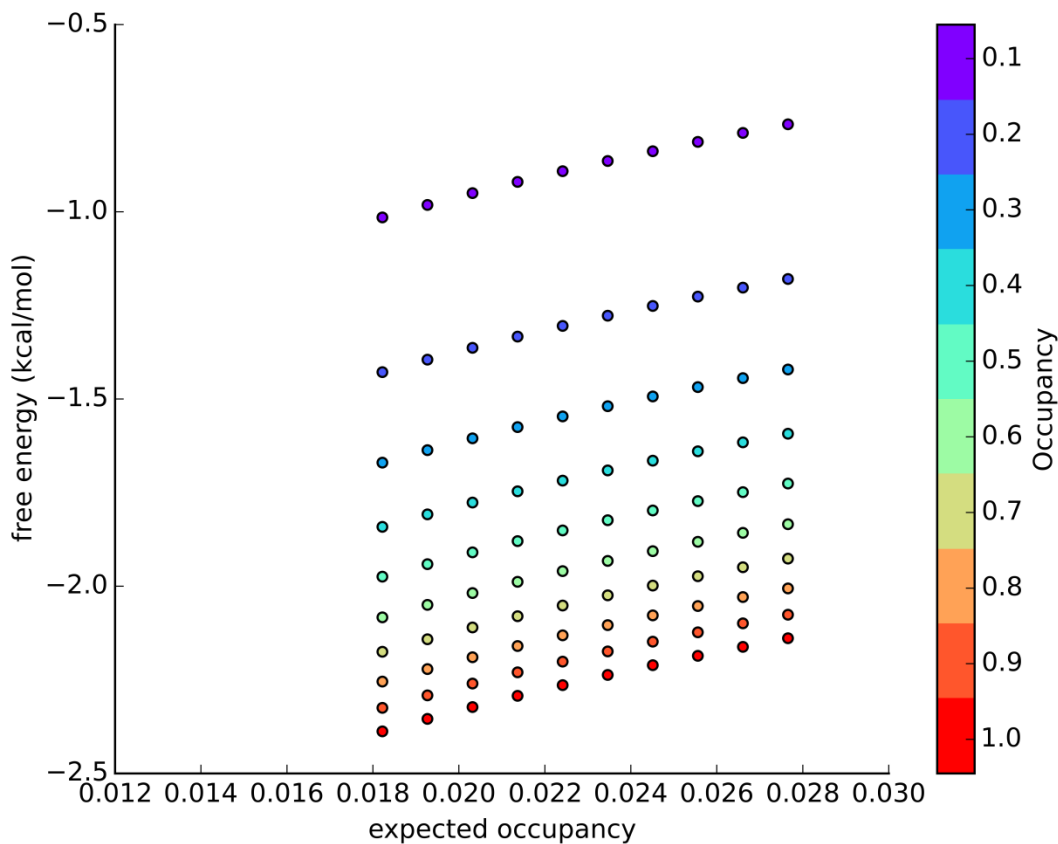


Figure 4-2. The relationship between free energy and expected occupancy is shown above using acetonitrile as an example. The free energy values for acetonitrile are calculated using equation (1) for observed occupancies ranging from 0.1 to 1 while varying the expected occupancy from 0.000046839 to 0.000071094. While the magnitude of the free energy values varies with the

expected occupancy (left to right), the difference in free energies (spacing within the columns) increase or decrease same the amount for each expected occupancy. Thus, the relative rankings between different occupancies remain the same.

4.4.2 Free energy calculations from mixed-solvent simulations

Several groups have used similar approaches for obtaining free energy changes with a ratio of observed and expected occupancies. However, the approach adopted differs from one group to another.

Barril and co-workers, in their use of isopropanol-based binary solvent simulations, calculate the binding free energy for the methyl and oxygen atoms of isopropanol separately (27). Volumes of the size of typical drug-like molecules are then created using clustering techniques by combining grid maps of the free energies for methyl and oxygen atoms of isopropanol. Using the argument that ligands of the size of drug molecules are not only involved in achieving binding affinity but also serve as a framework for the atoms to interact with the protein, the sum of the free energies of all the grid points within these drug molecule sized volumes is considered to be the maximal affinity achievable within that site/volume. Interestingly, the authors reveal that the ligand efficiencies (LE) for the methyl and oxygen groups of isopropanol frequently surpassed the limit of -1.5 kcal/mol per non-hydrogen atom observed by Kuntz and co-workers (56). However, using our method for calculating free energies, LE values never exceeded this limit (Figure 4-4). The maximum LE we found were for acetonitrile molecules at -0.65 kcal/mol per heavy atom (HA). The binding affinity of organic solvents to the protein surface is very weak, mM level, so a value like ours appears more reasonable. Acetonitrile's LE is in keeping with values desired from fragment screening.

Similarly, Mackerel and co-workers have developed "Site-Identification by Ligand Competitive Saturation" (SILCS), a cosolvent simulation technique that involves performing ternary solvent simulations of benzene, propane, and water (32). Free energies for ligands in SILCS are calculated separately for the benzene carbons, propane carbons, water hydrogens and oxygens,

using equation (1) without summing over the probe volume. The authors describe these free energies as “Grid Free Energies” (GFE). The GFE values obtained from benzene carbons correspond to interaction energies of aromatic atoms. Similarly, propane carbons, water hydrogens and oxygens correspond to aliphatic, donor, and acceptor atoms, respectively. Using these GFE values, the authors assign atom types to each ligand and evaluate a ligand's free energy by first bringing the ligand from a crystal structure into the frame of reference of a grid with these GFE values. The free energies of ligands are then computed by summing up the GFE values based on the atom type in the ligand and the corresponding GFE values on the grid.

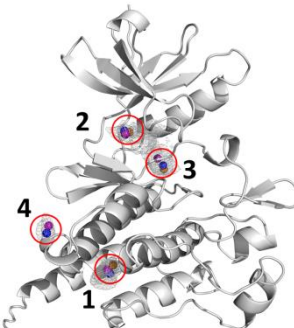
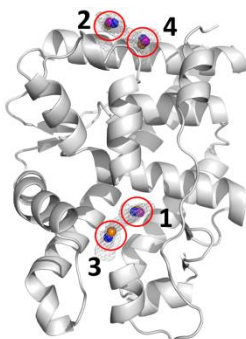
Bahar and GSK collaborators have also performed cosolvent simulations using a mixture of isopropanol, isopropyl amine, acetic acid, and acetamide. Free energies were derived from the maximum occupancy of grid points within the volume of a probe (51). Our approach for calculating free energies from MixMD simulations is along similar lines to the one proposed by Bahar and GSK collaborators, in that free energies should be calculated by taking into consideration the entire volume of a probe. We are of the firm belief that ΔG_{bind} values are the property of a whole molecule and cannot be decomposed to obtain meaningful information on the sub-atomic scale of the grids used by all approaches.

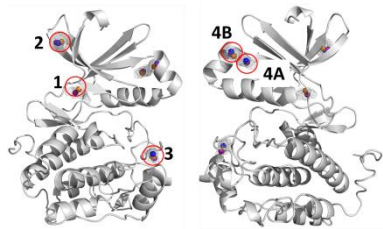
4.4.3 Free energies of MixMD binding sites calculated by occupancies

The free energies for acetonitrile, isopropanol, and pyrimidine were calculated using the aforementioned algorithm. Across all the protein targets, ΔG_{bind} for acetonitrile were lower compared to isopropanol and pyrimidine. Figure 4-3 shows the distribution of ΔG_{bind} for the top-10 probes from each binary simulation across all the protein targets. Interestingly, LE for these same probes were flipped; acetonitrile probes had higher LE. The LE for all these sites were well within the -1.5 kcal/mol limit established in a study by Kuntz and co-workers (56) and the -1.75 kcal/mol observed in our previous work (57). Using our approach, we have calculated ΔG_{bind} of the probe molecules within the active and allosteric binding sites on our test proteins. Their locations on the protein surface and their free energies are presented in Table 4-2. We have found it ideal to visualize MixMD binding sites contoured at 20σ contour using all atom binned maps; this revealed the full extent of the binding site mapped by MixMD probes.

These MixMD maps allow one to understand the all atom contacts of the probe molecules with the protein. However, our free energy calculations were performed on CoM binning. Thus we found instances where MixMD binding site accommodated multiple probes. For example in pdk1 kinase, site 4 (allosteric site) can be seen to bind multiple probes in distinct sub-sites. In the case of pdk1 kinase, site 4 was subdivided into 4A and 4B. Similar observations were made for site 1 (the allosteric site) in Glucokinase where two subsites (site 1A and 1B) could be seen.

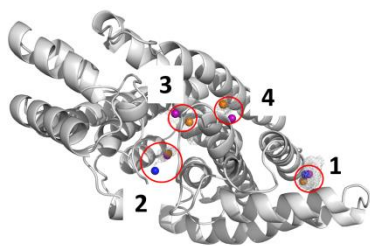
Table 4-2. The ΔG_{bind} of the probes acetonitrile, isopropanol, and pyrimidine within the top-four MixMD sites (identified using our all atom binning method) are presented for the protein targets Ablkinase, Androgen receptor, Pdk1 kinase, Farnesyl Pyrophosphate Synthase, Chk1 kinase, Glucokinase, and Protein Tyrosine Phosphatase 1B. On rare occasions, the binding site identified by MixMD accommodated more than one probe. These sites were further divided in to subsites A and B.

Protein	Site No.	Binding site Classification ^(a)	Acetonitrile (kcal/mol)	Isopropanol (kcal/mol)	Pyrimidine (kcal/mol)
 <p>Ablkinase</p>	1	A	-1.94	-1.92	-1.69
	2	C	-1.12	-1.55	-1.96
	3	C	-1.78	-2.07	-2.02
	4	O	--	-1.74	-1.82
 <p>Androgen receptor</p>	1	C	-1.68	-1.36	-1.65
	2	A	-1.47	-1.63	-1.95
	3	C	-1.46	-1.84	-1.37
	4	A	-1.46	-1.23	-1.8



Pdk1kinase

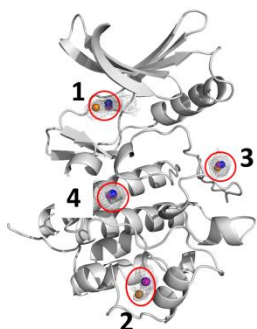
1	C	-0.85	--	-2.01
2	O	-1.69	-2.08	-1.86
3	O	-1.16	-1.75	-1.59
4A	A	-1.53	-1.51	-1.75
4B	A	-1.42	-1.78	-1.7



Farnesyl Pyrophosphate

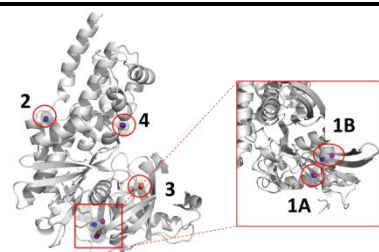
Synthase

1	O	-1.53	-1.81	-1.95
2	A	-1.43	-0.9	-1.28
3	C	-1.24	--	-1.17
4	C	-1.48	--	-0.78



Chk1kinase

1	C	-1.8	-1.62	-1.81
2	O	-1.41	-1.95	-1.96
3	O	-1.62	-1.76	-2.05
4	A	-1.85	-2.08	-2.05



Glucokinase

1A	A	-1.65	-1.86	-1.87
1B	A	--	-2.04	-1.62
2	O	-1.82	-1.78	-1.8
3	O	-1.19	--	-0.87
4	O	-1.19	-1.49	-1.6



1	O	-1.66	-2.07	-2.08
---	---	-------	-------	-------

2	A	--	-1.65	-1.83
3	O	-1.09	-1.22	-1.57
4	O	--	-1.47	-1.82

Protein Tyrosine

Phosphatase 1B

(a) Binding site classification followed C for competitive, A for allosteric, and O for the others.

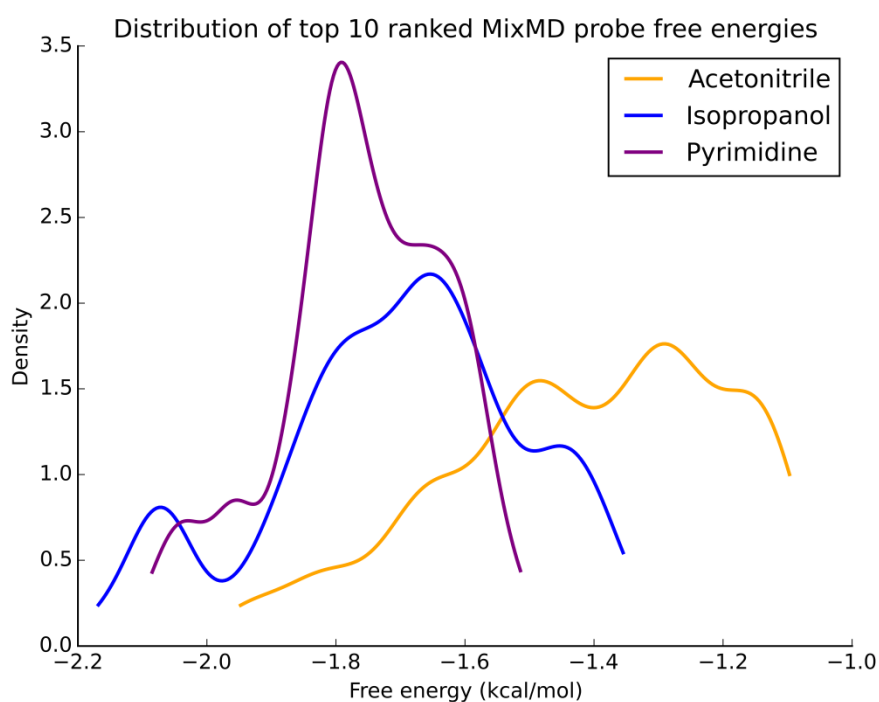


Figure 4-3. The normalized distribution profile of ΔG_{bind} for the top-10 MixMD probes is shown. Across the seven protein targets studied, binding free energies for isopropanol and pyrimidine were found to be more favorable than acetonitrile. Acetonitrile distribution is colored yellow, isopropanol distribution is colored purple, and pyrimidine distribution is colored purple.

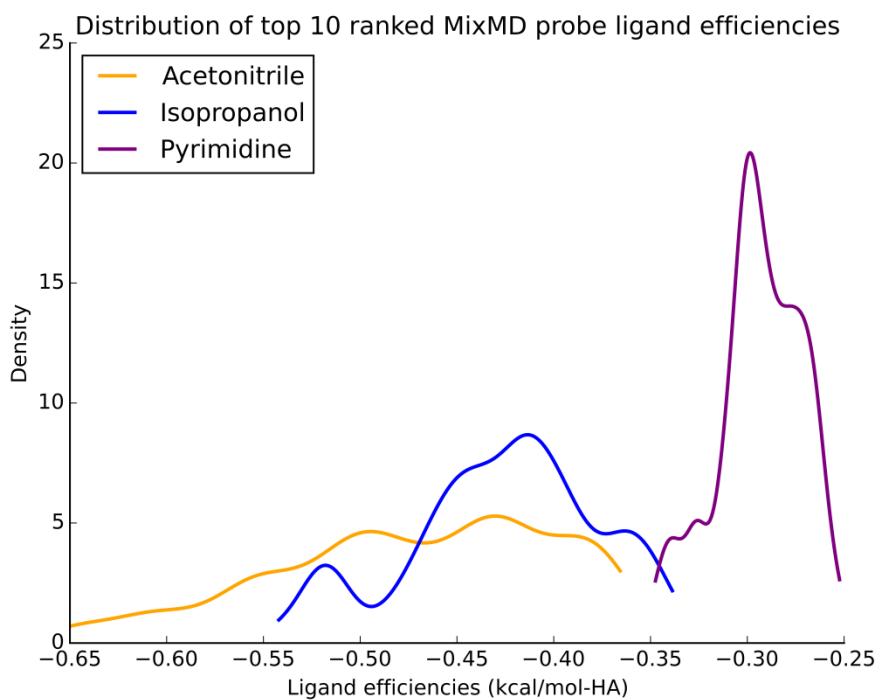


Figure 4-4. The ligand efficiencies for the top-10 probes from MixMD simulations of seven protein systems are presented in the units kcal/mol-HA. Across the seven protein targets studied, ligand efficiencies for acetonitrile were more favorable than isopropanol and pyrimidine. Acetonitrile distribution is colored yellow, isopropanol distribution is colored purple, and pyrimidine distribution is colored purple.

4.4.4 Ranking MixMD binding sites based on configurational entropy

The entropy of a probe in a site (ΔS_{site}) can be partitioned into

$$\Delta S_{\text{site}} = \Delta S_{\text{probe}} + \Delta S_{\text{trans}} \quad (2)$$

where ΔS_{probe} reflects the behavior of the probe within the site and ΔS_{trans} is the entropy of taking a probe from the freedom of occupying anywhere in the simulation box to occupying a site identified by the volume of the probe. As noted earlier, we define that site by a sphere centered at each high-occupancy point. That sphere definition is the same anywhere on the protein surface, so the translational entropy is the same for all sites. It simply reflects the

difference in the volume of the sphere vs the volume of the box: $\Delta S_{\text{trans}} = k \times \ln(\text{number of grid points in sphere}) - k \times \ln(\text{total number of grid points in the box})$. This dependence upon the box highlights that ΔS_{trans} is defined by the system setup, just like $\Delta G_{\text{bind}}(\text{max})$. However, it is commonly assumed that the value is basically the same for any probe to any protein because it just reflects translation of the CoM.

In calculating the difference in entropy between the sites ($\Delta\Delta S_{\text{site}}$), the ΔS_{trans} term cancels. The interesting comparison lies in the other $3N-3$ degrees of freedom sampled by the probe's atoms. While molecules in the bulk rotate freely, interactions with the protein impart a level of structure, limiting the probe's freedom. ΔS_{probe} is the difference between a probe evenly and freely sampling the sphere, $S_{\text{probe}}(\text{max})$, to the actual translational and rotational behavior of the probe seen during the simulations, S_{probe} . Here, we draw upon the concept of entropy as the density of states and use our grid points as shown in Figure 4-5. To simplify the analysis, we decomposed the probe into its non-hydrogen atoms and used the same binning routine from calculating free energies to count the atomic occupancies on the grid points in the sphere. Entropy of the probe is calculated using the Gibbs-Shannon equation (126), shown in equation (3). The probability of finding an atom at a particular grid point is determined by equation (4). The entropy measures obtained for each heavy atom are then combined as shown in equation (5) to give S_{probe} .

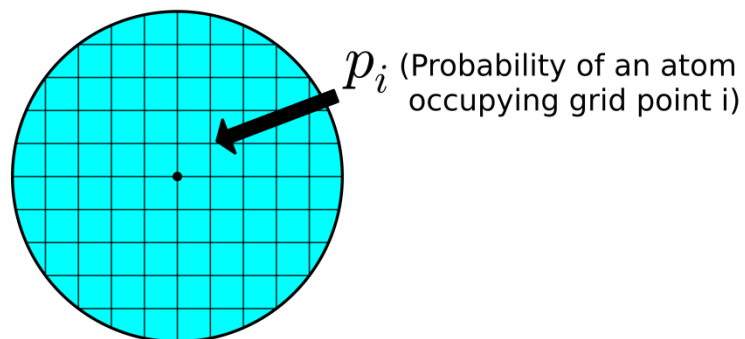


Figure 4-5. The concept of entropy as the density of states is applied within the volume of a probe sphere. Each grid point within the volume is considered a state. The probability of each state (p_i) for each heavy atom is calculated using equation (4).

$$S = -k \int p \times \ln(p) \quad (3)$$

$$p_i(HA) = \frac{\text{occupancy of HA at grid point } i}{\sum_j^{\text{sphere}} \text{occupancy of HA at grid point } j} \quad (4)$$

$$S_{HA} = -R \sum_i^{\text{sphere}} p_i(HA) \times \ln[p_i(HA)] \quad (5)$$

$$S_{\text{probe}} = \sum_{\text{heavy atoms}} S_{HA} \quad (6)$$

Under no constraint while freely exploring the box in the bulk solvent, each grid point is equally occupied, and one can establish an upper limit of entropy achievable within the volume of a probe. The $S_{\text{probe}}(\text{max})$ values possible under our framework are presented in equations (7) and (8) and listed for acetonitrile, isopropanol, and pyrimidine in Table 4-3. This maximal value may

be an over-estimate because the chemical structure of the probe imparts an inherent bias to sampling the grid. However, this inherent bias is the same in all sites; furthermore, $S_{probe}(max)$ drops out when calculating the difference between the sites, $\Delta\Delta S_{site}$.

$$S_{HA}(max) = -R \sum_i^{sphere} p_{bulk} \times \ln(p_{bulk}) \quad (7)$$

$$S_{probe}(max) = \sum_{heavy\ atoms} S_{HA}(max) \quad (8)$$

Table 4-3. Maximum entropy at 300K (in kcal/mol) for a freely rotating and translating probe molecule is calculated. Under such conditions every grid point within the volume of a probe will be occupied with equal probability (p_{bulk}).

Probe	No. of grid points in volume of probe (gpt)	p_{bulk} 1/(gpt)	- $TS_{probe}(max)$ (kcal/mol at 300K)
Acetonitrile	389	0.0025706940874	11.497
Isopropanol	515	0.00194174757282	15.329
Pyrimidine	619	0.0016155088853	22.993

4.4.5 Entropies across MixMD binding sites

In order to compare the configurational entropy of MixMD binding sites, we have computed the change in entropy of moving a probe molecule from the bulk into each binding site sphere. As one would expect, moving a freely rotating probe in the bulk to a binding site decreases the entropy and thus one should observe that such a change is unfavorable (but compensated by enthalpic gain). We have confirmed this behavior by computing the $-T\Delta S_{probe}$ for the top-50 probes ranked by free energy in all the allosteric protein systems which we simulated in MixMD. The distribution of $-T\Delta S_{probe}$ for the probes acetonitrile, isopropanol, and pyrimidine are shown in Figure 4-6. The high peaks close to zero show that many probe molecules tumble

close to the bulk behavior. Most importantly, none of the entropy changes are less than zero; this confirms our assumption that none of the probe molecules exceed the maximal entropy we have calculated in previous sections (Table 4-3).

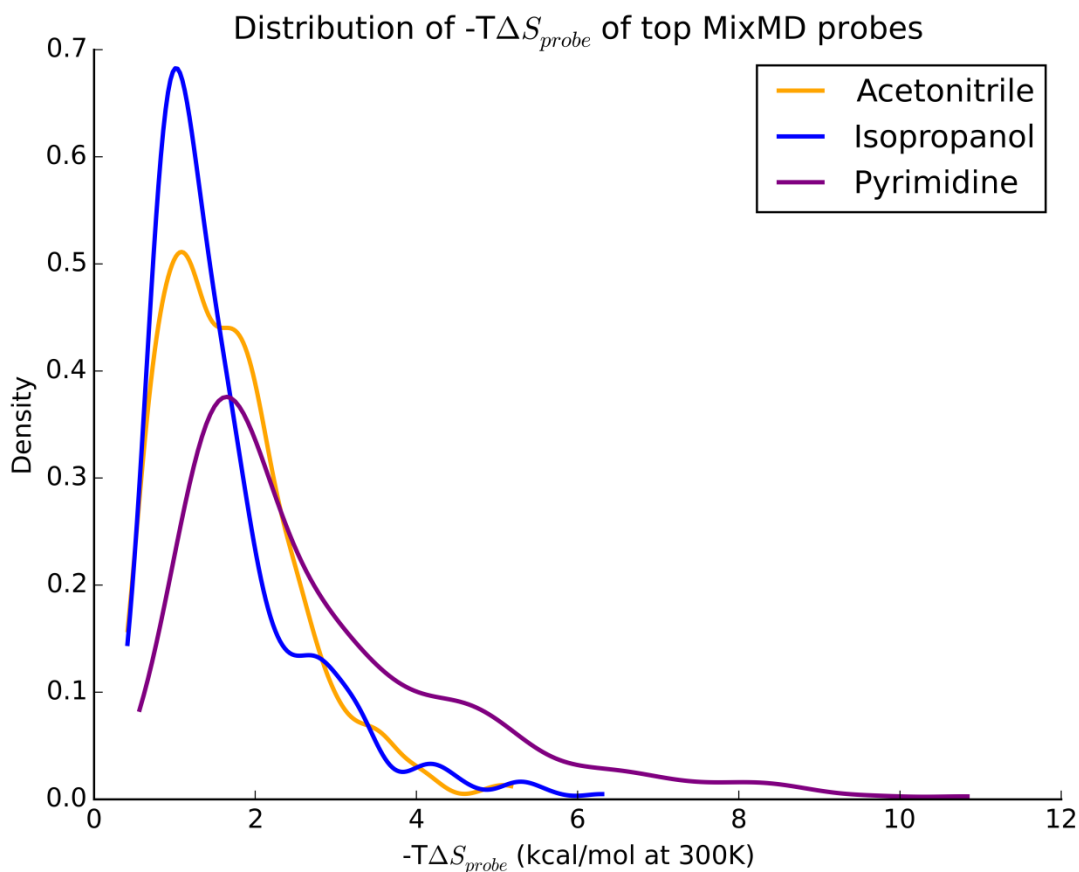


Figure 4-6. The distribution of $-T\Delta S_{probe}$ for the top-50 MixMD probes ranked by free energy are presented for acetonitrile (orange), isopropanol (blue), and pyrimidine (purple). As expected, moving from the bulk into the binding site where the probes are restricted is unfavorable, thus $-T\Delta S_{probe}$ values are positive.

4.4.6 Validating configurational entropies obtained from MixMD binding sites

Entropies measured using our approach report upon the local thermodynamic environment of an individual probe molecule and as such cannot be verified using experiments. It is important

to note that an experimental measure of a binding event also reflects the entropic costs paid by the protein and the reordering of the water around the binding site. While the effect on the protein may be partially observed from the order seen for the probes, the effects on water are very hard to estimate. More importantly, very subtle changes to ligands can result in significant and unexpected changes in water as the work of Klebe shows (127). It is unreasonable to assume that the water's behavior around the solvent probes is a good estimate of their behavior in the presence of a drug-like ligand.

Despite these limitations, these measures describe the structure/order of the probe's conformational sampling within the binding site, and one can in principle visualize the occupancies of the HA of the probe molecules to validate these findings. When visualizing the occupancies of the probe's HA, it is important to normalize the HA density within the volume of the probe, as we do in equation (3). This is necessary because, raw bin counts not only reflect upon the positional preference of a probe, but also on the duration a probe molecule has spent its time at a given location. By normalizing the occupancies to give densities of HA within the binding site sphere, one can separate the information needed for ΔG_{bind} to reflect each HA's contribution to ΔS_{probe} and analyze the density for any probe's configurational sampling within their binding site sphere. We have assessed this important metric using $-\Delta S_{\text{probe}}$ calculated for all the systems and MixMD probes used on our earlier study. The minimum, median, and maximum $-\Delta S_{\text{probe}}$ are presented for the probes acetonitrile, isopropanol, and pyrimidine in Table 4-4.

Table 4-4. The change in configurational entropy when moving a co-solvent from the bulk to the protein binding site were calculated using the top-fifty MixMD probes ranked by free energy from all seven allosteric systems. The minimum, median and maximum $-\Delta S_{\text{probe}}$ in this dataset was reported for each probe at 300K. The proteins to which these values belong along with the rank of the probe according to free energy are provided in brackets.

Probe	Minimum $-\Delta S_{\text{probe}}$ (kcal/mol at 300K)	Median $-\Delta S_{\text{probe}}$ (kcal/mol at 300K)	Maximum $-\Delta S_{\text{probe}}$ (kcal/mol at 300K)
Acetonitrile	0.42 (Ablkinase, rank 19)	1.56 (Pdk1kinase, rank 43)	5.18 (Glucokinase, rank 32)
Isopropanol	0.42 (FPPS, rank 3)	1.3 (Androgen receptor, rank 38)	6.31 (Glucokinase, rank 31)
Pyrimidine	0.57 (FPPS, rank 4)	2.18 (Androgen receptor, rank 32)	10.84 (FPPS, rank 43)

In order to make a proper comparison across the minimum, median, and maximum $-\Delta S_{\text{probe}}$, we have visualized the population density of each HA in the probe molecule at a contour level of 0.5% of the population in the binding site. In the case of acetonitrile, these densities are shown in Figure 4-7. The density of the nitrogen atom of acetonitrile is colored blue, whereas the densities of the central and terminal carbons of acetonitrile are colored cyan and brown. The CoM that defines the binding site of the probe molecule is shown as an orange colored sphere for reference. The maximum $-\Delta S_{\text{probe}}$ represents the most unfavorable transfer from the bulk to the protein binding site. As expected, in Figure 4-7A, the densities of the three atoms within the acetonitrile probe molecule are clearly visible at the atomic level. This demonstrates the restriction on the probe when bound to the site. When the density of the probe with the median $-\Delta S_{\text{probe}}$ is visualized in Figure 4-7B, one sees a lesser degree of structure. Clearly, the acetonitrile molecule is oriented with its nitrogen pointing up like the example in Figure 4-7A,

but some freedom is seen in the lateral movement. Figure 4-7C shows the probe with the minimum $-\Delta S_{\text{probe}}$ observed for acetonitrile, where the HA density around the CoM is disperse and overlapping. This is consistent with the idea that a low $-\Delta S_{\text{probe}}$ probe in this location is similar to the bulk environment and thus is freely rotating. In going from maximum to minimum $-\Delta S_{\text{probe}}$, there is a trend of decreasing structure/order of the probe molecules seen when visualizing the HA density. This is consistent with our theory.

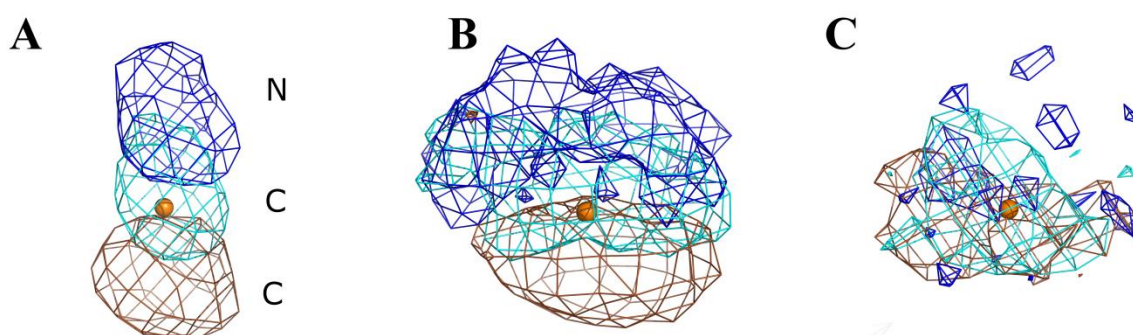


Figure 4-7. Acetonitrile HA densities are presented for the maximum, median, and minimum entropies reported in **Table 4-4**. The CoM that defines the binding site of the acetonitrile probe is shown as an orange sphere for reference. The normalized occupancies of all the atoms in A, B, and C are contoured at 0.005. The density of nitrogen atoms is colored blue, the density of the carbon atom in the middle of acetonitrile is colored cyan, and the density of the terminal carbon is colored brown. A) The $-\Delta S_{\text{probe}}$ is at a maximum, making this the most constrained probe in our dataset. Consequently, all atoms of the acetonitrile probe can be clearly seen in this case. B) The acetonitrile with the median $-\Delta S_{\text{probe}}$ shows some structure in the configurational sampling but also some latitude. C) Density for the case of minimum $-\Delta S_{\text{probe}}$ shows that the probe molecule at this location is freely rotating, and is close to the entropy of the bulk. As a result, the density is smeared out and overlapping.

Similar trends were observed for isopropanol. When the density of the probe molecule with the maximum $-T\Delta S_{\text{probe}}$ (Figure 4-8A) was visualized clear, structured density could be seen. The densities at the median $-T\Delta S_{\text{probe}}$ (Figure 4-8B) clearly show two conformations with the hydroxyl oxygen sampling between two hydrogen-bonding interactions. The minimum $-T\Delta S_{\text{probe}}$ (Figure 4-8C) follows similar trends as seen for acetonitrile. The same results were obtained for pyrimidine, where visualization of the densities for the maximum, median, and minimum $-T\Delta S_{\text{probe}}$ followed the established trend of increasing structure/order in the probe molecules.

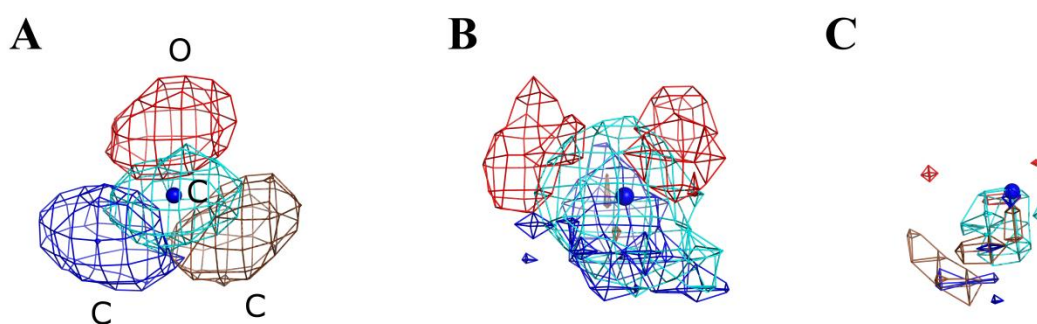


Figure 4-8. Normalized HA occupancies of isopropanol are presented for the maximum, median, and minimum entropies reported in **Table 4-4**. The CoM that defines the binding site of the isopropanol probe is shown as a blue sphere for reference. The density of all the atoms in A, B, and C are contoured at 0.005. The density of oxygen atoms is colored red, the density of the central carbon is colored cyan, and the two terminal carbons are colored blue and brown. A) The maximum $-T\Delta S_{\text{probe}}$ example is the most constrained probe in our dataset. Consequently, all atoms of the isopropanol probe can be clearly seen in this case. B) The $-T\Delta S_{\text{probe}}$ in this case is at the median of all processed sites, there is some structure in the probe molecule. Notably, the hydroxyl oxygen is sampling two hydrogen-bonding interactions. C) For the case of minimum $-T\Delta S_{\text{probe}}$, the molecule at this location is freely rotating, and is close to the entropy of the bulk. As a result, the density is smeared out and can only be seen partly.

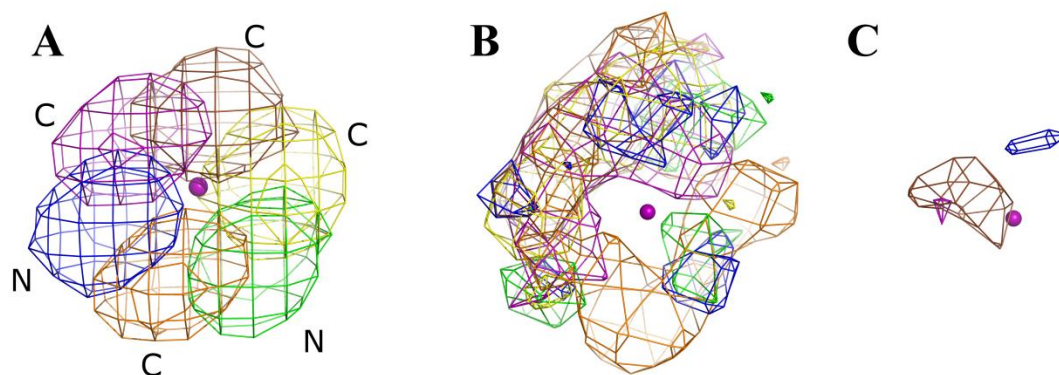


Figure 4-9. Pyrimidine per probe normalized density is presented for the maximum, median, and minimum entropies reported in **Table 4-4**. The CoM that defines the binding site of the pyrimidine probe is shown as a purple sphere for reference. The normalized occupancies of all the atoms in A, B, and C are contoured at 0.005. The density of two nitrogen atoms are blue and green, whereas the density of carbon atoms is colored brown, purple, yellow, and orange. A) In the maximum $-T\Delta S_{\text{probe}}$ case, the molecule is very constrained. Consequently, all atoms of the pyrimidine probe can be clearly seen in this case. B) The $-T\Delta S_{\text{probe}}$ in this case is at the median of all processed sites, there is some structure in the probe molecule. Notably, the molecule is rotating and giving HA densities with a torus shape. It appears that the nitrogens are sampling three locations, separated by roughly 120° . C) For the minimum $-T\Delta S_{\text{probe}}$ case, the probe molecule at this location is freely rotating, and is close to the entropy of the bulk. As a result, the density is smeared out and cannot be seen.

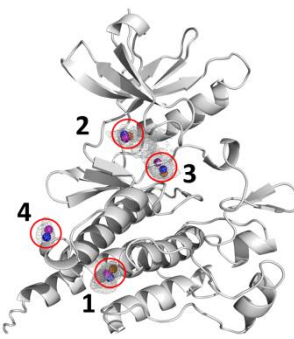
4.4.7 Comparison of entropies across protein targets

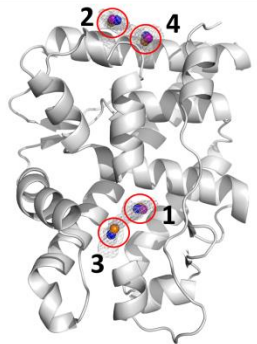
We wanted to specifically examine the probes within the top four MixMD binding sites, and the results are shown in Table 4-6. As expected, in general pyrimidine probe molecules were more restricted in their motions. Visual inspection across all results suggested that using a cutoff of 3 kcal/mol for $-T\Delta S$ is ideal to comment upon whether sites displayed a strong configurational bias or not. Using this metric, only a few acetonitrile and isopropanol molecules were identified. In Ablkinase, site 2 in the active site displayed conformational bias for all the probes.

In Androgen receptor, pyrimidine was the only probe with any $-\Delta S_{\text{probe}}$ above the 3 kcal/mol cutoff. In Chk1 kinase, higher $-\Delta S_{\text{probe}}$ was seen for the active-site (site 1) and peptide-substrate binding site (site 2). In FPPS, pyrimidine probe molecules were significantly ordered in sites 3 and 4 which are part of the active site. In Glucokinase, the pyrimidine in the ATP binding site (site 4) displayed significant order. One could speculate that this specificity in binding may be related to the structural features shared between pyrimidine and adenine. In Pdk1 kinase, with the exception of site 1, most of the sites did not exhibit any configurational preference. In PTP1B, site 3 and 4 which are at the protein-packing interface displayed high configurational bias for bound pyrimidine.

We have also investigated whether kinases that bind ATP display configurational specificity for pyrimidine, as it closely resembles the adenine ring of ATP. In Ablkinase site 1, Pdk1 kinase site 1, and Chk1 kinase site 1, an ATP molecule binds to perform the phosphorylation function of these proteins. In all these cases, a pyrimidine molecule binds with higher $-\Delta S_{\text{probe}}$. In fact, pyrimidine's entropic penalty was on average ~ 2 kcal/mol higher than that for acetonitrile and isopropanol for these sites.

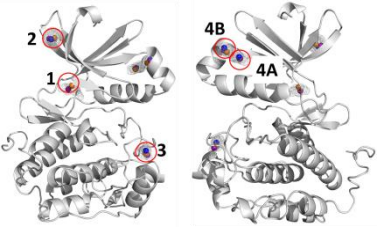
Table 4-5 The entropic penalties ($-\Delta S_{\text{probe}}$) of the MixMD binding sites are computed at 300K and are presented for the top-four MixMD binding sites.

Protein	Site No.	Binding site Classification ^(a)	Acetonitrile (kcal/mol)	Isopropanol (kcal/mol)	Pyrimidine (kcal/mol)
 <p>Ablkinase</p>	1	A	1.47	1.58	3.12
	2	C	2.45	2.91	3.72
	3	C	0.95	1.39	1.89
	4	O	--	1.07	1.74



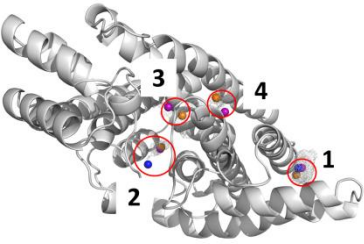
1	C	1.99	2.66	4.54
2	A	1.36	0.97	2.65
3	C	1.68	2.58	3.95
4	A	2.0	2.2	4.12

Androgen receptor



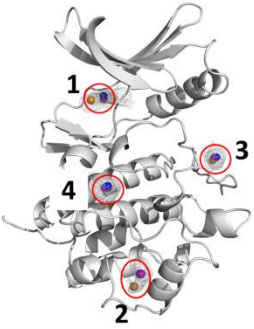
1	C	3.4	--	5.39
2	O	0.64	1.09	1.45
3	O	1.45	1.8	2.86
4A	A	0.92	0.93	1.19
4B	A	0.69	0.75	1.04

Pdk1kinase



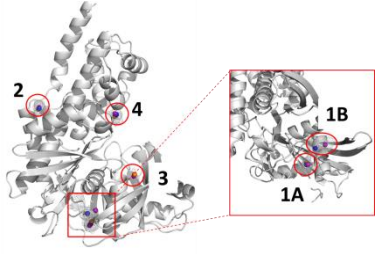
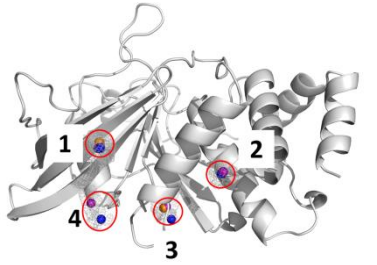
1	O	1.3	1.85	1.73
2	A	2.0	2.66	4.58
3	C	2.09	--	8.14
4	C	1.61	--	5.4

Farnesyl Pyrophosphate Synthase



1	C	2.01	1.76	3.61
2	O	2.53	4.08	3.44
3	O	1.31	1.11	1.87
4	A	1.03	1.15	1.38

Chk1kinase

	1A	A	1.66	2.22	2.95
	1B	A	--	1.71	1.72
	2	O	1.72	1.6	2.29
	3	O	3.74	--	6.49
	4	O	1.87	1.67	3.99
Glucokinase					
	1	O	1.68	1.82	2.46
	2	A	--	2.44	3.15
	3	O	2.79	2.39	8.3
	4	O	--	2.78	4.64
	Protein Tyrosine Phosphatase 1B				

(a) Binding site classification followed C for competitive, A for allosteric, and O for the others.

4.4.8 Conclusion

We have established a means of obtaining the free energy and entropy rankings based on MixMD simulations. The limitations of the free energy calculations were demonstrated. These limitations are universal to co-solvent MD simulations, and they call in to question other groups' rationale for trying to use co-solvent grids to establish a maximal free energy achievable for any/all drug-like molecules. Furthermore, a framework for calculating entropies is proposed and validated. In particular, we note that the entropies are only for the probe, not the whole system. The entropic effects on reordering water around protein-ligand complexes are very hard to estimate, and very subtle changes to ligands can result in significant and unexpected changes in water. It is unreasonable to assume that the water's behavior around the solvent probes is a good estimate of their behavior in the presence of a drug-like ligand.

4.5 Supplementary Information

The script for calculating a relative ranking of MixMD binding sites based on free energies and entropies are provided in Appendix C and Appendix D respectively.

Chapter 5. Hitting an Undruggable Target: A Blinded Test of MixMD on Heat Shock Protein 27

5.1 Abstract

The increasing availability of structural data at early stages of the drug discovery process, coupled with a lack of protein information to guide the design of drug molecules, has presented a significant challenge in the application of structure based drug design (SBDD). Mixed solvent molecular dynamics (MixMD) is a binary-solvent molecular dynamics technique designed to overcome these challenges while allowing full protein flexibility and explicit competition with water. In our most recent study with MixMD, we have established a protocol for identifying druggable binding sites on the protein surface using MixMD simulations with acetonitrile, isopropanol, and pyrimidine as probes. Here, we present the first successful blinded application of any mixed-solvent molecular dynamics method. MixMD simulations indicated the presence of druggable binding sites on the “undruggable” Heat Shock Protein 27 (Hsp27). Using a combination of NMR and high throughput screening studies; we have successfully verified our predicted binding sites. Furthermore, a direct comparison is made between MixMD results and NMR chemical shift data for Hsp27 in the presence of co-solvents. A striking level of agreement was found. Taken together, these studies demonstrate the utility of MixMD in SBDD.

5.2 Introduction

Genomics has given us many potential new drug targets. All targets cannot be pursued because of the high cost of bringing a drug to the market. Assessing a protein's druggability has gained increasing prominence. The perils of achieving potency by increasing the molecular size are well documented. Such attempts fail from a myriad of possible reasons including unfavorable physicochemical properties (128). Experimental techniques that assess a protein's druggability abound, like high throughput screening or NMR-based fragment screening (129). Computational techniques that provide similar measures would present an alternative, cost-effective approach to assessing such properties. Currently reported computational techniques that assess this important measure rely on static structures of protein (130–133). However with our increasing understanding of protein structure and function, including protein flexibility and competition with water are important components to include. To overcome these challenges, we and others have been developing computational chemistry techniques inspired by the seminal contribution by Ringe and coworkers involving multiple solvent crystal structures (MSCS) (134, 135). The MSCS method involves solving crystal structures of proteins with a variety of probe molecules. The locations on the protein surface that bind several different probes are considered to be "hotspots" that contribute disproportionately well to the affinity of a ligand. However, applying MSCS to every protein is not possible because many protein crystals are not amenable. Cosolvent MD simulations aim to provide the same information through computer modelling (103, 136–140). Some hallmarks of our approach are an emphasis on using water-miscible organic probe molecules, rigorous analysis and application of force field parameters for probes, and application of the technique with minimum to no reliance on prior knowledge of the system. Our mixed solvent molecular dynamics computational approach (MixMD) relies on performing binary solvent simulations of protein with water and organic, water-miscible probes (103, 136).

Our early work on MixMD focused on recapturing the location of organic probe molecules found on the protein surface in MSCS experiments (103, 136). We have further optimized the technique by decreasing the concentration of probe molecules from an initial concentration of

50% v/v to 5% v/v and accelerating the sampling by optimizing the MixMD simulation setup. In our most recent application of MixMD, we have established and optimized a protocol for identifying druggable binding sites on proteins using MixMD with acetonitrile, isopropanol, and pyrimidine as our water-miscible, organic probes. This approach was successful in capturing the active and allosteric sites on the proteins with a high signal to noise ratio. In addition to this successful application, we have discovered that sites that do not correspond to either active or allosteric site represented binding sites that have some form of biological/functional significance.

Until now, studies involving MixMD and similar techniques have only been tested in a retrospective manner. Such studies are important as they allow one to lay a strong foundation and provide means of optimizing protocols to decrease false positives. However in a real world setting, the location of druggable sites is unknown, especially for non-enzyme systems. Applying computational techniques where very little is known about the protein target is more challenging and realistic. Here, to the best of our knowledge, we provide the first successful application of such MD-based approaches in a blinded fashion with Heat shock protein-27 (Hsp27) as the test case. We have raised the question whether binding sites that display characteristics one would associate with druggability exist on the surface of Hsp27.

Hsp27 is a 27 kiloDalton protein that consists of a variable N- and C-terminal regions with a conserved alpha crystalline domain (ACD) in the middle. Hsp27 functions through a non-ATP-dependent chaperone mechanism and assists in the refolding of misfolded proteins. It is implicated in several diseases such as cancer, myocardial infarction, and cerebral ischemia (141–143). Hsp27 negatively regulates the apoptotic mechanism and is considered to be a potential anticancer drug target (144, 145). Mutations in Hsp27 have also been linked with Charcot-Marie-Tooth disease. Interestingly, most mutations causing this disease are located in the ACD region (146). The ACD region is primarily composed of beta sheets that provide a dimerization interface. These dimers form the building blocks for higher-order oligomers of varying sizes that mediate the biological functions of Hsp27 (141).

Current literature presents a very discouraging picture for targeting Hsp27 with small molecules. Having been a well-studied system for a long time, the only therapy being investigated under clinical trials is Apatorsen (also known as OGX-427) (147, 148). Apatorsen is an antisense oligonucleotide that functions indirectly by decreasing the production of Hsp27 as a result of silencing mRNA and is being pursued as an anticancer drug (147). Under such circumstances it is unclear if adequate efforts have not been invested in obtaining small molecules or if Hsp27 is indeed undruggable.

In this study, we focus our attention on the ACD region of Hsp27. We have applied MixMD using the NMR structure of the ACD domain of Hsp27 as our only guiding piece of information. We have applied our established protocol for identifying druggable binding sites on the protein surface. This was achieved using MixMD simulations of Hsp27 with acetonitrile, isopropanol, and pyrimidine as our water-miscible probes. Below, we describe the binding sites and the experimental data that supports our assertion that they are druggable.

5.3 Methods

5.3.1 MixMD simulation setup

The simulations were started from an NMR structure of Hsp27 ACD structure determined by Klevit and coworkers (unpublished data). This domain spanned from residues 80 to 178, based on the numbering of the full length protein. A 9Å layer of probe molecules was added around the protein using tleap in Amber Tools followed by the addition of TIP3P (109) water molecules necessary to create a 5% v/v ratio of probe to water. The force field parameters for the probes were obtained from our previous work (48) and simulations were carried out in AMBER 12 (124) using the FF99SB force field (111). The SHAKE algorithm (112) was used to restrain bonds to hydrogen atoms, and a time step of 2 fs was used to integrate the equations of motion. Particle Mesh Ewald approximation as implemented for the GPUs, PMEMCUDA (113) was used. Non-bonded interactions were limited to a 10 Å cutoff, and an Anderson Thermostat (149) was used to maintain temperature at 303 K. The simulations were carried out at 303 K to closely simulate conditions used in the experiments. Hsp27 was subjected to an equilibration protocol

to gradually increase the temperature and allow proper relaxation of all the atoms in the system. Using this approach, ten simulations of 20ns each were performed with Hsp27 and Acetonitrile, Isopropanol, and Pyrimidine separately. A total of 600ns of sampling was obtained.

5.3.2 Analyzing MixMD results

The last five nanoseconds of the ten runs for each probe (the final 50 ns) were analyzed. The location of all atoms in the probe were binned onto a $0.5 \text{ \AA} \times 0.5 \text{ \AA} \times 0.5 \text{ \AA}$ grid centered on the protein using the Ptraj (125) module of Amber Tools. The binned data was then converted to sigma values by subtracting the mean grid value from each grid point and dividing it by the standard deviation of all the grid points. This approach allowed us to visualize the propensity of a probe binding at a particular location akin to electron density. A high sigma value denotes heavily occupied sites. In order to assess different binding sites, we overlaid sigma maps from all three MixMD probe simulations. The binding sites were ranked using the same algorithm that successfully identified active and allosteric sites in our previous study. In short, this algorithm entails the examination of MixMD maps starting at high sigma value followed by subsequent dialing down of the sigma value. The sites that appear earlier during this process are the ones that are ranked higher. In addition to this criterion, we also required sites to be mapped by more than one probe.

5.3.3 Deriving free energies from MixMD simulations

The center of mass of all the probes was binned onto a grid of 0.5 \AA spacing using an in-house version of cpptraj. The binned data was converted to free energies for each probe. In this procedure, the grid points in the binned simulation data were sorted from highest to lowest based on grid count. Each local maximal grid point was the center of an enclosing sphere of the volume of the probe. The grid count was added for all the grid points to within the sphere to obtain the observed occupancy for the probe molecule. The observed occupancy was converted to free energy using equation (1) wherein the natural logarithm of the ratio of the observed and expected occupancy is multiplied by the Boltzmann constant and temperature. Expected

occupancies represent occupancies one would observe in the absence of the protein; these were obtained from our previous study (Chapter 4). Expected occupancy values used for calculating free energies were 0.002346102, 0.002911845, 0.002669823 for acetonitrile, isopropanol, and pyrimidine respectively. Since Hsp27 is a dimer and NMR studies do not distinguish between the two monomers, we have chosen to average the free energies across the monomers.

$$\Delta G_{bind} = -RT \ln \left(\frac{\sum_i \text{occupancy}(i)}{\sum_i \text{expected occupancy}} \right) \quad (1)$$

where i is the count of grid points within the probe's volume,

occupancy(i) is the value of each grid point, and expected occupancy is a constant.

5.3.4 NMR HSQC experiments with Hsp27

Our collaborators, Jason Gestwicki, Leah Makley, Rachel Klevit, and Ponni Rajagopal conducted the NMR experiments. HSQC spectra were acquired at 30°C on a 600 MHz Bruker Avance III spectrometer equipped with cryoprobe, running Topspin version 2.1, or a Bruker DRX500 with a QCI Z, axis gradient cryoprobe, running Topspin version 1.3. Spectra were acquired on samples containing 150-200 μ M Hsp27 core domain in 50 mM NaPi, pH 7.5, 100 mM NaCl at 30°C and always compared to solvent, controlled reference spectra. A 2% v/v ratio of acetonitrile, isopropanol, and pyrimidine were used. 256 scans were acquired per t1 value and spectral widths of 1500 Hz and 9615 Hz were used in the 1H and 15N dimensions, respectively. Processing and spectral visualization was performed using Sparky (150) and rNMR (151).

5.4 Results and Discussion

5.4.1 Mapping binding sites on HSP27 using MixMD

Binding sites on the Hsp27 protein surface were identified by viewing MixMD maps from acetonitrile, isopropanol, and pyrimidine simulations. Sites with the highest occupancies on the grid maps were the focus of our analysis. Each probe type is analyzed separately. After analysis,

the grids are overlaid for the druggability analysis. Binding sites identified by MixMD were required to be mapped by more than one type of probe. MixMD simulations revealed the presence of six potential binding sites on the protein surface of Hsp27. These six sites are shown in Figure 5-1 and are named sites 1, 2, 3, 4, 5, and 6 for convenience. The binding site identification process is depicted in Figure F-1. While there are minor differences in the maps generated from the three different probes, for the most part there is a striking agreement between the three probes and their preference for binding in these three sites. We have contoured the MixMD maps at 20σ in Figure 5-1 based on our earlier work that showed that this contour value allowed one to adequately visualize the binding sites on the protein surface. Moreover, the mapping of the binding sites appears consistently in both monomers of Hsp27 which supports adequate sampling in the simulations.

To compare and contrast the MixMD binding sites, we have computed the free energies from MixMD simulations. Across the three MixMD binding sites, acetonitrile bound weakly to the protein compared to isopropanol and pyrimidine. It is important to note that while acetonitrile free energies were relatively less favorable, its ligand efficiency were comparable to those of other probes Table 5-1. These ligand efficiencies were also within the theoretical limit of -1.5 kcal/mol established by Kuntz and co-workers (56) and our limit of -1.75 kcal/mol described in a recent work (57). MixMD simulations when contoured at 20σ revealed the extent of the binding sites. Our results from previous free energy calculations show that some MixMD binding sites (identified using all atom binned data) can bind multiple probes. MixMD contour plots do not allow one to examine binding sites at this level of detail. Thus such free energy calculations complement our method for detecting binding sites on proteins. For example, site 2 in Hsp27 could be seen bound to two molecules of acetonitrile (Figure 5-2B) one of which had a free energy of -1.08 kcal/mol whereas the other bound to the protein at -0.79 kcal/mol.

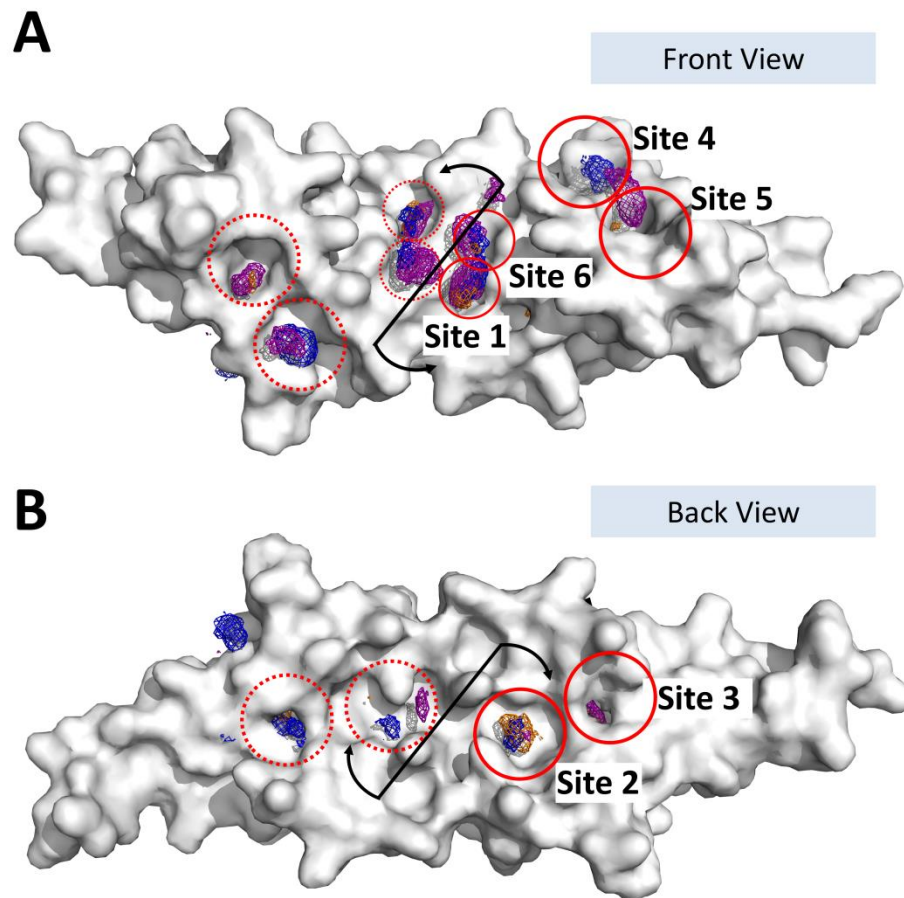


Figure 5-1. The all atom MixMD maps contoured at 20σ along with the average protein structure of Hsp27 are shown above. The protein surface is white, the isopropanol density is blue, the pyrimidine density is purple, and the acetonitrile density is orange. The binding sites identified by MixMD are marked with red circles, and their symmetric relationships are shown by solid and dashed red circles. A) The front view of Hsp27 with the sites 1, 4, 5, and 6 is shown. B) The back view of Hsp27 is shown with site 2 and 3.

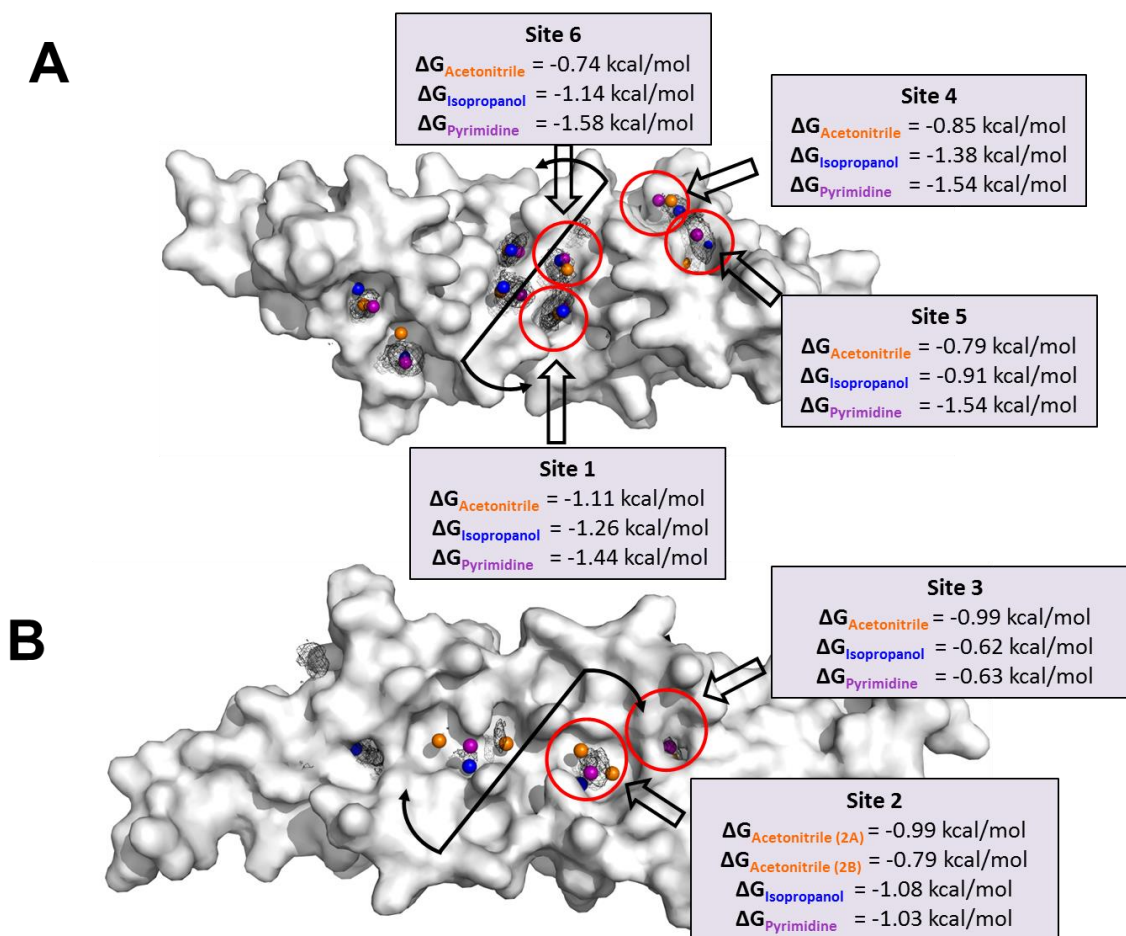


Figure 5-2. The symmetry averaged free energies of the top 6 MixMD sites are shown for acetonitrile, isopropanol, and pyrimidine. All solvent maps are contoured at 20σ and are shown as a black mesh. A) The front face of Hsp27 along with sites 1, 4, 5, and 6 are shown. B) The back face of hsp27 with sites 2 and 3 are shown.

Table 5-1. The free energies and ligand efficiencies of acetonitrile, isopropanol, and pyrimidine for the MixMD binding sites are presented below. For site 2, there were two acetonitrile probe molecules bound in the same site, as a result, the one with weaker binding is listed in parenthesis.

Site No.	Acetonitrile free energy (kcal/mol)	Acetonitrile ligand efficiency (kcal/mol-HA) ^(a)	Isopropanol free energy (kcal/mol)	Isopropanol ligand efficiency (kcal/mol-HA) ^(a)	Pyrimidine free energy (kcal/mol)	Pyrimidine ligand efficiency (kcal/mol-HA) ^(a)
Site 1	-1.11	-0.37	-1.26	-0.31	-1.44	-0.24
Site 2	-0.99 (-0.79)	-0.33 (-0.26)	-1.08	-0.27	-1.03	-0.17
Site 3	-0.99	-0.33	-0.62	-0.16	-0.63	-0.1
Site 4	-0.85	-0.28	-1.39	-0.35	-1.38	-0.23
Site 5	-0.79	-0.26	-0.91	-0.23	-1.54	-0.26
Site 6	-0.74	-0.25	-1.14	-0.28	-1.58	-0.26

(a) HA stands for non-hydrogen heavy atoms

5.4.2 Comparing 15N, 1H HSQC spectra of organic probe molecules to MixMD probe mapping

NMR chemical shift data for Hsp27 provided us with an avenue for direct comparison with MixMD results. Chemical shift perturbations (CSPs) in the presence of acetonitrile, isopropanol, and pyrimidine were obtained at 2% v/v concentration. These chemical shift perturbations were calculated from the NMR HSQC spectra of Hsp27 using equation (2).

$$\text{Chemical Shift Perturbation} = \sqrt{(H_{\text{Water}} - H_{\text{probe}})^2 + ((N_{\text{Water}} - N_{\text{probe}})/6.51)^2} \quad (2)$$

The CSPs were normalized with the total CSPs seen for all residues for that particular solvent experiment. Residues with normalized CSPs greater than the 0.02 average were considered to be significant. The comparison between MixMD results and HSQC experiments is shown in Figure 5-3 where NMR residues that shift in at least two cosolvent HSQC experiments are colored green and MixMD maps contoured at 20 σ are colored orange, blue, and purple for

acetonitrile, isopropanol, and pyrimidine respectively. Several residues are seen to shift near Site 4 and 5 (Figure 5-3A). However, threonine 61 was the only residues with significant NMR chemical shift near sites 1 and 6. On the back face of Hsp27 (Figure 5-3B) NMR chemical shifts could be seen for residues between sites 2 and 3. Barring site 1 and 6 on the front face, for the most part, a striking level of agreement could be seen between MixMD and NMR HSQC experiment.

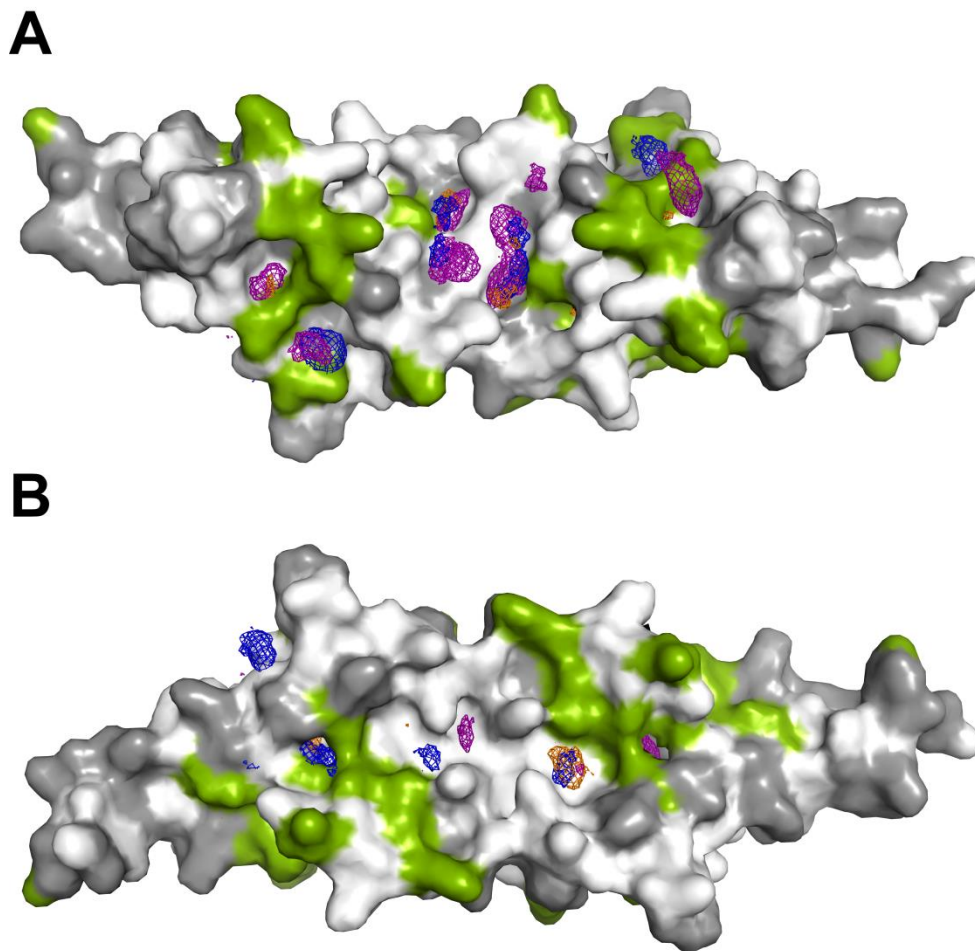


Figure 5-3 MixMD maps of Hsp27 with acetonitrile, isopropanol, and pyrimidine are shown contoured at 20σ . These MixMD maps are color coded as orange for acetonitrile, blue for isopropanol, and purple for pyrimidine. Hsp27 residues that shift in at least two different co-solvent NMR experiments are shown colored green. Residues missing assignment and prolines which are invisible to the NMR HSQC experiment are colored grey.

5.4.3 Experimental support from crystallography

Site 1 and site 6 mapped by MixMD is located in a groove at the dimer interface of Hsp27. This binding site is formed by beta 6 and 7 strands from both monomers. A disulfide-bond between two cysteine residues unique to Hsp27 in this region stitches the dimer together. Given the location of the site at the dimer interface, molecules binding in this region may have the potential to modulate the chaperone activity of Hsp27 by disrupting the dimer formation process. Interestingly, structural evidence from human alpha b crystallin (PDB ID: 2Y1Y)(152), a protein with similar fold as Hsp27 demonstrates the presence of unmodelled density that suggests the presence of a cosolvent in this region. However, no evidence for such cosolvent binding can be found in the only reported crystal structure of Hsp27(153).

MixMD site 2 maps a binding site formed at the interface of the dimer. Examination of crystal structure of sHSP reveals that cosolvent molecules bind in this region. In the crystal structure of the sHSP Heat shock protein beta-6 (HSPB6), a glycerol molecule is bound in site 2 (154). Similarly, in α -crystallin A, 2-Methylpentane-2,4-Diol is bound in site 3 (155). These examples provide experimental support for our prediction that site 2 is a relatively easy site to desolvate on the protein surface. MixMD probe simulations map two adjacent sites, site 4 and site 5. Crystallographic, NMR, and experimental evidence from several small heat shock proteins (sHSP) suggest that this site binds the LXL motif of the C-terminal region (153, 156). A recent crystal structure of Hsp27 in conjunction with a fragment of the c-terminus containing the LXL motif provides direct evidence for this interaction in Hsp27 (153). The two isoleucine residues from the c-terminal fragment occupy sites 4 and 5 in this structure (Figure 5-4). Mutations in the conserved LXL motif of α B-crystallin, a sHSP closely related to Hsp27, have been shown to affect its oligomerization state (157). This evidence suggests that targeting site 2 in Hsp27 through small molecules could provide a potential avenue for modulating its oligomerization state and function.

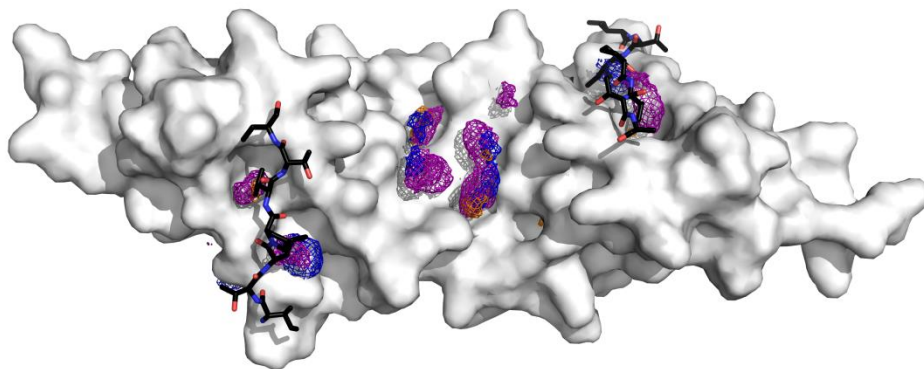


Figure 5-4. A recently reported crystal structure of the LXL motif of the c-terminal region of Hsp27 (PDB ID: 4MJH) is shown in black stick model overlaid with the MixMD maps. The leucine residues from the LXL motif can be seen to bind site 4 and site 5 thereby validating these sites identified via MixMD. MixMD density is contoured at 20σ and is color coded to represent acetonitrile (orange), isopropanol (blue), and pyrimidine (purple).

5.4.4 High Throughput Screening – Derived Inhibitors

A screening campaign initiated by our collaborator Gestwicki (158) using Differential scanning fluorimetry identified Captopril (159) as a hit that bound in site 1 and site 6 by means of a disulfide bond. Captopril is an Angiotensin Converting enzyme inhibitor (160). It was found to retain binding activity for Hsp27 even after mutating the cysteine to an alanine residue. These experimental studies present a strong connection between the ability of site 1 and 6 to bind drug-like molecules and the propensity of MixMD probes to bind in this region. Similar results were found for site 2, where a high throughput screen resulted in an unusually high hit rate of 0.4%. Further details on the experimental workup related to the identification and development of these small molecules will be the subject of a future paper from the Gestwicki group.

5.5 Conclusions

We have described the application of MixMD in a blinded fashion on Hsp27. MixMD identified six druggable binding sites on the protein surface. Using experimental verification, we have

identified sites 1, 2, and 6 can bind drug like matter. A recently reported crystal structure of Hsp27 with the c-terminal LXL motif, confirmed the importance of the site 4 and 5 as an interface for oligomerization. Furthermore, a high level of agreement was found when MixMD simulation results were compared with NMR chemical shift data. This study to the best of our knowledge provides details on the first successful application of a cosolvent-based molecular dynamics approach in a blinded fashion.

Chapter 6. Conclusions and Future Directions

6.1 Significant contributions of this thesis

The introduction of the thesis (Chapter 1) provided a brief overview of probe mapping based on two different approaches, one for static structures and another for MD simulations. Such probe mapping techniques aim to serve as an alternative to experimental approaches such as Multiple Solvent Crystal Structures. The major focus of my thesis included probe mapping based on two different approaches, one for static structures and another for MD simulations.

Chapter 2 focused on comparing the use of NMR and X-ray structures on the performance of our Multiple Protein Structures (MPS) receptor based pharmacophore models (12–17, 19). The MPS method incorporates protein flexibility through the use of many static snapshots of proteins obtained from experimental (X-ray, NMR) or computational (MD) methods. Using HIV-1 protease as a test case, previous work from our group showed that NMR structures provided MPS pharmacophore models with superior performance compared to crystal structures (17). In this thesis work, the MPS technique was applied to several other systems with NMR and X-ray structures. Test systems included Growth factor receptor bound protein 2 (Grb2), Src SH2 homology domain (Src-SH2), FK506-binding protein 1A (FKBP12), and Peroxisome proliferator-activated receptor- γ (PPAR- γ). The results from this work demonstrated that NMR pharmacophore models displayed superior performance, an observation in line with previous work from our lab. We note that pharmacophore models from NMR structures had fewer pharmacophore elements, and they represented only essential features observed in a diverse set of inhibitors/agonists. This chapter delved deeper to understand the origin and location of extraneous pharmacophore elements in X-ray pharmacophore models. In our analysis of MPS X-ray pharmacophore models, we note that such extraneous pharmacophore elements primarily lie at the periphery of the active site and arise as a result of an increased rigidity of the

protein likely from crystal packing effects. X-ray pharmacophore models retained performance for the most part when elements at the periphery of the active site were truncated, confirming this observation.

Chapter 3 presented our application of MixMD to allosteric systems. Previous work in the Carlson lab demonstrated that under conditions of full protein flexibility, accurate mapping of the co-solvent binding locations could be recaptured using MixMD (46). This thesis work, presented approaches for moving from identifying co-solvent locations with MixMD to mapping hotspots/binding sites on the protein surfaces. Drawing upon seminal contribution of MSCS by Dagmar Ringe and co-workers (3, 4, 161), we have identified the optimal approach to require sites to be mapped by more than one type of probe. It is important to note that binding sites were defined by overlapping locations of co-solvent molecules identified from MD simulations. These conditions were only met when each co-solvent simulation is run separately. This provided MixMD with a distinct advantage over other co-solvent based MD techniques that use multiple co-solvents in MD simulation. In performing MixMD simulations, we have used co-solvents such as acetonitrile, isopropanol and pyrimidine. This allowed us to map a range of interactions including hydrogen-bonding, hydrophobic, and aromatic interactions. An application of MixMD using this protocol successfully recaptured the location of competitive and allosteric sites on proteins starting from conformations with no allosteric ligands bound. Our design, setup, execution, and analysis of MixMD simulations made it amenable for use in prospective application. Given the increasing need for driving selectivity by targeting allosteric sites, especially in the field of protein kinases, the developments presented in this chapter will have a strong impact on the field.

Chapter 4 described our development of a suite of computational techniques to fully characterize MixMD binding sites. Most co-solvent based approaches decompose the atomic free energies onto a grid (27, 31). However, we believe that the free energy of binding (ΔG_{bind}) is a whole molecule property. In keeping with this philosophy, we have presented our procedure for calculating ΔG_{bind} for the probe molecules as a whole. This approach is similar to the one described by Bahar and co-workers (51). Furthermore, evidence was provided in

support of the argument that measures of ΔG_{bind} calculated using occupancies vary based on concentration and cannot be used to obtain maximal free energies of binding for sites on the protein surface as reported by other. Additionally, methods for calculating entropies from co-solvent based simulations were presented.

Chapter 5 presented the first successful prospective application of any mixed-solvent molecular dynamics. Most co-solvent simulation studies presented in the literature thus far have been retrospective in nature. It is worthwhile to note that MacKerell and co-workers used pharmacophore models derived from SILCS method to prescreen for molecules that can potentially bind $\beta 2\text{AR}$ (40). This screening was followed by other *in silico* methods such as docking to propose and verify molecules targeting $\beta 2\text{AR}$. Such studies present a step in the right direction towards realizing the potential of co-solvent simulations in SBDD. However, it could be argued that the binding site of $\beta 2\text{AR}$ is a well-known druggable target, and resorting to docking in the end throws away the advantage of cosolvent MD, reducing it to a standard approach on a known target. Here; by applying MixMD to the “undruggable” target Heat Shock Protein 27 (Hsp27), we have presented definitive proof that such simulations can be used to identify druggable binding sites in a prospective manner. Prompted by favorable results from MixMD, a high throughput screen conducted by Gestwicki and co-workers (unpublished data) identified drug-like molecules for MixMD binding sites in Hsp27. In addition to this important contribution, this chapter also presented a more direct comparison of MixMD simulation results with NMR chemical shifts for Hsp27 in 2% cosolvent mixtures! For the most part, a striking agreement could be seen between MixMD results and shifts in the NMR spectra when our co-solvents were added.

6.2 Future Directions

Hotspot identification using MD simulations is an appealing method as competition with water and protein flexibility is taken into account during the mapping process. Following are some potential applications of MixMD in the near future.

Converting MixMD maps to pharmacophore models. Latest developments in co-solvent simulations have focused on converting the information derived from simulations into pharmacophore models. Most method development in this direction has been very similar to our highly cited work on the MPS method (12–17, 19, 20). Moving forward, creating pharmacophore models by wedding our ideas of MPS and MixMD would be a natural progression for us.

Using MixMD simulations to assist in scoring and ranking ligands.

In addition to converting MixMD results to pharmacophore models, presenting the binding preference of MixMD probe molecules on a grid would allow one to score and rank ligands either from crystal structures or after a docking procedure. This ranking could be performed based on a score obtained from summing the occupancy at each grid point when it overlaps with an atom from the ligand. The occupancy used for summation would be obtained from different MixMD simulation based on an interaction type classification of the atoms of the ligands. As the preference for MixMD probes is a complex interplay between competition with water and favorable interaction energies with the protein, accounting for it through scoring and ranking would provide an extra dimension to current docking approaches that either lack means of dealing with solvation effects or treat it in a rudimentary fashion. Such a ranking procedure has been implemented by Mackerell and co-workers in their SILCS simulations (31–40, 162, 163). As SILCS simulations are performed using repulsion terms between cosolvents to drive adequate mixing, it remains to be seen if similar success can be achieved using MixMD simulations.

Identifying cryptic pockets using MixMD. Mixed-solvent simulations including our application of MixMD have demonstrated that cryptic pockets that open upon side-chain movement are achievable. However, it is yet to be determined if such simulations are enough to observe cryptic pockets that open upon large scale protein backbone motions. Methods that accelerate conformation sampling such as accelerated molecular dynamics (164) and metadynamics (165) are attractive alternatives to these problems and need to be investigated in conjunction with

MixMD. These cryptic pockets could then be exploited to drive selectivity between proteins where the orthosteric sites are similar.

Assessing the druggability of protein-protein interactions. Targeting protein-protein interactions using small molecules is challenging as these interactions are typically spread over a larger shallow surface area (166). Understanding whether sites that disproportionately contribute to binding exist and targeting them will be key in assessing the druggability and success rate in disrupting protein-protein interactions. In an application of MixMD to Farnesyl Pyrophosphate Synthase in this thesis work (Chapter 3), we note that protein-protein interaction interface was mapped very strongly. These results suggest that MixMD simulations can be used in the identification of protein-protein interaction sites. Furthermore, our application of MixMD on Heat Shock Protein 27 demonstrated that binding sites identified by MixMD can be targeted by drug-like molecules. Taken together, these results prompt the need to assess the utility of MixMD in prioritizing which protein-protein interactions to target with small molecules.

Expanding the range of probe molecules used with MixMD. In Chapter 3 - Chapter 5, probe molecules used for MixMD simulations were acetonitrile, isopropanol, and pyrimidine. These probes allowed use to capture a range of hydrophilic, hydrophobic, aromatic, and hydrogen bonding interactions. Recent work from our lab identified other organic water miscible probes suitable for MixMD (48). Additionally, in this thesis work, OPLS force field parameters were validated for acetate and methyl ammonium in order to map charged binding sites. With a wide variety of probe molecules at our disposal, it remains to be seen if MixMD simulations with different probes can be used to tailor the application for identifying charged binding site of functional significance, protein-protein interactions, and if using a different or extended set of probes would provide an added advantage in identifying druggable binding sites.

MixMD simulations with multiple probe molecules. Current protocols developed in Chapter 3 rely on performing MixMD simulations with each probe separately. This allowed us to detect druggable binding sites using approaches similar to MSCS wherein sites were required to be mapped by more than one probe. Mapping protein surfaces with MixMD simulations that use multiple probes simultaneously could be explored to obtain pharmacophore models from a

single MixMD simulation. Furthermore, synergistic effects that arise from two different co-solvents binding near each other could reveal further insights into the preference of the binding sites for different probe types that could be exploited in SBDD.

Exploring conformational dependence of MixMD simulations. In Chapter 3, we note the dependence of MixMD results on starting conformation. This was illustrated using the active and inactive conformations of ABL kinase. MixMD simulations starting from both ABL kinase conformations resulted in competitive and allosteric sites being mapped within the top-four sites. Interestingly, other high-ranking sites appeared in the peptide substrate binding region in the active conformation that were absent in the inactive conformation. These results are consistent with the biological function of the two different conformations, wherein the inactive conformation is not expected to bind peptide substrates. Our results serve as a starting point for extending this approach to other protein targets and confirming similar observation. Similar results across other systems would strengthen the argument for the use of MixMD to comment upon the importance of different conformations in the context of biological function.

Exploring Structure Activity Relationships with MixMD entropies. A relative ranking in terms of local entropies was obtained for MixMD binding sites using our method outlined in Chapter 4. These local entropy quantities were validated by visual inspection of normalized occupancies for acetonitrile, isopropanol, and pyrimidine. One could envision the use of such local entropy measures to assist in structure activity relationships. For example, MixMD binding sites with low entropies represent sites with no orientational preference. These locations could be available for substitution in ligands by any type of fragment that has similar physicochemical properties. However, MixMD sites with high entropies correspond to locations that are restricted in motion and make specific interactions with the protein. Substituting such location with other fragments might be difficult. The use of entropies in such a manner to guide SAR needs to be fully evaluated by comparison across several systems.

Appendices

Appendix A. Raw data for MPS Pharmacophore models

Table A-1 The residue and atom used for flooding the active site of proteins with probe molecules is shown for each protein.

Protein	Residue in Sequence (Highlighted in Red)	Atom used for MPS Flooding
Src SH2	MGSNKS ^K PKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSAAFAPAAAE PKLFGGFNSSDTVTSPQRAGPLAGGVTTFVALYDYESRTETDLSFKKGERLQIVNNTGEGD WWLAHSLSTGQGTGYIPSNYVAPSDSIQAEWYFGKITRRESERLLLNAENPRGTFLVRES ETTKGAYCLSVSDFDNAKGLNVKH ^Y KIRKLDSSGGFYITSRTQFNSLQQLVAYYSKHADGL CHRLTTCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFCGEVWMTWNGTTRVAIKTL KPGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPYIVTEYMSKGSLLDFLKGETGKY LRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARLIEDNEYT ARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLQDQVER GYRMPCPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQYQPGENL	CA
Grb2 SH2	MEAIKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGFIKPNYIEMKPHPW FFGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFLSVKFGNDVQH ^F KVLRDGAGKYFL WVVVFNSLNELVDYHRSTS ^V SRNQI ^F LRDIEQVPQPTYVQALFDFDPQEDGELGFRRG DFIHVMDNSDPNWWKGACHGQTGMFPRNYVTPVNRNV	CA
FKBP1 2	MGVQVETISPGDGRTPKRGQTCVVHYTGMLEDGKKFDSSRDRNKPFKFM ^L GKQEVIRGW EEGVAQMSVQRAKLTISP ^D YA ^Y GATGHPGIIPPHATLVFDVELLKE	OH

PPAR- γ	MGETLGDSPIDPESDSFTDTLSANISQEMTMVDTEMPFWPTNFGISSVDLSVMEDHSHSF DIKPFTTYVDFSSISTPHYEDIPFTRTPVVDYKYDLKLQEYQSAIKVEPASPPYYSEKT QLYNKPHEEPSNSLMAIECRVCGDKASGFHYGVHACEGCKGFFRRTIRLKLIDRCDLNC RIHKKSRNKQYCRFQKCLAVGMSHNAIRFGRMPQAEKEKLLAEISSDIDQLNPESADLR ALAKHLYDSYIKSFPLTKAKARAILTGKTTDKSPFVIYDMNSLMMGEDKIKFKHITPLQE QSKEVAIRIFQGCQFRSVEAVQEITEYAKSIPGFVNLDLNDQVTLTKYGVHEIITMLAS LMNKDGLVISEGQGFMTREFLKLRLKPFQDFMEPKFEFAVKFNALELDDSDLAIFIAVII LSGDRPGLLNKPIEDIQDNLLQALELQLKLNHPESQLFAKLLQKMTDLRQIVTEHVQL LQVIKKTETDMSLHPLLQEIYKDLY	O
------------	---	---

Table A-2 Crystal pharmacophore model coordinates and radius for Src SH2, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1O49.

Pharmacophore Element Type	x	y	z	RMSD, Å
Acceptor	23.045	25.309	15.744	0.78
Hydrophobic	22.494	19.634	15.611	1.14
Hydrophobic	21.056	20.259	19.264	0.82
Hydrophobic	16.246	28.319	17.404	1.02
Hydrophobic	19.733	12.946	22.01	0.59
Hydrophobic	18.206	26.191	25.499	1.03
Aromatic	21.118	15.93	22.328	0.71
Donor Acceptor	18.506	26.537	22.549	0.72
Donor Acceptor	20.04	28.426	15.815	1.14
Donor	20.88	16.386	24.425	0.64

Table A-3 Crystal pharmacophore model coordinates and radius for Src SH2 at a cutoff of 9Å, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1O49.

Pharmacophore Element Type	x	y	z	RMSD, Å
Hydrophobic	22.494	19.634	15.611	1.14
Hydrophobic	21.056	20.259	19.264	0.82
Hydrophobic	16.246	28.319	17.404	1.02
Hydrophobic	19.733	12.946	22.01	0.59
Hydrophobic	18.206	26.191	25.499	1.03
Aromatic	21.118	15.93	22.328	0.71
Donor Acceptor	18.506	26.537	22.549	0.72
Donor	20.88	16.386	24.425	0.64

Table A-4 Crystal pharmacophore model coordinates and radius for Src SH2 at a cutoff of 8Å, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1O49.

Pharmacophore Element Type	x	y	z	RMSD, Å
Hydrophobic	22.494	19.634	15.611	1.14
Hydrophobic	21.056	20.259	19.264	0.82

Hydrophobic	18.206	26.191	25.499	1.03
Aromatic	21.118	15.93	22.328	0.71
Donor Acceptor	18.506	26.537	22.549	0.72
Donor	20.88	16.386	24.425	0.64

Table A-5 NMR pharmacophore model coordinates and radius for Src SH2, the location of the pharmacophore model is relative to the NMR ensemble of FKBP12, PDB ID: 1FKR.

Pharmacophore Element Type	x	y	z	RMSD, Å
Acceptor	-0.944	9.605	6.074	0.81
Hydrophobic	-3.474	3.708	8.571	0.97
Hydrophobic	-4.927	8.771	3.624	1
Hydrophobic	2.594	4.854	8.569	1.21
Hydrophobic	0.697	-4.722	8.361	1.43
Donor Acceptor	-4.043	-3.668	11.248	1.15

Table A-6 Crystal pharmacophore model coordinates and radius for Grb2 SH2, the location of the pharmacophore model is relative to the crystal structure of Grb2 SH2, PDB ID: 1JYR.

Pharmacophore Element Type	x	y	z	RMSD, Å
Acceptor	32.185	26.561	18.443	0.69
Acceptor	28.961	28.686	14.664	0.67
Hydrophobic	32.212	22.944	20.033	0.99
Hydrophobic	37.887	24.645	13.197	0.82
Hydrophobic	38.017	19.467	18.022	0.99
Hydrophobic	31.909	27.788	14.595	0.85
Donor Acceptor	37.101	24.974	12.662	0.74
Donor Acceptor	36.867	17.705	18.149	0.74
Donor	38.794	24.65	12.34	0.74

Table A-7 Crystal pharmacophore model coordinates and radius for Grb2 SH2 at a cutoff of 8Å, the location of the pharmacophore model is relative to the crystal structure of Grb2 SH2, PDB ID: 1JYR.

Pharmacophore Element Type	x	y	z	RMSD, Å
Acceptor	32.185	26.561	18.443	0.69
Acceptor	28.961	28.686	14.664	0.67
Hydrophobic	32.212	22.944	20.033	0.99
Hydrophobic	37.887	24.645	13.197	0.82
Hydrophobic	38.017	19.467	18.022	0.99
Hydrophobic	31.909	27.788	14.595	0.85
Donor Acceptor	37.101	24.974	12.662	0.74
Donor	38.794	24.65	12.34	0.74

Table A-8 Crystal pharmacophore model coordinates and radius for Grb2 SH2 at a cutoff of 7Å, the location of the pharmacophore model is relative to the crystal structure of Grb2 SH2, PDB ID: 1JYR.

Pharmacophore Element Type	x	y	z	RMSD, Å
Acceptor	32.185	26.561	18.443	0.69
Acceptor	28.961	28.686	14.664	0.67
Hydrophobic	37.887	24.645	13.197	0.82
Hydrophobic	31.909	27.788	14.595	0.85
Donor Acceptor	37.101	24.974	12.662	0.74
Donor	38.794	24.65	12.34	0.74

Table A-9 NMR pharmacophore model coordinates and radius for Grb2 SH2, the location of the pharmacophore model is relative to the NMR ensemble of Grb2 SH2, PDB ID: 1X0N.

Pharmacophore Element Type	x	y	z	RMSD, Å
Hydrophobic	10.583	-7.886	-12.445	1.31
Hydrophobic	18.933	-4.414	-7.544	1.35
Hydrophobic	18.354	3.699	-11.588	0.99
Hydrophobic	14.746	-1.428	-14.61	1.36
Donor Acceptor	15.406	-4.32	-11.184	0.51
Donor Acceptor	17.648	-4.794	-7.216	0.57
Donor Acceptor	18.167	4.635	-11.144	1.08
Donor Acceptor	9.56	-6.35	-13.362	0.6

Table A-10 Crystal pharmacophore model coordinates and radius for FKBP12, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1FKB.

Elem. Type	X	Y	Z	RMSD, Å
Acceptor	4.575	6.275	15.78	0.57
Hydrophobic	2.937	6.679	12.549	0.99
Hydrophobic	8.444	11.583	11.511	0.7
Hydrophobic	13.271	8.494	9.785	0.7
Hydrophobic	3.842	7.424	18.287	1.04
Hydrophobic	12.767	11.235	16.237	1.01
Aromatic	-2.94	10.424	16.155	0.52
Aromatic	11.756	9.936	9.823	0.46
Donor Acceptor	12.235	7.771	10.307	0.57
Donor Acceptor	5.089	8.12	11.808	0.52
Donor	-1.936	8.814	17.561	0.37
Donor	6.678	11.9	11.446	0.83
Donor	9.412	13.152	9.461	1.11
Donor	11.61	11.466	18.357	0.82

Table A-11 Crystal pharmacophore model coordinates and radius for FKBP12 at a cutoff of 9Å, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1FKB.

Elem. Type	X	Y	Z	RMSD, Å
Acceptor	4.575	6.275	15.78	0.57
Hydrophobic	2.937	6.679	12.549	0.99
Hydrophobic	8.444	11.583	11.511	0.7
Hydrophobic	13.271	8.494	9.785	0.7
Hydrophobic	3.842	7.424	18.287	1.04
Hydrophobic	12.767	11.235	16.237	1.01
Aromatic	11.756	9.936	9.823	0.46
Donor Acceptor	12.235	7.771	10.307	0.57
Donor Acceptor	5.089	8.12	11.808	0.52
Donor	-1.936	8.814	17.561	0.37
Donor	6.678	11.9	11.446	0.83
Donor	9.412	13.152	9.461	1.11
Donor	11.61	11.466	18.357	0.82

Table A-12 Crystal pharmacophore model coordinates and radius for FKBP12 at a cutoff of 8Å, the location of the pharmacophore model is relative to the crystal structure of FKBP12, PDB ID: 1FKB.

Elem. Type	X	Y	Z	RMSD, Å
Acceptor	4.575	6.275	15.78	0.57
Hydrophobic	2.937	6.679	12.549	0.99
Hydrophobic	8.444	11.583	11.511	0.7
Hydrophobic	3.842	7.424	18.287	1.04
Hydrophobic	12.767	11.235	16.237	1.01
Aromatic	11.756	9.936	9.823	0.46
Donor Acceptor	12.235	7.771	10.307	0.57
Donor Acceptor	5.089	8.12	11.808	0.52
Donor	6.678	11.9	11.446	0.83
Donor	11.61	11.466	18.357	0.82

Table A-13 NMR pharmacophore model coordinates and radius for FKBP12, the location of the pharmacophore model is relative to the NMR ensemble of FKBP12, PDB ID: 1FKR.

Elem. Type	X	Y	Z	RMSD, Å
Hydrophobic	-46.867	-29.041	73.395	0.830
Hydrophobic	-40.231	-26.134	72.841	1.200
Hydrophobic	-48.158	-22.092	70.676	0.970
Donor Acceptor	-47.338	-27.331	72.338	0.670

Table A-14 Crystal pharmacophore model coordinates and radius for PPAR- γ , the location of the pharmacophore model is relative to the crystal structure of PPAR- γ , PDB ID: 1ZGY.

Pharmacophore Element Type	x	y	z	RMSD, Å
Acceptor	23.837	3.429	27.071	0.95
Hydrophobic	27.153	3.864	30.464	0.9
Hydrophobic	33.506	-4.124	26.833	1.03
Aromatic	32.508	0.178	25.149	0.77
Donor Acceptor	33.958	-4.405	25.491	0.98
Donor	21.755	1.116	24.65	0.69

Table A-15 Crystal pharmacophore model coordinates and radius for PPAR- γ at a cutoff of 8Å, the location of the pharmacophore model is relative to the crystal structure of PPAR- γ , PDB ID: 1ZGY.

Pharmacophore Element Type	x	y	z	RMSD, Å
Acceptor	23.837	3.429	27.071	0.95
Hydrophobic	27.153	3.864	30.464	0.9
Hydrophobic	33.506	-4.124	26.833	1.03
Aromatic	32.508	0.178	25.149	0.77
Donor Acceptor	33.958	-4.405	25.491	0.98

Table A-16 Crystal pharmacophore model coordinates and radius for PPAR- γ at a cutoff of 7Å, the location of the pharmacophore model is relative to the crystal structure of PPAR- γ , PDB ID: 1ZGY.

Pharmacophore Element Type	x	y	z	RMSD, Å
Hydrophobic	27.153	3.864	30.464	0.9
Hydrophobic	33.506	-4.124	26.833	1.03
Aromatic	32.508	0.178	25.149	0.77
Donor Acceptor	33.958	-4.405	25.491	0.98

Table A-17 NMR pharmacophore model coordinates and radius for PPAR- γ , the location of the pharmacophore model is relative to the NMR ensemble of PPAR- γ , PDB ID: 2QMV.

Pharmacophore Element Type	x	y	z	RMSD, Å
Hydrophobic	-1.264	-9.563	1.228	1.81
Hydrophobic	11.163	-3.189	2.634	1.48
Donor	4.161	-4.331	-1.857	1.57

Table A-18 PDB IDs and references of 22 SRC5H2 crystal structures used to create the Src SH2 X-ray MPS pharmacophore model.

1O44(167)	1O4Q(167)	1O4N(167)	1A08(168)	1O48(167)	1A1C(168)	1O4M(167)	1O4O(167)	1SHD(168)	1O42(167)
-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------

1O49(167)	1A1B(168)	1O4P(167)	1O47(167)	1O4H(167)	1O4B(167)	1O4L(167)	1A1E(168)	1O4R(167)	1O4J(167)
1O46(167)	1A07(168)								

Table A-19 PDB IDs and references of 18 GRB2 crystal structures used to create the Grb2 SH2 X-ray MPS pharmacophore model.

1BM2(169)	1BMB(169)	3N84(170)	1JYQ(171)	1TZE(172)	3IN7(173)	3MXV(174)	3IMJ(173)	3IMD(173)	1JYR(171)
3KFJ(173)	2HUW(175)	3N7Y(170)	3C7I(175)	3MXC(174)	3IN8(173)	1ZFP(176)	3N8M(170)		

Table A-20 PDB IDs and references of 22 PPARGAMMA crystal structures used to create the PPAR- γ X-ray MPS pharmacophore model.

2I4J(177)	1ZGY(178)	3GBK(179)	1RDT(180)	2ATH(181)	3G9E(182)	2PRG(183)	2HWR(184)	2Q8S(185)	1FM6(186)
1ZEO(187)	2HWQ(184)	2F4B(188)	2GTK(189)	2VV1(190)	2ZNO(191)	1I7I(192)	3CWD(193)	3IA6(194)	3HOD(195)
2VSR(190)	3CS8(196)								

Table A-21 PDB IDs and references of 20 FKBP12 crystal structures used to create the FKBP12 X-ray MPS pharmacophore model.

2FKE(197)	1NSG(198)	1FKB(199)	1FKJ(200)	1BKF(201)	2FAP(198)	1J4I(202)	1J4H(202)	1FKI(203)	1FKD(197)
1FKF(204)	1FKH(203)	1D7J(205)	1J4R(206)	1A7X(207)	1BL4(208)	1QPF(209)	1FKG(203)	1D7I(205)	3FAP(198)

Table A-22 PDB IDs and references of NMR ensembles used to create the NMR MPS pharmacophore models.

Protein	NMR ensemble PDB ID
Src SH2	1HCT(210)
Grb2 SH2	1XON(211)
FKBP12	1FKR(212)
PPAR- γ	2QMV(213)

Table A-23 ROC plot data for Src SH2 Crystal Pharmacophore model.

CR YS	Src SH2 High Affinity				Src SH2 Low Affinity				Decoy Molecules				High Affinity distance from (0,100)				Low Affinity distance from (0,100)			
	10 of 10	9 of 10	8 of 10	7 of 10	10 of 10	9 of 10	8 of 10	7 of 10	10 of 10	9 of 10	8 of 10	7 of 10	10 of 10	9 of 10	8 of 10	7 of 10	10 of 10	9 of 10	8 of 10	7 of 10
1.0 0x	0	0	0	0	0	0	0	0	0	0	0	0	10 0	10 0	10 0	10 0	10 0	10 0	10 0	10 0
1.3 3x	0	0	0	6.2 5	0	0	0	1.3 79	0	0	0	0.0 43	10 0	10 0	10 0	93. 75	10 0	10 0	10 0	98. 62 1
1.6 6x	0	0	0	6.2 5	0	0	0	8.2 76	0	0	0	0.3 91	10 0	10 0	10 0	93. 75 1	10 0	10 0	10 0	91. 72 5
2.0 0x	0	0	6.2 5	43. 75	0	0	3.4 48	37. 24 1	0	0	0.0 43	1.7 37	10 0	10 0	93. 75	56. 27 7	10 0	10 0	96. 55 2	62. 78 3
2.3 3x	0	0	18. 75	68. 75	0	0	10. 34 5	62. 06 9	0	0	0.2 61	5.6 88	10 0	10 0	81. 25	31. 76 3	10 0	10 0	89. 65 5	38. 35 5
2.6 6x	0	12. 5	31. 25	93. 75	0	2.0 69	20. 69	75. 17 2	0	0.0 87	1.3 46	11. 94 1	10 0	87. 5	68. 76 3	13. 47 8	10 0	97. 93 1	79. 32 1	27. 55
3.0 0x	0	12. 5	50	10 0	0	7.5 86	37. 24 1	77. 24 1	0	0.4 34	3.9 95	22. 62 3	10 0	87. 50 1	50. 15 9	22. 62 3	10 0	92. 41 5	62. 88 6	32. 09

Table A-24 ROC plot data for Src SH2 Crystal pharmacophore model at a cutoff of 9Å.

CTF9 A	Src SH2 High Affinity			Src SH2 Low Affinity			Decoy Molecules			High Affinity distance from (0,100)			Low Affinity distance from (0,100)		
	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8
1.00 x	0	0	0	0	0	0.69	0	0	0	100	100	100	100	100	99.31
1.33 x	0	0	0	0	0	2.75 9	0	0	0.08 7	100	100	100	100	100	97.241
1.66 x	0	0	6.25	0	0	13.1 03	0	0	0.65 1	100	100	93.752	100	100	86.899
2.00 x	0	6.25	18.7 5	0	2.06 9	28.2 76	0	0	2.86 6	100	93.75	81.301	100	97.931	71.781
2.33 x	0	12.5	50	0	8.27 6	45.5 17	0.04 3	0.60 8	7.99	100	87.502	50.634	100	91.726	55.066
2.66 x	0	12.5	81.2 5	2.06 9	19.3 1	59.3 1	0.04 3	1.82 4	14.7 63	100	87.519	23.864	97.931	80.711	43.285
3.00 x	0	25	87.5	2.06 9	31.0 34	71.7 24	0.21 7	4.34 2	23.9 69	100	75.126	27.033	97.931	69.103	37.068

Table A-25 ROC plot data for Src SH2 Crystal pharmacophore model at a cutoff of 8Å.

CTF8 A	Src SH2 High Affinity		Src SH2 Low Affinity		Decoy Molecules		High Affinity distance from (0,100)		Low Affinity distance from (0,100)	
	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6
1.00x	0	0	0	3.448	0	0.347	100	100.001	100	96.553
1.33x	0	6.25	0	13.103	0.043	1.259	100	93.758	100	86.906
1.66x	0	25	1.379	35.172	0.434	5.341	100.001	75.19	98.622	65.048

2.00x	6.25	81.25	7.586	56.552	0.955	12.245	93.755	22.394	92.419	45.141
2.33x	12.5	93.75	9.655	68.966	2.692	22.319	87.541	23.178	90.385	38.226
2.66x	18.75	93.75	20.69	75.172	5.037	34.26	81.406	34.825	79.47	42.31
3.00x	25	100	34.483	80	7.729	47.069	75.397	47.069	65.971	51.142

Table A-26 ROC plot data for Src SH2 NMR pharmacophore model.

NMR	Src SH2 High Affinity		Src SH2 Low Affinity		Decoy Molecules		High Affinity distance from (0,100)		Low Affinity distance from (0,100)	
	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6
1.00x	0	12.5	0	9.655	0	0.261	100	87.5	100	90.345
1.33x	6.25	25	2.069	43.448	0	1.389	93.75	75.013	97.931	56.569
1.66x	12.5	68.75	8.966	68.966	0.391	5.341	87.501	31.703	91.035	31.49
2.00x	25	100	26.207	77.241	1.476	14.54 6	75.015	14.546	73.808	27.01
2.33x	62.5	100	54.483	78.621	4.255	29.48 3	37.741	29.483	45.715	36.419
2.66x	93.75	100	70.345	80	10.59 5	46.54 8	12.301	46.548	31.491	50.663
3.00x	100	100	78.621	82.759	20.79 9	65.00 2	20.799	65.002	29.827	67.25

Table A-27 ROC plot data for Grb2 Crystal pharmacophore model.

CRYS	Grb2 SH2 High Affinity			Grb2 SH2 Low Affinity			Decoy Molecules			High Affinity distance from (0,100)			Low Affinity distance from (0,100)		
	9 of 9	8 of 9	7 of 9	9 of 9	8 of 9	7 of 9	9 of 9	8 of 9	7 of 9	9 of 9	8 of 9	7 of 9	9 of 9	8 of 9	7 of 9
1.00x	0	0	0	0	0	0	0	0	0	100	100	100	100	100	100
1.33x	0	0	4.91 8	0	0	5.15 5	0	0	0	100	100	95.082	100	100	94.845
1.66x	0	6.55 7	60.6 56	0	1.03 1	42.2 68	0	0.04 3	0.95 5	100	93.443	39.356	100	98.969	57.74
2.00x	3.27 9	34.4 26	90.1 64	0	28.8 66	77.3 2	0.04 3	0.30 4	3.82 1	96.721	65.575	10.552	100	71.135	23
2.33x	18.0 33	75.4 1	98.3 61	10.3 09	57.7 32	90.7 22	0.08 7	1.52	8.64 1	81.967	24.637	8.795	89.691	42.295	12.679
2.66x	57.3 77	90.1 64	100	41.2 37	82.4 74	92.7 84	0.34 7	4.12 5	15.1 11	42.624	10.666	15.111	58.764	18.005	16.746
3.00x	75.4 1	98.3 61	100	58.7 63	90.7 22	92.7 84	1.21 6	7.55 5	26.4	24.62	7.731	26.4	41.255	11.965	27.368

Table A-28 ROC plot data for Grb2 Crystal pharmacophore model at a cutoff of 8Å.

CTF8 A	Grb2 SH2 High Affinity			Grb2 SH2 Low Affinity			Decoy Molecules			High Affinity distance from (0,100)			Low Affinity distance from (0,100)		
	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8
1.00x	0	0	3.27 9	0	0	1.03 1	0	0	0.08 7	100	100	96.721	100	100	98.969
1.33x	0	0	52.4	0	0	39.1	0	0	1.04	100	100	47.552	100	100	60.834

x			59			75			2						
1.66x	0	18.033	81.967	0	16.495	84.536	0.043	0.434	6.036	100	81.968	19.016	100	83.506	16.6
2.00x	4.918	77.049	100	7.216	57.732	91.753	0.043	2.388	15.892	95.082	23.075	15.892	92.784	42.335	17.904
2.33x	42.623	88.525	100	26.804	82.474	92.784	0.261	6.079	30.048	57.378	12.986	30.048	73.196	18.55	30.902
2.66x	75.41	98.361	100	58.763	92.784	94.845	1.737	11.376	47.026	24.651	11.493	47.026	41.274	13.472	47.308
3.00x	91.803	100	100	78.351	92.784	96.907	4.386	21.45	65.697	9.297	21.45	65.697	22.089	22.631	65.77

Table A-29 ROC plot data for Grb2 Crystal pharmacophore model at a cutoff of 7Å.

CTF7 A	Grb2 SH2 High Affinity		Grb2 SH2 Low Affinity		Decoy Molecules		High Affinity distance from (0,100)		Low Affinity distance from (0,100)	
	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6
1.00x	0	21.311	0	24.742	0.043	1.346	100	78.701	100	75.27
1.33x	1.639	75.41	7.216	75.258	0.13	8.424	98.361	25.993	92.784	26.137
1.66x	21.311	95.082	29.897	93.814	1.216	21.928	78.698	22.473	70.114	22.784
2.00x	80.328	100	73.196	94.845	4.559	39.557	20.193	39.557	27.189	39.891
2.33x	96.721	100	89.691	97.938	10.03	53.495	10.552	53.495	14.383	53.535
2.66x	100	100	93.814	97.938	17.065	65.567	17.065	65.567	18.152	65.599
3.00x	100	100	95.876	98.969	25.488	73.6	25.488	73.6	25.819	73.607

Table A-30 ROC plot data for Grb2 NMR pharmacophore model.

NMR	Grb2 SH2 High Affinity			Grb2 SH2 Low Affinity			Decoy Molecules			High Affinity distance from (0,100)			Low Affinity distance from (0,100)		
	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8	8 of 8	7 of 8	6 of 8
1.00x	0	0	32.787	0	0	14.433	0	0	0.261	100	100	67.214	100	100	85.567
1.33x	0	18.033	86.885	0	8.247	72.165	0	0.174	3.083	100	81.967	13.472	100	91.753	28.005
1.66x	6.557	73.77	100	4.124	50.515	89.691	0.043	1.042	10.551	93.443	26.251	10.551	95.876	49.496	14.751
2.00x	47.541	91.803	100	36.082	78.351	92.784	0.304	3.43	22.492	52.46	8.886	22.492	63.919	21.919	23.621
2.33x	78.689	96.721	100	64.948	91.753	93.814	1.389	7.512	41.945	21.356	8.196	41.945	35.08	11.155	42.399
2.66x	85.246	100	100	79.381	92.784	94.845	2.996	14.546	61.572	15.055	14.546	61.572	20.836	16.238	61.787
3.00x	98.361	100	100	90.722	92.784	95.876	5.601	28.658	78.333	5.836	28.658	78.333	10.838	29.553	78.441

Table A-31 ROC plot data for FKBP12 Crystal Pharmacophore model.

CRYS	FKBP12 High Affinity	FKBP12 Low Affinity	Decoy Molecules	High Affinity distance from (0,100)	Low Affinity distance from (0,100)
------	----------------------	---------------------	-----------------	-------------------------------------	------------------------------------

Radius	14 of 14	13 of 14	12 of 14	11 of 14	14 of 14	13 of 14	12 of 14	11 of 14	14 of 14	13 of 14	12 of 14	11 of 14	14 of 14	13 of 14	12 of 14	11 of 14	14 of 14	13 of 14	12 of 14	11 of 14	
1.00x	0	0	0	0	0	0	0	0	0	0	0	0	10	10	10	10	10	10	10	10	10
1.33x	0	0	0	0	0	0	0	0	0	0	0	0	10	10	10	10	10	10	10	10	10
1.66x	0	0	0	0	0	0	0	0	0	0	0	0	10	10	10	10	10	10	10	10	10
2.00x	0	0	0	0	0	0	0	0	0	0	0	0	10	10	10	10	10	10	10	10	10
2.33x	0	0	0	0	0	0	0	0	0	0	0	0	10	10	10	10	10	10	10	10	10
2.66x	0	0	0	0	0	0	0	0	0	0	0	0	10	10	10	10	10	10	10	10	10
3.00x	0	0	0	0	0	0	0	0	0	0	0	0.129	10	10	10	10	10	10	10	10	10

CRYS	FKBP12 High Affinity				FKBP12 Low Affinity				Decoy Molecules				High Affinity distance from (0,100)				Low Affinity distance from (0,100)			
radius	10 of 14	9 of 14	8 of 14	7 of 14	10 of 14	9 of 14	8 of 14	7 of 14	10 of 14	9 of 14	8 of 14	7 of 14	10 of 14	9 of 14	8 of 14	7 of 14	10 of 14	9 of 14	8 of 14	7 of 14
1.00x	0	0	0	0	0	0	0	0	0	0	0	0	100	100	100	100	100	100	100	100
1.33x	0	0	1.282	12.821	0	0	0	0	0	0	0	0.215	100	100	98.718	87.179	100	100	100	100
1.66x	0	0	5.128	30.769	0	0	1.042	7.292	0	0	0.172	2.41	100	100	94.872	69.273	100	100	98.958	92.739
2.00x	0	0	19.231	47.436	0	0	6.25	14.583	0	0.043	1.42	7.659	100	100	80.781	53.119	100	100	93.761	85.76
2.33x	0	12.821	34.615	65.385	0	2.083	8.333	26.042	0.043	0.688	4.389	15.878	100	87.22	65.532	38.083	100	97.919	91.772	75.643
2.66x	5.128	32.051	48.718	80.769	1.042	6.25	13.542	42.708	0.258	2.668	8.434	28.485	94.872	68.001	51.971	34.369	98.958	93.788	86.88	63.983
3.00x	20.513	38.462	61.538	88.462	5.208	9.375	23.958	54.167	1.463	5.508	16.91	48.064	79.5	61.784	42.015	49.429	94.803	90.792	77.99	66.414

Table A-32 ROC plot data for FKBP12 Crystal Pharmacophore model at a cutoff of 9Å.

CTFA	FKBP12 High Affinity				FKBP12 Low Affinity				Decoy Molecules				High Affinity distance from (0,100)				Low Affinity distance from (0,100)			
Radius	13 of 13	12 of 13	11 of 13	10 of 13	13 of 13	12 of 13	11 of 13	10 of 13	N 13 of 13	12 of 13	11 of 13	10 of 13	13 of 13	12 of 13	11 of 13	10 of 13	13 of 13	12 of 13	11 of 13	10 of 13
1.00x	0	0	0	0	0	0	0	0	0	0	0	0	10	10	10	10	10	10	10	10
1.33x	0	0	0	0	0	0	0	0	0	0	0	0	10	10	10	10	10	10	10	10
1.66x	0	0	0	0	0	0	0	0	0	0	0	0	10	10	10	10	10	10	10	10
2.00x	0	0	0	0	0	0	0	0	0	0	0	0	10	10	10	10	10	10	10	10

2.3 3x	0	0	0	0	0	0	0	0	0	0	0	0.0 43	10 0	10 0	10 0	10 0	10 0	10 0	10 0	10 0
2.6 6x	0	0	0	3.8 46	0	0	0	1.0 42	0	0	0	0.2 58	10 0	10 0	10 0	96. 15 4	10 0	10 0	10 0	98. 95 8
3.0 0x	0	0	0	16. 66 7	0	0	0	5.2 08	0	0	0.1 29	1.4 2	10 0	10 0	10 0	83. 34 5	10 0	10 0	10 0	94. 80 3

CTF9 A	FKBP12 High Affinity			FKBP12 Low Affinity			Decoy Molecules			High Affinity distance from (0,100)			Low Affinity distance from (0,100)		
	9 of 13	8 of 13	7 of 13	9 of 13	8 of 13	7 of 13	9 of 13	8 of 13	7 of 13	9 of 13	8 of 13	7 of 13	9 of 13	8 of 13	7 of 13
1.00 x	0	0	0	0	0	0	0	0	0	100	100	100	100	100	100
1.33 x	0	1.28 2	12.8 21	0	0	0	0	0	0.21 5	100	98.718	87.179	100	100	100
1.66 x	0	5.12 8	30.7 69	0	1.04 2	7.29 2	0	0.17 2	2.41	100	94.872	69.273	100	98.958	92.739
2.00 x	0	19.2 31	47.4 36	0	5.20 8	14.5 83	0.04 3	1.42	7.57 3	100	80.781	53.107	100	94.803	85.752
2.33 x	12.8 21	34.6 15	65.3 85	2.08 3	8.33 3	26.0 42	0.64 5	4.38 9	15.8 78	87.181	65.532	38.083	97.919	91.772	75.643
2.66 x	30.7 69	48.7 18	80.7 69	6.25	13.5 42	42.7 08	2.66 8	8.34 8	28.3 13	69.282	51.957	34.227	93.788	86.86	63.906
3.00 x	38.4 62	60.2 56	88.4 62	9.37 5	22.9 17	54.1 67	5.42 2	16.8 24	48.0 21	61.776	43.158	49.388	90.787	78.898	66.383

Table A-33 ROC plot data for FKBP12 Crystal pharmacophore model at a cutoff of 8Å.

CT F8 A	FKBP12 High Affinity				FKBP12 Low Affinity				Decoy Molecules				High Affinity distance from (0,100)				Low Affinity distance from (0,100)			
	10 of 10	9 of 10	8 of 10	7 of 10	10 of 10	9 of 10	8 of 10	7 of 10	10 of 10	9 of 10	8 of 10	7 of 10	10 of 10	9 of 10	8 of 10	7 of 10	10 of 10	9 of 10	8 of 10	7 of 10
1.0 0x	0	0	0	0	0	0	0	0	0	0	0	0	100	10 0	10 0	10 0	100	10 0	10 0	10 0
1.3 3x	0	0	0	2.5 64	0	0	0	0	0	0	0	0.0 43	100	10 0	10 0	97. 43 6	100	10 0	10 0	10 0
1.6 6x	0	0	1.2 82	19. 23 1	0	0	1.0 42	2.0 83	0	0	0	0.4 73	100	10 0	98. 71 8	80. 77	100	10 0	98. 95 8	97. 91 8
2.0 0x	0	0	10. 25 6	34. 61 5	0	0	1.0 42	9.3 75	0	0	0.0 86	2.5 82	100	10 0	89. 74 4	65. 43 6	100	10 0	98. 95 8	90. 66 2
2.3 3x	0	0	25. 64 1	42. 30 8	0	0	5.2 08	15. 62 5	0	0	0.9 9	6.3 68	100	10 0	74. 36 6	58. 04 2	100	10 0	94. 79 7	84. 61 5
2.6 6x	0	3.8 46	33. 33	62. 82 1	0	0	7.2 92	31. 25	0	0.0 86	2.8 4	13. 33 9	100	96. 15 4	66. 72 7	39. 49 9	100	10 0	92. 75 1	70. 03 2
3.0 0x	0	20. 51 3	41. 02 6	76. 92 3	0	2.0 83	9.3 75	43. 75	0.0 43	0.6 45	5.9 38	25. 73 1	100	79. 49	59. 27 2	34. 56 3	100	97. 91 9	90. 81 9	61. 85 6

Table A-34 ROC plot data for FKBP12 NMR pharmacophore model.

NMR	FKBP12 High Affinity		FKBP12 Low Affinity		Decoy Molecules		High Affinity distance from (0,100)		Low Affinity distance from (0,100)	
	4 of 4	3 of 4	4 of 4	3 of 4	4 of 4	3 of 4	4 of 4	3 of 4	4 of 4	3 of 4
0.10x	0	39.744	0	10.417	0	0.775	100	60.261	100	89.586
0.20x	0	69.231	0	31.25	0	8.003	100	31.793	100	69.214
0.30x	0	80.769	0	48.958	0	21.601	100	28.921	100	55.425
0.40x	0	91.026	0	64.583	0	37.694	100	38.748	100	51.722
0.50x	2.564	98.718	2.083	82.292	0.258	57.616	97.436	57.63	97.917	60.276
0.60x	21.795	100	6.25	90.625	0.818	72.461	78.209	72.461	93.754	73.065
0.70x	37.179	100	9.375	91.667	2.022	82.229	62.854	82.229	90.648	82.65
0.80x	51.282	100	14.583	95.833	3.744	87.866	48.862	87.866	85.499	87.965
0.90x	64.103	100	17.708	96.875	6.756	91.738	36.527	91.738	82.569	91.791
1.00x	73.077	100	26.042	97.917	10.37	93.675	28.851	93.675	74.681	93.698
1.33x	88.462	100	54.167	100	29.819	97.547	31.973	97.547	54.679	97.547
1.66x	98.718	100	83.333	100	48.021	98.752	48.038	98.752	50.831	98.752
2.00x	100	100	97.917	100	63.941	99.269	63.941	99.269	63.975	99.269
2.33x	100	100	98.958	100	79.303	99.613	79.303	99.613	79.31	99.613
2.66x	100	100	98.958	100	88.812	99.656	88.812	99.656	88.818	99.656
3.00x	100	100	98.958	100	94.793	99.742	94.793	99.742	94.799	99.742

Table A-35 ROC plot data for PPAR- γ Crystal pharmacophore model.

CRYS	PPAR- γ High Affinity		PPAR- γ Low Affinity		Decoy Molecules		High Affinity distance from (0,100)		Low Affinity distance from (0,100)	
	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6	6 of 6	5 of 6
1.00x	0	0	0	0	0	0.279	100	100	100	100
1.33x	0	14.815	0	10.084	0	1.582	100	85.2	100	89.93
1.66x	0	48.148	0	30.252	0.047	5.863	100	52.182	100	69.994
2.00x	0	77.778	2.521	48.739	0.14	12.517	100	25.505	97.479	52.767
2.33x	1.852	90.741	5.882	69.748	1.07	21.638	98.154	23.536	94.124	37.194
2.66x	5.556	98.148	15.126	78.992	3.164	32.899	94.497	32.951	84.933	39.034
3.00x	18.519	98.148	42.017	93.277	5.351	44.998	81.657	45.036	58.229	45.497

Table A-36 ROC plot data for PPAR- γ Crystal pharmacophore model at a cutoff of 8Å.

CTF8	PPAR- γ High	PPAR- γ Low	Decoy	High Affinity distance from	Low Affinity distance from
------	---------------------	--------------------	-------	-----------------------------	----------------------------

A	Affinity		Affinity		Molecules		(0,100)		(0,100)	
	5 of 5	4 of 5	5 of 5	4 of 5	5 of 5	4 of 5	5 of 5	4 of 5	5 of 5	4 of 5
0.10x	0	0	0	0	0	0	100	100	100	100
0.20x	0	0	0	0	0	0	100	100	100	100
0.30x	0	0	0	1.681	0	0	100	100	100	98.319
0.40x	0	0	0	1.681	0	0.14	100	100	100	98.319
0.50x	0	11.111	0	1.681	0	0.419	100	88.89	100	98.32
0.60x	0	20.37	0	4.202	0	1.07	100	79.637	100	95.804
0.70x	0	46.296	0	15.966	0	2.839	100	53.779	100	84.082
0.80x	0	74.074	0	31.092	0	4.979	100	26.4	100	69.088
0.90x	0	81.481	0	49.58	0.047	7.538	100	19.994	100	50.98
1.00x	0	90.741	0	60.504	0.14	11.075	100	14.436	100	41.019
1.33x	14.815	98.148	8.403	87.395	0.838	25.361	85.189	25.429	91.601	28.321
1.66x	48.148	100	26.05	95.798	3.443	42.811	51.966	42.811	74.03	43.017
2.00x	77.778	100	47.059	97.479	9.167	59.516	24.039	59.516	53.729	59.569
2.33x	88.889	100	67.227	97.479	17.31	73.197	20.569	73.197	37.064	73.24
2.66x	96.296	100	78.151	97.479	26.989	82.922	27.242	82.922	34.724	82.96
3.00x	98.148	100	92.437	98.319	37.599	90.507	37.645	90.507	38.352	90.523

Table A-37 ROC plot data for PPAR- γ crystal pharmacophore model at a cutoff of 7Å.

CTF7A	PPAR- γ High Affinity	PPAR- γ Low Affinity	Decoy Molecules	High Affinity distance from (0,100)	Low Affinity distance from (0,100)
Radius	4 of 4	4 of 4	4 of 4	4 of 4	4 of 4
0.10x	0	0	0	100	100
0.20x	0	0	0	100	100
0.30x	0	0	0	100	100
0.40x	0	0	0.093	100	100
0.50x	1.852	0	0.233	98.148	100
0.60x	5.556	1.681	0.465	94.445	98.32
0.70x	20.37	9.244	1.303	79.641	90.765
0.80x	53.704	20.168	2.047	46.341	79.858
0.90x	66.667	33.613	3.118	33.479	66.46
1.00x	72.222	47.059	4.839	28.196	53.162
1.33x	96.296	85.714	13.774	14.263	19.845
1.66x	100	92.437	29.223	29.223	30.186
2.00x	100	93.277	44.393	44.393	44.899
2.33x	100	95.798	57.236	57.236	57.39
2.66x	100	96.639	68.218	68.218	68.301

3.00x	100	96.639	77.757	77.757	77.83
--------------	-----	--------	--------	--------	-------

Table A-38 ROC plot data for PPAR- γ NMR pharmacophore model.

NMR	PPAR- γ High Affinity	PPAR- γ Low Affinity	Decoy Molecules	High Affinity distance from (0,100)	Low Affinity distance from (0,100)
Radiu s	3 of 3	3 of 3	3 of 3	3 of 3	3 of 3
0.10x	0	0.84	0.14	100	99.16
0.80x	14.815	23.529	12.564	86.107	77.496
1.00x	22.222	35.294	21.266	80.633	68.111
1.33x	31.481	69.748	37.599	78.157	48.258
1.66x	35.185	84.034	54.863	84.917	57.139
2.00x	38.889	91.597	72.359	94.712	72.845
2.33x	38.889	93.277	85.389	105.004	85.653
3.00x	38.889	94.958	97.115	114.743	97.246

Appendix B. OPLS parameters for Acetate and Methyl Ammonium

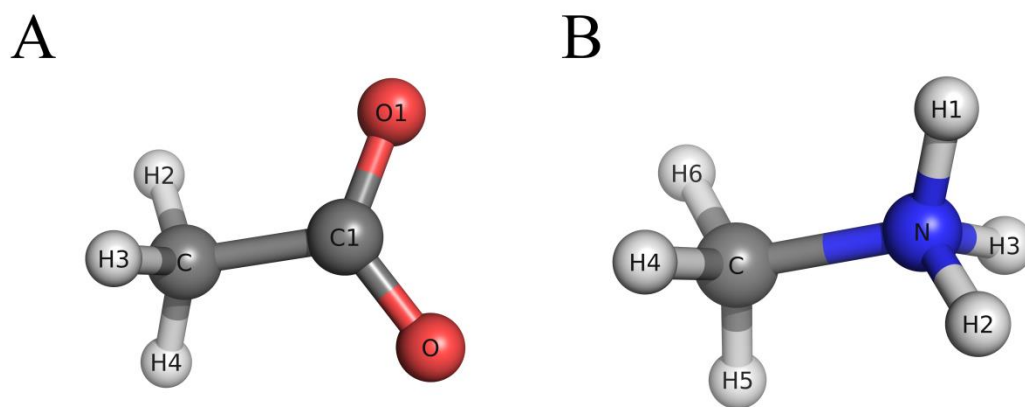


Figure B-1 The names of the atoms within the probes A) Acetate and B) Methyl ammonium used in MixMD are presented.

Table B-1 The OPLS force field parameters used for acetate and methyl ammonium in MixMD simulations are provided in the table below.

Molecule	Atom	Atom Type	q(e)	$\sigma(\text{\AA})$	ϵ (kcal/mol)
Acetate (ACT)	C	c3	-0.28	3.500	0.066
	H2-H4	hc	0.06	2.500	0.030
	C1	c	0.70	3.750	0.105
	O,O1	O	-0.80	2.960	0.210
Methyl Ammonium (MAI)	H1-H3	hn	0.33	0	0
	N	n4	-0.30	3.25	0.17
	C	c3	0.13	3.500	0.066
	H4-H6	hx	0.06	2.500	0.030

Appendix C. Python Script for Calculating MixMD Free energies

```
#!/usr/bin/env python
# Script used for calculating the free energy of probes using MixMD
# Author: Phani Ghanakota, Carlson Lab
# Contact Information: gphani@umich.edu

from __future__ import division
from optparse import OptionParser
import numpy as np
import os, math, sys

usage = "\n This program calculates the free energy of the probes from MixMD simulations\n This script works
with the xplor density output from cpptraj (AmberTools14)\
\nSAMPLE COMMAND\n\
python MixMD_Free_Energy_Calc.py --solvent ACN --xplor_maps 1zys_1P3_16.0-20.0ns_1P3_COM.xplor
--num_snapshots 25000\
"

parser = OptionParser(usage)
parser.add_option("-s", "--solvent", dest="solvent",
                  help="The solvent used in MixMD simulations whose free energies will be calculated", metavar="SOLV")
parser.add_option("-d", "--dir", dest="dir", default = "NONE",
                  help="The directory where the xplormaps are located, if not given, it assumes the maps are in the
current directory", metavar="DIR")
parser.add_option("-n", "--num_snapshots", dest="num_snapshots", type=int,
                  help="The number of MD snapshots used to create the xplor file", metavar="NUMSNAPSHOTS")
parser.add_option("-r", "--xplor_maps", dest="xplor_maps",
                  help="comma separated names of the xplor maps on which this script should run",
metavar="XPLORMAPS")
parser.add_option("-t", "--simulation_temp", dest="temp", default=300, type=float,
                  help="Temperature in Kelvin used for the MixMD simulations", metavar="TEMP")
parser.add_option("-o", "--outputfile", dest="outfile", default = "MixMD_energies.pdb",
                  help="The name of the output file into which the MixMD energies are saved", metavar="OUTFILE")
parser.add_option("-p", "--num_hotspots", dest="num_hotspots", default = 50, type=int,
                  help="The number of MixMD hotspots to save", metavar="HOTSPOT")

# CONSTANTS #####
# List of currently supported probes
probes = ["ACN", "IPA", "1P3"]
# Volume of the probe used for calculating occupancy of probe
probe_volume = {'ACN': 47.1564,
                'IPA': 68.7399,
                '1P3': 75.2784,
                'H2O': 16.5030}
# Expected Occupancy calculated using 7 systems upon which MixMD was performed
expected_occupancy_per_gpt = {'ACN': 0.000071094,
```

```

        'IPA':0.000051085,
        '1P3':0.000046839,
        'H2O':0.004111400}
# The script currently expects an xplor maps obtained from binning in a cube
#####

class XPLORHeader(object):
    """Simple class to represent an XPLOR file header
    """

    def __init__(self, headertext):
        """
        Arguments:
        - `headertext`:text string containing the header
        """

        # Here's an example header, for reference:
        #
        #This line is ignored
        # 1
        #rdparm generated grid density
        # 200 -99 100 200 -99 100 200 -99 100
        # 100.000 100.000 100.000 90.000 90.000 90.000
        #ZYX
        #-99

        """
        There is a discrepancy between the starting grid point reported in the header
        and after the ZYX in the xplor file output by cpptraj Ambertools 14 version.
        Cross checking with results from earlier calculations reveals that the value
        after ZYX is wrong and should be change to match what is mentioned in the
        header. Future updates to this script must be done if and when "DataIO_Xplor.cpp"
        file in cpptraj changes.
        """

        self.headertext = headertext
        headerlines = headertext.split('\n')
        assert len(headerlines) == 7
        assert headerlines[6] == '' # .split('\n') gives a blank entry
            # after the last carriage return.

        gridsize = [int(i) for i in headerlines[3].split()]
        numptsx, numptsy, numptsz = gridsize[0], gridsize[3], gridsize[6]
        firstx, firsty, firstz = gridsize[1], gridsize[4], gridsize[7]
        lastx, lasty, lastz = gridsize[2], gridsize[5], gridsize[8]

        gridlength = [float(i) for i in headerlines[4].split()]
        gridlenx, gridleny, gridlenz = gridlength[0], gridlength[1], gridlength[2]

        # Maybe this headerlines[4] is physical spacing and angles?
        # It sounds lke you don't actually need to know for your purposes.

```

```

something = [float(i) for i in headerlines[4].split()]
assert headerlines[5].strip() == 'ZYX'

# This assertion is necessary because I am not sure this
# algorithm might work for any grid box other than a cube!
assert(numptsx == numptyy == numptsz)
assert(gridlenx == gridleny == gridlenz)

# The length of each grid cube would be
grid_unit_length = gridlenx/numptsx

self.numptsx = numptsx
self.numptyy = numptyy
self.numptsz = numptsz
self.firstx = firstx
self.firsty = firsty
self.firstz = firstz
self.lastx = lastx
self.lasty = lasty
self.lastz = lastz
self.something = something
self.grid_unit_length = grid_unit_length

```

class XPLOrFile(object):

```

    """Read XPLOr ZYX data format into a dict
    In order to save memory only those points
    with no zero grid values will make it into the
    dictionary!
    """
    def __init__(self, fname):
        data = {}
        headertext = ""
        f = open(fname)
        # 3 junk lines
        for i in range(6):
            headertext = headertext + f.next()
        header = XPLOrHeader(headertext)

        # Now we actually read in the data.
        # I could be wrong, but I'm assuming the ZYX format means:
        # 1. write the Z value on a line by itself.
        # 2. For each Y value, write out the numptsx X values in groups of 6.
        zs = range(header.firstz,header.lastz+1)
        ys = range(header.firsty,header.lasty+1)
        xs = range(header.firstx,header.lastx+1)
        for z in zs:
            assert int(f.next()) == z+1
            for y in ys:
                values = []
                for i in range(int(np.ceil(header.numptyy/6))): # 6 values per line
                    values.extend([float(j) for j in f.next().split()])
                assert len(values) == header.numptsx

```

```

        for (xi,x) in enumerate(xs):
            data[(x,y,z)] = values[xi] # MODIFIED FROM BELOW TO ALLOW NON ZERO VALUES
                #if values[xi]:
                    # data[(x,y,z)] = values[xi]
        self.data = data
        self.header = header

class hotspot():
    """
    A class to hold information regarding MixMD hotspots
    """

    def __init__(self,grid_point):
        # The center remains the same regardless of probeocc or volocc
        self.gridx,self.gridy,self.gridz = grid_point
        self.realx = 0
        self.realy = 0
        self.realz = 0

        # Stats for the probeocc (This is for the volume of the probe)
        self.cum_gvalue = 0
        self.cum_enpts = 0 # This will be the number of grid points with value greater than the average
        self.cum_ngpts = 0
        self.cum_nzpts = 0
        self.cum_zpts = 0
        self.cum_nanpts = 0
        self.occ = 0
        self.free_energ = 0
        # Each spot in the spots list will have a tuple of the real x,y,z coordinates and the
        # grid bin count / num snapshots -> the occupancy of that grid point!
        self.spots = []

    def __lt__(self,other):
        #return self.voloccrad < other.voloccrad
        return self.occ < other.occ

    def get_enclosing_box_indices(k,r):
        """
        get all the indices that lie within an enclosing box
        """
        list_of_indices = []
        for x in range(k[0]-r,k[0]+r+1):
            for y in range(k[1]-r, k[1]+r+1):
                for z in range(k[2]-r,k[2]+r+1):
                    list_of_indices.append((x,y,z))
        return list_of_indices

    def write_pdb_hotspots(file_out_name,input_hotspot_list,probe,verbose=0):
        serno = 1
        resno = 1
        finalout = open(file_out_name, 'w')
        # We have to sort the input hotspot list based on the radius,
        # the smaller the radius the tighter the binding and the lesser the entropy!!

```

```

    if verbose:

        finalout.write("ATOM,SERNO,ATOM_NAME,RES_NAME,CHAIN_NAME,RES_NUM,X,Y,Z,OCC,BFACTOR,\tO
CC_ENERG,\tOCC,\tCUM_NGPTS,\tCUM_NZPTS,\tCUM_ZPTS,\tCUM_NANPTS,\tCUM_ENPTS\n")
        # The hotspots will be written starting with the probe with the most favourable free energy
        for spot in sorted(input_hotspot_list,key=lambda x: x.free_ener):
            #print "location (%s,%s,%s) --> volume occupancy %s"%(spot.realx,spot.realy,spot.realz,spot.volocc)
            if verbose == 0:
                finalout.write("%-6s%5d %-4s%3s %s%4d  %8.3f%8.3f%8.3f%6.2f
%f\n"%( "ATOM",serno,"XX", "UNX", "A",resno,

                                spot.realx,

                                spot.realy,

                                spot.realz,

                                0,

spot.free_ener))
            else:
                finalout.write("#REMARK SITE %03d\n"%resno)
                finalout.write("%-6s%5d %-4s%3s %s%4d
%8.3f%8.3f%8.3f%6.2f%6.2f\t%f\t%f\t%f\t%f\t%f\t%f\t%f\n"%( "ATOM",serno,"XX", "UNX", "A",resno,

                                spot.realx,

                                spot.realy,

                                spot.realz,

                                0,0,

spot.free_ener,

                                spot.occ,

                                spot.cum_ngpts,

                                spot.cum_nzpts,

                                spot.cum_zpts,

                                spot.cum_nanpts,

                                spot.cum_enpts))
                resno += 1
        finalout.close()

def generate_probe_occ_map_to_volocc(map,probe,num_snapshots,file_out_name):
    """
    This function performs the bulk of the free energy calculation from MixMD simulations
    """

```

```

grid = XPLOrFile(map)

#####
# Calculate the radius in grid dimensions
try:
    volume_of_sphere = probe_volume[probe]
except KeyError:
    print "error, there was a mistake in recognizing the probe"
    sys.exit()
    radius_of_sphere = math.pow(3 * volume_of_sphere / (4.0 * math.pi), 1/3.0)
    sqr_radius_of_sphere = math.pow(radius_of_sphere, 2)
    # This will make it faster to compare distances
    # Now we need to convert these into values that make sense in the grid dimensions!!
    # i.e., from angstroms to units in the grid box unit length
    grid_radius_of_sphere = radius_of_sphere/grid.header.grid_unit_length
    grid_sqr_radius_of_sphere = math.pow(radius_of_sphere/grid.header.grid_unit_length, 2)
    print "The radius of the sphere is %f"%radius_of_sphere
#####

gpts_loc = {} # This is the dictionary of the center of all the probes
              # that constantly gets updated as and when new probes are
              # created.
              # This dictionary holds hotspot objects

hotspot_list = [] # what is hotspot list??

for gpt in sorted(grid.data, key=lambda x: grid.data[x], reverse=True):

    one
    continue_flag = 0 # This flag is used to check if the newly created point clashes with an older

    cum_gvalue = 0 # The cumulative grid value, in older versions called final_grid_value
    cum_enpts = 0 # The number of points with occupancy
    cum_ngpts = 0 # The total no. of points that are required to add to 1
    cum_nzpts = 0 # The total number of non zero points
    cum_zpts = 0 # The total number of zero points
    cum_nanpts = 0 # The total number of removed points

    # There can be a possible scenario where during an earlier grid point (higher grid value),
    # the one near it is removed (i.e., added) and this does not get updated in the for loop above
    # so we need to check if that data point has been removed by checking if it has the value 'nan'
    if math.isnan(grid.data[gpt]):
        continue

    # We check to make sure that the new probe we create does not overlap with the old one!
    for chkpt in gpts_loc:
        #print "chkpt is ", chkpt
        if ((chkpt[0]-gpt[0])**2 + (chkpt[1]-gpt[1])**2 + (chkpt[2]-gpt[2])**2) < 4 *
grid_sqr_radius_of_sphere:
            continue_flag = 1
            break
    if continue_flag == 1:
        # We "continue" since creating a probe from this grid point would lead to two probes

```

```

# clashing with each other, so while we prevent the creation of a probe here we
# do not delete the grid point thereby making it available for use by probes
# created at other centers
continue # This continue is for the gpt under question

# I had to do an int on the radius_of_sphere, since, the grid points are integers,
# so in order to avoid confusion I converted it to int, instead making the grid points
# float would be better.

new_hotspot = hotspot(gpt) # what about the realx, really, realz????

for index in get_enclosing_box_indices(gpt,int(math.ceil(grid_radius_of_sphere))):
    # The get_enclosing_box_indices may also return some non
    # existent indices, but since we check to see if it is
    # the "data" dictionary, it shouldn't matter.
    # This scenario occurs more commonly for grid indices at the
    # corner of the entire grid! Since we deal with the top few sites
    # We will not encounter these out of the grid ones, however, we
    # may need to implement a variable to track these later on!

    if index in grid.data:
        # Need to check if this index point is inside the sphere
        if ((index[0]-gpt[0])**2 + (index[1]-gpt[1])**2 + (index[2]-gpt[2])**2) <=
grid_sqr_radius_of_sphere:
            cum_ngpts += 1
            # I honestly think we will never hit the first condition!
            if math.isnan(grid.data[index]):
                cum_nanpts += 1
                new_hotspot.spots.append((index[0] *
grid.header.grid_unit_length,
            index[1] * grid.header.grid_unit_length,
            index[2] * grid.header.grid_unit_length,
            0))
            elif grid.data[index] != 0:
                cum_nzpts += 1
                cum_gvalue += grid.data[index]
                if grid.data[index]/num_snapshots >
expected_occupancy_per_gpt[probe]:
                    cum_enpts += 1
            else:
                cum_zpts += 1

                new_hotspot.spots.append((index[0] * grid.header.grid_unit_length,
            index[1] * grid.header.grid_unit_length,
            index[2] * grid.header.grid_unit_length,
            grid.data[index]/num_snapshots))

            # we will make use of 'nan' to tell us that we have finished reading the
data point!

            if math.isnan(grid.data[index]):
                continue
            else:
                grid.data[index] = float('nan') # Points that we processed will

```


have the grid vaule of nan

```
#print gpt, " grid value ", cum_gvalue/num_snapshots
gpts_loc[gpt] = cum_gvalue
#print
"ngpts,nzpts,zpts,nanpts",cum_volocc_ngpts,cum_volocc_nzpts,cum_volocc_zpts,cum_volocc_nanpts
new_hotspot.cum_ngpts = cum_ngpts
new_hotspot.cum_gvalue = cum_gvalue
new_hotspot.cum_enpts = cum_enpts
new_hotspot.cum_nzpts = cum_nzpts
new_hotspot.cum_zpts = cum_zpts
new_hotspot.cum_nanpts = cum_nanpts
#print
"ngpts,nzpts,zpts,nanpts",new_hotspot.cum_volocc_ngpts,new_hotspot.cum_volocc_nzpts,new_hotspot.cum_vol
occ_zpts,new_hotspot.cum_volocc_nanpts
new_hotspot.occ = cum_gvalue/num_snapshots
new_hotspot.realx = gpt[0] * grid.header.grid_unit_length
new_hotspot.realy = gpt[1] * grid.header.grid_unit_length
new_hotspot.realz = gpt[2] * grid.header.grid_unit_length
new_hotspot.free_ener = -RT_VALUE *
math.log((new_hotspot.occ/(expected_occupancy_per_gpt[probe]*new_hotspot.cum_ngpts)),math.e)
hotspot_list.append(new_hotspot)

print "hotspot number %03d out of %03d is being
processed"%(len(hotspot_list),NUM_HOTSPOTS_TO_WRITE)

if len(hotspot_list) >= NUM_HOTSPOTS_TO_WRITE:
    break

# Now write all the hotspot information
write_pdb_hotspots(file_out_name,hotspot_list,probe)

if __name__ == '__main__':

    (options, args) = parser.parse_args()
    # Number of hotspots to report
    NUM_HOTSPOTS_TO_WRITE = options.num_hotspots

    RT_VALUE = 1.9872041 * 0.001 * options.temp # The units are kcal/mol and 10^-3 for the R is shown as
0.001

    # Verfiy that we support this probe
    probe = options.solvent
    if probe not in probes:
        sys.exit("The probe %s is currently not supported. Exiting...."%probe)

    # Process the xplor maps
    maplist = options.xplor_maps.split(",")
    if options.dir != "NONE":
        maplist = [os.path.join(options.dir,map) for map in maplist]

    print "List of xplor files being processed are .... ", maplist
```

```
for map in maplist:
    "Processing xplor file %s"%map
    if not os.path.exists(map):
        print "File %s does not exists skipping it"
        continue
    generate_probe_occ_map_to_volocc(map,probe,options.num_snapshots,options.outfile)
```

Appendix D. Python Script for Calculating MixMD Entropies

```
#!/usr/bin/env python
# Script used for calculating a relative ranking of the entropy of probes using MixMD
# Author: Phani Ghanakota, Carlson Lab
# Contact Information: gphani@umich.edu

from __future__ import division
from optparse import OptionParser
from scipy.stats.kde import gaussian_kde
import matplotlib as mpl
mpl.use('Agg')
import matplotlib.pyplot as plt
import numpy as np
import os, math, sys

usage = "\n This program normalizes the xplor map at the level of a single probe location\
\n It will take in a list of x,y,z coordinates in a file which correspond to the location of\
\n favourable sites, and normalizes the single atom xplor density map\
\n This script works with the xplor density output from cpptraj (AmberTools14)\
\n This script works on the xplor density files generated from single atom binning\
\n Input the free energy output file from MixMD_Free_Energy_Calc.py\
\n SAMPLE COMMAND\n\
python MixMD_Entropy_Per_Probe_Normalized_Maps.py --solvent ACN --hotspot_file=xyz.pdb --
xplor_maps normalized_per_probe_xplor --num_snapshots 25000\
"

parser = OptionParser(usage)
parser.add_option("-s", "--solvent", dest="solvent",
                  help="The solvent used in MixMD simulations", metavar="SOLV")
parser.add_option("-d", "--dir", dest="dir", default = "NONE",
                  help="The directory where the xplor map being processed is located, if not given, it assumes the maps
are in the current directory", metavar="DIR")
parser.add_option("-n", "--num_snapshots", dest="num_snapshots", type=int,
                  help="The number of MD snapshots used to create the xplor file", metavar="NUMSNAPSHOTS")
parser.add_option("-r", "--xplor_map", dest="xplor_map",
                  help="Name of the xplor map on which this script should run", metavar="XPLORMAPS")
parser.add_option("-o", "--outputfile", dest="outfile", default = "MixMD_entropies.pdb",
                  help="The name of the output file into which the MixMD entropies are saved", metavar="OUTFILE")
parser.add_option("-p", "--num_hotspots", dest="num_hotspots", default = 50, type=int,
                  help="The number of MixMD hotspots to save", metavar="HOTSPOT")
parser.add_option("-i", "--hotspotfile", dest="hotspot_file",
                  help="The file generated by the MixMD_Free_Energy_Calc.py that contains the energies of the probes
used in MixMD", metavar="ENERGFILE")

# CONSTANTS #####
```

```

# List of currently supported probes
probes = ["ACN", "IPA", "1P3"]
# Volume of the probe used for calculating occupancy of probe
probe_volume = {'ACN': 47.1564,
                'IPA': 68.7399,
                '1P3': 75.2784,
                'H2O': 16.5030}

expected_occupancy_per_gpt = {'ACN':0.000071094,
                              'IPA':0.000051085,
                              '1P3':0.000046839,
                              'H2O':0.004111400}

# The script currently expects an xplor maps obtained from binning in a cube
#####
def splitseq(seq,size):
    """ Split up seq in pieces of size

    Arguments:
    - `seq`:the sequence
    - `size`:the size of the chunks.

    In [34]: u.splitseq(range(30),10)
    Out[34]:
    [[0, 1, 2, 3, 4, 5, 6, 7, 8, 9],
     [10, 11, 12, 13, 14, 15, 16, 17, 18, 19],
     [20, 21, 22, 23, 24, 25, 26, 27, 28, 29]]

    In [35]: u.splitseq(range(34),10)
    Out[35]:
    [[0, 1, 2, 3, 4, 5, 6, 7, 8, 9],
     [10, 11, 12, 13, 14, 15, 16, 17, 18, 19],
     [20, 21, 22, 23, 24, 25, 26, 27, 28, 29],
     [30, 31, 32, 33]]
    """
    try:
        return [seq[i:i+size] for i in range(0, len(seq), size)]
    except ValueError:
        print "Cannot split this seq",seq
        print "Into this size",size
        raise

class XPLORHeader(object):
    """Simple class to represent an XPLOR file header
    """

    def __init__(self, headertext):
        """

        Arguments:
        - `headertext`:text string containing the header
        """

    # Here's an example header, for reference:

```

```

#
#This line is ignored
# 1
#rdparm generated grid density
# 200 -99 100 200 -99 100 200 -99 100
# 100.000 100.000 100.000 90.000 90.000 90.000
#ZYX
#-99

```

```

'''

```

There is a discrepancy between the starting grid point reported in the header and after the ZYX in the xplor file output by cpptraj Ambertools 14 version. Cross checking with results from earlier calculations reveals that the value after ZYX is wrong and should be change to match what is mentioned in the header. Future updates to this script must be done if and when "DataIO_Xplor.cpp" file in cpptraj changes.

```

'''

```

```

self.headertext = headertext
headerlines = headertext.split('\n')
assert len(headerlines) == 7
assert headerlines[6] == '' # .split('\n') gives a blank entry
                        # after the last carriage return.

```

```

gridsize = [int(i) for i in headerlines[3].split()]
numptsx, numptsy, numptsz = gridsize[0], gridsize[3], gridsize[6]
firstx, firsty, firstz = gridsize[1], gridsize[4], gridsize[7]
lastx, lasty, lastz = gridsize[2], gridsize[5], gridsize[8]

```

```

        gridlength = [float(i) for i in headerlines[4].split()]
        gridlenx, gridleny, gridlenz = gridlength[0], gridlength[1], gridlength[2]

```

```

# Maybe this headerlines[4] is physical spacing and angles?
# It sounds lke you don't actually need to know for your purposes.
something = [float(i) for i in headerlines[4].split()]
assert headerlines[5].strip() == 'ZYX'

```

```

        # This assertion is necessary because I am not sure this
        # algorithm might work for any grid box other than a cube!
        assert(numptsx == numptsy == numptsz)
        assert(gridlenx == gridleny == gridlenz)

```

```

        # The length of each grid cube would be
        grid_unit_length = gridlenx/numptsx

```

```

self.numptsx = numptsx
self.numptsy = numptsy
self.numptsz = numptsz
self.firstx = firstx
self.firsty = firsty
self.firstz = firstz
self.lastx = lastx

```

```

self.lasty = lasty
self.lastz = lastz
self.something = something
self.grid_unit_length = grid_unit_length

```

class XPLOrFile(object):

"""Read XPLOr ZYX data format into a dict

In order to save memory only those points
with no zero grid values will make it into the
dictionary!

"""

def __init__(self, fname):

```

    data = {}
    headertext = ""
    f = open(fname)
    # 3 junk lines
    for i in range(6):
        headertext = headertext + f.next()
    header = XPLOrHeader(headertext)

```

```

    # Now we actually read in the data.
    # I could be wrong, but I'm assuming the ZYX format means:
    # 1. write the Z value on a line by itself.
    # 2. For each Y value, write out the numptsx X values in groups of 6.
    zs = range(header.firstz,header.lastz+1)
    ys = range(header.firsty,header.lasty+1)
    xs = range(header.firstx,header.lastx+1)

```

```

    for z in zs:
        assert int(f.next()) == z+100
        for y in ys:
            values = []
            for i in range(int(np.ceil(header.numptsy/6))): # 6 values per line
                values.extend([float(j) for j in f.next().split()])
            assert len(values) == header.numptsx
            for (xi,x) in enumerate(xs):
                data[(x,y,z)] = values[xi] # MODIFIED FROM BELOW TO ALLOW NON ZERO VALUES as

```

Heather want's a count of this!!

```

                #if values[xi]:
                #    data[(x,y,z)] = values[xi]
            self.data = data
        self.header = header

```

def write(self,fname):

"""

Arguments:

- `fname`:name of file to write.

"""

```

f = file(fname,'w')
f.write(self.header.headertext)
# This is based on those same assumptions of how the ZYX

```

```

# format actually works, so it could easily be wrong.
zs = range(self.header.firstz,self.header.lastz+1)
ys = range(self.header.firsty,self.header.lasty+1)
xs = range(self.header.firstx,self.header.lastx+1)
for z in zs:
    fm = z + 100
    f.write('%8i\n'%fm)
    for y in ys:
        values = []
        for x in xs:
            values.append(self.data[(x,y,z)])
        for chunk in splitseq(values,6):
            line = ['%12.5f'%c for c in chunk]
            line = ''.join(line) + '\n'
            f.write(line)
f.close()

```

```
class hotspot():
```

```
'''
```

```
A class to hold information regarding MixMD hotspots
```

```
'''
```

```
def __init__(self,grid_point):
```

```
# The center remains the same regardless of probeocc or volocc
```

```
self.gridx,self.gridy,self.gridz = grid_point
```

```
self.realx = 0
```

```
self.realy = 0
```

```
self.realz = 0
```

```
# Stats for the probeocc (This is for the volume of the probe)
```

```
self.cum_gvalue = 0
```

```
self.cum_enpts = 0 # This will be the number of grid points with value greater than the average
```

```
self.cum_ngpts = 0
```

```
self.cum_nzpts = 0
```

```
self.cum_zpts = 0
```

```
self.cum_nanpts = 0
```

```
self.occ = 0
```

```
#self.free_energ = 0
```

```
# Each spot in the spots list will have a tuple of the real x,y,z coordinates and the
```

```
# grid bin count / num snapshots -> the occupancy of that grid point!
```

```
self.spots = []
```

```
def __lt__(self,other):
```

```
#return self.voloccrad < other.voloccrad
```

```
return self.occ < other.occ
```

```
def get_enclosing_box_indices(k,r):
```

```
'''
```

```
get all the indices that lie within an enclosing box
```

```
'''
```

```
list_of_indices = []
```

```
for x in range(k[0]-r,k[0]+r+1):
```

```

        for y in range(k[1]-r, k[1]+r+1):
            for z in range(k[2]-r, k[2]+r+1):
                list_of_indices.append((x,y,z))
    return list_of_indices

def entropy_ranking(file_out_name, input_hotspot_list, probe):
    """
    This section outputs the atom occupancy and entropy of the top x hotspots
    probability is defined as the occupancy at each grid point divided by the total
    occupancy of the probe volume it belongs to.....
    """
    finalout = open(file_out_name, 'w')
    finalout.write("%s%s%s%s\n"%( "X".center(14),
    "Y".center(14), "Z".center(14), "Occupancy[Atom]".center(14), " Entropy[sum(plnp)]".center(14)))
    for spot in input_hotspot_list:
        data = [ind_spot[-1] for ind_spot in spot.spots]
        entp = 0
        for datapoint in data:
            probability = (datapoint/spot.occ)
            if probability != 0:
                entp += probability*math.log(probability)
        print len(data)

    finalout.write("%14.3f%14.3f%14.3f%14.3f%20.6f\n"%(spot.realx, spot.realy, spot.realz, spot.occ, entp))
    finalout.close()

def plot_occupancy_hist(file_out_name, input_hotspot_list, probe):
    for spot_num, spot in enumerate(sorted(input_hotspot_list, key=lambda x: x.occ, reverse=True)):
        spot_num += 1

        data = [ind_spot[-1] for ind_spot in spot.spots]
        # generated a density class
        density = gaussian_kde(data)

        # set the covariance_factor, lower means more detail
        density.covariance_factor = lambda : .25
        density._compute_covariance()

        # generate a range of x values from min to max
        xs = np.linspace(min(data), max(data), 200)

        # fill y values using density class
        ys = density(xs)
        plt.plot(xs, density(xs))

        plt.ylim(0, 500)
        plt.xlim(0, 0.035)
        plt.title("%s %s"%(protein_dir[pdb], probe))
        plt.annotate("HOTSPOT number %02d\nSD %s"%(spot_num, np.std(data)), xy=(0.05, 0.85),
        xycoords='axes fraction')
        plt.savefig("%s_hist_hotspot_%02d.png"%(file_out_name, spot_num))

```



```

plt.clf()

def generate_probe_occ_map_to_volocc(map,probe,num_snapshots,list_of_sites,file_out_name):
    """
    This function performs the bulk of the calculations from MixMD simulations
    """

    grid = XPLOrFile(map)

    #####
    # Calculate the radius in grid dimensions
    try:
        volume_of_sphere = probe_volume[probe]
    except KeyError:
        print "error, there was a mistake in recognizing the 'probe'"
        sys.exit()
        radius_of_sphere = math.pow(3 * volume_of_sphere / (4.0 * math.pi), 1/3.0)
        sqr_radius_of_sphere = math.pow(radius_of_sphere, 2)
        # This will make it faster to compare distances
        # Now we need to convert these into values that make sense in the grid dimensions!!
        # i.e., from angstroms to units in the grid box unit length
        grid_radius_of_sphere = radius_of_sphere/grid.header.grid_unit_length
        grid_sqr_radius_of_sphere = math.pow(radius_of_sphere/grid.header.grid_unit_length, 2)
        print "The radius of the sphere is %f"%radius_of_sphere
    #####

    gpts_loc = {} # This is the dictionary of the center of all the probes
                 # that constantly gets updated as and when new probes are
                 # created.
                 # This dictionary holds hotspot objects

    hotspot_list = [] # what is hotspot list??

    #for gpt in sorted(grid.data, key=lambda x: grid.data[x], reverse=True):
    for gpt in
    [(int(pt[0]/grid.header.grid_unit_length),int(pt[1]/grid.header.grid_unit_length),int(pt[2]/grid.header.grid_unit_le
    ngth)) for pt in list_of_sites]:
        #print "Processing entropy for the grid point ", gpt

        continue_flag = 0 # This flag is used to check if the newly created point clashes with an older
one
        cum_gvalue = 0 # The cumulative grid value,in older versions called final_grid_value
        cum_enpts = 0 # The number of points with occupancy
        cum_ngpts = 0 # The total no. of points that are required to add to 1
        cum_nzpts = 0 # The total number of non zero points
        cum_zpts = 0 # The total number of zero points
        cum_nanpts = 0 # The total number of removed points

        # There can be a possible scenario where during an earlier grid point (higher grid value),
        # the one near it is removed (i.e., added) and this does not get updated in the for loop above
        # so we need to check if that data point has been removed by checking if it has the value 'nan'
        if math.isnan(grid.data[gpt]):
            continue

```

```

# We check to make sure that the new probe we create does not overlap with the old one!
for chkpt in gpts_loc:
    #print "chkpt is ", chkpt
    if ((chkpt[0]-gpt[0])**2 + (chkpt[1]-gpt[1])**2 + (chkpt[2]-gpt[2])**2) < 4 *
grid_sqr_radius_of_sphere:
    continue_flag = 1
    break
if continue_flag == 1:
    # We "continue" since creating a probe from this grid point would lead to two probes
    # clashing with each other, so while we prevent the creation of a probe here we
    # do not delete the grid point thereby making it available for use by probes
    # created at other centers
    continue # This continue is for the gpt under question

# I had to do an int on the radius_of_sphere, since, the grid points are integers,
# so in order to avoid confusion I converted it to int, instead making the grid points
# float would be better.

new_hotspot = hotspot(gpt) # what about the realx, realy, realz????

for index in get_enclosing_box_indices(gpt,int(math.ceil(grid_radius_of_sphere))):
    # The get_enclosing_box_indices may also return some non
    # existent indices, but since we check to see if it is
    # the "data" dictionary, it shouldn't matter.
    # This scenario occurs more commonly for grid indices at the
    # corner of the entire grid! Since we deal with the top few sites
    # We will not encounter these out of the grid ones, however, we
    # may need to implement a variable to track these later on!

    if index in grid.data:
        # Need to check if this index point is inside the sphere
        if ((index[0]-gpt[0])**2 + (index[1]-gpt[1])**2 + (index[2]-gpt[2])**2) <=
grid_sqr_radius_of_sphere:
            cum_ngpts += 1
            # I honestly think we will never hit the first condition!
            if math.isnan(grid.data[index]):
                cum_nanpts += 1
                new_hotspot.spots.append((index[0] *
grid.header.grid_unit_length,
index[1] * grid.header.grid_unit_length,
index[2] * grid.header.grid_unit_length,
0))
            elif grid.data[index] != 0:
                cum_nzpts += 1
                cum_gvalue += grid.data[index]
                if grid.data[index]/num_snapshots >
expected_occupancy_per_gpt[probe]:
                    cum_enpts += 1
            else:
                cum_zpts += 1

                new_hotspot.spots.append((index[0] * grid.header.grid_unit_length,

```

```

index[1] * grid.header.grid_unit_length,
index[2] * grid.header.grid_unit_length,
grid.data[index]/num_snapshots))

# we will make use of 'nan' to tell us that we have finished reading the
data point!

if math.isnan(grid.data[index]):
    continue
else:
    grid.data[index] = float('nan') # Points that we processed will
have the grid vaule of nan

#print gpt, " grid value ", cum_gvalue/num_snapshots
gpts_loc[gpt] = cum_gvalue
#print
"ngpts,nzpts,zpts,nanpts",cum_volocc_ngpts,cum_volocc_nzpts,cum_volocc_zpts,cum_volocc_nanpts
new_hotspot.cum_ngpts = cum_ngpts
new_hotspot.cum_gvalue = cum_gvalue
new_hotspot.cum_enpts = cum_enpts
new_hotspot.cum_nzpts = cum_nzpts
new_hotspot.cum_zpts = cum_zpts
new_hotspot.cum_nanpts = cum_nanpts
#print
"ngpts,nzpts,zpts,nanpts",new_hotspot.cum_volocc_ngpts,new_hotspot.cum_volocc_nzpts,new_hotspot.cum_vol
occ_zpts,new_hotspot.cum_volocc_nanpts
new_hotspot.occ = cum_gvalue/num_snapshots
new_hotspot.realx = gpt[0] * grid.header.grid_unit_length
new_hotspot.realy = gpt[1] * grid.header.grid_unit_length
new_hotspot.realz = gpt[2] * grid.header.grid_unit_length
hotspot_list.append(new_hotspot)

print "hotspot number %03d out of %03d is being
processed"%(len(hotspot_list),NUM_HOTSPOTS_TO_WRITE)

if len(hotspot_list) >= NUM_HOTSPOTS_TO_WRITE:
    break

# Now write the free energy, occupancy and entropy values for the hotspots
entropy_ranking(file_out_name,hotspot_list,probe)

normalized_grid = {} # This dict will contain the x,y,z coordinates in the grid dimension
# and the normalized bin count, we will zero out all the coordinates
# in the data array and write these value back into it and resave it
# The save file will have the name _ENTROPYNORM.xplor

# Zero out the entire grid
for key in grid.data:
    grid.data[key] = 0
for spot in hotspot_list:
    for ind_spot in spot.spots:

grid.data[(int(ind_spot[0]/grid.header.grid_unit_length),int(ind_spot[1]/grid.header.grid_unit_length),int

```

```
(ind_spot[2]/grid.header.grid_unit_length))) = ind_spot[3]/spot.occ
grid.write(os.path.basename(map)[:6] + "_ENTROPY_NORM.xplor")
#plot_occupancy_hist(file_out_name,hotspot_list,probe)
```

```
def list_of_grid_points_to_process(energy_file):
```

```
'''
```

```
This function will process the grid points sorted by free energy
```

```
This file is generated by the MixMD_Free_Energy_Calc.py
```

```
We will generate a tuple of xyz coordinates
```

```
The MixMD energy file format is .....
```

```
ATOM 1 XX UNX A 1 14.500 2.500 10.000 0.00 -3.011455
```

```
ATOM 1 XX UNX A 2 12.000 -9.500 -8.500 0.00 -2.696113
```

```
ATOM 1 XX UNX A 3 17.000 -3.000 5.000 0.00 -2.682999
```

```
ATOM 1 XX UNX A 4 13.000 -23.000 2.500 0.00 -2.588952
```

```
ATOM 1 XX UNX A 5 7.500 -20.500 -8.500 0.00 -2.558617
```

```
ATOM 1 XX UNX A 6 -13.000 0.000 14.000 0.00 -2.549849
```

```
'''
```

```
energ_coord_list = []
```

```
infile = open(energy_file, 'r')
```

```
for idx,line in enumerate(infile.readlines()):
```

```
    if idx+1 > NUM_HOTSPOTS_TO_WRITE:
```

```
        break
```

```
    x = float(line.strip().split()[-5])
```

```
    y = float(line.strip().split()[-4])
```

```
    z = float(line.strip().split()[-3])
```

```
    energ_coord_list.append((x,y,z))
```

```
infile.close()
```

```
#print energ_coord_list
```

```
return energ_coord_list
```

```
if __name__ == '__main__':
```

```
(options, args) = parser.parse_args()
```

```
# Number of hotspots to report
```

```
NUM_HOTSPOTS_TO_WRITE = options.num_hotspots
```

```
num_snapshots = options.num_snapshots
```

```
# Verfiy that we support this probe
```

```
probe = options.solvent
```

```
if probe not in probes:
```

```
    sys.exit("The probe %s is currently not supported. Exiting..."%probe)
```

```
if not options.hotspot_file:
```

```
    sys.exit("You need to specify the file with energies")
```

```
# Process the xplor maps
```

```
map = options.xplor_map
```

```
if options.dir != "NONE":
```

```
    map = os.path.join(options.dir,map)
```

```
print "xplor files being processed is .... ", map
```

```
if not os.path.exists(map):
```

```
    print "File %s does not exists skipping it"
```

```
        sys.exit()
    generate_probe_occ_map_to_volocc(map,probe,num_snapshots,list_of_grid_points_to_process(options.
hotspot_file),options.outfile)
```

Appendix E. MixMD maps for Allosteric Systems

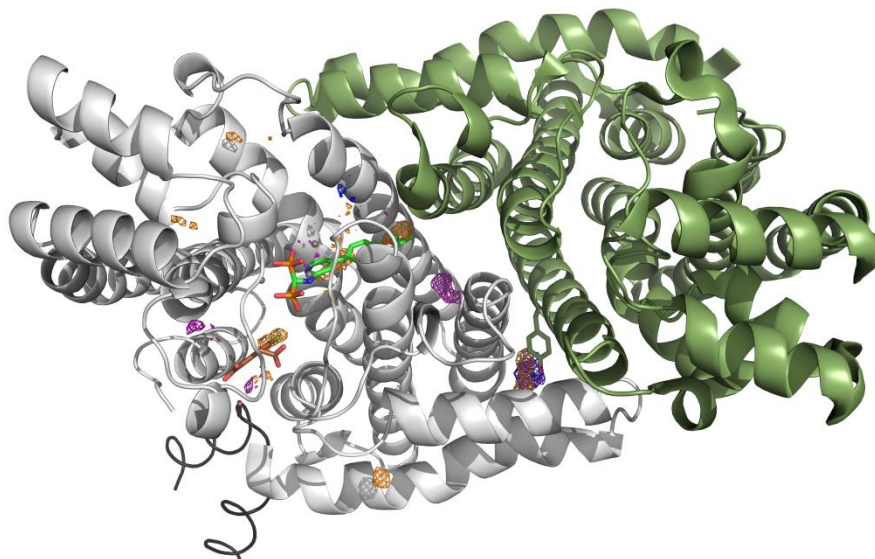


Figure E-1 MixMD maps of Farnesyl Pyrophosphate Synthase (FPPS) contoured at 35σ are shown with examples (where available) of molecules from the PDB database bound in probe mapped locations on the protein surface. FPPS functions as a dimer and a second copy of the dimer counterpart is shown in green with a tyrosine residue rendered as a stick model to illustrate the overlap of this residue with the MixMD maps. A protein packing interface rendered as cartoon is shown using PDB ID: 2P1C (Black). The allosteric and competitive ligands are shown for reference using the crystal structures PDB ID: 3N5J – Brown (Allosteric ligand) and PDB ID: 4DEM – Green (Competitive ligand).

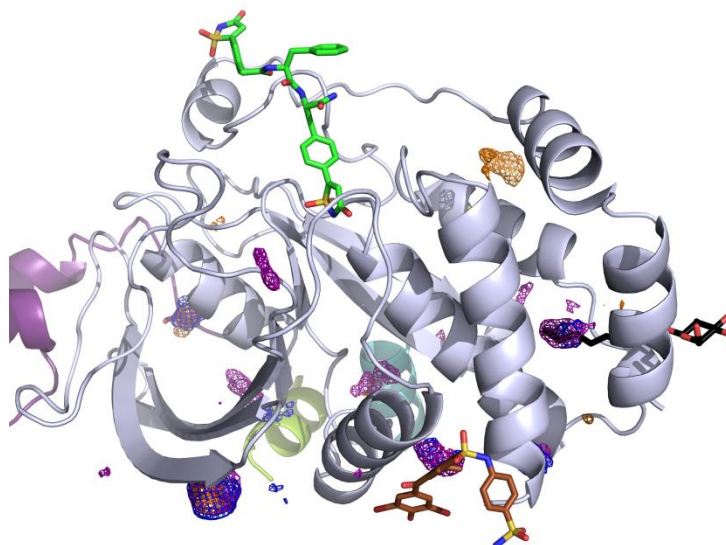


Figure E-2 MixMD maps of Protein Tyrosine Phosphatase 1B (PTP1B) contoured at 35σ are shown with examples (where available) of molecules from the PDB database bound in probe mapped locations on the protein surface. The different protein packing interfaces and examples of cosolvent molecules known to bind PTP1B and mapped by MixMD are color coded as follows, PDB ID: 4GRY – Pea, PDB ID: 2CMC – Cyan, PDB ID: 2CMB – Black, PDB ID: 1GWZ– Purple, PDB ID: 1T49 – Brown (Allosteric ligand) and PDB ID: 2CMB – Green (Competitive ligand).

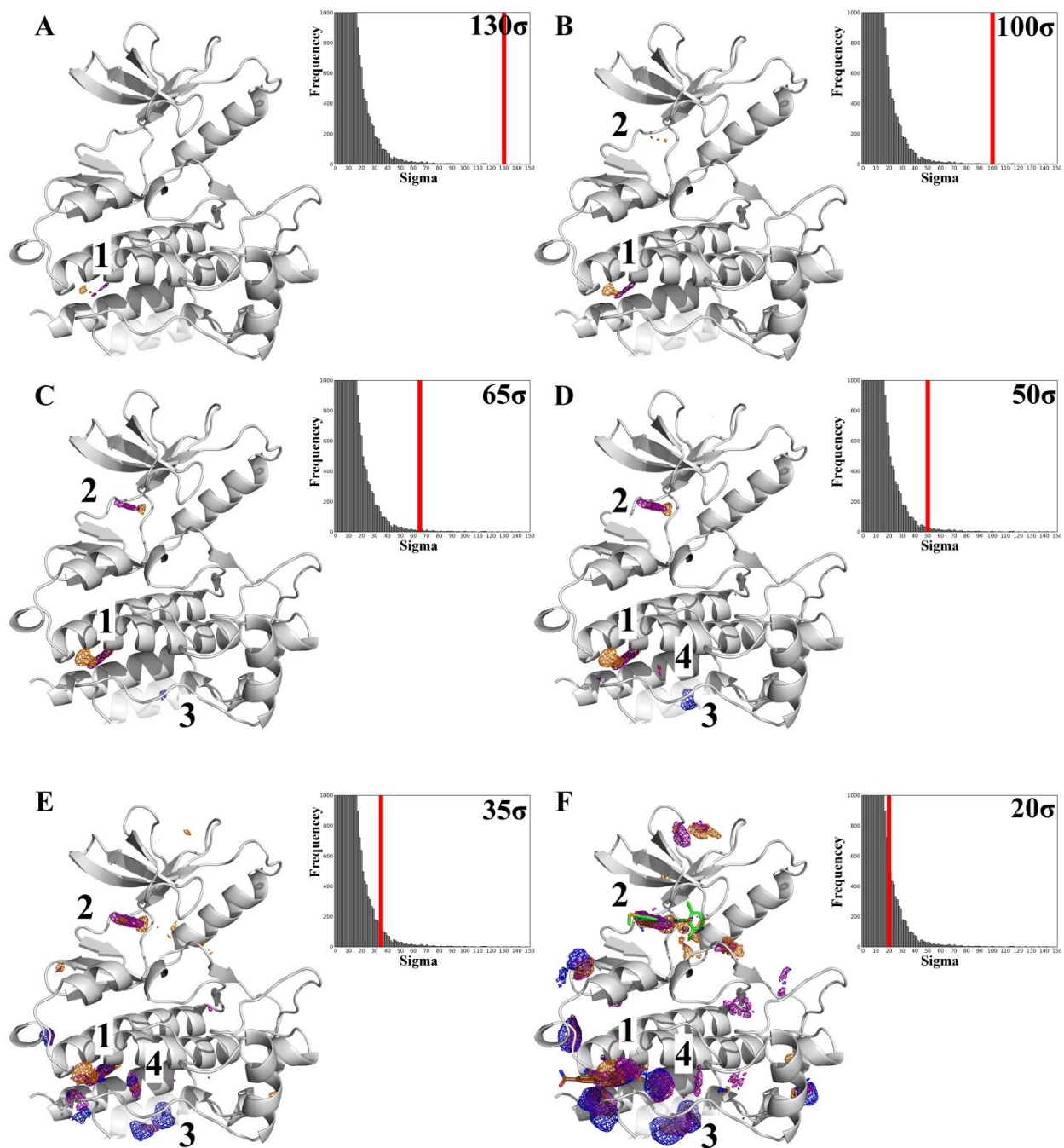


Figure E-3 The ranking of the top-four sites is shown for MixMD simulations starting from the active conformation of Abl Kinase (PDB ID: 1M52) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 130σ to 20σ (Shown in figures A – F).

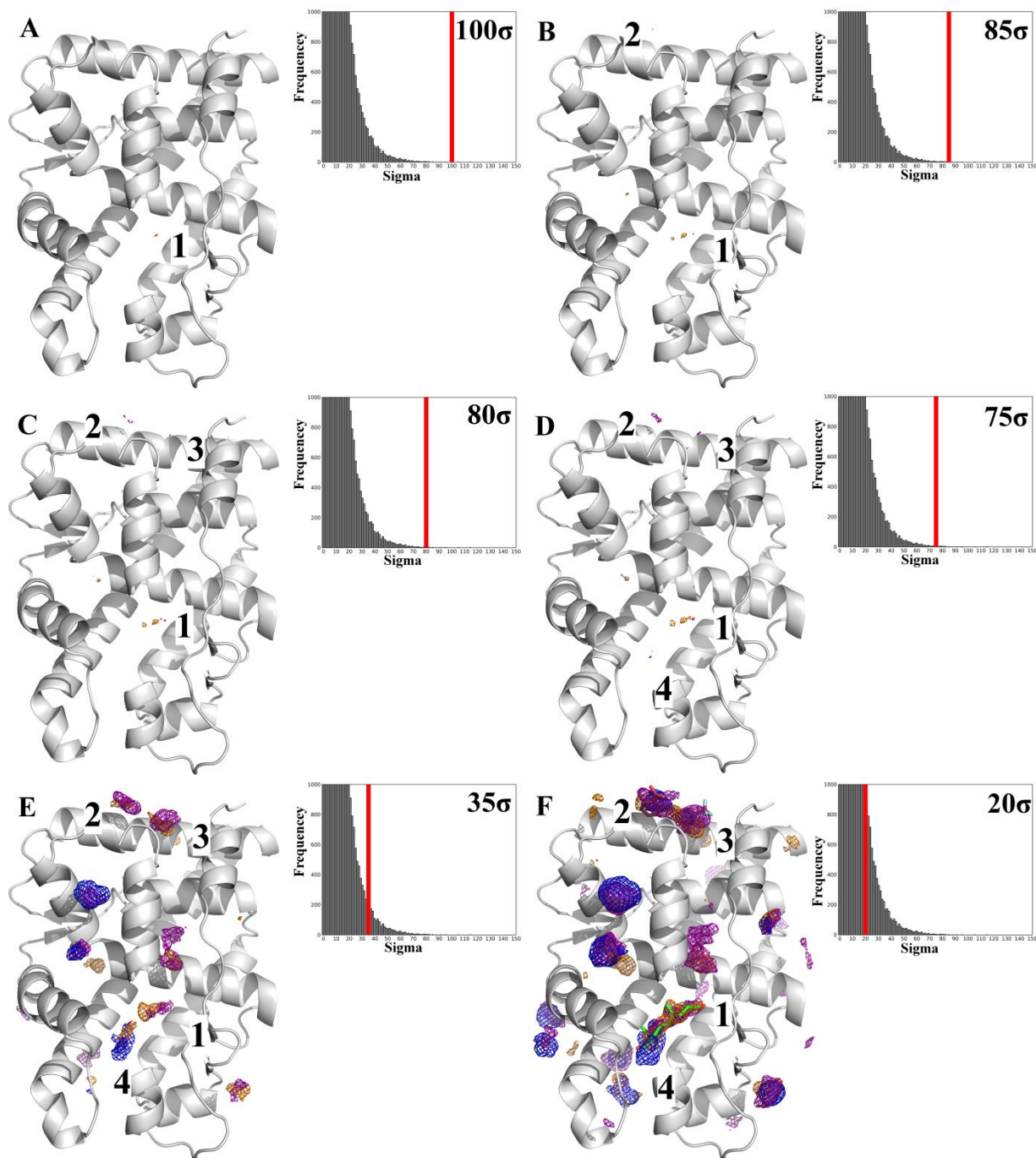


Figure E-4 The ranking of the top-four sites is shown for MixMD simulations starting from Androgen Receptor (PDB ID: 2AM9) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 100σ to 20σ (Shown in figures A – F).

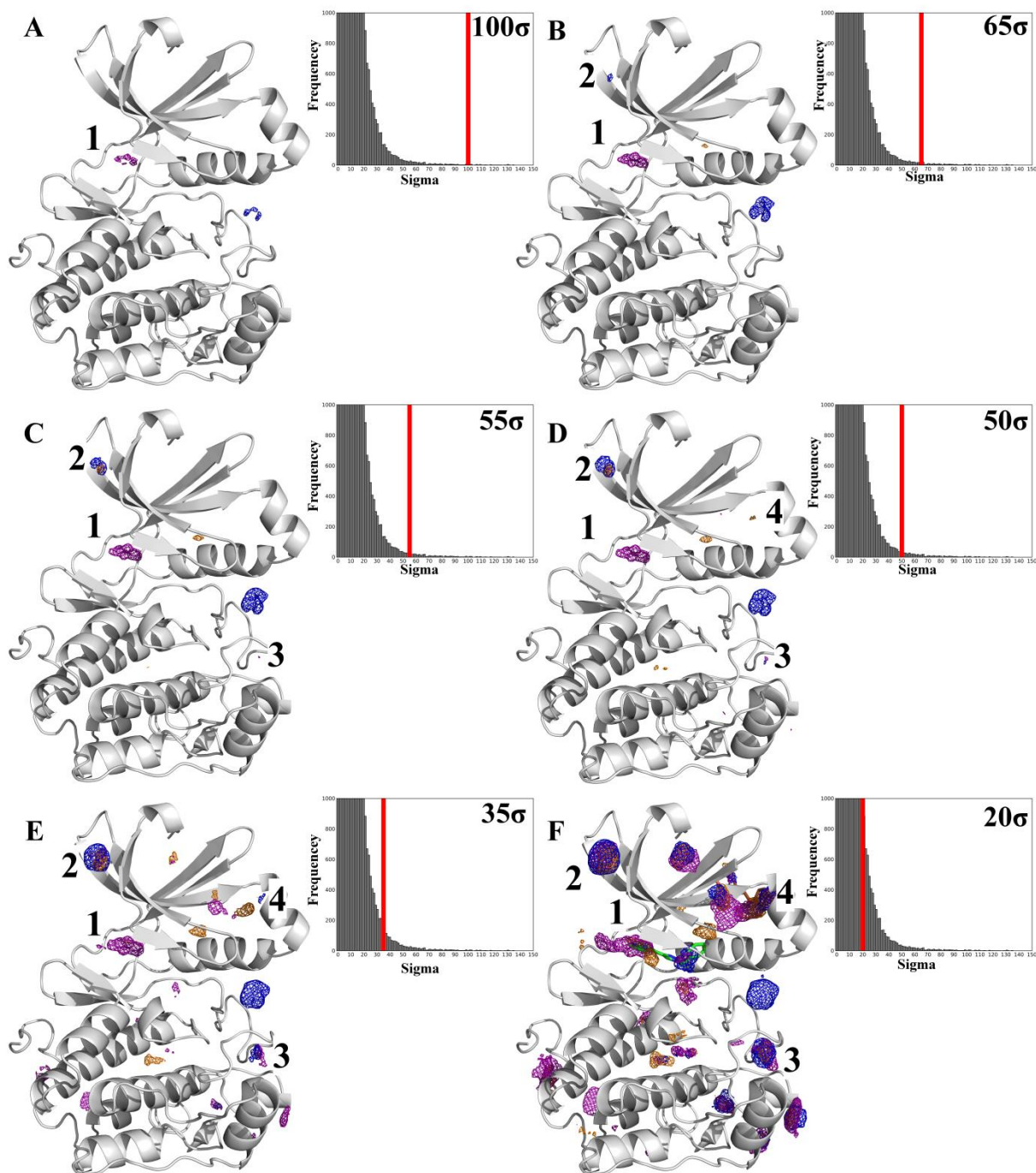


Figure E-5 The ranking of the top-four sites is shown for MixMD simulations starting from Pdk1 Kinase (PDB ID: 3RCJ) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 100σ to 20σ (Shown in figures A – F).

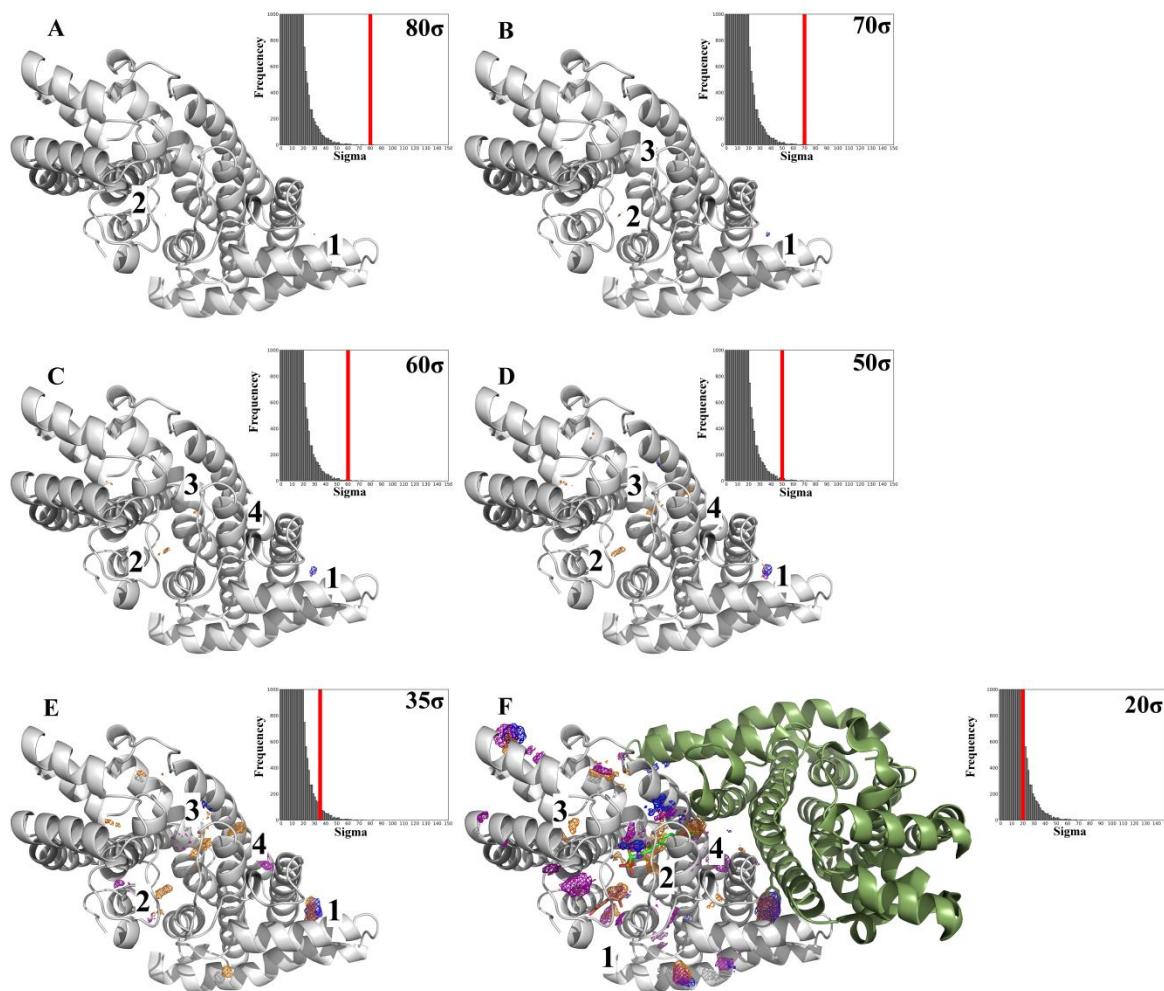


Figure E-6 The ranking of the top-four sites is shown for MixMD simulations starting from Farnesyl Pyrophosphate Synthase (PDB ID: 4DEM) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 80σ to 20σ (Shown in figures A – F).

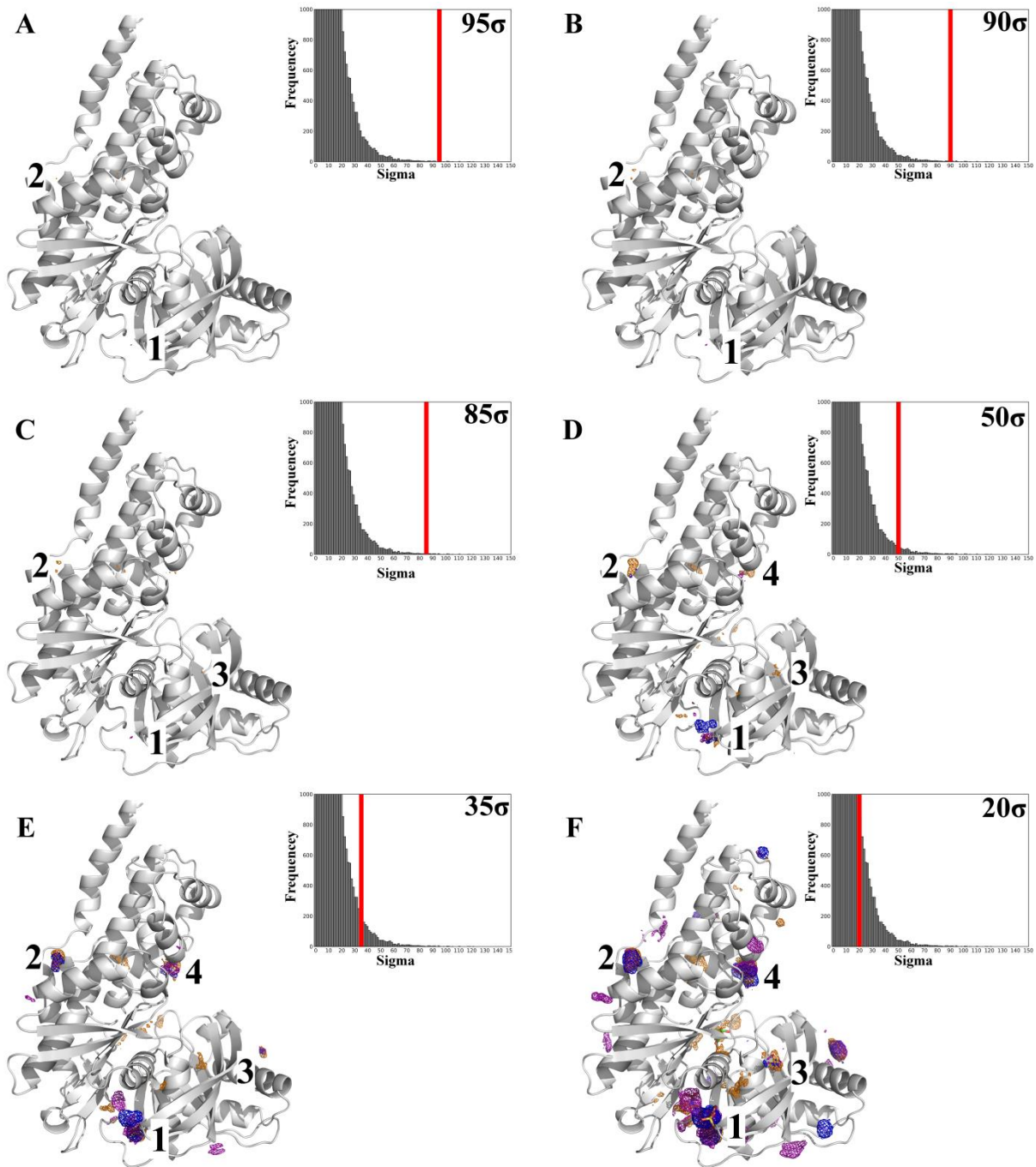


Figure E-7 The ranking of the top-four sites is shown for MixMD simulations starting from Glucokinase (PDB ID: 3IDH) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 95σ to 20σ (Shown in figures A – F).

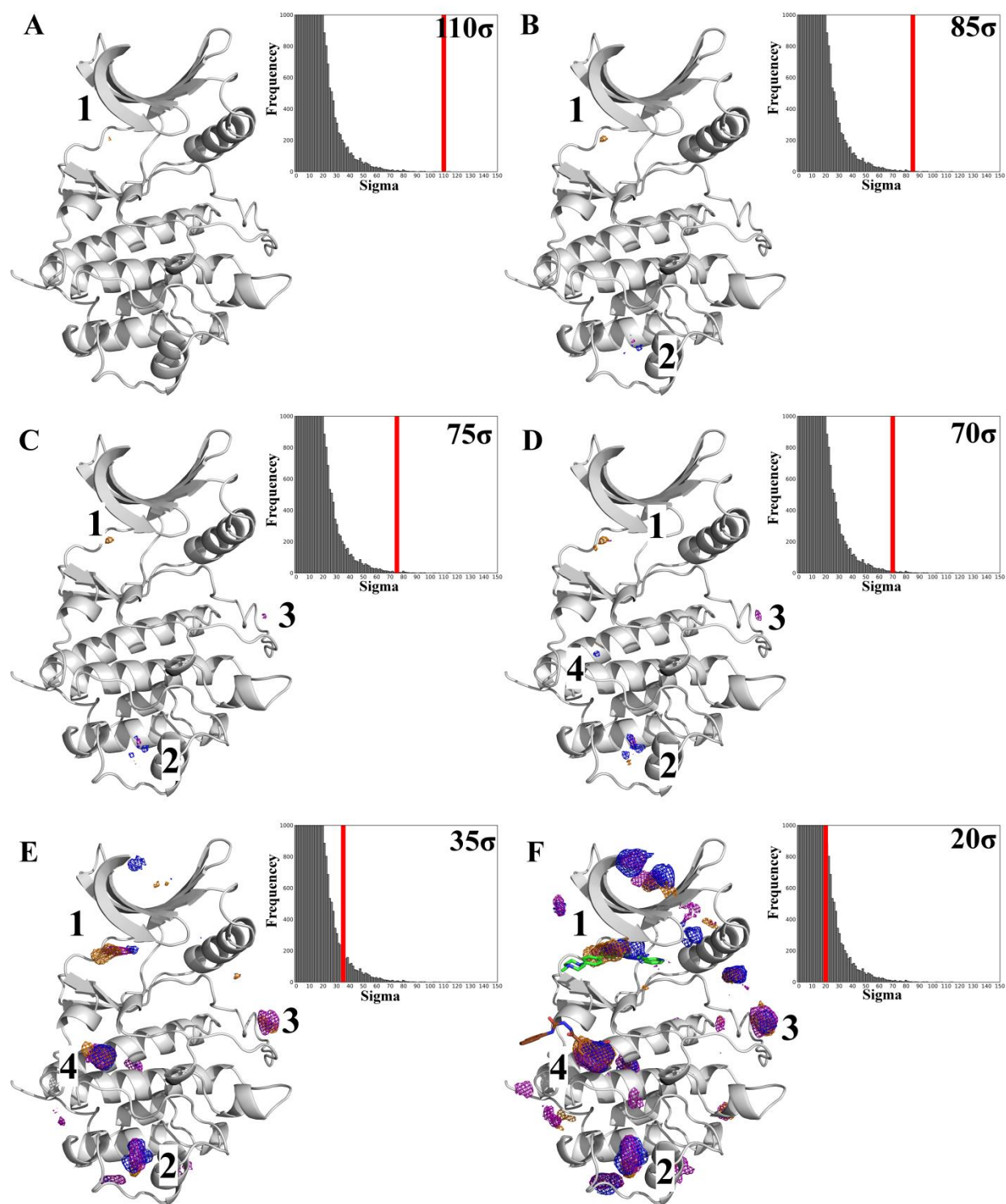


Figure E-8 The ranking of the top-four sites is shown for MixMD simulations starting from CHK1 Kinase (PDB ID: 1ZYS) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple)

probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 110σ to 20σ (Shown in figures A – F).

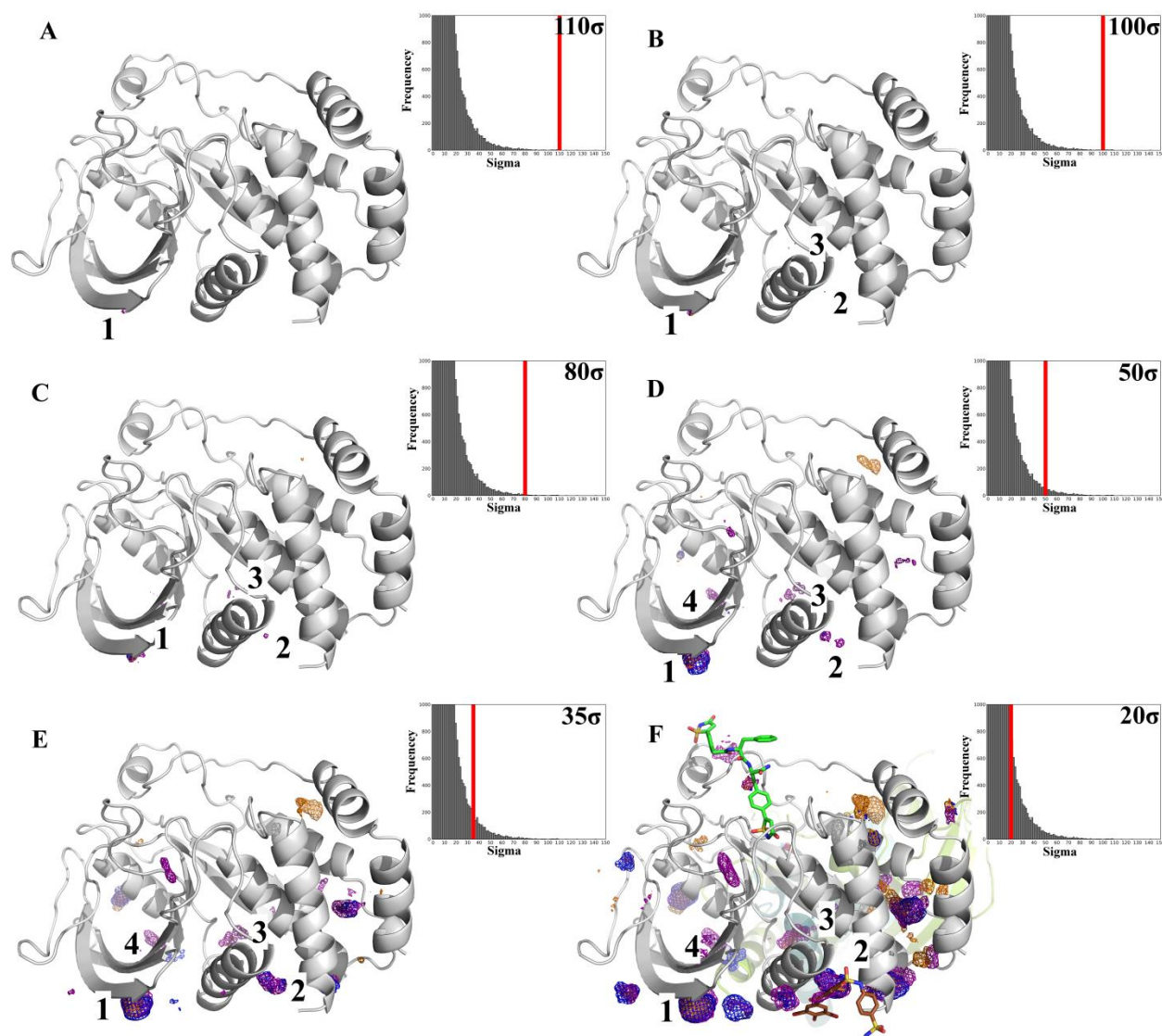


Figure E-9 The ranking of the top-four sites is shown for MixMD simulations starting from Protein Tyrosine Phosphatase 1B (PDB ID: 2CMB) using acetonitrile (orange), isopropanol (blue) and pyrimidine (purple) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 110σ to 20σ (Shown in figures A – F).

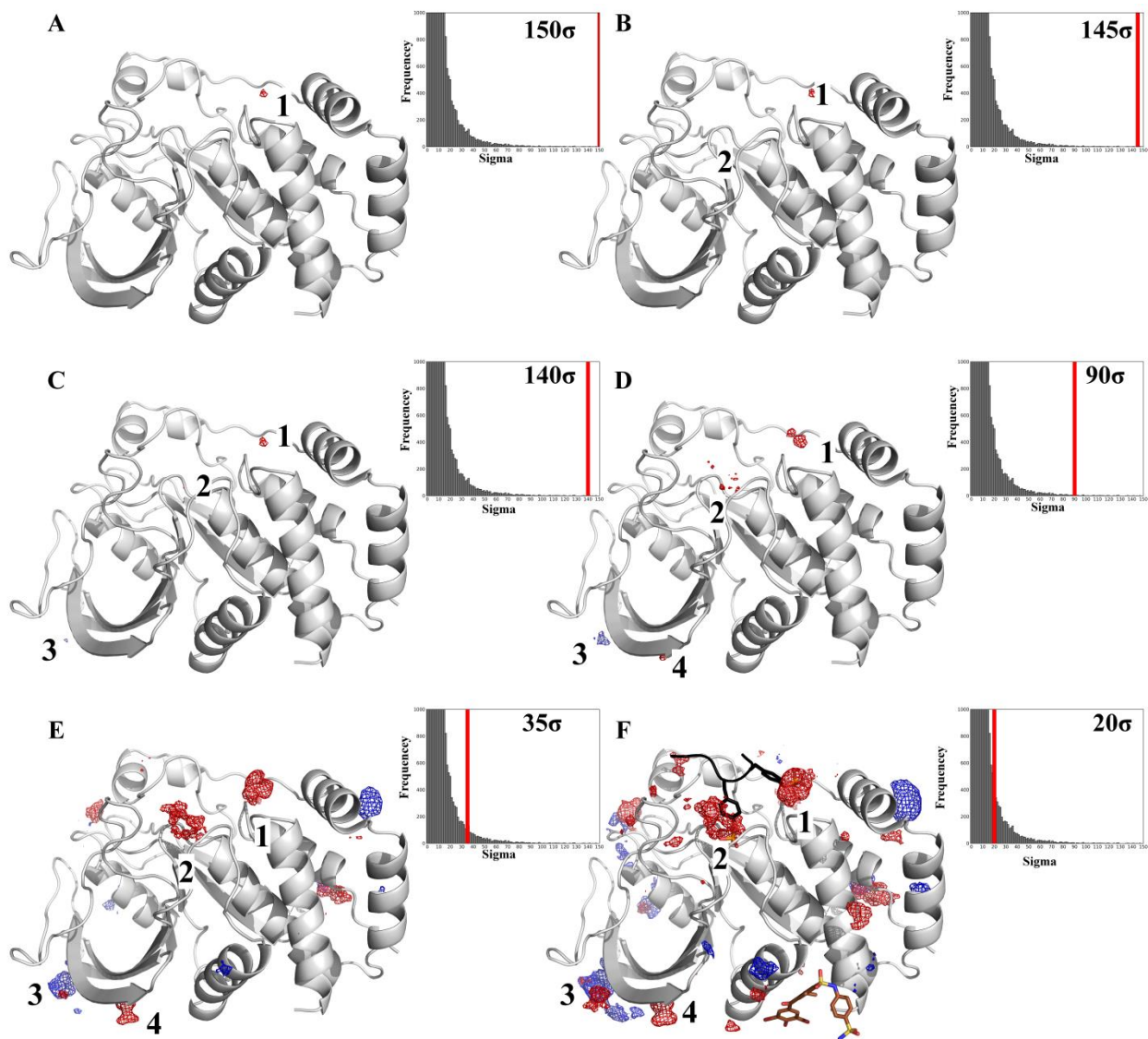


Figure E-10 The ranking of the top-four sites is shown for MixMD simulations starting from Protein Tyrosine Phosphatase 1B (PDB ID: 2AM9) using acetate (red) and methyl ammonium (blue) probes. This is achieved by contouring MixMD maps at decreasing sigma values starting from 150σ to 20σ (Shown in figures A – F).

Appendix F. MixMD maps for Heat Shock Protein 27

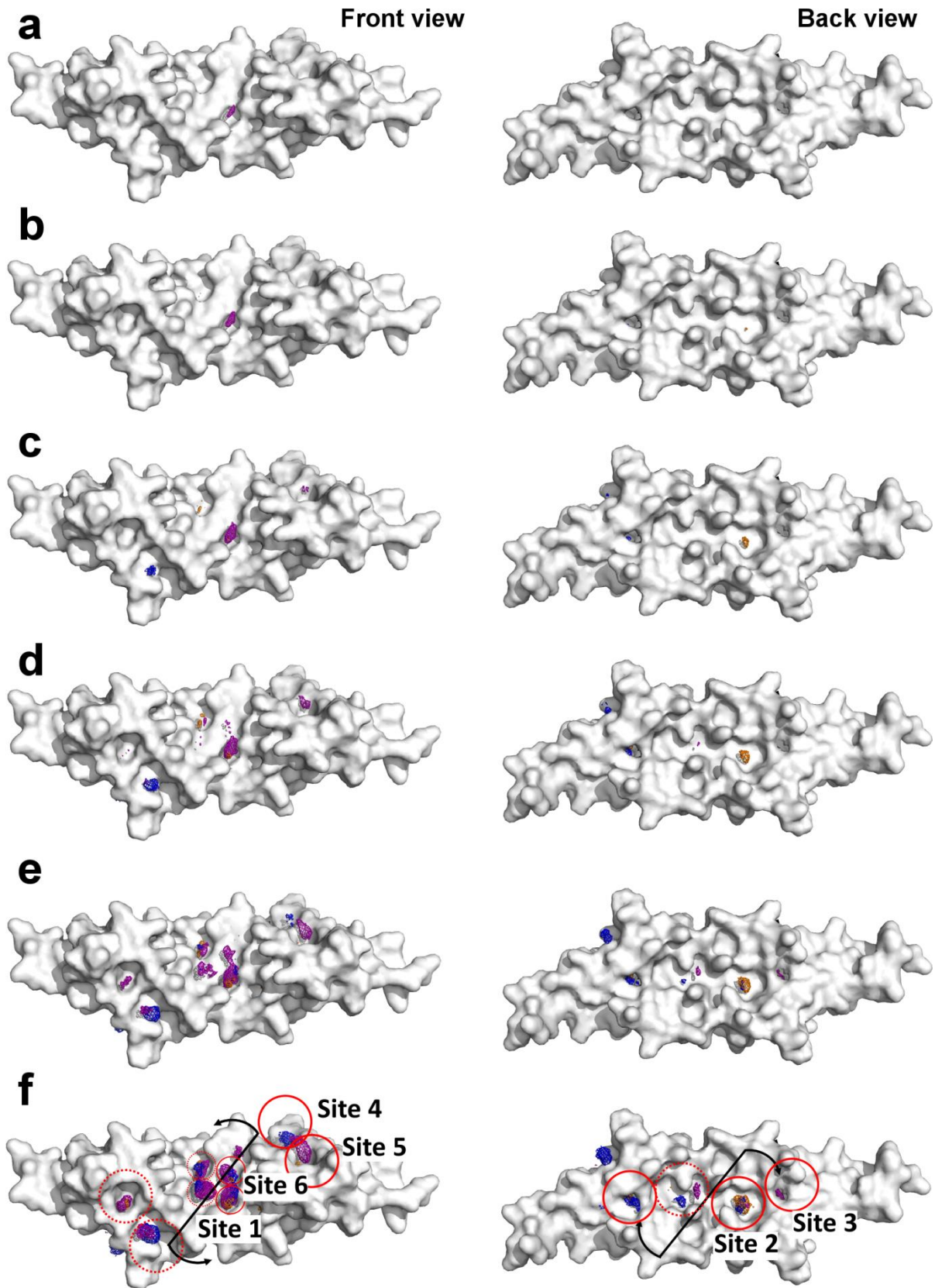


Figure F-1 A detailed description of the MixMD protocol used to predict and rank binding sites is shown. The MixMD maps are decreased gradually from high sigma values to low sigma values. This process is depicted using the following sigma cutoff values a) 55σ b) 45σ c) 35σ d) 30σ e) 25σ and f) 20σ .

Appendix G. NMR data for Heat Shock Protein 27

Table G-1 The Hsp27 NMR chemical shift perturbation (CSP) normalized for each cosolvent are shown for acetonitrile, isopropanol, and pyrimidine. Residues that shift significantly are defined as those CSPs above the normalized average of 0.02 and are highlighted in pink.

Residue	Acetonitrile	Isopropanol	Pyrimidine
7	0.043	0.047	0.02
11	0.022	0.025	0.009
16	0.048	0.053	0.019
17	0.049	0.053	0.026
19	0.006	0.004	0.006
21	0.013	0.019	0.028
23	0.021	0.015	0.001
24	0.01	0.003	0.005
25	0.003	0.013	0.009
26	0.024	0.003	0.002
27	0.038	0.045	
29	0.003	0.032	0.035
30	0.009	0.001	0.012
31	0.004	0.005	0.011
33	0.01	0.02	0.05
34	0.026	0.028	0.027
35	0.025	0.005	0.04
36	0.002	0.003	0.017
37	0.036	0.005	0.022
38	0.006	0.011	0.005
39	0.003	0.012	0.004
41	0.012	0.002	0.013
42	0.03	0.007	0.011
43	0.003	0.016	0.019
45	0.016	0	0.003
46	0.009	0.006	0.003
48	0.029	0.023	0.007
49	0.003	0.067	0.051
50	0.007	0.024	0.006
51	0.009	0.007	0.002

52	0.022	0.064	0.114
53	0.032	0.014	0.002
54	0.024	0.027	0.02
55	0.006	0.008	0.009
56	0.016	0.01	0.035
58	0.03	0.013	0.003
59	0.008	0.004	0.002
60	0.021	0.005	0.008
61	0.003	0.008	0.025
62	0.005	0.019	0.024
63	0.026		0.001
64	0.032	0.016	0.021
65	0.002	0.01	0.021
66	0.022	0.028	0.027
69	0.005	0.003	0.004
70	0.002	0.01	0.01
71	0.019	0.022	0.017
73	0.011	0	0.001
75	0.008	0	0.009
76	0.027	0.022	0.008
77	0.02	0.02	0.045
78	0.017	0.017	0.009
79	0.01	0.018	0.039
80	0.014	0.003	0.01
82	0.004		
83	0.002	0.008	0.007
84	0.015	0.011	0.004
85	0.007	0.003	0.004
86	0.009	0.009	0.006
87	0.003	0.021	0.017
88	0.037	0.006	0.007
89	0.008	0.016	0.005
91	0.008	0.013	0.007
94	0.004	0.007	0.01
98	0.002	0.009	0.011

References

1. Fitzpatrick PA, Ringe D, Klibanov AM (1994) X-Ray Crystal Structure of Cross-Linked Subtilisin Carlsberg in Water vs Acetonitrile. *Biochem Biophys Res Commun* 198(2):675–681.
2. Yennawar NH, Yennawar HP, Farber GK (1994) X-ray Crystal Structure of γ -Chymotrypsin in Hexane. *Biochemistry (Mosc)* 33(23):7326–7336.
3. Allen KN, et al. (1996) An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *J Phys Chem* 100(7):2605–2611.
4. Mattos C, et al. (2006) Multiple Solvent Crystal Structures: Probing Binding Sites, Plasticity and Hydration. *J Mol Biol* 357(5):1471–1482.
5. English AC, Done SH, Caves LSD, Groom CR, Hubbard RE (1999) Locating interaction sites on proteins: The crystal structure of thermolysin soaked in 2% to 100% isopropanol. *Proteins Struct Funct Bioinforma* 37(4):628–640.
6. English AC, Groom CR, Hubbard RE (2001) Experimental and computational mapping of the binding surface of a crystalline protein. *Protein Eng* 14(1):47–59.
7. Dechene M, Wink G, Smith M, Swartz P, Mattos C (2009) Multiple solvent crystal structures of ribonuclease A: An assessment of the method. *Proteins Struct Funct Bioinforma* 76(4):861–881.
8. Miranker A, Karplus M (1991) Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins Struct Funct Bioinforma* 11(1):29–34.
9. Caflich A, Miranker A, Karplus M (1993) Multiple copy simultaneous search and construction of ligands in binding sites: application to inhibitors of HIV-1 aspartic proteinase. *J Med Chem* 36(15):2142–2167.
10. Elber R, Karplus M (1990) Enhanced sampling in molecular dynamics: use of the time-dependent Hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *J Am Chem Soc* 112(25):9161–9175.
11. Miller M, et al. (1989) Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 Å resolution. *Science* 246(4934):1149–1152.
12. Carlson HA, Masukawa KM, McCammon JA (1999) Method for Including the Dynamic Fluctuations of a Protein in Computer-Aided Drug Design. *J Phys Chem A* 103(49):10213–10219.
13. Heather A. Carlson, et al. (2000) Developing a Dynamic Pharmacophore Model for HIV-1 Integrase. *J Med Chem* 43(11):2100–2114.

14. Meagher KL, Carlson HA (2004) Incorporating Protein Flexibility in Structure-Based Drug Discovery: Using HIV-1 Protease as a Test Case. *J Am Chem Soc* 126(41):13276–13281.
15. Meagher KL, Lerner MG, Carlson HA (2006) Refining the Multiple Protein Structure Pharmacophore Method: Consistency across Three Independent HIV-1 Protease Models. *J Med Chem* 49(12):3478–3484.
16. Bowman AL, Lerner MG, Carlson HA (2007) Protein Flexibility and Species Specificity in Structure-Based Drug Discovery: Dihydrofolate Reductase as a Test System. *J Am Chem Soc* 129(12):3634–3640.
17. Damm KL, Carlson HA (2007) Exploring Experimental Sources of Multiple Protein Conformations in Structure-Based Drug Design. *J Am Chem Soc* 129(26):8225–8235.
18. Bowman AL, Nikolovska-Coleska Z, Zhong H, Wang S, Carlson HA (2007) Small Molecule Inhibitors of the MDM2-p53 Interaction Discovered by Ensemble-Based Receptor Models. *J Am Chem Soc* 129(42):12809–12814.
19. Lerner MG, Bowman AL, Carlson HA (2007) Incorporating Dynamics in E. coli Dihydrofolate Reductase Enhances Structure-Based Drug Discovery. *J Chem Inf Model* 47(6):2358–2365.
20. Lerner MG, Meagher KL, Carlson HA (2008) Automated clustering of probe molecules from solvent mapping of protein surfaces: new algorithms applied to hot-spot mapping and structure-based drug design. *J Comput Aided Mol Des* 22(10):727–736.
21. Damm KL, Ung PMU, Quintero JJ, Gestwicki JE, Carlson HA (2008) A poke in the eye: Inhibiting HIV-1 protease through its flap-recognition pocket. *Biopolymers* 89(8):643–652.
22. Ung PM-U, Dunbar JB, Gestwicki JE, Carlson HA (2014) An Allosteric Modulator of HIV-1 Protease Shows Equipotent Inhibition of Wild-Type and Drug-Resistant Proteases. *J Med Chem* 57(15):6468–6478.
23. Jorgensen, W. L. (2000) *BOSS* (Yale University: New Haven, CT).
24. Landon MR, et al. (2009) Detection of ligand binding hot spots on protein surfaces via fragment-based methods: application to DJ-1 and glucocerebrosidase. *J Comput Aided Mol Des* 23(8):491–500.
25. Buhrman G, et al. (2011) Analysis of Binding Site Hot Spots on the Surface of Ras GTPase. *J Mol Biol* 413(4):773–789.
26. Hall DH, et al. (2011) Robust Identification of Binding Hot Spots Using Continuum Electrostatics: Application to Hen Egg-White Lysozyme. *J Am Chem Soc* 133(51):20668–20671.
27. Seco J, Luque FJ, Barril X (2009) Binding Site Detection and Druggability Index from First Principles. *J Med Chem* 52(8):2363–2371.
28. Barril X (2013) Druggability predictions: methods, limitations, and applications. *Wiley Interdiscip Rev Comput Mol Sci* 3(4):327–338.

29. Alvarez-Garcia D, Barril X (2014) Relationship between Protein Flexibility and Binding: Lessons for Structure-Based Drug Design. *J Chem Theory Comput* 10(6):2608–2614.
30. Alvarez-Garcia D, Barril X (2014) Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. *J Med Chem* 57(20):8530–8539.
31. Guvench O, MacKerell AD Jr (2009) Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput Biol* 5(7):e1000435.
32. Raman EP, Yu W, Guvench O, MacKerell AD (2011) Reproducing Crystal Binding Modes of Ligand Functional Groups Using Site-Identification by Ligand Competitive Saturation (SILCS) Simulations. *J Chem Inf Model* 51(4):877–896.
33. Foster TJ, MacKerell AD, Guvench O (2012) Balancing target flexibility and target denaturation in computational fragment-based inhibitor discovery. *J Comput Chem* 33(23):1880–1891.
34. Raman EP, Vanommeslaeghe K, MacKerell AD (2012) Site-Specific Fragment Identification Guided by Single-Step Free Energy Perturbation Calculations. *J Chem Theory Comput* 8(10):3513–3525.
35. Raman EP, Yu W, Lakkaraju SK, MacKerell AD (2013) Inclusion of Multiple Fragment Types in the Site Identification by Ligand Competitive Saturation (SILCS) Approach. *J Chem Inf Model* 53(12):3384–3398.
36. Yu W, Lakkaraju SK, Raman EP, Jr ADM (2014) Site-Identification by Ligand Competitive Saturation (SILCS) assisted pharmacophore modeling. *J Comput Aided Mol Des* 28(5):491–507.
37. Lakkaraju SK, Raman EP, Yu W, MacKerell AD (2014) Sampling of Organic Solutes in Aqueous and Heterogeneous Environments Using Oscillating Excess Chemical Potentials in Grand Canonical-like Monte Carlo-Molecular Dynamics Simulations. *J Chem Theory Comput* 10(6):2281–2290.
38. Yu W, Lakkaraju SK, Raman EP, Fang L, MacKerell AD (2015) Pharmacophore Modeling Using Site-Identification by Ligand Competitive Saturation (SILCS) with Multiple Probe Molecules. *J Chem Inf Model* 55(2):407–420.
39. Raman EP, MacKerell AD (2015) Spatial Analysis and Quantification of the Thermodynamic Driving Forces in Protein–Ligand Binding: Binding Site Variability. *J Am Chem Soc* 137(7):2608–2621.
40. Lakkaraju SK, et al. (2015) Mapping Functional Group Free Energy Patterns at Protein Occluded Sites: Nuclear Receptors and G-Protein Coupled Receptors. *J Chem Inf Model* 55(3):700–708.
41. Basse N, et al. (2010) Toward the Rational Design of p53-Stabilizing Drugs: Probing the Surface of the Oncogenic Y220C Mutant. *Chem Biol* 17(1):46–56.
42. Yang C-Y, Wang S (2010) Computational Analysis of Protein Hotspots. *ACS Med Chem Lett* 1(3):125–129.
43. Yang C-Y, Wang S (2011) Hydrophobic Binding Hot Spots of Bcl-xL Protein–Protein Interfaces by Cosolvent Molecular Dynamics Simulation. *ACS Med Chem Lett* 2(4):280–284.

44. Yang C-Y, Wang S (2012) Analysis of Flexibility and Hotspots in Bcl-xL and Mcl-1 Proteins for the Design of Selective Small-Molecule Inhibitors. *ACS Med Chem Lett* 3(4):308–312.
45. Yang C-Y (2015) Identification of Potential Small Molecule Allosteric Modulator Sites on IL-1R1 Ectodomain Using Accelerated Conformational Sampling Method. *PLoS ONE* 10(2):e0118671.
46. Lexa KW, Carlson HA (2011) Full Protein Flexibility Is Essential for Proper Hot-Spot Mapping. *J Am Chem Soc* 133(2):200–202.
47. Lexa KW, Carlson HA (2013) Improving Protocols for Protein Mapping through Proper Comparison to Crystallography Data. *J Chem Inf Model* 53(2):391–402.
48. Lexa KW, Goh GB, Carlson HA (2014) Parameter Choice Matters: Validating Probe Parameters for Use in Mixed-Solvent Simulations. *J Chem Inf Model* 54(8):2190–2199.
49. Huang D, Caflisch A (2011) Small Molecule Binding to Proteins: Affinity and Binding/Unbinding Dynamics from Atomistic Simulations. *ChemMedChem* 6(9):1578–1580.
50. Huang D, Rossini E, Steiner S, Caflisch A (2014) Structured Water Molecules in the Binding Site of Bromodomains Can Be Displaced by Cosolvent. *ChemMedChem* 9(3):573–579.
51. Bakan A, Nevins N, Lakdawala AS, Bahar I (2012) Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *J Chem Theory Comput* 8(7):2435–2447.
52. Tan YS, et al. (2012) Using Ligand-Mapping Simulations to Design a Ligand Selectively Targeting a Cryptic Surface Pocket of Polo-Like Kinase 1. *Angew Chem* 124(40):10225–10228.
53. Tan YS, Spring DR, Abell C, Verma C (2014) The Use of Chlorobenzene as a Probe Molecule in Molecular Dynamics Simulations. *J Chem Inf Model* 54(7):1821–1827.
54. Prakash P, Hancock JF, Gorfe AA (2015) Binding hotspots on K-ras: Consensus ligand binding sites and other reactive regions from probe-based molecular dynamics analysis. *Proteins Struct Funct Bioinforma*:n/a–n/a.
55. Ho WC, et al. (2006) High-resolution structure of the p53 core domain: implications for binding small-molecule stabilizing compounds. *Acta Crystallogr D Biol Crystallogr* 62(12):1484–1493.
56. Kuntz ID, Chen K, Sharp KA, Kollman PA (1999) The maximal affinity of ligands. *Proc Natl Acad Sci* 96(18):9997–10002.
57. Smith RD, Engdahl AL, Dunbar JB, Carlson HA (2012) Biophysical Limits of Protein–Ligand Binding. *J Chem Inf Model* 52(8):2098–2106.
58. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking Sets for Molecular Docking. *J Med Chem* 49(23):6789–6801.
59. Molecular Operating Environment (2010) (Chemical Computing Group Inc., Montreal, Canada).

60. Ewing TJ, Makino S, Skillman AG, Kuntz ID (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* 15(5):411–428.
61. Morris GM, et al. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* 30(16):2785–2791.
62. Hu B, Lill MA (2012) Protein Pharmacophore Selection Using Hydration-Site Analysis. *J Chem Inf Model* 52(4):1046–1060.
63. Trott O, Olson AJ (2010) AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31(2):455–461.
64. Gilson MK, Given JA, Bush BL, McCammon JA (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys J* 72(3):1047–1069.
65. Hamelberg D, McCammon JA (2004) Standard Free Energy of Releasing a Localized Water Molecule from the Binding Pockets of Proteins: Double-Decoupling Method. *J Am Chem Soc* 126(24):7683–7689.
66. Halgren T (2007) New Method for Fast and Accurate Binding-site Identification and Analysis. *Chem Biol Drug Des* 69(2):146–148.
67. Halgren TA (2009) Identifying and Characterizing Binding Sites and Assessing Druggability. *J Chem Inf Model* 49(2):377–389.
68. Filippakopoulos P, et al. (2012) Histone Recognition and Large-Scale Structural Analysis of the Human Bromodomain Family. *Cell* 149(1):214–231.
69. Schmidtke P, Bidon-Chanal A, Luque FJ, Barril X (2011) MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics* 27(23):3276–3285.
70. Carlson HA (2002) Protein flexibility and drug design: how to hit a moving target. *Curr Opin Chem Biol* 6(4):447–452.
71. Lexa KW, Carlson HA (2012) Protein flexibility in docking and surface mapping. *Q Rev Biophys* 45(03):301–343.
72. Bernstein FC, et al. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112(3):535–542.
73. Chen VB, et al. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66(Pt 1):12–21.
74. Halgren TA (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J Comput Chem* 17(5-6):490–519.
75. The PyMOL Molecular Graphics System (Schrödinger, LLC).

76. Damm KL, Carlson HA (2006) Gaussian-Weighted RMSD Superposition of Proteins: A Structural Comparison for Flexible Proteins and Predicted Protein Structures. *Biophys J* 90(12):4558–4573.
77. Warr WA (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J Comput Aided Mol Des* 23(4):195–198.
78. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL Keys for Use in Drug Discovery. *J Chem Inf Model* 42(6):1273–1280.
79. Comprehensive Medicinal Chemistry Database (2003) (MDL Information Systems, San Leandro, CA,).
80. OEGraphSim (2010) (OpenEye Scientific Software, Inc., Santa Fe, NM).
81. OMEGA (2010) (OpenEye Scientific Software, Inc., Santa Fe, NM).
82. Xu W, Harrison SC, Eck MJ (1997) Three-dimensional structure of the tyrosine kinase c-Src. *Nature* 385(6617):595–602.
83. Zhou S, et al. (1993) SH2 domains recognize specific phosphopeptide sequences. *Cell* 72(5):767–778.
84. Eck MJ, Shoelson SE, Harrison SC (1993) Recognition of a high-affinity phosphotyrosyl peptide by the Src homology-2 domain of p56lck. *Nature* 362(6415):87–91.
85. Rahuel J, et al. (1996) Structural basis for specificity of GRB2-SH2 revealed by a novel ligand binding mode. *Nat Struct Mol Biol* 3(7):586–589.
86. Brown EJ, et al. (1994) A mammalian protein targeted by G1-arresting rapamycin-receptor complex. *Nature* 369(6483):756–758.
87. Liu J, et al. (1991) Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes. *Cell* 66(4):807–815.
88. Hamilton GS, Steiner JP (1998) Immunophilins: Beyond Immunosuppression. *J Med Chem* 41(26):5119–5143.
89. Schreiber SL (1991) Chemistry and biology of the immunophilins and their immunosuppressive ligands. *Science* 251(4991):283–287.
90. Tontonoz P, Spiegelman BM (2008) Fat and Beyond: The Diverse Biology of PPAR γ . *Annu Rev Biochem* 77(1):289–312.
91. Nolte RT, et al. (1998) Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor-[gamma]. *Nature* 395(6698):137–143.
92. Bradshaw JM, Grucza RA, Ladbury JE, Waksman G (1998) Probing the “Two-Pronged Plug Two-Holed Socket” Model for the Mechanism of Binding of the Src SH2 Domain to Phosphotyrosyl Peptides: A Thermodynamic Study†. *Biochemistry (Mosc)* 37(25):9083–9090.

93. Tsai C-J, del Sol A, Nussinov R (2008) Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play. *J Mol Biol* 378(1):1–11.
94. Nussinov R, Tsai C-J (2013) Allostery in Disease and in Drug Discovery. *Cell* 153(2):293–305.
95. Ma L, et al. (2009) Selective activation of the M1 muscarinic acetylcholine receptor achieved by allosteric potentiation. *Proc Natl Acad Sci* 106(37):15950–15955.
96. Hardy JA, Wells JA (2004) Searching for new allosteric sites in enzymes. *Curr Opin Struct Biol* 14(6):706–715.
97. Hardy JA, Lam J, Nguyen JT, O'Brien T, Wells JA (2004) Discovery of an allosteric site in the caspases. *Proc Natl Acad Sci U S A* 101(34):12461–12466.
98. Reynolds KA, McLaughlin RN, Ranganathan R (2011) Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell* 147(7):1564–1575.
99. McClendon CL, Friedland G, Mobley DL, Amirkhani H, Jacobson MP (2009) Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles. *J Chem Theory Comput* 5(9):2486–2502.
100. Bakan A, Nevins N, Lakdawala AS, Bahar I (2012) Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *J Chem Theory Comput* 8(7):2435–2447.
101. Allen KN, et al. (1996) An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins†. *J Phys Chem* 100(7):2605–2611.
102. Mattos C, Ringe D (1996) Locating and characterizing binding sites on proteins. *Nat Biotechnol* 14(5):595–599.
103. Lexa KW, Carlson HA (2011) Full Protein Flexibility Is Essential for Proper Hot-Spot Mapping. *J Am Chem Soc* 133(2):200–202.
104. Lexa KW, Carlson HA (2013) Improving Protocols for Protein Mapping through Proper Comparison to Crystallography Data. *J Chem Inf Model* 53(2):391–402.
105. Seco J, Luque FJ, Barril X (2009) Binding Site Detection and Druggability Index from First Principles. *J Med Chem* 52(8):2363–2371.
106. Guvench O, MacKerell AD (2009) Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput Biol* 5(7):e1000435.
107. Raman EP, Yu W, Guvench O, MacKerell AD (2011) Reproducing Crystal Binding Modes of Ligand Functional Groups Using Site-Identification by Ligand Competitive Saturation (SILCS) Simulations. *J Chem Inf Model* 51(4):877–896.
108. Kollman (2010) AMBER 11.

109. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926–935.
110. Lexa KW, Goh GB, Carlson HA (2014) Parameter Choice Matters: Validating Probe Parameters for Use in Mixed-Solvent Simulations. *J Chem Inf Model* 54(8):2190–2199.
111. Hornak V, et al. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct Funct Bioinforma* 65(3):712–725.
112. Ryckaert J-P, Ciccotti G, Berendsen HJ. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* 23(3):327–341.
113. Götz AW, et al. (2012) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J Chem Theory Comput* 8(5):1542–1555.
114. Iwata S, Kamata K, Yoshida S, Minowa T, Ohta T (1994) T and R states in the crystals of bacterial L-lactate dehydrogenase reveal the mechanism for allosteric control. *Nat Struct Mol Biol* 1(3):176–185.
115. Ohta T, Yokota K, Minowa T, Iwata S (1992) Mechanism of allosteric transition of bacterial L-lactate dehydrogenase. *Faraday Discuss* 93(0):153–162.
116. Brenke R, et al. (2009) Fragment-based identification of druggable “hot spots” of proteins using Fourier domain correlation techniques. *Bioinformatics* 25(5):621–627.
117. Barr AJ (2010) Protein tyrosine phosphatases as drug targets: strategies and challenges of inhibitor development. *Future Med Chem* 2(10):1563–1576.
118. De Munter S, Köhn M, Bollen M (2013) Challenges and Opportunities in the Development of Protein Phosphatase-Directed Therapeutics. *ACS Chem Biol* 8(1):36–45.
119. Blaskovich MAT (2009) Drug Discovery and Protein Tyrosine Phosphatases. *Curr Med Chem* 16(17):2095–2176.
120. Cheng AC, et al. (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 25(1):71–75.
121. Hansen SK, et al. (2005) Allosteric Inhibition of PTP1B Activity by Selective Modification of a Non-Active Site Cysteine Residue[†]. *Biochemistry (Mosc)* 44(21):7704–7712.
122. Tonks NK (2003) PTP1B: From the sidelines to the front lines! *FEBS Lett* 546(1):140–148.
123. McQuarrie DA (2000) *Statistical mechanics* (University Science Books, Sausalito, Calif).
124. Case DA, et al. (2012) *AMBER 12* (University of California, San Francisco) Available at: <http://ambermd.org/>.

125. Roe DR, Cheatham TE (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* 9(7):3084–3095.
126. Suárez D, Díaz N (2015) Direct methods for computing single-molecule entropies from molecular simulations. *Wiley Interdiscip Rev Comput Mol Sci* 5(1):1–26.
127. Klebe G (2015) Applying thermodynamic profiling in lead finding and optimization. *Nat Rev Drug Discov* 14(2):95–110.
128. Hopkins AL, Keserü GM, Leeson PD, Rees DC, Reynolds CH (2014) The role of ligand efficiency metrics in drug discovery. *Nat Rev Drug Discov* 13(2):105–121.
129. Hajduk PJ, Huth JR, Fesik SW (2005) Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. *J Med Chem* 48(7):2518–2525.
130. Cheng AC, et al. (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 25(1):71–75.
131. Schmidtke P, Barril X (2010) Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *J Med Chem* 53(15):5858–5867.
132. Halgren TA (2009) Identifying and Characterizing Binding Sites and Assessing Druggability. *J Chem Inf Model* 49(2):377–389.
133. Krasowski A, Muthas D, Sarkar A, Schmitt S, Brenk R (2011) DrugPred: A Structure-Based Approach To Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *J Chem Inf Model* 51(11):2829–2842.
134. Allen KN, et al. (1996) An experimental approach to mapping the binding surfaces of crystalline proteins. *J Phys Chem* 100(7):2605–2611.
135. Organic solvents identify specific ligand binding sites on protein surfaces : Abstract : Nature Biotechnology (1997) *Nat Biotech* 15(3):264–268.
136. Lexa KW, Carlson HA (2013) Improving protocols for protein mapping through proper comparison to crystallography data. *J Chem Inf Model* 53(2):391–402.
137. Seco J, Luque FJ, Barril X (2009) Binding Site Detection and Druggability Index from First Principles. *J Med Chem* 52(8):2363–2371.
138. Guvench O, Mackerell AD Jr (2009) Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput Biol* 5(7):e1000435.
139. Bakan A, Nevins N, Lakdawala AS, Bahar I (2012) Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *J Chem Theory Comput* 8(7):2435–2447.
140. Tan YS, et al. (2012) Using Ligand-Mapping Simulations to Design a Ligand Selectively Targeting a Cryptic Surface Pocket of Polo-Like Kinase 1. *Angew Chem Int Ed* 51(40):10078–10081.

141. Arrigo A-P, et al. (2007) Hsp27 (HspB1) and α B-crystallin (HspB5) as therapeutic targets. *FEBS Lett* 581(19):3665–3674.
142. Efthymiou CA, et al. (2004) Heat shock protein 27 protects the heart against myocardial infarction. *Basic Res Cardiol* 99(6):392–394.
143. Latchman DS (2005) HSP27 and cell survival in neurones. *Int J Hyperthermia* 21(5):393–402.
144. Nagaraja GM, Kaur P, Asea A (2012) Role of human and mouse HspB1 in metastasis. *Curr Mol Med* 12(9):1142–1150.
145. Charette SJ, Lavoie JN, Lambert H, Landry J (2000) Inhibition of Daxx-Mediated Apoptosis by Heat Shock Protein 27. *Mol Cell Biol* 20(20):7602–7612.
146. Datskevich PN, Nefedova VV, Sudnitsyna MV, Gusev NB (2012) Mutations of small heat shock proteins and human congenital diseases. *Biochem Mosc* 77(13):1500–1514.
147. Hotte SJ, et al. (2010) Phase I trial of OGX-427, a 2'-methoxyethyl antisense oligonucleotide (ASO), against heat shock protein 27 (Hsp27): Final results. *J Clin Oncol* 28:15s(suppl; abstr 3077). Available at: <http://meetinglibrary.asco.org/content/44263-74> [Accessed February 10, 2015].
148. Baylot V, et al. (2011) OGX-427 inhibits tumor progression and enhances gemcitabine chemotherapy in pancreatic cancer. *Cell Death Dis* 2(10):e221.
149. Andrea TA, Swope WC, Andersen HC (1983) The role of long ranged forces in determining the structure and properties of liquid water. *J Chem Phys* 79(9):4576–4584.
150. Goddard TD, Kneller DG *SPARKY 3* (University of California, San Francisco).
151. Lewis IA, Schommer SC, Markley JL (2009) rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn Reson Chem* 47(S1):S123–S126.
152. Clark AR, Naylor CE, Bagn ris C, Keep NH, Slingsby C (2011) Crystal Structure of R120G Disease Mutant of Human α B-Crystallin Domain Dimer Shows Closure of a Groove. *J Mol Biol* 408(1):118–134.
153. Hochberg GKA, et al. (2014) The structured core domain of B-crystallin can prevent amyloid fibrillation and associated toxicity. *Proc Natl Acad Sci* 111(16):E1562–E1570.
154. Weeks SD, et al. (2014) Molecular structure and dynamics of the dimeric human small heat shock protein HSPB6. *J Struct Biol* 185(3):342–354.
155. Laganowsky A, et al. (2010) Crystal structures of truncated alphaA and alphaB crystallins reveal structural mechanisms of polydispersity important for eye lens function. *Protein Sci* 19(5):1031–1043.
156. Delbecq SP, Jehle S, Klevit R (2012) Binding determinants of the small heat shock protein, α B-crystallin: recognition of the “IxI” motif. *EMBO J* 31(24):4587–4594.

157. Hilton GR, et al. (2013) C-terminal interactions mediate the quaternary dynamics of α B-crystallin. *Philos Trans R Soc B Biol Sci* 368(1617):20110405.
158. Makley LN (2014) Chemical Approaches for “Undruggable” Targets: The Discovery of Ligands for Small Heat Shock Proteins. Ph.D. (University of Michigan, United States -- Michigan). Available at: <http://search.proquest.com.proxy.lib.umich.edu/dissertations/docview/1615822829/abstract/939CB8A027E9432EPQ/1?accountid=14667> [Accessed May 8, 2015].
159. Journal TP, 2009 17 Apr From snake venom to ACE inhibitor — the discovery and rise of captopril. *Pharm J*. Available at: <http://www.pharmaceutical-journal.com/news-and-analysis/news/from-snake-venom-to-ace-inhibitor-the-discovery-and-rise-of-captopril/10884359.article> [Accessed April 27, 2015].
160. Cushman DW, Cheung HS, Sabo EF, Ondetti MA (1977) Design of potent competitive inhibitors of angiotensin-converting enzyme. Carboxyalkanoyl and mercaptoalkanoyl amino acids. *Biochemistry (Mosc)* 16(25):5484–5491.
161. Mattos C, Ringe D (1996) Locating and characterizing binding sites on proteins. *Nat Biotechnol* 14(5):595–599.
162. Wenbo Yu, Olgun Guvench, Alexander D MacKerell (2013) Computational approaches for the design of protein-protein interaction inhibitors. *Understanding and Exploiting Protein-Protein Interactions as Drug Targets*, Future Science Book Series. (Future Science Ltd), pp 90–102. Available at: <http://www.futuremedicine.com/doi/abs/10.4155/ebo.13.141> [Accessed April 8, 2015].
163. Klon AE ed. (2015) Site Identification by Ligand Competitive Saturation (SILCS) Simulations for Fragment-Based Drug Design - Springer. *Methods in Molecular Biology*. (Springer New York). Available at: http://link.springer.com.proxy.lib.umich.edu/protocol/10.1007/978-1-4939-2486-8_7 [Accessed April 8, 2015].
164. Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J Chem Phys* 120(24):11919–11929.
165. Laio A, Parrinello M (2002) Escaping free-energy minima. *Proc Natl Acad Sci* 99(20):12562–12566.
166. Wells JA, McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* 450(7172):1001–1009.
167. Lange G, et al. (2003) Requirements for Specific Binding of Low Affinity Inhibitor Fragments to the SH2 Domain of pp60Src Are Identical to Those for High Affinity Binding of Full Length Inhibitors. *J Med Chem* 46(24):5184–5195.
168. Charifson PS, et al. (1997) Peptide Ligands of pp60c-src SH2 Domains: A Thermodynamic and Structural Study. *Biochemistry (Mosc)* 36(21):6283–6293.
169. Ettmayer P, et al. (1999) Structural and Conformational Requirements for High-Affinity Binding to the SH2 Domain of Grb21. *J Med Chem* 42(6):971–980.

170. DeLorbe JE, Clements JH, Whiddon BB, Martin SF (2010) Thermodynamic and Structural Effects of Macrocyclic Constraints in Protein–Ligand Interactions. *ACS Med Chem Lett* 1(8):448–452.
171. Nioche P, et al. (2002) Crystal structures of the SH2 domain of grb2: highlight on the binding of a new high-affinity inhibitor. *J Mol Biol* 315(5):1167–1177.
172. Rahuel J, et al. (1996) Structural basis for specificity of Grb2-SH2 revealed by a novel ligand binding mode. *Nat Struct Biol* 3(7):586–589.
173. DeLorbe JE, et al. (2009) Thermodynamic and Structural Effects of Conformational Constraints in Protein–Ligand Interactions. Entropic Paradox associated with Ligand Preorganization. *J Am Chem Soc* 131(46):16758–16770.
174. Das S, Raychaudhuri M, Sen U, Mukhopadhyay D (2011) Functional Implications of the Conformational Switch in AICD Peptide upon Binding to Grb2-SH2 Domain. *J Mol Biol* 414(2):217–230.
175. Benfield AP, et al. (2006) Ligand Preorganization May Be Accompanied by Entropic Penalties in Protein–Ligand Interactions. *Angew Chem Int Ed* 45(41):6830–6835.
176. Rahuel J, et al. (1998) Structural basis for the high affinity of amino-aromatic SH2 phosphopeptide ligands. *J Mol Biol* 279(4):1013–1022.
177. Pochetti G, et al. (2007) Insights into the Mechanism of Partial Agonism. *J Biol Chem* 282(23):17314–17324.
178. Li Y, et al. (2005) Structural and biochemical basis for selective repression of the orphan nuclear receptor liver receptor homolog 1 by small heterodimer partner. *Proc Natl Acad Sci U S A* 102(27):9505–9510.
179. Lin C-H, et al. (2009) Design and Structural Analysis of Novel Pharmacophores for Potent and Selective Peroxisome Proliferator-activated Receptor γ Agonists. *J Med Chem* 52(8):2618–2622.
180. Haffner CD, et al. (2004) Structure-Based Design of Potent Retinoid X Receptor α Agonists. *J Med Chem* 47(8):2010–2029.
181. Mahindroo N, et al. (2005) Novel Indole-Based Peroxisome Proliferator-Activated Receptor Agonists: Design, SAR, Structural Biology, and Biological Activities. *J Med Chem* 48(26):8194–8208.
182. Bénardeau A, et al. (2009) Aleglitazar, a new, potent, and balanced dual PPAR α/γ agonist for the treatment of type II diabetes. *Bioorg Med Chem Lett* 19(9):2468–2473.
183. Nolte RT, et al. (1998) Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor- γ . *Nature* 395(6698):137–143.
184. Mahindroo N, et al. (2006) Structural Basis for the Structure–Activity Relationships of Peroxisome Proliferator-Activated Receptor Agonists. *J Med Chem* 49(21):6421–6424.

185. Casimiro-Garcia A, et al. (2008) Effects of modifications of the linker in a series of phenylpropanoic acid derivatives: Synthesis, evaluation as PPAR α/γ dual agonists, and X-ray crystallographic studies. *Bioorg Med Chem* 16(9):4883–4907.
186. Gampe RT Jr, et al. (2000) Asymmetry in the PPAR γ /RXR α crystal structure reveals the molecular basis of heterodimerization among nuclear receptors. *Mol Cell* 5(3):545–555.
187. Shi GQ, et al. (2005) Design and Synthesis of α -Aryloxyphenylacetic Acid Derivatives: A Novel Class of PPAR α/γ Dual Agonists with Potent Antihyperglycemic and Lipid Modulating Activity. *J Med Chem* 48(13):4457–4468.
188. Mahindroo N, et al. (2006) Indol-1-yl Acetic Acids as Peroxisome Proliferator-Activated Receptor Agonists: Design, Synthesis, Structural Biology, and Molecular Docking Studies. *J Med Chem* 49(3):1212–1216.
189. Kuhn B, et al. (2006) Structure-based design of indole propionic acids as novel PPAR α/γ co-agonists. *Bioorg Med Chem Lett* 16(15):4016–4020.
190. Itoh T, et al. (2008) Structural basis for the activation of PPAR[γ] by oxidized fatty acids. *Nat Struct Mol Biol* 15(9):924–931.
191. Oyama T, et al. (2009) Adaptability and selectivity of human peroxisome proliferator-activated receptor (PPAR) pan agonists revealed from crystal structures. *Acta Crystallogr D Biol Crystallogr* 65(Pt 8):786–795.
192. Cronet P, et al. (2001) Structure of the PPAR α and γ ligand binding domain in complex with AZ 242; ligand selectivity and agonist activation in the PPAR family. *Struct Lond Engl* 1993 9(8):699–706.
193. Li Y, et al. (2008) Molecular recognition of nitrated fatty acids by PPAR[γ]. *Nat Struct Mol Biol* 15(8):865–867.
194. Casimiro-Garcia A, et al. (2009) Synthesis and evaluation of novel α -heteroaryl-phenylpropanoic acid derivatives as PPAR α/γ dual agonists. *Bioorg Med Chem* 17(20):7113–7125.
195. Fracchiolla G, et al. (2009) New 2-Aryloxy-3-phenyl-propanoic Acids As Peroxisome Proliferator-Activated Receptors α/γ Dual Agonists with Improved Potency and Reduced Adverse Effects on Skeletal Muscle Function. *J Med Chem* 52(20):6382–6393.
196. Li Y, Kovach A, Suino-Powell K, Martynowski D, Xu HE (2008) Structural and Biochemical Basis for the Binding Selectivity of Peroxisome Proliferator-activated Receptor γ to PGC-1 α . *J Biol Chem* 283(27):19132–19139.
197. Becker JW, et al. (1993) FK-506-binding protein: three-dimensional structure of the complex with the antagonist L-685,818. *J Biol Chem* 268(15):11335–11339.
198. Liang J, Choi J, Clardy J (1999) Refined structure of the FKBP12-rapamycin-FRB ternary complex at 2.2 Å resolution. *Acta Crystallogr D Biol Crystallogr* 55(Pt 4):736–744.

199. Van Duyne GD, Standaert RF, Schreiber SL, Clardy J (1991) Atomic structure of the rapamycin human immunophilin FKBP-12 complex. *J Am Chem Soc* 113(19):7433–7434.
200. Wilson KP, et al. (1995) Comparative X-ray structures of the major binding protein for the immunosuppressant FK506 (tacrolimus) in unliganded form and in complex with FK506 and rapamycin. *Acta Crystallogr D Biol Crystallogr* 51(4):511–521.
201. Itoh S, DeCenzo MT, Livingston DJ, Pearlman DA, Navia MA (1995) Conformation of FK506 in X-ray structures of its complexes with human recombinant FKBP12 mutants. *Bioorg Med Chem Lett* 5(17):1983–1988.
202. Sun F, et al. (2003) Design and Structure-Based Study of New Potential FKBP12 Inhibitors. *Biophys J* 85(5):3194–3201.
203. Holt DA, et al. (1993) Design, synthesis, and kinetic evaluation of high-affinity FKBP ligands and the X-ray crystal structures of their complexes with FKBP12. *J Am Chem Soc* 115(22):9925–9938.
204. Van Duyne GD, Standaert RF, Karplus PA, Schreiber SL, Clardy J (1991) Atomic structure of FKBP-FK506, an immunophilin-immunosuppressant complex. *Science* 252(5007):839–842.
205. Burkhard P, Taylor P, Walkinshaw MD (2000) X-ray structures of small ligand-FKBP complexes provide an estimate for hydrophobic interaction energies. *J Mol Biol* 295(4):953–962.
206. Dubowchik GM, et al. (2001) 2-Aryl-2,2-difluoroacetamide FKBP12 ligands: synthesis and X-ray structural studies. *Org Lett* 3(25):3987–3990.
207. Schultz LW, Clardy J (1998) Chemical inducers of dimerization: the atomic structure of FKBP12-FK1012A-FKBP12. *Bioorg Med Chem Lett* 8(1):1–6.
208. Clackson T, et al. (1998) Redesigning an FKBP-ligand interface to generate chemical dimerizers with novel specificity. *Proc Natl Acad Sci U S A* 95(18):10437–10442.
209. Becker JW, et al. (1999) 32-Indolyl Ether Derivatives of Ascomycin: Three-Dimensional Structures of Complexes with FK506-Binding Protein. *J Med Chem* 42(15):2798–2804.
210. Xu RX, et al. (1995) Solution structure of the human pp60c-src SH2 domain complexed with a phosphorylated tyrosine pentapeptide. *Biochemistry (Mosc)* 34(7):2107–2121.
211. Ogura K, et al. (2008) Solution structure of the Grb2 SH2 domain complexed with a high-affinity inhibitor. *J Biomol NMR* 42(3):197–207.
212. Michnick SW, Rosen MK, Wandless TJ, Karplus M, Schreiber SL (1991) Solution structure of FKBP, a rotamase enzyme and receptor for FK506 and rapamycin. *Science* 252(5007):836–839.
213. Riepl H, et al. (2005) Sequential Backbone Assignment of Peroxisome Proliferator-Activated Receptor- γ Ligand Binding Domain. *J Biomol NMR* 32(3):259–259.