Protein Flexibility In Structure-Based Drug Design

by

Katrina Walden Lexa

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Medicinal Chemistry)
in the University of Michigan
2011

Doctoral Committee:

Professor Heather A. Carlson, Chair
Professor Shaomeng Wang
Assistant Professor Hashim M. Al-Hashimi
Assistant Professor Jason E. Gestwicki
Assistant Professor Oleg V. Tsodikov

To my Momma and Papa

Thank you for always believing

Acknowledgements

I am so very grateful to all of the people who have lent me support and guidance along the way. First, I have to thank George Shields, my undergraduate professor who introduced me to computational chemistry. My mentor Heather Carlson was central to my decision to pursue a PhD in Medicinal Chemistry at Michigan and she has been instrumental in my development as a computational chemist. Her support and patience has been critical to my advancement and I will always be thankful for the opportunity. I am also deeply appreciative of the advice and helpful suggestions from the members of my dissertation committee.

I owe the members of the Carlson Group a huge debt of gratitude for the countless impromptu discussions that have broadened my skillset. I was lucky to overlap with Kelly Damm and Steve Spronk, whose passion for science was infectious. Mark Benson, Michael Lerner, and Richard Smith were always willing to help, especially with programming questions. Jim Dunbar Jr., Nickolay Khazanov, and Peter Ung were never too busy to bounce ideas around with me or read manuscripts. I owe Richard Smith a special debt of gratitude for having the patience to have me as his cubby mate over the years, especially when things in lab and life were chaotic. Further, Nickolay Khazanov shared my love for coffee and food and was always ready to listen.

I would like to thank the faculty and staff in the Medicinal Chemistry department, as our tightly-knit family has provided constant support and assistance. Hank Mosberg and Anna Mapp have always been willing to listen and/or give advice when I needed to talk something through. Both Oleg Tsodikov and our collaborator Jeanne Stuckey were instrumental in the forward progress of my research and I value all of the time they took to educate me further about x-ray crystallography. Maria Herbel has been a fantastic resource and has always been ready to assist me with any student life questions that arose. I would also like to thank Dennis Gilbert, for taking me under his wing and allowing me to continue writing as a graduate student. I am incredibly grateful for the

financial support I have received towards my education from the Pharmocological Sciences Training Program and the American Foundation for Pharmaceutical Education Pre-doctoral Fellowship.

Finally, I would be remiss not to thank all of my friends and family who have made this journey with me. The friendship of Erin Smith, Mollie Wright, Stephanie Godleski, Kate Neff, Fernanda Burke, Caleb Bates, James Patrone, Kelly Damm, Bob Rarig, Amy Payeur, Ahleah Rohr Daniel, Trey Porter, Antek Wong-Foy, Stefanie Stachura, Jon Mortison, John Henssler, Matt Rohr Daniel, Chris Avery, Ron Jenkins, and Kyle Heslip (in no particular order), has made these past few years a joy, and I cannot thank any of them enough for their enduring support and friendship. I am especially grateful to Caleb Bates, who has endured these past few years with patience and much-needed perspective. I would also like to thank my parents and my brother, as they have been with me from the very first and have provided me with encouragement, solace, support, laughter, and the best food a girl could ask for.

Table of Contents

Dedication………………………………………………………………………………. ii

Acknowledgements…………………………………………………………………iii

List of Figures……………………………………….….………………………viii

List of Tables……………………………..…….……………………………xxii

Abstract………………………………...……………….…………….….…. xxiv

Chapter 1: Introduction 1

1.1 Specific Aims 1

1.2 Theory of Molecular Mechanics 2

1.3 Structure-based Drug Design 4

1.4 Protein Flexibility 6

1.5 Protein-Ligand Binding 7

Relaxation Methods 12

Conformational generation on the fly 16

Grid-based ensembles 19

Structure-based ensembles 23

1.6 *De Novo* Approaches through Fragment-Based Drug Design 32

1.7 Limitations of Structure-Based Drug Design 36

1.8 HIV-1 Protease as a Computational Model 37

1.9 Summary 39

Chapter 2: Full Protein Flexibility is Essential for Proper Hot-Spot Mapping................ 41

2.1 Introduction.. 41

2.2 Methods... 42

2.3 Results and Discussion 43

2.4 Conclusion 52

Chapter 3: How to Compare Computational Probe Mapping to Crystal Structures......... 53

3.1 Introduction.. 53

3.2 Methods... 56

3.3 Results and Discussion 58

3.4 Conclusions... 79

Chapter 4: Binding to the Open Conformation of HIV-1 Protease 81

4.1 Introduction.. 81

4.2 Methods... 82

4.3 Results and Discussion 83

4.4 Conclusion 99

Chapter 5: Clarifying Allosteric Control of Flap Conformations in the 1TW7 Crystal Structure of HIV-1 Protease... 101

5.1 Introduction.. 101

5.2 Methods... 102

5.3 Results and Discussion 107

5.4 Conclusion 115

Appendix: Exploring the Potential for Allostery at the Elbow Region of HIV-1 Protease..…………………………………………………………………..115

Introduction……………………………………………………………………..116

Methods.. 120

Discussion 124

Conclusion 134

Bibliography……………………………………………………………………136

List of Figures

Figure 1-1: The conformational variation in holo and apo BACE structures illustrates structural differences frequently seen across multiple crystal structures of the same protein. 8

Figure 1-2: The same protein crystallized with different ligands. This simple two ligand-one protein set demonstrates the influence of structure on the shape of the binding pocket. 9

Figure 1-3: MUSIC simulations of benzene probes (gray) to the surface of a HIVp monomer (purple). Five hundred probes were flooded onto the protein surface, minimized, and then clustered together. Parent probes were identified for each monomer conformation, the HIVp monomers were aligned through a weighted-RMSD function, then the parent probes were clustered together and retained when at least half of the protein conformations contained a probe at that site. This resulted in low-energy probe clusters (colored red through purple) of benzene that could be combined with ethane and methanol clusters to develop a pharmacophore model.. 24

Figure 1-4: The typical semi-open flap conformation for HIVp with each distinct region of the protein shown in a different color. The catalytic aspartic acids (Asp25/25') are shown in stick form. The flaps are shown in dark blue, residues 43-58; the flap tips in yellow, residues 49-52; the fulcrum in orange, residues 11-22; the cantilever in lime green, residues 59-72; the active site in cyan, residues 23-30; the dimer interface in blue, residues 1-5 and 95-99; the elbow in hot pink, residues 35-42; the helix in light blue, residues 86-90; the 80s loop or wall turn in red, residues 79-84; and the nose in violet, residues 6-10. 38

Figure 2-1: Results from unrestrained vs. restrained protein simulations using CCN and water to solvate HEWL (white). The single hotspot identified by MSCS is shown in stick form; CCN (cyan). The probe density from the fully restrained simulation is shown in orange, from backbone-restrained in green, and from fully flexible in blue. Many incorrect local minima are seen in green and orange, but only the correct position dominates the simulation of the fully flexible protein in blue. 44

Figure 2-2: MixMD data for the fully flexible simulation agrees well with densities from MSCS experiments using CCN and water. All snapshots were superimposed on the crystal structure of HEWL (white surface). MixMD density for water is shown as a red mesh, and crystallographic waters are colored black for B-factors below 33 Å and yellow above 33 Å. The MSCS coordinates for CCN are shown in stick form (cyan with its electron density as a mesh, $2F_O$-F_C shown at 1.5σ); the highest occupancy for CCN in the fully flexible simulation matches perfectly and is shown in solid blue surface. Unsatisfied electron density in the crystal structure (positive F_O-F_C shown at 3σ) is shown in solid purple. (A) Water maps within the protein highlight interior waters conserved in the crystal structure and reproduced in our simulation. The large astrices (*) denote two highly occupied regions of the interior water map that correspond to unfulfilled density in the crystal structure. (B) Maps of the protein surface show good correspondence between crystallographic and MixMD densities for both CCN and well resolved waters. For the 16 crystallographic waters with B-factors below 33 Å (black spheres) five occur at symmetry-packing interfaces. MixMD misses four of the five, which is expected because the contacts are not present in our simulation. The other 11 best-resolved waters are well reproduced. 45

Figure 2-3**:** The maximal occupancy positions over the last 2 ns for each independent simulation of HEWL in pre-equilibrated 50% w/w CCN and water. Only the fully flexible system shows convergence of the five simulations in agreement with experiment. Individual runs are indicated by color (red, orange, yellow, green, and blue). 47

Figure 2-4**:** Combined results from the last 2 ns of all 5 simulations (10 ns of sampling) with flexible HEWL (white surface), comparing the layered solvent to the pre-equilibrated protocol. The probe density is shown in blue mesh for 50% w/w pre-equilibrated solvent and in orange mesh for 50% w/w layered water and CCN. The highest sampled densities overlap exactly and agree well with the position of CCN in the 2LYO crystal structure... 48

Figure 2-5: The positions of maximal occupancy on the CCN grid were calculated over the last 2 ns for each independent simulation around a fully flexible HEWL (white). **A**) Results for the individual simulations based on pre-equilibrated solvent are shown. All five positions are in excellent agreement with one another and the CCN molecule in the 2LYO structure. The second (orange) and fourth (green) run results lie beneath the first (red) and fifth (blue) run results. **B**) Results for the individual simulations based on layered solvent are shown, with the orientation skewed slightly from A to show the green site (run 3) that does not agree. Four of the five positions are in agreement with one another and the crystallographic CCN. The first run result (red) lies beneath the fifth run result (blue). 49

Figure 2-6: Results from the three different solvation protocols with fully flexible HEWL and CCN as the organic probe. The probe density calculated by combining the last 2 ns of the five independent simulations (to give 10 ns of sampling) is shown as a purple mesh for 90% w/w CCN, as a blue mesh for 50%, and as a green mesh for 10%. Results from all three simulations are nearly identical and reproduce the crystallographic position of CCN. 50

Figure 2-7: The water density calculated by combining the last 2 ns of the five independent simulations (giving 10 ns of sampling) is shown. The water density is shown as a green mesh for 10% w/w CCN (A), as a blue mesh for 50% (B), and as a purple mesh for 90% (C). The separate solvent densities are overlaid in (D). Crystallographic waters are colored black for B-factors below 33 Å and yellow above 33 Å. Although the 10% and 90% environments reproduce the CCN position, they do not give appropriate mapping for the water. In A, too many equivalent positions for water are seen without identifying the water with higher B-factor (yellow). In C, there are too few water locations.. 51

Figure 3-1: The MSCS of elastase+IPA[228] is shown in violet (right) with a neighboring symmetry partner in cyan (left). The symmetry-related equivalent IPA in both structures are circled in red. One of the bound IPA probes in the symmetry partner is located within 4 Å of the protein in the initial unit cell. When the probe site is support by the refined electron density, it is necessary that it receive special consideration during evaluation of the simulation data. Probes at the contact interfaces may be irreproducible, but when they are observed, the probe may map to either or both of the circled sites on the protein.. 60

Figure 3-2: Refined density maps calculated for several of the MSCS structures. The refined maps were based on the coordinate file for the MSCS structure, with probes and close-contact waters/ions removed. The positive F_O-F_C density is contoured in green; the negative F_O-F_C density is contoured in red at 3.0 σ. This illustrates regions where crystallographic probes are justified or not justified, respectively. In A, the positive F_O-F_C density from the MSCS of elastase+IPA (2FOF) clearly agrees with the probe placement, where IPA.1001 and IPA.1003 are located in the active site with good agreement between coordinates and density (IPA.1002 is located near the crystal interface). The positive F_O-F_C density in B indicates that the site for IPA.905A in RNase A+IPA (3EV2) may best correspond with a water molecule. In C, the F_O-F_C density along the active site of thermolysin+IPA (7TLI) indicates that the site for IPA.2001 is justified while the sites for IPA.2005 and IPA.2008 might best support a water molecule as opposed to a probe. In D, the F_O-F_C density in p53 core (2IOM) illustrates that this probe is justified in the crystal structure. 62

Figure 3-3: MixMD data combined for the last 5 ns of 5 individual 50-ns runs of fully flexible HEWL in 50% w/w IPA and water. All snapshots were superimposed on the starting crystal structure (PDB ID 1YL0, white surface). The MSCS coordinate positions for IPA.2836 and IPA.2837 are shown in cyan as ball-and-stick. A) Simulation IPA density does map the IPA.2836 site using a 2-ns window after 10 ns of simulation, however many other spurious minima are mapped as well. B-F) Using a 5-ns window, we find that the local minima are eliminated over time as the simulation converges to the maximally occupied hot spot. 66

Figure 3-4: MixMD data combined for the last 5 ns of 5 individual 50-ns runs of fully flexible HEWL in 50% w/w IPA and water. All snapshots were superimposed on the starting crystal structure (PDB ID 1YL0, white surface). Low B-factor (<33 Å) water sites are shown in black and the other water sites are shown in yellow. The MSCS coordinate positions for IPA.2836 and IPA.2837 are shown in cyan as ball-and-stick. Simulation water density overlays well with both A) the crystallographic water sites and B) the secondary IPA probe, IPA.2837.. 67

Figure 3-5: MixMD data combined for the last 5 ns for 5 individual 50 ns runs of fully flexible p53 core in 50% w/w IPA and WAT. All snapshots were superimposed on the original crystal structure (PDB ID 2IOM, white surface). MixMD density is shown in blue for the IPA probe. The MSCS coordinate position for IPA.3001 is shown in cyan as ball-and-stick. The maximally occupied site for simulation IPA density clearly agrees with the crystallographic probe position. A) Simulation IPA density does map the IPA.3001 site using a 2-ns window after 10 ns of simulation, however a few other spurious minima are mapped as well. B-E) Using a 5-ns window, we find that the local minima are eliminated over time as the simulation quickly converges to the maximally occupied hot spot. 68

Figure 3-6: MixMD data combined for the last 5 ns of 5 individual 50-ns runs of fully flexible elastase in 50% w/w IPA and WAT. All snapshots were superimposed on the original crystal structure (PDB ID 2FOF, white surface). The MSCS coordinate positions for IPA.1001, IPA.1002, and IPA.1003 are shown in cyan as ball-and-stick. The crystal interface probe site IPA.1002x is shown in gray. Alternate probes from other MSCS by the same authors are shown in blue. Density contour levels were chosen such that the top eight hot spots were represented at a “hide dust” level of two. A) Results from performing FTMAP against PDBID 2FOF, with functional group clusters shown in pink. B) All probe sites from MSCS with elastase. C) Simulation IPA density from the last 5ns of 50 total ns from independent runs 1-5 (purple) and D) the last 5ns of 50 total ns from independent runs 6-10 (green) do not perform as well as the results from all 10 runs. E) The maximally occupied site for simulation IPA density from the last 5 ns of 50 total ns from 10 independent runs (blue) corresponds with all three the crystallographic probe positions as well as two other MSCS sites. F) Simulation IPA density from the last 5 ns of 20 ns total simulation time over 10 individual runs performs better than the IPA density after 50 ns over 5 runs. 72

Figure 3-7: MixMD data combined for the last 5 ns of 5 individual 50-ns runs of fully flexible thermolysin in 50% w/w IPA and WAT. All snapshots were superimposed on the original crystal structure (PDB ID 7TLI, white surface). The MSCS coordinate positions for IPA.2001, IPA.2006, and IPA.2008 are shown in cyan as ball-and-stick while the unjustified probe sites are shown in gray. The crystal interface probe sites are shown in green. Density contour levels were chosen such that the top eight hot spots were represented, and a “hide dust” level of three was used. A) Results from performing FTMAP against PDBID 7TLI, with functional group clusters shown in pink. B) Simulation IPA density from the last 5ns of 50 total ns from independent runs 1-5 (purple) and C) the last 5ns of 50 total ns from independent runs 6-10 (green) do not perform as well as the results from all 10 runs. D) The maximally occupied site for simulation IPA density from the last 5 ns of 50 total ns from 10 independent runs (blue) corresponds with all three the crystallographic probe positions as well as two other MSCS sites. E) Simulation IPA density from the last 5 ns of 20 ns total simulation time over 10 individual runs performs better than the IPA density after 50 ns over 5 runs. 75

Figure 3-8: MixMD data combined for the last 5 ns of 5 individual 50 ns runs of fully flexible RNase A in 50% w/w IPA (blue) and water (water). All snapshots were superimposed on the crystal structure (PDB ID 3EV2, white surface). Crystallographic waters are colored black for B-factors < 33 Å and in yellow for B-factors > 33 Å. Water molecules that were identified along the crystallographic boundary are not shown. The MSCS coordinate positions for IPA.905A, IPA.917A, IPA.902B, and IPA.903B are shown in cyan as ball-and-stick. MSCS with other functional groups obtained by the same authors are shown in blue, FTMap sites are shown in pink. A) The MSCS probe sites from 3EV2 chain B. B) The consensus MSCS probe sites with a variety of different solvent probes. C) The cluster sites from performing FTMAP against chain B of 3EV2. D) The simulation density for the probe occupancy over the final 5 ns of the combined runs. E) The simulation density for the water occupancy of the final 5 ns of the combined runs. 78

Figure 4-1: A) The crystallized HIV-1 protease uniquely bound by two identical inhibitors, with pose α colored in grey and pose β in black. B) The crystal structure 3BC4 with Damm compound 1[103] (black) bound at the eye site. C) The 5-Nitroindole fragment (black) crystallized in the eye site by Perryman et al.[242] D) A

2-dimensional representation of the pyrrolidine inhibitor that was co-crystallized with 3BC4. For the following figures, we have used a convention of orienting the complex so that a naphthyl occupies the eye position on the right (ie. Monomer 2). We are labeling the monomers as “1” and “2” instead of “A” and “B” to avoid confusion with the α and β notation for the ligands... 81

Figure 4-2: The overall RMSD calculated for the core (all residues but the flaps 43-58/43'-58') of HIVp over the length of the production run. The core remains stable for the duration of the trajecories for each protein-ligand complex αβ (A), β (B), α (C), α’ (D). 84

Figure 4-3: The distance calculated between the center of mass (COM) of the two flaps 43-58/43'-58' over the length of the production run. The flaps are less stable for the duration of the trajectories for complexes αβ (A), β (B), as compared to α (C), α’ (D). 85

Figure 4-4: The distance between the catalytic Asp25 and the flap tip Ile50 for monomers 1 and 2 of HIVp over the length of the production runs. The flap-to-active-site distances indicate a wide range of motion for both monomers (1-2), with one majority flap conformation seen for simulations of αβ (A) and α (C), while β (B) and α’ (D) are characterized by several equally population flap conformations. The typical closed (14 Å) and semi-open (18 Å) distances are indicated with a black line. 86

Figure 4-5: The distance between the flap tip Cα to the 80s loop Cα for monomers 1 and 2 over the length of the production run. The distance fluctuates mainly between 10-20 Å throughout our simulations, but on average is much higher for the trajecories of αβ (A) and β (B), while α (C) and α’ (D) occupy a narrower range of distances. The typical closed (15 Å) and semi-open (10 Å) distances are indicated with a black line... 87

Figure 4-6: The overall RMSD calculated for each ligand within its protein-ligand complex, yielding the RMSD for α-only in HIVp+αβ (A), β-only in HIVp+αβ (B), β in HIVp+β (C), α in HIVp+α (D), and α’ in HIVp+α’ (E) over the length of the production run. Trajectories were first fit to the Cα core of the 3BC4 crystal structure... 89

Figure 4-7: The distance from the center of mass (COM) of the catalytic aspartic acids 25/25’ to the pyrrole nitrogen on the ligand, for α-only in HIVp+αβ (A), β-only in HIVp+αβ (B), β in HIVp+β (C), α in HIVp+α (D), and α’ in HIVp+α’ (E). The green line in C clearly depicts the run in which the ligand flips into the active site (the α position), as well as the run with a partial flip.. 90

Figure 4-8: Representative structures from the MD simulation of the HIVp+αβ complex, taken from the last 5ns of each 25ns trajectory. The α ligand is shown in green, the β ligand is shown in black, the S1/S1’ site is shown in yellow, and the S2/S2’ site is shown in purple. The conformational families demonstrate the instability of the 2:1 bound complex. The pyrrolidine ligands find a wide variety of ways to interact with the protease flaps, S1/S1’, S2/S2’, and/or the eye site... 92

Figure 4-9: Representative structures from the MD simulation of the HIVp+β complex, taken from the last 5ns of each 25ns trajectory. The β ligand is shown in black, the S1/S1’ site is shown in yellow, and the S2/S2’ site is shown in purple. The conformational variance of the β ligand is exceptionally high. The β ligand does not preferentially interact with the binding site or eye site, except in the case of (B). This family is the result of one run, where the ligand has moved to bind at the active site, displaying a standard α pose, with contact in one eye site and the S2 pocket......... 94

Figure 4-10: Representative structures from the MD simulation of the HIVp+α complex, taken from the last 5ns of each 25ns trajectory. The α ligand is shown in green, the S1/S1' site is shown in yellow, and the S2/S2' site is shown in purple. The conformational families for the α ligand illustrate its strong preference for forming one interaction between the naphthyl ring and the eye site, while the other naphthyl ring flips to interact at the S1/S1' or S2/S2' site, and the pyrrole maintains a hydrogen-bonding interaction with the catalytic aspartic acids............................... 96

Figure 4-11: The overall RMSD from the crystal pose calculated for each naphthyl ring of the ligand in HIVp+α over the length of the production run. Trajectories were first fit to the Cα core of the 3BC4 crystal structure. Each color represents a single production run; and denotes the same run for each plot. A) highlights the RMSD of the first naphtyhl ring over time, (B) highlights the RMSD of the second naphthyl ring over time. As noted in figure 1, we have used the convention of labeling monomers 1 and 2 based on the behavoir of the ligand, where better agreement with the initial position in the eye is oriented to the right in the figures and labeled as monomer 2 in the graphs. An RMSD of 6.2-7.9 Å indicates occupation of the S2/S2' site, while an RMSD of 7.8-10.1 Å indicates occupation of the S1/S1' site........... 97

Figure 4-12: A) The initial minimized conformation of the α' ligand. B-F) Representative structures from the 200ns MD simulation of the HIVp-α' complex. The α' ligand is shown in cyan, the S1/S1' site is shown in yellow, and the S2/S2' site is shown in purple. Although the simulations were initiated with the naphthyl rings occupying traditional subsites of the active site, one naphthyl ring moves to form interactions at the eye site over the course of all eight independent simulations. The second ring remains in contact with the S1 or S2 site. 98

Figure 5-1: The 1TW7 structure of HIVp, showing crystal contacts between neighboring flap tips and the elbow. The flap-tip residues in direct contact with the elbow cleft (residues 49-52) are shown in black. 101

Figure 5-2: (A) The GIGG sequence makes no contacts into the base of the elbow cleft, but (B) using D-Ile50 and D-Phe51 (green) increases the peptide's contact with the sides and bottom of the cleft by ~100 Å^2. 102

Figure 5-3: Distance restraints employed in the LD and MD simulations of the GIGG-HIVp complex. The Ac-G1-I2-G3-G4-NMe ligand was held in the elbow cleft with restraints woven between the elbow and the cantilever regions: A) I2(O) – Q61(O) = 6.17Å, B) I2(N) – R41(N) = 4.83 Å, C) I2(N) – D60(N) = 7.55 Å, D) G3(O) – V62(O) = 10.42 Å, E) G3(N) – P39(O) = 4.08 Å. These values are upper limits of the allowed distances; there is no penalty for forming closer contacts. This prevents dissociation but allows for some freedom in sampling and adaptation outside the crystalline environment of 1TW7. 104

Figure 5-4: The rmsf of the backbone heavy atoms is presented for the simulations of restrained GIGG-HIVp and apo HIVp (noted 1TW7). The restrained MD is often the most restricted, but the restrained LD is often the least - flaps are highly mobile in each simulation. 106

Figure 5-5: Percent of native contacts over the course of the unrestrained LD for (A,B) GIGG or (C,D) DifG bound to HIVp 108

Figure 5-6: The 3-ns snapshot of runs 2 and 4 from the unrestrained LD simulations of the GIGG-HIVp complex. These images demonstrate that while some of the native contacts are retained, the tetrapeptide is no longer associating with the elbow pocket (compare to contacts in Figure 5-1). 108

Figure 5-7: Representative cluster families display the conformations sampled in 85% of each simulation. Each set is overlaid to the crystal structure conformation of 1TW7 (in black): (A) LD of the restrained GIGG-HIVp, (B) LD of unrestrained apo HIVp, (C) MD of restrained GIGG-HIVp, and (D) MD of unrestrained apo HIVp. The structures are overlaid by the Cα of the core residues (all residues except the flaps 43-58 and 43'-58'), and the RMSD of the flap region is noted. 113

Figure 5-8: A wide degree of conformational sampling is seen in the LD simulations, whether GIGG is present or not. Both flaps of the restrained GIGG-HIVp (yellow and green lines) and unrestrained apo HIVp (red and black lines) simulations are shown. The snapshots were overlaid with respect to the Cα atoms of the core of the dimer in 1TW7, and the RMSD of only the flap Ca (residues 43-58 and 43'-58') are shown above 114

Figure 5-9: The distance from the flap tips Ile50/50' to the catalytic Asp25/25' during the implicit-solvent LD of (A) the restrained, GIGG-HIVp complex and (B) the unrestrained, apo HIVp. The individual flaps have different colors in each plot. Horizontal, black lines note the distances seen in different conformational states: the skewed-open 1TW7 is 18.8 Å, the semi-open 1HHP is 17.2 Å, and the closed 1PRO is 14.1 Å. 114

Figure 5-10: The activity of HIVp in the presence of 250 μM peptides is given: DifG (blue circle) and the model peptide for GIGG (Ac-GIGGK, orange triangle). The IC50 curve of the control PepA is also shown (purple circles). 115

List of Tables

Table 1-1: Techniques for incorporating full protein flexibility into docking approaches. 11

Table 1-2: Studies including full receptor flexibility after docking. 13

Table 1-3: Studies based on SC-Flex. Caveats for each method are similar in that flexibility of the side chains alone has limited success in describing receptor motion for most binding events........ 15

Table 1-4: Docking studies that utilized full receptor flexibility during the performance of IFD. 17

Table 1-5: Studies including full receptor flexibility through grid-based ensemble docking........ 21

Table 1-6: Studies including full receptor flexibility through a structural ensemble for docking........ 26

Table 2-1: The extent of convergence for HEWL in MixMD, as determined by the ratio of co-solvent to water molecules at the box edges. Values of Obs/Exp near 1.0 indicate complete and unbiased sampling at the edges of the box. 47

Table 2-2: The extent of convergence for fully flexible HEWL in MixMD, as determined by the ratio of CCN to water at the box edges. Simulations of 90% w/w CCN were problematic, and the behavior at the edges of the box show that they do not demonstrate even, unbiased sampling (the Obs/Exp value is much larger than 1.0).48

Table 3-1: Analysis of the refined cycles in Buster of electron density for each MSCS probe site in the noted crystal structure. The real-space correlation coefficient from the EDS is shown in parentheses. 59

Table 3-2: Probe-to-water ratio at the box edges over the last 5 ns of simulation time, using 5 runs. The expected ratios differ slightly due to system setup. Upon introduction of the protein into a 50% w/w pre-equilibrated solvent box, a random number of water and probes are removed from the system to avoid clashes with the protein. This can shift the probe:water ratio, thus slightly altering the composition of solvent in the system. 64

Table 5-1: Snapshots from the implicit-solvent LD simulations were clustered into 20 conformational families. Below, Cα-rmsd is used to compare each representative to the dimer of HIVp in the 1TW7 structure. The "RMSD All" is the standard comparison by global overlay of all residues in the dimer; the other two metrics are obtained by overlaying only the core of HIVp to the 1TW7 structure and then measuring RMSD for the flaps and core separately (measurement, not additional overlays). The flap residues are 43-58 and 43'-58', and all other residues are included in the core. 110

Table 5-2: Snapshots from the explicit-solvent MD simulations were clustered into 20 conformational families. Below, Cα-rmsd is used to compare each representative to the dimer of HIVp in the 1TW7 structure. The "RMSD All" is the standard comparison by global overlay of all residues in the dimer; the other two metrics are obtained by overlaying only the core of HIVp to the 1TW7 structure and then measuring RMSD for the flaps and core separately (measurement, not additional overlays). The flap residues are 43-58 and 43'-58', and all other residues are included in the core. 111

Abstract

Structure-based drug design (SBDD) is defined as the use of three-dimensional structural data to advance lead development and optimization studies. Many SBDD projects have used a rigid protein structure to represent the receptor target in order to gain greater throughput with minimal computational time. However, numerous studies have illustrated the significant influence protein flexibility exerts upon binding predictions. Inclusion of protein flexibility has become essential due to the need for ligands with novel scaffolds and unique modes of action that combat increasing rates of drug resistance and decreasing approval of clinical candidates. Additionally, accurate modeling of protein flexibility may reveal unknown allosteric sites and increase the number of viable lead compounds for a given target.

Previously, Carlson *et al*. incorporated structural flexibility into pharmacophore modeling through the development of the multiple protein structure (MPS) method (2000). This technique was the first computational-mapping algorithm to identify experimentally-validated lead compounds. Probe mapping is a common computational technique for identifying potential binding pockets along a protein surface. However, the efficacy of most methods has been limited by neglecting desolvation penalties. To broaden the impact of our studies, we have developed an improved technique for probe mapping, Mixed Solvent Molecular Dynamics (MixMD), which extends our MPS approach by simultaneously incorporating flexibility *and* solvent competition. This technique has been validated on the canonical hen egg-white lysozyme system and has been generalized across a series of pharmaceutically-relevant targets. MixMD can be used to develop accurate pharmacophores of druggable hot spots through the incorporation of several different probe types.

As a complement to our methodology development, we have specifically targeted protein flexibility in another canonical protein system. HIV-1 Protease (HIVp) is an exceptional test case due to the abundance of structural data available, its importance as a pharmaceutical target, and

its potential for allosteric regulation. Three allosteric sites have been hypothesized for HIVp: the elbow site, the eye site, and the dimer interface. We have used MD simulations to probe the allosteric control possible at the elbow and eye sites by small molecules. Our studies have identified important features for designing effective allosteric inhibitors of HIVp.

Chapter 1

Introduction

1.1 Specific Aims

The specific aims of my study encompassed 1) the development of an improved method for mapping flexible targets, based upon our achievements with the Multiple Protein Structure (MPS) method and 2) the investigation of potential allosteric regulation of HIV-1 Protease (HIVp).

My underlying hypothesis was that better techniques for structure-based drug design (SBDD) enable the discovery of ligands with new scaffolds and novel modes of action. We expected that through the incorporation of protein flexibility and new surface-mapping methods, an improvement would be seen in terms of site-prediction accuracy as well as identification of potential allosteric sites.

My long-term goal has been to advance the current state of SBDD through studies that improve our understanding of the impact of protein flexibility on binding. Through successful mapping studies of pharmaceutically-relevant targets using the MPS method, the Carlson group has demonstrated that receptor flexibility enables results to surpass the known limitations of rigid models; my research has built from that success.

1.2 Theory of Molecular Mechanics

Computer technology has rapidly progressed, enabling the simultaneous advance of computational tools for basic and applied science. Medicinal chemistry has particularly benefitted from technological advances that allowed detailed analyses of binding characteristics in pharmaceutically-relevant target systems. Several computational methods have been developed for studying macromolecular systems, ranging from quantum mechanics (QM) to molecular mechanics (MM). Although QM provides information on energetic interactions at the atomic scale at the highest possible level of accuracy, their huge computational cost has rendered them infeasible for large systems (greater than 100 atoms). MM is based on statistical mechanics theory; therefore, it can be used to predict the energy of biomolecular systems with reasonable accuracy. In MM, the potential energy for all atoms in the system is calculated from the sum of the covalent (bond, angle, dihedral) and noncovalent (van der Waals, dipole, coloumbic) contributions, according to a user-specified force field.

Force fields contain the applied energy function and the set of parameters used to describe all atoms within the system. These parameters are usually derived from high quality experimental and QM data that represent a typical biomolecular system (proteins, nucleic acids, lipids, small organic molecules). Numerous parameter sets exist for each MM program, but combining or transferring parameter sets across force fields and/or programs is inadvisable.

The stochastic Monte Carlo (MC) method and the deterministic molecular dynamics (MD) method are two forms of MM simulation.[1] MC is reliant on a distribution of probabilities to randomly sample the potential energy surface of a system, while MD integrates over Newton's equations of motion to dynamically sample the energy landscape: $F_i = m_i a_i = -\nabla_i V \rightarrow -\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2}$. Here, F_i is the force exerted on atom *i*, m_i the atomic mass of *i*, a_i the acceleration of *i*, and V the potential energy of the system. The general equation to describe MD simulations, including those performed in AMBER, is given by the potential function[2]:

$$V(R) = \sum_{bonds} K_r \left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}\left(1 + \cos[n\phi - \gamma]\right) + \sum_{i<j}^{atoms} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j}^{atoms} \frac{q_i q_j}{\varepsilon r_{ij}}$$

These terms define the bond-stretching, angle-bending, torsion-rotation, van der Waals (vdW), and electrostatic forces respectively. Bond stretching and angle bending are described by a harmonic oscillator function, torsions by a periodic function, vdW interactions by the Lennard-Jones potential, and electrostatics by Coulomb's Law.[1] Usually, the positions of the atoms are given by a crystal or NMR structure placed within a computer-generated solvent box. Velocities are assigned to the starting xyz coordinates based on a Maxwell-Boltzmann distribution, where the magnitude is proportional to the user-defined temperature. The potential energy is calculated based on the atomic coordinates, and the individual atoms accelerate/decelerate according to the resultant force. Time is advanced by a user-specified timestep (usually 2 fs), and each step calculates new atomic positions, velocities, and forces according to the Verlet integrator[3]. The appropriate timestep is dependent on the fastest motions within the system; therefore, bonds to hydrogen are typically restrained[4] to allow a larger timestep.

Since MC simulations involve random-step searches of the energy landscape, they cannot provide a time-dependent view of the system at a given temperature. However, MD is capable of describing of the system's energetics over a simulation time. True system dynamics are not observed in most MD because coupling to a temperature bath scales velocities at various intervals. Assuming the veracity of the ergodic hypothesis, which states that all available microstates of the system will be sampled given an "adequate" timescale,[5] both MC and MD should give appropriate distributions of states if run long enough. To optimize sampling, it is frequently desirable to perform multiple shorter simulations instead of one long simulation.

In the Carlson Lab, MD is applied to biomolecular systems through the program AMBER (Assisted Model Building with Energy Refinement).[6] Under a nonpolarizable force field, individual atoms are represented as spheres with assigned vdW radii and a constant net charge. The basic force field used within AMBER is described by the potential energy function. Several force fields are available for use with MD simulations, and the choice of which to implement is dependent on efficacy, performance, and the system under study. When the original FF94 parameter set was shown to overstabilize α-helices, continual evolution of the force field over time led to the development of FF99SB.[7] Studies comparing computational results with experimental data strongly support the accuracy of this parameter set for most proteins in AMBER, thus we have chosen to apply FF99SB in our protein-ligand systems.

In simulations of a biomolecular complex in solution, the explicit inclusion of water molecules can greatly impact simulation time. In some cases, exploration of the motion of the biomolecule is more important than observing specific solute-solvent interactions; the solvent can be represented implicitly by a continuum model. A number of continuum solvent models have been developed, the most prevalent being the generalized Born (GB) model.[8] Our calculations involving the generalized Born approach involved the use of a modified set of parameters published by Onufriev *et al*, $GB_{OBC.}$[9]

The generalized Born/surface area model mimics the effect of solvent through the addition of two terms to the "vacuum" potential energy function previously noted as the basic force field for molecular mechanics: $\Delta G_{sol} = \sum_{ij}(1-\frac{1}{\varepsilon})(q_i q_j / f_{GB}(r_{ij})) + A\sum_i \sigma_i$.[10] These terms define the polar component of the solvation free energy and the non-polar contribution, which is proportional to the surface area of molecule A. The term f^{gb} is a smoothing function, which depends upon the atomic radii and interatomic distances r_{ij}. The advantage to implicit solvation is that it allows calculations to be performed at a reduced expense because solvent motion does not need to be explicitly calculated. Additionally, the lack of viscous drag from the solvent allows the molecule of interest to move more quickly through the potential energy space available. In order to more accurately account for explicit solvent-solute interactions, Langevin dynamics (LD) were employed to account for the random buffeting of the solute by the solvent. The Langevin equation is based on Newtonian physics and models continuum solvent interactions through the relationship;

$$m_i \frac{d^2 r_i}{dt^2} = F_i\{r_i(t)\} - \xi \frac{dr_i(t)}{dt} + R_i(t)$$

where R is defined as the random force and ξ as the frictional coefficient.[1]

1.3 Structure-based Drug Design

Structure-based drug design (SBDD) involves the use of three-dimensional structural data to advance lead identification and subsequent optimization for drug discovery. The exponential growth of the PDB and improvement in homology modeling techniques makes SBDD applicable to an ever-growing number of pharmaceutically-relevant targets with a three-dimensional structure available. SBDD studies are generally centered upon at least one of the following three goals: prediction of binding modes, prediction of binding affinities, and prediction of novel binding partners. Depending on these goals and available data, SBDD can involve different approaches, which are often separated into docking techniques and *de novo* design. Docking is implemented to predict the binding mode and affinity of small molecules. When docking is applied to a large compound database, it is referred to as virtual screening (VS) and often centered on the prediction of new ligands with high affinity for a target molecule to enrich the compound set for experimental testing. *De novo* design is performed with the intent of predicting new compounds in novel chemical space. Fragment-mapping techniques that use functional groups to probe binding sites are often applied for this type of approach. Over the past 20 years, these SBDD studies have produced viable leads, enabling the development of successful clinical drugs and leading to the extensive implementation of SBDD in medicinal chemistry research.[11-13]

X-ray crystallography and NMR studies have clearly demonstrated conformational differences between many receptors' holo (bound) and apo (unbound) states. Sampling ligand conformations is straightforward; most SBDD protocols now include ligand flexibility (with on-the-fly sampling being superior to a rigid set of pre-generated conformations), yet this is insufficient for the most accurate results. Despite data demonstrating the influence of protein flexibility on ligand binding, most SBDD efforts still rely upon a static receptor structure because of the resources required to account for its many degrees of freedom. Although a few proteins can have their binding potential represented with a single, fixed conformation, for most systems the information presented by a rigid structure is simply inadequate.

In a comparative study of 10 docking programs and 37 scoring functions, no single method outperformed the others when performing rigid docking of diverse compounds to a set of eight proteins.[14] The scoring functions were unable to accurately predict binding affinity or relatively rank the compounds. Ginalski and co-authors evaluated the binding predictions and scoring results for 1,300 protein-ligand complexes from PDBbind 2007 with Surflex, LigandFit, Glid, GOLD, FlexX, eHiTS, and AutoDock.[15] The authors found that the programs achieved a mean $\text{RMSD}_{\text{Top-Sscore}}$ that ranged from 2.77 Å (GOLD) to 4.37 Å (FlexX) for the docked poses, none of the scoring functions were able to achieve a reasonable correlation between the pose score and the experimental activity. Several recent reviews have been published that discuss the deficiencies of existing scoring functions for docking.[14,16-20] Klebe and coworkers hypothesized that there is an intimate link between docking and scoring, postulating that correct prediction of accurate binding geometries will simultaneously solve the scoring problem.[21-23] Therefore, while we acknowledge the limitations of current scoring functions, we focus on the variety of techniques for incorporating protein flexibility in SBDD.

A number of reviews have included some consideration of the impact of protein flexibility on drug discovery.[16,21,24-34] Many of these reviews concentrated on only a subset of protein flexibility methods or emphasized a particular technique. Here, we examine a variety of techniques that account for flexibility in protein-ligand binding, not only studies where the primary focus is on docking but also studies that concentrate on binding-site mapping.

1.4 Protein Flexibility

1.4.1. Conformational variability

Our understanding of receptor-ligand binding has significantly advanced from the original lock-and-key model proposed by Fischer.[35] Early experimental studies showed that the act of ligand binding influences the protein conformation, referred to as conformational induction or induced fit.[36] Another model of ligand binding is conformational selection, wherein the ligand chooses a binding partner from among available states in the conformational ensemble, thereby shifting the population distribution.[37-42] Recently, several papers have been published examining the evidence for whether receptor binding occurs because of induced fit or conformational selection. Sullivan and Holyoak studied the kinetics of phosphoenolpyruvate carboxykinase and concluded that induced fit and conformational selection were not mutually exclusive, rather they were complementary avenues for binding.[43] Weikl and von Deuster developed equations of binding kinetics using a four-state protein-ligand complex to distinguish between induced and selected fit.[44] Hammes *et al*. examined the binding pathways for dihydrofolate reductase (DHFR) and flavodoxin and determined that a mixed binding mechanism was most likely.[45] They noted that the relative importance of induced fit or conformational selection for a particular case could be analyzed by comparing the reaction path flux. Both mechanisms of binding will produce the same result; it is important only that some mechanism of receptor conformational change be incorporated in docking simulations. This is particularly useful in SBDD because it implies that computationally inexpensive methods that include protein flexibility should correctly predict binding modes.

1.4.2. Allostery

Protein-ligand flexibility upon binding is crucial for proper understanding of allosteric regulation. Nussinov and coauthors postulated that most proteins exist in an ensemble of states, and thus most proteins have the potential for allostery.[46] Based on experimental literature, they demonstrated that the binding of an allosteric ligand shifted the population of conformational substates, thereby influencing the ability of other ligands to bind an alternate site.

Drug resistance is a growing problem that calls for new approaches to drug therapy. By exploring allosteric control in protein targets, we can find new modes of action and hence, overcome emergent resistance, and develop cocktails of drugs to improve treatment.

1.5 Protein-Ligand Binding

1.5.1. The Cross Docking Problem

Research into protein flexibility and allostery has lent support to the importance of representing multiple states in binding studies (Figure 1-1). Mobley and Dill noted that binding free energy (ΔG_{bind}) and entropy are influenced by the shape and width of the entire conformational landscape, rather than a single rigid pose.[47] Murray *et al.* examined approaches for using rigid receptors in docking studies.[48] When the authors attempted to dock a known ligand into a protein structure solved in the presence of a different ligand (referred to as cross-docking), they found that the active site was biased towards to the native ligand (Figure 1-2). A variety of differences in the surface of the binding site were identified for the same protein solved with different ligands or in the absence of a ligand. Movement was observed in the backbone, side chain (both dependent and independent of the backbone Cα), and active site metals. As a consequence, active sites were biased towards a particular ligand type, negatively impacting docking efforts. The resultant misdocking could not be overcome without accounting for critical conformational shifts. Najmanovich *et al.* examined the percentage of residues that actually undergo a change upon ligand binding, based on two non-redundant datasets containing paired holo and apo protein structures from the PDB.[49] No significant correlation was found between backbone movement and side chain flexibility, but Lys, Arg, Gln, and Met were identified as the residues most likely to demonstrate side-chain rearrangement upon ligand binding. Najmanovich *et al.* showed that 60-70% of the binding site undergoes some change in side chain orientation. Taken together, these studies clearly illustrated that the lack of protein flexibility in typical SBDD efforts severely limits the identification of true ligands and accurate docked poses.

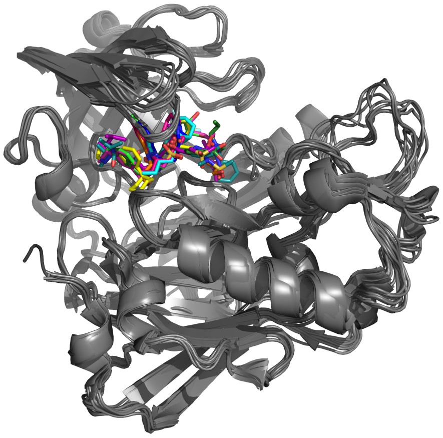


Figure 1-1: The conformational variation in holo and apo BACE structures illustrates structural differences frequently seen across multiple crystal structures of the same protein.

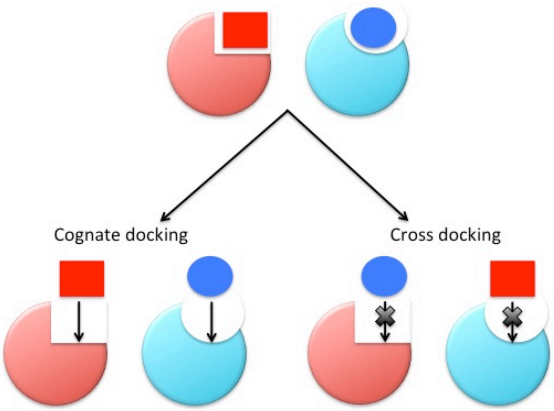


Figure 1-2: The same protein crystallized with different ligands. This simple two ligand-one protein set demonstrates the influence of structure on the shape of the binding pocket.

An underlying assumption in rigid docking efforts is that the complexed state is the lowest free energy state. However, the ligand-bound state is not always the lowest energy state for either the protein or the ligand. Smith *et al*. were first to examine conformational similarity, they compared the four available structures of interleukin-4, two from NMR and two from crystallography.[50] They found that differences in the experimental methods clearly affected the structure, and that the two NMR structures were more similar to the crystal structures than to each other. Further, they noted that the exposed regions of the surface, particularly the highly flexible loops, showed the largest conformational difference among the structures. Eyal *et al*. evaluated a set of 659 structure pairs from the PDB and showed that experimental variation in protein structures certainly exists, as a result of refinement, crystal packing, and/or crystallization conditions.[51] Chopra *et al*. employed MD on a set of 75 proteins to explore the role of solvent in structural determination and concluded that solvent effects have an enormous influence on the final refined protein structure.[52]

1.5.2. Sources of Structural Information

Use of a set of similar conformations only generates a finite amount of information on potential binding partners. Additional receptor space should be explored, especially in scenarios where alternate binding modes are of interest. X-ray crystallography, NMR structures, homology models, normal mode analysis (NMA), and molecular mechanics simulations are all potential sources of structural diversity. Advances in technology have allowed for significant increases in available structural information for a vast number of receptor targets. The Protein Data Bank (PDB) now contains over 74,601 structures (as of 7/19/11), up from 13,605 structures in 2000.[53] Several research groups, as well as thousands of computer users, have dedicated their computer power to distributed structure prediction of proteins with Folding@HOME, POEM@HOME, and Rosetta@HOME. Martin-Renom *et al*. estimated that almost one-third of all known protein sequences can have their structure predicted via homology modeling efforts.[54] Schafferhans and Klebe found that accuracy increases when the results of several different homology modeling programs are combined.[55] While predictive models are not as accurate as structures solved by x-ray crystallography or NMR, they enable the study of targets that are difficult to determine experimentally. No consensus has been reached on the best source of data for including structural flexibility in SBDD, and a number of studies clearly demonstrate the difficulty and variability in assessing appropriate conformations.[56-60]

The following sections are organized based on different techniques for including protein flexibility in docking and *de novo* design (Table 1-1). Soft docking allows limited discovery of new conformation space while docking with flexible side chains (SC-Flex) samples observed conformational space. More recent techniques incorporate ensembles of structures to allow for a greater measure of native flexibility or to reveal new conformations. By relying on several protein conformations, new chemical space can be explored, allosteric sites can be discovered, and more accurate binding conformations can be generated. The accepted metric for a well-docked pose is ≤ 2.0 Å RMSD from the experimentally determined structure. Protein flexibility can be incorporated into a binding-site model before docking is initiated or it can be allowed post-docking through refinement of the bound complex. Docking can be performed against an average structure or ensemble, a series of rigid-receptor conformations (semi-flex), or with conformations generated “on the fly” (induced-fit docking, IFD).

Table 1-1: Techniques for incorporating full protein flexibility into docking approaches.

Approach	Method for incorporation	Advantages	Limitations
Refinement	Flexibility introduced after docking through reduced or all-atom modeling with molecular dynamics or Monte Carlo minimization.	Fast docking to rigid receptor enables searching through vast compound libraries	Unlikely to generate as much structural diversity as the other methods, hard to move beyond known binding space
Average or unified structure	Ensemble averaging through use of a unified structure or grid representation. Can also occur through the selection of conformational subsets from a rotameric library, May also involve generation of receptor conformations based on ligand poses.	Can find novel binding mechanisms, orphan sites, and explore new receptor conformations	Discovery of "paradoxical inhibitors" that bind only to averaged conformation but not a native structure
Serial docking	Docking performed iteratively to a rigid ensemble of structures, conformational variation of the receptor ensemble typically comes from the inclusion of several X-ray crystallography, NMR, homology model, PCA-derived, NMA-derived or MD-derived structures.	Can allow for discovery of novel binding modes	Ensemble generation and parallel docking can be time-consuming, not usually appropriate for screening large libraries Structural variation can increase false positives
Conformations on the fly	Receptor conformational changes are explicitly modeled during docking.	Allows receptor conformation to change during interaction with ligand for optimal binding, can be quite accurate	Can generate conformations that are not experimentally accessible Can be quite time-intensive, not necessarily appropriate for virtual screening

Experimental methods for SBDD inherently include protein-ligand flexibility. However, lead development and design that is based solely on experimental data for ligand binding can be time-consuming and prohibitively expensive since every lead optimization step must be tested. Fragment-based drug design (FBDD) has enabled the application of experimental methods to searching the interaction potential along the protein surface with small organic probes. The use of fragment-based studies to SBDD enables the identification of ligands with high affinity for targets that have been traditionally difficult to characterize. These insights have greatly influenced computational approaches to protein flexibility, where fragment mapping can be applied to flexible receptors with functional group probes to find binding hot spots and identify specific interaction potentials. The use of molecular dynamics and/or probe minimizations to search for important binding sites with small probes has frequently applied in SBDD projects because the results enable the development of a pharmacophore model for use in virtual screening studies.

1.5.3. Early Protein-Ligand Flexibility in Docking

Soft Docking

It was not until the mid-90s that researchers began to tackle the problem of receptor degrees of freedom in SBDD. The original technique involved soft docking, which accounted for minimal backbone movement, and side-chain flexibility through attenuation of the Lennard-Jones repulsion term between the rigid protein and ligand.[61] This allowed for some penetration between the ligand and protein, and was followed by a rigid-body minimization. Soft docking increased the number of high scoring hits compared to the use of a single structure for docking. A study of T4 lysozme and aldose reductase showed that soft docking was superior to docking to a static structure.[62] However, the authors also found that serial docking to rigid receptors in a structural ensemble identified six novel compounds that soft docking had ranked poorly. Of these six compounds, four were experimentally verified and one had a low micromolar IC_{50} value.

Soft docking is a fast and easy method for including some protein flexibility into docking studies, and as a result is it included as a step in many of the methods presented below.

Relaxation Methods

Refinement of the docked complex is another simple approach that adjusts for protein flexibility by modeling induced-fit effects. The incorporation of protein flexibility on the “back end” can only be done when the docking technique was based on an all-atom structure; it cannot be performed on a grid, or in any other situation where the protein is not explicitly represented. MC or MD simulations are a popular choice for refinement because they enable the optimization of docked poses investigation of solvent effects, examination of kinetic stability, and prediction of ΔG_{bind} from physics-based scoring functions. Refinement is frequently performed as a final step in many of the docking approaches discussed here; Table 1-2 lists protein-ligand docking methods that limit the inclusion of full flexibility to the refinement stage.[63]

Table 1-2: Studies including full receptor flexibility after docking.

Method	Target	Flexibility	Results	Caveats	Author
FDS	15 cases from GOLD test set	Rigid-pro directed by hydrogen bonding Results clustered (clique finding technique), ~5 poses subjected to SC-Flex via MC with GB/SA, soft potential, & SA	Rigid-pro reproduced exp. pose for 13 cases Refinement with SC-Flex reproduced exp. pose for 11 cases Accounted for continuum solvation For full-flex, RMSD of pose closest to crystal conformation (majority ranked 1) was 0.78 to 3.81 Å to the crystal structure	No consideration of the effects of bound water Minimal conformation change Required 30-40 hours for a single run	Taylor *et al,* 2003
ADAM/ BLUTO	18 cases for native docking; 22 for cross docking	Docked to binding site as vdW-offset grid Post-docking optimization of ligand and protein side chains (within 7 Å) of the binding site	Top-ranked poses for thymidine kinase from ADAM/BLUTO were more accurate (RMSD 0.52-1.89 Å) than rigid docking results from GOLD (RMSD 0.49-3.11 Å), DOCK (RMSD 0.82-9.62 Å), Surflex (RMSD 0.74-3.84 Å), Glide (RMSD 0.35-4.22 Å), FlexX (RMSD 0.78-13.30 Å), and ADAM (RMSD 0.49-3.11 Å) RMSD of top-ranked model for all cases from ADAM/BLUTO ranged from 0.43 to 2.66 Å for cross-docking RMSD of top-ranked model for all cases from ADAM alone ranged from 0.67 to 6.31 Å for cross-docking	Most relevant for studying local changes in binding site conformation	Mizutani *et al*., 2006

AMMOS	Estrogen Receptor, Neuramindase	Five options to refine docked complex, from full to zero protein flexibility Minimized results and reranked Assessed impact of AMMOS force field minimization vs. MM94 and Tffs on results for four known ligands	Initial docking in DOCK6 RMSD of refined structure to the crystal conformation ranged from 1.03 to 2.12 Å for AMMOS, 1.01 to 2.39 Å for MMFF94s, 1.17 to 1.57 Å for Tff Enrichment increased by AMMOS refinement from 40% to 60% overall, with actives retrieved from the top 3-5% of the data set	Extent of minimization was very limited (2x500 cycles) Only identifies minor conformational shifts	Pencheva *et al.*, 2008
Enhanced molecular docking	Human prion protein	Used metadynamics to refine docked complexes	Calculated dissociation ΔG was 7.8-8.6 kcal/mol while experimental dissociation ΔG was 7.5 kcal/mol Predicted multiple binding sites Affinities agreed with NMR experiment	Computationally intensive	Kranjc *et al.,* 2009

1.5.4. Docking with Flexible Side Chains

Early efforts at “on the-fly” docking focused on the side chains, with the use of a rotamer library based on backbone dihedral angles to describe protein flexibility. In 1993, rotamer libraries were first used to predict side-chain conformations while studying the protein folding problem.[64] In 1994, the use of side-chain rotamer libraries was extended to protein-ligand docking with the examination of trypsin-benzamidine and antibody McPC 603-phosphocholine binding through a modified version of AMBER 4.0.[65] The study restricted the protein backbone completely while the side chains within 10 Å of the ligand were allowed to sample a set of discrete rotameric states. Leach found that the presence of the ligand revealed additional accessible conformational states by modulating the protein’s potential energy surface.

Similar side-chain only methods for modeling protein flexibility in docking are presented in Table 1-3, including AutoDock 4.0[66], Dynasite/GOLD[67], FlexX[68], ICM/MC[69], Mining Minima[70], Skelgen[71,72], SLIDE[73-75], and SOFTSPOTS/PLASTIC[76].

Table 1-3: Studies based on SC-Flex. Caveats for each method are similar in that flexibility of the side chains alone has limited success in describing receptor motion for most binding events.

Method	Target	Flexibility	Results	Author
Rotamer Library	Trypsin, McPC 603	Rotameric states sampled for all side chains within 10 Å of binding site	Limited by the lack of a solvation term RMSD of closest structure to the bound pose from the crystal was 0.7 Å for trypsin, 0.8 Å for McPC 603	Leach, 1994
ICM/MC	Leucine zipper helices	Applied internal coordinated modeling (ICM) and MC minimization to all side chains at 600K	Lowest energy conformation of leucine zipper had RMSD of 1.18 Å to crystal pose	Abagyan *et al.*, 1994
FlexX	19 cases	Greedy incremental construction of the ligand into the active site from base ligand fragment	RMSD of pose closest to crystal conformation (majority ranked 1) ranged 0.48 to 1.20 Å	Rarey *et al.*, 1996
SOFT SPOTS/ PLASTIC	Thymidylate synthase	Identified variation based on structural comparison (or binding site analysis) Disregarded polar residues Retained hydrophobic residues and loops for potential adaptation Subjected 3 residues to rotamer variation Minimized docked pose	Minimized, remodeled complex achieved reasonable energy scores for two potent inhibitors Found -51.5 kcal /mol for ligand BW1843U89 and -49.7 kcal /mol for CB3717 for cross-docking with SC-Flex Found -32.8 kcal/mol for CB3717 and -44.7 kcal/mol for BW1843U89 with rigid-pro Scores from native docking were -56.5 kcal/mol for BW1843U89 and -52.9 kcal/mol for CB3717	Anderson *et al.*, 2001
Mining Minima	Set of 18 protein-ligand crystal structures	On-the-fly optimization of the 5 nearest side chains (to the x-ray ligand) or hydroxyl-only hydrogens on the fly Three docking runs per system Native docking	RMSD from SC-Flex of lowest energy pose ranged from 0.34 to 2.32 Å to the crystal orientation for 8 known cases SC-Flex improved accuracy relative to rigid-pro for 4 cases while it diminished accuracy relative to rigid-pro for the other 4 SC-Flex highlighted important SC movement in hypothetical protein case	Kairys & Gilson, 2002
Skelgen	MMP-1, Acetylcholin-esterase	Random transitions of side chain χ angles were found to be preferable to a rotamer library due to the variation in composition &	SC-Flex RMSD of pose closest to crystal conformation (majority ranked 1) were 1.0-1.3 Å (native), 1.3-1.4 Å (non-native), 1.4 Å (apo) Rigid-pro RMSD of pose closed	Alberts *et al.*, 2005a

		construction of rotamer libraries	to crystal conformation (majority ranked 1) was 0.7-1.0 Å (native), 4.4-4.5 Å (non-native), 5.6 Å (apo)	
SLIDE	20 cases	Ligand anchor fragment, induced fit ligand/side chain rotation based on mean field theory Binding site represented as interaction points Docked to the apo pose Included active site solvation Minimal rotation hypothesis: side chains will move as little as necessary to form complex	RMSD was ≤ 2.5 Å to the crystal orientation for all known ligands Specifically intended for systems without large global rearrangements (required apo to holo RMSD of ≤ 0.5 Å)	Zavodszky *et al.*, 2005
AutoDock 4.0	188 native cases, 87 HIVp cross-docking cases	Selected side-chain flexibility, receptor represented as a grid	SC-Flex docked small native molecules well SC-Flex successful for most large inhibitors of HIVp, failed relative to rigid-pro for small inhibitors >50% of the cases (3.5Å)	Morris *et al.*, 2009

Allowing side-chain flexibility was less resource intensive than full flexibility methods and it enabled some conformational variability through the exploration of low energy orientations of side chains. However, incorporating side-chain flexibility has failed in cases of proteins with large-scale hinge or loop rearrangements or even backbone-dependent movement of side chains, neither of which can be taken into account by SC-Flex.

1.5.5. Induced Fit Docking

Conformational generation on the fly

Significant progress has been made since the initial development of SC-Flex. Sampling motion of the side chains “on the fly” increases the potential energy space but can still overlook global conformational shifts. Allowing for conformational changes “on the fly” during docking can be a highly accurate technique for modeling bound poses of protein-ligand complexes. Induced fit docking (IFD) is important because it can allow the docking simulation to search new conformational space, however this sampling of receptor and ligand degrees of freedom is also quite computationally intensive, which limits its application in large-scale virtual screening studies.

The first "on the fly" method that expanded beyond conformational sampling of side chains was based on DOCK 4.0.[77] The authors included the conformational ensemble as an adjustable variable in the IFD algorithm as opposed to performing serial docking across an entire ensemble. In addition to the usual six degrees of freedom for ligand flexibility, the authors added a protein conformation term, which allowed the protocol to simultaneously optimize the ligand conformation and choose the best protein structure for binding. Ten separate protein ensembles were used for development and were obtained from the set by Cavasotto and Abagyan[78], the set by Claussen *et al.*[79], and two HIV-1 protease (HIVp) ensembles. They achieved a success rate of 93% when the five highest-ranked poses were included using a threshold for success of 2.5 Å RMSD from the experimental pose. Their method ran for a similar amount of time as rigid-pro and was significantly faster than traditional techniques for serial docking to rigid receptors. Other IFD methods for including protein flexibility include 4D Docking[80], FITTED[81-83], GLIDE/PRIME[84,85], PC-RELAX[86,87], REMD/PLOP[88], and SCaRE[89], as presented below in Table 1-4.

Table 1-4: Docking studies that utilized full receptor flexibility during the performance of IFD.

Method	Target	Flexibility	Results	Caveats	Author
GLIDE/ PRIME	21 cases	Soft-potential docking to RR Ala replacement ≤ 3 residues Top 20 complexes refined & redocked	Avg RMSD = 5.5 Å for rigid-pro, 1.4 Å for IFD	Very time intensive, limited to local motion	Sherman *et al,* 2006
Conform-ational ensemble as docking variable	10 cases	Conformational ensemble included as a variable during docking with DOCK4.0 Selected each representative side chain within the active site based on which had the greatest distance from the reference	Ensemble docking was the same speed as rigid-pro Ensemble docking had 67-100% success based on pose and score Rigid-pro had 23-87% success in cross-docking based on pose and score	Use of a unified representation of the receptor can result in the identification of high-scoring false positives	Huang and Zou, 2007

		sphere points from SPHGEN Bound complex was optimized with SIMPLEX Success defined as RMSD from crystal pose of ≤ 2.5 Å and an energy score > the native docking	Sequential rigid-pro had 33-100% success based on pose and score		
SCaRE	16 cross-docking pairs	Optimal results with Ala replacement of 2 neighboring side-chain pairs Ligand docked to gapped conformation Clustered, optimized, and refined complex with original side chains	With pocket boundaries equal to all residues ≤ 5 Å from the ligand, best RMSD for a docked pose to the crystal structure (with majority rank 1) was 0.7-2.0 Å (90% success) Rigid-pro 50% success rate	Limited view of active site flexibility Time consuming	Bottegoni *et al.*, 2008
PC-RELAX	set of CDK2 structures	PCA/ANM-derived soft mode docking with no flexibility, side chain flexibility, backbone flexibility, or full flexibility	Starting with apo structure, found best RMSD to the crystal structure of 0.8-1.8 Å (rank ranges 1-9) for fully flexible versus 0.5 to 5.2 Å (rank ranges 1-18) for rigid-pro	Time intensive relative to rigid-pro, cannot capture highly flexible motion	May & Zacharias 2008
REMD/ PLOP	6 cases	REMD used to generate low energy loop conformations After clustering, loops refined with PLOP Lowest energy conformer used for docking with GLIDE	RMSD of docked pose between holo and predicted structure was 1.4-12.5 Å RMSD of docked pose between holo and crystal structure was 1.0-2.5 Å RMSD of docked pose between holo and apo structure was 1.8-13.2 Å	Limited by efficiency of REMD for generating loop conformations	Wong & Jacobson, 2008
FITTED 2.6	5 protein-ligand sets	Modified GA for receptor chromosome Allowed switching between conformations, side chains in the binding site, and/or	Success rate based on RMSD of docked pose to crystal structure was 79% for native rigid-pro, 56% for cross-docking rigid-pro, 67% for	Flexible docking was unable to select the exact experimental complex structure While computational	Corbeil *et al.*, 2009

		water positions Can perform rigid-pro, semi-flex, or full-flex	SC-Flex, and 67% for flexible docking Notable speed increase over previous versions	speed increased, accuracy decreased between FITTED 1.5 and 2.6	
4D Docking	267 nonredundant structures	Ensemble assembled onto 4D grid based on binding potential and superposition During docking, could switch receptor conformation as well as ligand conformation	4D docking 73% success rate with 3-8 conformers when cognate receptor wasn't included For the same scenario, semi-flex had 71.1% success rate 4D docking was 4x faster than semi-flex	Performance decreased with > 8 conformers 4D docking marginally less successful than semi-flex	Bottegoni *et al.*, 2009

1.5.6. Ensemble Docking

Ensemble docking differs from IFD in that protein flexibility is accounted for prior to the actual docking. Although the studies by Huang and Zou[90] as well as Bottegoni *et al.*[80] used a pre-existing ensemble of conformations, they sampled those conformations on the fly, and so they are included above in the IFD section. Two different types of methods exist for representing receptor flexibility during docking: grid-based ensembles and structure-based ensembles. Frequently, alternate protein conformations are represented on a two-body potential grid, enabling fast, inexpensive docking simulations. Flexibility can also be incorporated into binding predictions through the sequential docking to structures in a conformational ensemble or docking to an averaged/united receptor structure. Due to the time required for initial ensemble generation and repeated rigid-pro, sequential docking is typically the most computationally intensive approach. However, it avoids the discovery of receptor-ligand complexes that are physically impossible (paradoxical ligands), which are sometimes seen in the results from docking to an average structure.

Grid-based ensembles

The first docking method to use a composite grid was performed through DOCK3.5 in order to evaluate the impact of representing conformational variability as a structural ensemble on binding pose results.[91] With HIVp, ras p21, retinol binding protein, and uteroglobin as their test cases, the authors compared the capacity of standard grids to energy- and geometry-weighted average grids to reproduce known binding poses and affinities. They found that docking scores were sensitive to the grid spacing and threshold parameters used to define the composite grid. The performance of the geometry-weighted grid was less dependent on the protein conformations that were available than the energy-weighted, to the extent that the geometry-weighted grid placed known ligands for HIVp was within the top 21%, while the energy-weighted grid placed them in the top 33%. This occurred because the geometry-weighted grid ignored the repulsive potentials between flexible atoms, while the energy-weighted grid simply smoothed the repulsive potential, which can result in the retrieval of paradoxical inhibitors.

The multiple copy simultaneous search (MCSS) methodology was the earliest computational approach for mapping binding sites with functional group probes.[92] The authors simultaneously minimized or quenched the probes by MD to the binding site of the hemagglutinin protein to identify favorable minima and found some correspondence between the positions of the minima and the functional groups on the ligand, sialic acid. While the authors noted that their method could account for a limited amount of side chain flexibility by including multiple copies of the side chains, the published work was performed against a rigid structure.

Stultz and Karplus explored the influence of protein flexibility on the results from grid-based MCSS where five different protocols for placing two functional group probes, methanol and methyl ammonium, were used to search the binding surface of HIVp.[93] MCSS calculations were performed for *1*: a minimized crystal structure, *2*: a conformation generated from quenched MD of the initial structure, *3*: an unrestrained structure, *4*: the output of *1* subjected to quenched MD with functional groups restrained and multiple side-chain copies, and *5*: a different rigid crystal structure of HIVp. The authors found that their results with protocols *4* and *5* yielded more favorable interaction energies than the reference protocol *1*. Protocol *4* gave valuable information for ligand design, while the different low-energy minima from protocol *5* supported the idea of performing MCSS against an ensemble of structures and comparing the results, therefore the authors suggested a combination of *4* and *5*. However, one of the limitations of MCSS and other grid-based methods is that they do not account for entropy or solvation during surface mapping, which hampers discrimination between druggable and irrelevant minima. Furthermore, as Schubert and Stultz point out in their review of MCSS, functionality maps are difficult to combine across different structures of a protein.[94]

The incorporation of protein flexibility through docking to an interaction grid that represents the receptor ensemble is a common approach for SBDD, and additional methods are presented in Table 1-5, including those based on AutoDock 3.0[95], Flog[96], GRID/CPCA[97], IFREDA[78], ISCD[98], and sets of consensus structures[99].

Table 1-5: Studies including full receptor flexibility through grid-based ensemble docking.

Method	Target	Flexibility	Results	Caveats	Author
DOCK3.5 composite grids	4 cases	Geometry-averaged and energy-weighted grids	Geometry-averaged grids gave results that were less influenced by input conformation Geometry-weighted average grid yielded RMSD of 0.4-1.6 Å Energy-weighted average found RMSD of 0.4-4.0 Å	Recovers local minima in addition to the binding site Can identify paradoxical inhibitors	Knegtel *et al.*, 1997
Grid-based MCSS	HIVp	Used quenched MD to generate new conformations or mapped a different conformation from the reference	Local optimization of selected probes from MCSS improved interaction energy Protocols 2 and 3 yielded functionality maps that were the most divergent from the reference Protocol 4 yielded the most low-energy minima	Recovers local minima in addition to the binding site Very minor amount of protein flexibility included	Stultz and Karplus, 1999
GRID/ CPCA	3 serine proteases	Docking performed to crystal structures with GRID Used a scaling weight to normalize the probes Consensus PCA	Validated as a potential tool for enhancing ligand selectivity to a specific target Able to predict important cation-π and hydrophobic interactions Highlighted ways	Contour plots Only gives probe position, not orientation	Kastenholz *et al*, 2000

		gave contour plots of MIF	enzyme selectivity could be incorporated in ligand design		
Flog	*L. casei* DHFR, murine COX2	Used snapshots from short MD for averaged grid-based docking Used SIMPLEX optimization of the ligand in the binding site Calculated binding energy with post-docking MD free energy perturbation	Weight-averaged grids performed better than the grid of the average or static crystal structure, selecting 8 leads within the top 1% of the database, compared to 6 leads (crystal) or 7 leads (average) out of the 16 possible leads Average and weight-averaged grids required less than half the CPU hours with optimization than using the crystal structure grid	Automated process, likely to result in incorrect results for certain complexes Could identify paradoxical binders	Broughton 2000
AutoDock 3.0	HIVp (21)	Combined multiple receptor conformations onto interaction energy grid Compared mean, minimum, clamped, and energy-weighted grids Retained structural waters during docking	RMSD in energies was 1.34 kcal/mol for clamped, 1.43 kcal/mol for energy weighted, 7.69 kcal/mol for mean, 6.07 kcal/mol for minimum grids Weight-averaged grids performed best, retrieving the correct conformation 87% of the cases	Limited ability to map conformational shifts Could identify paradoxical binders	Osterberg *et al,* 2002
IFREDA	PK (33), 1000 compound virtual screen	Multiple conformers generated by repeated flexible docking with known ligands Serial-dock used for screening	Average accuracy of 70% in cross-docking with a threshold of 2.5 Å *De novo* structure generation for eight complexes found that most of the lowest energy ligands were within 2 Å of the native state	Unable to map backbone shifts Inadequate solvation model	Cavasotto & Abagyan, 2004
ISCD	Aldose reductase, thrombin, trypsin	Three single confomer grids were joined Repulsive layers between the individual grids	Top 15 ranked ligands had an RMSD to the crystal structure < 1.4 Å for aldose reductase Joined grids for thrombin and trypsin, docking found that 7	Limited capacity for structural ensemble Could identify paradoxical inhibitors	Zentgraf *et al.*, 2006

			of 9 ligands preferentially docked to thrombin (cluster rank 1) The other 2 ligands were selective for trypsin and bound trypsin (cluster rank 1) Computationally inexpensive relative to typical ensemble docking		
Consensus structures	HIV-RT (47)	Normalized B-factors, ligand-induced displacement, and consensus grids represented binding cavity	Identified 2 novel interaction sites Results supported by experimental work by Sweeney *et al.*, 2008 Proposed combining theories of NNRTI activity	Requires available experimental structural information for consensus calculation	van Westen *et al.*, 2010

Structure-based ensembles

The original technique for structure-based ensemble docking was introduced by Carlson *et al*. as the dynamic pharmacophore model[100], later termed the multiple protein structure, MPS, method.[101] Two pharmacophores for HIV integrase were generated, a static model based on a rigid crystal structure and a dynamic model based on MD snapshots initiated from the same crystal structure. The MPS method mapped the dynamic binding surface of the protein with common functional groups by running a series of multi-unit search for interacting conformers (MUSIC) simulations (Figure 1-3), wherein hundreds of probes were used to flood the protein surface and then minimized by Monte Carlo (MC) sampling. During MUSIC simulations, all of the probe-probe interactions were ignored, which allowed probes to interact with the rugged binding potential of the protein and cluster into hot spot positions along the protein surface. These cluster sites were used to develop a pharmacophore for virtual screening, and where the rigid model was unable to locate any of the test case inhibitors, the dynamic pharmacophore model not only identified known inhibitors, but it also predicted new inhibitors that were confirmed by experiment. The MPS method is one of the few ensemble-based methods that has been successfully used to find novel leads with demonstrated activity.[102,103]

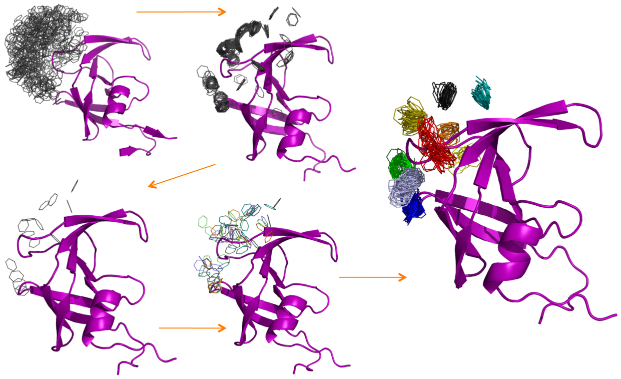


Figure 1-3: MUSIC simulations of benzene probes (gray) to the surface of a HIVp monomer (purple). Five hundred probes were flooded onto the protein surface, minimized, and then clustered together. Parent probes were identified for each monomer conformation, the HIVp monomers were aligned through a weighted-RMSD function, then the parent probes were clustered together and retained when at least half of the protein conformations contained a probe at that site. This resulted in low-energy probe clusters (colored red through purple) of benzene that could be combined with ethane and methanol clusters to develop a pharmacophore model.

McCammon and co-authors developed the relaxed complex scheme (RCS) as a technique for incorporating receptor flexibility prior to docking.[104-106] Initially the method was developed to generate conformational ensembles of the *apo* state through MD, which were then used in AutoDock to screen compound mini-libraries and score the receptor-ligand complexes. The 2003 study by Lin *et al*. examined FKBP-12 by RCS and implemented a final refinement step with MM/PBSA[107] to yield final results. Further updates to this method have included extension to virtual screening and reduction of the conformational ensemble to a representative configurational set[108]. RCS was successfully used to identify cryptic binding pockets and/or lead inhibitors of HIVp[109], avian influenza neuraminidase[110], acetylcholine binding protein[111], DNA polymerase β[112], MDM2/MDMX[113], and cruzain[114].

RosettaLigand was introduced as an extension of the protein-protein docking program, RosettaDock.[115] In the 2006 paper, side chains were repacked during docking through the implementation of a rotamer library while the vdW repulsion term was softened and ligand conformations were perturbed with MC. Their benchmark set included 100 complexes for native and 10 for cross-docking. They found a success rate of 80% for native docking and 70% for cross-docking. The initial reliance of RosettaLigand on side-chain motion as the only representation of protein flexibility was altered by the inclusion of full receptor flexibility obtained from multiple crystal structures.[116,117] Although RosettaLigand has generated good results, with protein flexibility included it requires 40-80 processor hours to generate a bound pose for the receptor-ligand complex and therefore is too time-consuming for use with large compound libraries.

To determine the best process for developing a structural ensemble, Rueda *et al.* used ICM to examine the same set of 1068 conformations from 99 pharmaceutically-relevant proteins that was used in the 4D docking study.[118] The authors judged performance in virtual screening based on the "area under the ROC plot curve" (AUC) and found that holo conformations yielded improved AUC values compared to apo conformations. It was interesting that the authors found that ensembles of holo plus apo conformations did not significantly improve results compared to holo-only ensembles. In cases without a holo structure, the authors recommended the use of an apo structure with the largest pocket volume. The authors proposed the use of a ligand-guided approach to find the optimal protein conformation, but stated that docking to an ensemble would be an acceptable substitute in the absence of ligand activity data. Several groups have found that there tends to be a single receptor conformation that grants the best performance in docking studies.[118,119]

Normal mode analysis (NMA) has been frequently used as a tool for examining protein flexibility relevant to ligand binding.[74,86,87,120-125] It is important to note that low-frequency modes identify large-scale motions, while high-frequency modes identify small-scale motion. Rueda *et al.* showed that they could use NMA on elastic network models for all heavy atoms of the receptor in order to derive an ensemble of protein conformations and improve results from rigid-pro cross-docking in ICM.[126] Fourteen proteins, each with two structures and their cognate ligands, were used as the benchmark test set. Optimal results involved the generation of 100 different conformations from NMA on all heavy-atoms within 10Å of the ligand-binding site. Larger conformational ensembles, such as those with 200 members, negatively affected false positive rates. Another NMA-based method for conformational selection was based on cyclin dependent kinase-2 (CDK-2), because it had a sufficient amount of available experimental data on ligand binding for validation of the NMA protocol.[127] Unlike most other NMA-based methods, the authors included all of the protein atoms in the calculation in order to better represent conformational changes and found that too many protein conformations negatively impacted docking results through the recovery of false positives.

Of the variety of techniques that have been developed to improve binding predictions for protein-ligand complexes, the use of a conformational ensemble is perhaps the most common approach. Representative cases and novel methods based on structure-based ensemble docking are listed in Table 1-6 and include modifications to DOCK[128], docking to a large set of conformations[119,129], NMA and MD-based approaches[125,130-132], structure prediction[133], the ensemble reduction method[134], F-DycoBlock[135], FIRST[136], FIRST/ROCK/SLIDE[137], Fleksy[138], FlexX-Ensemble[79,139], FlipDOCK[140,141], Flo98[142], MCSA-PCR[143], MD+LigBuilder[144,145], PhE/SVM[146-148], QSiCR[149,150], a reduced receptor ensemble[151], Skelgen[152], and Surflex[153-155].

Table 1-6: Studies including full receptor flexibility through a structural ensemble for docking.

Method	Target	Flexibility	Results	Caveats	Author
FlexX-Ensemble	105 cases	Averaged highly conserved regions Retained orientations of flexible regions as a set	FlexX-Ensemble yielded 66.7% success (40 hits) compared to 63.3% with FlexX within an RMSD of < 2 Å for the first 10 solutions	May find ligands that are not compatible with the "real" ensemble May not identify the lowest RMSD complex	Claussen *et al.*, 2001
MCSA-PCR	Major histocompatibility complex	Ligand grown into binding pocket with each SA cycle Generated pseudo-crystallographic density map	Identified optimal ligand position and important waters based on simulation density Side-chain RMSD of VSV8 peptides from explicit simulation was 2.15 Å for N/C termini-restraints, 2.61 Å for Cα restraints, and 2.41 Å for backbone restraints	High computational effort, not suitable for VS	Ota & Agard, 2001
F-Dyco Block	HIVp, COX-2	Split ligands into smaller building blocks Performed locally enhanced sampling and	Recovered the ligand 3/3 results for HIVp when full flexibility was allowed (except secondary-structure hydrogen bonds),	Simple clustering Dependent on threshold value and cluster	Zhu *et al.*, 2001

		multiple-copy MD with a dynamic connecting algorithm Tested grid approximation & 4 ways to manage receptor flexibility	compared to only recovering the correct complex for 1 ligand when backbone-flexibility was restrained Recovered the correct ligand in 7/76 results for COX-2	position	
MD+ Lig Builder	Alanine Racemase	Dynamic pharmacophore modeling similar to MPS, but can simultaneously search with multiple probes LigBuilder mapped surface properties of each conformation (11 total) after MD	Dynamic pharmacophore model identified 34 hits out of the set of 43 known binders, compared to the 27 identified by the static model	Very brief phase of MD for conformation generation	Mustata & Briggs, 2002
RCS	FKBP-12	Conformations generated by MD Rapid docking in AutoDock Refinement of high scoring complexes with MM/PBSA	Correctly ranked the crystallographic pose as the highest rank with MM/PBSA for all three cases	Can be time-consuming Short MD simulation (2ns) used for conformer generation	Lin *et al.*, 2003
Modified DOCK	T4 lysozyme mutant	Calculated interaction energy between ligand and flexible regions independently Used results to create average receptor representation for VS	Found 18 new hits Conformation's internal energy important for final ranking Conformational ensemble retrieved approximately 79% of ligands in the top 1.5% of the database, compared to the holo conformation retrieving 77% and the apo conformation retrieving 54% of the ligands	Conformational changes caused by ligand binding not fully predicted by method	Wei *et al.*, 2004
FIRST/ ROCK/ SLIDE	CypA, estrogen receptor	Analyzed flexibility potential (FIRST) and sampled rotational bonds to generate receptor ensemble (ROCK) SLIDE accounted	Representative ensemble similar to NMR structures Captured important hydrogen bonds present in the receptor-ligand complex	Compared score distributions, no individual assessment of relative likelihood/ energy for generated	Zavodszky *et al.*, 2004

		for side-chain flexibility and solvation		conformers	
Docking to a Structural Ensemble	Hsp90 (57), CDK2 (34)	Compared full-flex to rigid-pro	For Hsp90, the use of pharmacophoric restraints improved docking results, with multiple cavity docking having 86% average success compared to single cavity docking having 57% average success For CDK2, single best cavity gave 64% success, 6 best gave 94%, entire ensemble gave 76%	Incorrectly assumed negligible difference in conformational free energy Use of multiple cavities for virtual screening was limited by bad poses with good scores	Barril & Morley, 2005
Low-frequency Cα NMA	cAPK Kinase	NMA generated alternative conformations based on relevance to loop plasticity Minimized side chains while bound to non-native ligand	Improved docking and enrichment scores relative to rigid-pro RMSD of the top ranked pose to the native pose was 0.6-4.0 Å, with one outlier at 7.0 Å for docking to the NMA conformer and was 0.4-10.2 Å for docking to the crystal structure	Used known binders, potentially biasing available lead space	Cavasotto *et al.*, 2005
Elastic Network Model	6 protein-ligand cases	Low frequency NMA to optimize docked receptor-ligand complexes and identify new conformational space Used eigenvectors of *apo* protein	Improved the global coordinate RMSD by ≤ 3.2 Å to the predicted complex to the experimental pose Refinement with restraints on mode amplitude performed much better than MD or NM minimization, or unrestrained scanning	Most successful for systems where flexibility can be represented well in only a few modes	Lindahl & Delarue, 2005
Skelgen	Estrogen receptor-α (7)	*De novo* design through serial-dock Ligands generated from fragment design Two different pharmacophoric	Identified 4 novel lead compounds Best had IC_{50}=340nm	Dependent on pharmacophoric constraints Time and material intensive	Firth-Clark *et al.*, 2006

		constraints applied Selected top 25 complexes for each constraint set, yielding 350 compounds for testing			
Explicit solvent MD/ Auto Dock	p38 MAPK (5000 MD snapshots)	Performed 3 simulations of 60-ns each with MD Used snapshots from MD simulation for docking with AutoDock	Found 2 novel conformations of DFG motif Successfully docked 5 known inhibitors Conformers from simulation identified cryptic binding sites	Extensive MD simulation (390ns total) render this technique impractical for some SBDD applications	Frembgen-Kesner & Elcock, 2006
Flo98	*C. hominis* DHFR, *T. gondii* DHFR	SA of active site MC search for optimal bound pose Averaged energy of 25 lowest energy protein-ligand complexes Developed homology model for TgDHFR	Calculated binding affinity was 72.9% correlated to experiment (ChDHFR) Identified alternate binding site 50.2% correlation between docking and activity (TgDHFR)	Enabled limited flexibility of binding site residues only during global MC docking	Popov *et al.*, 2007
PhE/ SVM	Ether-à-go-go Related Gene	Receptor flexibility through a pharmacophore ensemble from 26 training set & 13 test set compounds	Potential to enable protein plasticity even when structural information is inadequate for other methods	Can not identify large structural shifts	Leong, 2007
Predictive Modeling	Androgen Receptor	Predicted receptor conformation from backbone conformation Optimized side chains Repeated selection ~15 times to discriminate agonist and antagonist bind	Potentially identified drug repurposing candidates for antiandrogenic activity, based on cross-docking and competitive binding assays	Time intensive Requires *a priori* knowledge & similar protein structures for initial predictive modeling	Bisson *et al.* 2007
Fleksy	35 cases	Generated ensemble from soft potential backbone-dependent rotamer exploration Used FlexX-Ensemble with soft docking Flexible	Successful flexible cross-docking for 78% of the cases, compared to 44% success for FlexX (rigid-pro)	Cannot handle large changes in backbone conformation No consideration of solvent effects	Nabuurs *et al*, 2007

		optimization of complex with Yasara			
Flip DOCK	HIVp, Protein Kinase A	Flexibility Tree data structure represented conformational subspace Enabled full-flex with AutoDock	Divide-and-conquer GA yielded best results FlipDOCK outperformed rigid-pro 93.5% to 72% in cross-docking	Each degree of freedom must be selected by hand	Zhao & Sanner, 2007 & 2008
QSiCR	CDK2, p38 MAPK	Used known binders to generate protein flexibility Built conformers from MACCS keys and topological details	Average $R^2 = 0.73$ for active site size and distances of CDK2 Average cross-validated $R^2 = 0.60$ for predicted backbone conformational changes of p38 MAPK	Does not fully represent available interaction space Highly dependent on ligand training set	Subramanian *et al*. 2006 & 2008
Reduced Receptor Ensemble	Avian flu neuramindase	Similar to MPS Generated MD ensemble Flooded ensemble with probes (CS-Map), created pharmacophore	Predicted novel hot spots potentially relevant for *de novo* ligand design	No experimental support of sites or energy ranking of fragment clusters	Landon *et al.*, 2008
Ensemble Reduction Method	DHFR	Generated MD ensemble Used relative difference distance to define optimal conformations	Conservation of essential distances between binding residues best for selecting representative ensemble	Maintaining conserved core distances may limit exploration of important large-scale shifts	Bolstad & Anderson, 2009
Rosetta Ligand	85 cases	Included receptor flexibility through multiple conformations	Average RMSD of pose closest to crystal conformation (majority ranked 1) ranged from 1.43-2.44 Å for flexible backbone docking compared to 2.39-2.49 Å for rigid docking to JNK3 kinase and urokinase	Not as easily applicable to virtual screening	Davis *et al.*, 2009
EN-NMA	14 cases	Used NMA to generate 100 distinct conformers for heavy atoms with 10 Å of the binding site	From NMA-based conformers, able to select 20 of 28 cross-docking cases amount top 5	The larger conformational ensemble of 200 structures increased false positive rates	Rueda *et al.*, 2009
Implicit	6 Protein	Docked via 40	Best results: distance-	One solvation	Huang &

Solvent MD	Kinases & Phosphat-ases	simulated annealing cycling simulations for each protein-peptide complex Restrained protein Cα	dependent dielectric solvation model docking & GB molecular volume scoring, based on total system energy	model alone (GB molecular volume versus distance-dependent dielectric) not always correct	Wong, 2009
Surflex	85 cognate cases, 8 cross-docking cases	Updated to allow structural ensembles and post-docking refinement Created idealized ligand to model active site Ligand built in through incremental construction/combination	61% success after docking, 67% after refinement and rescoring, 75% after using the best of 2 pose families	Moderate computational expense For some cases, ligand sub-fragments guiding docking could have a negative impact	Jain, 2009
Docking to a Structural Ensemble	BACE, cAbl	Examined enrichment from ensemble docking compared to rigid-pro for choosing the inhibitors over mimetic decoys Used five different strategies for combining receptor structures into an ensemble Docking performed in Glide	Area-under-curve (AUC) for rigid BACE structures ranged from 0.688 to 0.778 Construction of an ensemble from receptors which demonstrated optimal rigid-pro success against different chemotypes yielded the best performing ensemble	Technique was based on improving enrichment against known inhibitors rather than exploring conformational space Did not account for solvation or energy of the receptor	Craig, *et al*. 2010
ICM	99 cases	Studied the impact of using holo versus apo protein conformations for virtual screening based on AUC scores	Found that there is one receptor conformation within the ensemble that gives the optimal performance Holo conformations outperformed apo conformations in most cases	The authors noted that ensemble docking performed better than a single average structure, with 0.88 ± 0.18 mean AUC versus 0.78 ± 0.22, but those ranges overlap significantly	Rueda *et al*., 2010
All atom NMA	CDK2	Used all of the protein atoms for conformational selection Pursued relevant	Over all cases, successful docking occurred for 54.3% (holo crystal structure), 58.1%	As seen in many of the ensemble-based methods, care	Sperandio *et al*., 2010

		structures from the 20-25 lowest modes Required a deviation of at least 1 Å, but no more than 8 Å from the minimized crystal structure for new conformations	(apo crystal structure) 42.8% (minimized apo crystal struct-ure) 70.4% (best NMA structure) 55.7% (second-best NMA structure), and 55.2% (third best NMA structure)	was required in conformer selection to avoid discovery of too many false positives NMA is not applicable to small local motions or large domain shifts	
Explicit solvent MD/ Glide	Reverse transcriptase (10,000), W191G (7,500)	Generated conformational ensemble from MD of holo and apo crystal structures Docked ligand/decoy set to conformers in Glide	Found that MD could be used to move a conformation into a predictive range for docking No correlation between MD run and AUC Identified a correlation between the average predictive power and the average flexibility of the binding site, such that highly flexible sites had less utility for docking	A broadly-applicable protocol for the application of multiple receptor structures for use in docking is still a distant goal No single feature can be used to pick out conformations May require extensive knowledge of the system	Nichols *et al.*, 2011

1.6 *De Novo* Approaches through Fragment-Based Drug Design

Several interesting experimental methods that naturally account for protein flexibility have greatly influenced computational SBDD. The burgeoning field of FBDD has allowed for the identification of low-molecular weight binders, which allow for the identification of difficult binding sites.[156] First introduced by the multiple solvent crystal structure (MSCS)[157] and SAR-by-NMR[158] techniques, FBDD can be performed via NMR or x-ray crystallography, and provides a different set of hits from HTS experiments. Screening proteins with a fragment collection both verifies the druggability of a target system and identifies fragments that can be linked and optimized to develop a viable lead compound.[159] Importantly, hits from FBDD have been experimentally validated and identified novel compounds shown to work in the clinic.[160,161]

The MSCS technique was originally introduced as a crystallography tool for the characterization of the binding potential of pancreatic elastase. Other groups have also used MSCS to probe ligand interactions in thermolysin[162,163], subtilisin[164-166], RNase A[167,168], p53 core[169], lysozyme[170], and H-Ras[171]. During crystallization, the water and organic solvent acted as probes of binding affinity because the organic solvent, which was chosen to represent a common functional group, displaces bound waters only at locations with favorable affinity for the particular interaction type presented by the protein. MSCS can be performed with a variety of probes, thus enabling the results from the protein crystallography experiment to be superimposed in order to identify regions of consensus binding.

When Joseph-McCarthy *et al.* compared their results from studies with the original MCSS (non-grid-based) to results from MSCS, they found that for traditional MCSS it was important to consider occupancy at a first site in order to accurately predict occupancy at a second binding site.[172] English *et al.* compared the use of MSCS, GRID, and MCSS through the exploration of the binding surface of thermolysin with acetone, acetonitrile, isopropanol, and phenol acting as functional group probes. The authors found that in comparison with MSCS, MCSS and GRID overestimated electrostatic interactions, retrieved an overabundance of local energy minima, and could not provide detailed descriptions of the interactions between the protein and probes.[163]

Although both SAR-by-NMR and MSCS allow for a larger amount of protein flexibility then current methods for docking do, they are also both material and time-intensive. Furthermore, as discussed by English *et al.*, MSCS can be quite slow and probe choice is limited by the fragility of protein crystals.[163] Computational solvent mapping is an complementary approach to experimental fragment-binding studies, however most computational approaches do not account for either the impact of protein flexibility or proper solvation effects, which can lead of poor mapping results. Over the past two years, several groups have developed new methodology for including flexibility and solvent competition in computational fragment mapping. While the results of any fragment-based study can be translated into a consensus pharmacophore model and used in virtual screening applications, the primary objective of MSCS and computational solvent mapping has been the correct identification of potential binding sites.

Locus Pharmaceuticals developed a method based on a grand-canonical (GC) MC simulation for pharmacophore development.[173-175] Both studies from Clark *et al.* were performed against a static structure. Their later study concentrated on deriving accurate ΔG_{bind}, while their earlier study used 10 simulated annealing runs of GC-MC to explore the protein surface of thermolysin with 2 probes and of T4 lysozyme with 14 probes. Based on the simulated annealing runs, GB/SA was used to predict binding affinities and Clark *et al.* found that their results retrieved some of the hot spots observed in published MSCS data. Moore discussed the use of torsion-space dynamics to generate protein flexibility in combination and general outcomes from the use of a 500-member fragment library to search target surfaces with GC-MC, however no specific details were given.

Vadja and co-workers developed a fast algorithm for searching protein surfaces with small organic probes that was based on a fast Fourier Transform correlation approach (FTMAP).[176] In a manner akin to MPS, billions of solvent probes are minimized to the protein surface and then clustered based on a simple greedy algorithm. The probe clusters demonstrate the position and orientation of proposed hot spots along the target, where the largest cluster is defined as the maximal hot spot location; the second largest is the second most-important site, and etc. Although FTMAP is another method for computational solvent mapping that does not inherently include protein flexibility, receptor dynamics could be modeled through serial searches over multiple conformations.

A novel method for probe mapping incorporated solvent competition and protein flexibility through MD.[177] Based on MSCS, the authors used isopropyl alcohol (IPA) and water together to perform a single MD simulation over 16 nanoseconds (ns). The use of binary solvent MD inherently accounted for solvation effects and protein flexibility. This method was applied to three proteins with experimental MSCS results and five pharmaceutically relevant receptors that have not been studied by MSCS; simulation probe occupancy was used to differentiate binding sites and predict ΔG_{bind}. This method was much more computationally demanding than other computational solvent-mapping techniques like FTMAP,[176] but may be more competent at distinguishing between druggable and nondruggable sites because of the use of flexibility data. Yang and Wang used this same solvent-mapping technique to study hot spot mapping against thermolysin together with a double-decoupling method which provided a more rigorously calculation of ΔG_{bind} at potential hot spots.[178] It is important to note that the analysis of probe occupancy to locate binding sites necessarily assumes that Boltzmann sampling occurred during simulation, therefore careful selection of the trajectory length required for convergence is exceedingly important but frequently neglected in method development.

Guvench and MacKerell published a similar method, Site Identification by Ligand Competitive Saturation (SILCS), where small molecule probes were used to examine the binding surface.[179] Their technique used all-atom MD simulations of protein in a box of propane, benzene, and water as the explicit solvent probes. Ten independent simulations over five nanoseconds were generated for analysis and the mapping results were represented on a 1Å x 1Å x 1Å grid of volume occupancy. SILCS was capable of reproducing known binding interactions for the test case, BCL-6 oncoprotein, and a follow-up study further validated the approach on seven proteins from five different protein families.[180] In their follow up study, MacKerell and co-workers extended the simulation time to 10 runs of 20 ns each and scored their occupancy results based on a normalized calculation of the observed:expected occupancy, which they termed grid free energy (GFE). When compared to the positions of known ligands, their GFE was able to select for the crystallographic pose of the bound ligand in several protein families. However, in the only available figure of an entire protein surface, SILCS is clearly shown to preferentially map many irrelevant minima before identifying the binding site.

We have developed a computational method that achieves hot-spot-mapping results that are similar to available experimental data; Mixed-solvent Molecular Dynamics (MixMD). Our multiple protein structure (MPS) method[101] for pharmacophore development demonstrated success in mapping protein systems for drug design[102,103] and MixMD complements MPS while simultaneously allowing protein flexibility and probe competition with water. MixMD was inspired by the MSCS technique, but incorporates more explicit conformational sampling. Since MSCS results identified specific electron density that can be attributed to the placement of organic solvent along the protein surface, we specifically compare grids of solvent occupancy from simulation to that of experimental electron density. Using many, short MD simulations of protein in 50% weight/weight (w/w) mixtures of acetonitrile and water, we successful validated MixMD based on the available MSCS structures.[181] Similar results were obtained with isopropanol (in preparation).

1.7 Limitations of Structure-Based Drug Design

Most published methods for flexible protein-ligand docking are based on limited sets for benchmarking results. The use of a large test set for methods development is crucial for demonstrating performance across a range of different targets. Furthermore, docking to a flexible protein is more resource and time-intensive than docking to a rigid receptor and in many cases ensemble docking is unrealistic for screening huge compound libraries. Through the careful development of methods for docking and scoring, the computational requirements for accurate simulation may be sufficiently lessened to allow for more widespread implementation of flexibility.

The use of multiple conformations for docking is limited in that several protein conformations may decrease the selectivity of lead compounds by increasing the false positive rate. Using combinations of features from different conformations may also lead to the creation of a ligand with high affinity for an average receptor structure that is not experimentally accessible, a so-called “paradoxical inhibitor”. As a result, there is the potential for introduction of bias through user intervention in many of these methods for flexible docking.

Due to the rugged landscape of most proteins, not every conformation that is included in a low-energy ensemble will adequately represent its true binding potential. This has led many recent studies to dedicate their focus to identifying the optimal method for selection of only the relevant protein conformations. For this process of conformer selection and weighting to succeed, it is crucial that the internal energy of the individual protein conformations be included in the scoring process (though this is more difficult to properly calculate). Furthermore, it is necessary that the scoring functions used in fully flexible procedures be as accurate as possible in order to determine the most physically-realistic results.

Experimental FBDD approaches like SAR-by-NMR and MSCS are time-consuming, costly, and dependent upon the receptor size and its potential for crystallization. On the other hand, computational FBDD approaches can be just as time-consuming and are heavily dependent on proper parameterization and analysis. The trend towards estimating binding affinity from computational solvent grids has shown only partial success when compared with experimental affinities. Computational FBDD shows great promise as a tool for including flexibility and probe-water competition explicitly in a "docking" experiment; however, implementation of these methods requires careful consideration of the underlying chemistry to ensure reliable results.

1.8 HIV-1 Protease as a Computational Model

Since infection with Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome (HIV/AIDS) emerged as a global health threat in the early 1980s[182], it has become one of the most devastating diseases worldwide.[183] The most recent estimates (2009) from WHO/UNAIDS reported that 33.3 million people were living with HIV/AIDS, 2.6 million were newly infected, and 1.8 million had died due to AIDS.[184] Infection with HIV/AIDS results in immune system failure, eventually resulting in death caused by a secondary illness. Although there is no cure, HIV/AIDS is usually treated by a combination of highly active antiretroviral therapies (HAART). These drugs target various stages of the viral life cycle, including viral enzymes like HIV integrase, protease, and reverse transcriptase. While there are a number of therapeutics for people living with HIV, many of these therapies are becoming less effective because of drug-resistant viral mutations.

HIVp is widely recognized as an important pharmaceutical target for the treatment of HIV, the lentivirus that causes AIDS. HIVp is critical to the continued viral life-cycle; it cleaves the precursor proteins gag and gag-pol, yielding mature HIV proteins including integrase, reverse transcriptase, protease, and the viral matrix, capsid, and nucleoproteins.[185,186] If HIVp is inactivated, the virus cannot undergo maturation and does not become infectious. Despite the existence of ten protease inhibitors (PIs) in clinical use[187], there is little variety in their mechanism of action. Typical PIs are pseudosymmetric and compete with the substrate for binding at the base of the active site. Unfortunately, even the most recent PIs suffer from taxing side-effects and poor pharmacokinetic properties.[188,189] In addition, HIVp is highly mutagenic, particularly in the region surrounding the active site, resulting in the rapid development of drug resistance.[190-192] Consequently, finding new PIs that can bypass or treat multi-drug resistance (MDR) is essential.

1.5.1 HIVp structure

HIVp is an attractive target for SBDD due to the wealth of structural information available from x-ray crystallography, NMR, and MD. HIVp is a C2 symmetric homodimer aspartyl protease with two catalytic aspartic acids, residues 25/25', at the base of the binding site. The active site is loosely covered by two glycine-rich α-hairpins, or "flaps" (Figure 1-4).[193-195] The conformational behavior of the flap region of HIVp has been extensively studied in the last few years, as reviewed by Hornak and Simmerling.[196] These flaps are highly mobile and can occupy a range of conformations. NMR and x-ray crystallography studies of HIVp demonstrate that the protease can occupy three different conformational states: open, semi-open, and closed.[197-199] In the apo form, the thermodynamically-favored state is thought to be the semi-open conformation, in which the flaps are loosely positioned over the active site cavity restricting ligand entry.[200-202] Upon binding to a ligand, the flaps close down over the active site, shifting their position by 5-7 Å.[203] Differences in flap mobility are a potential contribution to the mechanism of MDR for certain HIVp mutants.[204]

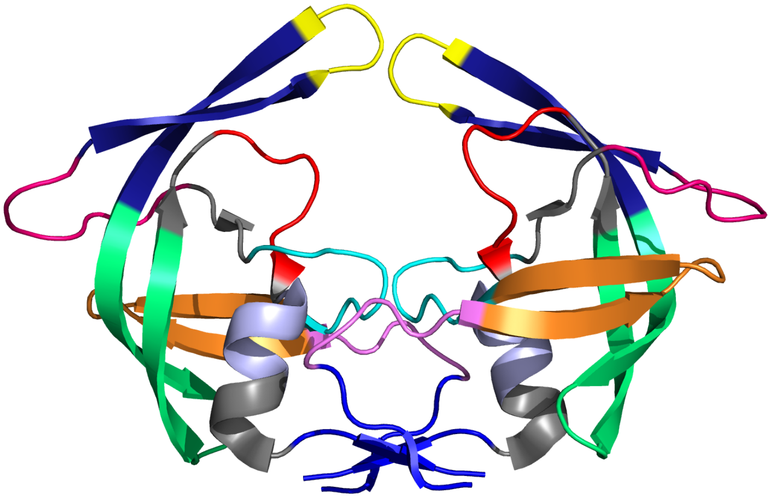


Figure 1-4: The typical semi-open flap conformation for HIVp with each distinct region of the protein shown in a different color. The catalytic aspartic acids (Asp25/25') are shown in stick form. The flaps are

shown in dark blue, residues 43-58; the flap tips in yellow, residues 49-52; the fulcrum in orange, residues 11-22; the cantilever in lime green, residues 59-72; the active site in cyan, residues 23-30; the dimer interface in blue, residues 1-5 and 95-99; the elbow in hot pink, residues 35-42; the helix in light blue, residues 86-90; the 80s loop or wall turn in red, residues 79-84; and the nose in violet, residues 6-10.

1.5.2 HIVp inhibition

The flexibility of the flaps and their connection to protease conformation and activity has been examined at length.[196] A coarse-grained (CG) MD study by Trylska *et al*. found that the substrate waits in proximity to the binding site for the flaps to open long enough so that the substrate can enter the active site.[205] Another CG dynamics simulation demonstrated that when the flaps are semi-open or almost closed the small cyclic urea inhibitor XK263 can enter the active site from the side.[206] However, the same study also showed that a peptide substrate must sample the surface of the protein until it encounters an opening event. It is generally held that most ligands, particularly the peptide substrate, can only access the active site through the open conformation.[205-207] Therefore, it is possible that a new class of PIs could be developed to treat HIV through the restriction of flap mobility. Most of the mutations seen in HIVp occur around the binding site,[208] so by targeting flexibility of the protease with an allosteric inhibitor, existing drug resistance could be avoided. There are three different allosteric sites with the potential to control HIVp allosterically: the elbow, the eye[103], and the dimer interface. By controlling motion at these sites, the position of the protease flaps may be altered, thus affecting access of the peptide substrate to the binding site. Development of a new strategy to treat HIV infection by targeting HIVp would lead to the first new class of inhibitors since the introduction of dimerization inhibitors more than twenty years ago.[209,210]

1.9 Summary

The major areas addressed in this thesis include protein flexibility, allosteric control, and hot-spot prediction.

Chapters 2 and 3 focus on computational probe mapping through MD simulations of proteins in binary solvent. Probe mapping is a commonly used computational technique for identifying potential binding pockets along a protein surface, but it typically relies on gas phase calculations against a rigid protein. Chapter 2 discusses the development and validation of a new method, mixed-solvent molecular dynamics (MixMD), that simultaneously allows for the probe's competition with water and protein flexibility. Chapter 3 highlights the generalization of the MixMD approach to a broad range of pharmaceutically relevant protein targets. Additionally, the correct implementation of solvent parameters as well as the impact of simulation time on convergence is examined.

Chapters 4 and 5 consider the use of explicit-solvent molecular dynamics (MD) to examine the potential for allosteric regulation between small molecules and HIVp. Chapter 4 presents an in-depth study of the interactions between HIVp and symmetric pyrrolidine inhibitors that complement the semi-open conformation of HIVp. This work dove-tails with previous studies in the Carlson Lab suggesting the eye site as a possible alternative to traditional competitive inhibitors. Chapter 5 discusses my initial studies in the Carlson Lab into the impact of small molecules at the elbow region. The work detailed in Appendix A is an extension of these HIVp-elbow studies, where we employed positional restraints and small beads to examine the structural properties of a potential compound that could target HIVp as an allosteric regulator. While the studies presented in Chapters 2-5 represent work that has been published or is in preparation for publication, the study in Appendix A was not published.

Chapter 2

Full Protein Flexibility is Essential for Proper Hot-Spot Mapping

2.1 Introduction

We have developed a new protocol for using Mixed Solvent MD (MixMD) to identify important hot spots. Our multiple protein structure (MPS) method[211-213] for creating binding-site pharmacophore models based on conformational ensembles has demonstrated success in mapping protein systems for drug design.[102,103] MixMD expands the MPS concept while combining ideas from MSCS to simultaneously allow protein flexibility and competition between probes and water.

Several similar efforts have incorporated MSCS concepts into a computational method, but each has notable limitations. FTMap[176] is modeled after MSCS, but while it can be used with ensembles like MPS[151], neither ligand nor on-the-fly protein flexibility is used during probe mapping. A recent study from Seco *et al*. utilized MD with mixed water and isopropanol to detect binding sites and predict potential druggability.[177] However, the method was unable to reproduce many known binding sites. SILCS is a mapping method that incorporates a ternary solvent system (benzene, isopropanol, and water) with MD to map sites.[179,180] Therefore, these methods are in their infancy and require significant development to provide a robust tool for SBDD. Here, we present initial findings based on our MixMD protocol that should have significant impact on solvent mapping approaches.

Hen egg-white lysozyme (HEWL) is a canonical model system that allows for appropriate testing and validation of MixMD to identify hot spots. A MSCS of HEWL was produced using acetonitrile (CCN) as the organic solvent.[170] The high quality electron density available for this structure allows for an accurate assessment of MixMD data. Below, we demonstrate how occupancy grids for both the probe and water can be directly compared to electron density.

2.2 Methods

The starting structure of HEWL in CCN and water (2LYO)[170] was obtained from the PDB[214]. We performed all-atom MD simulations of the HEWL protein in the presence of multiple solvents using standard procedures for *sander* in AMBER10[10] at 300K. Pre-equilibrated solvent boxes with an even distribution of 50% weight/weight (w/w) CCN and water were used. Simulation set-up was completed in tLeAP using the ff99SB force field[7], TIP3P water[215], neutralizing ions, a 10 Å vdw cutoff, and CCN parameters from Grabuleda *et al*[216]. A time step of 2 fs was implemented, temperature was controlled through an Anderson thermostat[217], and SHAKE was applied. Three different protocols for protein flexibility were evaluated for proper sampling and convergence: all-atom restrained, backbone restrained, and fully flexible HEWL. Five independent simulations with 10 ns of production time each were performed for every system, initiated from the same solvent configuration. Though it might enhance sampling to have alternate starting locations for solvent in each simulation, it would make it more difficult for us to properly evaluate convergence in the simulations.

The stability of the protein core was verified on the basis of low backbone RMSD over the course of the trajectory, as measured with *ptraj*. The presence of mixed solvent did not destabilize the unrestrained protein on the timescale examined.

Five independent trajectories of 10 ns each were performed, but for the analysis we combined the last 2 ns of each simulation to provide the most "equilibrated" 10 ns of the available 50 ns of simulation time. To calculate the solvent grids, the central C2 atom of acetonitrile and the oxygen atoms for water were binned into 0.5 Å × 0.5 Å × 0.5 Å volume elements, which spanned the entire box. The electron density from the initial crystal structure and the grids from MD simulation were compared in Chimera[218] to examine the ability of MixMD to recover known hydration sites and binding sites of CCN.

The CCP4i suite[219] was used to produce electron density grids for the solvent probe density in the original crystallographic study. The *ptraj* function in AMBERTOOLS1.2 was used to generate grids of solvent density from simulation data. Prior to grid output, each trajectory was centered, imaged, and aligned to the protein core backbone from the 2LYO crystal structure.

Convergence in the simulations was examined with several measures. The first test compared the grids between the five independent simulations (the last 2 ns from each that were used to make the 10 ns of sampling noted above were calculated separately). Difference grids were calculated to determine differences in the density data. The second test compared the position of maximal occupancies for each of the five separate simulations. The 99% maximum occupancy calculation yields the grid points with the highest occupied solvent density.

Third, the influence of the protein should be minimal at the box edges. Therefore, the ratio of probe to water on the distal points of the occupancy grids should simply approach the ratio of the total number of probes to water in the simulation. Occupancy grids within 5 Å of each box boundary were examined. The ratio of probe occupancy to water occupancy was compared to the ratio of probes and water in the initial system set up. Occupancy within 5 Å from the box boundary for each plane was used to calculate this ratio. Correspondence between the two different ratios implies adequate sampling has occurred (values near 1.0). The calculation of this ratio was performed through a python script in Chimera.

2.3 Results and Discussion

The positions of the solvent from the *sander* trajectories were converted into occupancy grids using *ptraj*. In this way, we were able to directly compare our solvent "density" results to electron density data obtained in the crystallography study. This allowed for an equivalent comparison of solvent positions during simulation with solvent occupancy from crystal studies, which is a more even assessment than simply using the solvent coordinates given. (In the figures below, crystallographic coordinates for CCN and water are often used in place of electron density to avoid the confusion of overlaying many grids.) Technically, the equivalent data to crystallographic density would be an occupancy grid based on all atoms of the simulation (protein, water, CCN, and counter ions), but we have made the simplification of examining only solvent-occupancy grids.

Our initial simulation used mobile solvent and a fixed protein; we aimed to establish a minimum sampling time required for the solvent to reproduce the MSCS results. We assumed that the mapping would identify the position for CCN and that longer sampling times would be required as more flexibility was allowed for the protein. Instead, we were surprised to find that our simulation of the rigid protein converged to multiple, trivial minima (Figure 2-1). Though the CCN hot spot in the crystal structure was mapped with weak occupancy, it was equal to and less than many incorrect sites. When we added side-chain flexibility (backbone still fixed), a variety of incorrect sites were again located, but the correct location was not. Only when full protein flexibility was allowed was the correct location for the CCN hot spot found and the trivial minima eliminated.

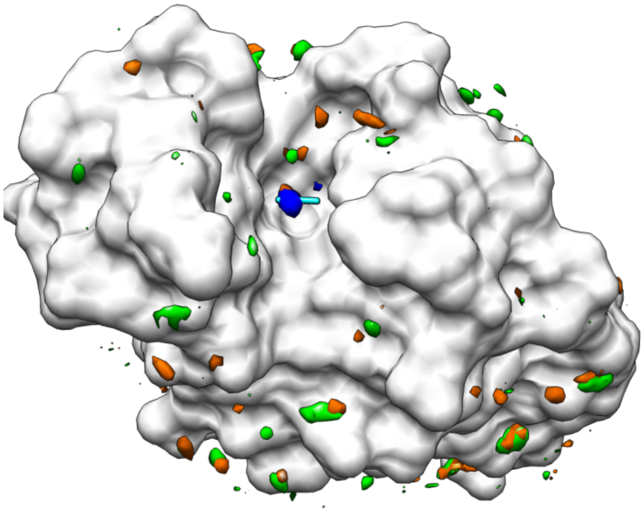


Figure 2-1: Results from unrestrained vs. restrained protein simulations using CCN and water to solvate HEWL (white). The single hotspot identified by MSCS is shown in stick form; CCN (cyan). The probe density from the fully restrained simulation is shown in orange, from backbone-restrained in green, and from fully flexible in blue. Many incorrect local minima are seen in green and orange, but only the correct position dominates the simulation of the fully flexible protein in blue.

It appears that the numerous local minima obtained when performing gas-phase minimizations of probe molecules are not an artifact of the vacuum; they are an artifact of using a rigid protein conformation. A rugged landscape is observed, even in the presence of mobile solvent and side chains. The abundant local minima cannot be distinguished from the binding site, and probe mapping cannot successfully differentiate between irrelevant and druggable hot spots. With full receptor flexibility included, MixMD appropriately reproduces the one hot-spot binding site seen in the crystallographic data for CCN. The agreement between simulation data and experimental electron density validates MixMD as an accurate mapping tool (Figure 2-2).

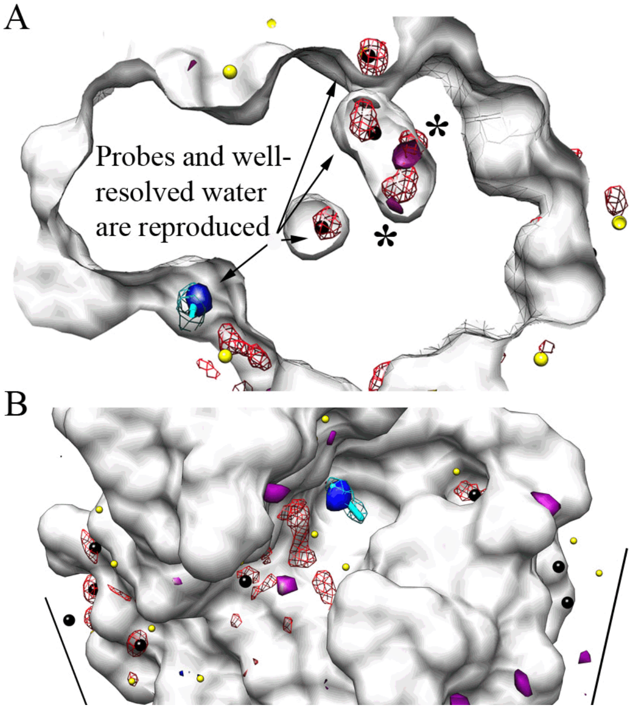


Figure 2-2: MixMD data for the fully flexible simulation agrees well with densities from MSCS experiments using CCN and water. All snapshots were superimposed on the crystal structure of HEWL (white surface). MixMD density for water is shown as a red mesh, and crystallographic waters are colored black for B-factors below 33 Å and yellow above 33 Å. The MSCS coordinates for CCN are shown in stick form (cyan with its electron density as a mesh, $2F_O$-F_C shown at 1.5σ); the highest occupancy for CCN in the fully flexible simulation matches perfectly and is shown in solid blue surface. Unsatisfied electron density in the crystal structure (positive F_O-F_C shown at 3σ) is shown in solid purple. (A) Water maps within the protein highlight interior waters conserved in the crystal structure and reproduced in our simulation. The large astrices (*) denote two highly occupied regions of the interior water map that correspond to unfulfilled density in the crystal structure. (B) Maps of the protein surface show good correspondence between crystallographic and MixMD densities for both CCN and well resolved waters. For the 16 crystallographic waters with B-factors below 33 Å (black spheres) five occur at symmetry-packing interfaces. MixMD misses four of the five, which is expected because the contacts are not present in our simulation. The other 11 best-resolved waters are well reproduced.

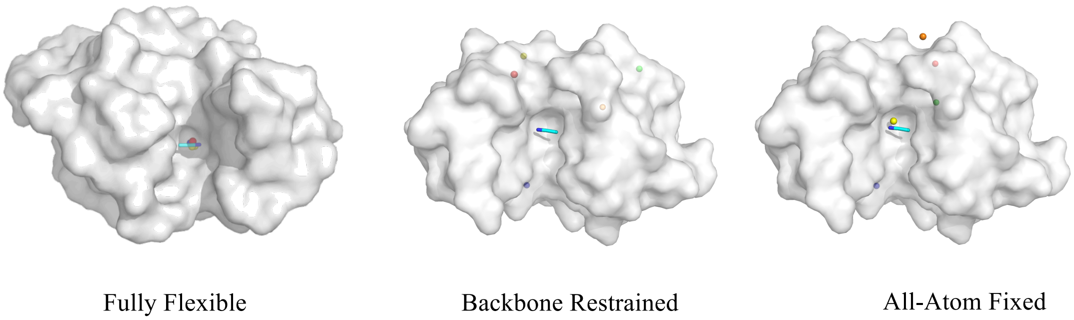
In addition to the CCN hot spot, MixMD reproduced the locations of low-B-factor water (<33 Å). The only locations that were not reproduced were on surfaces of the protein that were involved in crystallographic contacts (Figure 2-2B). A few locations were seen where significant water occupancy in the interior of the protein did not correlate with water coordinates in the crystal structure, but those locations were in excellent agreement with unfulfilled density in the crystal structure (Figure 2-2A). The location of positive density on the Fo-Fc map may in fact correspond to water positions.

While not all unfulfilled density will correspond to solvent molecules, the locations identified by MixMD water maps may indicate positions where water should have been placed.

2.3.1. Convergence of Sampling

Though the 10-ns sampling time used in the simulations is relatively short by current standards, it is important to stress that long trajectories are inappropriate in mixed solvent. Modest timescales are needed: long enough to allow solvent equilibration and convergence, but short enough to avoid possible unfolding of the protein. Furthermore, an accurate MD technique built on short timescales makes this method more accessible for practical application in a pharmaceutical setting.

We calculated the maximal occupancy location of each probe type during each individual simulation using the *ptraj* grid utility. These positions for CCN over the last 2 ns of production were compared between independent simulations of the same initial system (Figure 2-3). Excellent convergence is seen across the five independent MixMD of the fully flexible HEWL. However, the individual simulations of the rigid and backbone-fixed simulations did not agree on a common location for the CCN hot spot, reflecting a propensity for solvent molecules to become trapped within local minima along the protein surface. For the fully flexible simulation, these points were all within <1 Å, which is within the limits of accuracy when using a 0.5 Å grid. Not only did the locations agree with one another, they were in excellent agreement with the position for CCN in the crystal structure. In contrast, there was no agreement between the five independent MixMD simulations of the rigid and backbone-fixed HEWL. Those simulations also failed to identify the correct location for the CCN hot spot, except for one trajectory of the rigid HEWL.



Fully Flexible Backbone Restrained All-Atom Fixed

Figure 2-3: The maximal occupancy positions over the last 2 ns for each independent simulation of HEWL in pre-equilibrated 50% w/w CCN and water. Only the fully flexible system shows convergence of the five simulations in agreement with experiment. Individual runs are indicated by color (red, orange, yellow, green, and blue).

To further compare sampling, we calculated the ratio of the number of solvent probes to water molecules at the edges of the box. Far from the protein, there should be no bias between the solvents, and the ratio of their occupancies should approach N_p/N_w (ratio of the number of CCN probes to the number of water in the simulation).[220] All systems demonstrated good convergence according to this metric with the fully flexible system being the least biased (Table 2-1,

Table 2-2). The fact that CCN and water exchange freely at the box edge indicates that the mixed solvent system inherently samples evenly, but the pronounced differences at the protein surface indicate that solvent molecules become trapped and poorly sample the rugged potential surface of a rigid or semi-rigid protein.

Table 2-1: The extent of convergence for HEWL in MixMD, as determined by the ratio of co-solvent to water molecules at the box edges. Values of Obs/Exp near 1.0 indicate complete and unbiased sampling at the edges of the box.

	Pre-equil 50% Fully Flexible	Pre-equil 50% Backbone Restrained	Pre-equil 50% All Atoms Fixed
$(N_p/N_w)_{Obs}$	0.4501	0.4664	0.4709
$(N_p/N_w)_{Exp}$	0.4490	0.4490	0.4490
Obs/Exp	1.0024	1.0388	1.0488

Table 2-2: The extent of convergence for fully flexible HEWL in MixMD, as determined by the ratio of CCN to water at the box edges. Simulations of 90% w/w CCN were problematic, and the behavior at the edges of the box show that they do not demonstrate even, unbiased sampling (the Obs/Exp value is much larger than 1.0).

	Pre-equil 50% w/w	Layered 50% w/w	Layered 10% w/w	Layered 90% w/w
$(Np/Nw)_{Obs}$	0.4501	0.4619	0.0477	4.7453
$(Np/Nw)_{Exp}$	0.4490	0.4314*	0.0489	3.6563
Obs/Exp	1.0024	1.0707	0.9755	1.2978

* The layered 50% box contained a few more water molecules than the pre-mixed box, giving it a slightly different expected (Np/Nw).

2.3.2. Initial preparation of the mixed solvent environment

The results above were obtained with a pre-equilibrated 50% w/w solution, but we have also examined other choices for the mixed solvent environment. Two protocols for initiating the mixed solvent box were compared. The first used the pre-equilibrated 50% w/w mixed solution, providing an even distribution of both solvents (data shown above). The second method aimed to reproduce the MSCS experiment where the CCN has to compete water off the surface of the protein. The waters were placed in a shell around the protein, and the CCN were placed outside the water shell, resulting in a layered solvent environment.

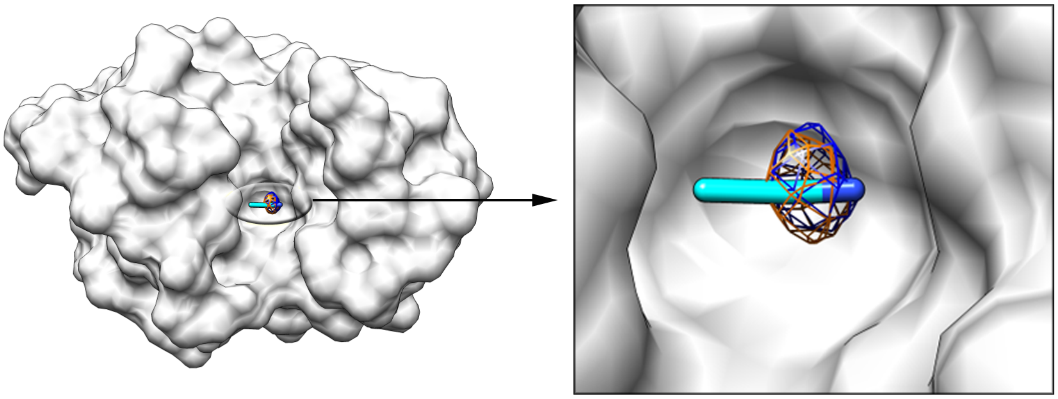


Figure 2-4: Combined results from the last 2 ns of all 5 simulations (10 ns of sampling) with flexible HEWL (white surface), comparing the layered solvent to the pre-equilibrated protocol. The probe density is shown in blue mesh for 50% w/w pre-equilibrated solvent and in orange mesh for 50% w/w layered water and CCN. The highest sampled densities overlap exactly and agree well with the position of CCN in the 2LYO crystal structure.

Densities of CCN were in good agreement between the two solvent protocols (Figure 2-4). Maximal occupancy positions were used to compare coordinates of the experimental probes to simulation probes. For simulations of fully flexible HEWL, we found that the layered solvent produced a maximally occupied location 0.8 Å from the crystallographic C2 atom of CCN. The pre-equilibrated, evenly mixed solvent produced a maximally occupied location 0.9 Å from the crystallographic C2 atom of CCN. These maximally occupied locations were 0.5 Å away from each other. Again, this is within the limits of error of our grids for calculating the occupancy maps. It appears that either protocol may be appropriate for 50% w/w CCN and water (**Error! Reference source not ound.**), but the layered solvent showed a slight disagreement in the convergence of the five independent simulations (Figure 2-5).

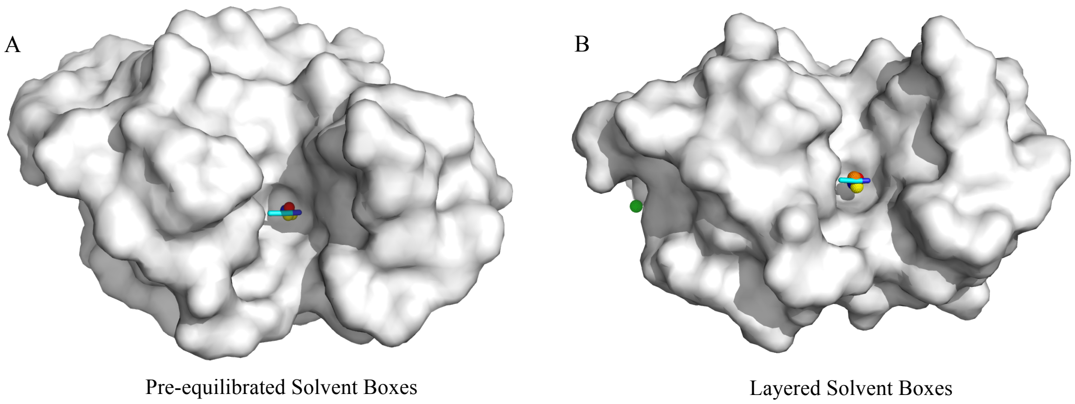


*Figure 2-5: The positions of maximal occupancy on the CCN grid were calculated over the last 2 ns for each independent simulation around a fully flexible HEWL (white). **A**) Results for the individual simulations based on pre-equilibrated solvent are shown. All five positions are in excellent agreement with one another and the CCN molecule in the 2LYO structure. The second (orange) and fourth (green) run results lie beneath the first (red) and fifth (blue) run results. **B**) Results for the individual simulations based on layered solvent are shown, with the orientation skewed slightly from A to show the green site (run 3) that does not agree. Four of the five positions are in agreement with one another and the crystallographic CCN. The first run result (red) lies beneath the fifth run result (blue).*

We have also examined 90% and 10% w/w mixed solutions of water and CCN to determine whether maps are more accurate when more or fewer probes are present. Both 90% and 10% mixtures identified the correct hot spot for CCN (Figure 2-6). However, we found that the 50% mixtures gave better water maps and more complete sampling than either 90% or 10% mixtures of CCN and water (Figure 2-7).

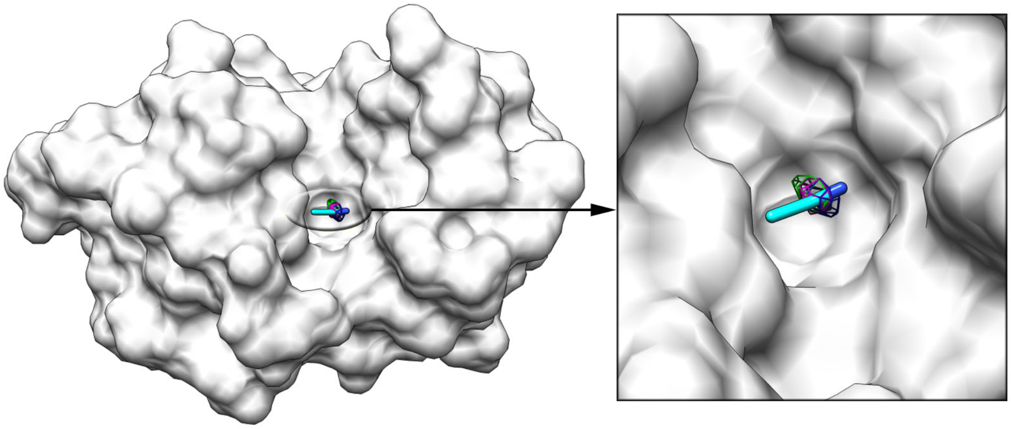
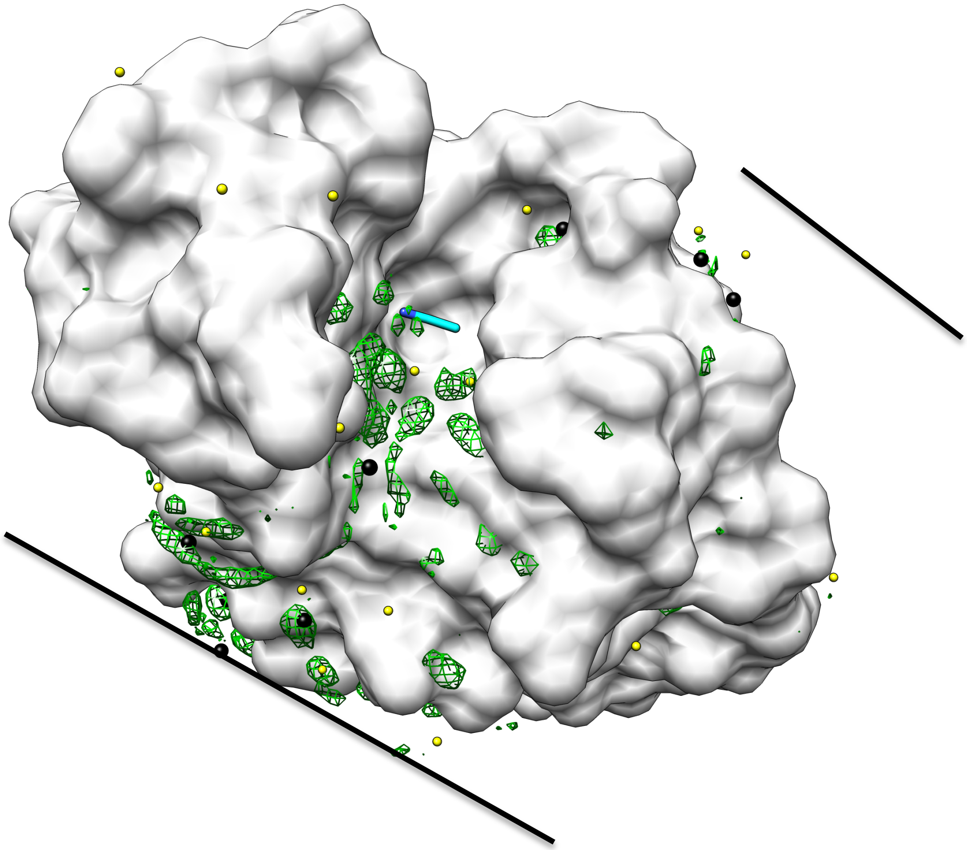


Figure 2-6: Results from the three different solvation protocols with fully flexible HEWL and CCN as the organic probe. The probe density calculated by combining the last 2 ns of the five independent simulations (to give 10 ns of sampling) is shown as a purple mesh for 90% w/w CCN, as a blue mesh for 50%, and as a green mesh for 10%. Results from all three simulations are nearly identical and reproduce the crystallographic position of CCN.



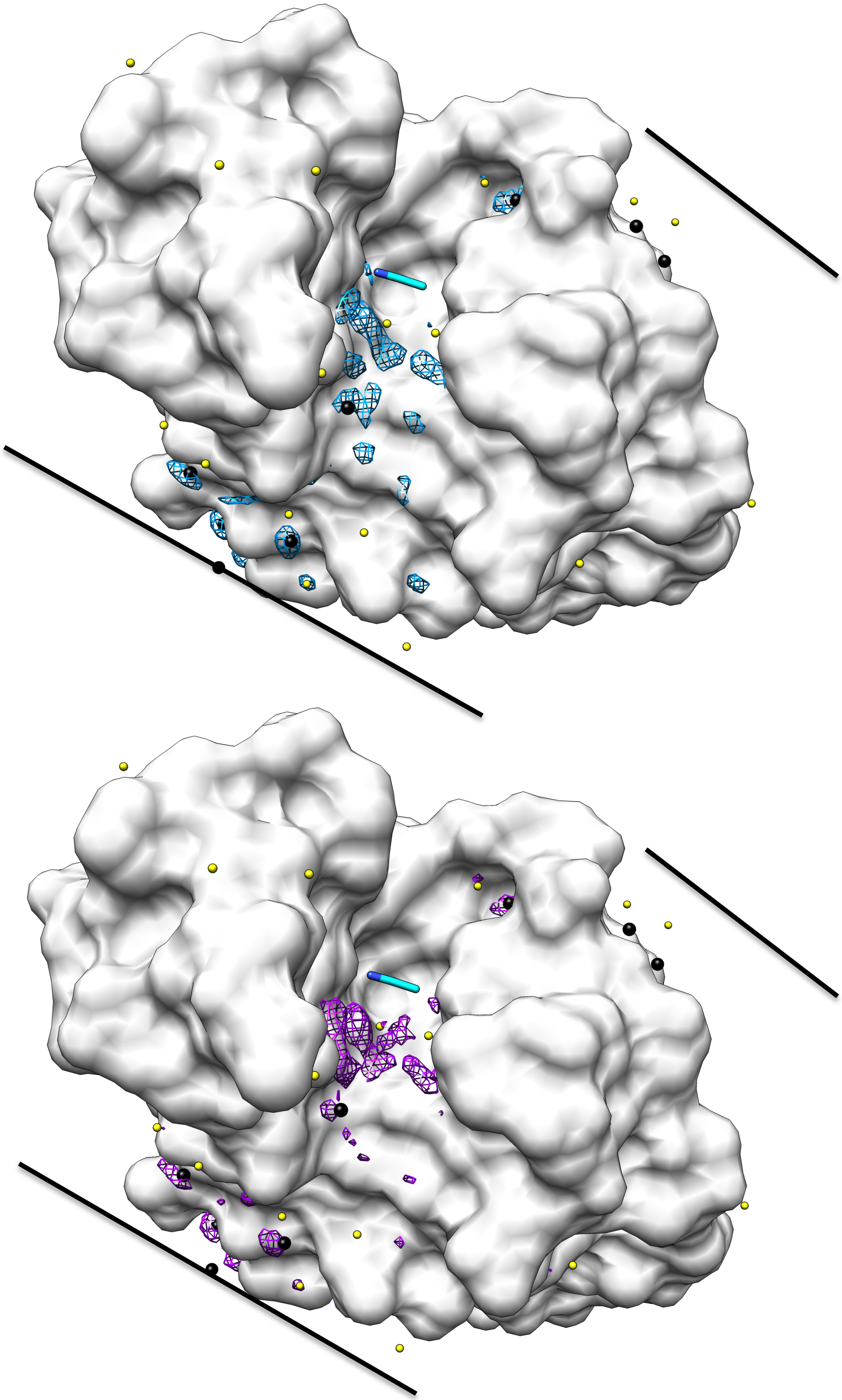


Figure 2-7: The water density calculated by combining the last 2 ns of the five independent simulations (giving 10 ns of sampling) is shown. The water density is shown as a green mesh for 10% w/w CCN (A), as a blue mesh for 50% (B), and as a purple mesh for 90% (C). The separate solvent densities are overlaid in

(D). Crystallographic waters are colored black for B-factors below 33 Å and yellow above 33 Å. Although the 10% and 90% environments reproduce the CCN position, they do not give appropriate mapping for the water. In A, too many equivalent positions for water are seen without identifying the water with higher B-factor (yellow). In C, there are too few water locations.

2.4 Conclusion

Our results demonstrate the need to include protein flexibility to achieve valid hot-spot mapping. MixMD simulations have been successfully performed to determine the correct mapping procedure for locating truly relevant binding minima. MixMD was capable of locating hot spots for the CCN solvent probe, and it identified crystallographic waters with the lowest B-factors, crystal contact waters, and locations where water could have been modeled into the structure (unsatisfied density in the F_O-F_C map). The information contained within individual MixMD trajectories can be combined into a consensus model retaining only the consistently important mapped sites. We have shown that only through the incorporation of protein flexibility and appropriate solvent competition can viable mapping results be obtained.

This work was published as***:*** Lexa KW and Carlson HA. Full Protein Flexibility is Essential for Proper Hot Spot Mapping. *J Am Chem Soc* **2011**, *133*, 200-202.

Chapter 3

How to Compare Computational Probe Mapping to Crystal Structures

3.1 Introduction

The development of FBDD as a compliment to traditional SBDD has allowed discovery efforts to probe binding potential across a broad range of protein structures, including many traditionally difficult cases. FBDD was first introduced as a viable tool for drug discovery in 1996 with the SAR-by-NMR and multiple solvent crystal structure (MSCS) methods.[157,158,221] In the first SAR-by-NMR study of protein-ligand binding, small organic molecules with low molecular weight were screened for binding affinity to FKBP and their bound positions were determined from ^{15}N-HSQC spectra.[158] The authors identified low-affinity binding sites along the protein surface, optimized the resultant hits, and then linked the fragments together, thereby achieving ligands with nanomolar affinity for FKBP. In the original MSCS study, cross-linked crystals of elastase were soaked in a solution of CCN, allowing CCN to compete off water and form favorable interactions along the protein surface. Then, the crystals were washed to remove unbound CCN, enabling the crystallographers to identify optimal binding sites for amphiphilic nitrogen-containing fragments. The authors concluded that the MSCS results from several organic probe types could be superimposed and translated into a template, or pharmacophore, for drug design. Both x-ray crystallography and NMR techniques allow for the identification of hot spots, defined as sites where a small number of protein residues confer most of the free energy of binding.[222,223] These hot spots signify regions where important protein-ligand interactions may be formed and thus represent a tool for the design of novel therapeutics.

FBDD has been effectively implemented in the development of clinical candidates.[224,225] However, many combinations of linked fragments may be required before identification of a viable lead compound. Computational techniques can be used to complement the existing experimental methods for FBDD by offering a less costly approach to the examination of potential binding interactions in the target system. Several groups have recently developed computational techniques that incorporate concepts from MSCS into MD simulations. The first of these computational approaches to probe mapping used a single 16 ns MD simulation of protein solvated by a box of 20% volume/volume (v/v) IPA and water to identify hot spots based on grid occupancies and calculated binding free energies.[177] The second MD-based approach, SILCS, relied upon a ternary solvent box of 1M benzene, 1M propane, and water to calculate probe occupancies and map the potential energy surface.[179,180] However, each of these methods has been limited by the identification of irrelevant local minima with equal weight as the true binding site(s). With this problem, extrapolation to cases where the answer is unknown becomes very difficult. An improved sampling protocol that reduces spurious minima is essential for robust application. Spurious local minima are also a common problem among the traditional computational probe-mapping techniques like GRID[91] and MCSS[92].

The FTMAP algorithm[176] deserves mention because it was also based on experimental methods for fragment mapping. It applies a fast, Fourier-Transform approach to correlate the docking of billions of probe conformations to a rigid protein. While it does not dynamically sample protein flexibility like the MD-based methods, it does not identify as many spurious minima as other probe-mapping methods.

In our development of MixMD, we focused on providing a computational tool complementary to the MSCS approach that would preferentially locate the most relevant hot spots along a protein surface.[181] MSCS studies can be difficult to perform with fragile crystals or at a high concentration of organic solvent, and results can be influenced by the crystallization conditions. Computational studies avoid these limitations and enable the detailed study of protein-probe interactions. Our original MixMD study was the first computational technique to definitively show the need to include full protein flexibility to reduce extraneous minima and correctly map a protein surface. Our initial studies of HEWL in CCN and water demonstrated the utility of MixMD for hot-spot mapping, and we would like to incorporate additional functional groups, to permit consensus pharmacophore modeling of putative binding sites. We were particularly focused on the impact of length and number of simulations on solvent behavior and protein structure. To determine the appropriate MixMD approach that would be applicable to systems where the binding site(s) are unknown, we have performed MixMD for a range of protein cases with the most common MSCS probes: IPA and CCN.

Hot-spot data are available for IPA and CCN with the following proteins: elastase, HEWL, p53 core, RNase A, subtillisin, and thermolysin. These probes were originally selected for MSCS based on their miscibility with water, their interaction type, and the ease of distinguishing their crystallographic density from that of water.[221,226] Due to the subjective nature of assigning experimental density to specific atoms, the most valid comparison between simulation and MSCS is between occupancy grids and experimental density data. As a result, we focused primarily on cases with electron density data available: elastase+IPA, HEWL+CCN, HEWL+IPA, p53 core+IPA, RNase A+IPA, and thermolysin+IPA. Structure factor files were not available for elastase+CCN (coordinates obtained from author[221]), subtillisin+CCN (1SCB[166]), or thermolysin+CCN (1FJU[163]) but our simulations of these systems showed that our results agreed with the known binding sites of these proteins. These proteins vary in size and active-site composition, thus providing a variety of interaction types to explore through MixMD to develop the most robust protocol with the greatest potential for application to new proteins.

3.2 Methods

HEWL+CCN was the focus of our previous study.[181] So we extended CCN+water to new systems and added IPA as a probe. We concentrated on the systems with experimental density data available for this study. The starting structures for elastase+IPA (2FOF[227]), HEWL+IPA (1LY0[228]), p53 core+IPA (2IOM[229]), RNase A+IPA (3EV2[167]), and thermolysin+IPA (7TLI[226]) were obtained from the PDB[214]. Crystallographic waters and structural ions were retained to maintain their stabilization of the protein conformation, but all probe molecules and crystallographic ions were removed. We emphasize that no simulations were initiated with the probe in the experimental conformation. Molprobity[230] was used to check the side-chain orientations of ASN, GLN, and HIS, and these results were confirmed by a visual examination.

The crystallographic structure of RNase A contained two copies of the protein within the asymmetric unit. Chain B was selected for use in our simulations because chain A was missing several residues. Furthermore, chain B contained two IPA molecules located in the active site, while chain A contained two IPA molecules that were located along a packing interface.

Parameters for IPA were based on the OPLS-AA parameters for pure alcohols from Jorgensen *et al*.[231] This choice was based on an in-depth exploration of available solvent parameters, discussed in a forthcoming publication. Parameters for CCN were obtained from Kollman and co-authors and implemented as described in our established MixMD protocol.[181]

The atomic coordinates from the MSCS studies were used for each system. Using the AMBER10/AMBERTOOLS 1.2 package[10] and FF99SB parameter set[7], hydrogens were added to the protein by *tLeAP* and minimized in *sander*. The protein was solvated in an 18-Å pre-equilibrated box of 50% w/w probe and TIP3P water[215], and then ions were added to neutralize the system charge. Simulations used a 2-fs timestep, SHAKE[4] to restrain bonds to hydrogen, and a 10-Å cutoff for Particle Mesh Ewald[232] approximations of long-range vdW interactions. The initial velocities of each independent simulation were generated from a different random number seed. Temperature was regulated through an Anderson thermostat.[217] Each system underwent 250 cycles of steepest-descent minimization followed by 4750 cycles of conjugate-gradient minimization with the protein fixed. Then, each system was gradually heated from 10K to 300K over 80 ps while the protein was gently restrained by a harmonic force constant of 10 kcal/mol*Å. These restraints were gradually removed over 500 ps of equilibration until the protein was fully flexible. Five independent 50-ns simulations were performed simultaneously for each MixMD system; for elastase and thermolysin, a total of 10 independent 50-ns simulations were generated.

Since our binary-solvent simulations commenced from the crystal coordinates of the MSCS structure, we could use a common frame of reference to accurately compare our simulation data with the crystallographic density data when available. The results of our MD simulations were analyzed using *ptraj,* a module within the AMBERTOOLS package. The final 5 ns of each individual run were read into *ptraj* and combined to represent the converged system density data. The simulation data was first imaged and fit to the crystal conformation by Cα-RMS; then the solvent occupancies of water and probe were calculated using the grid command with a 0.5 Å x 0.5 Å x 0.5 Å spacing over the entire box. To correlate with experimental data, occupancies were examined for each solvent heavy atom, each solvent residue, both solvent types together, and all atoms. The occupancy grid for all atoms is the equivalent of the electron density in the crystal structure, but it removes the identity of which atoms are populating the grid points. Contour levels for probe density were chosen such that the first 5 maximally-occupied regions were visible (i.e. the 5 hottest spots) in order to extrapolate from our results to the best approach for analysis of systems where the answer is unknown.

3.3 Results and Discussion

3.3.1. Electron density

The most accurate comparison between experiment and computation involves a joint examination of the MSCS electron density and the occupancy grids from the MixMD simulation. Although the MSCS probes were specifically chosen to aid in identification of crystallographic density, placement of probes is necessarily subjective. Electron density cannot be unambiguously assigned in many cases, so occupancy grids for solvent should not be directly compared to *atomic coordinates*. It is often unclear whether density represents the position of a probe or a water molecule, and the placement of one of these versus the other can affect the final density map.

Electron density maps are typically presented as $2F_O$-F_C and F_O-F_C maps, where F_O refers to the observed phasing information and F_C refers to the calculated phasing information. The $2F_O$-F_C map illustrates true structure features, while the F_O-F_C map is a difference map of areas where F_O does not match F_C. When examined at a sufficiently high level ($\geq 2.5\sigma$), noise in the F_O-F_C map is minimized, and it shows positive peaks that indicate structure features which may be missing in the model as well as negative peaks that represent features in the model which are in error. The atomic coordinates for all probes must be compared to the refined F_O-F_C map to ensure that each is supported by appropriate electron density (Table 3-1).

Table 3-1: Analysis of the refined cycles in Buster of electron density for each MSCS probe site in the noted crystal structure. The real-space correlation coefficient from the EDS is shown in parentheses.

Protein-Probe System	Unsupported Probe Density	Supported Probe Density	Supported at an Interface
HEWL (1YL0)		2836 (0.934)	2837 (0.741)
p53 core (2IOM)		3001 (0.889)	
RNase A (3EV2)	905A (0.934)*, 917A (0.881), 903B (0.485)*	902B (0.723)	
Elastase (2FOF)		1001 (0.823), 1003 (0.871)	1002 (0.898), 1002x
Thermolysin (7TLI)	2004 (0.838), 2005 (0.742), 2007 (0.703), 2008 (0.693)*	2001 (0.837), 2002 (0.872)	2003 (0.494), 2006 (0.921)
	*Density does not support a probe site, but is instead a potential water site		

There are several caveats inherent to this comparison. Crystallographic density may include artificial features influenced by experimental conditions and crystal contacts. As Allen *et al*. acknowledged in their MSCS study of elastase+CCN[221], several of the CCN probes were involved in crystal-packing interactions and were not believed to indicate true hot spots for binding. Therefore, probes near crystal-packing interfaces required additional scrutiny (Figure 3-1). Depending upon the protein system, it may not be possible to reproduce the probe location at an interface because it is not favorable in the solution phase simulated in MixMD.

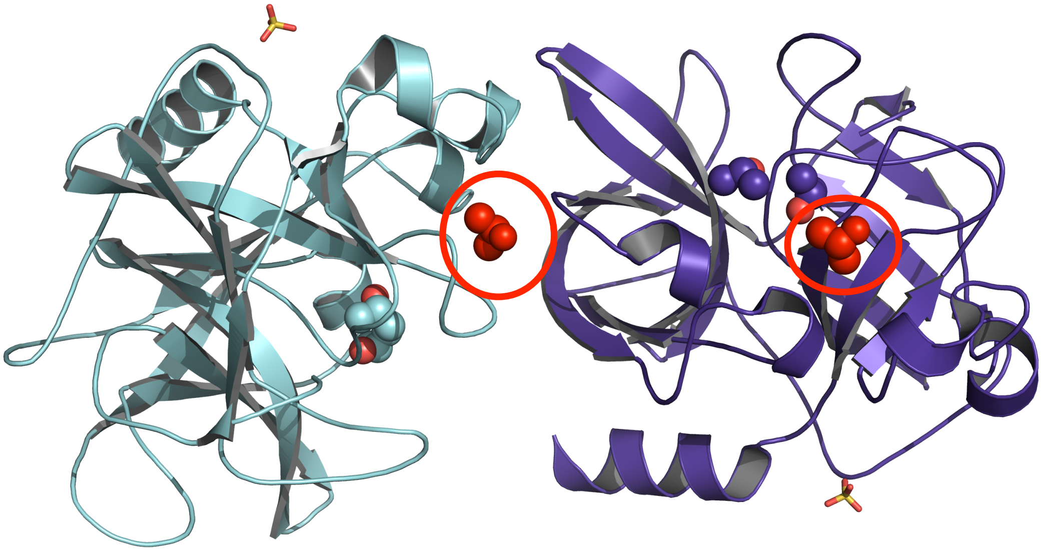


Figure 3-1: The MSCS of elastase+IPA[228] is shown in violet (right) with a neighboring symmetry partner in cyan (left). The symmetry-related equivalent IPA in both structures are circled in red. One of the bound IPA probes in the symmetry partner is located within 4 Å of the protein in the initial unit cell. When the probe site is support by the refined electron density, it is necessary that it receive special consideration during evaluation of the simulation data. Probes at the contact interfaces may be irreproducible, but when they are observed, the probe may map to either or both of the circled sites on the protein.

3.3.2. Identifying probes with appropriate density

It must be stressed that crystallographic structures can be subject to phase bias/ experimental uncertainty, and accurate comparisons between theory and experiment can only be made when the limitations of both are understood. To properly confirm that a probe site is supported with appropriate density, the F_O-F_C map must be regenerated with that particular probe excluded from the set of atomic coordinates. This removes the bias of the probe on the determination of the density. This procedure was important for confirming hot spots in the crystal structure, and it was just as important for our analysis of any “spurious” sites observed in the occupancy grids from MixMD. In some cases, “spurious” sites actually reflected the position of a valid probe site in the neighboring unit cell. Thus, it was important to examine the surrounding symmetry partners present in the crystalline environment for probe sites in contact with the central protein. When refined density maps were generated, several outcomes were possible. The probe and/or probes along the crystalline interface could be well-placed in the positive F_O-F_C density, the F_O-F_C density at the probe site appeared to support placement of a water molecule, the probe was placed in negative F_O-F_C density, or there was not sufficient density to justify placement of any molecule (Figure 3-2).

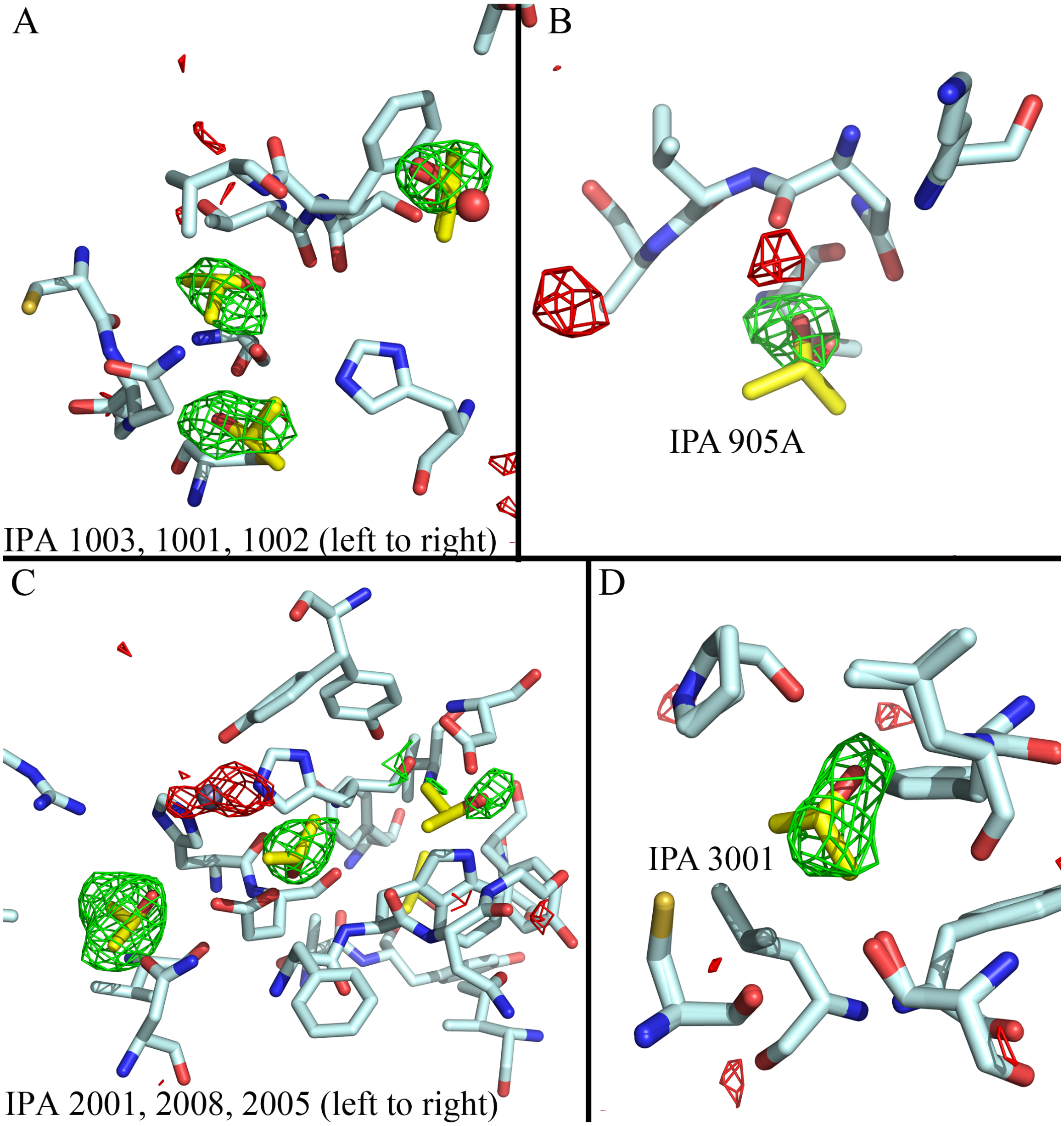

A
B
IPA 905A
IPA 1003, 1001, 1002 (left to right)
C
D
IPA 3001
IPA 2001, 2008, 2005 (left to right)

Figure 3-2: Refined density maps calculated for several of the MSCS structures. The refined maps were based on the coordinate file for the MSCS structure, with probes and close-contact waters/ions removed. The positive F_O-F_C density is contoured in green; the negative F_O-F_C density is contoured in red at 3.0 σ. This illustrates regions where crystallographic probes are justified or not justified, respectively. In A, the positive F_O-F_C density from the MSCS of elastase+IPA (2FOF) clearly agrees with the probe placement, where IPA.1001 and IPA.1003 are located in the active site with good agreement between coordinates and density (IPA.1002 is located near the crystal interface). The positive F_O-F_C density in B indicates that the site for IPA.905A in RNase A+IPA (3EV2) may best correspond with a water molecule. In C, the F_O-F_C density along the active site of thermolysin+IPA (7TLI) indicates that the site for IPA.2001 is justified while the sites for IPA.2005 and IPA.2008 might best support a water molecule as opposed to a probe. In D, the F_O-F_C density in p53 core (2IOM) illustrates that this probe is justified in the crystal structure.

CCP4i[233] was used to derive crystallographic density for each MSCS probe and to rotate maps of the simulation density into the crystal structure orientation. In CCP4i, the pdb and structure factor files were used to derive the mtz file, which was then refined through one complete round of Buster[234] (five cycles). To verify the probe locations in MSCS, each probe was removed from the structure individually. The density in these figures has been generated through a refinement cycle and is therefore *not equivalent* to the ready-made mtz file that is electronically-available through the Uppsala Electron Density Server[235]. Chimera[218], COOT[236], and PyMol[237] were used for visualization of results. Instead of displaying an additional density layer in the figures, presentation was simplified by showing the MSCS probes whose positions were confirmed by the F_O-F_C density from Buster[234] in cyan and probes without adequate density in gray as ball-and-stick coordinates.

3.3.3. Appropriate Simulation Length: Adequate Sampling and Convergence

We were interested in establishing the appropriate simulation length for MixMD to allow widespread application of hot-spot mapping, including cases where the active site might be unknown. In order to determine the appropriate methodology to be used with a protic probe, which should diffuse more slowly than an aprotic probe, we first had to define appropriate system behavior. All of the mixed-solvent simulations were stable over the ns timescales examined here. Long timescales should reveal unfolding of the protein and are less desirable. To assess whether adequate mixing of the two solvents had occurred, we examined the number ratio of probe to water (N_p/N_w) at the edges of the box. In a properly mixed system, N_p/N_w at the edges of the box will correspond with

N_p/N_w of the whole system, since the presence of the protein should not impact the distribution of solvent at the box edges. We found that this was indeed the case within 5 Å of the box edges; therefore, our binary solvents were mixed appropriately (Table 3-2).

Table 3-2: Probe-to-water ratio at the box edges over the last 5 ns of simulation time, using 5 runs. The expected ratios differ slightly due to system setup. Upon introduction of the protein into a 50% w/w pre-equilibrated solvent box, a random number of water and probes are removed from the system to avoid clashes with the protein. This can shift the probe:water ratio, thus slightly altering the composition of solvent in the system.

	Elastase+ IPA	Thermolysin + IPA	HEWL+ IPA	RNaseA + IPA	p53 core + IPA
$(Np/Nw)_{Obs}$	0.2114	0.2104	0.2276	0.2181	0.2489
$(Np/Nw)_{Exp}$	0.2212	0.2222	0.2219	0.2379	0.2208
Obs/Exp	0.9557	0.9480	1.0255	0.9168	1.1271
Np (simulation)	1949	2301	1459	1455	1847
Nw (simulation)	8810	10355	6576	6116	8365

HEWL was chosen as our initial validation protein because it is a simple canonical system with a wealth of available structural and binding data. IPA is a protic solvent with a slower diffusion rate than CCN and we were interested in comparing our results from HEWL+CCN to the HEWL+IPA system. The MSCS of HEWL contained only a single CCN bound at the active site, but two bound probes were identified in the MSCS of HEWL+IPA: IPA.2836 was bound in the active site and IPA.2837 was located on the opposite surface at the crystal interface.

Although our previous MixMD study used the final 2 ns from a set of 10-ns simulations to identify hot spots, we have found that a wider window and longer simulation time was needed for obtaining converged data with a protic solvent. To calculate the optimal window size, we examined the data that resulted from combining the final 2-ns, 5-ns, 10-ns, 20-ns, 30-ns, and 40-ns segments of a 50-ns simulation. For example, the last 5 ns from five independent 50-ns trajectories of HEWL+IPA were read into *ptraj* and combined by the grid function to yield solvent occupancies for 25 ns of total simulation time. For both HEWL and p53 core, we found that the final 5 ns of trajectory data best represented the probe sites from the converged simulation. Therefore, all of the data presented in this study has used a final-5-ns timeframe for the calculation of simulation occupancy grids.

Length of the simulation was also examined. We analyzed the solvent occupancy from the first 10 ns (2-ns window), 20 ns (5-ns window for the following), 30 ns, 40 ns, and 50 ns. We found that by 50 ns the local minima had been sufficiently reduced such that only the relevant hot spots were located. Based on the proximity of the second IPA to the crystal packing interface and the small spherical density at that site (Table 3-1), we had hypothesized that IPA.2837 could be a water site, and our data supported this. We found that the maximal probe density converged to the location of the probe at the active site (Figure 3-3). Sizable water density was located the second IPA site and most of the low B-factor crystallographic waters from the MSCS (Figure 3-4). In cases where the refined density map indicates that the organic probe might actually be a water molecule, simulation density can aid in the proper assignment. Mapping water density and examining lifetimes of water at sub-sites have been established techniques for some time; our ability to map low B-factor waters demonstrates that MixMD does retain this functionality. This was shown in our previous paper on HEWL+CCN.[182] Therefore, we focus our discussion of water occupancy to only the probe sites where the experimental density appears to support a water molecule.

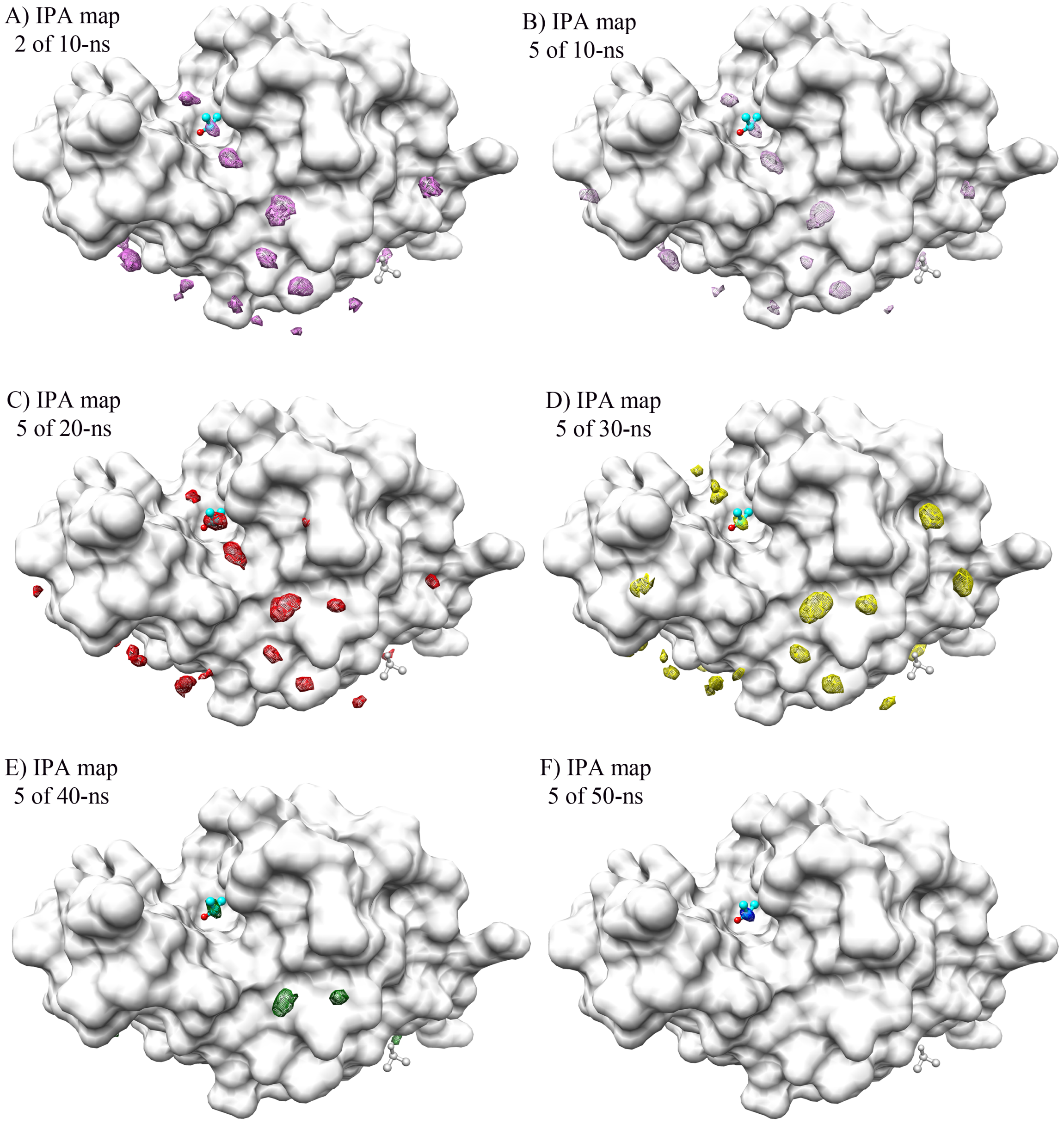


Figure 3-3: MixMD data combined for the last 5 ns of 5 individual 50-ns runs of fully flexible HEWL in 50% w/w IPA and water. All snapshots were superimposed on the starting crystal structure (PDB ID 1YL0, white surface). The MSCS coordinate positions for IPA.2836 and IPA.2837 are shown in cyan as ball-and-stick. A) Simulation IPA density does map the IPA.2836 site using a 2-ns window after 10 ns of simulation, however many other spurious minima are mapped as well. B-F) Using a 5-ns window, we find that the local minima are eliminated over time as the simulation converges to the maximally occupied hot spot.

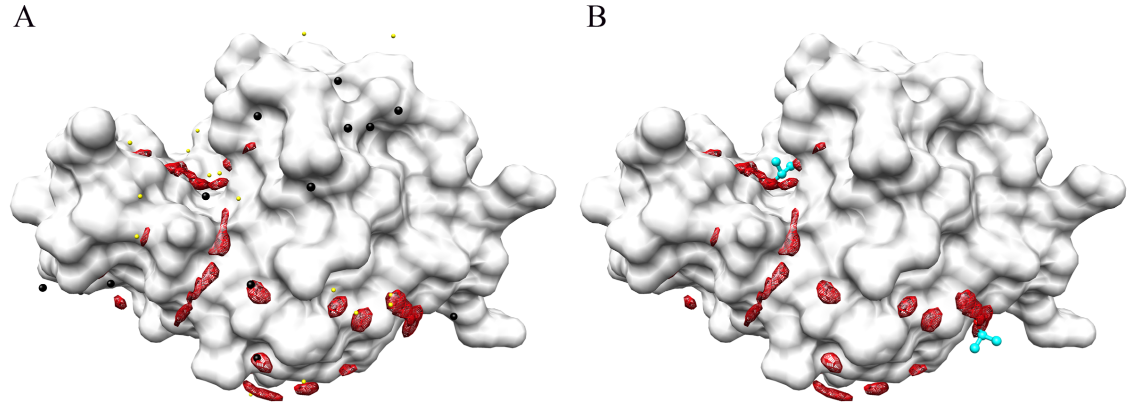


Figure 3-4: MixMD data combined for the last 5 ns of 5 individual 50-ns runs of fully flexible HEWL in 50% w/w IPA and water. All snapshots were superimposed on the starting crystal structure (PDB ID 1YL0, white surface). Low B-factor (<33 Å) water sites are shown in black and the other water sites are shown in yellow. The MSCS coordinate positions for IPA.2836 and IPA.2837 are shown in cyan as ball-and-stick. Simulation water density overlays well with both A) the crystallographic water sites and B) the secondary IPA probe, IPA.2837.

To determine the general applicability of 5 x 50-ns simulations as the optimal trajectory length for sampling with a protic solvent, we also applied MixMD to the MSCS of p53 core domain (PDB ID 2IOM). Although the crystal structure was missing residues 92-96 and 285-294, these residues are part of the disordered domain linkers and were not expected to impact mapping results. Experimentally, a single IPA molecule was bound to p53 core at the active site, and this was validated by the refined F_O-F_C density from Buster. As we found for HEWL+IPA, examination of 5-ns windows over the course of the trajectory showed that 50-ns of simulation time was sufficient for mapping hot spots with IPA while minimizing local minima along the p53 core surface (Figure 3-5). The crystallographic site matched our maximally occupied location from the simulation IPA density, demonstrating the capacity of MixMD to capture important binding sites in DNA-binding proteins.

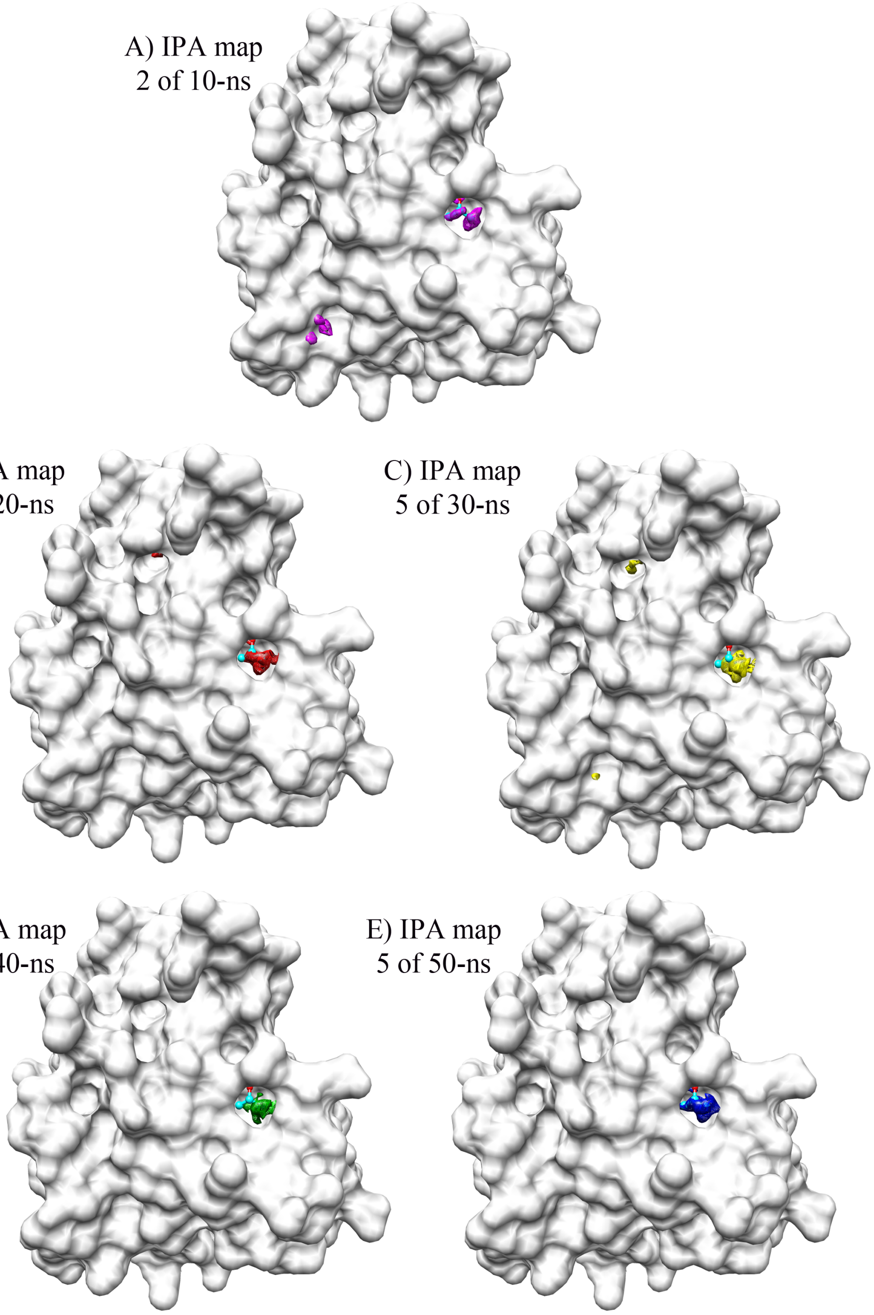


Figure 3-5: MixMD data combined for the last 5 ns for 5 individual 50 ns runs of fully flexible p53 core in 50% w/w IPA and WAT. All snapshots were superimposed on the original crystal structure (PDB ID 2IOM, white surface). MixMD density is shown in blue for the IPA probe. The MSCS coordinate position for IPA.3001 is shown in cyan as ball-and-stick. The maximally occupied site for simulation IPA density clearly agrees with the crystallographic probe position. A) Simulation IPA density does map the IPA.3001 site using a 2-ns window after 10 ns of simulation, however a few other spurious minima are mapped as well. B-E) Using a 5-ns window, we find that the local minima are eliminated over time as the simulation quickly converges to the maximally occupied hot spot.

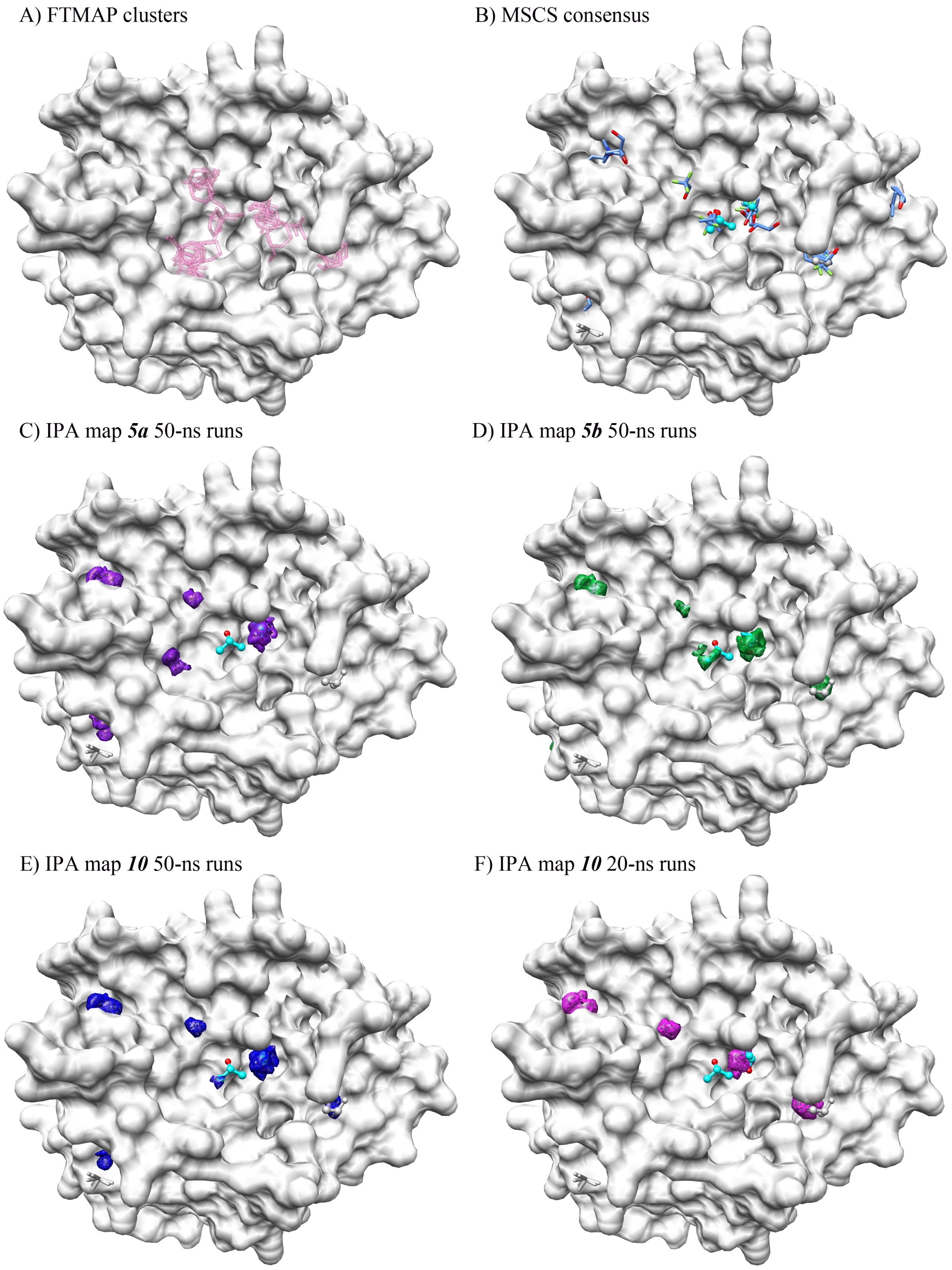
Through MixMD simulations of HEWL+IPA and p53 core+IPA, we showed that we could map the protein hot spots correctly without simultaneous identification of numerous irrelevant minima. We could show other minima by decreasing the contour value, but that does not change the fact that the correct location is clearly identified by the highest-occupancy grid points. Our studies based on these two proteins indicated an optimal simulation time of 5 independent simulations of 50 ns and an analysis window of 5 ns when running MixMD with a protic solvent.

3.3.4. Determining the Appropriate Number of Simulations

To study the consequence of the number of independent simulations, we compared the results from 5 runs of 50 ns each to 10 runs of 50 ns each for the larger protein systems, elastase and thermolysin. The success of MD simulations is firmly based on the assumption of the ergodic hypothesis: given adequate sampling, all relevant states will be reached. We prefer many shorter simulations run in parallel to increase sampling instead of one long simulation because they increase the potential space that may be explored. This is particularly true for computational studies of solvent mapping, where effective mapping must be balanced with an efficient use of computational cycles. Furthermore, long simulations should eventually unravel the protein. Many short simulations focus the maps on the more biologically relevant conformation of the protein. However, we need to determine if it is better for probe mapping.

MSCS of elastase were solved in the presence of acetone, CCN, benzene with IPA, cyclohexane with IPA, dimethylformamide, ethanol, IPA, and trifluoroethanol. The large amount of available crystallographic data enabled us to compare the results from MixMD of elastase+IPA/CCN in water to the experimental data for both IPA/CCN as well as other functional group probes. The MSCS structure of elastase+IPA contained three probe molecules; two were bound in the active site (IPA.1001 and IPA.1003), and the third was located near the crystal-packing interface (IPA.1002). Inspection of the symmetry partners revealed that IPA.1002 could also be located on an opposite face of elastase (referred to as IPA.1002x). Results from density refinement in CCP4i showed that all four locations were supported by the F_O-F_C density (Table 3-1).

Our MixMD results for elastase+IPA were highly similar to the MSCS data for its corresponding crystal structure as well as the consensus MSCS data of many probe types in many crystal structures. Overlaying all of the MSCS structures and examining their electron densities of the probes indicated that IPA.1001 was the most populated consensus site, which confirmed our identification of IPA.1001 as the maximally occupied hot spot. The other IPA positions in the binding site were also identified within the top 8 hot spots. We found that solvent occupancies calculated from the middle of 10 independent trajectories (15-20 ns) resulted in the better mapping when compared to the occupancies from the final 5 ns (45-50 ns) of 5 runs (Figure 3-6). Although 5 simulations performed over 50 ns were sufficient to show convergence, doubling the number of simulations decreased the amount of simulation required: 250 ns were required for 5 runs compared to 200 ns for 10 runs. After the first 20 ns for each of the 10 runs, the hot spot at IPA.1001 had been identified as the highest populated site and continued to be the primary hot spot for the duration of simulation time. Examination of the top eight hot spots showed that IPA.1001, IPA.1002, and IPA.1003 were each mapped, in addition to a pocket close to IPA.1002x and several hot spots from other MSCS results, including sites with justified electron density for ethanol, dimethylformamide, and IPA. The other binding sites that were identified by probe or water density included crystallographic water positions and pockets near hydrogen-bonding residues Arg36 and Arg54. Although structure factor files were not available for elastase+CCN, our MixMD simulations of this system preferentially located the active site as well as several MSCS probe locations.



A) FTMAP clusters

B) MSCS consensus

C) IPA map ***5a*** 50-ns runs

D) IPA map ***5b*** 50-ns runs

E) IPA map ***10*** 50-ns runs

F) IPA map ***10*** 20-ns runs

Figure 3-6: MixMD data combined for the last 5 ns of 5 individual 50-ns runs of fully flexible elastase in 50% w/w IPA and WAT. All snapshots were superimposed on the original crystal structure (PDB ID 2FOF, white surface). The MSCS coordinate positions for IPA.1001, IPA.1002, and IPA.1003 are shown in cyan as ball-and-stick. The crystal interface probe site IPA.1002x is shown in gray. Alternate probes from other

MSCS by the same authors are shown in blue. Density contour levels were chosen such that the top eight hot spots were represented at a "hide dust" level of two. A) Results from performing FTMAP against PDBID 2FOF, with functional group clusters shown in pink. B) All probe sites from MSCS with elastase. C) Simulation IPA density from the last 5ns of 50 total ns from independent runs 1-5 (purple) and D) the last 5ns of 50 total ns from independent runs 6-10 (green) do not perform as well as the results from all 10 runs. E) The maximally occupied site for simulation IPA density from the last 5 ns of 50 total ns from 10 independent runs (blue) corresponds with all three the crystallographic probe positions as well as two other MSCS sites. F) Simulation IPA density from the last 5 ns of 20 ns total simulation time over 10 individual runs performs better than the IPA density after 50 ns over 5 runs.

Like elastase, thermolysin has also been used extensively for MSCS studies[163,226] and well as other computational FBDD studies[177,238,239], allowing us to compare our MixMD results to a wealth of experimental and computational data. The results from our simulations of thermolysin lent strong support to our emphasis on comparing simulation occupancy grids to the experimental density instead of to the crystallographic coordinates. Although several of the MSCS systems we have discussed had more than one solvent probe along the protein surface, the refined electron density has not always given a clear indication that the coordination position of the probe was justified. This was particularly true for thermolysin+IPA; although eight IPA molecules were specified in the PDB file, comprehensive analysis of the pre-generated density map from the Uppsala Electron Density Server[235], the refined structure factor file, and the crystal contact region showed that *none* of the IPA molecules were placed in justified locations (Table 1). In fact, the F_O-F_C density map from the refined crystallography data did not indicate the existence of any probe sites at 2.5 σ. A number of the solvent-mapping procedures that have been published have used the probe positions in thermolysin to validate their technique, which further emphasizes the necessity of understanding the limitations of crystallographic data prior to performing validation analyses.

Although the refined F_O-F_C density did not support the placement of probes along the binding surface of thermolysin in this particular crystal structure, which was solved at 90% IPA, this structure had been solved in concentrations ranging from 2 to 100% IPA by the same authors. Refinement of the electron density map for the 100% IPA crystal structure (PDB ID 8TLI) illustrates that some probe positions are justified based on structure factors in the MSCS at the highest probe concentration. The RMSD between 7TLI and 8TLI is equal to 0.25 Å, allowing reasonable comparison between the MSCS density results for the two structures. Thus, the thermolysin was not the same as the RNase A case, and further exploration of simulation occupancy results was appropriate. Also, the strong emphasis on thermolysin as a test system for fragment binding made it an interesting case for developing our simulation protocol. On the basis of comparisons to the electron density data from 8TLI, we continued our investigation of the optimal trajectory length for MixMD of proteins with large binding surfaces. Refined F_O-F_C density revealed that the only potentially justified probe sites were IPA.2001, IPA.2006, and IPA.2008. The maximally occupied sites for IPA molecules correctly agreed with the position of the greatest probe density in the active site, the position of IPA.2001. We identified several other hot spots in the thermolysin active site, supported by the density for IPA.2008 and alternate TYR157 position A. As we saw for elastase, the use of 10 runs of 50 ns each improved mapping results compared to 5 runs of 50 ns each (Figure 3-7).

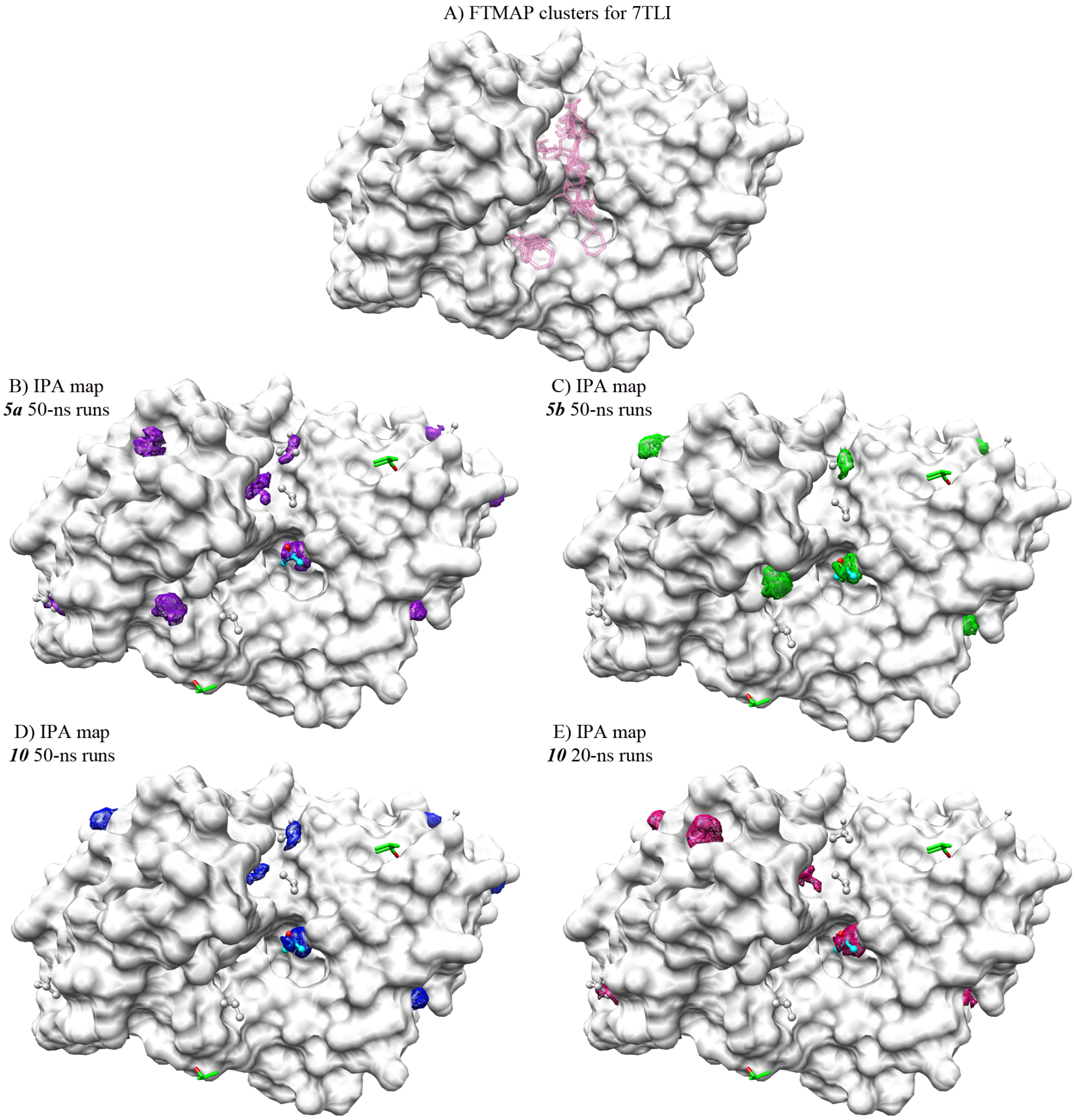


Figure 3-7: MixMD data combined for the last 5 ns of 5 individual 50-ns runs of fully flexible thermolysin in 50% w/w IPA and WAT. All snapshots were superimposed on the original crystal structure (PDB ID 7TLI, white surface). The MSCS coordinate positions for IPA.2001, IPA.2006, and IPA.2008 are shown in cyan as ball-and-stick while the unjustified probe sites are shown in gray. The crystal interface probe sites are shown in green. Density contour levels were chosen such that the top eight hot spots were represented, and a "hide dust" level of three was used. A) Results from performing FTMAP against PDBID 7TLI, with functional group clusters shown in pink. B) Simulation IPA density from the last 5ns of 50 total ns from independent runs 1-5 (purple) and C) the last 5ns of 50 total ns from independent runs 6-10 (green) do not perform as well as the results from all 10 runs. D) The maximally occupied site for simulation IPA density from the last 5 ns of 50 total ns from 10 independent runs (blue) corresponds with all three the crystallographic probe positions as well as two other MSCS sites. E) Simulation IPA density from the last 5 ns of 20 ns total simulation time over 10 individual runs performs better than the IPA density after 50 ns over 5 runs.

Our results for the appropriate number of simulation runs to perform for MixMD were particularly relevant to a common assumption in studies of hot-spot mapping, which holds that individual simulation densities should converge to the correct hot spot. We found this assumption to be invalid. Individual simulations all showed some occupancies at the principal sites of probe interaction, but they diverged in their identification of the global minimum. They also diverged in their positions for spurious sites. Combining the simulation data best represented the true binding potential of the protein and eliminated irrelevant minima. We found that for both elastase and thermolysin, solvent occupancy grids from the 10 runs outperformed simulation occupancy from only 5 runs, where performance was defined as the number of crystallographic probe sites found within the top-ranked hot spots. We focused our examination of hot spots on the top eight probe sites that were well-occupied by simulation density and required that they correspond with most of the crystallographic hot spots to demonstrate success.

3.3.5. Allowing for Conformational Change

There are occasions when a MixMD map may not reproduce the MSCS probe coordinates. In cases where conformational rearrangement of the protein alters the solvent accessible surface area in the binding site, the available interactions may change relative to the original crystal structure. The crystal structure of RNase A was solved such that a salt bridge was formed between Asp121 and a neighboring residue in a symmetry partner. When MD simulations were performed in the absence of the crystalline environment, this salt bridge no longer existed and Asp121 shifted towards the active site. This caused the binding cavity to narrow, with a new distance of 10.2 Å from Val43-

Asp121 compared to an original distance of 11.6 Å. However, the interactions available at these sites was altered over the course of simulation due to a median RMSD shift of 4.54 Å from the crystal structure, according to the final 5 ns from 5 independent simulations of 50-ns each for RNase A+IPA.

RNase A binds and cleaves RNA substrate in a deep cleft flanked by the two catalytic residues, His12 and His119, through a transition state complex that is stabilized by the presence of three basic residues: Lys9, Lys41, and Lys66. The MSCS of RNAse A (chain B) had coordinates for two IPA probe molecules at the active site, each with B-factors over 60 Å. IPA.902 was principally associated with His12 and Thr45, while IPA.903 interacted with His12 through a bridging water (Wat921). The refined map of 3EV2 from Buster, obtained using the original PDB structure with the IPA molecules removed individually, showed that the density for IPA.902 was more compelling as a probe site than the density for IPA.903. We found that the F_O-F_C density of the active site indicated a justified probe site at IPA.902B and the expected density for a water molecule at IPA.903B.

The maximally occupied site for probe occupancy from simulation IPA density was located in a deep pocket next to Asp121 (Figure 3-7), where the probe was able to form a stable hydrogen-bonding interaction due to the conformation shift that occurred in 4 of the 5 simulations. The second site identified by probe density overlapped with the position of a cytosolic RNase inhibitor bound across the active site (PDB ID 3MWQ[240]). Without experimental confirmation of the structural changes observed in our simulation, it is not possible to judge the validity of these findings.

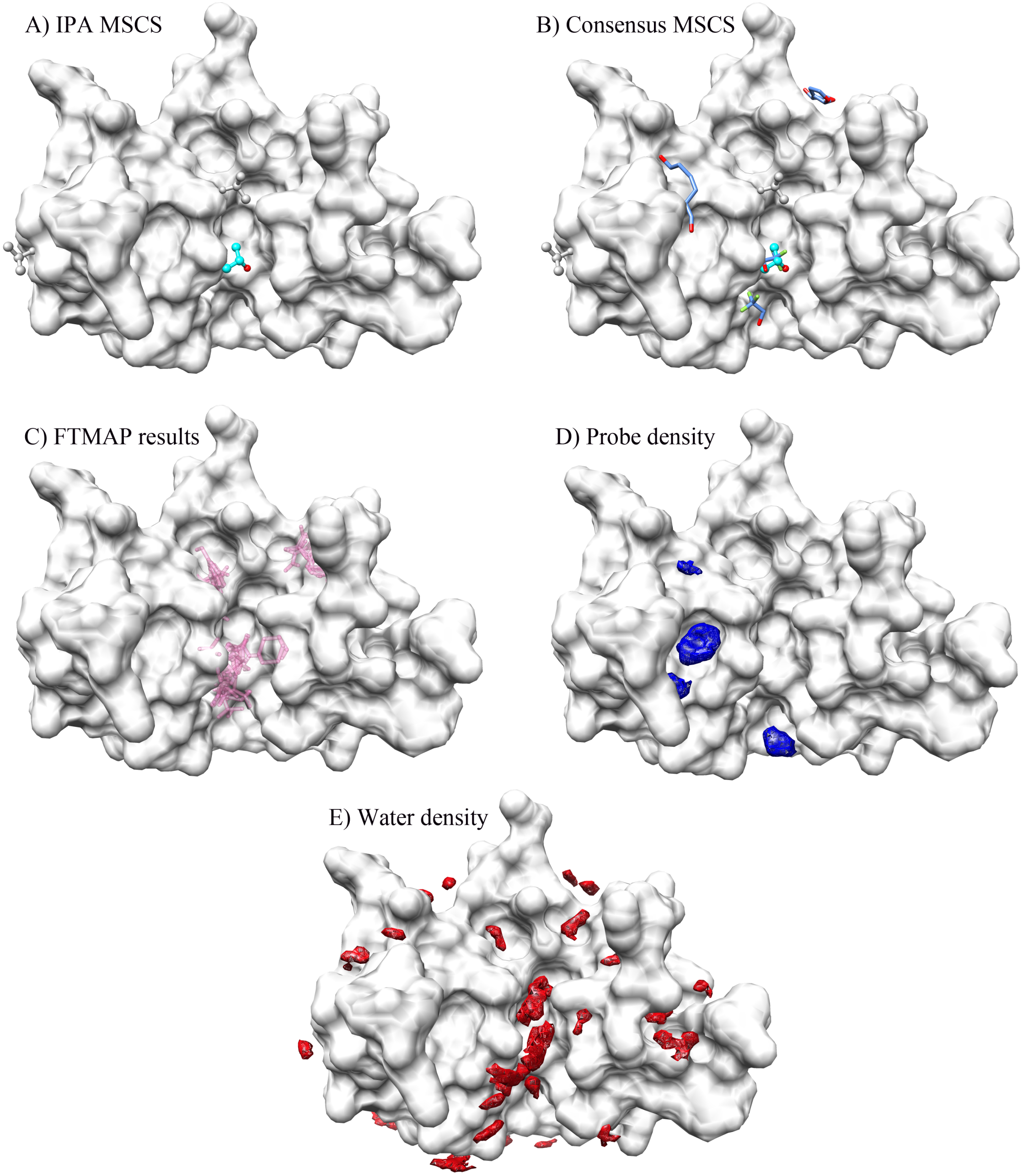


Figure 3-8: MixMD data combined for the last 5 ns of 5 individual 50 ns runs of fully flexible RNase A in 50% w/w IPA (blue) and water (water). All snapshots were superimposed on the crystal structure (PDB ID 3EV2, white surface). Crystallographic waters are colored black for B-factors < 33 Å and in yellow for B-factors > 33 Å. Water molecules that were identified along the crystallographic boundary are not shown. The MSCS coordinate positions for IPA.905A, IPA.917A, IPA.902B, and IPA.903B are shown in cyan as ball-and-stick. MSCS with other functional groups obtained by the same authors are shown in blue, FTMap sites are shown in pink. A) The MSCS probe sites from 3EV2 chain B. B) The consensus MSCS probe sites with a variety of different solvent probes. C) The cluster sites from performing FTMAP against chain B of 3EV2. D) The simulation density for the probe occupancy over the final 5 ns of the combined runs. E) The simulation density for the water occupancy of the final 5 ns of the combined runs.

We found that our occupancy grid for water mapped the binding site, including the position of IPA.902 and a secondary binding site between IPA.902B and IPA.903B, which was well-positioned to take advantage of contacts to His12 and Lys41. Over the course of our MD simulation, the conformational shift caused by movement of Asp121 caused the β-sheet that contained His119 to move downward, which oriented His119 into the position occupied by Wat921 in the crystal structure. This affected mapping of the binding site, and the secondary site is likely the equivalent of the density assigned to IPA.903B in the crystal structure. Our combined solvent occupancy (IPA+water) mapped the bound position of IPA.917A of 3EV2, which lies along the protein surface in a shallow binding pocket near Thr60. Only sixteen of the crystallographic water positions were within 5 Å of the protein and had B-factors of < 33 Å; all but one of these positions was strongly mapped by our water occupancy grids.

These results for RNase A + IPA indicated that a lower concentration of probe may be required for mapping in order to correctly identify the active site. It is unclear based on the available data whether or not the conformation observed in our simulation is valid for this protein system, or whether the solvent concentration has influenced the conformational ensemble towards an unfolded state. As a result, it would be interesting to determine the impact of a lower concentration of solvent on conformational sampling and probe mapping in our simulations.

3.4 Conclusions

Our results demonstrated the utility of MixMD for identifying the hot spots present in a broad range of pharmaceutically-relevant receptor targets. We have shown that both protic and aprotic solvent types can be used in MixMD to specifically model the consensus sites for binding as a complementary technique to experimental FBDD. Evaluation of the F_O-F_C omit density generated based on the refined structure with probe(s) removed showed the importance of thoroughly examining the experimental data in a test set to ensure fair and accurate comparison. In all cases, true hot spots in the crystal structure were located by the simulation density. The additional hot spots that were identified by solvent occupancies from MixMD were judged to be reasonable based on other crystal structures, the surrounding binding surface, and analysis of the crystalline interface. Our study has determined the optimal simulation length and run number for mapping probe sites within overabundant local minima. Probe concentration should be explored in order to determine its influence on true conformational change or partial unfolding.

We have stressed the importance of performing careful validation studies when developing an approach to solvent mapping like MixMD. We have examined both the optimal simulation length and number needed for broadly applying MixMD studies to diverse protein cases. Our study found that an analysis of computational solvent mapping that is based on simulation density *can* result in identification of binding sites that are not overshadowed by overabundant local minima.

This work is in preparation for submission: Lexa KW and Carlson HA. **2011**.