CONFORMATIONAL SAMPLING OF BINDING POCKET AND PREDICTING BINDING FREE ENERGIES

By

Nupur Bansal

A DISSERTATION

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

Chemistry—Doctor of Philosophy

2018

ABSTRACT

CONFORMATIONAL SAMPLING OF BINDING POCKET AND PREDICTING BINDING FREE ENERGIES

By

Nupur Bansal

In order to correctly predict protein-ligand binding poses and free energies, it is essential to accurately take into account receptor flexibility. However, incorporating it into even the smallest region, for example, the binding site of a protein is computationally demanding. Even if this task can be accomplished there is a risk of running into false positives due to the enormous conformational space involved. So, it is the interplay between sampling and scoring. Nonetheless, to better mimic experiments, structure-based drug design methods need to identify and then incorporate the most populated receptor states in any docking and scoring campaign.

This work addresses the development of a novel tool that has been implemented to incorporate receptor flexibility into the ligand-binding domain of a protein. This method enumerates conformational states on an energy landscape in a computationally tractable manner. The algorithm treats molecules at an atom pair level and uses a distance-based coordinate system, where each selected distance is associated with a pair-potential value selected from a look-up table. With a collection of conformations in hand, we then perform on-the-fly local partition function estimations on each of the "seed structures" using the Movable Type (MT) method to estimate the associated free energy changes. This strategy helps to simultaneously generate relevant structures with the most favorable free energies. We initially applied our side chain flexibility method to a set of 159 protein-ligand systems and the docking score of Glide and our

in-house Movable Type based scoring improved over the crystal docking score. Later we applied it to study the active site loop transitions seen in the Streptavidin-biotin system.

To my wonderful parents for their unconditional love and unwavering support

ACKNOWLEDGEMENTS

I convey my sincere thanks to my Ph.D. advisor, Professor Kenneth M. Merz Jr. for giving me an opportunity to learn from him. He is a great mentor who taught me a good deal about research and life too. His guidance correctly molded my scientific outlook and helped me in finishing up my projects. Words are not sufficient to describe my gratitude for him. All I can say is without his support and encouragement; I wouldn't have made this far.

I would like to thank Dr. Zheng Zheng for collaborating with me on my research projects and guiding me. His mentorship and friendship helped me in learning things in a much easier way.

I would also like to thanks my committee members Professor Katharine C. Hunt, Professor Robert I. Cukier, and Professor Heedeok Hong for giving me numerous suggestions on my projects, taking out their time to attend meetings, providing valuable feedback and comments.

Special thank goes to former and current Merz group members whose affection and non-scientific discussions made my Ph.D. life really easy. Thank you Doris, Jun, Lin, and Anthony! I would also like to thank Dr. Nihan M. Ucisik, Dr. Mona Minkara, Dr. Pengfei Li, Dr. Arkajyoti Sengupta and Dr. David Cerutti for answering numerous questions (scientific and otherwise) and their kind friendship. I would extend my thanks to Dr. Dhruva Chakravorty for helping me in the initial phase of my research life.

I would also like to thank HPCC facility at Michigan State University and HPC facility at the University of Florida for providing me with the computational resources. I would also like to thank the chemistry staff for answering my endless list of administrative questions and taking care of my needs very efficiently.

I would like to extend special thanks to my Master's advisor, Dr. Neelanjana Sengupta for teaching me basic principals and ethics of doing research. I cannot thank her enough.

Countless thanks to my parents for keeping their faith in me and loving me unconditionally. I thank my brother and sister for always being there for me. They are my support system and have always stood by me. I would like to thank their spouses and my two lovely nieces. I sincerely thank God for giving me such a nice family and this opportunity in life. A big shout out to Prashant for sticking with me through thick and thin. Last, but definitely not the least, I would like to thank my crazy friends for their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AMBER	Assisted Model Building with Energy Refinement
CADD	Computer Aided Drug Design
FF	Force Field
FEP	Free Energy Perturbation
FES	Free Energy Surface
GA	Genetic Algorithm
GAFF	General Amber Force Field
LJ	Lennard-Jones
MC	Monte Carlo
MCI	Monte Carlo Integration
MD	Molecular Dynamics
MM	Molecular Mechanics
MT	Movable Type
MT_{Flex}	Movable Type with Flexible side chains
$MT_{Flex\text{-}B}$	Movable Type with Flexible backbone and side chains
MUE	Mean Unsigned Error
NMR	Nuclear Magnetic Resonance
PBC	Periodic Boundary Conditions
PDB	Protein Data Bank
PME	Particle Mesh Ewald
PMF	Potential of Mean Force

QM	Quantum Mechanics
RESP	Restrained Electrostatic Potential
RMSD	Root Mean Squared Deviation
TC	Thermodynamic Cycle
TI	Thermodynamic Integration
TS	Transition State
US	Umbrella Sampling
VDW	Van der Waals
WHAM	Weighted Histogram Analysis Method

CHAPTER 1

Introduction

†Adapted from Bansal, N.; Merz, K. M., *Introducing receptor flexibility in protein-ligand binding associations.* Manuscript in preparation.

This dissertation aims to discuss the novel in-house tools developed to incorporate receptor flexibility in the active site of the proteins, which we call MT_{Flex} for including side chain flexibility and $MT_{Flex\text{-}B}$ for including both side-chains and backbone flexibility. The first chapter provides a brief introduction of the current methods used to introduce flexibility in the binding site of the protein in the protein-ligand binding mechanism. It provides a retrospective and prospective view of the available tools and their performances in the blind challenges hosted by the community. The second chapter describes the development of both MT_{Flex} (side chain flexibility) and $MT_{Flex\text{-}B}$ (side-chains and backbone flexibility) methods and the procedure to calculate the free energies to predict the binding affinities and conformational free energies. The third chapter entails the application of MT_{Flex} on the active site of 159 protein-ligand complexes. The fourth chapter presents the application of $MT_{Flex\text{-}B}$ on the eight residue long loop of Streptavidin from the streptavidin-biotin complex. Finally, the fifth chapter summarizes the findings of our work and discusses the future prospects.

1.1 Overview

In order to correctly predict protein-ligand binding poses and free energies, it is essential to accurately take into account receptor flexibility. However, incorporating it into even the smallest region, for example, the binding site of a protein, is computationally demanding. Even if this task can be accomplished there is a risk of running into false positives due to the enormous conformational space involved. This thesis addresses computational approaches that have been implemented to incorporate receptor flexibility into the ligand-binding domain of a protein. The introduction is not limited to receptor flexibility methods used in molecular docking studies. Prospective validation of several docking and other comprehensive tools has also been discussed

largely within the scope of blind challenges conducted by the D3R and CSAR organizations. Based on our analysis to date, we conclude that the community is moving forward by fine-tuning several computational approaches but that the statistical uncertainties in the sampling and scoring accuracy still need to be improved.

The plasticity of proteins is closely associated with several important molecular recognition processes including protein-protein and protein-ligand binding events, enzyme catalysis, allosteric control, and bio-molecular assembly[1]. Conformational changes linked to ligand binding have been characterized by the induced fit and conformational selection models[2]. The induced fit model states that external perturbations like ligand binding induces conformational changes in the receptor forcing it to the *holo* conformation while the conformational selection model proposes that the ligand selectively picks one of the conformations out of a pool of a pre-existing conformational ensemble[1]. Both of the theories are plausible for different systems and have been supported experimentally[1]. Several experiments suggest that both models play a significant role in the protein-ligand binding process[3, 4].

These peculiarities of the mechanistic details of protein-ligand binding, pose an extremely difficult yet interesting challenge for the computer-aided drug discovery (CADD) community. The problem is further complicated by the enormous conformational space available to proteins, which increases the chances of encountering false positives. Using only available crystal structures offers a simple solution to the problem but at the cost of adding inaccuracies in binding affinity prediction associated with the lack of inclusion of receptor flexibility[5]. Indeed, considering only the crystal structure is fundamentally incorrect as proteins are inherently flexible and undergo a variety of conformational changes ranging from vibrational fluctuations to

large-scale domain motions upon ligand binding. Some motions are so large in both the *apo* and *holo* states that the typical time scales used in *in silico* studies are too short to capture them. The importance of receptor flexibility in the CADD field has been emphasized multiple times[1, 6-11]. Accurate prediction of binding free energies and binding mode prediction rely heavily on the accurate accounting of receptor flexibility[11]. To this end, it is crucial to incorporate both ligand-induced and inherent receptor flexibility when estimating protein-ligand binding affinities[2, 11].

This chapter presents an overview of the computational methods available to explore receptor flexibility as a result of protein-ligand binding in the field of structure-based drug design (SBDD). There are several reviews that focus only on receptor flexibility in terms of molecular docking[12-18], but this introduction will cover the general methods used to model receptor flexibility. Our emphasis will be on methods exploring receptor flexibility in the ligand-binding region of proteins. As proteins undergo a variety of motions, different levels of flexibility modeling are required for different protein target systems. For instance, His64 in Human carbonic anhydrase II is the only residue found in alternate conformations on binding with three very similar inhibitors thereby accounting for variations in binding affinity[19]. While kinases exhibit flexibility in terms of loop rearrangements and large lobe motions delimiting the active site region[13]. To include flexibility in drug design applications simulation-based algorithms and docking based methods have been reported. However, these two approaches are not necessarily mutually exclusive. Multiple docking methods use MD or MC simulations or their variants to generate an ensemble of protein conformations and then perform ensemble based docking. Similarly, some simulation-based algorithms perform docking as the first step to obtain the optimum protein-ligand bound conformation and then generate a conformational ensemble.

Although it is quite difficult to segregate the various approaches, we have tried to categorize methods into docking based, simulation-based or other. Methods based on docking have been categorized based on the level of receptor flexibility employed, which are then further parsed into sub-sections. A section is devoted to the assessment of prospective validation of emerging computational methods utilizing receptor flexibility.

1.2 Receptor Flexibility in Docking Methods

1.2.1 Implicit flexibility

Several docking methods incorporate receptor flexibility without including it explicitly. One way of introducing implicit receptor flexibility in docking algorithms is by enlarging the binding pocket. The simplest way to accomplish this is by lowering the repulsive part of the potential energy function using a soft docking approach[20]. This approach qualitatively addresses very small conformational changes (up to 1Å) in the binding pocket region of the protein. Fundamentally this method works by reducing the penalty to binding afforded by the repulsive term in the Lennard Jones[20]. For example, Ferari *et al* applied this approach in their virtual screening study of T4 lysozyme and aldol reductase and found it afforded better results[21]. The advantage of this strategy is that it is computationally efficient but it can only include very subtle conformational changes limited to one or very few side-chains in the proteins.

Several variants of the soft docking approach have been employed recently where the "softened" van der Waals (vdw) term is combined with structural refinement. Soft docking has

also been coupled with Monte Carlo minimization to include both receptor and ligand flexibility[22]. Mizutani *et al* enlarged the protein cavity by offsetting the vdw radii and then optimized the structure in the subsequent step[23]. There are more advanced hybrid approaches involving soft docking as the first step in a docking campaign and these are discussed further below[24, 25]. Another simple way to enlarge the binding pocket area is by mutating the amino acid side chains to alanine (so-called "Alanine scanning")[26] which has the effect of enlarging the binding pocket. In this way the energy landscape becomes "smoother" resulting in an enhancement of the conformational sampling of the ligand. This approach is quite fast but has the drawback that it may result in false positives requiring further refinement[24].

1.2.2 Side chain and limited backbone flexibility

The methods listed in this section handle larger conformational changes but are still localized, as flexibility is localized to a few select residues in the binding site of the protein. To obtain a sense of the scope of the problem, consider a typical ligand-binding site for a drug like molecule, which is generally delineated by twelve to twenty amino acid side chains thereby approximating to dozens of rotatable torsions. In this class of methods the side chains undergo rearrangement with only limited or no backbone flexibility. A number of methods have been devised to explore this class of partial flexibility using both induced-fit and conformational selection approaches. An example is illustrated in Figure 1.1 to illustrate the level of flexibility introduced by incorporating side chain flexibility in active site residues. The active site displayed belongs to PDBID: 2XDX with the bound ligand highlighted in yellow color. The flexible side-chains (shown in pink, tan, purple, *etc*.) are superimposed on the crystal active site. Incorporating side chain flexibility is less computationally expensive and can be treated independently while backbone flexibility typically cannot[27]. For some target systems side chain flexibility is

sufficient to represent conformational changes induced by ligand binding, but for some not. To address this methods that incorporate full side-chain with limited backbone flexibility have been introduced. To reduce the computational expense of this model it is typically applied only to a few targeted residues in the binding pocket to explore their available conformational space. This category of method can be further binned into discrete (rotamer library) and continuous approaches to incorporate partial receptor flexibility.

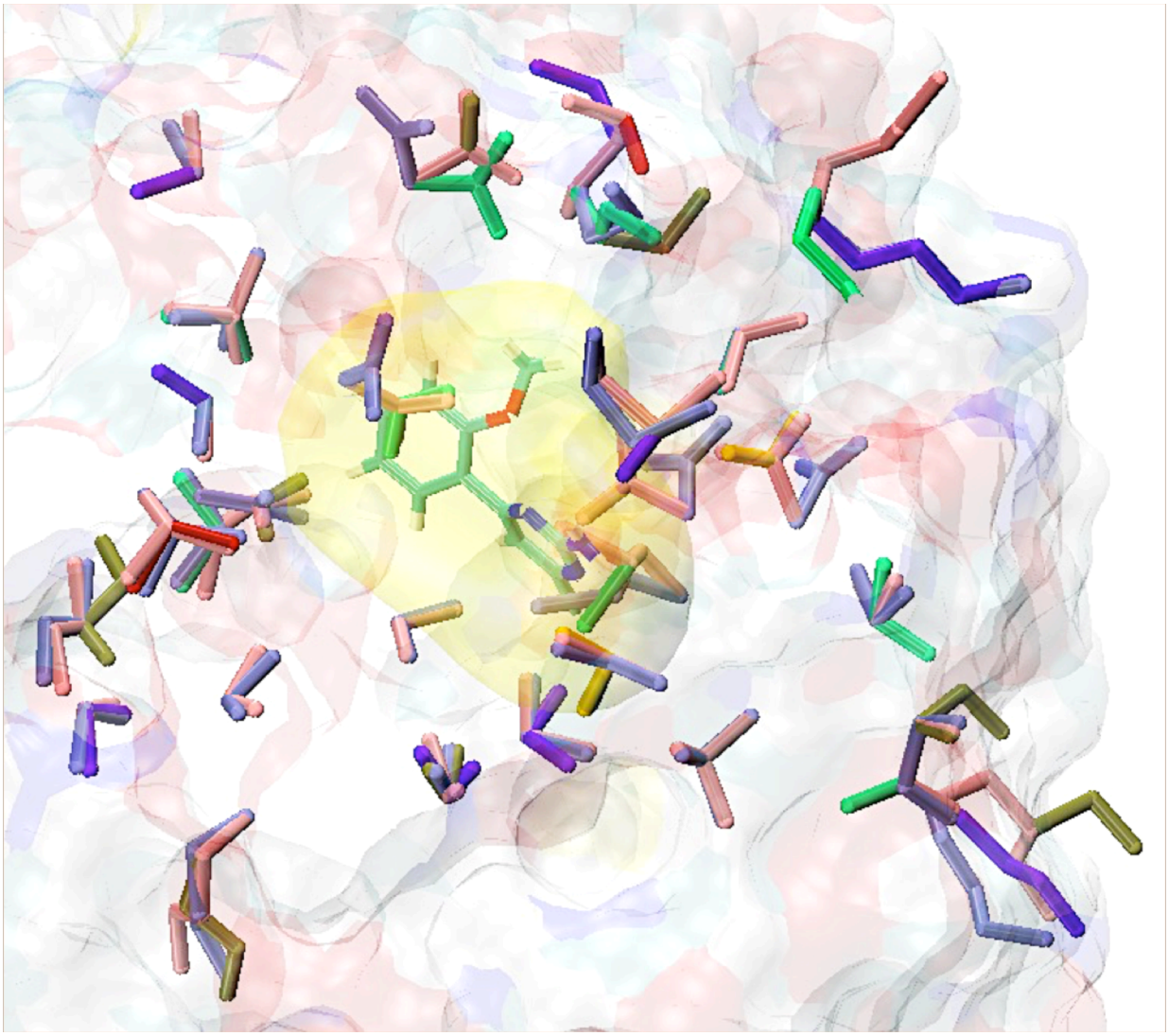

Figure 1.1 A model example to show side-chain flexibility in the active site residues of a protein (PDBID: 2XDX) with the ligand bound. The residues of the crystal active site are shown in ice blue color with the ligand highlighted in yellow background. The flexible side-chains (shown in pink, tan, purple, *etc.*) are superimposed on the crystal active site.

1.2.2.1 Rotamer Library exploration

To explore side chain rotatable bonds, rotamer libraries are generated that contain collections of experimentally preferred rotameric states (i.e. low energy conformations) of amino-acid side chains. Leach developed the side chain specific rotamer library in order to explore the conformational space of the side-chains of the protein receptor. Given a specific orientation of ligand, the global energy minimum conformation of the side chain was found within the provided energy cutoff. He observed that most of the side-chain conformations had no concerted motions and were independent from each other[28, 29]. The large solution space offered by the flexible side-chains was explored using A* and dead-end elimination. Hartman *et al.* also explored discrete rotamer libraries for generating side-chains conformations in the binding site region and used a knowledge based scoring function called ROTA for scoring the protein-ligand interactions. They incorporated multiple side chain conformations of protein and investigated its effect on the docking efficiency by validating it against the screening DUD[30] database[31]. ROSETTADOCK uses a backbone dependent rotamer library boosted with additional side-chain dihedral angles χ_1 and χ_2 rotamers. In their iterative procedure, the side-chain conformations are optimized via first substituting, at each position, the allowable rotamers and then using a quasi-Newton minimization for refinement of torsion angles. Flexibility is incorporated in all the amino acids up to the second shell of ligand-binding region[32].

Schumann *et al.* employed rotamer libraries to generate the receptor side chain conformations and then used the cheapest-path algorithm to achieve a fast and thorough optimization of the receptor structures[33]. MedusaDock uses STROLL (stochastic rotamer library of ligands) and simultaneously models flexibility of both the receptor side-chains and the

ligand molecule[34]. The Fleksy algorithm generates side-chain conformers at selected binding site residues using a backbone-dependent rotamer library[35]. It employs "interaction sampling" to explore the ambiguous orientations of the Asn, Gln and His side chains, different tautomers of His, and the rotational freedom of the thiol group in Cys and the hydroxyl groups in Ser, Thr, Tyr. This interaction sampling approach alleviates the problem of managing protonation states and the handling of different side chain variations observed by the FlexE approach. GalaxyDock pre-defines the residues within the binding site region and accounts for their flexibility by using global optimization. The side-chain flexibility version1 of GalaxyDock performed better than RosettaLigand and SCARE when flexible residues were pre-specified[36]. GalaxyDock2 uses the same energy function for scoring as GalaxyDock but constructs side-chains more quickly using Voronoi diagrams of protein atoms[37]. Both the versions of GalaxyDock estimated the RMSD of the predicted ligand within 2 Å as compared to the experimental structures for ~80-87% of the cases. This performance was comparable to several other leading docking algorithms such as AutoDock (versions 3 and 4), SCARE, RosettaLigand and FLIPDock[17, 36, 37].

1.2.2.2 Continuous approaches

Rotamer libraries are simple to use and have low computational costs but their main drawback is that the method is discrete. The sampling is restricted and biased by the contents of the rotamer library itself. To address this issue methods have been developed which explore selected rotatable bonds beyond simply using discrete rotamer libraries. For example Specitope and Slide are docking tools developed in the Kuhn lab[38-40], which carry out very fast and low-resolution docking using ligands from a large database. They use soft-docking approach to dock the ligand into the protein receptors. Both the receptor and ligand flexibility is incorporated

during the post-docking optimization step by rotating single bonds of either the ligand and the protein side-chain in order to resolve clashes. Rotations are incorporated using Mean-field theory such that there is an improvement in shape complementarity and the collisions are resolved.

ICM (internal coordinate modeling) performs global energy optimization of active side chains and small molecules together using a "double energy" Monte Carlo minimization procedure. It is called double energy scheme as the energies calculated during local minimization and after minimization are both used in the Metropolis selection criteria. In this procedure, torsion angles of the ligand and protein side-chain within 7Å of the binding site are randomly changed. The receptor conformations are generated using a biased probability Monte Carlo procedure (BPMC)[41]. Using a continuous probability distribution function for a given conformational subspace (e.g., side-chain torsional angles or phi-psi), the method chooses a new random position completely unbiased of the previous position which is then locally minimized in the torsional angle space[42].

The mining minima method offers no limitation on the number of rotatable bonds to be considered as continuous degree of freedom[43]. It is computationally expensive because it calculates the configurational integral of the protein-ligand complex. The user defines the binding site region (also called "the live set"), which is considered flexible while the other part of receptor ("real set") is held rigid and untreated. Initially only side-chains were made flexible, but in subsequent versions limited backbone flexibility was also included[44]. The energy optimization is carried out simultaneously for the flexible side chains and the ligand. Conformation search algorithm is used to search the local minima for different binding modes. The binding affinity predictions for 24 HIV-1 protease inhibitors and 20 inhibitors of phosphodiesterase 10a were successfully studied with this method[44]. Within the family of

continuous approaches, methods can be further categorized as on-the-fly (during the docking process) and ensemble based (pre-generation of conformational ensemble) methods.

1.2.2.2.1. On-the-fly methods

SCARE (SCan Alanines and Refine) is an induced fit docking protocol in which the algorithm scans the neighboring side-chains pairs and mutates them to Alanine to fit the ligand[26]. Afterwards, the pocket residues are completely optimized for side-chains with limited backbone flexibility. The algorithm does not rely on the knowledge of location of the binding pocket, geometry of binding ligand or the extent of flexible receptor regions. Induced fit docking (IFD) uses a softened potential to allow for modest steric clashes in the first round of Glide docking[45, 46] with a rigid receptor and then PRIME is used to refinement the protein[24].

FlipDock searches the conformational space by using a divide and conquer approach combined with a very powerful genetic algorithm (GA). The conformational space of both the ligands and receptors is represented by using Flexibility Tree (FT) database[47]. Flexibility Tree (FT) is a tree-like computational data structure, which allows for the hierarchical and multi-resolution encoding of sub-spaces of a protein's conformational space[48]. The FT structure combines and nests a wide variety of motions such as shear, screw, hinge, twist, rotameric side chains, normal mode and essential dynamics in a straightforward manner.

POPSS (Pose Prediction using shape similarity) method, as the name suggests uses the shape similarity approach to place the ligand optimally in the receptor structure. After the ligand is placed, side-chain repacking is performed followed by Monte Carlo minimizations to refine the docked complex[49].

1.2.2.2.2. Ensemble Based

FlexE docking method relies on the united protein description obtained from the superimposed structures of the ensemble[50]. It implies that upon superposition, similar parts of the structures are merged together while dissimilar parts stand out and can be treated as separate alternatives. The concept is quite similar to the rotamer library approach. All atoms are selected within 6.5 Å of any part of the ligand within the binding pocket to consider flexibility. The initial structures were taken from PDBs but in principle can be taken from MD simulation, rotamer libraries or homology modeling. The algorithm is fast as multiple receptor conformations from the ensemble are treated simultaneously and sometimes even new structures are formed by the combination of several structures. The flexibility is incorporated by selection of a combination of partial structures (from the provided input structures) suited best for a particular ligand based on the scoring function. It considers full side chain flexibility and even loop flexibility to some extent. Although it is an ensemble-based approach, the protein flexibility is incorporated during ligand placement stage and not during ligand optimization. A problem of handling different side chain variations and protonation states was observed in FlexE algorithm[51]. It is able to perform very well for the side-chain and some small variations in loops. However, any large movements are not predicted accurately. MT_{Flex} algorithm introduced by our group also generates an ensemble of side chain conformations within the 6Å cutoff of the bound ligands on an energy landscape using a pair potential look up table. The best ligand mode predictions for a set of 159 systems were ~1.31 Å deviated from the native pose as opposed to the RMSD of 2.3 Å for the best ligand modes docked in the native structure. Our method calculates free energies by utilizing a Monte Carlo integration (MCI) scheme for simulation of local partition function. It

was observed that apart from pose prediction, scoring also improved for ligand docked in flexible side-chain conformers as compared to the ligands docked in the crystal structure[52]. ALiBERO (Automatic Ligand-guided Backbone Ensemble Receptor Optimization) algorithm incorporates flexibility by generating either the best or a "team" of complimentary pockets best suited for the particular ligand. It takes single or multiple receptor structures as input along with a ligand training dataset. Receptor ensembles are created iteratively followed by the virtual screening and MC refinement until they converge to the fitness function. The complimentary pockets (either best or the team) are then selected based on the ligands provided[53]. AutoDock is a widely used docking algorithm that allows specifications of flexible side-chains. It was successfully applied to study the conformational variability in the HIV protease's binding site by several side chains reorganization and a water molecule. However, the problem persists as it has a hardcoded limit of including flexibility in ~32 rotatable bonds which is easily surpassed when the side chains in the receptor are made flexible[54]. AutoDockFR (AutoDock for flexible Receptors) supersedes FlipDock by using a very powerful and efficient genetic algorithm along with the advanced motion descriptors of FT for simulating partial side chain flexibility. The novel genetic algorithm can allow explicit side chain flexibility for upto ~14 residues in the binding site region as well as very limited backbone flexibility[55]. Higher docking success rates were achieved by using AutoDockFR and implementing receptor flexibility in the binding site region of the experimentally determined *apo* receptor conformations.

1.2.3. Large conformational changes and loop rearrangement

MD or MC simulations provide a very detailed representation of molecular flexibility. However, large domain motions occur beyond the time scales of current conventional MD

simulations. The flexibility of DFG loops in Kinases was implemented using mean field approach. DFG-in loops of various kinases were converted to type-II bound state (or DFG-out state) using a general deterministic modeling protocol[56]. DOLPHIN models were prepared from DFG-in known structures by removal of all the atoms of Phenylalanine residues on the DFG loop and the consecutive 4 residues following Phe. In the second step, the side-chain atoms of removed Phe and the backbone atoms of the other removed residues except for Gly were used to generate pharmacophore like field. These models represent the average physico-chemical profile of the loop. The density map of the DOLPHIN model was combined with the standard ICM receptor maps for docking.

These models performed exceptionally well in both ligand docking and in-silico activity profiling when validated against a kinase-ligand benchmark database. A large set of methods has been developed using ensemble approach.

1.2.3.1. Ensemble approach

Several methods attempt to use multiple low energy loop conformations in docking. These methods usually dwell on the conformational selection model and aim to mimic conformational changes that occur in the protein region upon ligand binding[57]. The conformational selection model assumes the pre-existence of all the protein conformations in solution and states that upon ligand binding, the *holo*-like conformation is stabilized. The encoding hypothesis is an extension of the conformational selection model, which states that the *apo* form of protein encodes all the necessary fluctuations critical for the ligand bound *holo* conformation[2, 58-60]. Such methods rely on collective degrees of freedom where most atoms

move relative to each other. Collective degree of freedom captures only the dominant motions of protein derived from the native degrees of freedom[15].

1.2.3.1.1. Collective degree of freedom

Collective degrees of freedom enable modeling of huge protein motions as they capture conformational variations in terms of changes in all or part of native degrees of freedom. These methods assume that perturbations in the *apo* form of the protein such as ligand binding can lead to the *holo* form and all these essential protein fluctuations/perturbations are encoded in the collective degrees of freedom. Ikeguchi and co-workers used linear response theory where they modeled ligand binding as an external perturbation to the *apo* form of the protein. They applied their method to the ferric binding protein, F1-ATPase and citrate synthetase and observed that large-scale changes between the *apo* and *holo* forms of proteins (up to 15 Å Cα) was predicted on the basis of 5 collective modes[61]. However, some portions of the predicted *holo* structures differed from the experimental structure by 2-5 Å. Normal mode analysis (NMA) was used by Cavassato *et al.* on the *apo* form of cAMP dependent kinase to model the loop flexibility coupled to ligand binding[62]. Normal modes are Eigen vectors of the hessian matrix and have been shown to represent most motions of the protein at low frequency (< 30 cm^{-1})[62-66]. To increase the computational efficiency, the collective variables of NMA derived from an elastic network model (ENM) have also been used[67-69]. In ENM, protein structures are modeled as elastic networks in which amino acids are represented by Cα atoms and the Cα pairs are connected by uniform springs within a pre-defined distance cutoff[70]. The connectivity matrix of inter-residue contacts is used to derive vibrational vectors and frequencies, which is used to assess the protein flexibility. Dietzan *et al.* investigated the applicability of Cα-ENMs normal modes for the

binding-pocket region in protein-small molecule docking[71]. Using low-lying normal modes from *apo* structures, they reproduced the Cα trace of *holo* partner proteins for the 433 *apo/holo* pairs contained in the Astex[72] data sets. They then assessed the docking capability based on the number of modes used to represent the *holo* structure. However, they found that even for cognate docking, the use of NMA was limited and it was difficult to find a generalized rule to define the number of necessary modes. An alternative approach to NMA is essential dynamics (ED). In ED, atomic coordinates are used to build the covariance matrix and a principal component analysis of this matrix yields the eigenvalues (squared magnitude) and eigenvectors (principal components), where the top principal component corresponds to the most significant conformational change. In contrast to NMA, PCA does not rest on the assumption of a harmonic potential[73]. Cukier *et al.* studied the adenylate kinase by applying PCA analysis to its *apo* structure[59]. As the substrates AMP and Mg^{2+}-ATP bind, the LID (ligand) and AMP binding domains of adenylate kinase undergo huge conformational change leading to the closed form of the binding pocket. With the help of PCA analysis, they discerned the presence of 12 modes that encoded the conformational change of the LID domain upon ligand binding. However, the modes did not encode the motion associated with the AMP binding domain. Their study reinforced the encoding hypothesis as only the modes associated with the ligand-binding domain were encoded in the *apo* form.

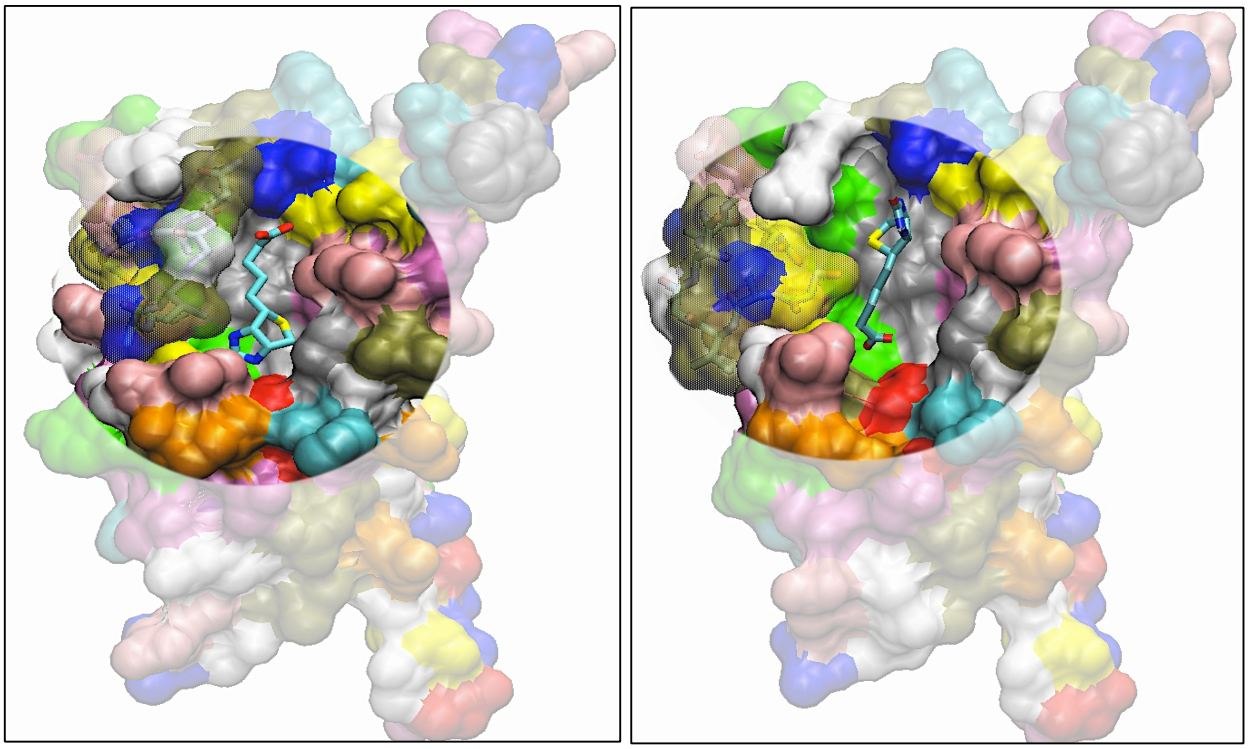

Figure 1.2 The closed (left) and the open (right) state of the $loop_{3\text{-}4}$ of Streptavidin monomer are highlighted. The effect of loop transitioning can be clearly seen on the size of the binding pocket and also on the preferred binding mode of Biotin.

1.2.3.1.2. Native degree of freedom

The $loop_{3\text{-}4}$ in the streptavidin-biotin system was recently studied using $MT_{Flex\text{-}B}$ by Bansal et al.[74] The algorithm generated an ensemble of loop conformations where both *apo* and *holo* forms of the eight-residue loop were generated with the best loop conformation having a RMSD of 1.6 Å with respect to available crystal structures of the two forms. An example to illustrate the closed and open state of the Streptavidin $loop_{3\text{-}4}$ is shown in Figure 1.2. The opening and closing of the loop is highly affecting the size of binding site as shown in the highlighted region in Figure 1.2. It can also be seen that the preferred binding mode of Biotin is quite different in the open and closed states of loop reinforcing the importance of receptor

flexibility in docking algorithms. Wong *et al* also studied "latch" loops (i.e. loops covering the ligand in the *holo* state and switches to an open conformation in the unbound state)[75]. Their method used a combination of replica exchange molecular dynamics (REMD) simulations with harmonic restraints to sample the loop conformations in an unbound state, followed by clustering to identify conformational substates and, finally performed docking against a representative structure from each cluster. For loops spanning up to 15 residues, it yielded a RMSD of ~2 Å with respect to the known *holo* structures. Although harmonic restraints were applied to limit the sampling of the loop, the method cannot *a priori* determine the number of degrees of freedom required to capture the possible rearrangements. Bansal's and Wong's study both support and provide a computational validation of the conformational selection model.

Flick *et al* implemented a multistage loop reconstruction algorithm where they predefined segments of backbone and changed all the dihedral angles of the backbone simultaneously. A global loop closure search was applied to close the loop and the resulting conformation was then locally minimized using a modified steepest descent algorithm. If after minimization, the loop was closed successfully with the gap < 0.001 nm then the resulting conformation was preserved in the acceptance criteria. This method was applied to study the *apo-holo* transitions in the ATP binding loop of ERK2 kinase and adenosine dependent protein kinase (PKA)[76]. Recently a new algorithm called ED/MD (essential dynamics/molecular dynamics) was developed for generating perturbed ensembles representing ligand induced binding site flexibility[77]. It is a hybrid method that uses collective degrees of freedom to enhance the sampling. The first thirty modes obtained by a PCA analysis of MD trajectories were perturbed in the ligand-binding domain. Perturbations were accounted by Lennard-Jones term and then ED/MD simulations[78]

were carried out. These ensembles led to superior docking performance when compared to single structure or conventionally derived MD ensembles.

1.3 Simulation based methods

Apart from the docking based methods, another way of introducing receptor flexibility is through the use of simulation-based methods. These methods are not necessarily independent of the docking based approach as several docking based methods use simulations to collect an ensemble of conformations before performing docking. Some of the methodologies using simulation based conformation generation methods were already covered in the ensemble based docking section. In this section, we will briefly go over the enhanced sampling approaches and the novel tools employed in pathway-based free energy methods to incorporate receptor flexibility. Several extant review articles discuss simulation based methods[6, 10, 11, 58, 79, 80].

1.3.1 Enhanced sampling

Short MD simulations are known to be susceptible to get trapped in the local minima on a free energy landscape[10, 11]. Recent advances in this area to overcome simulation length shortcomings come by implementation of GPU accelerated computing[81], specialized supercomputers such as Anton[82, 83], or using the cloud[84]. Another solution is to introduce bias into the potential or by using enhanced sampling methods such as metadynamics, umbrella sampling, temperature accelerated replica exchange, Morkov state models, and Hamiltonian based accelerated MD[10, 80]

1.3.2. Receptor flexibility in pathway free energy methods

Pathway free energy methods (alchemical and PMF based) in combination with MD or MC simulations can accurately estimate free energies. Unfortunately, the procedure requires extensive conformational sampling and is computationally demanding. Recently, enhanced sampling methods have been added to standard pathway based free energy methods to facilely subsume conformational changes associated with ligand binding[85-89]. Mobley *et al* used the "confine-and-release" framework in conjunction with umbrella sampling methods to study the binding site of T4-lysozyme involving a conformational change[86]. aMD was coupled with thermodynamic integration simulations to improve free energy convergence[87]. Independent trajectory thermodynamic integration (IT-TI) was used to study the flexible loop regions of the H5N1 avian influenza virus neuraminidase interacting with peramivir[85]. aMD was applied on selective dihedrals of neuraminidase and free energy calculations were performed to obtain the converged free energy of binding[88]. FEP/REST was used to study the apolar cavity of T4 lysozyme L99A and it was shown that binding free energies are sensitive to protein reorganization in the binding pocket region[89].

1.4 Other algorithms

Apart from incorporating conformational sampling in docking and simulation based methods, there are loop prediction algorithms that are based on the conformational selection model. Danielson *et al*. used CorLps program to generate *holo* like protein conformations from the *apo* form of the protein in the presence of the ligand[90]. Loop prediction was performed on the following three systems: GART (a six residue loop segment), CYP119 (two nine residue loop regions), and enolase (an eleven residue loop region)[90]. The energetically favorable loop

conformational ensemble was generated, which was then filtered by clustering and further refined and the remaining top 100 poses were ranked based on the DFIRE scoring function.

1.5 Binding affinity predictions in docking methods

Once the conformational ensemble containing the most relevant structure is generated, the focus is shifted to how one can predict reliable binding affinities. Although great success has been achieved in retrospective pose prediction and virtual screening strategies, binding affinity predictions still pose a daunting challenge[14, 79, 91, 92].

A wide range of scoring function strategies have implemented protein flexibility and achieved some level of improvement in either pose prediction or binding affinity or both[24, 32, 36, 93-99]. Recently, Zheng et al. developed a novel tool for free energy calculations called the movable type (MT) sampling method. It directly calculates the Helmholtz free energy by estimating the local partition function using the principles of statistical mechanics. Bansal *et al*. used this approach in their MT_{Flex} strategy for side-chain flexibility, which allowed for a better prediction of the binding free energy (over 159 protein-ligand systems an RMSE of ~2.72 kcal/mol versus 3.4 kcal/mol using only the crystal structure)[52]. However, one common thread in all of the strategies discussed so far is that all the evaluations are performed retrospectively. Our tools are clearly honed and perfected to perform better when we know the answer *a priori*, but what about the cases where we don't. The real challenge is to evaluate our methods against the prospective targets and this is discussed below.

1.6 Prospective validation of docking methods with receptor flexibility

Retrospective assessments are important but nonetheless suffer from inherent bias. The knowledge of the correct answer is embedded in the construction of the problem, which often leads to skewing of the results. Hence, there is no substitute for making truly prospective predictions[100]. Prospective validations are the gold standard to accurately judge the predictive ability of methods, as they leave no scope for fitting of parameters and other human interventions or biases. In this regard, several blind community challenges initiated by CSAR and further continued by the D3R organizations have been run since 2009 for the prospective assessment of ligand pose prediction and ranking protein-ligand binding affinities[101-105]. Grand challenges (GC) hosted by the D3R organization offered the community a range of prospective validations of docking pose prediction[103, 104]. These challenges included receptor targets with flexible binding pockets.

In the 2015 GC1 challenge, HSP90 and MAP4K4 were the receptor targets chosen to perform docking and binding affinity prediction. All submissions that afforded good binding mode predictions incorporated receptor flexibility in some manner in their modeling efforts. The Camacho lab used Smina (default parameters) and the AutoDock Vina scoring function to dock into HSP90 and MAP4K4[106]. They used multiple receptor/ligand co-crystal structures as binding templates for performing minimization and docking from the known complex structures and obtained ligand RMSDs of 0.32 Å and 1.6 Å for the target systems. POPSS method, which incorporates 3D shape similarity and included side chain repacking and protein minimization, also performed well in the D3R prospective challenge with median RMSD for their top prediction of 0.73 Å and for 2.87Å[107] for HSP90 and MAP4K4, respectively. DockBench tool also performed reasonably well for both the targets. It also used existing crystal structures to

obtain training set to perform ensemble based docking[108]. GRIM graph matching method also predicted the best binding mode with reasonable accuracy by performing docking using several crystal protein structures with and without conserved water molecules[109].

The D3R evaluators established that overall the best pose-prediction methods were less associated with a single docking algorithm, rather with a "similarity docking" approach[104]. Its worth noting that several pose predictions for HSP90 were within 2 Å RMSD which was not the case for MAP4K4. HSP90 is a structurally well-characterized system, while MAP4K4 is less so allowing participants the ability to pre-evaluate the capability of their chosen approach. The MAP4K4 compound collection was also more challenging with ~1/3 coming from a congeneric series while for HSP90 the compounds were all from a congeneric series. MAP4K4 was also simply more challenging due to the flexibility of the classic P-loop found in kinases.

In Grand Challenge2 as well, a rather flexible bile acid receptor target called FXR was used. Totrov and group predicted the lowest mean RMSD (of all the submissions) of 1.95 Å by using a new hybrid ligand/receptor structure-based docking method called LigBEnd[110]. The Cournia group used a combination of docking and physics based methods and attained the lowest median RMSD of 0.99 Å out of all the submissions[111]. Most of the leading docking methods were combined with the knowledge of existing PDB data, which made it quite difficult to discern the performance of individual methods. In fact, when organizers excluded the most similar ligands from evaluations, the performance of the known docking methods became much worse. Literature search for similar compounds proved to be handy except for one of the ligands (FXR34) despite of having a similarity coefficient of 0.83 with the available systems. It happened because the common available known structures for that ligand had appreciable different binding modes and different protein conformations, which misguided the challengers

relying on the known crystal structures[103]. In the first stage of both the Grand Challenges, there were very few methods that behaved consistently well for all the targets. This problem was realized in CSAR 2014 challenge as well where the major task was to find the methods that performed consistently well for all the three targets provided by the organizers[101]. Only 34% of the methods had RMSD of < 2 Å for all the provided targets. A few methods that consistently edged the prediction of binding mode did not do well on the predictions of binding affinity. Ranking of ligands based on affinity was an issue for nearly all the blind challenged conducted so far[101-105]. In GC1, the highest correlation of the submitted rankings with respect to the experimental rankings was quite poor with the values of 0.32 for HSP90 and 0.48 for MAP4K4[104]. The highest Kendall Tau attained for affinity ranking in GC2 was 0.45 despite of the knowledge of crystal structures. The only encouraging aspect observed in Kendall's Tau trend for GC2 was that all the values were positive depicting the lack of randomness in the protocols. Nevertheless, no particular approach was observed to edge both the docking and scoring challenges showing that the results are statistically insignificant. Apart from these blind challenges, prospective testing of cross docking was performed by Jain lab on a set of 10 pharmaceutically relevant targets using a total of 949 ligands[112]. They used a cross docking benchmark dataset "PINC" to follow a knowledge-guided docking protocol. Protein pocket similarity was used to perform ensemble docking using Surflex-Dock. Out of all the predictions, correct poses were identified nearly 90% of the time.

Understanding protein-ligand binding mechanism is a challenging problem and to solve it using a generalized method that handles all target systems equally well has still not been identified. A good example illustrating the problem involves the participation of RosettaLigand with full ligand flexibility and receptor backbone flexibility in theSAMPL1 blind challenge. In

this challenge participants were asked as to dock nearly 100 ligands into the JNK-3 kinase and urokinase-type plasminogen activator. Initially, they did not incorporate full backbone flexibility in their challenge submissions, but reevaluated their results retrospectively where they included full flexibility[113]. The addition of full backbone flexibility benefitted the JNK3 challenge yielding more predictions within a 2 Å RMSD of the native ligand and lowering the structural RMSD on average. However, implementing backbone flexibility for urokinase opened avenues for mistakes since the receptor does not undergo significant conformational changes upon binding. Overall, docking of some compounds fare better but several fared worse than the original submission for the urokinase challenge. The problem of structural flexibility of the binding pocket residues is clearly important, but there are several other issues that pose problems like accounting for conserved water molecules, accurate sampling of ligand conformations, missing residues, solvation models, scoring functions, *etc*. Another major bottleneck in the field is the lack of a plethora of high-resolution experimental protein-ligand structures. Some of the concerns associated with the field of binding mode prediction, affinity estimation and virtual screening are listed here[114] For the more precise estimate of binding affinities, it is crucial to consider the configurational entropy contributed by the conformational substates of proteins, but the conundrum is that these additional components are also adding inaccuracies into the calculated binding affinity. The hard part is to decipher whether the problem lies with the scoring functions or with the sampling strategies or both. In a nutshell, these blind challenges for protein-ligand docking and affinity rankings are helping the community to move forward but to clearly pinpoint the underlying cause of the error in any particular method is still a major issue.

1.7 Discussion

Receptor flexibility is one of the major challenges faced by the CADD community in binding mode prediction and accurate prediction of binding free energies. Recent advancements in the field of SBDD have accounted for receptor flexibility in the binding pocket. Computational methods are being continuously improved as not all methods work for all the protein systems. The computational approaches accounting for receptor flexibility in docking based, simulation based and other protein loop prediction algorithms have been discussed. As most of the groups validate their methodologies retrospectively, we discussed prospective validation of extant methods based on blind challenges hosted by CSAR and D3R. Although we are seeing significant improvement in retrospective evaluations, prospective validations are still statistically challenged. It is difficult to *a priori* anticipate the level of receptor flexibility needed for a particular target system, however it is certain that some level of receptor motion is important to account for the entropy associated with the binding processes[115]. Considering the configurational entropy contributed by receptor conformations is extremely important for the accurate estimation of binding affinities, however probably adding this component is adding more inaccuracies in the free energy calculations. The problem remains with the scoring function or the sampling strategy is hard to pinpoint as of yet.

REFERENCES

REFERENCES

1. Boehr, D.D., R. Nussinov, and P.E. Wright, *The role of dynamic conformational ensembles in biomolecular recognition.* Nature Chemical Biology, 2009. **5**(11): p. 789-796.

2. Lill, M.A., *Efficient Incorporation of Protein Flexibility and Dynamics into Molecular Docking Simulations.* Biochemistry, 2011. **50**(28): p. 6157-6169.

3. James, L.C. and D.S. Tawfik, *Structure and kinetics of a transient antibody binding intermediate reveal a kinetic discrimination mechanism in antigen recognition.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(36): p. 12730-12735.

4. Ma, B.Y. and R. Nussinov, *Enzyme dynamics point to stepwise conformational selection in catalysis.* Current Opinion in Chemical Biology, 2010. **14**(5): p. 652-659.

5. Huang, S.Y. and X.Q. Zou, *Advances and Challenges in Protein-Ligand Docking.* International Journal of Molecular Sciences, 2010. **11**(8): p. 3016-3034.

6. Carlson, H.A., *Protein flexibility and drug design: how to hit a moving target.* Current Opinion in Chemical Biology, 2002. **6**(4): p. 447-452.

7. Nichols, S.E., R.V. Swift, and R.E. Amaro, *Rational Prediction with Molecular Dynamics for Hit Identification.* Current Topics in Medicinal Chemistry, 2012. **12**(18): p. 2002-2012.

8. Teague, S.J., *Implications of protein flexibility for drug discovery.* Nature Reviews Drug Discovery, 2003. **2**(7): p. 527-541.

9. Csermely, P., R. Palotai, and R. Nussinov, *Induced fit, conformational selection and independent dynamic segments: an extended view of binding events.* Trends in Biochemical Sciences, 2010. **35**(10): p. 539-546.

10. Sinko, W., S. Lindert, and J.A. McCammon, *Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design.* Chemical Biology & Drug Design, 2013. **81**(1): p. 41-49.

11. Feixas, F., et al., *Exploring the role of receptor flexibility in structure-based drug discovery.* Biophysical Chemistry, 2014. **186**: p. 31-45.

12. Antunes, D.A., D. Devaurs, and L.E. Kavraki, *Understanding the challenges of protein flexibility in drug design.* Expert Opinion on Drug Discovery, 2015. **10**(12): p. 1301-1313.

13. Cavasotto, C.N. and R.A. Abagyan, *Protein flexibility in ligand docking and virtual screening to protein kinases.* Journal of Molecular Biology, 2004. **337**(1): p. 209-225.

14. Guedes, I.A., C.S. de Magalhaes, and L.E. Dardenne, *Receptor-ligand molecular docking.* Biophys Rev, 2014. **6**(1): p. 75-87.

15. Teodoro, M.L., G.N. Phillips, and L.E. Kavraki, *Understanding protein flexibility through dimensionality reduction.* Journal of Computational Biology, 2003. **10**(3-4): p. 617-634.

16. Yuriev, E., M. Agostino, and P.A. Ramsland, *Challenges and advances in computational docking: 2009 in review.* Journal of Molecular Recognition, 2011. **24**(2): p. 149-164.

17. Yuriev, E., J. Holien, and P.A. Ramsland, *Improvements, trends, and new ideas in molecular docking: 2012-2013 in review.* Journal of Molecular Recognition, 2015. **28**(10): p. 581-604.

18. Yuriev, E. and P.A. Ramsland, *Latest developments in molecular docking: 2010-2011 in review.* Journal of Molecular Recognition, 2013. **26**(5): p. 215-239.

19. Smith, G.M., et al., *Positions of His-64 and a Bound Water in Human Carbonic-Anhydrase-Ii Upon Binding 3 Structurally Related Inhibitors.* Protein Science, 1994. **3**(1): p. 118-125.

20. Jiang, F. and S.H. Kim, *Soft Docking - Matching of Molecular-Surface Cubes.* Journal of Molecular Biology, 1991. **219**(1): p. 79-102.

21. Ferrari, A.M., et al., *Soft docking and multiple receptor conformations in virtual screening.* Journal of Medicinal Chemistry, 2004. **47**(21): p. 5076-5084.

22. Apostolakis, J., A. Pluckthun, and A. Caflisch, *Docking small ligands in flexible binding sites.* Journal of Computational Chemistry, 1998. **19**(1): p. 21-37.

23. Mizutani, M.Y., et al., *Effective handling of induced-fit motion in flexible docking.* Proteins-Structure Function and Bioinformatics, 2006. **63**(4): p. 878-891.

24. Sherman, W., et al., *Novel procedure for modeling ligand/receptor induced fit effects.* Journal of Medicinal Chemistry, 2006. **49**(2): p. 534-553.

25. Venkatraman, V. and D.W. Ritchie, *Flexible protein docking refinement using pose-dependent normal mode analysis.* Proteins-Structure Function and Bioinformatics, 2012. **80**(9): p. 2262-2274.

26. Bottegoni, G., et al., *A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE).* Journal of Computer-Aided Molecular Design, 2008. **22**(5): p. 311-325.

27. Totrov, M. and R. Abagyan, *Flexible ligand docking to multiple receptor conformations: a practical alternative.* Current Opinion in Structural Biology, 2008. **18**(2): p. 178-184.

28. Leach, A.R., *Ligand Docking to Proteins with Discrete Side-Chain Flexibility.* Journal of Molecular Biology, 1994. **235**(1): p. 345-356.

29. Leach, A.R. and A.P. Lemon, *Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm.* Proteins-Structure Function and Bioinformatics, 1998. **33**(2): p. 227-239.

30. Huang, N., B.K. Shoichet, and J.J. Irwin, *Benchmarking sets for molecular docking.* Journal of Medicinal Chemistry, 2006. **49**(23): p. 6789-6801.

31. Hartmann, C., I. Antes, and T. Lengauer, *Docking and scoring with alternative side-chain conformations.* Proteins-Structure Function and Bioinformatics, 2009. **74**(3): p. 712-726.

32. Meiler, J. and D. Baker, *ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility.* Proteins-Structure Function and Bioinformatics, 2006. **65**(3): p. 538-548.

33. Schumann, M. and R.S. Armen, *Systematic and efficient side chain optimization for molecular docking using a cheapest-path procedure.* Journal of Computational Chemistry, 2013. **34**(14): p. 1258-1269.

34. Ding, F., S.Y. Yin, and N.V. Dokholyan, *Rapid Flexible Docking Using a Stochastic Rotamer Library of Ligands.* Journal of Chemical Information and Modeling, 2010. **50**(9): p. 1623-1632.

35. Wagener, M., J. de Vlieg, and S.B. Nabuurs, *Flexible Protein-Ligand Docking Using the Fleksy Protocol.* Journal of Computational Chemistry, 2012. **33**(12): p. 1215-1217.

36. Shin, W.H. and C. Seok, *GalaxyDock: Protein-Ligand Docking with Flexible Protein Side-chains.* Journal of Chemical Information and Modeling, 2012. **52**(12): p. 3225-3232.

37. Shin, W.H., et al., *GalaxyDock2: Protein-Ligand Docking Using Beta-Complex and Global Optimization.* Journal of Computational Chemistry, 2013. **34**(30): p. 2647-2656.

38. Schnecke, V., et al., *Screening a peptidyl database for potential ligands to proteins with side-chain flexibility.* Proteins-Structure Function and Genetics, 1998. **33**(1): p. 74-87.

39. Schnecke, V. and L.A. Kuhn, *Virtual screening with solvation and ligand-induced complementarity.* Perspectives in Drug Discovery and Design, 2000. **20**(1): p. 171-190.

40. Schnecke, V. and L.A. Kuhn, *Database screening for HIV protease ligands: the influence of binding-site conformation and representation on ligand selectivity.* Proc Int Conf Intell Syst Mol Biol, 1999: p. 242-51.

41. Abagyan, R. and M. Totrov, *Biased Probability Monte-Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins.* Journal of Molecular Biology, 1994. **235**(3): p. 983-1002.

42. Totrov, M. and R. Abagyan, *Flexible protein-ligand docking by global energy optimization in internal coordinates.* Proteins-Structure Function and Genetics, 1997: p. 215-220.

43. David, L., R. Luo, and M.K. Gilson, *Ligand-receptor docking with the Mining Minima optimizer.* Journal of Computer-Aided Molecular Design, 2001. **15**(2): p. 157-171.

44. Chen, W., et al., *Modeling Protein-Ligand Binding by Mining Minima.* Journal of Chemical Theory and Computation, 2010. **6**(11): p. 3540-3557.

45. Friesner, R.A., et al., *Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.* Journal of Medicinal Chemistry, 2004. **47**(7): p. 1739-1749.

46. Halgren, T.A., et al., *Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening.* Journal of Medicinal Chemistry, 2004. **47**(7): p. 1750-1759.

47. Zhao, Y. and M.F. Sanner, *FLIPDock: Docking flexible ligands into flexible receptors.* Proteins-Structure Function and Bioinformatics, 2007. **68**(3): p. 726-737.

48. Zhao, Y., D. Stoffler, and M. Sanner, *Hierarchical and multi-resolution representation of protein flexibility.* Bioinformatics, 2006. **22**(22): p. 2768-2774.

49. Kumar, A. and K.Y.J. Zhang, *A pose prediction approach based on ligand 3D shape similarity.* Journal of Computer-Aided Molecular Design, 2016. **30**(6): p. 457-469.

50. Claussen, H., et al., *FlexE: Efficient molecular docking considering protein structure variations.* Journal of Molecular Biology, 2001. **308**(2): p. 377-395.

51. Polgar, T. and G.M. Keseru, *Ensemble docking into flexible active sites. Critical evaluation of FlexE against JNK-3 and beta-secretase.* Journal of Chemical Information and Modeling, 2006. **46**(4): p. 1795-1805.

52. Bansal, N., Z. Zheng, and K.M. Merz, *Incorporation of side chain flexibility into protein binding pockets using MTflex.* Bioorganic & Medicinal Chemistry, 2016. **24**(20): p. 4978-4987.

53. Rueda, M., M. Totrov, and R. Abagyan, *ALiBERO: Evolving a Team of Complementary Pocket Conformations Rather than a Single Leader.* Journal of Chemical Information and Modeling, 2012. **52**(10): p. 2705-2714.

54. Osterberg, F., et al., *Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock.* Proteins-Structure Function and Genetics, 2002. **46**(1): p. 34-40.

55. Ravindranath, P.A., et al., *AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility.* Plos Computational Biology, 2015. **11**(12).

56. Kufareva, I. and R. Abagyan, *Type-II Kinase Inhibitor Docking, Screening, and Profiling Using Modified Structures of Active Kinase States.* Journal of Medicinal Chemistry, 2008. **51**(24): p. 7921-7932.

57. Ma, B.Y., et al., *Folding funnels and binding mechanisms.* Protein Engineering, 1999. **12**(9): p. 713-720.

58. Carlson, H.A. and J.A. McCammon, *Accommodating protein flexibility in computational drug design.* Molecular Pharmacology, 2000. **57**(2): p. 213-218.

59. Cukier, R.I., *Apo Adenylate Kinase Encodes Its Holo Form: A Principal Component and Varimax Analysis.* Journal of Physical Chemistry B, 2009. **113**(6): p. 1662-1672.

60. Lou, H.F. and R.I. Cukier, *Molecular dynamics of apo-adenylate kinase: A principal component analysis.* Journal of Physical Chemistry B, 2006. **110**(25): p. 12796-12808.

61. Ikeguchi, M., et al., *Protein structural change upon ligand binding: Linear response theory.* Physical Review Letters, 2005. **94**(7).

62. Cavasotto, C.N., J.A. Kovacs, and R.A. Abagyan, *Representing receptor flexibility in ligand docking through relevant normal modes.* Journal of the American Chemical Society, 2005. **127**(26): p. 9632-9640.

63. Brooks, B. and M. Karplus, *Normal-Modes for Specific Motions of Macromolecules - Application to the Hinge-Bending Mode of Lysozyme.* Proceedings of the National Academy of Sciences of the United States of America, 1985. **82**(15): p. 4995-4999.

64. Hayward, S., A. Kitao, and H.J.C. Berendsen, *Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme.* Proteins-Structure Function and Genetics, 1997. **27**(3): p. 425-437.

65. Hinsen, K., A. Thomas, and M.J. Field, *Analysis of domain motions in large proteins.* Proteins-Structure Function and Genetics, 1999. **34**(3): p. 369-382.

66. Tama, F., et al., *Building-block approach for determining low-frequency normal modes of macromolecules.* Proteins-Structure Function and Genetics, 2000. **41**(1): p. 1-7.

67. May, A. and M. Zacharias, *Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: Evaluation on kinase inhibitor cross docking.* Journal of Medicinal Chemistry, 2008. **51**(12): p. 3499-3506.

68. Petrone, P. and V.S. Pande, *Can conformational change be described by only a few normal modes?* Biophysical Journal, 2006. **90**(5): p. 1583-1593.

69. Rueda, M., G. Bottegoni, and R. Abagyan, *Consistent Improvement of Cross-Docking Results Using Binding Site Ensembles Generated with Elastic Network Normal Modes.* Journal of Chemical Information and Modeling, 2009. **49**(3): p. 716-725.

70. Tirion, M.M., *Large amplitude elastic motions in proteins from a single-parameter, atomic analysis.* Physical Review Letters, 1996. **77**(9): p. 1905-1908.

71. Dietzen, M., et al., *On the Applicability of Elastic Network Normal Modes in Small-Molecule Docking (vol 52, pg 844, 2012).* Journal of Chemical Information and Modeling, 2014. **54**(12): p. 3453-3453.

72. Hartshorn, M.J., et al., *Diverse, high-quality test set for the validation of protein-ligand docking performance.* Journal of Medicinal Chemistry, 2007. **50**(4): p. 726-741.

73. Amadei, A., A.B.M. Linssen, and H.J.C. Berendsen, *Essential Dynamics of Proteins.* Proteins-Structure Function and Genetics, 1993. **17**(4): p. 412-425.

74. Bansal, N., et al., *The Role of the Active Site Flap in Streptavidin/Biotin Complex Formation.* Journal of the American Chemical Society, Submitted

75. Wong, S. and M.P. Jacobson, *Conformational selection in silico: Loop latching motions and ligand binding in enzymes.* Proteins-Structure Function and Bioinformatics, 2008. **71**(1): p. 153-164.

76. Flick, J., F. Tristram, and W. Wenzel, *Modeling loop backbone flexibility in receptor-ligand docking simulations.* Journal of Computational Chemistry, 2012. **33**(31): p. 2504-2515.

77. Chaudhuri, R., et al., *Application of Drug-Perturbed Essential Dynamics/Molecular Dynamics (ED/MD) to Virtual Screening and Rational Drug Design.* Journal of Chemical Theory and Computation, 2012. **8**(7): p. 2204-2214.

78. Carrillo, O., C.A. Laughton, and M. Orozco, *Fast Atomistic Molecular Dynamics Simulations from Essential Dynamics Samplings.* Journal of Chemical Theory and Computation, 2012. **8**(3): p. 792-799.

79. Kokh, D.B., R.C. Wade, and W. Wenzel, *Receptor flexibility in small-molecule docking calculations.* Wiley Interdisciplinary Reviews-Computational Molecular Science, 2011. **1**(2): p. 298-314.

80. Rocchia, W., M. Masetti, and A. Cavalli, *Enhanced Sampling Methods in Drug Design.* Physico-Chemical and Computational Approaches to Drug Discovery, 2012(23): p. 273-301.

81. Buch, I., T. Giorgino, and G. De Fabritiis, *Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(25): p. 10184-10189.

82. Shan, Y.B., et al., *How Does a Drug Molecule Find Its Target Binding Site?* Journal of the American Chemical Society, 2011. **133**(24): p. 9181-9183.

83. Scarpazza, D.P., et al., *Extending the generality of molecular dynamics simulations on a special-purpose machine.* Ieee 27th International Parallel and Distributed Processing Symposium (Ipdps 2013), 2013: p. 933-945.

84. De Paris, R., et al., *wFReDoW: A Cloud-Based Web Environment to Handle Molecular Docking Simulations of a Fully Flexible Receptor Model.* Biomed Research International, 2013.

85. Lawrenz, M., R. Baron, and J.A. McCammon, *Independent-Trajectories Thermodynamic-Integration Free-Energy Changes for Biomolecular Systems: Determinants of H5N1 Avian Influenza Virus Neuraminidase Inhibition by Peramivir.* Journal of Chemical Theory and Computation, 2009. **5**(4): p. 1106-1116.

86. Mobley, D.L., J.D. Chodera, and K.A. Dill, *Confine-and-release method: Obtaining correct binding free energies in the presence of protein conformational change.* Journal of Chemical Theory and Computation, 2007. **3**(4): p. 1231-1235.

87. de Oliveira, C.A.F., D. Hamelberg, and J.A. McCammon, *Coupling accelerated molecular dynamics methods with thermodynamic integration simulations.* Abstracts of Papers of the American Chemical Society, 2009. **238**.

88. Wereszczynski, J. and J.A. McCammon, *Using Selectively Applied Accelerated Molecular Dynamics to Enhance Free Energy Calculations.* Journal of Chemical Theory and Computation, 2010. **6**(11): p. 3285-3292.

89. Lim, N.M., et al., *Sensitivity in Binding Free Energies Due to Protein Reorganization.* Journal of Chemical Theory and Computation, 2016. **12**(9): p. 4620-4631.

90. Danielson, M.L. and M.A. Lill, *Predicting flexible loop regions that interact with ligands: The challenge of accurate scoring.* Proteins-Structure Function and Bioinformatics, 2012. **80**(1): p. 246-260.

91. Forrey, C., J.F. Douglas, and M.K. Gilson, *The fundamental role of flexibility on the strength of molecular binding.* Soft Matter, 2012. **8**(23): p. 6385-6392.

92. Kolb, P. and J.J. Irwin, *Docking screens: right for the right reasons?* Curr Top Med Chem, 2009. **9**(9): p. 755-70.

93. Abagyan, R., M. Totrov, and D. Kuznetsov, *Icm - a New Method for Protein Modeling and Design - Applications to Docking and Structure Prediction from the Distorted Native Conformation.* Journal of Computational Chemistry, 1994. **15**(5): p. 488-506.

94. Knegtel, R.M.A., I.D. Kuntz, and C.M. Oshiro, *Molecular docking to ensembles of protein structures.* Journal of Molecular Biology, 1997. **266**(2): p. 424-440.

95. Sotriffer, C.A. and I. Dramburg, *"In situ cross-docking" to simultaneously address multiple targets.* Journal of Medicinal Chemistry, 2005. **48**(9): p. 3122-3125.

96. Yang, C.Y., R.X. Wang, and S.M. Wang, *M-score: A knowledge-based potential scoring function accounting for protein atom mobility.* Journal of Medicinal Chemistry, 2006. **49**(20): p. 5903-5911.

97. Craig, I.R., J.W. Essex, and K. Spiegel, *Ensemble Docking into Multiple Crystallographically Derived Protein Structures: An Evaluation Based on the Statistical Analysis of Enrichments.* Journal of Chemical Information and Modeling, 2010. **50**(4): p. 511-524.

98. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility.* Journal of Computational Chemistry, 2009. **30**(16): p. 2785-2791.

99. Neves, M.A.C., M. Totrov, and R. Abagyan, *Docking and scoring with ICM: the benchmarking results and strategies for improvement.* Journal of Computer-Aided Molecular Design, 2012. **26**(6): p. 675-686.

100. Leach, A.R., B.K. Shoichet, and C.E. Peishoff, *Prediction of protein-ligand interactions. Docking and scoring: Successes and gaps.* Journal of Medicinal Chemistry, 2006. **49**(20): p. 5851-5855.

101. Carlson, H.A., et al., *CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma.* Journal of Chemical Information and Modeling, 2016. **56**(6): p. 1063-1077.

102. Damm-Ganamet, K.L., et al., *CSAR Benchmark Exercise 2011-2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series.* Journal of Chemical Information and Modeling, 2013. **53**(8): p. 1853-1870.

103. Gaieb, Z., et al., *D3R Grand Challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies.* Journal of Computer-Aided Molecular Design, 2018. **32**(1): p. 1-20.

104. Gathiaka, S., et al., *D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions.* Journal of Computer-Aided Molecular Design, 2016. **30**(9): p. 651-668.

105. Smith, R.D., et al., *CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions.* Journal of Chemical Information and Modeling, 2011. **51**(9): p. 2115-2131.

106. Ye, Z.F., et al., *Optimal strategies for virtual screening of induced-fit and flexible target in the 2015 D3R Grand Challenge.* Journal of Computer-Aided Molecular Design, 2016. **30**(9): p. 695-706.

107. Kumar, A. and K.Y.J. Zhang, *Prospective evaluation of shape similarity based pose prediction method in D3R Grand Challenge 2015.* Journal of Computer-Aided Molecular Design, 2016. **30**(9): p. 685-693.

108. Salmaso, V., et al., *DockBench as docking selector tool: the lesson learned from D3R Grand Challenge 2015.* Journal of Computer-Aided Molecular Design, 2016. **30**(9): p. 773-789.

109. Slynko, I., et al., *Docking pose selection by interaction pattern graph similarity: application to the D3R grand challenge 2015.* Journal of Computer-Aided Molecular Design, 2016. **30**(9): p. 669-683.

110. Lam, P.C.H., R. Abagyan, and M. Totrov, *Ligand-biased ensemble receptor docking (LigBEnD): a hybrid ligand/receptor structure-based approach.* Journal of Computer-Aided Molecular Design, 2018. **32**(1): p. 187-198.

111. Athanasiou, C., et al., *Using physics-based pose predictions and free energy perturbation calculations to predict binding poses and relative binding affinities for FXR ligands in the D3R Grand Challenge 2.* Journal of Computer-Aided Molecular Design, 2018. **32**(1): p. 21-44.

112. Cleves, A.E. and A.N. Jain, *Knowledge-guided docking: accurate prospective prediction of bound configurations of novel ligands using Surflex-Dock.* Journal of Computer-Aided Molecular Design, 2015. **29**(6): p. 485-509.

113. Davis, I.W. and D. Baker, *ROSETTALIGAND Docking with Full Ligand and Receptor Flexibility.* Journal of Molecular Biology, 2009. **385**(2): p. 381-392.

114. Jain, A.N. and A. Nicholls, *Recommendations for evaluation of computational methods.* Journal of Computer-Aided Molecular Design, 2008. **22**(3-4): p. 133-139.

115. Cournia, Z., B. Allen, and W. Sherman, *Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations.* Journal of Chemical Information and Modeling, 2017. **57**(12): p. 2911-2937.

CHAPTER 2

Methodology

†Reprinted (adapted) with permission from Bansal, N.; Zheng, Z.; Merz, K. M., Incorporation of side chain flexibility into protein binding pockets using MTflex. *Bioorgan Med Chem* **2016**, 24, 4978-4987.
Reprinted (adapted) with permission from Bansal, N.; Zheng, Z.; Song, F. L.; Pei, J.; Merz, K. M., The Role of the Active Site Flap in Streptavidin/Biotin Complex Formation. *J Am Chem Soc* **2018,** DOI 10.1021/jacs.8b00743

2.1 Abstract

This chapter delineates the motivation and methodologies used in our study. Sampling and scoring are two key factors to improve the overall free energy estimated by a computational method. We generate conformations with the bigger purpose of improving the overall binding free energy estimation by our movable type free energy method. The conformers are generated by novel methods, which we call MT_{flex} and $MT_{Flex\text{-}b}$. The conformational ensemble is generated on an energy scale using pre-tabulated one-dimensional databases of distance versus energies. MT_{flex} generates conformations for the side-chain only keeping the backbone rigid while $MT_{Flex\text{-}b}$ extends the concept to include both backbone and side chain flexibility. However, both the methods generate conformations using same mechanics. Free energies are further calculated by using our in-house method called Movable type free energy method

2.2 Introduction

Our unique and trivially parallel free energy estimation procedure follows a two-step procedure. In the first step, the significant uncorrelated configurational states are assembled on a molecular energy landscape using our conformation search algorithm for receptor flexibility- MT_{flex} and $MT_{Flex\text{-}b}$. The molecular conformations are generated at an atom-pair level using a distance-based coordinate system, where each selected distance is associated with the pair-potential value selected from a pre-built look up table, which helps us to rapidly generate structures on an energy landscape. In this way, we obtain several uncorrelated molecular conformations of the ensemble on an energy surface in a rapid manner by performing smart or guided dihedral scans using our potentials.

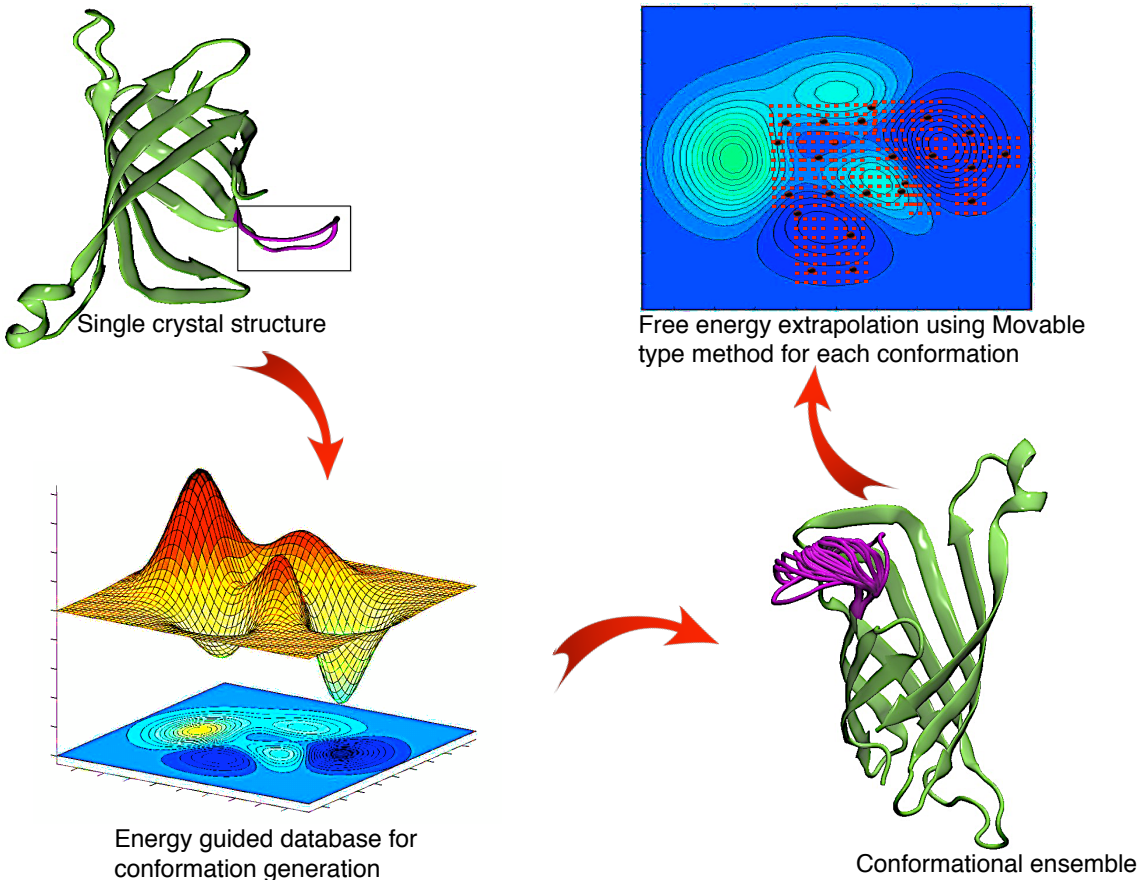

Figure 2.1 Pictorial representation of our free energy estimation procedure. The conformations generated by $MT_{Flex\text{-}B}$ are generated on an energy surface (shown in the heat map plot on the left). Each "seed structure" is fed to the Movable type method (shown on the right), which performs the local sampling around the initial conformation as explained with the red dots on the right heat map plot.

The generated pool of molecular conformations served as the "seed structures", which were fed to the MT free energy method for extrapolation of the local partition function and plotting the complete energy landscape. Figure 2.1 shows the pictorial representation of our free energy simulation procedure.

2.3 Conformation generation

Our conformation generation strategy is analogous to 3-D printing. In three-dimensional space, any coordinate system requires three variables to locate a particle in the domain of definition *e.g.* x, y, z in the Cartesian coordinate system and r, θ, φ in the spherical coordinate system. Locating one particle requires three distance variables associated with three reference points. A coordinate translation from the Cartesian coordinate system is shown in equation 2.1, with the particle coordinate (x, y, z) translated to (d_1, d_2, d_3) using three reference points (x_1, y_1, z_1), (x_2, y_2, z_2) and (x_3, y_3, z_3).

$$\begin{aligned} d_1 &= \sqrt{(x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2} \\ d_2 &= \sqrt{(x - x_2)^2 + (y - y_2)^2 + (z - z_2)^2} \\ d_3 &= \sqrt{(x - x_3)^2 + (y - y_3)^2 + (z - z_3)^2} \end{aligned} \tag{2.1}$$

The reference points, like the origin in the Cartesian coordinate system, can be randomly defined within the domain of definition. Nonetheless, due to the association of our pair potentials to the distances, the reference points used herein are defined as atoms with known locations.

In each step a new atom is added to the growing molecular ensemble based on the solution of equation 4 using three different reference atoms (see Figure 2.2). Every selected distance is associated with a pair potential value chosen from a pair potential lookup table indexed with the atom pair type at discretized distances. Theoretically, the selection of the distances (d_1, d_2, d_3) could be random once it satisfies the solution of equation 2.1. However, random selection yields too many unfavorable structures appearing in the final molecular ensemble collection, which suggests the importance of using the pair potentials (as Boltzmann weighted factors – See the upper-right of Figure 2.2) as a reference for the selection of appropriate distances. Moreover, any three atoms with known locations in the growing molecular ensemble can be selected as the

reference points during each step of the construction, resulting in a much larger sample size during the computation. Briefly, for each of the distances (d_1, d_2, d_3) between the new atom and the reference atoms, a sampling or "blurring" range is set around the selected distance to create a vector of energies as a function of r, which is the candidate pool for future structure sampling. The vector of one-dimensional energies (atom pairwise energy component to the molecular ensemble) is then expanded into a matrix with a user defined size, by randomly "scrambling" the order of the original energy vectors repeatedly and then "tiling" these disordered vectors into a fix-sized matrix. The numerical integral is performed by point-wise multiplication through all the three one-dimensional energy matrices according to the three reference atoms (equations 2.2 to 2.4).

$$\not{z}_\alpha = \begin{bmatrix} \vdots \\ z_\alpha(r_0 - 2\Delta r) \\ z_\alpha(r_0 - \Delta r) \\ z_\alpha(r_0) \\ z_\alpha(r_0 + \Delta r) \\ z_\alpha(r_0 + 2\Delta r) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ e^{-\beta E_\alpha(r_0 - 2\Delta r)} \\ e^{-\beta E_\alpha(r_0 - \Delta r)} \\ e^{-\beta E_\alpha(r_0)} \\ e^{-\beta E_\alpha(r_0 + \Delta r)} \\ e^{-\beta E_\alpha(r_0 + 2\Delta r)} \\ \vdots \end{bmatrix} \Rightarrow scramble\left(\not{z}_\alpha\right) = \begin{bmatrix} z_\alpha(r_0 - 8\Delta r) \\ z_\alpha(r_0 + 3\Delta r) \\ z_\alpha(r_0 + \Delta r) \\ \vdots \\ z_\alpha(r_0) \end{bmatrix} \tag{2.2}$$

$$\not{Z}_\alpha = \underbrace{\left.\begin{bmatrix} scramble(\not{z}_\alpha)_1 & scramble(\not{z}_\alpha)_{i+1} & \cdots & scramble(\not{z}_\alpha)_{n-i+1} \\ scramble(\not{z}_\alpha)_2 & scramble(\not{z}_\alpha)_{i+2} & & scramble(\not{z}_\alpha)_{n-i+2} \\ \vdots & \vdots & \ddots & \vdots \\ scramble(\not{z}_\alpha)_i & scramble(\not{z}_\alpha)_{2i} & \cdots & scramble(\not{z}_\alpha)_n \end{bmatrix}\right\} m \text{ rows}}_{n \text{ columns}} \tag{2.3}$$

$$\not{Z}_A = \not{Z}_1 \bullet \not{Z}_2 \bullet \not{Z}_3 \tag{2.4}$$

The random disordered permutations to each one-dimensional energy matrix is meant to maximize the variety of energy combinations at different distances, and the fixed-size matrix multiplication is to maintain a computationally tractable sampling size. As shown in equation 2.2, the row number m of the matrices is defined as the least multiplier of all the atom pairwise vector sizes in the molecular system under study in order to identically size all atom pairwise energy vectors with different sampling or "blurring" ranges for different pair potentials, while the column number n is a user defined number defining the sampling size to satisfy a convergent ensemble through the different pair potentials. Hence, the matrix Z_A in equation 2.4 after the assembly through the 3 pair-potential matrices includes $m \times n$ possibilities to locate the new atom for each distance set (d_1, d_2, d_3) that have been selected. This candidate pool selection for just one single atom would be a burden for the construction of a molecular structure due to the exponential increase in the sampling size with the number of atoms. To address this, the selection of each distance set (d_1, d_2, d_3) is constrained within a limited range according to the pair potential significance (see upper-right of Figure 2.2).

The method used herein is analogous to the conformational search algorithm for small molecules recently developed by our group.[99] Bonds and angles are restricted to their well-depth location due to their extremely narrow energy ranges. Distances for torsions and non-bond contacts are chosen at their most favorable regions to give precedence to the most relevant contacts. For instance when coming to a step involving the addition of a new hydrogen bond donor/acceptor to the growing ensemble, other than considering all possible torsion distributions of this atom, a 3 Å vicinity is searched for new locations to satisfy any potential hydrogen bond contacts. Atoms forming a disulphide bond are not allowed to move during the side chain generation. The generated poses are accepted only if they don't clash with the rest of the

structure. A geometric cutoff of 2.8 Å is set as the collision criteria for the non-bonded heavy atoms. This procedure narrows down our conformational search while sampling in the appropriate regime. An example illustrating the parallel printing scheme of our method is shown in the left-hand side of Figure 2.2.

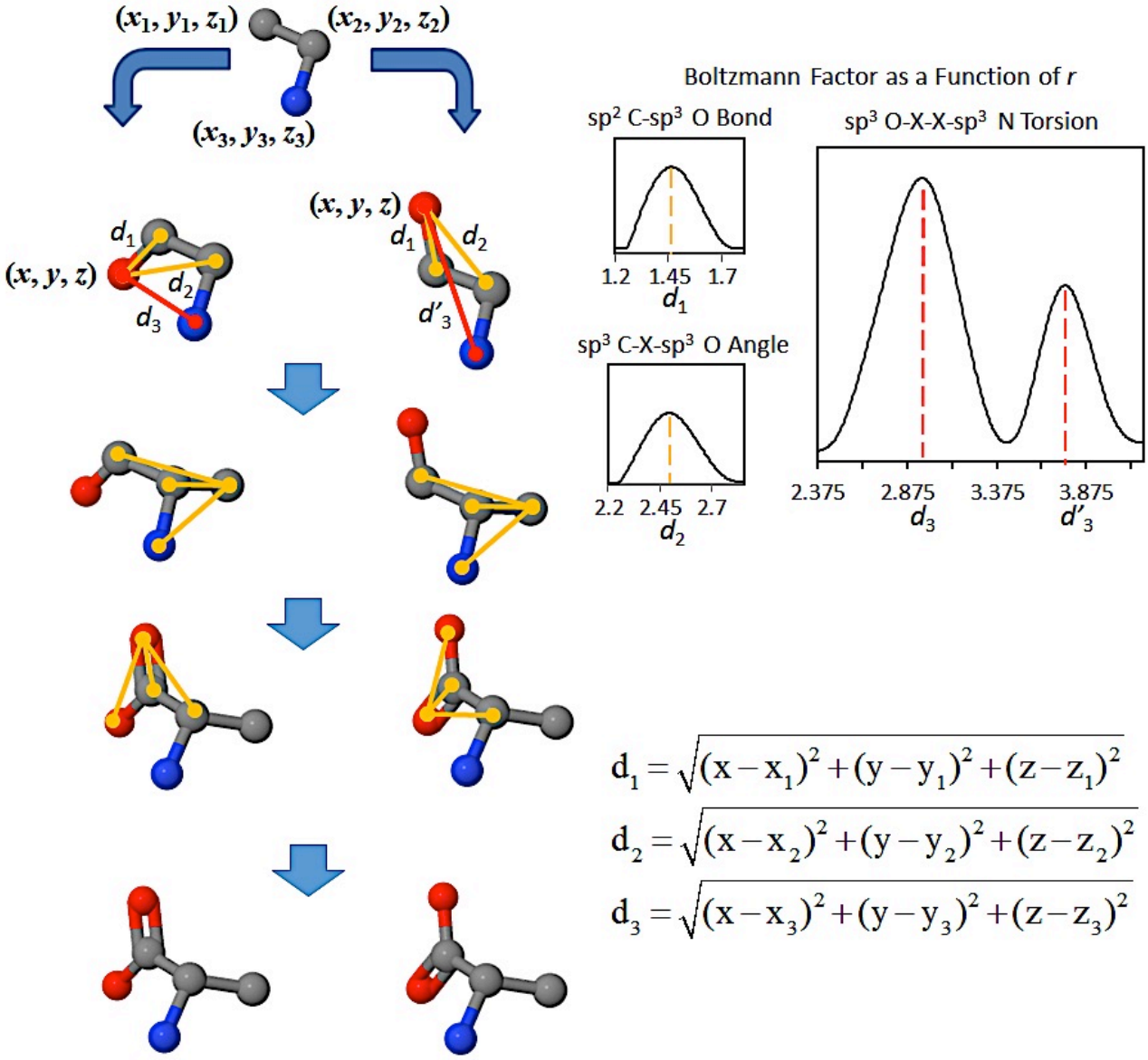

Figure 2.2. An example illustrating the parallel printing scheme used to construct alanine. The printing starts from three atoms with known locations. In each step a new atom is added to the system fixed by three distances constrained by the pair potentials from the lookup table. Bonds and angles are fixed to their minimum well depths due to their extremely narrow distribution ranges. Distances for the torsion angles and non-bonded interactions (not used in the alanine example) are selected based on a range of points around the energy minima (only minima are selected in this figure for simplicity).

2.4 Side chain flexibility

The construction of side chain conformers for one residue was independent of the other residues of the binding pocket; *i.e*. the three reference points to locate a particle were taken from the same residue. In this way, the first side chain atom to be rotated for each residue was C_γ, where backbone amide N, C_α and C_β served as the three reference atoms. Because the exploration of side chain conformations was done per residue steric clashes for each residue were checked after looping over all residues with respect to all other residues of the binding pocket and the remaining residues in the protein. Following this strategy, conformers were generated for 17 out of a total of 20 amino acids (Gly, Pro and Ala were skipped). Though conceptually similar, MT_{flex} differs from the rotamer library based methods in that new atoms are not built according to a structural database but on a local free-energy landscape defined by the potential utilized.

2.5 Loop flexibility

Conformational space available to loop grows exponentially with the loop length, so we generated the loop from both N and C termini separately. This loop buildup procedure bears similarity to Jacobson et al strategy [1], where loop was generated from both ends and closed in the middle. Our loop building strategy is different in a way that backbone and side-chain atoms for each residue were constructed altogether. This helped us in eliminating those backbone positions immediately, which couldn't fit side chain atoms. Several restraints were also applied during structural sampling. The entire loop was generated in the presence of the rest of the protein structure so the Vander Waal collision restraints served as the primary source of screening. A geometric restraint of 2.8 Å was used for the non-bonded heavy atoms and 2.3 Å

for the hydrogen bond forming atoms. Steric clashes within the generated loop structure were also identified using the same geometric restraint. The atoms forming a disulphide bond were also kept intact during the side chain formation of the generated loop structure. Chirality at α-carbon position was considered for all the amino acids for the backbone generation. β-carbon chirality was considered for isoleucine and threonine for the generation of side chains conformers. Ring planarity of aromatic amino acids was also enforced. The sampling size was kept at a manageable level in this way. The loop closure of the generated loop fragments was obtained using the following procedure. The N-termini loop fragments were generated from 1 to i^{th} loop residues, where the i^{th} residue corresponds to one of the middle residues of the loop. Similarly, the C-termini loop fragments were generated from M to $(i+1)^{th}$ residues, where M is the last residue of the loop. See Figure 2.3 for reference. Both ends were tried to meet by finding the pair from either end falling in the optimum bond distance for the carbonyl carbon of the i^{th} residue and backbone Nitrogen of $(i+1)^{th}$ residue. The optimum distance was chosen from our pre-built look up table of distance versus energies. Apart from the bond distance restraint, the geometric criterion for checking Van der Waals collisions between the connecting residues was also imposed. In this way, the loop fragments, which were unable to close or had steric clashes, were pruned from the ensemble on the fly.

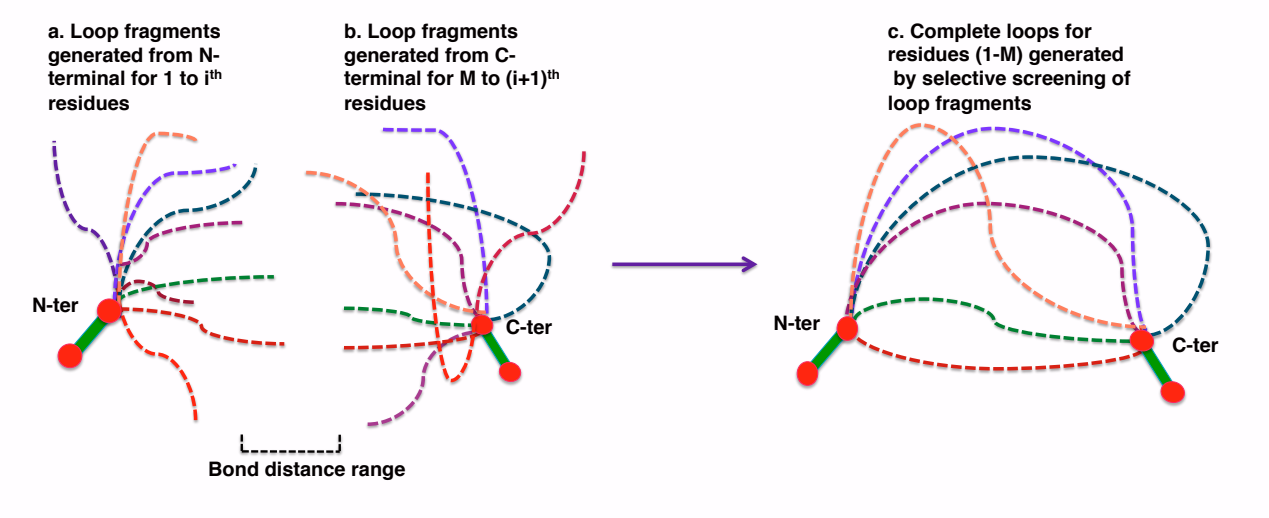

Figure 2.3 Pictorial representation of our loop closure strategy. Loop fragments are generated from both N and C terminals individually. The individual loop fragments are then combined from both ends to form complete loops.

2.6. Free energy estimation

The Movable Type (MT) method numerically simulates local partition functions utilizing a Monte Carlo integration (MCI) scheme, given an initial structure, from a canonical ensemble. The MCI method is a widely used numerical approach for free energy calculation.[2, 3] By simulating the integral of the canonical partition function instead of generating enthalpy and entropy values separately, the MCI method allows for the avoidance of expensive and poorly converging entropy calculations.

$$A = -RT\ln\left[Z_M\right] = -RT\ln\left[\int_D e^{-\beta E_M(\tau)}d\tau\right] \approx -RT\ln\left[V\frac{\sum_i^N e^{-\beta E_i(r)}}{N}\right] \qquad (2.5)$$

The development of the MT algorithm is inspired by the idea of the MCI approach expressed in equation 2.5, where the Helmholtz free energy is simulated using the average of the sampled energy states multiplied by the actual sampling volume. The distinctive feature of the

MT method is that it numerically simulates the average of the local partition function given a defined sampling volume centered around an initial structure, instead of searching among actual physical structures within that defined volume. In our first MT algorithm publication, a matrix-based random sampling strategy combining every atom pairwise potential against each target molecular system was introduced, in which all pairwise potentials are regarded as orthogonal and a total random combination among the atom pairwise distances were performed within a small range of sampling (±0.5 Å for every atom pairwise distances). The generated hyper-dimensional energy states were associated with pre-modeled structural weighting factors and averaged over their sampling magnitude C^N, where C is the defined sampling range and N is the pairwise contact number. This approach is to simulate the average of the actual physical energy states using the more easily constructed virtual states.

Given an N-particle physical space, a quantitative description of the ensemble volume is written as:

$$V = \int \cdots \int_D d\tau_1 \cdots d\tau_N \tag{2.6}$$

where, τ_1 to τ_N are the coordinates of all the particles and D is the domain of definition for all of the particle coordinates. The ensemble volume is under exponential growth as the number of dimensions increases. The MT algorithm uses a distance-based coordinate system in order to better estimate the ensemble volume, and importantly, this approach is well suited to the MT method, where each distance is associated with a pair potential value chosen from a pre-built lookup table allowing us to simultaneously generate the structures and energies for a given system. [4]

The feature of our method that sets it apart from other traditional methods is that it simultaneously generates the structure and its local free energy by simulating the partition

function within a defined ensemble volume by multiplying the pair potential matrices as summarized in equations 2.2-2.4 through all atom pairwise potentials in the molecular system. Atom pairwise energies stored in the final partition function matrix (Z_M) represent the virtual hyper-dimensional energy states given the defined sampling range regarding each generated conformation. The free energy is then calculated using the ensemble volume and Z_M.

$$Z_M = Z_1 \bullet Z_2 \bullet Z_3 \bullet \cdots \bullet Z_n \tag{2.7}$$

$$A = -RT \ln \left[V \left\langle e^{-\beta E(\tau)} \right\rangle \right] = -RT \ln \left[V \frac{Z_M}{m \times n} \right] \tag{2.8}$$

where, β is the Boltzmann constant, $E(\tau)$ is the molecular energy as a function of geometric variable τ, m and n represent the number of rows and columns of the Z-matrix. For further detailed explanation of the free energy calculation and the solvation model, please see Zheng *et al.* and Pan *et al*.[6, 7, 8]

2.7. Solvation free energy

Solvation free energy is calculated using the implicit solvation model KECSA-Movable Type Implicit Solvation Model (KMTISM).[5] KMTISM is a semi-continuum solvation model, which places water molecules around the solute as isotropic rigid balls with the van der Waals radii of 1.6 Å. The water molecules are placed in incremented isometric layers of 0.005 Å up to 8 Å from the solute's van der Waals surface. The solvation free energy change for the protein-ligand binding process is separated into two components, representing the translation of the solutes and water molecules during the binding process. During the protein-ligand binding in solution, the ligand replaces a number of water molecules in the protein-binding site, with

respect to the volume occupied by the ligand. At the same time, these water molecules lost upon ligand binding are transferred into bulk. Figure 2.4 highlights the pictorial representation of the change in solvation free energy during the protein-ligand binding process.

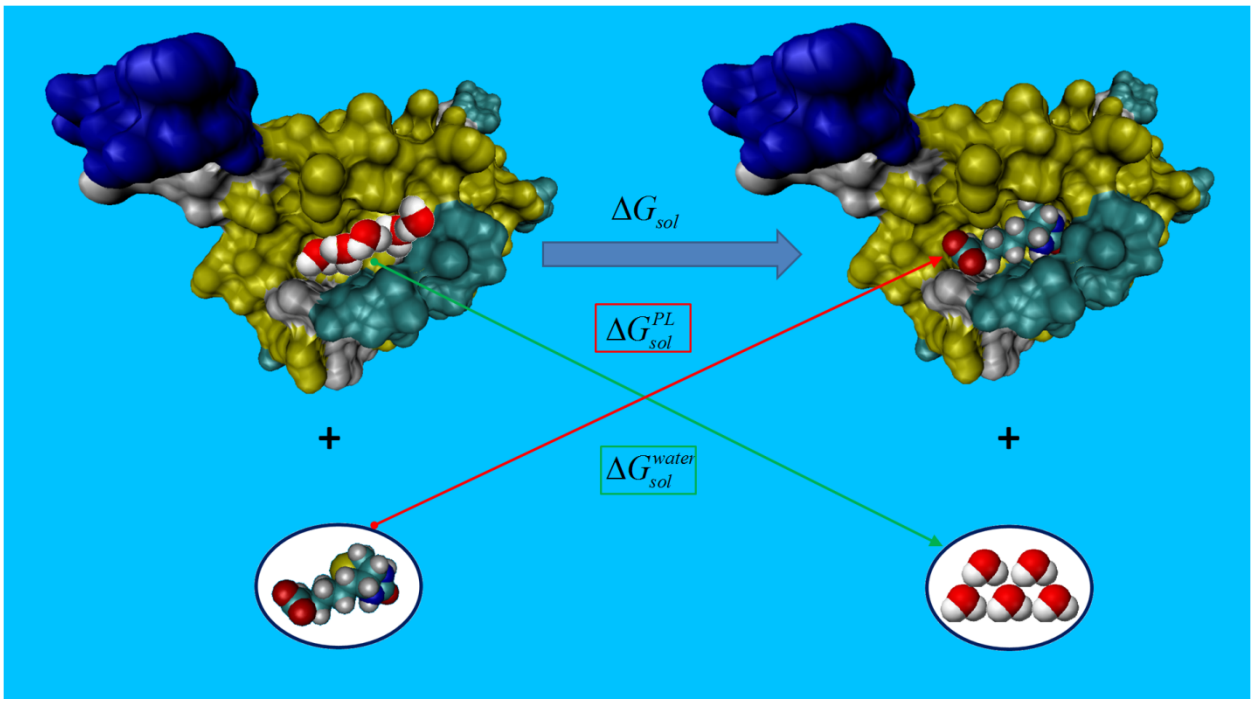

Figure 2.4 Illustration of the solvation free energy change during protein-ligand binding. The red arrow shows the desolvation and placement of the ligand into the protein-binding site. This process replaces water molecules bound in the protein-binding site that are displaced by the ligand volume. The green arrow shows the dissociation and solvation of these water molecules during the binding procedure.

Given these two simultaneous processes, the solvation free energy change is calculated using the following equation:

$$\Delta G_{sol} = \Delta G_{sol}^{PL} + \Delta G_{sol}^{water} \tag{2.9}$$

where,

$$\Delta G_{sol}^{PL} = \frac{1}{-\beta}\log\left[\sum_i \exp\left(-\beta\left(E_{sol}^{PL}\right)_i\right)\right] - \frac{1}{-\beta}\log\left[\sum_j \exp\left(-\beta\left(E_{sol}^{P}\right)_j\right)\right] - \frac{1}{-\beta}\log\left[\sum_k \exp\left(-\beta\left(E_{sol}^{L}\right)_k\right)\right] \quad (2.10)$$

ΔG_{sol}^{PL} represents the solvation free energy change with respect to the solute-solvent interactions, and:

$$\Delta G_{sol}^{water} = \frac{1}{-\beta}\log\left[\sum_i \exp\left(-\beta\left(E_{sol}^{water}\right)_i\right)\right] - \frac{1}{-\beta}\log\left[\sum_j \exp\left(-\beta\left(E_{receptor}^{water}\right)_j\right)\right] \quad (2.11)$$

ΔG_{sol}^{water} represents the solvation free energy change with respect to removing water molecules at the binding site and placing them into solvent. E_{sol}^{PL}, E_{sol}^{P} and E_{sol}^{L} are interaction potential energies for the protein-ligand complex, protein and ligand in contact with solvent respectively; E_{sol}^{water} and $E_{receptor}^{water}$ are interaction potential energies for the water molecules being replaced representing their contacts with solvent and with the protein receptor, respectively. Boltzmann factors for all the sampled states are summed to simulate the corresponding partition functions. Free energy changes are then calculated using the logarithms of the summed Boltzmann factors using the mechanics described above for the free energy calculation.

REFERENCES

REFERENCES

1. Jacobson, M.P., et al., *A hierarchical approach to all-atom protein loop prediction.* Proteins-Structure Function and Bioinformatics, 2004. **55**(2): p. 351-367.

2. Valleau, J.P. and D.N. Card, *Monte-Carlo Estimation of Free-Energy by Multistage Sampling.* Journal of Chemical Physics, 1972. **57**(12): p. 5457-&.

3. Jorgensen, W.L. and C. Ravimohan, *Monte-Carlo Simulation of Differences in Free-Energies of Hydration.* Journal of Chemical Physics, 1985. **83**(6): p. 3050-3054.

4. Zheng, Z. and K.M. Merz, *Development of the Knowledge-Based and Empirical Combined Scoring Algorithm (KECSA) To Score Protein-Ligand Interactions.* Journal of Chemical Information and Modeling, 2013. **53**(5): p. 1073-1083.

5. Zheng, Z., et al., *KECSA-Movable Type Implicit Solvation Model (KMTISM).* Journal of Chemical Theory and Computation, 2015. **11**(2): p. 667-682.

6. Zheng, Z., M.N. Ucisik, and K.M. Merz, Jr., *The Movable Type Method Applied to Protein-Ligand Binding.* J Chem Theory Comput, 2013. **9**(12): p. 5526-5538.

7. Pan, L.L., et al., *Free Energy-Based Conformational Search Algorithm Using the Movable Type Sampling Method.* Journal of Chemical Theory and Computation, 2015. **11**(12): p. 5853-5864.

8. Zheng, Z., et al., *KECSA-Movable Type Implicit Solvation Model (KMTISM).* Journal of Chemical Theory and Computation, 2015. **11**(2): p. 667-682.

CHAPTER 3

Incorporation of Side Chain Flexibility into Protein Binding Pockets using MT_{flex}

† Reprinted (adapted) with permission from Bansal, N.; Zheng, Z.; Merz, K. M., Incorporation of side chain flexibility into protein binding pockets using MTflex. *Bioorgan Med Chem* **2016**, 24, 4978-4987.

3.1 Abstract

The plasticity of active sites plays a significant role in drug recognition and binding, but the accurate incorporation of "receptor flexibility" remains a significant computational challenge. Many approaches have been put forward to address receptor flexibility in docking studies by generating relevant ensembles on the energy surface, but, herein, we describe a method (Movable Type with flexibility (MT_{flex})) that generates ensembles on the more relevant free energy surface in a computationally tractable manner. This novel approach enumerates conformational states based on side chain flexibility and then estimates their relative free energies using the MT methodology. The resultant conformational states can then be used in subsequent docking and scoring exercises. In particular, we demonstrate that using the MT_{flex} ensembles improves MT's ability to predict binding free energies over docking only to the crystal structure.

3.2 Introduction

Gaining a deep understanding of the interactions involved in the protein-ligand binding process has been of great interest for many decades because of its importance in a wide range of biochemical processes occurring inside living organisms. To explain this process several theories have been put forward of which the most widely accepted ones include: lock and key[1], induced fit[2], and the conformational selection[3] model. The former dates back to 1854, when Emil Fischer first proposed the lock and key model where the active site was viewed as structurally immutable. His theory remained popular until experiments reported changes in the receptor structure upon ligand binding.[4-8] From these observations, Koshland proposed the induced fit model while Nussinov proposed the conformational selection model, which incorporated

receptor flexibility. The induced fit model proposed that ligand binding induced conformational changes in the receptor; the conformational selection model assumes that the protein has an ensemble of states and that the ligand selectively picks one of these.[9, 10] Altogether experimental and theoretical considerations have modified our perception of how a ligand recognizes a biological receptor.

Today, proteins are considered as inherently flexible that possess a range of intrinsic motions.[5] The inherent flexibility of proteins is attributed to the numerous non-covalent interactions that exist in their native states. They possess a variety of motions, which range from vibrational fluctuations up to large backbone rearrangements, which taken altogether generate a conformational ensemble for a given protein system.[5, 11-15] After much debate, it has been posited that proteins exist in multiple conformations in solution while available crystal structures represent just one member of the entire conformational ensemble. [5, 9, 11, 16-18]

Computational chemists have expended enormous intellectual capital to capture the plasticity of proteins.[19, 20] The conformational space spanned by proteins can be quite substantial making addressing this issue a highly challenging problem.[21-25] Even in the face of these challenges, numerous methods have been developed to account for protein flexibility in protein-ligand binding studies using simulations or docking methodologies.[26] Soft-receptor docking methods represent some of the first attempts to address this issue. These methods enumerated a very limited or localized flexibility by softening the van der Waals potentials, which allow for small steric clashes between the protein and ligand molecule.[27, 28] Rotamer library exploration is another way to approach partial flexibility by including side chain motions on the basis of allowable states of the rotatable bonds of select active site residues.[29-32] Ensemble docking is another class of docking methods that employs multiple receptor

conformations to account for protein flexibility prior to docking. These multiple conformations are generated by either computer simulations based on Monte Carlo, conventional or accelerated molecular dynamics[33-36], Normal Mode Analysis[37-40], homology models[41], or by collecting experimental structures using NMR or X-Ray[42-44]. Few of the known docking algorithms that use ensemble based docking are Autodock[45, 46], ITScore[47, 48], IFREDA[49], MedusaDock[50], RosettaBackrub[51], FlexE[52], BP-Dock[53] and MDock[44]. Induced fit docking (IFD) is another major category of docking algorithms that account for a certain degree of receptor flexibility. Few examples of docking algorithms that use IFD are Autodock Vina[54], Autodock4[45, 55], GOLD[56, 57], GLIDE[58], RosettaLigand[59, 60], ICM[39, 49], FLIPDock[61, 62], DOCK6.0[63], SLIDE[64-66], ReFlexin[67], GalaxyDock[68], FITTED[69-71], PC-RELAX[72, 73], Flesky[74], FiberDock[75], hinge belt docking algorithms[76-79] and Adaptive BP-Dock.[80] Other methods include introducing flexibility in the receptor structures using hybrid methods like the linear interaction energy (LIE)[81, 82], Relaxed complex scheme (RCS)[83, 84], dynamic pharmacophore modeling[85, 86] and single step perturbation.[87-89] An extensive overview of all receptor flexibility methods and docking software packages is provided in recent reviews.[26, 90-94] One drawback of docking methods is the difficulty of obtaining a converged partition function within a limited structural ensemble. Monte Carlo and molecular dynamics simulations (and their variations, *e.g.*, Markov Chain Monte Carlo, replica exchange, *etc.*) can delineate more of the relevant ensemble, but in general, at an increased computational cost.

Recently we have described a novel free energy calculation method called the Movable Type (MT) method.[95] Using its mechanics of parallel conformation generation, we have developed MT_{flex} and an associated program for fast protein conformation generation and energy

sampling. The present manuscript introduces a method to sample the side chain conformations within a protein active site via the MT_{flex} methodology generating an ensemble of structures that can be subsequently used in ensemble docking studies.

3.3 Results and Discussion

In the current study, we applied the MT_{flex} method to side-chain conformational ensemble generation against *holo* protein crystal structures, which was then followed by docking and scoring to the resultant structure. In particular, we explored the following: (1) side chain flexibility and protein free energy variations with fixed backbone conformations in the binding pocket area; (2) binding mode predictions and binding free energy predictions using protein structures with different side chain conformations. The validation benchmark set involved 159 protein-ligand crystal structures from the PDBbind v2014 core database after excluding proteins having metal ions in the binding pockets area.[96, 97] The core dataset consists of high-resolution structures with a wide distribution of pK_d values.

3.3.1. Generation of the protein side chain conformations and the relative free energies using MT_{flex}

The binding pocket's residues were identified as those lying in a 6-Å vicinity of the co-crystal native ligands. All the heavy atoms in a residue were used to identify a binding pocket. If a residue had one or more than one heavy atoms (be it a side chain or backbone atom) within 6 Å of the native ligand, then the entire residue was considered to be a part of the binding pocket. Isoleucine and threonine are chiral at the β-carbon position, which is the first heavy atom of the

side chains of all amino acids except Glycine. Chirality was considered for these two amino acids for the generation of side chains conformers using MT_{flex}. Ring planarity of aromatic amino acids was also enforced. Disulphide bonds were also kept intact. Steric clashes within the generated binding pocket and with the rest of the protein were avoided by the use of a geometric restraint of 2.3 Å for the non-bonded heavy atoms with the capacity to form hydrogen bonds and 2.8 Å for the non-bonded non-hydrogen heavy atoms. The MT_{flex} conformations were generated *in-vacuo* (no explicit or implicit water model).

The MT_{flex} code was written using MATLAB R2015a software.[98] The code was run using a single CPU processor on the Intel14 cluster provided by the High Performance Computer Center (HPCC) facility at MSU. Each node consists of two 2.5 GHz 10-core Intel Xeon Ivy Bridge E5-2670v2 processors and 512 GB memory per node. The performance of the code varied based on several factors including the- i) type of residue, ii) number of residues in the binding pocket, and iii) the crowding of the binding pocket. We did an averaging of the computation time required to generate MT_{flex} conformers for the 17 amino acids (excluding Gly, Pro and Ala) over 159 systems. Table 3.2 lists the average CPU time (in seconds) required to generate the MT_{flex} conformers for each amino acids. Currently, we are in the process of transitioning the code to C++ to facilitate distribution and to improve the computational performance.

The application of MT_{flex} with this criteria and restraints generated a varied number of conformers for each system depending on the nature of the binding pocket. The magnitude of the number of conformations varies from ~10 conformations for the crowded and compact binding pockets to thousands of poses. In total, we generated 38397 conformers for 159 systems. After generation of the conformers, we used a RMSD criterion of 0.5 Å to select the final set of

receptor conformations. The first conformation was always selected. For the remaining conformations, their RMSDs were compared to all the preceding selected conformations. If a conformation's RMSD was <= 0.5 Å with respect to any of the already selected conformations, it was rejected otherwise it was accepted. The list containing the numbers of MT_{flex} conformers retained for each system is provided in Table 3.3 of the supporting information.

The RMSDs (Å) of the MT_{flex} generated conformers were calculated with respect to the crystal structure for all systems. RMSDs were obtained by an in-house code to make sure that MT_{flex} conformers were orientated for the best fit to the crystal structure. The code was written using MATLAB R2015a. [98] The details of the algorithm are provided in Section 3.5.1 of the supporting information. The top panel of Figure 3.1 shows the deviation of the MT_{flex} conformers from each crystal structure in the validation set. It is clear that more conformations were generated for more flexible binding pockets (RMSD distribution range as the gap between the green line and the blue line shown in Figure 3.1). The distribution of the lowest RMSD conformers (the blue line in top panel of Figure 3.1) is from 0.05 Å to 1.15 Å representing how MT_{flex} reproduces structures relative to the crystal structures in the validation set.

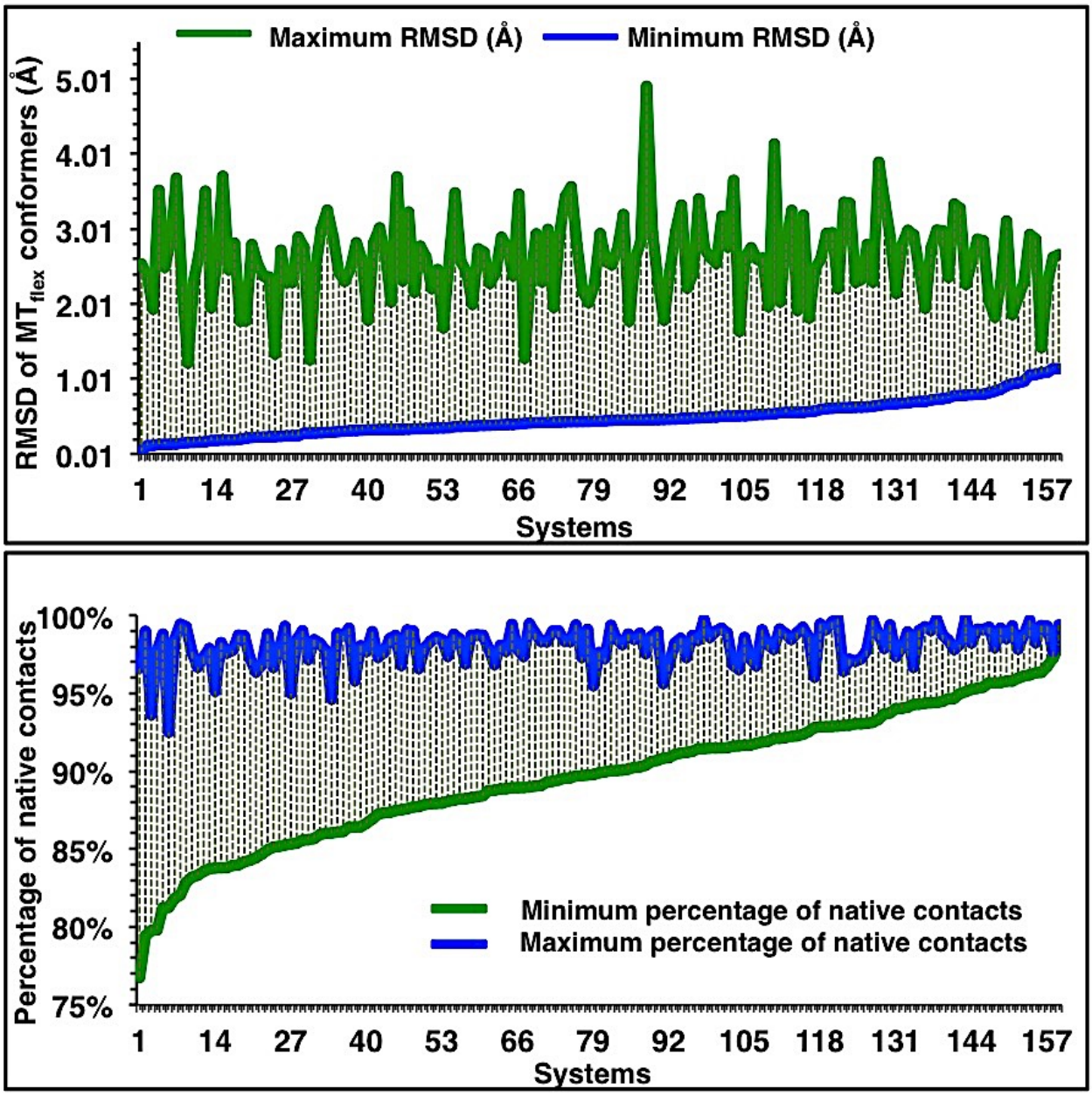

Figure 3.1 The top panel represents the minimum, maximum and RMSD range of MT_{flex} conformers generated for all 159 systems. Dark blue curve denotes the lowest RMSD, Green curve denotes the highest RMSD conformer and black stripes represent the RMSD ranges of MT_{flex} conformers for all 159 systems. The lower panel represents the percentage (%) of native contacts in MT_{flex} conformers with respect to the crystal-binding pocket. Dark blue curve shows the maximum %, Green curve symbolizes the minimum % and black stripes represent the range of native contacts in MT_{flex} conformers.

In general, binding pockets do not afford much space to work with in terms of reorienting the side chains. Indeed, any large-scale movement of a side chain would be rejected because of

van der Waals collisions with the rest of the protein. All of the systems studied had their lowest RMSDs less than 1.15 Å with the minimum value being ~0.05 Å.

However, for some systems the side chain RMSDs were higher with the maximum as high as 4.91Å. On closer examination, we observed that the source of the large RMSD values was the rotation of aromatic side chains in the receptor's binding pocket. RMSD is a good criterion to measure the deviation from a native structure, but we realized that they were biased somewhat due to the rotation of aromatic side chain in the binding pocket. Hence, in order to further understand the overall mobility of the binding pocket, we also calculated the percentage of native contacts in MT_{flex} conformers with respect to the binding pocket of the crystal structure (refer to the lower panel of Figure 3.1). The percentage of native contacts reflects the number of native contacts retained in the MT_{flex} conformers over the total number of contacts present in the crystal protein structure. The number of native contacts was calculated with a distance cutoff of 6.0 Å using the CPPTRAJ utility of Amber tools (2015).[99] It can be seen from the lower panel of Figure 3.1 that the maximum percentage of native contacts for all 159 systems of the validation dataset is within 92.6-99.6%, while the minimum percentage is as low as 76.7%. Given the fixed backbone structures, the gap between the minimum and the maximum percentage of native contacts indicates the magnitude of flexibility of the side chains in the binding pockets according to the MT_{flex} method.

Shown in Figure 3.2, two protein structures, with PDBIDs 3UEU and 1R5Y, were selected from the validation dataset as examples to illustrate the positioning of the side chains produced by the MT_{flex} method. These two structures had the largest gaps in the percentage of native contacts (according to the distances between the blue and green lines in Figure 3.1), while at the same time the MT_{flex} generated conformations covered the native structures for the two

proteins (lowest 0.50 Å RMSD and 0.17 Å RMSD and maximum native contacts of 97% for 3UEU and 99% for 1R5Y, respectively). The largest deviations for these two proteins have RMSDs of 2.8Å for 3UEU and 3.51Å for 1R5Y, with the corresponding percentage of native contacts being 76.7% for 3UEU and 79.4% for 1R5Y. Table S2 lists the minimum and maximum RMSDs for each system along with the PDBID of each system.

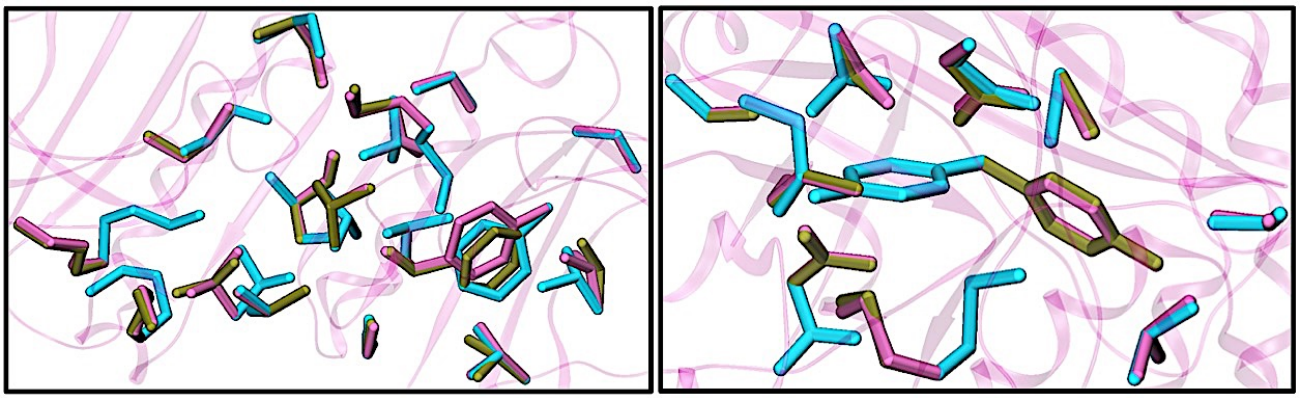

Figure 3.2 It shows the side chain of binding pocket region of PDB ID 3UEU (left) and 1R5Y (right). The crystal-binding pocket represented in pink licorice conformation is superimposed with the MT_{flex} conformer with the maximum percentage of native contacts (tan) and MT_{flex} conformer with minimum percentage of native contacts (cyan color).

The highest and lowest free energy differences among the protein conformations were obtained determined and the top panel of Figure 3.3 illustrates the range of free energies obtained for MT_{flex} binding pocket conformations relative to the *holo* crystal conformation. The blue curve represents the free energy difference between the most "relaxed" (lowest free energy) MT_{flex} binding pocket conformation and the *holo* crystal conformation and the red curve between the highest free energy and the crystal structure for all 159 systems in the validation dataset. Approximately, 68.5% of the systems in the validation dataset a MT_{flex} generated conformation gave a free energy lower than the crystal structure, while for 31.5% of the systems all MT_{flex} generated conformations had higher free energies than the crystal structure. To dig into this

more, we lumped the 68.5% into Region 1 and the rest in Region 2. We estimated the net difference in free energies between the highest free energy and the lowest free energy MT_{flex} conformation for each system of the validation dataset and observed that the average difference of free energies for all the systems in Region 1 is 4.31 kcal/mol and for Region 2 is 3.9 kcal/mol. We also calculated the minimum and maximum RMSDs (Å) of the MT_{flex} conformations for each system (plotted in the lower panel of Figure 3.3) and found out that the net RMSD difference between the minimum and maximum RMSD conformations in Region 1 is 2.21 Å while in Region 2 is 1.99 Å. The narrower range of free energy and the constricted RMSD difference between minimum and maximum RMSD conformations in Region 2 suggests conformational sampling is restricted in Region 2. The restricted conformational sampling suggests two possible scenarios-a) the crystal-binding pockets in Region 2 are not flexible in nature or, b) partial relaxation of the crystal binding pocket with only side chain flexibility is not sufficient for the systems in Region 2. Alternatively, issues with our statistical energy function can give the observed behavior. Because there are no hard and fast rules governing whether or not pocket sampling across a series of proteins should yield a certain percentage of lower or higher free energies conformations for a given pocket we cannot fully assess the performance of our approach.

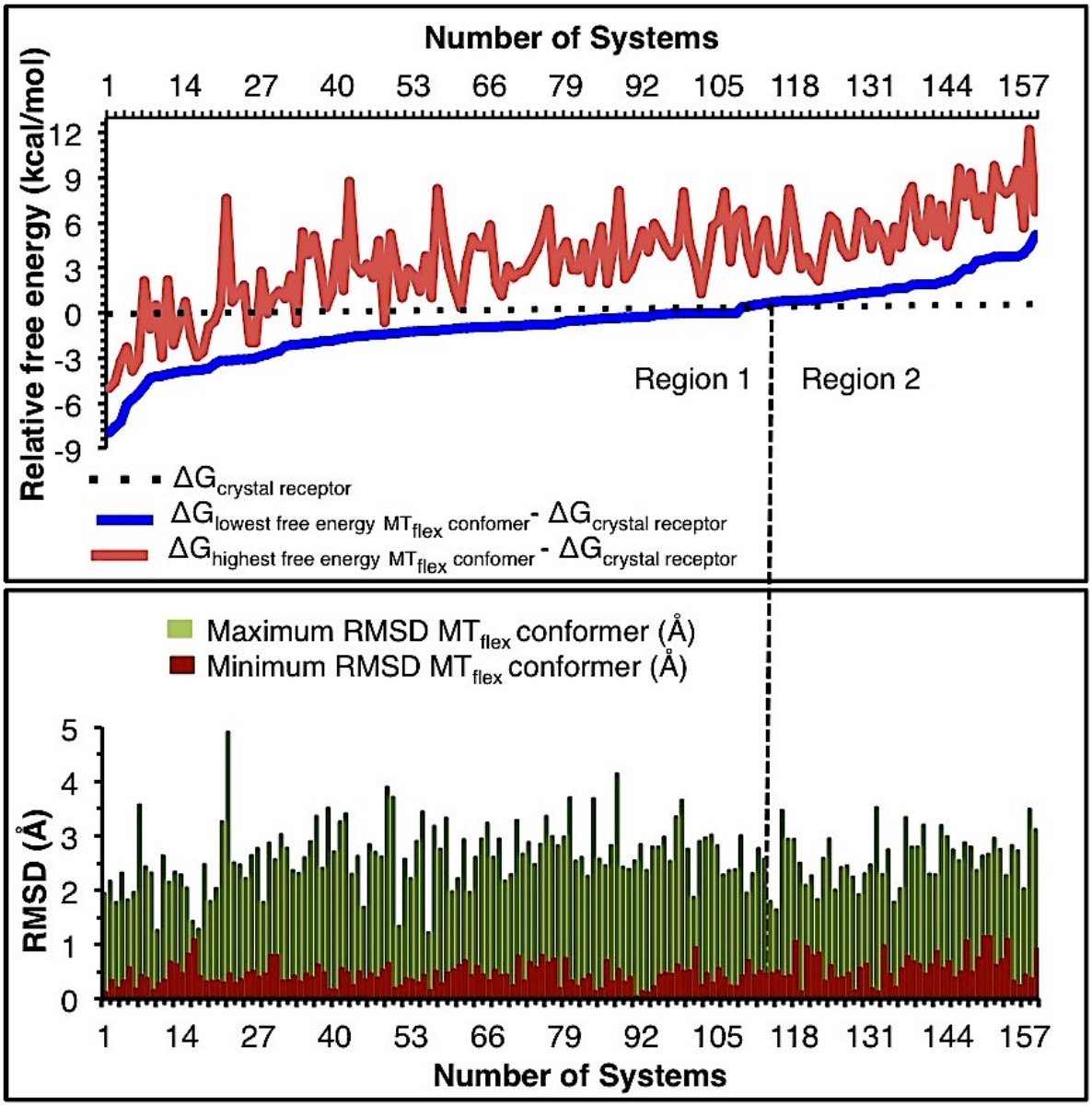

Figure 3.3 The upper panel represents the relative free energy (kcal/mol) between the lowest free energy MT_{flex} conformer and the crystal structure (blue curve) and the highest free energy MT_{flex} conformer and crystal structure (red curve) for all 159 systems in the validation dataset. The lower panel represents the corresponding RMSDs (Å) of the minimum RMSD MT_{flex} conformer (purple) superimposed with the maximum RMSD MT_{flex} conformer (green).

3.3.2. Ligand docking and scoring

After generating the MT_{flex} conformers, the next task was to dock the ligands into their respective binding pocket. Ligands were docked exhaustively into all of the receptor poses

generated by MT_{flex} for each of the 159 systems using GLIDE version 6.1 in the Schrodinger 2013-3 suite.[100-102] The Protein Preparation wizard utility, accessible from the Maestro interface of the Schrodinger 2013-3 suite was used to process the receptor structures.[100, 103] Protonation was performed at a pH-7.0.[104, 105] The hydrogen atom positions were optimized using the OPLS 2005 force field. The optimization was done in the absence of ligand.[106, 107] Crystal waters, if present, were removed from the receptor structures. Minimization was done only on the hydrogen atoms and the heavy atoms were kept fixed. Docking was done using the standard precision (SP) methodology of Glide version 6.1. The identical procedure was followed to dock ligands into the crystal structure for the purpose of comparison. The top five scored docked ligand poses were retained. In the current study, the Glide SP method was applied to ensure that a certain number of binding modes were been generated to cover as broad a range as possible of the conformational space for the following free energy simulations.

We obtained the ligand heavy atom RMSD (Å) after docking the native ligand into the binding pocket from the crystal structure and also into the MT_{flex} binding pocket(s) for all 159 systems. These RMSDs were calculated with respect to the native ligand geometry in the crystal structure. The ligand heavy atom RMSD values were obtained from the Glide report generated after docking the ligands into the respective binding pockets. Figure 3.4 gives a detailed pictorial analysis of the comparison of ligand RMSD (Å) between the ligand docked into the crystal structure and the ligand docked into the MT_{flex} generated binding pockets for all 159 systems. Ligand RMSDs reduced for ~ 78% of the systems after docking the native ligands into the MT_{flex} binding pockets as compared to the docked poses in the crystal structure binding pocket. The average RMSD of ligands over 159 systems docked into the MT_{flex} binding pocket was reduced to around 1.31 Å as opposed to the ligands docked into the crystal binding pockets with the

average being around ~2.11 Å. The reduction in average RMSD (Å) of the ligands supports the notion that inclusion of receptor flexibility is important in computer aided drug design campaigns.

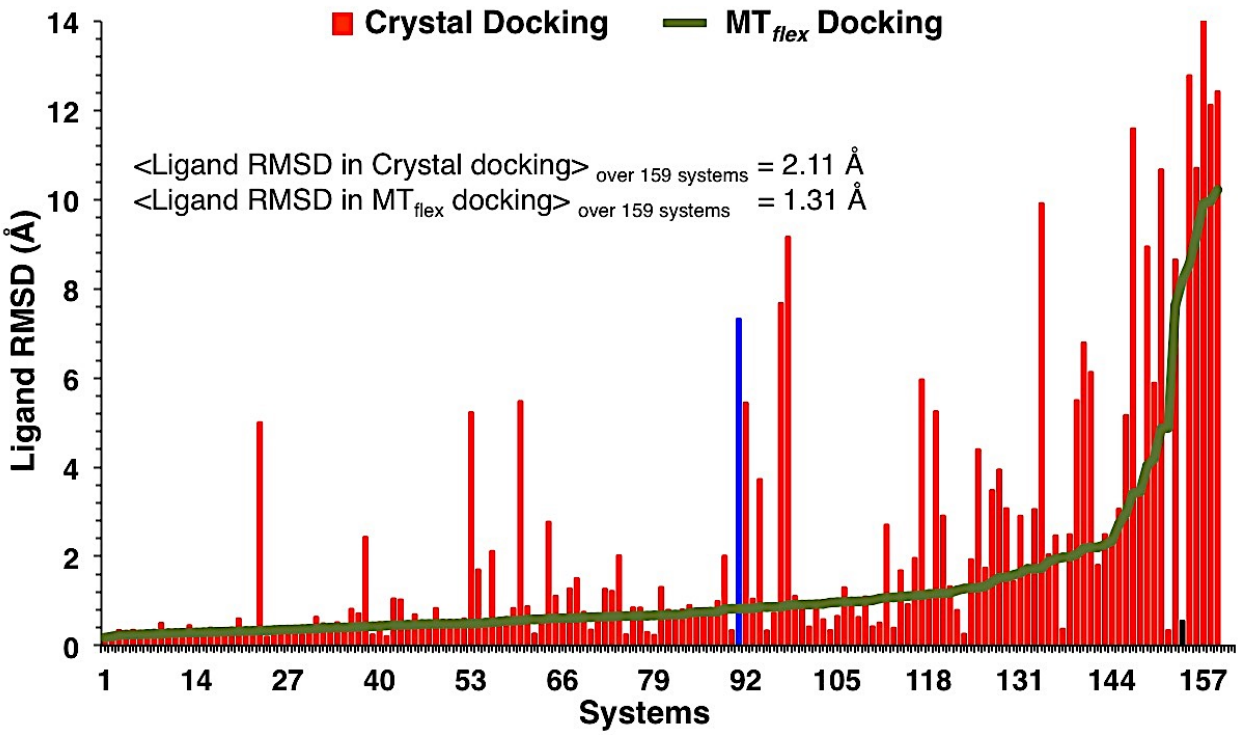

Figure 3.4 Ligand RMSD (Å) for all 159 systems after docking the native ligand into the crystal receptor's binding pocket (red) and into the MT_{flex} binding pocket (green). Lowest RMSD Conformations are used for comparison in both the datasets. The blue, yellow and the black columns represent the PDBIDs: 2R23, 3PXF and 1U33, which will be discussed in text.

A table containing the RMSD values (Å) of the minimum RMSD docked ligand conformers in both the MT_{flex} and binding pockets from the crystal structure for all 159 systems is provided in supporting information as Table 3.4. It can be seen from Figure 3.4 that for some of the systems, the RMSD difference is quite large (~7-8 Å) and two such cases are highlighted in Figure 3.4 in blue (PDBID 2R23) and black (PDBID 1U33). These two PDBIDs were picked to show the contradictory affect of docking ligand in the MT_{flex} binding pocket.

Figure 3.5 shows the superimposed images of the ligand in the crystal structure (orange), lowest RMSD docked ligand pose into the crystal structure binding pocket (blue) and the lowest RMSD docked ligand pose in the MT_{flex} binding pocket (green) for both PDBIDs. For PDBID 2R23 (left), the ligand docked in crystal structure is quite far apart from the ligand position in the crystal structure with an RMSD of ~7.3 Å. However, the ligand docked in the MT_{flex} binding pocket is nearly coincident with the ligand in the crystal structure with a RMSD of 0.83 Å. The situation is reversed for PDBID 1U33 (right side of Figure 3.5), where the lowest RMSD crystal structure docked ligand pose is well aligned to the ligand in the crystal structure with a RMSD of 0.54 Å while the lowest RMSD MT_{flex} docked ligand pose is 8.20 Å.

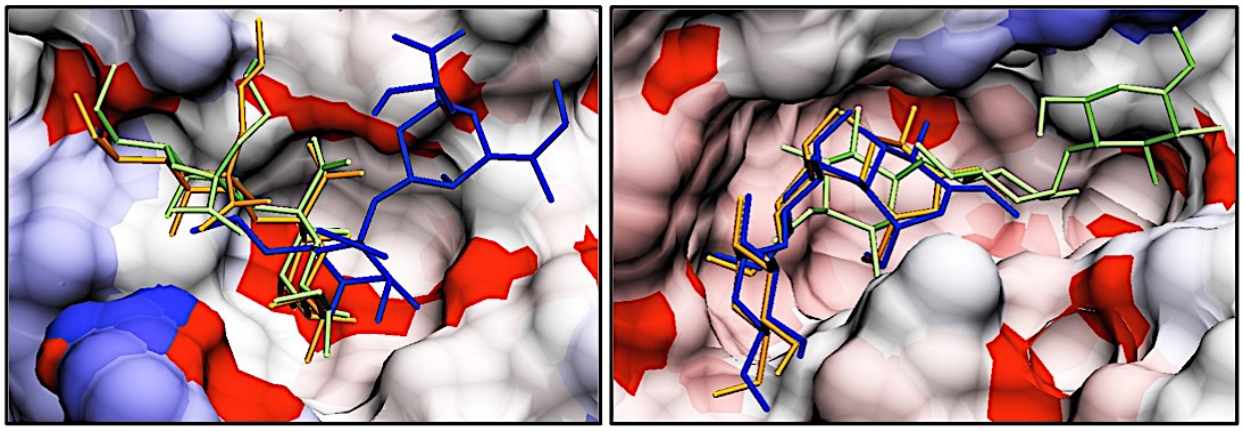

Figure 3.5 It shows the crystal complex ligand (orange), lowest RMSD docked ligand pose in the crystal binding pocket (blue) and the lowest RMSD docked ligand pose in the MT_{flex} binding pocket (green) superimposed onto each other for PDB IDs: 2R23 and 1U33 (from left to right). The background represents the binding pocket area of the respective protein.

We visualized the 2-D ligand interaction diagram (LID) in Maestro to look for the protein-ligand interactions in the crystal structure, ligand docked into the crystal structure and the ligand docked into MT_{flex} derived binding pocket for both PDBIDs. LIDs are shown in Figure 3.9 for PDBID: 2R23 (top panel) and 1U33 (lower panel). In both the panels, the left image displays

the protein ligand interactions found in the crystal structure, the center image is for the ligand docked into the crystal structure and the right image is for the ligand docked into the MT_{flex} derived binding pocket. We can establish from the LIDs that Glide is looking for a maximum number of contacts between the binding site residues and ligands in order to find an optimized position for the bound ligand. The number of contacts between the protein from the crystal structure and the docked ligand for PDB ID 2R23 (see top panel) are far more than the number of contacts found in the crystal structure. We anticipate that this is why the ligand docked in the protein from the crystal structure has a large RMSD from the position of the ligand found in the crystal structure. However, the side chains of the Arginine (Arg) residue (Resid 30) and Asparagine (Asn) (Resid 52 of chain D) are rotated in the MT_{flex} binding pocket (refer to the left panel of Figure 3.10) yielding a net RMSD of 1.86 Å for the MT_{flex} binding pocket with respect to the pocket found in the crystal structure. These rotations have established a new H-bond between the side chain of the Asn and the ligand and have restored the original salt bridge between the Arg and the ligand, keeping the position of the docked ligand nearly the same as found for the ligand in the crystal complex. For PDBID 1U33 (lower panel), Glide identifies very similar interactions between the docked ligand and crystal protein as that of the crystal complex, thereby finding the optimized position of the docked ligand almost overlapping with the ligand in crystal complex with the RMSD of 0.54 Å. But in the MT_{flex} binding pocket, even for the lowest RMSD MT_{flex} conformer (0.33 Å), the side chain of the Glutamate (Glu) residue (Resid 233) is tilted slightly such that it is blocking the position of the native ligand pose found in the crystal structure (see the right panel of Figure 3.10). Due to the slight change of the Glu residue coupled with few other side chain rotations, the docked position of the ligand for the MT_{flex} binding pocket is far from that found for the ligand in the crystal structure.

For further comparison, we correlated the Glide scores of the docked crystal and MT_{flex} ligands with the experimental binding affinities (pK_d/pK_i). To be consistent with the energy unit used by Glide, we converted experimental pK_d/pK_i values to ΔG in kcal/mole. We compared Glide's score for the best docking solution for both cases across all 159 systems. Correlation of Glide's score for the top docking solution in the crystal structure generated a Pearson's R of 0.39 and a Spearman's rank correlation coefficient (Spearman's rho) of 0.43 while the correlation of Glide's top docking score across all MT_{flex} conformations generated a Pearson's R of 0.47 and a Spearman's rho of 0.50. The Root mean square error (RMSE) (kcal/mol) reduced from 3.21 to 2.70 kcal/mol and the Mean unassigned error (MUE) (kcal/mol) from 2.34 to 2.17. The table listing all the statistical parameters for the best solution using Glide's score is provided in Table 3.5 of the supporting information. The correlation between Glide's top docking score for the crystal receptor and for the MT_{flex} conformers generated a Spearman's Rho of 0.82 and Pearson's R of 0.86. Overall, the Glide score for docking the ligand into MT_{flex} binding pocket improved in terms of both error and correlation when compared to docking the ligand into the crystal structure. By this analysis, it was observed that MT_{flex} generated multiple conformers serve as better input than a single crystal structure "seed" on which ligand scoring and docking is performed.

We also re-ranked Glide's docking solutions based on the relative free energy scale obtained for the MT_{flex} conformations in the ligand-unbound states. After re-ranking, we obtained the correlation between the Glide's score for the lowest free energy MT_{flex} conformer (in the ligand-unbound state) and the experimental binding affinities (the statistical parameters are present in Table 3.5). The correlation was quite poor with Pearson's R and Spearman Rho values of 0.37 and 0.38, respectively. The free energy errors were also higher with a RMSE of

3.17 kcal/mol and MUE of 2.43 kcal/mol. All the statistical parameters for correlation were poor as compared to the Glide's best score for crystal and MT_{flex} docking, which is understandable because the ranking of conformations on the free energy scale can be entirely different before and after ligand binding. A ligand does not necessarily have the strongest affinity towards the lowest free energy receptor conformation. It can bind with more affinity to a higher free energy conformation on a free energy scale and change the overall ranking of the conformations in the bound state as compared to the ligand unbound state. Figure 3.6 shows the relative free energy between the MT_{flex} lowest free energy conformation in the ligand bound state and the MT_{flex} lowest free energy conformations in the ligand-unbound state for all 159 systems of the validation dataset. The MT_{flex} lowest free energy conformation in the ligand-unbound state is the reference state (0 kcal/mol) for all systems. Based on our analysis, we found that only for ~31% of the systems, the MT_{flex} lowest free energy conformer in the ligand bound state was also the lowest free energy MT_{flex} conformer in the unbound state. However, no particular trend was observed. To give a pictorial overview of the free energy scales, we include Figure 3.11 in supporting information. It shows the position of the MT_{flex} lowest free energy conformation in the ligand bound state on the MT_{flex} free energy scale generated for four random systems from our validation dataset (PDBIDs: 10gs, 2fvd, 2qbr and 2zjw). It can be also be seen from Figure 3.11 that the position of the lowest free energy conformation in the ligand bound state (marked as a red square) is different for each of the systems and does not show any particular preference to a particular rank on the free energy scale.

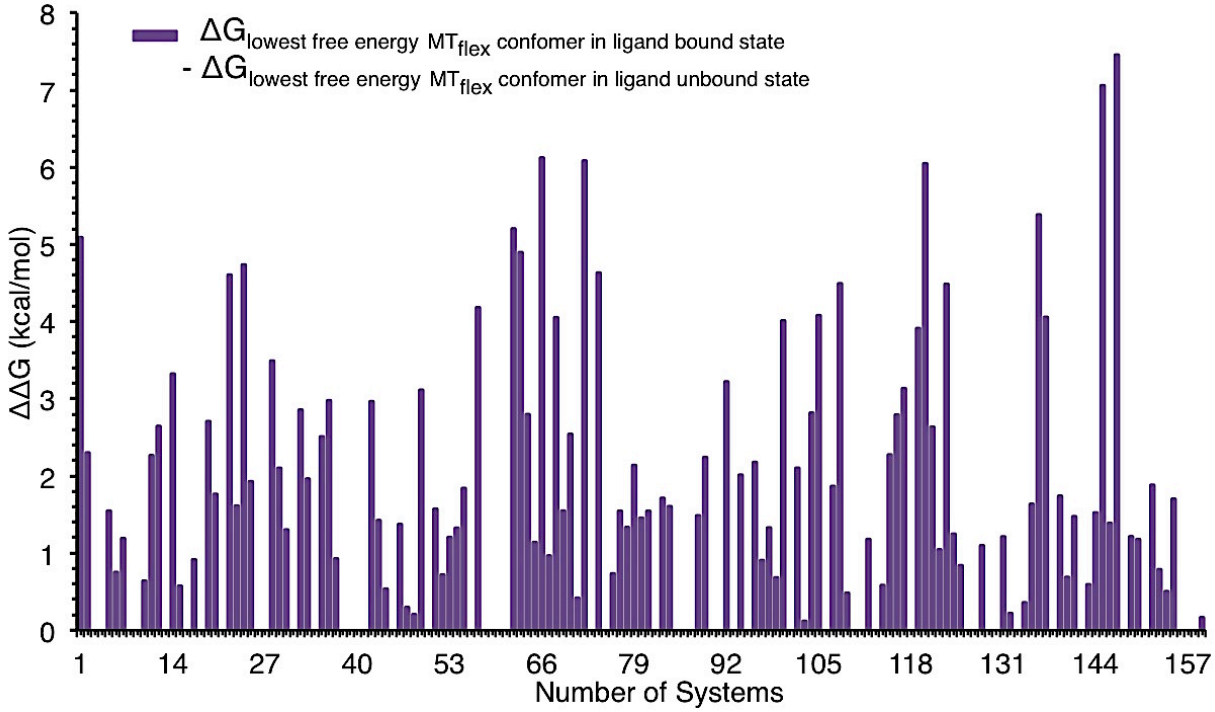

Figure 3.6 The position of the MT_{flex} lowest free energy conformation in the ligand bound state on the free energy scale obtained for the MT_{flex} conformations in the ligand-unbound state is shown. The lowest free energy MT_{flex} conformation in the unbound state is used as reference (set to 0 kcal/mol).

3.3.3. Multi seed versus rigid receptor free energy calculations

Using the MT method, we calculated the binding affinity (ΔG in kcal/mole) of all the protein-ligand complexes obtained after docking ligands into the MT_{flex} generated binding pockets and also into the crystal structure pockets. We also calculated the binding affinity of the crystal protein-ligand complex for the entire validation dataset and incorporated it as one of the conformers of the MT_{flex} ensemble. For the sake of simplicity, we will denote the experimental protein-ligand complex as P_CL_C, the crystal protein-docked ligand complex as P_CL_D and the MT_{flex} protein docked ligand complex as $P_{MT}L_D$.

We calculated binding affinities of $<P_CL_D>$ (average over all the docked solutions of the crystal protein docked ligand complex) and $<P_{MT}L_D>$ (average over all the docked solutions of

each MT_{flex} conformer) for each of the 159 systems from the validation dataset and correlated them with the experimental ΔG values. The binding affinity prediction for $<P_CL_D>$ yielded a Pearson's R of 0.58 and Spearman's Rho of 0.60. The correlation of the average binding affinity for $P_{MT}L_D$ including the full range ($<P_{MT}L_D>$) showed an improvement over the crystal structure docked structures in all statistical parameters including Pearson's R, Spearman's Rho, RMSE and MUE (see Table 3.1). The correlation coefficient, Pearson's R showed a modest improvement with a value of 0.65 and Spearman's Rho with a value of 0.69. The errors decreased with the RMSE dropping from 3.37 to 2.83 kcal/mol and the MUE went from 2.74 to 2.27 kcal/mol. The RMSE and MUE drops are significant observations because they affirm the relevance of our sampling space. Figure 3.7 shows the superposition of the binding affinities (kcal/mol) of $<P_CL_D>$ and $<P_{MT}L_D>$ along with the experimental binding affinities for all 159 systems in the validation set. We observed that for approximately 63.3% of the validation dataset, the binding affinities of the $<P_CL_D>$ complexes lie within ±3 kcal/mol of the experimental binding affinities while the percentage increases to 75.4% for the $<P_{MT}L_D>$ complexes. The overall improvement in $<P_{MT}L_D>$ binding affinity prediction over the $<P_CL_D>$ complex highlights the significance of using a multi-seed strategy over single receptor structures in MT based binding affinity prediction.

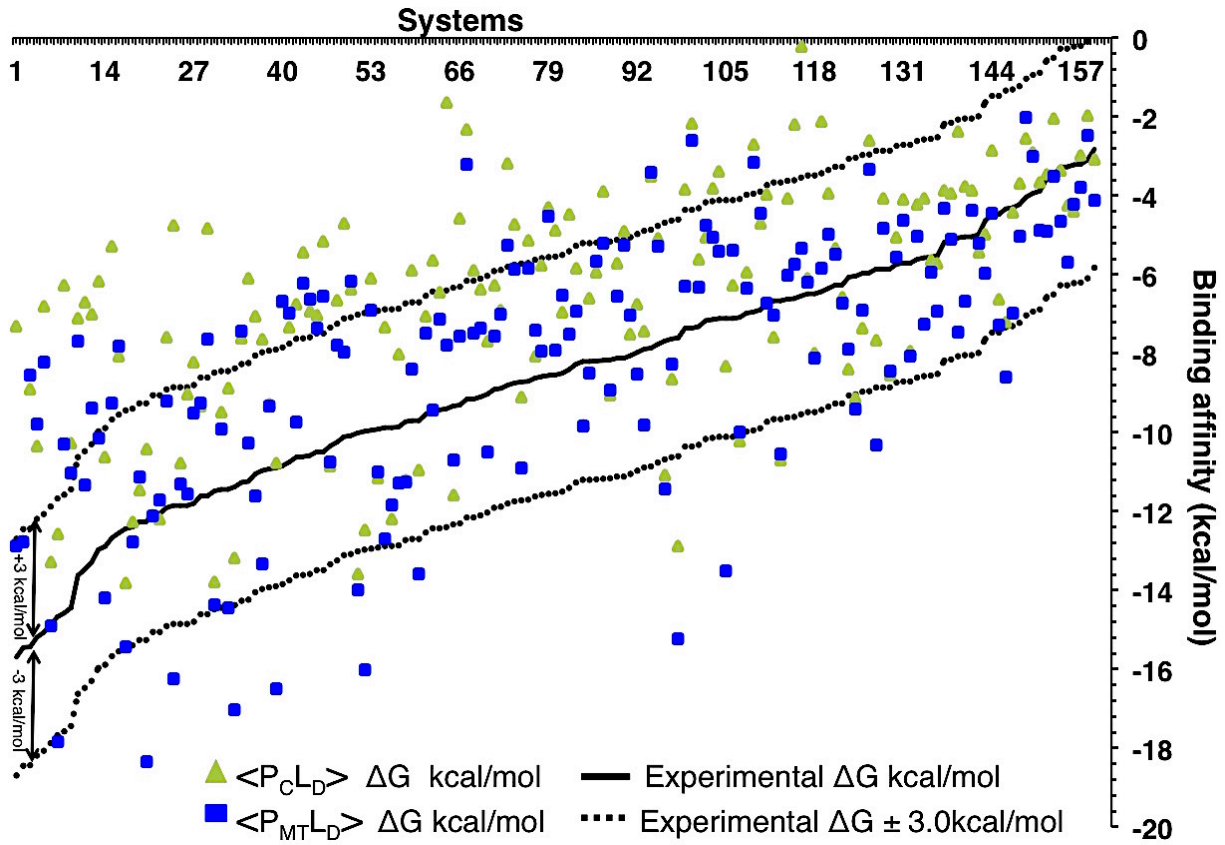

Figure 3.7 It shows the absolute binding affinities (kcal/mol) of $<P_CL_D>$ and $<P_{MT}L_D>$ superimposed with the experimental binding affinity (kcal/mol). ± 3kcal/mol error window is extended for the experimental binding affinities.

As already described in the method section, our method is based on a Monte Carlo integration. The accuracy of any MCI approach relies on the selection of the structural ensemble. The improvement of the $<P_{MT}L_D>$ binding affinity prediction validates that our sampling set does possess some of the key states of the conformational ensemble beyond the crystal structure. Nonetheless, some less significant conformations are also generated and the overall calculation would benefit from their elimination. The goal here was to find a converged ensemble out of the larger pool of conformers generated by MT_{flex}. To prune out the high-energy conformers and/or the very low populated states from our sampling space, we calculated the free energy difference ($\Delta\Delta G_{bind}$) between all the docked solutions of each MT_{flex} conformer in the $P_{MT}L_D$ complexes

($P_{MT}L_{D\text{-}All}$) and the P_CL_C for each system using equation 3.1, where n refers to the number of complexes in $P_{MT}L_{D\text{-}All}$ for each system in the validation dataset.

$$\Delta\Delta G^{n}_{bind} = \Delta G^{n}_{P_{MT}L_D} - \Delta G_{P_CL_C} \quad (3.1)$$

The free energy difference provided us with a scale to rank our conformations from the lowest to highest binding affinities with respect to the crystal structure. The crystal protein ligand complex was incorporated into the $P_{MT}L_{D\text{-}All}$ ensemble as one of the conformers. The free energy scale for all the 159 systems of the validation dataset is represented in Figure 3.8. The lowest free energy MT protein ligand complex ($P_{MT}L_{D\text{-}Lowest}$) for each system is highlighted by the orange curve. Based on the free energy scale, approximately 70.4% of the systems in the validation dataset have lower free energy $P_{MT}L_{D\text{-}Lowest}$ conformations (lower free energy binding modes) than the crystal complex. The observed free energy differences are not remarkable owing to the fact that we only introduced side chain flexibility. However, we hypothesize that if we add backbone flexibility on top of side chain flexibility, we would likely span a broader free energy range. We compared the binding affinity of the lowest free energy MT protein ligand complex ($P_{MT}L_{D\text{-}Lowest}$) by correlating it with the experimental binding affinity (kcal/mol). The correlation parameter for $P_{MT}L_{D\text{-}Lowest}$ showed a slight improvement over the prediction for the $<P_CL_D>$ complex with the Pearson's R incrementing from 0.58 to 0.64 and Spearman's Rho from 0.60 to 0.68. The errors were reduced substantially with the RMSE reducing from 3.37 to 2.72 kcal/mol and the MUE dropping from 2.74 to 2.14 kcal/mol. In comparison to the $<P_{MT}L_D>$, the correlation trends for $P_{MT}L_{D\text{-}Lowest}$ did not show any improvement but both the RMSE and MUE (kcal/mol) reduced.

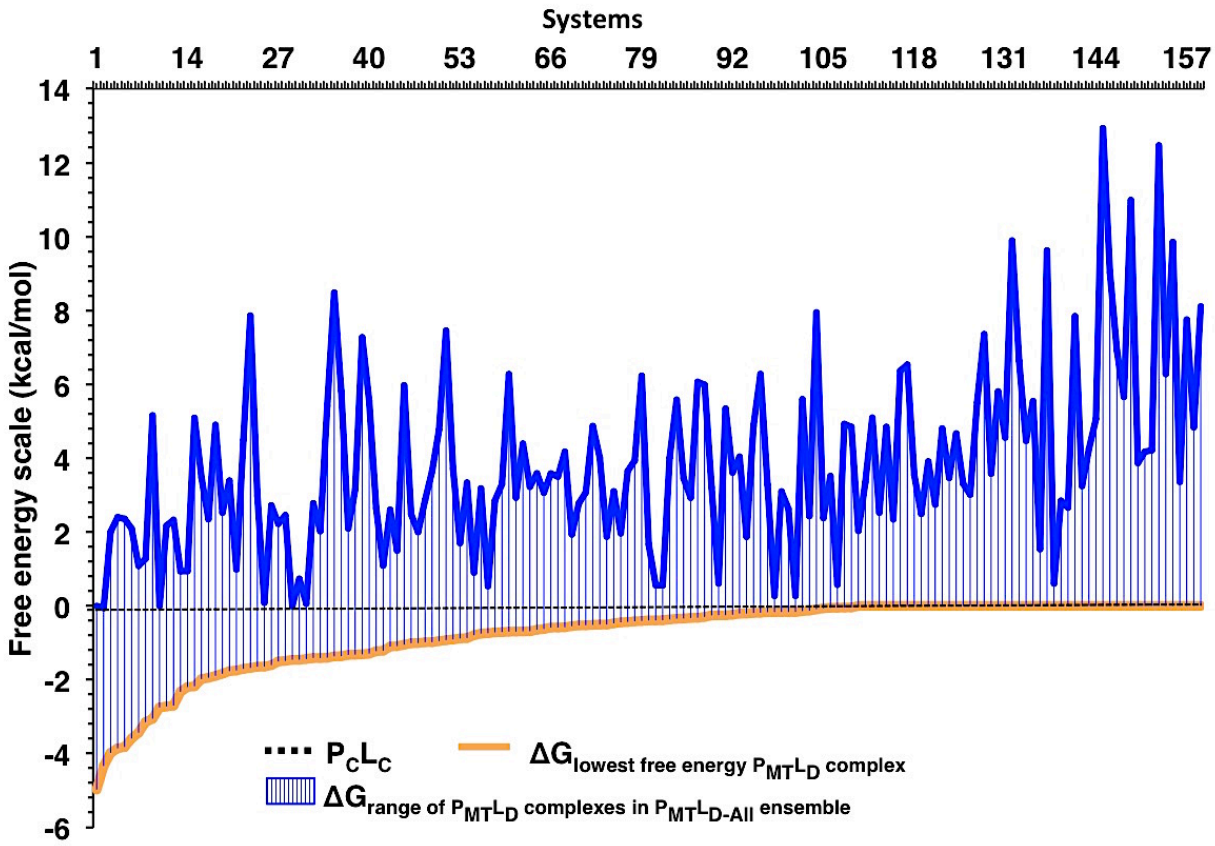

Figure 3.8 Free energy scale of $P_{MT}L_D$ complex for all 159 systems in the validation dataset. Approximately ~70.4% of the systems have lower free energy than the crystal complex.

Based on the free energy difference scale, we also selected several subsets of the conformations and correlated them with the experimental binding affinities. The subsets were formed by collecting PL complexes starting from the lowest up to a selected free energy difference endpoint from the crystal structure. We calculated the average binding affinity for all of the subsets and referred to them as $P_{MT}L_D$-set 1 to set 5, where $P_{MT}L_D$-set1 represents ΔG's up to 1 kcal/mole higher in free energy than the free energy of the crystal structure, set 2 denotes ΔG's up to 2 kcal/mole higher in free energy and so on. The selection of set 1 to set 5 is a qualitative approach to explore how the ensemble affects the computed results. In terms of correlation, $P_{MT}L_D$ set 1-5 generated nearly identical trends with the Pearson's R correlation coefficient being ~0.65 and the Spearman's Rho being ~0.69. The errors increased slightly on

going from $P_{MT}L_D$ set 1-5 as more noise was added to the ensemble. The errors had small differences between set 1 and 2 and between set 3,4 and 5, so we have only shown $P_{MT}L_D$ set 2 and set 4 in Table 3.1 to give a general sense of the outcome. It can be seen from Table 3.1 that the lowest free energy MT complex ($P_{MT}L_{D\text{-}Lowest}$) gives the lowest RMSE and MUE (kcal/mol) from the experimental binding affinities. After that there is a slight increase in both RMSE and MUE as the cut off is increased for all of the $P_{MT}L_D$ ensembles. For the $P_{MT}L_D$ set 1-5, the error reduction is better in comparison to the binding affinity prediction found for $<P_{MT}L_D>$ (including the full range). Since, it is an ensemble-based study, error reductions are a more accurate measure to assess the precision of any method rather than the correlation coefficients.

Table 3.1 Statistical data for the ΔG'S evaluated using the crystal protein docked ligand complex, MT_{flex} lowest free energy complex, MT_{flex} with conformations selected up to 2 kcal/mol and 4 kcal/mol higher in binding affinity than the crystal structure and MT_{flex} average (including the full range) relative to the experimental binding affinity.

	Pearson's R	Spearman's rho	RMSE (kcal/mol)	MUE (kcal/mol)
$<P_CL_D>$	0.58	0.60	3.37	2.74
$P_{MT}L_{D\text{-}Lowest}$	0.64	0.68	2.72	2.14
$P_{MT}L_{D\text{-}set2}$	0.65	0.69	2.75	2.21
$P_{MT}L_{D\text{-}set4}$	0.64	0.69	2.77	2.23
$< P_{MT}L_D >$	0.65	0.69	2.83	2.27

The results for $P_{MT}L_D$ ensembles are certainly far better both in terms of error and correlation than single docking into a receptor site (P_CL_D). Overall, the statistical comparison of the $P_{MT}L_{D\text{-}Lowest}$ and $P_{MT}L_D$ sets with the experimental binding affinity enhanced the improvement over the $<P_{MT}L_D>$ results (including the full range) indicating that by eliminating the low probability conformations we can improve the accuracy of the modeling. This analysis further strengthens our hypothesis that with MT_{flex}, we can obtain better binding modes, which are difficult to attain with single receptor docking.

An ensemble of conformations can be easily generated with the major sampling methods available, but the problem lies in the identification of the conformations that have the lowest free energy and, hence, contribute the most to the ensemble. With MT_{flex}, the identification of the lowest free energy conformations becomes feasible. The improved correlation and reduced RMSE (kcal/mole) after elimination of the higher energy conformations further illustrates the importance of generating relevant conformations. It also nicely demonstrates the significance of incorporating receptor flexibility yielding multiple seed structures that can be used in MT scoring. Moreover this multi-seed strategy shows significant improvement over traditional rigid/crystal structure docking.

3.4 Conclusions

Proteins undergo a wide variety of motions ranging from ultrafast vibrations to long-range backbone motions. All these motions help in generating multiple conformations of the protein, thus giving an ensemble of structures. Consideration of all the relevant receptor poses is extremely important in the field of structure based drug design to garner a detailed picture of the binding pocket and its interactions with guest molecules. MT_{flex}, unlike other sampling methods[15, 19, 20, 27, 28, 30-33, 36, 45, 58-60, 67, 84, 108], generate conformations on a free energy surface and as we show these structures seem to impart relevance to binding free energy prediction. We calculate the Boltzmann factor contribution of each torsional interaction, which allows us to create the ensemble on a free energy surface. Thus, along with the crystal structure, we can include several relevant (*i.e.*, low free energy) receptor poses as the starting point or seed for calculating protein-ligand binding free energies.

By using MT_{flex} generated multiple conformers, we have shown that Glide generates

better binding modes (lower RMSD conformers with respect to the ligand in crystal complex) and improved correlation with experimental binding affinities by using MT_{flex} generated multiple conformers rather than a single crystal structure "seed". Apart from Glide scoring, we also used our MT scoring method and validated that scoring results improve via the inclusion of the MT_{flex} generated ensemble. Overall, via side chain conformational sampling, we have shown that docking with multiple receptor poses or seeds gives better correlation with the experimental data than docking with a single structure. This viewpoint is consistent with the consensus docking viewpoint, but with one scoring function generating multiple samples. [109]

The two major motions that occur in the protein ligand binding process are side chain and backbone movements. Herein, we have only incorporated side-chain flexibility via generation of conformations on the free energy surface. The addition of backbone flexibility should also improve the quality of our computed free energies of binding, but offer computational complexities beyond side chain sampling.

3.5 Acknowledgement

NB would like to thank Mr. Pengfei Li for fruitful discussions. The authors would like to thank high performance computing center (HPCC) at Michigan State University for providing computational resources.

3.6 Supplementary information

Table 3.2 Average CPU time (in seconds) to generate MT_{flex} conformers for each Amino acid

Amino acid (3 letter code)	CPU time (seconds)
Ile	2.11
Leu	30.83
Val	3.40
Phe	222.24
Trp	223.55
Tyr	217.52
Asp	37.80
Glu	1008.62
Gln	784.53
Arg	5449.91
His	5.36
Lys	2649.07
Ser	1.00
Thr	2.67
Cys	4.52
Met	1140.16
Asn	165.75

Table 3.3 Minimum and Maximum RMSDs of MT_{flex} conformers along with the number of conformers retained for all 159 systems of the validation dataset.

PDBID	Minimum RMSD (Å)	Maximum RMSD (Å)	Number of MT_{flex} conformers retained
10gs	0.49	3.41	255
1a30	0.20	2.47	10
1bcu	0.47	4.91	4
1e66	0.29	1.27	3
1f8b	1.07	2.88	10
1f8c	1.15	2.67	14
1f8d	1.15	2.63	15
1gpk	0.68	2.15	4
1h23	0.80	2.87	19
1hnn	0.43	2.94	18
1igj	0.79	3.29	4
1jyq	0.56	3.26	40

Table 3.3 (cont'd)

1kel	0.43	1.97	2
1lbk	0.43	2.95	19
1lol	0.47	2.97	15
1loq	0.71	2.46	35
1lor	0.23	2.54	31
1mq6	0.79	3.34	36
1n1m	0.56	2.04	7
1n2v	0.14	3.69	20
1nvq	0.40	2.42	12
1o3f	0.29	2.76	6
1o5b	0.25	2.30	20
1oyt	0.16	2.32	15
1p1q	0.87	2.29	4
1q8t	0.51	2.77	4
1q8u	0.67	3.36	18
1qi0	0.36	2.47	22
1r5y	0.17	3.51	22
1sqa	0.63	2.34	6
1u1b	0.34	2.04	14
1u33	0.33	1.80	8
1uto	0.39	2.70	7
1vso	0.95	1.87	4
1w3k	0.34	2.32	12
1w3l	0.35	2.17	7
1w4o	0.40	2.77	49
1xd0	0.35	2.64	33
1yc1	0.14	2.84	13
1z95	0.40	2.39	40
1zea	0.43	2.37	2
2brb	1.10	2.28	528
2cbj	0.54	4.14	116
2cet	0.35	2.22	14
2d3u	0.34	3.70	30
2fvd	0.60	2.61	81
2g70	0.63	2.30	12
2gss	0.34	3.23	53
2hb1	0.39	2.74	17
2iwx	0.29	3.02	20
2j62	0.45	2.95	34
2j78	0.45	2.26	3

Table 3.3 (cont'd)

2jdm	0.98	2.30	4
2jdu	0.24	2.37	12
2jdy	0.05	2.54	16
2obf	0.46	2.74	37
2ole	0.62	2.96	59
2p4y	0.96	2.09	6
2pcp	0.58	1.82	20
2pq9	0.44	2.03	3
2qbp	0.24	2.73	23
2qbr	0.38	3.49	148
2qft	0.50	2.55	12
2qmj	0.36	1.69	39
2r23	0.83	2.05	19
2v00	0.19	2.58	12
2vl4	0.19	1.96	18
2vo5	0.21	1.78	8
2vot	0.12	1.94	29
2vvn	0.38	2.01	12
2vw5	0.44	2.80	8
2w66	0.23	2.80	18
2wbg	0.52	1.65	3
2wca	0.20	2.82	4
2wtv	0.61	2.95	12
2x00	0.44	3.57	61
2x0y	0.34	3.02	20
2xb8	1.06	2.94	10
2xbv	0.43	2.31	4
2xdl	0.52	2.76	18
2xnb	0.75	2.98	17
2xys	0.32	2.48	17
2y5h	0.30	2.90	11
2yfe	0.46	3.20	15
2yge	0.36	2.61	7
2yki	0.12	2.37	5
2ymd	0.38	2.44	4
2zjw	0.63	3.37	80
2zwz	0.46	1.78	4
2zx6	0.63	2.22	2
2zxd	0.48	2.22	24
3acw	0.25	2.29	6

Table 3.3 (cont'd)

3ag9	0.73	2.76	14
3ao4	0.64	2.80	13
3b68	0.44	2.17	10
3bfu	0.47	2.98	18
3bpc	1.09	1.43	4
3cft	0.20	3.71	176
3cj2	0.44	3.44	35
3coy	0.40	2.90	29
3cyx	0.58	2.48	8
3dxg	0.29	2.51	35
3e93	0.15	2.25	11
3ebp	0.74	2.99	5
3f3a	0.22	1.79	2
3f3c	0.16	1.23	2
3f3e	0.24	1.34	3
3fk1	0.69	2.80	18
3fv1	0.64	2.30	3
3g0w	0.46	2.53	8
3g2n	0.50	2.63	20
3g2z	0.47	1.80	4
3gbb	0.69	2.99	5
3gcs	0.45	2.58	11
3ge7	0.56	2.82	16
3gnw	0.32	2.83	14
3gy4	0.43	3.00	14
3huc	0.51	3.66	32
3imc	0.66	3.90	47
3ivg	0.51	3.18	125
3jvs	0.48	3.33	53
3k5v	0.49	2.41	7
3kgp	0.14	3.52	7
3l4u	0.85	1.83	6
3l4w	0.42	1.29	14
3l7b	0.54	2.62	59
3mss	0.34	2.67	14
3myg	0.57	3.19	21
3n7a	0.33	2.83	12
3n86	0.92	3.12	7
3nox	0.80	2.27	26
3nq3	0.76	2.37	10

Table 3.3 (cont'd)

3ov1	0.46	2.60	12
3owj	0.80	2.57	24
3pe2	0.47	2.29	34
3pww	0.56	1.93	3
3pxf	0.68	2.88	60
3s8o	0.38	2.57	7
3su2	0.17	2.71	25
3su3	0.46	2.84	23
3su5	0.71	2.93	16
3u9q	0.80	2.85	10
3udh	0.52	2.64	8
3ueu	0.49	2.80	66
3uex	0.39	2.29	5
3uo4	0.41	3.47	47
3uri	0.31	2.31	9
3utu	0.30	3.26	24
3zso	0.26	2.90	10
3zsx	0.31	2.43	21
4de1	0.54	1.98	4
4de2	0.54	2.61	4
4des	0.47	2.45	12
4dew	0.14	2.50	17
4djr	0.71	1.95	6
4djv	0.35	2.78	6
4g8m	0.33	2.59	11
4gid	0.23	2.38	7
4gqq	0.63	3.35	28

Table 3.4 Minimum RMSDs of the docked ligand conformers in both the crystal and MT_{flex} binding pockets of all 159 systems of the validation dataset.

PDBID	Crystal ligand RMSD (Å)	MT_{flex} RMSD (Å)
10gs	0.84	0.67
1a30	2.90	1.63
1bcu	0.34	0.25
1e66	0.49	0.40
1f8b	0.33	0.86
1f8c	0.25	1.28
1f8d	0.58	0.93

Table 3.4 (cont'd)

1gpk	0.39	0.30
1h23	5.49	2.05
1hnn	1.04	0.46
1igj	0.92	1.11
1jyq	9.91	1.73
1kel	1.18	1.16
1lbk	0.53	0.51
1lol	2.70	1.08
1loq	0.83	0.49
1lor	0.19	0.34
1mq6	0.50	1.05
1n1m	0.56	0.49
1n2v	2.00	0.82
1nvq	0.25	0.25
1o3f	1.21	0.65
1o5b	1.02	0.46
1oyt	0.69	0.47
1p1q	0.81	0.41
1q8t	0.99	0.99
1q8u	0.54	0.54
1qi0	5.88	4.19
1r5y	5.00	0.34
1sqa	0.59	0.56
1u1b	8.94	4.03
1u33	0.54	8.20
1uto	1.26	0.65
1vso	3.93	1.52
1w3k	0.25	0.20
1w3l	0.16	0.29
1w4o	7.67	0.87
1xd0	8.65	7.63
1yc1	0.40	0.36
1z95	0.38	0.38
1zea	12.78	8.56
2brb	0.59	0.49
2cbj	1.67	1.09
2cet	0.51	0.48
2d3u	0.47	0.47
2fvd	0.36	0.30
2g70	0.42	1.02

Table 3.4 (cont'd)

2gss	3.05	1.72
2hb1	1.10	0.92
2iwx	0.19	0.18
2j62	0.29	0.29
2j78	0.63	0.38
2jdm	3.07	1.53
2jdu	0.29	0.67
2jdy	0.33	0.83
2obf	1.09	1.00
2ole	2.46	1.94
2p4y	0.79	0.69
2pcp	0.51	0.40
2pq9	0.63	0.99
2qbp	1.95	1.12
2qbr	0.98	0.76
2qft	1.32	1.18
2qmj	5.44	0.83
2r23	7.32	0.83
2v00	1.50	0.62
2vl4	0.50	0.26
2vo5	0.48	0.40
2vot	1.70	0.49
2vvn	0.26	0.26
2vw5	0.39	0.48
2w66	0.28	0.32
2wbg	2.38	2.35
2wca	5.16	2.97
2wtv	0.84	0.67
2x00	0.49	0.49
2x0y	0.29	0.31
2xb8	0.39	1.08
2xbv	0.32	0.24
2xdl	5.24	1.17
2xnb	0.83	0.56
2xys	0.33	0.41
2y5h	0.33	0.33
2yfe	0.61	0.61
2yge	0.67	0.76
2yki	1.92	1.29
2ymd	0.44	0.29

Table 3.4 (cont'd)

2zjw	3.72	0.86
2zwz	0.26	0.59
2zx6	2.03	1.88
2zxd	0.27	0.35
3acw	0.35	0.64
3ag9	14.84	9.92
3ao4	0.34	0.31
3b68	0.93	0.93
3bfu	1.04	0.83
3bpc	11.59	3.41
3cft	0.56	0.49
3cj2	0.40	0.36
3coy	2.77	0.61
3cyx	5.47	0.56
3dxg	6.12	2.20
3e93	0.60	0.59
3ebp	0.34	4.88
3f3a	0.87	0.58
3f3c	0.41	0.92
3f3e	0.71	0.42
3fk1	3.47	1.43
3fv1	0.76	0.75
3g0w	0.20	0.45
3g2n	0.74	0.63
3g2z	3.60	3.45
3gbb	1.10	0.61
3gcs	0.24	0.43
3ge7	0.37	1.98
3gnw	0.34	0.24
3gy4	0.23	0.68
3huc	0.51	0.64
3imc	0.34	0.26
3ivg	5.96	1.13
3jvs	2.01	0.66
3k5v	2.48	1.99
3kgp	0.94	0.92
3l4u	1.73	1.33
3l4w	1.80	2.20
3l7b	0.51	0.44
3mss	2.49	2.26

Table 3.4 (cont'd)

3myg	1.78	1.72
3n7a	0.24	0.66
3n86	0.72	0.87
3nox	0.75	0.75
3nq3	3.05	2.73
3ov1	0.79	1.24
3owj	5.22	0.49
3pe2	0.25	0.37
3pww	10.70	9.21
3pxf	9.15	0.90
3s8o	0.90	0.73
3su2	0.28	0.28
3su3	0.36	0.36
3su5	0.33	0.33
3u9q	1.27	0.62
3udh	0.59	0.32
3ueu	2.90	1.17
3uex	1.43	1.58
3uo4	0.71	0.69
3uri	12.42	10.22
3utu	2.11	0.52
3zso	0.80	0.69
3zsx	4.39	1.30
4de1	0.65	0.97
4de2	1.30	0.68
4des	2.42	0.43
4dew	6.78	2.18
4djr	10.67	4.86
4djv	0.29	0.28
4g8m	1.29	0.98
4gid	12.12	9.97
4gqq	0.34	0.96

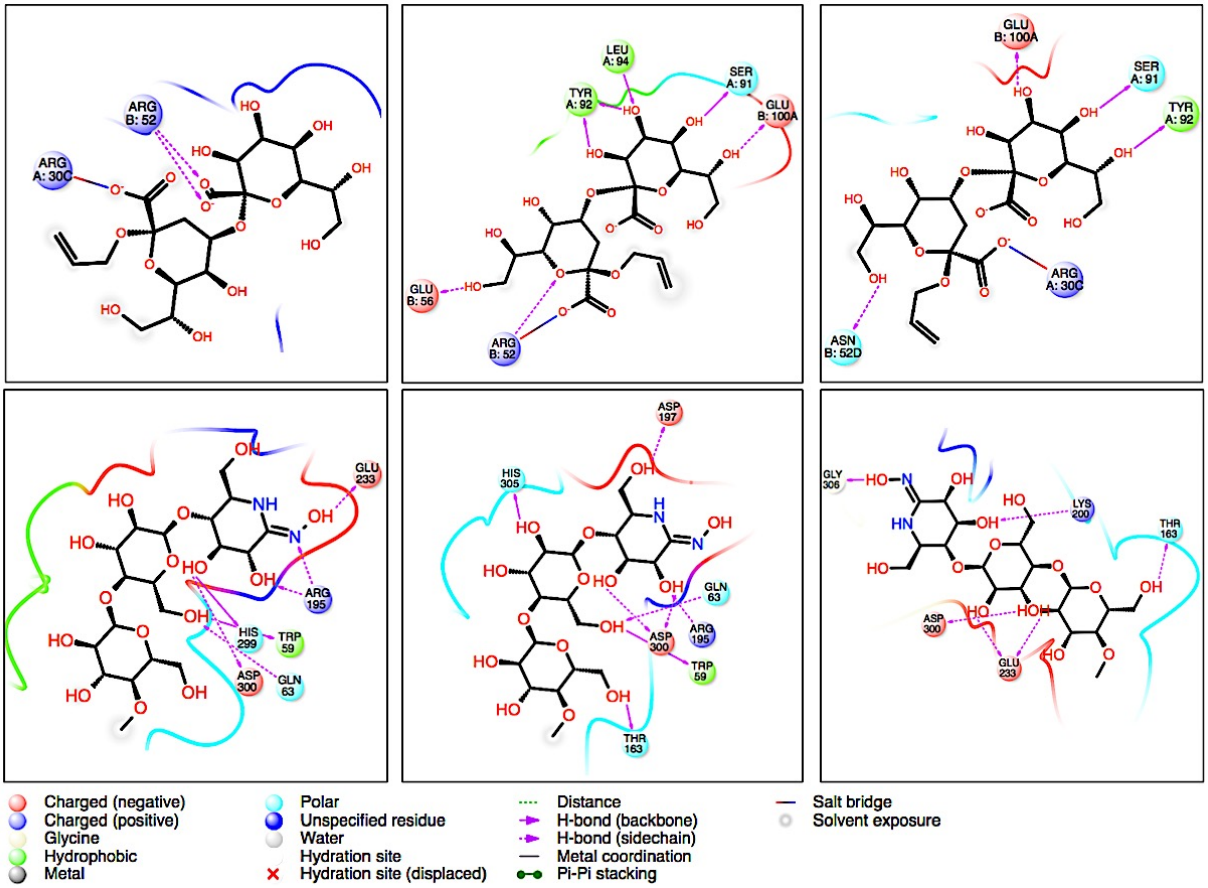

Figure 3.9 Ligand Interaction diagrams of PDBIDs: 2R23 (top panel) and 1U33 (lower panel). Crystal complex is shown on the left side, crystal protein docked ligand (center) and the MT_{flex} protein docked ligand in right.

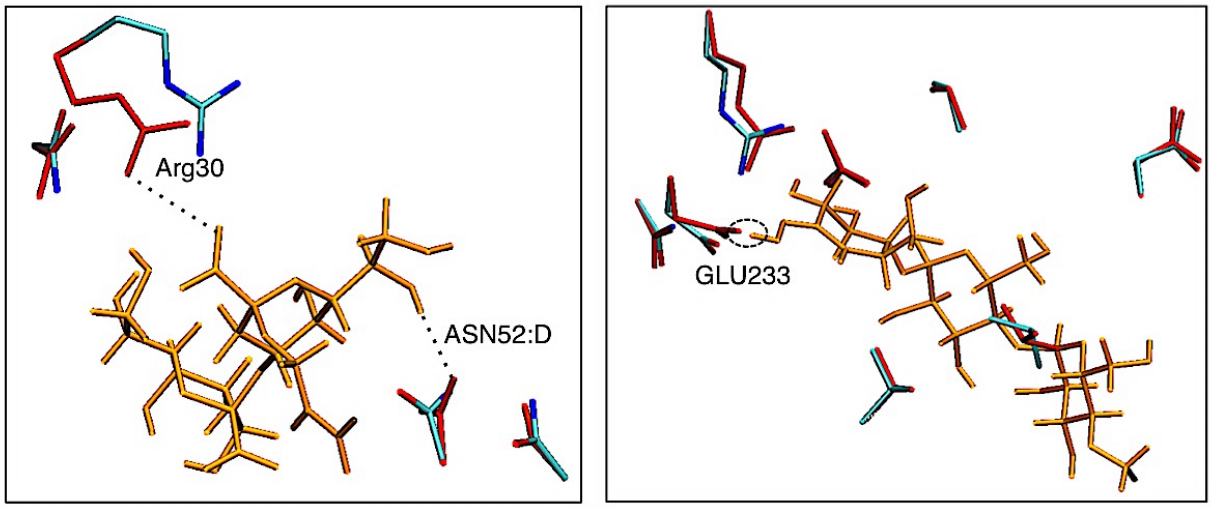

Figure 3.10 The superposition of MT_{flex} generated binding pocket (Red) with the crystal binding pocket (coloured) along with the native ligand (orange) for PDBIDs 2R23 (left) and 1U33(right). For PDBID 2R23, Arg30 and Asn52 of chain D interact with the native ligand restoring the position of docked ligand in the MT_{flex} binding pocket. For PDBID 1U33, the side chain of GLU233 in the MT_{flex} binding pocket clashes with an atom of native ligand.

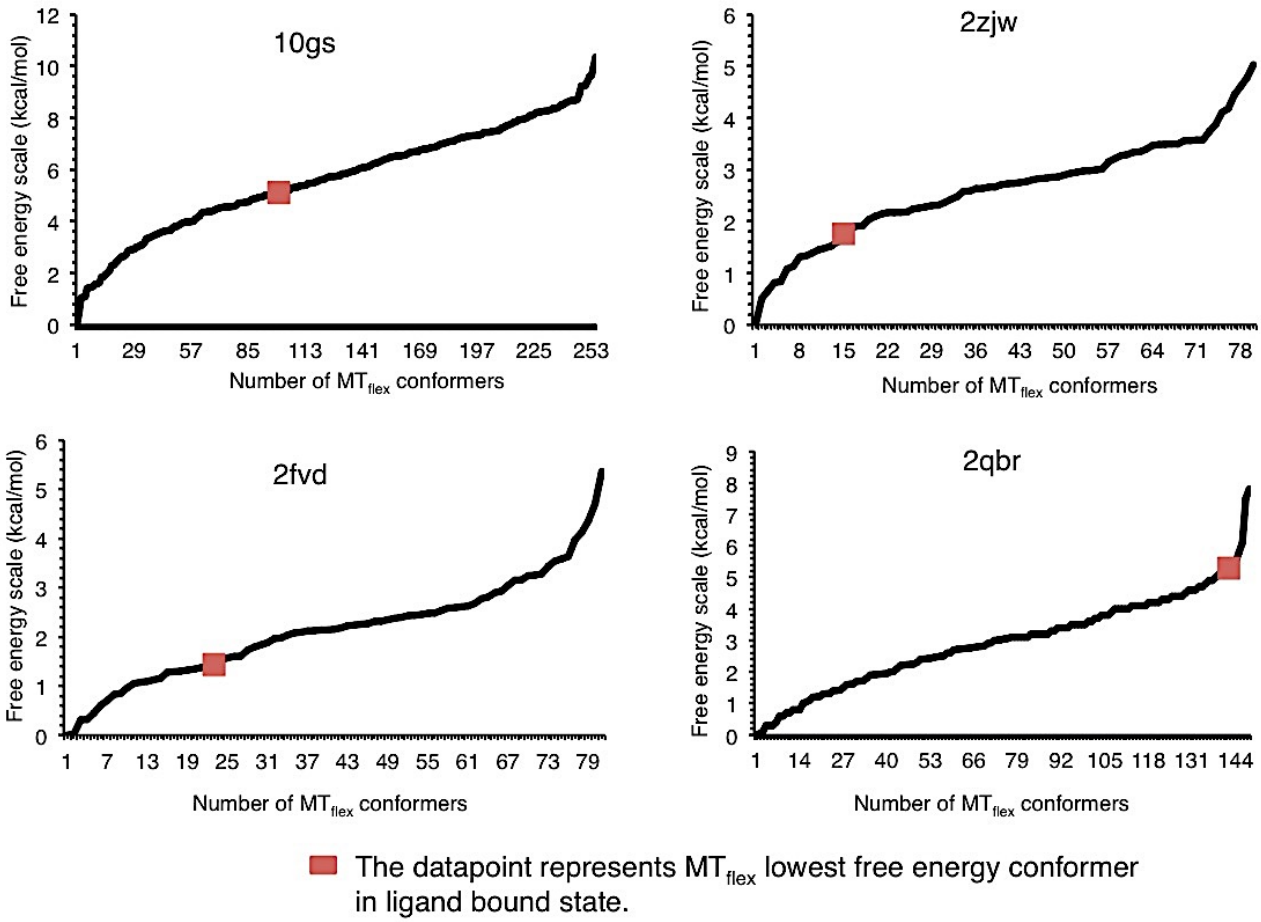

Figure 3.11 A plot showing the position of MT_{flex} lowest free energy conformation in ligand bound state (marked in red square) on the free energy scale generated by MT_{flex} conformers in ligand-unbound state for four systems from our validation dataset. PDBIDs of the systems are: 10gs, 2fvd, 1qbr and 2zjw.

Table 3.5 Correlation of Glide's best score (top docking solution) for crystal protein, Glide's best score for MT_{flex} conformers across whole conformations and Glide's score for the lowest free energy MT_{flex} conformation in the ligand-unbound state to the experimental binding affinities.

	Spearman's Rho	Pearson's R	RMSE (kcal/mol)	MUE (kcal/mol)
Glide's best score- Crystal	0.43	0.39	3.21	2.34
Glide's best score-MT_{flex}	0.50	0.47	2.71	2.17
Glide's score-MT_{flex} lowest free energy	0.38	0.37	3.17	2.43

3.6.1 RMSD in-house code

For calculating RMSDs of the MT_{flex} conformers from the crystal structure, we wanted to do best fitting of the structures by aligning them and minimizing the rotations. We were especially concerned about the two aromatic amino acids - Phenylalanine and Tyrosine, which have benzene and 4-hydroxy benzene as the side chain and have an axis of symmetry.

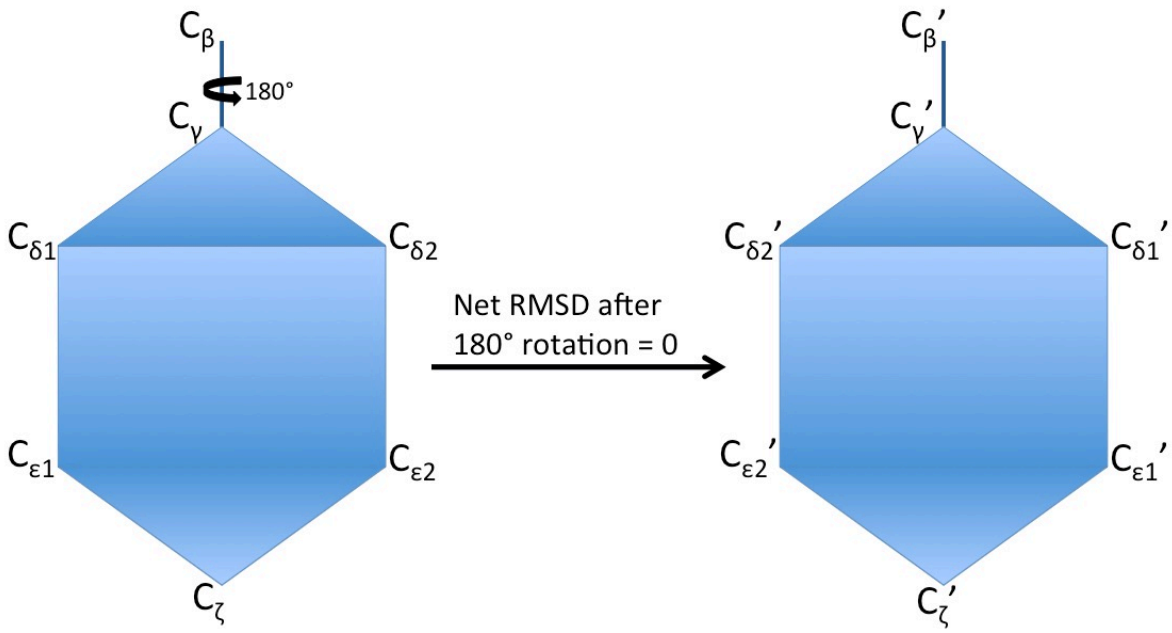

Figure 3.12 Side chain rotation of Phenylalanine residue.

If these residues undergo a 180° rotation around C_β atom, the net RMSD of the side chain would be zero. Please see the attached figure 3.12 for Phenylalanine. Similar situation would happen for Tyrosine. Our in-house code incorporated a feature, which would swap $C_{\delta1}$' with $C_{\delta2}$' and $C_{\varepsilon1}$' with $C_{\varepsilon2}$', if the Euclidean distance between $C_{\delta1}$ and $C_{\delta1}$' is greater than $C_{\delta1}$ with $C_{\delta2}$'. Another feature of our code is that it includes only those heavy atoms of side chains in doing averaging, which deviates from their position in the crystal structure. The code was written as a simple script in MATLAB software.

Algorithm:

If

$$||C_{\delta 1} - C_{\delta 1}'|| > ||C_{\delta 1} - C_{\delta 2}'||$$

then,

$$C_{\delta 1,(new)}' = C_{\delta 2}' \text{ and } C_{\delta 2,(new)}' = C_{\delta 1}'$$
$$C_{\varepsilon 1\ (new)} = C_{\varepsilon 2}' \text{ and } C_{\varepsilon 2\ (new)} = C_{\varepsilon 1}'$$

elseif

$$||C_{\delta 1} - C_{\delta 1}'|| <= ||C_{\delta 1} - C_{\delta 2}'||$$

then,

$$C_{\delta 1,(new)}' = C_{\delta 1}' \text{ and } C_{\delta 2,(new)}' = C_{\delta 2}'$$
$$C_{\varepsilon 1\ (new)} = C_{\varepsilon 1}' \text{ and } C_{\varepsilon 2\ (new)} = C_{\varepsilon 2}'$$

REFERENCES

REFERENCES

1. Koshland, D.E., *The Key-Lock Theory and the Induced Fit Theory.* Angewandte Chemie-International Edition, 1994. **33**(23-24): p. 2375-2378.

2. Koshland, D.E., *Correlation of Structure and Function in Enzyme Action.* Science, 1963. **142**(359): p. 1533-&.

3. Ma, B.Y., et al., *Folding funnels and binding mechanisms.* Protein Engineering, 1999. **12**(9): p. 713-720.

4. Fitzgerald, P.M.D., et al., *Crystallographic Analysis of a Complex between Human-Immunodeficiency-Virus Type-1 Protease and Acetyl-Pepstatin at 2.0-a Resolution.* Journal of Biological Chemistry, 1990. **265**(24): p. 14209-14219.

5. Gutteridge, A. and J. Thornton, *Conformational changes observed in enzyme crystal structures upon substrate binding.* Journal of Molecular Biology, 2005. **346**(1): p. 21-28.
6. Jorgensen, W.L., *Rusting of the Lock and Key Model for Protein-Ligand Binding.* Science, 1991. **254**(5034): p. 954-955.

7. Vanduyne, G.D., et al., *Atomic-Structure of Fkbp-Fk506, an Immunophilin-Immunosuppressant Complex.* Science, 1991. **252**(5007): p. 839-842.

8. Weber, P.C., et al., *Structural Origins of High-Affinity Biotin Binding to Streptavidin.* Science, 1989. **243**(4887): p. 85-88.

9. Boehr, D.D., R. Nussinov, and P.E. Wright, *The role of dynamic conformational ensembles in biomolecular recognition (vol 5, pg 789, 2009).* Nature Chemical Biology, 2009. **5**(12): p. 954-954.

10. Teague, S.J., *Implications of protein flexibility for drug discovery.* Nat Rev Drug Discov, 2003. **2**(7): p. 527-41.

11. Feixas, F., et al., *Exploring the role of receptor flexibility in structure-based drug discovery.* Biophys Chem, 2014. **186**: p. 31-45.

12. Gerstein, M., A.M. Lesk, and C. Chothia, *Structural Mechanisms for Domain Movements in Proteins.* Biochemistry, 1994. **33**(22): p. 6739-6749.

13. Gunasekaran, K. and R. Nussinov, *How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding.* Journal of Molecular Biology, 2007. **365**(1): p. 257-273.

14. Lesk, A.M. and C. Chothia, *Mechanisms of Domain Closure in Proteins.* Journal of Molecular Biology, 1984. **174**(1): p. 175-191.

15. Najmanovich, R., et al., *Side-chain flexibility in proteins upon ligand binding.* Proteins-Structure Function and Genetics, 2000. **39**(3): p. 261-268.

16. Frauenfelder, H., S.G. Sligar, and P.G. Wolynes, *The Energy Landscapes and Motions of Proteins.* Science, 1991. **254**(5038): p. 1598-1603.

17. Sorensen, J., et al., *Molecular Docking to Flexible Targets.* Molecular Modeling of Proteins: 2nd Edition, 2015. **1215**: p. 445-469.

18. Du, X., et al., *Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods.* International Journal of Molecular Sciences, 2016. **17**(2).

19. Abagyan, R. and M. Totrov, *High-throughput docking for lead generation.* Current Opinion in Chemical Biology, 2001. **5**(4): p. 375-382.

20. Carlson, H.A., *Protein flexibility and drug design: how to hit a moving target.* Current Opinion in Chemical Biology, 2002. **6**(4): p. 447-452.

21. Antunes, D.A., D. Devaurs, and L.E. Kavraki, *Understanding the challenges of protein flexibility in drug design.* Expert Opinion on Drug Discovery, 2015. **10**(12): p. 1301-1313.

22. B-Rao, C., J. Subramanian, and S.D. Sharma, *Managing protein flexibility in docking and its applications.* Drug Discovery Today, 2009. **14**(7-8): p. 394-400.

23. Henzler-Wildman, K. and D. Kern, *Dynamic personalities of proteins.* Nature, 2007. **450**(7172): p. 964-972.

24. Kokh, D.B., R.C. Wade, and W. Wenzel, *Receptor flexibility in small-molecule docking calculations.* Wiley Interdisciplinary Reviews-Computational Molecular Science, 2011. **1**(2): p. 298-314.

25. Spyrakis, F., et al., *Protein Flexibility and Ligand Recognition: Challenges for Molecular Modeling.* Current Topics in Medicinal Chemistry, 2011. **11**(2): p. 192-210.

26. Sousa, S.F., et al., *Protein-Ligand Docking in the New Millennium - A Retrospective of 10 Years in the Field.* Current Medicinal Chemistry, 2013. **20**(18): p. 2296-2314.

27. Ferrari, A.M., et al., *Soft docking and multiple receptor conformations in virtual screening.* Journal of Medicinal Chemistry, 2004. **47**(21): p. 5076-5084.

28. Jiang, F. and S.H. Kim, *Soft Docking - Matching of Molecular-Surface Cubes.* Journal of Molecular Biology, 1991. **219**(1): p. 79-102.

29. Jain, T., D.S. Cerutti, and J.A. McCammon, *Configurational-bias sampling technique for predicting side-chain conformations in proteins.* Protein Science, 2006. **15**(9): p. 2029-2039.

30. Leach, A.R., *Ligand Docking to Proteins with Discrete Side-Chain Flexibility.* Journal of Molecular Biology, 1994. **235**(1): p. 345-356.

31. Yang, A.Y.C., P. Kallblad, and R.L. Mancera, *Molecular modelling prediction of ligand binding site flexibility.* Journal of Computer-Aided Molecular Design, 2004. **18**(4): p. 235-250.

32. Alberts, I.L., N.P. Todorov, and P.M. Dean, *Receptor flexibility in de novo ligand design and docking.* Journal of Medicinal Chemistry, 2005. **48**(21): p. 6585-6596.

33. Sinko, W., S. Lindert, and J.A. McCammon, *Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design.* Chemical Biology & Drug Design, 2013. **81**(1): p. 41-49.

34. Frembgen-Kesner, T. and A.H. Elcock, *Computational sampling of a cryptic drug binding site in a protein receptor: Explicit solvent molecular dynamics and inhibitor docking to p38 MAP kinase.* Journal of Molecular Biology, 2006. **359**(1): p. 202-214.

35. Kua, J., Y.K. Zhang, and J.A. McCammon, *Studying enzyme binding specificity in acetylcholinesterase using a combined molecular dynamics and multiple docking approach.* Journal of the American Chemical Society, 2002. **124**(28): p. 8260-8267.

36. Armen, R.S., J.H. Chen, and C.L. Brooks, *An Evaluation of Explicit Receptor Flexibility in Molecular Docking Using Molecular Dynamics and Torsion Angle Molecular Dynamics.* Journal of Chemical Theory and Computation, 2009. **5**(10): p. 2909-2923.

37. Orry, A.J.W., R.A. Abagyan, and C.N. Cavasotto, *Structure-based development of target-specific compound libraries.* Drug Discovery Today, 2006. **11**(5-6): p. 261-266.

38. Atilgan, C., et al., *Manipulation of Conformational Change in Proteins by Single-Residue Perturbations.* Biophysical Journal, 2010. **99**(3): p. 933-943.

39. Cavasotto, C.N., J.A. Kovacs, and R.A. Abagyan, *Representing receptor flexibility in ligand docking through relevant normal modes.* Journal of the American Chemical Society, 2005. **127**(26): p. 9632-9640.

40. Gerek, Z.N. and S.B. Ozkan, *Change in Allosteric Network Affects Binding Affinities of PDZ Domains: Analysis through Perturbation Response Scanning.* Plos Computational Biology, 2011. **7**(10).

41. Gerek, Z.N. and S.B. Ozkan, *A flexible docking scheme to explore the binding selectivity of PDZ domains.* Protein Science, 2010. **19**(5): p. 914-928.

42. Carlson, H.A. and J.A. McCammon, *Accommodating protein flexibility in computational drug design.* Molecular Pharmacology, 2000. **57**(2): p. 213-218.

43. Damm, K.L. and H.A. Carlson, *Exploring experimental sources of multiple protein conformations in structure-based drug design.* Journal of the American Chemical Society, 2007. **129**(26): p. 8225-8235.

44. Huang, S.Y. and X.Q. Zou, *Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking.* Proteins-Structure Function and Bioinformatics, 2007. **66**(2): p. 399-421.

45. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility.* Journal of Computational Chemistry, 2009. **30**(16): p. 2785-2791.

46. Osterberg, F., et al., *Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock.* Proteins-Structure Function and Genetics, 2002. **46**(1): p. 34-40.

47. Huang, S.Y. and X.Q. Zou, *An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function.* Journal of Computational Chemistry, 2006. **27**(15): p. 1876-1882.

48. Huang, S.Y. and X.Q. Zou, *An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials.* Journal of Computational Chemistry, 2006. **27**(15): p. 1866-1875.

49. Cavasotto, C.N. and R.A. Abagyan, *Protein flexibility in ligand docking and virtual screening to protein kinases.* Journal of Molecular Biology, 2004. **337**(1): p. 209-225.

50. Ding, F. and N.V. Dokholyan, *Incorporating Backbone Flexibility in MedusaDock Improves Ligand-Binding Pose Prediction in the CSAR2011 Docking Benchmark.* Journal of Chemical Information and Modeling, 2013. **53**(8): p. 1871-1879.

51. Lauck, F., et al., *RosettaBackrub-a web server for flexible backbone protein structure modeling and design.* Nucleic Acids Research, 2010. **38**: p. W569-W575.

52. Claussen, H., et al., *FlexE: Efficient molecular docking considering protein structure variations.* Journal of Molecular Biology, 2001. **308**(2): p. 377-395.

53. Bolia, A., Z.N. Gerek, and S.B. Ozkan, *BP-Dock: A Flexible Docking Scheme for Exploring Protein Ligand Interactions Based on Unbound Structures.* Journal of Chemical Information and Modeling, 2014. **54**(3): p. 913-925.

54. Trott, O. and A.J. Olson, *Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading.* Journal of Computational Chemistry, 2010. **31**(2): p. 455-461.

55. Morris, G.M., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* Journal of Computational Chemistry, 1998. **19**(14): p. 1639-1662.

56. Jones, G., et al., *Development and validation of a genetic algorithm for flexible ligand docking.* Abstracts of Papers of the American Chemical Society, 1997. **214**: p. 154-COMP.

57. Kallblad, P. and P.M. Dean, *Efficient conformational sampling of local side-chain flexibility.* Journal of Molecular Biology, 2003. **326**(5): p. 1651-1665.

58. Sherman, W., et al., *Novel procedure for modeling ligand/receptor induced fit effects.* Journal of Medicinal Chemistry, 2006. **49**(2): p. 534-553.

59. Davis, I.W. and D. Baker, *ROSETTALIGAND Docking with Full Ligand and Receptor Flexibility.* Journal of Molecular Biology, 2009. **385**(2): p. 381-392.

60. Meiler, J. and D. Baker, *ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility.* Proteins-Structure Function and Bioinformatics, 2006. **65**(3): p. 538-548.

61. Zhao, Y. and M.F. Sanner, *FLIPDock: Docking flexible ligands into flexible receptors.* Abstracts of Papers of the American Chemical Society, 2007. **233**: p. 152-152.

62. Zhao, Y. and M.F. Sanner, *Protein-ligand docking with multiple flexible side chains.* Journal of Computer-Aided Molecular Design, 2008. **22**(9): p. 673-679.

63. Allen, W.J., et al., *DOCK 6: Impact of New Features and Current Docking Performance.* Journal of Computational Chemistry, 2015. **36**(15): p. 1132-1156.

64. Schnecke, V., et al., *Screening a peptidyl database for potential ligands to proteins with side-chain flexibility.* Proteins-Structure Function and Genetics, 1998. **33**(1): p. 74-87.

65. Zavodszky, M.I. and L.A. Kuhn, *Side-chain flexibility in protein-ligand binding: The minimal rotation hypothesis.* Protein Science, 2005. **14**(4): p. 1104-1114.

66. Zavodszky, M.I., et al., *Scoring ligand similarity in structure-based virtual screening.* Journal of Molecular Recognition, 2009. **22**(4): p. 280-292.

67. Leis, S. and M. Zacharias, *ReFlexIn: A Flexible Receptor Protein-Ligand Docking Scheme Evaluated on HIV-1 Protease.* Plos One, 2012. **7**(10).

68. Shin, W.H. and C. Seok, *GalaxyDock: Protein-Ligand Docking with Flexible Protein Side-chains.* Journal of Chemical Information and Modeling, 2012. **52**(12): p. 3225-3232.

69. Corbeil, C.R., P. Englebienne, and N. Moitessier, *Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0.* Journal of Chemical Information and Modeling, 2007. **47**(2): p. 435-449.

70. Corbeil, C.R., et al., *Docking Ligands into flexible and solvated macromolecules. 2. Development and application of FITTED 1.5 to the virtual screening of potential HCV polymerase inhibitors.* Journal of Chemical Information and Modeling, 2008. **48**(4): p. 902-909.

71. Corbeil, C.R. and N. Moitessier, *Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs.* Journal of Chemical Information and Modeling, 2009. **49**(4): p. 997-1009.

72. Zacharias, M., *Rapid protein-ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: Binding of FK506 to FKBP.* Proteins-Structure Function and Bioinformatics, 2004. **54**(4): p. 759-767.

73. Zacharias, M., *Combining elastic network analysis and molecular dynamics simulations by Hamiltonian replica exchange.* Journal of Chemical Theory and Computation, 2008. **4**(3): p. 477-487.

74. Nabuurs, S.B., M. Wagener, and J. De Vlieg, *A flexible approach to induced fit docking.* Journal of Medicinal Chemistry, 2007. **50**(26): p. 6507-6518.

75. Mashiach, E., R. Nussinov, and H.J. Wolfson, *FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking.* Nucleic Acids Research, 2010. **38**: p. W457-W461.

76. Mashiach, E., et al., *An integrated suite of fast docking algorithms.* Proteins-Structure Function and Bioinformatics, 2010. **78**(15): p. 3197-3204.

77. Sandak, B., H.J. Wolfson, and R. Nussinov, *Flexible docking allowing induced fit in proteins: Insights from an open to closed conformational isomers.* Proteins-Structure Function and Genetics, 1998. **32**(2): p. 159-174.

78. Schneidman-Duhovny, D., et al., *Geometry-based flexible and symmetric protein docking.* Proteins-Structure Function and Bioinformatics, 2005. **60**(2): p. 224-231.

79. Schneidman-Duhovny, D., R. Nussinov, and H.J. Wolfson, *Automatic prediction of protein interactions with large scale motion.* Proteins-Structure Function and Bioinformatics, 2007. **69**(4): p. 764-773.

80. Bolia, A. and S.B. Ozkan, *Adaptive BP-Dock: An Induced Fit Docking Approach for Full Receptor Flexibility.* Journal of Chemical Information and Modeling, 2016. **56**(4): p. 734-746.

81. Almlof, M., B.O. Brandsdal, and J. Aqvist, *Binding affinity prediction with different force fields: examination of the linear interaction energy method.* J Comput Chem, 2004. **25**(10): p. 1242-54.

82. Aqvist, J., C. Medina, and J.E. Samuelsson, *A new method for predicting binding affinity in computer-aided drug design.* Protein Eng, 1994. **7**(3): p. 385-91.

83. Amaro, R.E., R. Baron, and J.A. McCammon, *An improved relaxed complex scheme for receptor flexibility in computer-aided drug design.* Journal of Computer-Aided Molecular Design, 2008. **22**(9): p. 693-705.

84. Lin, J.H., et al., *Computational drug design accommodating receptor flexibility: The relaxed complex scheme.* Journal of the American Chemical Society, 2002. **124**(20): p. 5632-5633.

85. Carlson, H.A., et al., *Developing a dynamic pharmacophore model for HIV-1 integrase.* Journal of Medicinal Chemistry, 2000. **43**(11): p. 2100-2114.

86. Meagher, K.L. and H.A. Carlson, *Incorporating protein flexibility in structure-based drug discovery: Using HIV-1 protease as a test case.* Journal of the American Chemical Society, 2004. **126**(41): p. 13276-13281.

87. Liu, H.Y., A.E. Mark, and W.F. vanGunsteren, *Estimating the relative free energy of different molecular states with respect to a single reference state.* Journal of Physical Chemistry, 1996. **100**(22): p. 9485-9494.

88. Oostenbrink, C. and W.F. Van Gunsteren, *Single-step perturbations to calculate free energy differences from unphysical reference states: Limits on size, flexibility, and character.* Journal of Computational Chemistry, 2003. **24**(14): p. 1730-1739.

89. Zagrovic, B. and W.F. van Gunsteren, *Computational analysis of the mechanism and thermodynamics of inhibition of phosphodiesterase 5A by synthetic ligands.* Journal of Chemical Theory and Computation, 2007. **3**(1): p. 301-311.

90. Grinter, S.Z. and X.Q. Zou, *Challenges, Applications, and Recent Advances of Protein-Ligand Docking in Structure-Based Drug Design.* Molecules, 2014. **19**(7): p. 10150-10176.

91. Lexa, K.W. and H.A. Carlson, *Protein flexibility in docking and surface mapping.* Quarterly Reviews of Biophysics, 2012. **45**(3): p. 301-343.

92. Mobley, D.L. and K.A. Dill, *Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get".* Structure, 2009. **17**(4): p. 489-498.

93. Sliwoski, G., et al., *Computational Methods in Drug Discovery.* Pharmacological Reviews, 2014. **66**(1): p. 334-395.

94. Durrant, J.D. and J.A. McCammon, *Computer-aided drug-discovery techniques that account for receptor flexibility.* Current Opinion in Pharmacology, 2010. **10**(6): p. 770-774.

95. Zheng, Z., M.N. Ucisik, and K.M. Merz, Jr., *The Movable Type Method Applied to Protein-Ligand Binding.* J Chem Theory Comput, 2013. **9**(12): p. 5526-5538.

96. Wang, R.X., et al., *The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures.* Journal of Medicinal Chemistry, 2004. **47**(12): p. 2977-2980.

97. Wang, R.X., et al., *The PDBbind database: Methodologies and updates.* Journal of Medicinal Chemistry, 2005. **48**(12): p. 4111-4119.

98. *MATLAB R2015a.* MATLAB R2015a, The Mathworks Inc., Natick, Massachusetts, United States, 2015.

99. Roe, D.R. and T.E. Cheatham, *PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data.* Journal of Chemical Theory and Computation, 2013. **9**(7): p. 3084-3095.

100. Friesner, R.A., et al., *Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.* Journal of Medicinal Chemistry, 2004. **47**(7): p. 1739-1749.

101. Friesner, R.A., et al., *Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes.* Journal of Medicinal Chemistry, 2006. **49**(21): p. 6177-6196.

102. Halgren, T.A., et al., *Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening.* Journal of Medicinal Chemistry, 2004. **47**(7): p. 1750-1759.

103. Sastry, G.M., et al., *Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments.* Journal of Computer-Aided Molecular Design, 2013. **27**(3): p. 221-234.

104. Olsson, M.H.M., et al., *PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK(a) Predictions.* Journal of Chemical Theory and Computation, 2011. **7**(2): p. 525-537.

105. Rostkowski, M., et al., *Graphical analysis of pH-dependent properties of proteins predicted using PROPKA.* Bmc Structural Biology, 2011. **11**.

106. Jorgensen, W.L. and J. Tiradorives, *The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin.* Journal of the American Chemical Society, 1988. **110**(6): p. 1657-1666.

107. Kaminski, G.A., et al., *Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides.* Journal of Physical Chemistry B, 2001. **105**(28): p. 6474-6487.

108. Totrov, M. and R. Abagyan, *Flexible protein-ligand docking by global energy optimization in internal coordinates.* Proteins-Structure Function and Genetics, 1997: p. 215-220.

109. Houston, D.R. and M.D. Walkinshaw, *Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context.* Journal of Chemical Information and Modeling, 2013. **53**(2): p. 384-390.

CHAPTER 4

The Role of the Active Site Flap in Streptavidin/Biotin Complex Formation

† Reprinted (adapted) with permission from Bansal, N.; Zheng, Z.; Song, F. L.; Pei, J.; Merz, K. M., The Role of the Active Site Flap in Streptavidin/Biotin Complex Formation. *J Am Chem Soc* **2018,** DOI 10.1021/jacs.8b00743

4.1 Abstract

Obtaining a detailed description of how active site flap motion affects substrate or ligand binding will advance structure-based drug design (SBDD) efforts on systems including the kinases, HSP90, HIV protease, ureases, *etc*. Through this understanding we will be able to design better inhibitors and better proteins that have desired functions. Herein we address this issue by generating the relevant configurational states of a protein flap on the molecular energy landscape using an approach we call $MT_{Flex\text{-}b}$ and then following this with a procedure to estimate the free energy associated with the motion of the flap region. To illustrate our overall workflow, we explored the free energy changes in the Streptavidin/Biotin system upon introducing conformational flexibility in $loop_{3\text{-}4}$ in the Biotin un-bound (*apo*) and bound (*holo*) state. The free energy surfaces were created using the Movable Type free energy method and for further validation we compared them to potential of mean force (PMF) generated free energy surfaces using MD simulations employing the FF99SBILDN and FF14SB force fields. We also estimated the free energy thermodynamic cycle using an ensemble of closed-like and open-like end states for the ligand unbound and bound states and estimated the binding free energy to be ~ –16.2 kcal/mol (experimental -18.3 kcal/mol). The good agreement between $MT_{Flex\text{-}b}$ in combination with the MT method with experiment and MD simulations supports the effectiveness of our strategy in obtaining unique insights into the motions in proteins that can then be used in a range of biological and biomedical applications.

4.2 Introduction

The rapid and chemically accurate estimation of the free energy change involved when a ligand binds to a biochemical system is the Holy Grail of structure based drug design[1-3]. Several free energy methods ranging from end-point approach to pathway based free energy calculation have been explored computationally. The so-called end-point methods used to estimate free energies (*e.g.,* docking, MMPBSA and MMGBSA) are computationally inexpensive but they typically rely on a single static structure and often ignore receptor flexibility[4, 5]. Pathway free energy methods require extensive sampling to estimate the free energies. The pathway free energy methods can be broadly categorized into alchemical and potential of mean force approaches. Alchemical free energy methods use an "alchemical" pathway of non-physical intermediate states to build a thermodynamic cycle and compute the free energy differences between the end states[6-8]. Free energy perturbation[9] (FEP) and thermodynamic integration[10] (TI) are two of the most commonly used alchemical free energy methods. Another pathway method to estimate free energies is the Potential of mean force (PMF) approach[11-13]. Umbrella sampling coupled with the WHAM (weighted histogram analysis method) analysis is one of the most commonly used PMF approaches[14]. The major bottleneck of the PMF approach is the computational cost associated with the conformational sampling. Moreover, the pathway free energy methods, in contrast to end-points methods, are computationally intensive and are limited by simulation timescales. Finally, all methods are affected by the accuracy of the underlying force field and improvements in these remain an active area of research[6, 15-17].

Conformational sampling of the relevant receptor "states" and ligand poses is essential in order to advance the field of structure based drug design (SBDD). Conformational flexibility is pivotal to several important molecular recognition processes including protein-protein and protein-ligand binding events, enzyme catalysis, allosteric control, and biomolecular assembly[18-20]. In many protein structures the conformational flexibility is often attributed to the loop regions, which serve as the "connecting segments" between two defined secondary structures or different domains[20-23]. But, loops are not mere connectors. They participate in crucial functions, including ligand binding events[24-26], enzyme catalysis[27] and molecular recognition processes[28-30]. Often, crystal structures of proteins are reported with loops missing because their conformational flexibility makes it difficult to resolve them in X-ray crystallographic experiments[31]. In many instances, these missing loops mediate the biological function of a protein; hence, computational chemists have developed a range of loop modeling algorithms to address this issue[32-92]. Sampling methods based on MC and MD simulations have been used extensively[79, 93-99]. Several, inverse kinematic methods with Monte Carlo sampling scheme have been developed to simulate loop flexibility and domain motion. These methods explore the entire torsional space (ϕ, ψ, ω) by designing Monte Carlo moves and select or reject a move based on a metropolis algorithm. The conformational energy in these methods is usually explored using force fields based methods[100-104]. Rosetta also uses Metropolis Monte Carlo sampling of ϕ and ψ for local refinement of protein structures. It models large-scale backbone conformational changes by exchanging the backbone conformations of peptide fragments collected from homologous sequences in the PDB and a backbone dependent rotamer library for inclusion of side chain flexibility. The Rosetta suite uses an all atom energy function, which calculates the weighted sum of energy terms in Rosetta Energy units. The scoring function

is a combination of physics based and statistical terms[105]. Conceptually, our approach is similar to using rotamer library based methods but our atoms are added not in terms of structural database rather on a local energy landscape defined by the potential. Moreover, unlike other methods, we report free energies rather than energies. Detailed discussions of loop modeling methods and algorithms can be found in the extant literature[106-109].

A combination of both flexibility and rigidity are crucial for the protein function[110]. Flexibility of an amino acid is usually defined by its atomic temperature factors of B-factor[111]. Highly flexible regions are generally characterized by more hydrophilic residues and are enriched with higher charged residues while the less flexible regions tend to be high in hydrophobic residues and have a reduced number of charged residues[112]. Recently, a H192P point mutation on one of the cofactor binding loops in E.coli *Thermus thermophilus* improved the stability of the protein by two-fold at elevated temperature compared to wild type at 60º C[113]. Surface loops are known to undergo functionally relevant conformational changes. Experiments have shown that often these conformational changes are related to external perturbations including ligand binding[114]. These conformations pre-exist but the relative population shifts upon the binding event[115-117]. This leads to discrete open or closed states or multiple different states in the protein separated by an energy barrier. They may have a relatively low free energy barrier, which favors fast interconversion between different states or they may have a relatively high barrier, thereby, making the interconversion relatively slow[118]. If the barrier height is too high, it becomes more computationally challenging for sampling methods to capture the entire conformational space spanned by the loop requiring the use of biasing methods.

The accuracy of free energy calculations in part depends on the extent and accuracy of the conformational sampling of the relevant chemical space. With the continued advancement in computational hardware and software, molecular dynamics (MD) simulations have reached the millisecond milestone, but a sufficient number of uncorrelated configurations is still not guaranteed[3]. Monte Carlo sampling, on the other hand, does not suffer from the problem of uncorrelated configurations, however, convergence is still an issue. Moreover, the computational cost associated with "sufficient" conformational sampling with both of these methods is a major bottleneck for proteins due to their size and the number of degrees of freedom[3]. Apart from the computational challenges faced when sampling relevant conformational space, the accuracy of current force fields can be a concern as well[99].

In this spirit, we introduce a new procedure to estimate the free energies in biochemical systems via extrapolation of the local partition function for a set of pre-generated conformational ensembles. The conformational ensemble of the targeted flexible region of the protein is generated via a sampling method we call $MT_{Flex\text{-}b}$, which treats the molecule at an atom pair level and utilizes atom pairwise one-dimensional knowledge based potentials to enumerate the molecular conformations on an energy surface. The free energies are then estimated by extrapolating the local partition function of the generated "seed structures" from the conformational ensemble using the MT method[119].

We have applied our method to study the free energy changes observed in Streptavidin by accommodating conformational flexibility in the $loop_{3\text{-}4}$ region (in the streptavidin monomer) in the Biotin bound (*holo*) and un-bound (*apo*) states. This loop is highly mobile and undergoes an open to closed transition upon biotin binding[120-123]. Currently, we have incorporated flexibility into the loop region only, but the workflow is straightforward and can be applied to

larger regions of a given protein as needed. Using $MT_{Flex\text{-}b}$, we obtained 21,295 unique backbone conformations of the eight-residue long $loop_{3\text{-}4}$ where we observe examples of closed-like, open-like and several intermediate loop conformations. The lowest Cα RMSD conformations generated by $MT_{Flex\text{-}b}$ (with respect to the closed and open crystal structures) were similar to their corresponding crystal conformations with Cα RMSDs of ~1.6 Å. Using an ensemble of closed-like and open-like $MT_{Flex\text{-}b}$ generated conformations as end states in the ligand unbound and bound states, we obtained a free energy thermodynamic cycle that gives us detailed insights into biotin binding and the role flap motion plays in binding. We further explored the energy landscape by generating a free energy surface (FES) using the MT method and further validated it by generating a PMF via umbrella sampling followed by analysis by WHAM using force field based MD simulation methodologies.

The conformational search was run from both loop ends in two separate jobs each using a single CPU on the Laconia cluster available at the High Performance Computer Center (HPCC) facility at MSU. Conformation generation using this code base took ~7days to generate over 11 million different conformations of the eight residue long loop of Streptavidin. The free energy calculations using the MT method took ~20 minutes per seed using a MATLAB code on a single CPU**.** Since the conformations were pre-generated, the free energy calculations were run in a trivially parallel fashion using the HPCC facility at MSU.

4.3 Results and Discussion: Streptavidin-biotin

The streptavidin-biotin complex is biotechnologically an interesting protein-ligand system with one of the most remarkable binding affinities known[122-125]. Streptavidin is a homo tetramer with an eight-residue loop between β–strand 3 and strand 4 (also known as $loop_{3\text{-}}$

$_4$) in all four monomers. This loop$_{3\text{-}4}$ region is flexible and is known to undergo a transition from an open to closed state upon biotin binding[121, 124, 126]. The closed and open states of the streptavidin monomer with biotin bound in the binding pocket are shown in Figure 4.1. The backbone RMSD (Å) between the closed and open states of loop$_{3\text{-}4}$ is ~ 8.6 Å. The motion of the loop can be characterized by the distance between the C-alpha carbon of residue 49 (roughly the mid-point of loop$_{3\text{-}4}$) in the closed and open state, which is 12.7Å. This loop is highly mobile and its interaction with biotin has been extensively explored in several experimental and computational studies[123, 124, 126-129]. Recently, the conformational dynamics of loop$_{3\text{-}4}$ was studied by Song *et al* using MD simulations and enhanced sampling methods[129]. Using accelerated MD, they were able to study the transition of loop$_{3\text{-}4}$ between the open and closed states in the presence of biotin in the monomer of streptavidin but conventional MD was not able to observe the loop transition at the timescales employed. Because they studied the monomer the binding pocket was incomplete due to absence of an important hydrophobic/p-stacking interaction contributed by TRP120 from an adjacent monomer. This residue is important in stabilizing the streptavidin-biotin complex and how its neglect affects the outcome of the study of Song *et al* was not addressed[124, 130, 131]. To the best of our knowledge, no other loop modeling algorithm or sampling strategy has elucidated the loop$_{3\text{-}4}$ transitions of streptavidin. This motivated us to apply our method to study the free energy changes associated with the loop motion in streptavidin in the presence and absence of biotin by generating multiple loop$_{3\text{-}4}$ conformational states.

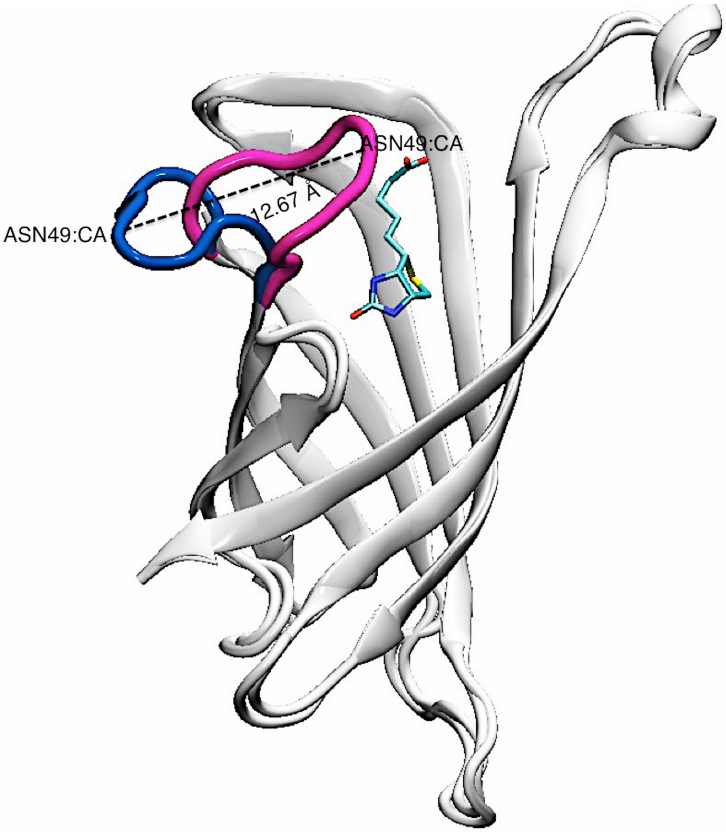

Figure 4.1 Cartoon representation of crystal Streptavidin monomer in *holo* (Biotin bound - in licorice) and *apo* states is superimposed on top of each other. The closed state of the $loop_{3\text{-}4}$ (Residues 45 to 52) pertaining to the *holo* conformation is highlighted in mauve color while the open loop is shown in blue color. The distance between the Cα of Residue 49 of the closed and open state is approx. 12.67Å.

The conformational ensemble for $loop_{3\text{-}4}$ was generated for a monomer (chain A) of the streptavidin *holo* structure (PDBID 1mk5) in the context of the tetramer. We preferred to use the *holo* conformation to the *apo* since biotin is present in the binding pocket, which helped us to calculate the protein-ligand binding affinities directly without having to resort to molecular docking and introduce additional errors into the free energy calculations pertaining to closed like conformations. Moreover, the crystal structures for the *apo* and *holo* conformations of the streptavidin protein are quite similar, except for the $loop_{3\text{-}4}$ region (the monomer structures are shown in Figure 4.1). Excluding the $loop_{3\text{-}4}$ region, the heavy atom RMSD (Å) between the

experimental *apo* and *holo* structures of streptavidin is ~1 Å with a backbone RMSD of less than 0.6 Å. The initial loop structure from the crystal structure isn't relevant when using $MT_{Flex\text{-}b}$, because of the nature of its conformation generation strategy. Only, the environment surrounding the loop residues is important since it defines van der Waals collisions and determines other important interactions (see chapter 2 for details). The rest of the protein including monomers B, C and D and rest of chain A (excluding the $loop_{3\text{-}4}$ region) were kept rigid. The biotin ligands in all the sub-units of streptavidin were also kept fixed. Biotin and the rest of the protein (the tetramer excluding the loop region of chain A) were kept rigid in order to check for van der Waals collisions. The presence of biotin added an additional layer of steric clashes, which were also identified during loop generation using the geometric restraint of 2.8 Å for the non-bonded heavy atoms and 2.3 Å for non-bonded hydrogen atoms. The total number of backbone and side-chain conformations generated using $MT_{Flex\text{-}b}$ for $loop_{3\text{-}4}$ of streptavidin, in the presence of biotin, was ~11 million conformations, for a total of 21,295 unique backbone loop conformations (see Table 4.2). The conformation generation calculations required ~7 days using our in-house MATLAB script. CPPTRAJ[132] and VMD[133] software were employed for analysis and visualization of the structures.

4.3.1 RMSD

The structural Cα RMSD (Å) of the $MT_{Flex\text{-}b}$ generated loop conformations was calculated with respect to the closed loop conformation from streptavidin (chain A from PDB ID 1mk5) using CPPTRAJ, a utility program in the AMBER suite[132]. The $MT_{Flex\text{-}b}$ loop conformation structurally closest to the closed state of the loop from streptavidin (closed-like conformation) had a Cα RMSD of 1.63 Å and the open-like conformation (structurally most similar to the open

loop in the crystal structure) had a Cα RMSD of 1.61 Å. Figure 4.2 highlights the closed-like and open-like (lowest RMSD (Å)) loop conformations superimposed with the closed and open states of $loop_{3\text{-}4}$ from the crystal structures.

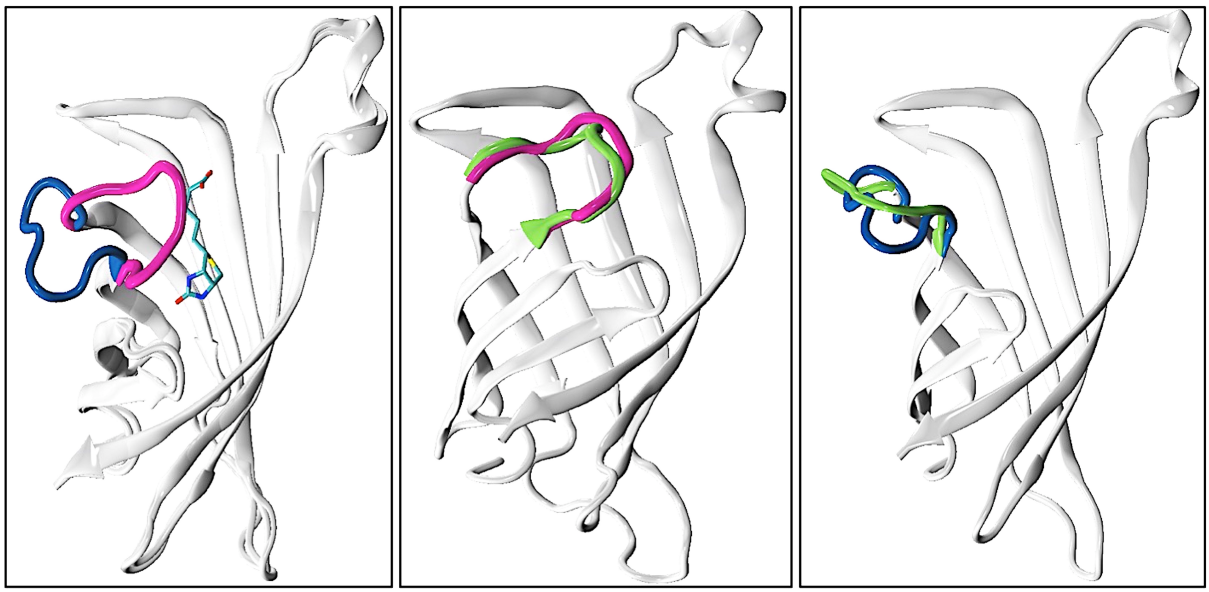

Figure 4.2 The left box displays the closed (biotin-bound) and open states of streptavidin superimposed on top of each other with the $loop_{3\text{-}4}$ highlighted in mauve for the closed and blue for the open conformation. The center image shows the lowest RMSD (Å) loop conformation (shown in green) generated by $MT_{Flex\text{-}b}$ superimposed on the *holo* crystal structure (closed loop), and the right image shows the lowest RMSD $MT_{Flex\text{-}b}$ conformation superimposed on the *apo* crystal conformation (open loop).

Using our conformation sampling strategy, we were able to map both the closed and open loop conformations from the available crystal structures.

Apart from the closed-like and open-like conformations, we obtained quite a few intermediate loop conformations generated by $MT_{Flex\text{-}b.}$ Figure 4.3 shows the Cα RMSD (Å) of all the generated $MT_{Flex\text{-}b}$ loop conformations with respect to the experimentally observed closed loop of streptavidin. The generated $MT_{Flex\text{-}b}$ loop conformers were arranged in an increasing order of structural RMSD (Å) with respect to the closed state of $loop_{3\text{-}4}$ from the streptavidin crystal structure. As can be seen from Figure 4.3, the $MT_{Flex\text{-}b}$ loop conformations traverse a

RMSD range of ~ 10 Å from the closed state covering all the major conformations including the closed, open and several intermediate states. Examples of conformational states explored by $MT_{Flex\text{-}b}$ are shown in Figure 4.3 in green. The experimental closed and open states of the loop are superimposed on the generated loop conformation in Figure 4.3 for comparison.

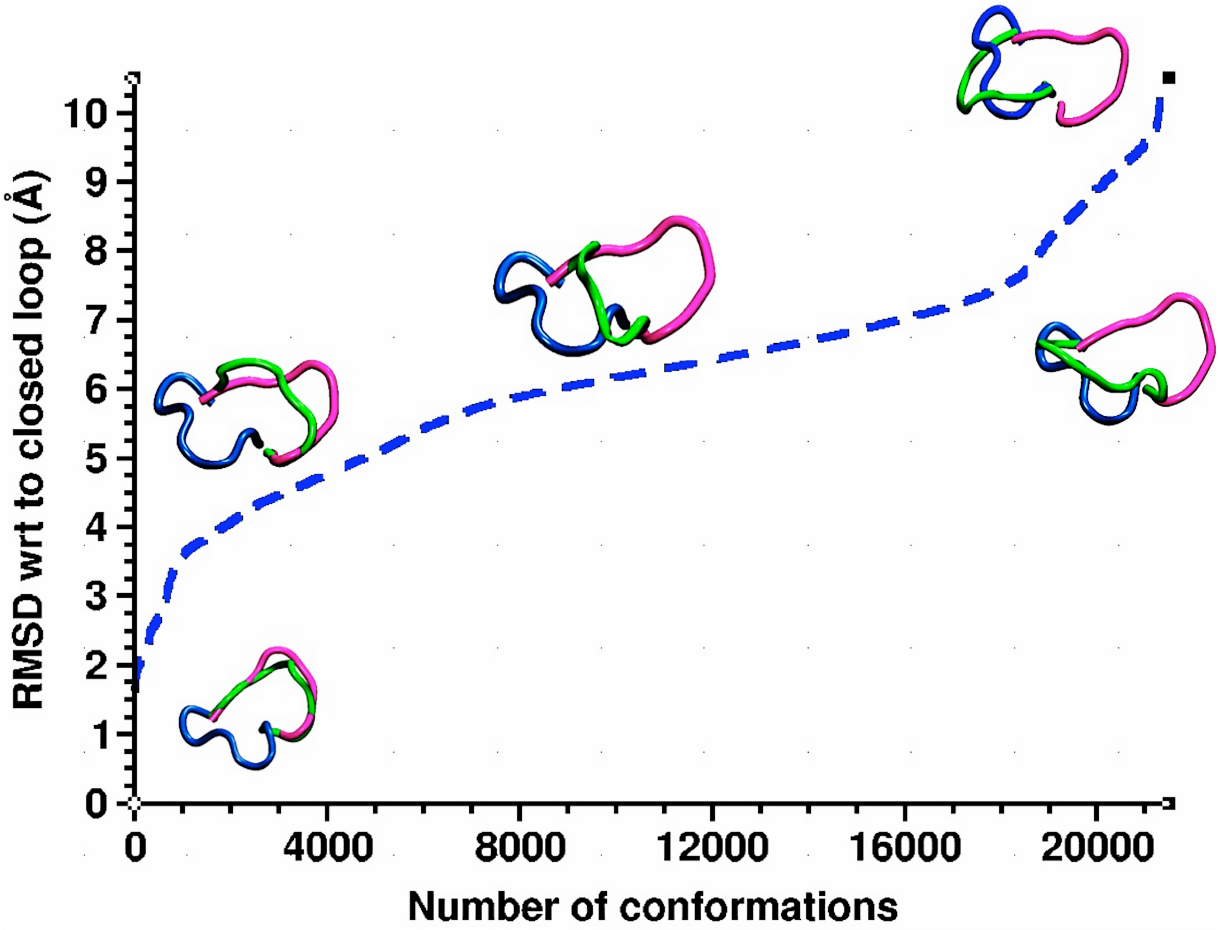

Figure 4.3 The Cα RMSD (Å) of the $MT_{Flex\text{-}b}$ loop conformations with respect to the closed state of $loop_{3\text{-}4}$ in streptavidin. The x-axis represents the total number of generated backbone loop conformations. For the sake of structural comparison, crystal-closed (pink) and open (blue) states of the loop are superimposed on the generated $MT_{Flex\text{-}b}$ loop conformation (green).

4.3.2 MT Free Energy Surface and MD Potential of Mean Force Studies

Two-dimensional relative free energy surfaces were assembled using MT in order to assess the free energy differences between the different loop states. The closed to open transition

of the loop was characterized by the separation between Cα atoms of residues Gly48 and Ile30 (X-axis), and Asn49 and Leu109 (Y-axis). The positions of the residues chosen for the reaction coordinates are shown in Figure 4.9. For the free energy surface (FES) construction, the entire reaction pathway was divided into a number of windows with a grid size of 0.5 Å. The $MT_{Flex\text{-}b}$ generated conformations were arranged in windows based on the two reaction coordinates. Our docking protocol[119] was applied to obtain the optimized ligand position for each protein conformation in all the windows. The crystal closed and open states of the loop were also included in the FES construction. The free energies were calculated in the *holo* (ligand bound) and the *apo* (free) states in the aqueous phase for each of the windows in a parallel fashion using the MT method with a sampling range of ±0.25 Å. Please note that the closed and open states of $loop_{3\text{-}4}$ in the ligand bound and free states were considered only for one of the monomers (monomer A) while the remaining three monomers were always considered in the native closed state with biotin bound in the binding pocket.

Figure 4.4 highlights the free energy profile for the *apo* and *holo* conformations obtained using this procedure. The entire FES construction has been subdivided into open (O, closed (C and intermediate (I regions. These regions are pinpointed on the 2-D heat map plot as well. For the *apo* state (shown in the left image of Figure 4.4), it can be seen that the free energy gradually increases on going from the open to the closed state of the loop with the closed state being higher in energy than the rest of the conformations. The open state of the loop is estimated to be around 10.5 kcal/mol more stable than the closed state. For the *apo* state FES, we also identified a region where the free energy was ~3 kcal/mol lower than the open state (shown on the upper right part of the *apo* FES in Figure 4.4). On visualization, we found that this region has conformations quite similar to the crystal open loop with Cα RMSDs of ~3.2 Å. The structure corresponding to

the local minima of the closed *apo* state had a ~2.5 Å RMSD from the closed loop seen in the crystal structure. The free energy difference between the local minimums in the *apo* state predicted via MT-FES is 9.1 kcal/mol. The loop conformations are shown on the *apo* FES in Figure 4.4. For the *holo* FES, the closed state of the loop is clearly the lowest free energy region on the entire landscape. It is approximately 12.0 kcal/mol lower in free energy than the open state of the loop. In the *holo* state, MT is again identifying a local minima for the open state of the loop, which is around ~0.7 kcal/mol lower in free energy and has a 2.5 Å RMSD from the open loop seen in the crystal structure which gives a free energy difference between the open and the closed state of -11.3 kcal/mol. The local minimums for the open state observed in both the *apo* and the *holo* states are almost similar with a backbone RMSD of 1.07Å. The superimposed open loop conformations predicted by MT in the *apo* and the *holo* states are shown in Figure S3.

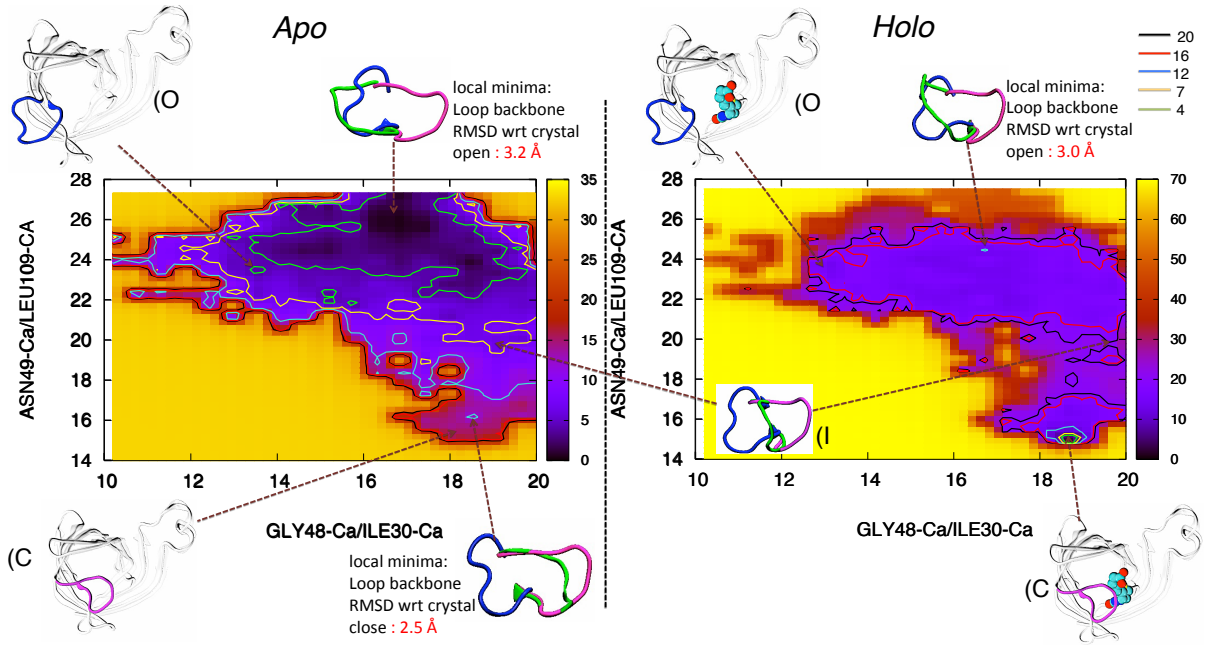

Figure 4.4 Relative free energy (kcal/mol) heat map for *apo* (left) and *holo* (right) streptavidin obtained using the Cα distance between ASN49/LEU109 and GLY48/ILE30 as reaction coordinate.

We also performed two 90 ns MD simulations starting from the experimental open and

closed state and overlaid the simulation results onto the MT generated FES using the same set of reaction coordinates (shown in Figure 4.5). The MD simulations details are described in Section 4.6.1 of the supporting information. In order to make a robust comparison, MD simulations were carried out using the ff99SBILDN[134] and ff14SB[135] force fields. Figure 4.5 highlights the MD snapshots generated by FF14SB simulations superimposed on the FES generated using MT. From the *apo* MT FES superimposed with the MD snapshots in Figure 4.5, we observe that the MD generated data points are most densely populated in the open region of the PMF with a horizontal spreading across the surface covering the low energy region observed by MT in the *apo* state. This shows that the MD simulation samples similar regions as predicted by the MT FES. For *holo*, MD generated data points are highly concentrated in the closed region of the *FES*, which is also identified as the lowest free energy point by MT in the ligand bound case. Similar trends are observed in the case of MD snapshots generated by FF99SBILDN simulations for both the *apo* and *holo* simulations as shown in Figure 4.11. Both the force fields behaved similarly and populated nearly the same regions in both the *apo* and *holo* simulations. Comparison between the two force fields is shown in Figure 4.12, which shows the superimposed MD snapshots obtained by both force fields for both the *apo* and *holo* simulations. As can be seen from the figure, the regions sampled by the two force fields are nearly on top of each other with small differences in the case of the *apo* simulation, highlighting the similarity of the two force fields in terms of their structural preferences.

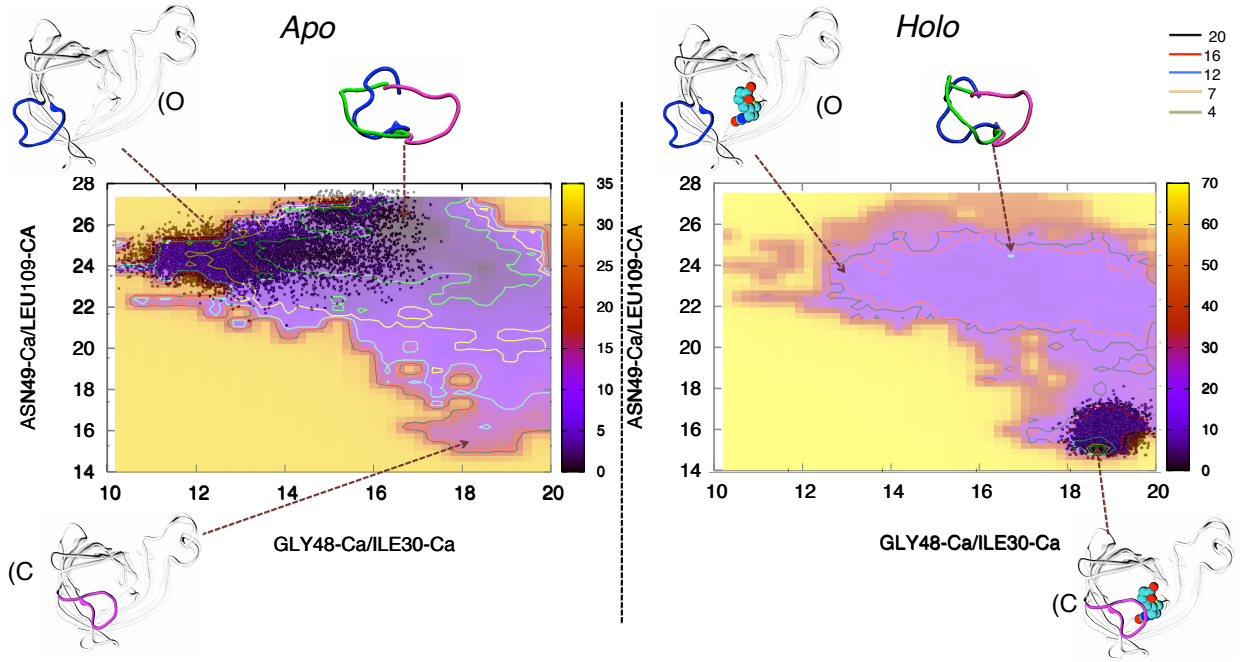

Figure 4.5 Relative free energy (kcal/mol) heat map of *apo* (left) and *holo* (right) streptavidin obtained using the Cα distance between Asn49/Leu109 and Gly48/Ile30 superimposed with the MD snapshots generated using FF14SB force fields. MD snapshots are highlighted in black color as scattered points in both the left and right figures with the rest of the PMF slightly faded in the background.

Further validation of the MT FES was performed using umbrella sampling (US) coupled with the weighted histogram analysis method (WHAM)[136, 137] in order to study the free energy change upon loop transitions in both the *apo* and the *holo* systems. The PMF simulations were also carried out using both the FF99SBILDN and FF14SB force fields. Both the PMF simulations ran in parallel and took roughly 10 days each to finish using a single K80 GPU processor in the MSU HPCC facility. The PMF simulations were initiated from the closed state to obtain the *holo* PMF and from the open state for the *apo* MD PMF. Initially, a diagonal scan was performed along the chosen set of reaction coordinates used to study the transition between the open and the closed states of the loop. After that, a detailed scan was performed successively in both the X and Y directions using every conformation generated from the diagonal scan. This

scanning strategy was adopted from our previous work on metallochaperones[138]. The procedure to generate the MD-PMF along with the necessary MD parameters and related details are provided in the 4.6.1 and 4.6.2 sections of the supporting information.

The PMF study shows significant differences between the two force fields. Figure 4.6 shows the MD-PMF obtained using the FF14SB force field for both the *apo* and *holo* simulations. The *apo* MD-PMF shows a free energy difference of 10.5 kcal/mol between the crystal open and crystal-closed state with the open state seen crystallographically as the global minimum. A local minimum was observed near the crystal-closed state (structure highlighted in the left image in Figure 4.6). The structural RMSD (Å) between the local minimum is ~2.85 Å from the crystal-closed state. This RMSD analysis only considers the loop region, but there was a large change in the orientation of the loop relative to the rest of the protein (see Figure 4.13). So, when the RMSD calculation is carried out via alignment on the entire protein, the loop backbone RMSD is estimated to be ~3.8 Å. The local minimum observed for the open state was quite similar to the experimental open state structure with a backbone RMSD of 0.41 Å. The free energy difference between the local minima in the *apo* state is roughly 9.4 kcal/mol. The free energy difference obtained from the MT-FES surface between the local minima of the closed and open states in the *apo* state is quite close (~9.1 kcal/mol). The *holo* MD-PMF (right image in Figure 4.6) estimates a free energy difference of 16 kcal/mol between the crystal closed and open state. A local minimum was identified here as well near the crystal open state. The backbone RMSD (Å) of the local minimum from the crystal structure of the open state is ~1.85Å. The total RMSD including the change in orientation was ~4.7 Å (structures shown in Figure 4.14). The local minimum observed for the closed state is almost same as the crystal closed state with the backbone loop RMSD of 0.67 Å. The free energy difference between the local minima in the

holo state is 11.7 kcal/mol. This free energy difference is quite similar to the free energy difference of 11.0 kcal/mol obtained from the MT-FES between the local minima and the experimental closed state. Since the free energy differences for the predicted minima were so similar between MT-FES and MD FF14SB PMF methodologies, we were also interested in comparing the structural differences between these two methods. Figure 4.15 highlights the superimposed structures obtained for the *apo* open, *apo* closed and *holo* open states predicted from the MT FES and the MD PMF (FF14SB) methods. The structures for the *apo* closed state have a Cα RMSD of 2.64 Å for the loop region. For the *apo* open state, the structure predicted by MT is curled inwards towards the protein relative to the MD FF14SB with a resultant Cα RMSD of 3.85 Å. The structure of the *holo* open state for MT is also curled inwards and is significantly different from the FF14SB structure with a RMSD of 5.90Å. MT-FES and FF14SB holo closed state is essentially identical to the experimental structure (see Figure 4.6 for details). Table 4.5 and 4.6 in the supporting information show the comparison of the phi-psi main chain dihedral between the crystal structure (closed and open state) with the dihedrals of the local minima obtained by $MT_{Flex\text{-}b}$ and from the MD FF14SB PMF simulations.

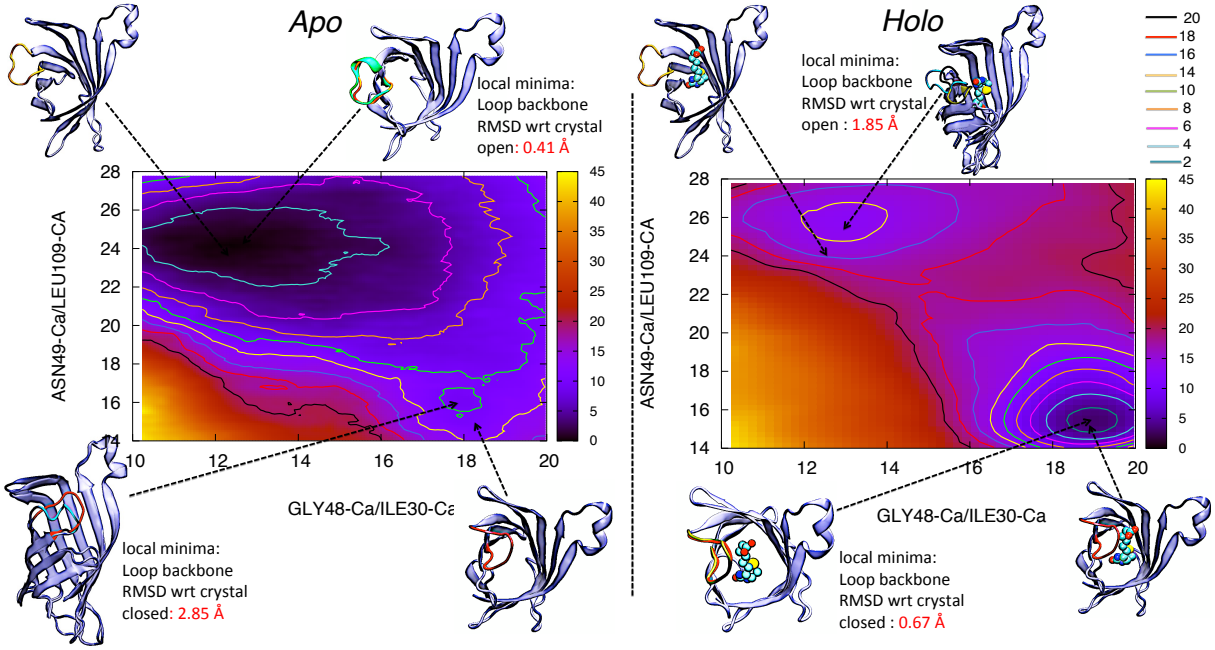

Figure 4.6 Relative free energy (kcal/mol) heat map for *apo* (left) and *holo* (right) streptavidin obtained by using umbrella sampling with the FF14SB force fields using the Cα distances between Asn49/Leu109 (y-axis) and Gly48/Ile30 (x-axis) as the two reaction coordinates.

From the *holo* MD FF14SB PMF plot in Figure 4.6 (right panel), a transition state can be clearly seen between the open and closed loop state. We further analyzed the structural and free energy differences between the transition state and the closed and open loop state to obtain some pathway details. Figure 4.7 schematically shows the one-dimensional relative free energy differences obtained using the open minima, closed minima and the transition state as observed in the FF14SB PMF simulations for the *holo* case. Residues forming hydrogen bond are highlighted in yellow color and residues forming vander Waal interactions are shown in green color in the figure 4.7.

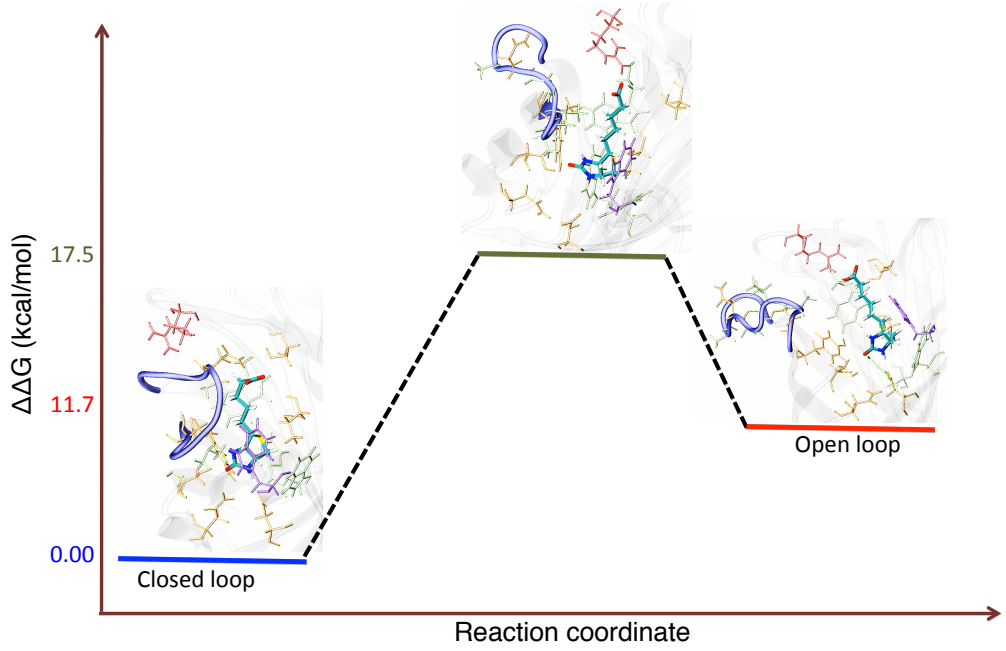

Figure 4.7 Computed changes in the relative free energy on going from the closed to the open *holo* state. The closed loop is the global minima in the *holo* state and the open loop is ~11 kcal/mol higher in free energy. The observed transition state barrier is ~17.5 kcal/mol on going from the closed to open loop.

In the *holo* state, the closed loop is the global minima and is favored because of numerous non-covalent interactions between biotin and streptavidin. The ureido oxygen of biotin has a hydrogen bond (H-bond) with the side chains of Ser27, Asn23 and Tyr43. The side chains of Ser45 and Asp128 form H-bonds with the hydrogen atoms attached to the two ureido nitrogen atoms. The sulfur from biotin also forms a H-bond with the side chain of Thr90. The carboxylate end of biotin forms H-bonds with the amide backbone of Asn49 and the side chain of Ser88. Stabilization by van der Waals interaction come from Trp70, Trp92, Trp105 and Trp120 (provided by the adjacent monomer). Apart from these Trp residues, three residues from the loop, Val47, Gly48 and Ala50 also provide hydrophobic interactions. Table 4.3 summarizes all the residues involved in H-bond and van der waal interactions with biotin.

The activation barrier observed by FF14SB MD PMF simulations in the *holo* state is ~17.5 kcal/mol. In the transition state, the loop (highlighted in purple ribbon in Figure 4.19) has partially opened up causing changes in the interactions present. Due to the extra space provided by partial loop opening, the carboxylate end of biotin has straightened up instead of being curled inwards while the H-bond between Asn49 and biotin is absent. The comparison of biotin's position in closed and transition state is shown in the left image of Figure 4.19. The carboxylate of biotin is slightly straightened and pushed out forming two H-bonds with the side chains of Arg84 (red in Figure 4.19). The first half of the loop is still folded in, but the side-chain of Ser45 has rotated breaking the H-bond with the ureido nitrogen of biotin. The H-bond between Asp128 and the other ureido nitrogen is also broken mainly due to the rotation of the side chain of Asp128. However, the H-bonds between the ureido oxygen and Tyr43 and Ser27 remain intact. The distance between Trp120 (from the adjacent monomer) and the ureido ring of biotin has also increased reducing the magnitude of the hydrophobic interaction. From the left image of Figure 4.19, Trp120 in the closed loop state (purple) is much closer to the biotin ring relative to its position in the transition state (black). As the loop opens up completely, the H-bond between biotin and loop residues Ser-45 and Asn49 remain severed. The carboxylate end of biotin is still forming 2 H-bonds with the side-chains of Arg84. The hydrophobic interactions between Trp120 and the ureido ring are lost as can be seen from the right image of Figure 4.19. The side-chain of Tyr43 moves closer to the ureido nitrogen thus forming a H-bond between O -Tyr and the H-ureido nitrogen along with the ureido oxygen. The hydrogen bond between Asp128 and the ureido nitrogens is re-established. The open loop is further stabilized by several intermolecular interactions with the protein itself. There is a H-bond observed between the side chain of Ser45

and backbone of Ser52, the backbones of Ser45 and Gly48, the backbone of Ser52 and the side chain of Asn81 and the side chains of Ser52 and Asn81.

The MD PMF obtained using the FF99SBILDN force fields is shown in Figure 4.16. In the *apo* PMF simulations by, the local minimum observed for the open state has an RMSD of 0.79 Å from experiment (shown in the left image in Figure 4.16). The local minimum for the closed state has a backbone RMSD of 2.36 Å with respect to experiment when the loops are aligned to each other. When the alignment is over the entire protein, the loop backbone RMSD is 4.64 Å (highlighted in Figure 4.17). The free energy difference between the crystal open and closed loop in the *apo* state is ~15 kcal/mol while the free energy difference between the observed minima is 14.6 kcal/mol. The *holo* state (right image in Figure 4.16) obtained using FF99SBILDN predicts a free energy difference of ~22 kcal/mol between the crystal open and closed states. The local minimum observed for the closed state is quite close to the crystal-closed state with a RMSD of 0.54 Å. The minimum in the case of *holo* PMF simulations has a backbone RMSD of 2.12 Å when aligned to the loop seen in the crystal structure and 2.78 Å when aligned to the entire protein (shown in Figure 4.18). The free energy difference observed between the two minima in the *holo* system is 21.3 kcal/mol.

With these PMFs in hand the free energy of binding in the open and closed states (exp = -18.26 kcal/mol) can be extrapolated from the free energy differences observed in the *apo* and *holo* states (via construction of a thermodynamic cycle). Using this approach, the free energy differences between open and closed loop states predicted using the FF99SBILDN force field are: free energy of binding to the closed state is -33.3 kcal/mol and to the open state is +3.7 kcal/mol. The former represents a significant overestimation of the experimental binding affinity, while the latter is an underestimation. The free energy differences observed using the FF14SB

force field are in better accord with experiment and with the free energy differences observed using MT FES. The free energy differences generated via different methodologies are summarized in Table 4.1.

Table 4.1 Summary of free energy differences (in kcal/mol) between open and closed states of the loop in *apo* and *holo* states obtained with MT-free energy surface (MT-FES), MT-thermodynamic cycle (MT-TC), and MD using FF99SBILDN and FF14SB force fields. The differences are reported for the crystal open and loop position and minima obtained by the MT and MD methods.

ΔG (kcal/mol)	MT-FES	MT-TC	MD-FF14SB	MD-FF99SBILDN
$\Delta G^{O->C}_{Apo\text{-}crystal}$	10.5	10.4	10.5	15
$\Delta G^{O->C}_{Holo\text{-}crystal}$	–12.0	–12.0	–16	–22
$\Delta G^{O->C}_{Apo\text{-}minima}$	9.1	9.0	9.4	14.6
$\Delta G^{O->C}_{Holo\text{-}minima}$	–11.3	–11.3	–11.7	-21.3

The observed free energy differences between the closed and open states obtained by MD-FF99SBILDN, MD-FF14SB and MT are largely different due to the different energy functions employed by the methods. Sampling effects, *etc*., also play a role, but the energy functions used are almost certainly the largest source of variance. Overall, the free energy trends are qualitatively similar in both (MT and MD) methods with the closed state being the lowest in free energy in the *holo* state and the least stable in the *apo* state. We performed a more detailed analysis of our free energy calculations to figure out the reason of stability of open loop in the *apo* state and stability of closed loop in *holo* state. Table 4.4 highlights the individual components of free energy differences for both the *apo* and *holo* states. It can be seen from the table that for the *apo state*, the net free energy change in the gas phase favors the closed loop over the open loop by 1.2 kcal/mol but the solvation free energy change is highly favorable for the open state (-11.6 kcal/mol) as compared to the closed state favoring the open state when the

ligand is not bound in the solution phase. So, based on our results we observe that the loop prefers the open state to closed state by 10.4 kcal/mol mainly due to a favorable solvation free energy in the open state.

In the *holo* state as well, the solvation free energy is highly favorable for the open loop when compared to the closed loop but the streptavidin-biotin interactions are much stronger for the closed loop (~24 kcal/mol) making the closed loop more favorable in the *holo* state by ~12 kcal/mol.

The free energy differences observed between PMF obtained using FF14SB and MT-FES method are comparable in both the *apo* and *holo* states. In the *apo* state, both FF14SB and MT-FES estimate a free energy difference of 10.5 kcal/mol between the closed and open state seen experimentally. For the *holo* case, the free energy difference between the experimentally observed open and closed states is 16 kcal/mol for the former and ~12 kcal/mol for the latter. But, the free energy differences between the local minimums obtained via MT-FES and FF14SB PMF are quite similar (see Table 4.1). The fact that MT gives comparable free energy differences demonstrates its relevance as a relatively efficient technique for estimating free energies.

4.3.3 Thermodynamic free energy cycle

The binding free energy of the streptavidin-biotin system in the aqueous phase was estimated using a thermodynamic cycle linking the open and the closed states of the $loop_{3\text{-}4}$ of streptavidin protein in the *holo* (ligand bound) and the *apo* (free) states as the end states (see Figure 4.8). The end states were represented by selecting an ensemble of closed-like and open-like conformations from the MT-FES windows containing the experimental closed and open

state of the loop. The free energies for each of the $MT_{Flex\text{-}b}$ conformations in the selected ensembles were calculated in the *holo* (ligand bound) and the *apo* (free) states using the MT method in the aqueous phase as described in the Methods section.

The free energies of the end states (open *apo*, open *holo*, closed *apo* and closed *holo*) for the thermodynamic cycle were estimated by taking the lowest free energy conformation for each of the end states individually. From the thermodynamic cycle (Figure 4.8a), it can be seen that the open loop in the *apo* state is ~10.4 kcal/mol more stable than the closed loop. The trend is in modest agreement with the Song *et al* MD study, which estimated the stability of open loop to be ~5 kcal/mol more stable than the closed loop in the *apo* state[129]. In the *holo* state, the closed loop is estimated to be more stable (as expected) than the open state (by 12 kcal/mol). The free energy differences between the open and closed states are consistent with those estimated by the MT-FES method (see Table 4.1 for a summary). A slight difference in the relative free energy differences exist because the *apo* state in the TC accounts for the free energy contributions from the free states of both protein and ligand while in MT-FES, the *apo* state FES correspond to only the free state of the protein.

Apart from the relative stability of the open and closed loops in the ligand un-bound and bound states; we can also estimate the free energy change upon ligand binding. The free energy change upon biotin binding to the open loop state is estimated to be –4.2 kcal/mol (as shown in Figure 4.8a). A study by Chu *et al* in 1998 estimated that the binding affinity of Streptavidin-Biotin was –10 kcal/mol after deleting $loop_{3\text{-}4}$ via circular permutation[120]. Based on this analysis, we are underestimating the binding affinity of biotin in the open loop state and overestimating the free energy change upon loop ordering in the ligand bound state (by ~6 kcal/mol). The red and the blue arrows shown in Figure 4.8a represent the two paths that can be

taken to estimate the free energy of binding. The estimated binding free energy of biotin taking either path is ~ –16.2 kcal/mol. Our prediction for binding free energy change is in reasonable accord with the experimental binding affinity of –18.3 kcal/mol[139, 140].

$$\Delta G_{binding} = \Delta G^{O->C}_{unbound} + \Delta G^{C}_{bind} = \Delta G^{O}_{bind} + \Delta G^{O->C}_{bound} \quad (4.1)$$

$$\Delta G_{binding} = -16.2 \text{ kcal/mol}$$

The high binding affinity of biotin to streptavidin is attributed to three major factors: extensive hydrogen bond network of streptavidin with the ureido and carboxylate groups of biotin, hydrophobic interactions between several residues and biotin and the closure of the $loop_{3-4}$ over the Biotin binding site[123, 127, 141, 142]. Out of all the hydrophobic interactions, Trp120 from an adjacent monomer (highlighted in Figure 4.8) forms a key hydrophobic interaction with biotin bound (Trp120-biotin interactions are formed between monomers A and D, and, B and C). Its importance in the binding free energy calculations has been shown previously [124, 130, 131]. Hence, reinforcing our use of the tetrameric structure in our calculations.

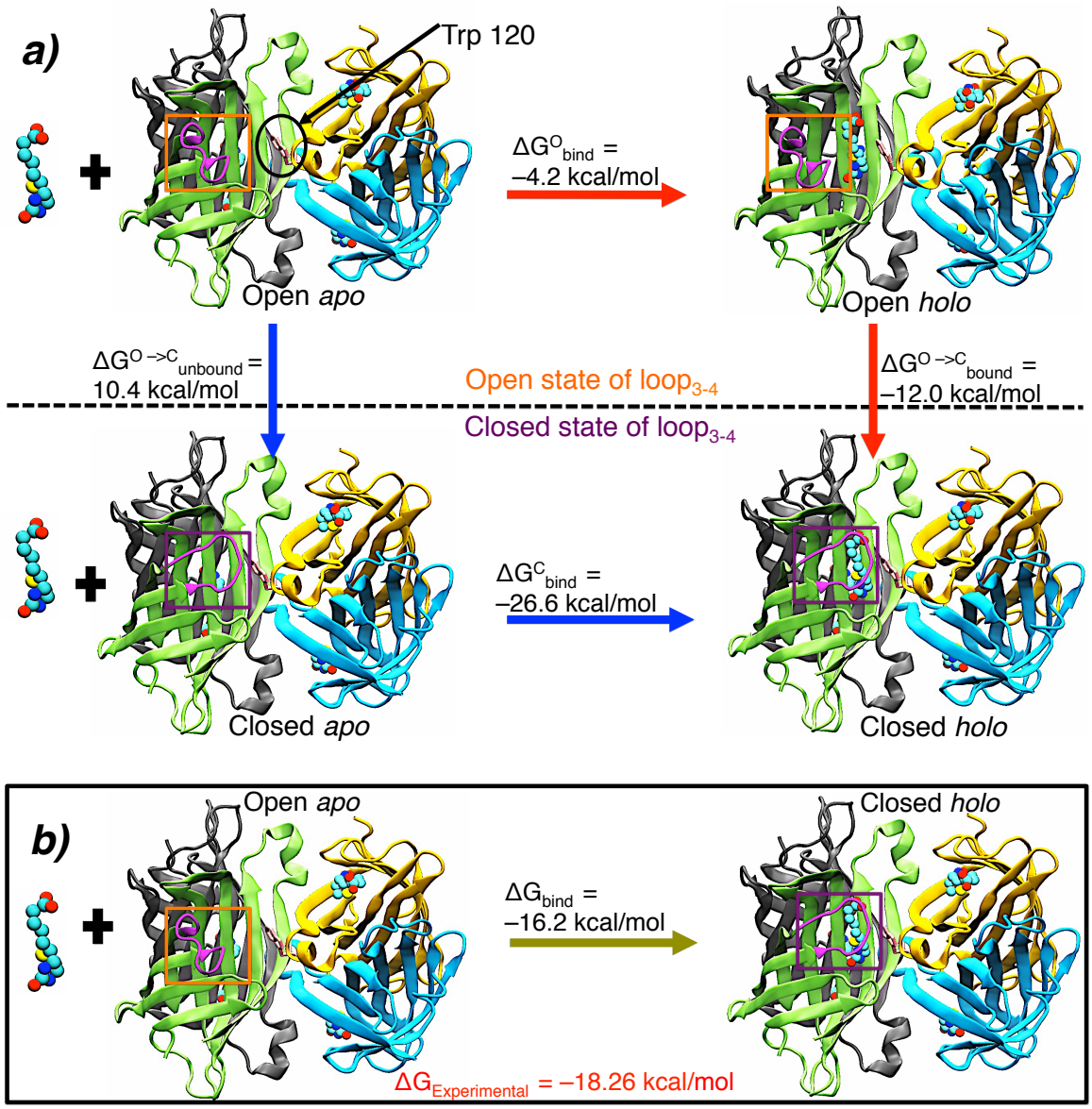

Figure 4.8 a) Detailed representation of thermodynamic free energy cycle for binding in solution phase for the streptavidin-biotin system. b) The net free energy change upon biotin binding and loop closure in the solution phase.

4.4 Conclusions

Rapidly obtaining accurate free energies remains a daunting problem, which if solved will have a major impact on structure based drug design[1-3]. Despite the many attempts to address this issue, we still need fast and accurate methods to obtain free energies for biochemical

processes[6, 7]. To a large extent, the current state of affairs is rooted in our use of compute intensive conformational sampling strategies. The conformational sampling issue is especially notable for large biomolecules like proteins[82, 84, 114]. Often, proteins possess a highly flexible loop or intrinsically disordered region (IDR)[143-145], which may undergo conformational changes of functional relevance. Therefore estimation of the energetic cost of loop motions in molecular recognition processes is of great import. To fully understand the free energy changes induced due to loop conformational changes, it is important to model the equilibrium flexibility of a protein loop, which remains a computationally intensive problem using our current sampling and free energy protocols.

To address this ongoing challenge, we introduce a novel approach for relatively fast and accurate estimation of free energies. The significant configurational states of the protein are sampled by using $MT_{Flex\text{-}b,}$ which uses 1-D atom pairwise databases to generate molecular conformations. Each of the generated molecular conformations serves as the "seed structure" for the MT method, which estimates the free energies by performing local partition function estimations. Using the parallel mechanics of the MT method, this strategy provides us with a trade off between sampling speed and accuracy.

This procedure was applied to study the free energy changes associated with the movement of the $loop_{3\text{-}4}$ region within streptavidin in both the free and bound states. The large conformational ensemble of the $loop_{3\text{-}4}$ region generated by $MT_{Flex\text{-}b}$ contained both closed-like and open-like loop conformations with a best structural RMSD of ~1.6 Å with respect to experiment. We generated the MT FES and found that the closed state of the loop is more stable in the *holo* state with the open loop being least stable. Similarly, in the *apo* state, the trend is reversed and the closed loop is found to be least stable. We also carried out MD PMF

simulations using umbrella sampling and obtained similar trends for both the *apo* and *holo* PMFs using the FF99SBILDN and FF14SB force fields relative to the MT FES. Even though the MT FES is less continuous than the MD PMF the trend in free energy is similar in both the MT and MD PMF plots. In particular, the free energy differences predicted by the FF14SB PMF simulations and MT-FES method were in good accord, while FF99SBILDN provide less internally consist free energy results. Clearly, the reparameterization carried out by the Simmerling group (creating FF14SB) has made significant improvements to the modeling of the processes highlighted herein[135]. To summarize, the free energy difference observed by MD-FF14SB between closed and open state in the *apo* state is 10.5 kcal/mol, which is exactly the same as the free energy difference estimated by MT FES. In the *holo* state, the free energy difference between the experimental closed and open states differ by ~4kcal/mol between the estimates provided by MD FF14SB and MT FES. But, the free energy differences between the minima estimated by the two methods are quite comparable at ~11kcal/mol. Our free energy components estimate that solvation is playing a key role in the stability of the open loop in the *apo* state. In the *holo* state as well, solvation is significantly favoring the open loop, but the contacts between Streptavidin and Biotin are stabilizing the closed loop much more favorably making it ~12 kcal/mol more stable than the open loop.

Using a free energy thermodynamic cycle, we observed that the relative stability of the open loop in the *apo* state is -10.4 kcal/mol with respect to the closed state, while in the *holo* state; the closed loop conformation was estimated to be -12 kcal/mol more stable than the open loop. The binding free energy change was also obtained and was estimated to be –16.2 kcal/mol, which is in reasonable agreement with the experimental binding affinity of –18.3 kcal/mol.

Give the success of the MT based approach, we can envision this method being applied to

other examples of free energy changes accompanied by the conformational mobility of loops in other protein systems including. As the size of the loops increases the complexity and the computational expense for the conformational search would concomitantly increase. However, with advancement in GPU and CPU processors and parallel computation coupled with enhanced software abilities, all of these problems are addressable using our approach.

4.5 Acknowledgement

This work was supported in part by Michigan State University through computational resources provided by the Institute for Cyber-Enabled Research.

4.6 Supporting information

Table 4.2 The total number of backbone loop conformations and the (backbone + sidechain) conformations generated by $MT_{Flex\text{-}b}$ in the presence of Biotin bound in the active site of crystal streptavidin monomer.

	Number of conformations in the presence of Biotin
Backbone conformations	21,295
Backbone and side chain conformations	11546486

Figure 4.9 Cartoon representation of Streptavidin crystal monomer (Chain A) with the $loop_{3\text{-}4}$ in the closed conformation. The reaction coordinates are distances between Cα atoms of Resid30 and 48, and between Resid109 and 49.

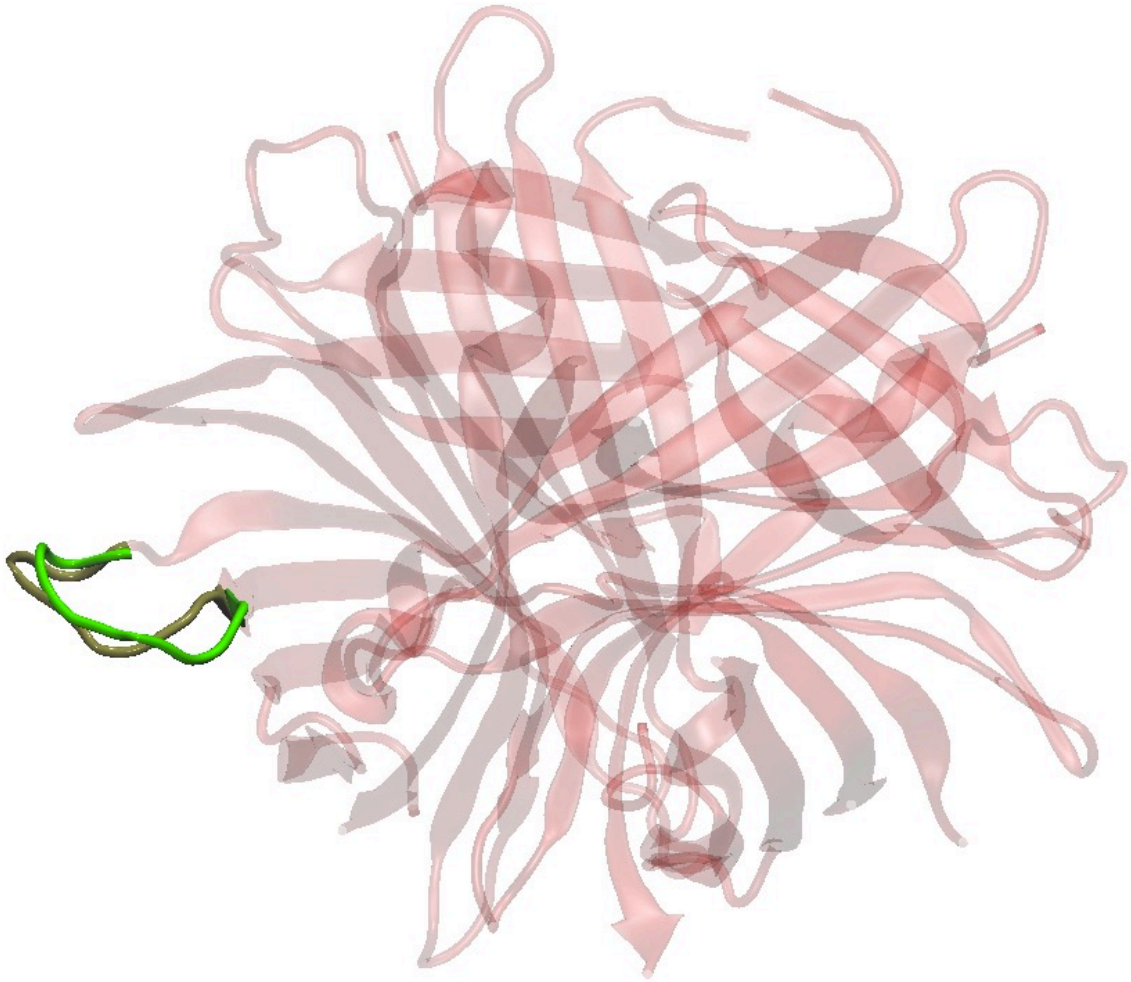

Figure 4.10 Superimposed conformations of open minima observed in MT-*apo* state (shown in tan color) and MT-holo state (highlighted in green color). The observed Cα RMSD is ~1.07 Å.

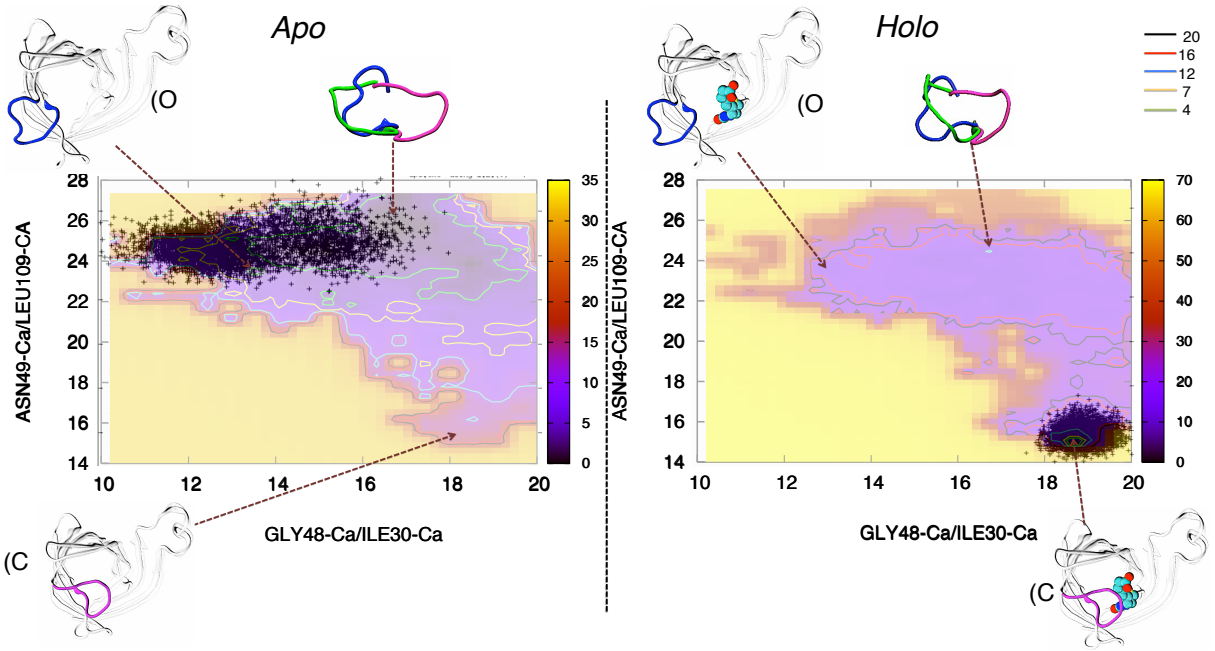

Figure 4.11 Relative free energy (kcal/mol) heat map of *apo* (left) and *holo* (right) streptavidin obtained using the Cα distance between Asn49/Leu109 and Gly48/Ile30 superimposed with the MD snapshots generated using FF99SBILDN force fields. MD snapshots are highlighted in black color as scattered points in both the left and right figures with the rest of the PMF slightly faded in the background.

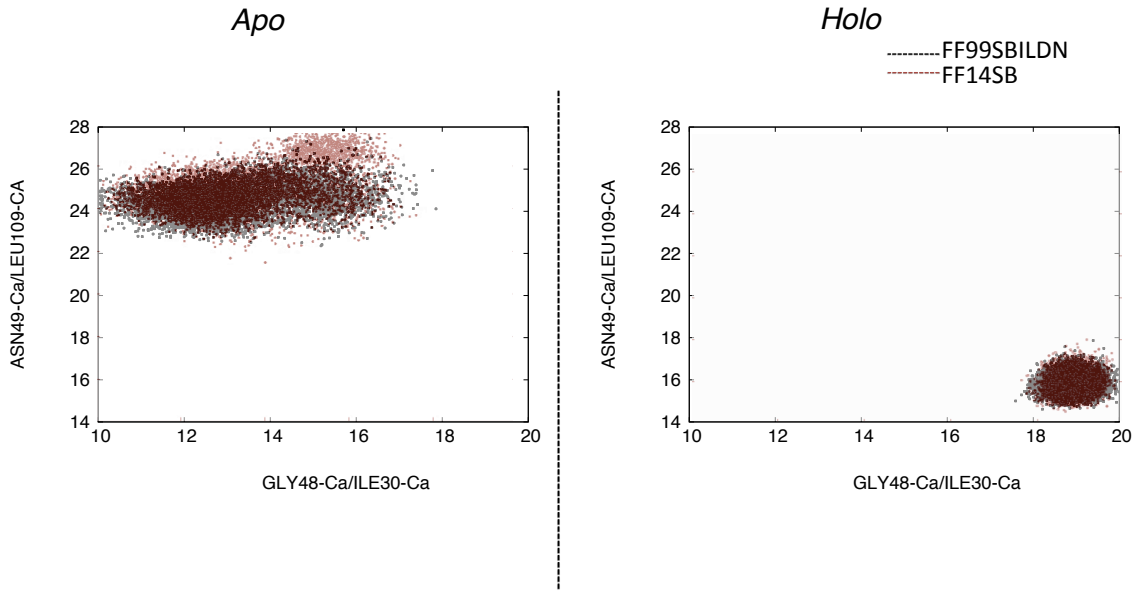

Figure 4.12 Superimposed MD snapshots generated using FF99SBILDN (shown as black scattered points) and FF14SB (pink scattered points) force fields obtained using the Cα distance between Asn49/Leu109 and Gly48/Ile30

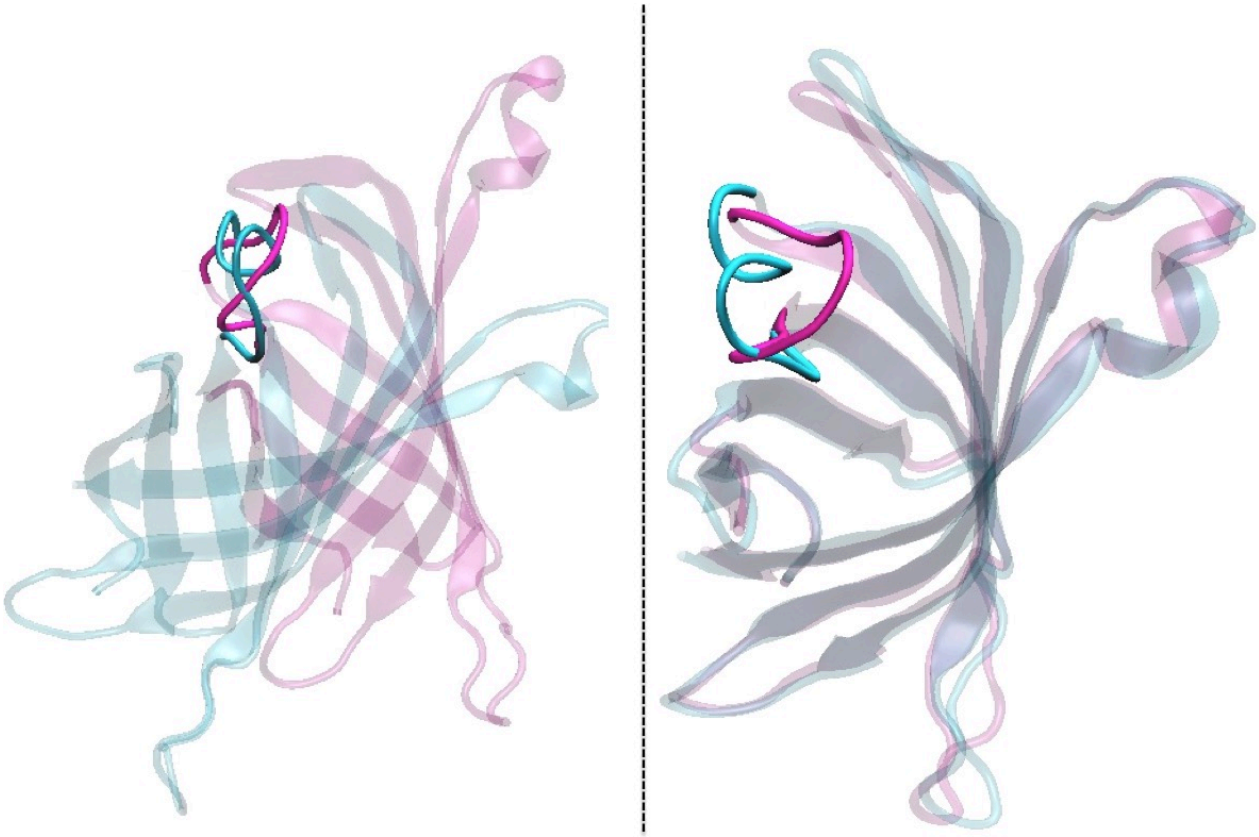

Figure 4.13 The crystal-closed state (shown in pink) superimposed with the *apo*-closed minima obtained by MD-FF14SB force field (highlighted in aqua color). The left panel shows the alignment based on only the loop region while the right panel shows the alignment based on the entire protein monomer.

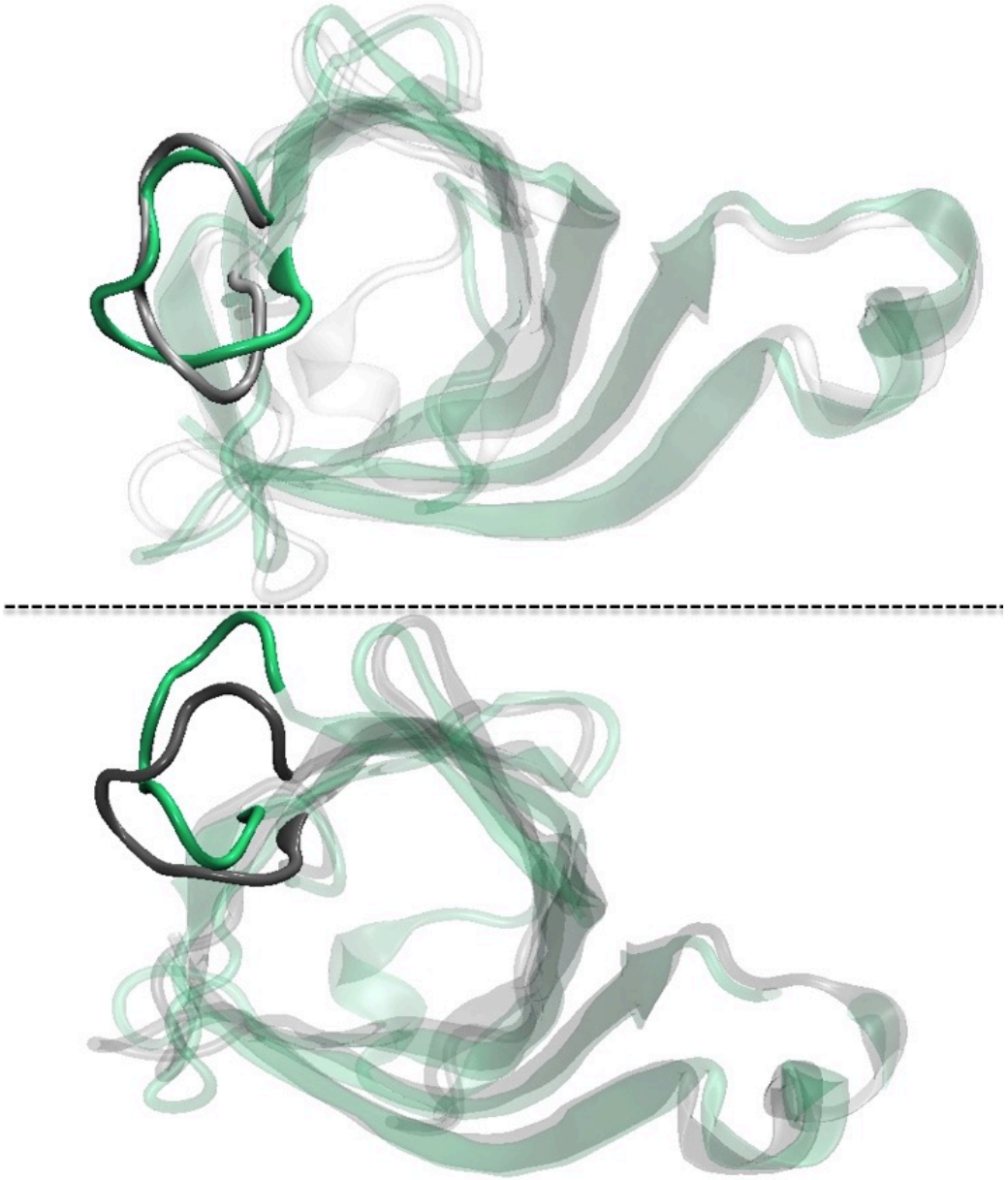

Figure 4.14 The crystal open state (shown in brown) superimposed with the *holo*-open minima obtained by MD-FF14SB force field (highlighted in green color). The upper panel shows the alignment based on only the loop region while the lower panel shows the alignment based on the entire protein monomer.

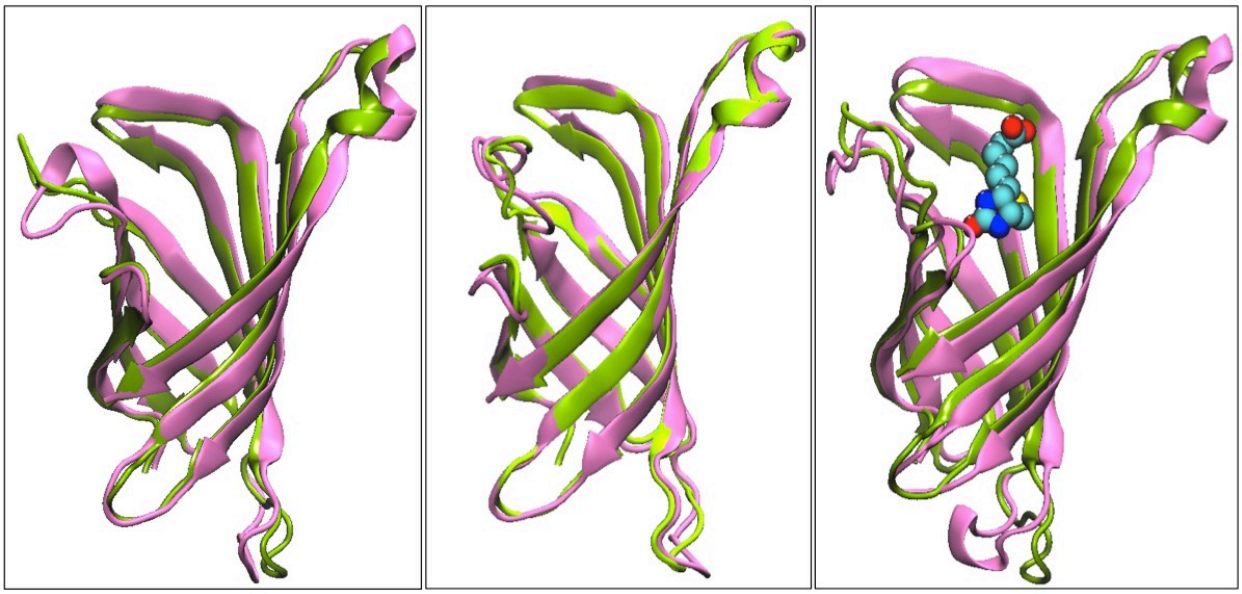

Figure 4.15 The superimposed minima conformations predicted by MT (green) and MD-FF14SB (mauve) methods for *apo* open minima, *apo* closed minima and *holo* open minima (from left to right).

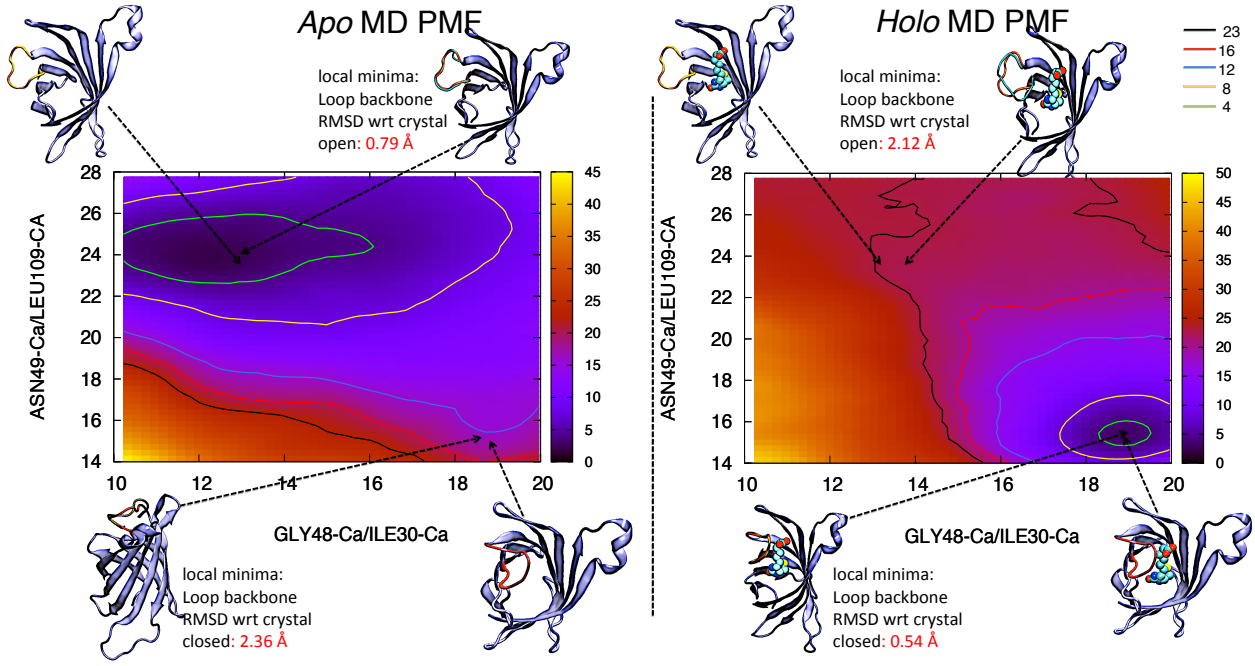

Figure 4.16 Relative free energy (kcal/mol) heat map for *apo* (left) and *holo* (right) streptavidin obtained by using umbrella sampling with FF99SBILDN force fields using the Cα distances between ASN49/LEU109 and GLY48/ILE30 as the two reaction coordinates.

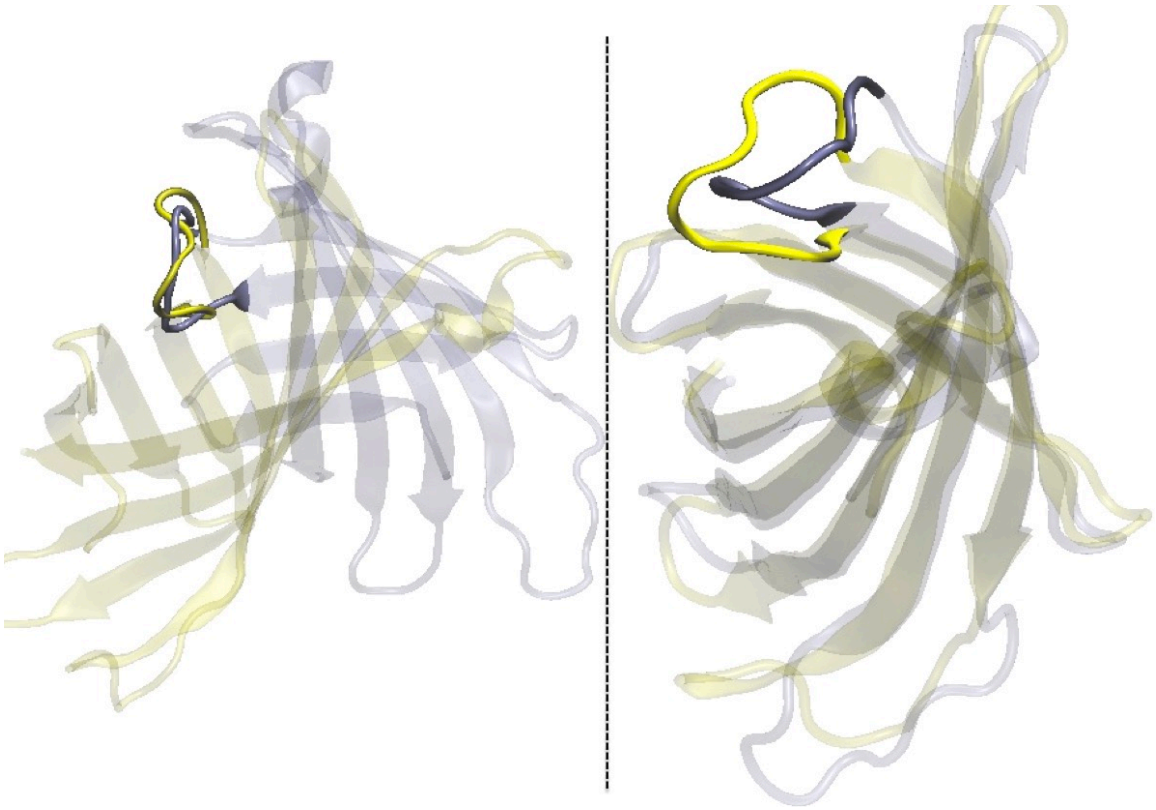

Figure 4.17 The crystal-closed state (shown in blue) superimposed with the *apo*-closed minima obtained by MD-FF99SBILDN force field (highlighted in yellow color). The left panel shows the alignment based on only the loop region while the right panel shows the alignment based on the entire protein monomer.

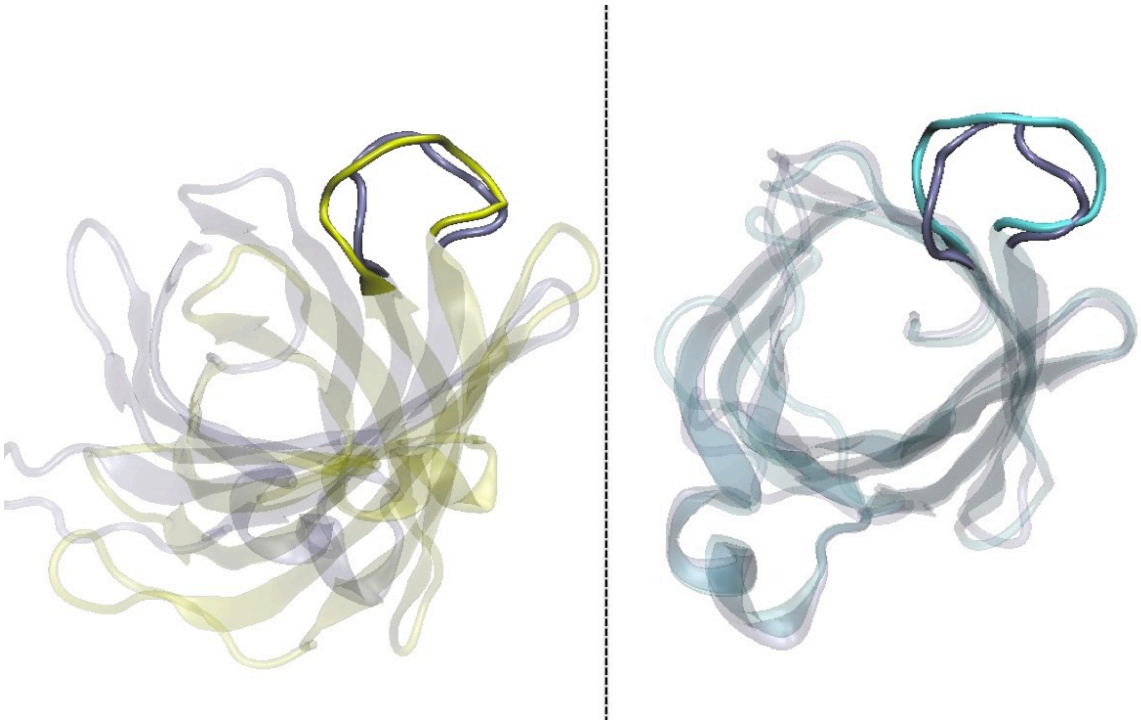

Figure 4.18 The crystal-open state (shown in blue) superimposed with the *holo*-open minima obtained by MD-FF99SBILDN force field (highlighted in yellow and cyan color). The left panel shows the alignment based on only the loop region while the right panel shows the alignment based on the entire protein monomer.

Table 4.3 List of the residues involved in forming hydrogen bond and vander waal interactions with Biotin. The residues color-coded in red are part of the $loop_{3\text{-}4}$.

	Hydrogen-bond interactions				Vander-waal interactions	
Loop state	Ureido oxygens	Ureido Nitrogens	Biotin Sulfur	Carboxylate end	Trp120	Other hydrophobic interactions
Closed	Ser27, Asn23 and Tyr43	Ser45 and Asp128	Thr90	Amide backbone of Asn49 and side chain of Ser88	Yes	Trp70, Trp92, Trp105, Ala46, Val47, Gly48, Ala50
Transition	Ser27 and Tyr43	-	Thr90	2 H-bonds with Arg84	Partial	Trp70, Trp92, Trp105, Ala46
Open	Ser27 and Tyr43	Tyr43 and Asp128	Thr90	2 H-bonds with Arg84	Lost	Trp70, Trp92, Trp105

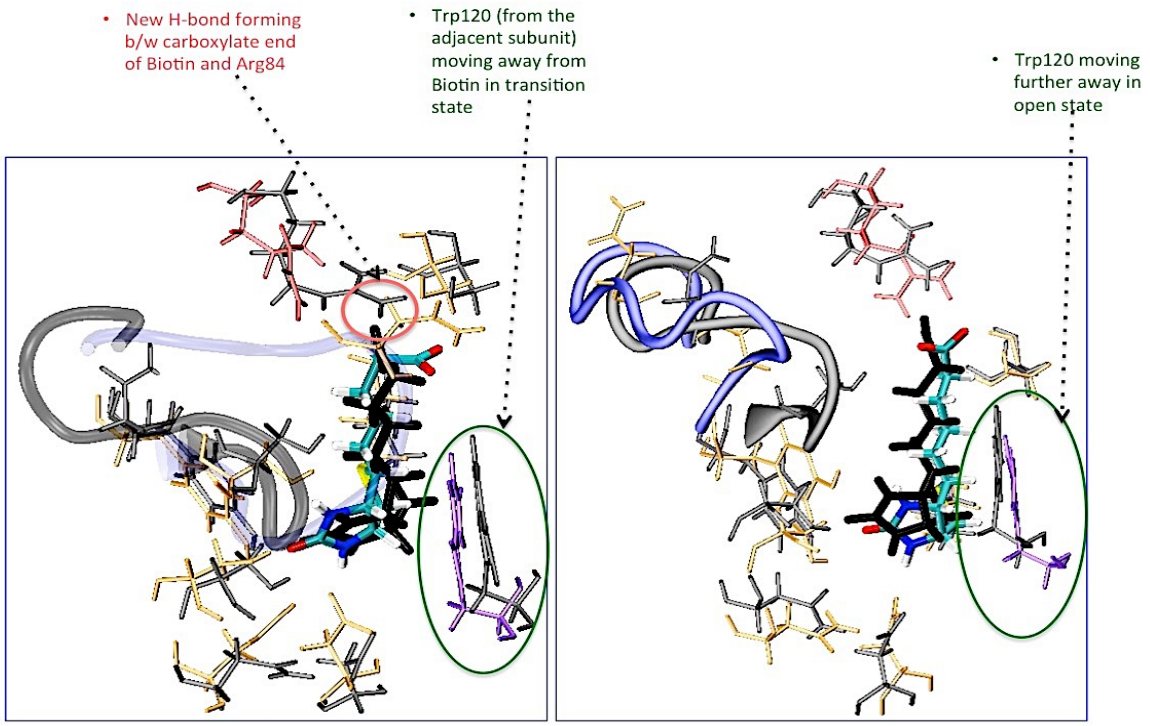

Figure 4.19 The active site of streptavidin with Biotin shown for the closed loop (left) and open loop (right) states superimposed with the structure of the active site at the transition state (shown in black in both plots). Residues forming hydrogen bonds are highlighted in orange, residues forming van der Waals interactions are shown in green, Trp120 from sub-unit D is shown in purple and Arg84 in red.

Table 4.4 The free energy difference between open and closed loop state with the free energy components in the form of the change in the protein torsion free energy, the change in the protein non-covalent interactions, the free energy change in the protein-ligand interactions, they change in the solvation free energy, the net free energy change in gas phase and in solution phase. All free energy values are in kcal/mol.

	$\Delta\Delta G_{Torsion\text{-}protein}$	$\Delta\Delta G_{Intra\text{-}protein\ (Non\text{-}covalent)}$	$\Delta\Delta G_{Inter\text{-}ProteinLigand}$	$\Delta\Delta G_{solvation}$	Net change $\Delta\Delta G$ (Gas phase)	Net change $\Delta\Delta G$ (Solution phase)
Apo state	1.8	-0.6	0	-11.6	1.2	-10.4
Holo state	-1.0	2.1	23.3	-12.4	24.4	12.0

Table 4.5 Comparison of main chain phi-psi dihedrals of all the loop residues of the experimental closed structure with the closed minima structures obtained using $MT_{Flex\text{-}b}$ and from the FF14SB MD simulations. The differences are shown as positive or negative from a reference value of 0º.

		$MT_{Flex\text{-}b}$ close minima	FF14SB close minima
Res 45	phi	+1.37	+15.15
	psi	+10.26	+13.97
Res 46	phi	+71.55	+10.61
	psi	+33.87	+1.11
Res 47	phi	+70.34	+3.27
	psi	+21.30	+5.24
Res 48	phi	–160.41	+14.43
	psi	–32.50	+4.26
Res 49	phi	–115.34	+20.67
	psi	–137.97	+89.59
Res 50	phi	+59.84	+89.75
	psi	+17.29	+1.26
Res 51	phi	+2.30	+5.73
	psi	+31.30	+13.13
Res 52	phi	+22.10	+5.77
	psi	–52.41	–33.37

Table 4.6 Comparison of main chain phi-psi dihedrals of all the loop residues of the experimental open structure with the open minima structures obtained using $MT_{Flex\text{-}b}$ and from the FF14SB MD simulations. The differences are shown as positive or negative from a reference value of 0º.

		$MT_{Flex\text{-}b}$ open minima	FF14SB open minima
Res 45	phi	+35.33	+21.01
	psi	–93.95	+19.33
Res 46	phi	+30.28	+12.37
	psi	–44.62	+31.99
Res 47	phi	+18.29	+11.61
	psi	+23.13	+9.44
Res 48	phi	–75.16	+16.22
	psi	+95.92	+20.37
Res 49	phi	-57.22	+3.69
	psi	+34.18	+3.91
Res 50	phi	–94.64	+2.36
	psi	+123.47	+8.79
Res 51	phi	+14.86	+15.11
	psi	+44.38	+26.04
Res 52	phi	+38.06	+11.49
	psi	+28.54	+8.39

4.6.1 MD-PMF methodology

Starting from the *holo* crystal closed structure and *apo* crystal open state, we performed PMF simulations that transited to the open or closed state through a series of windows where the two "reaction coordinate" distances increased/decreased by the window size successively[138]. The window size for the X and Y coordinate was chosen to be 0.32 Å and 0.46 Å, respectively. There were a total of 32 windows in this diagonal scan. For each window, 1ns equilibration and 2ns sampling was performed, and the data points were stored every 200fs. The starting structure of each window came from the final sampling snapshot of the previous window. After that, the last snapshot of each of the 32 diagonal windows was used as starting structure to sample along the X-axis while maintaining the Y coordinate distance. Again, data points were collected every 200fs and total sampling was 2ns for each window, with the starting structure coming from the final sampling snapshot from the previous window. This time, there were total 32 parallel runs and for each run there were 34 successive windows along the X-axis. The force constant for both "reaction coordinate" was 60 kcal/(mol*Å^2) for all the US windows. The two-dimensional WHAM was utilized to obtain the free energy profile[137].

4.6.2 MD simulation details

The simulations were performed using the FF99Sbildn[134] and FF14SB[135] force fields of the AMBER14 software package[146]. Each protein structure was solvated in a periodically replicated octahedral box using the TIP3P water model. SHAKE was used to constrain bonds with hydrogen atoms and a time step of 2fs was used during all the MD runs. The Particle mesh Ewald method[147-149] was used to treat long-range electrostatic interactions and a cutoff of 12.0 Å was used for the non-bonded interactions.[150] The counter-ions were

added to neutralize the system and the parameters for the ions were taken from IOD parameter set[151]. The parameters for Biotin were obtained using the GAFF force field[152]. HF/6-31G* level theory was used to optimize the Biotin structure and charges were fitted by using RESP fitting at MP2/6-31G* level[153].

REFERENCES

REFERENCES

1. Gilson, M.K. and H.X. Zhou, *Calculation of protein-ligand binding affinities.* Annual Review of Biophysics and Biomolecular Structure, 2007. **36**: p. 21-42.

2. Michel, J. and J.W. Essex, *Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations.* Journal of Computer-Aided Molecular Design, 2010. **24**(8): p. 639-658.

3. Shirts, M.R., D.L. Mobley, and J.D. Chodera, *Alchemical Free Energy Calculations: Ready for Prime Time?* Annual Reports in Computational Chemistry, Vol 3, 2007. **3**: p. 41-59.

4. Halperin, I., et al., *Principles of docking: An overview of search algorithms and a guide to scoring functions.* Proteins-Structure Function and Genetics, 2002. **47**(4): p. 409-443.

5. Yuriev, E. and P.A. Ramsland, *Latest developments in molecular docking: 2010-2011 in review.* Journal of Molecular Recognition, 2013. **26**(5): p. 215-239.

6. Chodera, J.D., et al., *Alchemical free energy methods for drug discovery: progress and challenges.* Current Opinion in Structural Biology, 2011. **21**(2): p. 150-160.

7. Christ, C.D., A.E. Mark, and W.F. van Gunsteren, *Feature Article Basic Ingredients of Free Energy Calculations: A Review.* Journal of Computational Chemistry, 2010. **31**(8): p. 1569-1582.

8. Woo, H.J. and B. Roux, *Calculation of absolute protein-ligand binding free energy from computer simulations.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(19): p. 6825-6830.

9. Zwanzig, R.W., *High-Temperature Equation of State by a Perturbation Method .1. Nonpolar Gases.* Journal of Chemical Physics, 1954. **22**(8): p. 1420-1426.

10. Kirkwood, J.G., *Statistical mechanics of fluid mixtures.* Journal of Chemical Physics, 1935. **3**(5): p. 300-313.

11. Buch, I., S.K. Sadiq, and G. De Fabritiis, *Optimized Potential of Mean Force Calculations for Standard Binding Free Energies.* Journal of Chemical Theory and Computation, 2011. **7**(6): p. 1765-1772.

12. Essex, J.W., et al., *Monte Carlo simulations for proteins: Binding affinities for trypsin-benzamidine complexes via free-energy perturbations.* Journal of Physical Chemistry B, 1997. **101**(46): p. 9663-9669.

13. Gumbart, J.C., B. Roux, and C. Chipot, *Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy?* Journal of Chemical Theory and Computation, 2013. **9**(1): p. 794-802.

14. Roux, B., *The Calculation of the Potential of Mean Force Using Computer-Simulations.* Computer Physics Communications, 1995. **91**(1-3): p. 275-282.

15. Jorgensen, W.L., *Free-Energy Calculations - a Breakthrough for Modeling Organic-Chemistry in Solution.* Accounts of Chemical Research, 1989. **22**(5): p. 184-189.

16. Klimovich, P.V., M.R. Shirts, and D.L. Mobley, *Guidelines for the analysis of free energy calculations.* Journal of Computer-Aided Molecular Design, 2015. **29**(5): p. 397-411.

17. Shirts, M.R. and D.L. Mobley, *An introduction to best practices in free energy calculations.* Methods Mol Biol, 2013. **924**: p. 271-311.

18. Bajorath, J. and S. Sheriff, *Comparison of an antibody model with an X-ray structure: The variable fragment of BR96.* Proteins-Structure Function and Genetics, 1996. **24**(2): p. 152-157.

19. Streaker, E.D. and D. Beckett, *Ligand-linked structural changes in the Escherichia coli biotin repressor: The significance of surface loops for binding and allostery.* Journal of Molecular Biology, 1999. **292**(3): p. 619-632.

20. Nestl, B.M. and B. Hauer, *Engineering of Flexible Loops in Enzymes.* Acs Catalysis, 2014. **4**(9): p. 3201-3211.

21. Bahar, I., et al., *Global Dynamics of Proteins: Bridging Between Structure and Function.* Annual Review of Biophysics, Vol 39, 2010. **39**: p. 23-42.

22. Meireles, L., et al., *Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins.* Protein Science, 2011. **20**(10): p. 1645-1658.

23. Papaleo, E., et al., *The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery.* Chemical Reviews, 2016. **116**(11): p. 6391-6423.

24. Saraste, M., P.R. Sibbald, and A. Wittinghofer, *The P-loop--a common motif in ATP- and GTP-binding proteins.* Trends Biochem Sci, 1990. **15**(11): p. 430-4.

25. Slesinger, P.A., Y.N. Jan, and L.Y. Jan, *The S4-S5 loop contributes to the ion-selective pore of potassium channels.* Neuron, 1993. **11**(4): p. 739-49.

26. Stuart, D.I., et al., *Alpha-lactalbumin possesses a novel calcium binding loop.* Nature, 1986. **324**(6092): p. 84-7.

27. Steichen, J.M., et al., *Structural Basis for the Regulation of Protein Kinase A by Activation Loop Phosphorylation.* Journal of Biological Chemistry, 2012. **287**(18): p. 14672-14680.

28. Bernstein, L.S., et al., *RGS2 binds directly and selectively to the M1 muscarinic acetylcholine receptor third intracellular loop to modulate Gq/11alpha signaling.* J Biol Chem, 2004. **279**(20): p. 21248-56.

29. Ciarapica, R., et al., *Molecular recognition in helix-loop-helix and helix-loop-helix leucine zipper domains - Design of repertoires and selection of high affinity ligands for natural proteins.* Journal of Biological Chemistry, 2003. **278**(14): p. 12182-12190.

30. Kiss, C., et al., *Antibody binding loop insertions as diversity elements.* Nucleic Acids Research, 2006. **34**(19).

31. Fukuchi, S., et al., *Intrinsically disordered loops inserted into the structural domains of human proteins.* Journal of Molecular Biology, 2006. **355**(4): p. 845-857.

32. Michalsky, E., A. Goede, and R. Preissner, *Loops In Proteins (LIP) - a comprehensive loop database for homology modelling.* Protein Engineering, 2003. **16**(12): p. 979-985.

33. Greer, J., *Model for Haptoglobin Heavy-Chain Based Upon Structural Homology.* Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences, 1980. **77**(6): p. 3393-3397.

34. Kwasigroch, J.M., J. Chomilier, and J.P. Mornon, *A global taxonomy of loops in globular proteins (vol 259, pg 855, 1996).* Journal of Molecular Biology, 1996. **261**(5): p. 673-673.

35. Ring, C.S., et al., *Taxonomy and Conformational-Analysis of Loops in Proteins.* Journal of Molecular Biology, 1992. **224**(3): p. 685-699.

36. Tramontano, A. and A.M. Lesk, *Common Features of the Conformations of Antigen-Binding Loops in Immunoglobulins and Application to Modeling Loop Conformations.* Proteins-Structure Function and Genetics, 1992. **13**(3): p. 231-245.

37. Zhang, H.Y., et al., *A fast and efficient program for modeling protein loops.* Biopolymers, 1997. **41**(1): p. 61-72.

38. Cui, M., M. Mezei, and R. Osman, *Prediction of protein loop structures using a local move Monte Carlo approach and a grid-based force field.* Protein Engineering Design & Selection, 2008. **21**(12): p. 729-735.

39. de Bakker, P.I.W., et al., *Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field*

with the generalized born solvation model. Proteins-Structure Function and Genetics, 2003. **51**(1): p. 21-40.

40. Rapp, C.S. and R.A. Friesner, *Prediction of loop geometries using a generalized born model of solvation effects.* Proteins-Structure Function and Genetics, 1999. **35**(2): p. 173-183.

41. Spassov, V.Z., P.K. Flook, and L. Yan, *LOOPER: a molecular mechanics-based algorithm for protein loop prediction.* Protein Engineering Design & Selection, 2008. **21**(2): p. 91-100.

42. Jacobson, M.P., et al., *A hierarchical approach to all-atom protein loop prediction.* Proteins-Structure Function and Bioinformatics, 2004. **55**(2): p. 351-367.

43. Rapp, C.S., et al., *Prediction of protein loop geometries in solution.* Proteins-Structure Function and Bioinformatics, 2007. **69**(1): p. 69-74.

44. Zhu, K., et al., *Long loop prediction using the protein local optimization program.* Proteins-Structure Function and Bioinformatics, 2006. **65**(2): p. 438-452.

45. Felts, A.K., et al., *Prediction of protein loop conformations using the AGBNP implicit solvent model and torsion angle sampling.* Journal of Chemical Theory and Computation, 2008. **4**(5): p. 855-868.

46. Danielson, M.L. and M.A. Lill, *Predicting flexible loop regions that interact with ligands: The challenge of accurate scoring.* Proteins-Structure Function and Bioinformatics, 2012. **80**(1): p. 246-260.

47. Fogolari, F. and S.C.E. Tosatto, *Application of MM/PBSA colony free energy to loop decoy discrimination: Toward correlation between energy and root mean square deviation.* Protein Science, 2005. **14**(4): p. 889-901.

48. Sellers, B.D., et al., *Toward better refinement of comparative models: Predicting loops in inexact environments.* Proteins-Structure Function and Bioinformatics, 2008. **72**(3): p. 959-971.

49. Zhou, H.Y. and Y.Q. Zhou, *Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction.* Protein Science, 2002. **11**(11): p. 2714-2726.

50. Yang, Y.D. and Y.Q. Zhou, *Specific interactions for ab initio folding of protein terminal regions with secondary structures.* Proteins-Structure Function and Bioinformatics, 2008. **72**(2): p. 793-803.

51. Park, J. and K. Saitou, *ROTAS: a rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures.* Bmc Bioinformatics, 2014. **15**.

52. Shen, M.Y. and A. Sali, *Statistical potential for assessment and prediction of protein structures.* Protein Science, 2006. **15**(11): p. 2507-2524.

53. Rata, I.A., Y.H. Li, and E. Jakobsson, *Backbone Statistical Potential from Local Sequence-Structure Interactions in Protein Loops.* Journal of Physical Chemistry B, 2010. **114**(5): p. 1859-1869.

54. Liang, S.D., et al., *Protein Loop Modeling with Optimized Backbone Potential Functions.* Journal of Chemical Theory and Computation, 2012. **8**(5): p. 1820-1827.

55. Galaktionov, S., G.V. Nikiforovich, and G.R. Marshall, *Ab initio modeling of small, medium, and large loops in proteins.* Biopolymers, 2001. **60**(2): p. 153-168.

56. Burke, D.F. and C.M. Deane, *Improved protein loop prediction from sequence alone.* Protein Engineering, 2001. **14**(7): p. 473-478.

57. Fiser, A., R.K.G. Do, and A. Sali, *Modeling of loops in protein structures.* Protein Science, 2000. **9**(9): p. 1753-1773.

58. Schueler-Furman, O., et al., *Progress in modeling of protein structures and interactions.* Science, 2005. **310**(5748): p. 638-42.

59. Mandell, D.J., E.A. Coutsias, and T. Kortemme, *Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling.* Nature Methods, 2009. **6**(8): p. 551-552.

60. Rohl, C.A., et al., *Modeling structurally variable regions in homologous proteins with rosetta.* Proteins-Structure Function and Bioinformatics, 2004. **55**(3): p. 656-677.

61. Simons, K.T., et al., *Ab initio protein structure prediction of CASP III targets using ROSETTA.* Proteins-Structure Function and Bioinformatics, 1999: p. 171-176.

62. Xiang, Z.X., C.S. Soto, and B. Honig, *Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(11): p. 7432-7437.

63. Jamroz, M. and A. Kolinski, *Modeling of loops in proteins: a multi-method approach.* Bmc Structural Biology, 2010. **10**.

64. Mas, M.T., et al., *Modeling the Anti-Cea Antibody Combining Site by Homology and Conformational Search.* Proteins-Structure Function and Genetics, 1992. **14**(4): p. 483-498.

65. Martin, A.C.R., J.C. Cheetham, and A.R. Rees, *Modeling Antibody Hypervariable Loops - a Combined Algorithm.* Proceedings of the National Academy of Sciences of the United States of America, 1989. **86**(23): p. 9268-9272.

66. vanVlijmen, H.W.T. and M. Karplus, *PDB-based protein loop prediction: Parameters for selection and methods for optimization.* Journal of Molecular Biology, 1997. **267**(4): p. 975-1001.

67. Deane, C.M. and T.L. Blundell, *CODA: A combined algorithm for predicting the structurally variable regions of protein models.* Protein Science, 2001. **10**(3): p. 599-612.

68. Sudarsanam, S., et al., *Modeling Protein Loops Using a Phi-I+1, Psi-I Dimer Database.* Protein Science, 1995. **4**(7): p. 1412-1420.

69. Fernandez-Fuentes, N., J. Zhai, and A. Fiser, *ArchPRED: a template based loop structure prediction server.* Nucleic Acids Research, 2006. **34**: p. W173-W176.

70. Lyskov, S., et al., *Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE).* Plos One, 2013. **8**(5).

71. Hildebrand, P.W., et al., *SuperLooper-a prediction server for the modeling of loops in globular and membrane proteins.* Nucleic Acids Research, 2009. **37**: p. W571-W574.

72. Chys, P. and P. Chacon, *Random Coordinate Descent with Spinor-matrices and Geometric Filters for Efficient Loop Closure.* Journal of Chemical Theory and Computation, 2013. **9**(3): p. 1821-1829.

73. Lopez-Blanco, J.R., et al., *RCD plus : Fast loop modeling server.* Nucleic Acids Research, 2016. **44**(W1): p. W395-W400.

74. Choi, Y. and C.M. Deane, *FREAD revisited: Accurate loop structure prediction using a database search algorithm.* Proteins-Structure Function and Bioinformatics, 2010. **78**(6): p. 1431-1440.

75. Park, H., et al., *Protein Loop Modeling Using a New Hybrid Energy Function and Its Application to Modeling in Inaccurate Structural Environments.* Plos One, 2014. **9**(11).

76. Ko, J., et al., *The FALC-Loop web server for protein loop modeling.* Nucleic Acids Research, 2011. **39**: p. W210-W214.

77. Fiser, A. and A. Sali, *ModLoop: automated modeling of loops in protein structures.* Bioinformatics, 2003. **19**(18): p. 2500-2501.

78. Li, Y.H., I. Rata, and E. Jakobsson, *Sampling Multiple Scoring Functions Can Improve Protein Loop Structure Prediction Accuracy.* Journal of Chemical Information and Modeling, 2011. **51**(7): p. 1656-1666.

79. Liu, P., et al., *A Self-Organizing Algorithm for Modeling Protein Loops.* Plos Computational Biology, 2009. **5**(8).

80. Shehu, A., C. Clementi, and L.E. Kavraki, *Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations.* Proteins-Structure Function and Bioinformatics, 2006. **65**(1): p. 164-179.

81. Cortes, J., et al., *A path planning approach for computing large-amplitude motions of flexible molecules.* Bioinformatics, 2005. **21**: p. I116-I125.

82. Lee, A., I. Streinu, and O. Brock, *A methodology for efficiently sampling the conformation space of molecular structures.* Physical Biology, 2005. **2**(4): p. S108-S115.

83. Mamonova, T., et al., *Protein flexibility using constraints from molecular dynamics simulations.* Physical Biology, 2005. **2**(4): p. S137-S147.

84. Thorpe, M.F. and M. Lei, *Macromolecular flexibility.* Philosophical Magazine, 2004. **84**(13-16): p. 1323-1331.

85. Amato, N.M., K.A. Dill, and G. Song, *Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures.* Journal of Computational Biology, 2003. **10**(3-4): p. 239-255.

86. Apaydin, M.S., et al., *Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion.* Journal of Computational Biology, 2003. **10**(3-4): p. 257-281.

87. Lavalle, S.M., et al., *A randomized kinematics-based approach to pharmacophore-constrained conformational search and database screening.* Journal of Computational Chemistry, 2000. **21**(9): p. 731-747.

88. Abagyan, R. and M. Totrov, *Biased Probability Monte-Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins.* Journal of Molecular Biology, 1994. **235**(3): p. 983-1002.

89. Abagyan, R.A. and M.M. Totrov, *Biased probability Monte Carlo as a powerful global energy optimization method for biomolecular structure prediction.* Abstracts of Papers of the American Chemical Society, 1996. **211**: p. 35-COMP.

90. Zhu, K. and T. Day, *Ab initio structure prediction of the antibody hypervariable H3 loop.* Proteins-Structure Function and Bioinformatics, 2013. **81**(6): p. 1081-1089.

91. Bruccoleri, R.E. and M. Karplus, *Prediction of the Folding of Short Polypeptide Segments by Uniform Conformational Sampling.* Biopolymers, 1987. **26**(1): p. 137-168.

92. Shenkin, P.S., et al., *Predicting Antibody Hypervariable Loop Conformation .1. Ensembles of Random Conformations for Ring-Like Structures.* Biopolymers, 1987. **26**(12): p. 2053-2085.

93. Unger, R., *The genetic algorithm approach to protein structure prediction.* Applications of Evolutionary Computation in Chemistry, 2004. **110**: p. 153-175.

94. Ring, C.S. and F.E. Cohen, *Conformational Sampling of Loop Structures Using Genetic Algorithms.* Israel Journal of Chemistry, 1994. **34**(2): p. 245-252.

95. Collura, V., J. Higo, and J. Garnier, *Modeling of Protein Loops by Simulated Annealing.* Protein Science, 1993. **2**(9): p. 1502-1510.

96. Olson, M.A., M. Feig, and C.L. Brooks, *Prediction of protein loop conformations using multiscale Modeling methods with physical energy scoring functions.* Journal of Computational Chemistry, 2008. **29**(5): p. 820-831.

97. Wu, M.G. and M.W. Deem, *Efficient Monte Carlo methods for cyclic peptides.* Molecular Physics, 1999. **97**(4): p. 559-580.

98. Shukla, D., et al., *Activation pathway of Src kinase reveals intermediate states as targets for drug design.* Nature Communications, 2014. **5**.

99. Raval, A., et al., *Refinement of protein structure homology models via long, all-atom molecular dynamics simulations.* Proteins-Structure Function and Bioinformatics, 2012. **80**(8): p. 2071-2079.

100. Coutsias, E.A., et al., *A kinematic view of loop closure.* Journal of Computational Chemistry, 2004. **25**(4): p. 510-528.

101. Dinner, A.R., *Local deformations of polymers with nonplanar rigid main-chain internal coordinates.* Journal of Computational Chemistry, 2000. **21**(13): p. 1132-1144.

102. Hayward, S. and A. Kitao, *Monte Carlo Sampling with Linear Inverse Kinematics for Simulation of Protein Flexible Regions.* Journal of Chemical Theory and Computation, 2015. **11**(8): p. 3895-3905.

103. Hoffmann, D. and E.W. Knapp, *Polypeptide folding with off-lattice Monte Carlo dynamics: The method.* European Biophysics Journal with Biophysics Letters, 1996. **24**(6): p. 387-403.

104. Nilmeier, J., et al., *Assessing Protein Loop Flexibility by Hierarchical Monte Carlo Sampling.* Journal of Chemical Theory and Computation, 2011. **7**(5): p. 1564-1574.

105. Kaufmann, K.W., et al., *Practically Useful: What the ROSETTA Protein Modeling Suite Can Do for You.* Biochemistry, 2010. **49**(14): p. 2987-2998.

106. Friesner, R.A., et al., *Computational methods for high resolution prediction and refinement of protein structures.* Current Opinion in Structural Biology, 2013. **23**(2): p. 177-184.

107. Li, Y., *Conformational sampling in template-free protein loop structure modeling: an overview.* Comput Struct Biotechnol J, 2013. **5**: p. e201302003.

108. Shehu, A. and L.E. Kavraki, *Modeling Structures and Motions of Loops in Protein Molecules.* Entropy, 2012. **14**(2): p. 252-290.

109. Soto, C.S., et al., *Loop modeling: Sampling, filtering, and scoring.* Proteins-Structure Function and Bioinformatics, 2008. **70**(3): p. 834-843.

110. Vihinen, M., *Relationship of Protein Flexibility to Thermostability.* Protein Engineering, 1987. **1**(6): p. 477-480.

111. Karplus, P.A. and G.E. Schulz, *Prediction of Chain Flexibility in Proteins - a Tool for the Selection of Peptide Antigens.* Naturwissenschaften, 1985. **72**(4): p. 212-213.

112. Radivojac, P., et al., *Protein flexibility and intrinsic disorder.* Protein Science, 2004. **13**(1): p. 71-80.

113. Morris, P., et al., *Impact of cofactor-binding loop mutations on thermotolerance and activity of E-coli transketolase.* Enzyme and Microbial Technology, 2016. **89**: p. 85-91.

114. Boehr, D.D., R. Nussinov, and P.E. Wright, *The role of dynamic conformational ensembles in biomolecular recognition.* Nature Chemical Biology, 2009. **5**(11): p. 789-796.

115. Kumar, S., et al., *Folding and binding cascades: Dynamic landscapes and population shifts.* Protein Science, 2000. **9**(1): p. 10-19.

116. Ma, B.Y., et al., *Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations.* Protein Science, 2002. **11**(2): p. 184-197.

117. Tsai, C.J., et al., *Folding funnels, binding funnels, and protein function.* Protein Science, 1999. **8**(6): p. 1181-1190.

118. Monod, J., J. Wyman, and J.P. Changeux, *On the Nature of Allosteric Transitions: A Plausible Model.* J Mol Biol, 1965. **12**: p. 88-118.

119. Zheng, Z., M.N. Ucisik, and K.M. Merz, *The Movable Type Method Applied to Protein-Ligand Binding.* Journal of Chemical Theory and Computation, 2013. **9**(12): p. 5526-5538.

120. Chu, V., et al., *Thermodynamic and structural consequences of flexible loop deletion by circular permutation in the streptavidin-biotin system.* Protein Science, 1998. **7**(4): p. 848-859.

121. Freitag, S., et al., *Structural studies of the streptavidin binding loop.* Protein Science, 1997. **6**(6): p. 1157-1166.

122. Le Trong, I., et al., *Streptavidin and its biotin complex at atomic resolution.* Acta Crystallographica Section D-Biological Crystallography, 2011. **67**: p. 813-821.

123. Stayton, P.S., et al., *Streptavidin-biotin binding energetics.* Biomolecular Engineering, 1999. **16**(1-4): p. 39-44.

124. Weber, P.C., et al., *Structural Origins of High-Affinity Biotin Binding to Streptavidin.* Science, 1989. **243**(4887): p. 85-88.

125. Izrailev, S., et al., *Molecular dynamics study of unbinding of the avidin-biotin complex.* Biophysical Journal, 1997. **72**(4): p. 1568-1581.

126. General, I.J. and H. Meirovitch, *Relative stability of the open and closed conformations of the active site loop of streptavidin.* Journal of Chemical Physics, 2011. **134**(2).

127. Miyamoto, S. and P.A. Kollman, *Absolute and Relative Binding Free-Energy Calculations of the Interaction of Biotin and Its Analogs with Streptavidin Using Molecular-Dynamics Free-Energy Perturbation Approaches.* Proteins-Structure Function and Genetics, 1993. **16**(3): p. 226-245.

128. O'Sullivan, V.J., et al., *Development of a Tetrameric Streptavidin Mutein with Reversible Biotin Binding Capability: Engineering a Mobile Loop as an Exit Door for Biotin.* Plos One, 2012. **7**(4).

129. Song, J.N., et al., *Functional Loop Dynamics of the Streptavidin-Biotin Complex.* Scientific Reports, 2015. **5**.

130. Hendrickson, W.A., et al., *Crystal-Structure of Core Streptavidin Determined from Multiwavelength Anomalous Diffraction of Synchrotron Radiation.* Proceedings of the National Academy of Sciences of the United States of America, 1989. **86**(7): p. 2190-2194.

131. Wu, S.C. and S.L. Wong, *Engineering soluble monomeric streptavidin with reversible biotin binding capability.* Journal of Biological Chemistry, 2005. **280**(24): p. 23225-23231.

132. Roe, D.R. and T.E. Cheatham, *PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data.* Journal of Chemical Theory and Computation, 2013. **9**(7): p. 3084-3095.

133. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics & Modelling, 1996. **14**(1): p. 33-38.

134. Lindorff-Larsen, K., et al., *Improved side-chain torsion potentials for the Amber ff99SB protein force field.* Proteins-Structure Function and Bioinformatics, 2010. **78**(8): p. 1950-1958.

135. Maier, J.A., et al., *ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB.* J Chem Theory Comput, 2015. **11**(8): p. 3696-713.

136. Kumar, S., et al., *The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .1. The Method.* Journal of Computational Chemistry, 1992. **13**(8): p. 1011-1021.

137. Grossfield, A., *WHAM:the weighted histogram analysis method*: http://membrane.urmc.rochester.edu/content/wham.

138. Chakravorty, D.K., et al., *Metal Ion Capture Mechanism of a Copper Metallochaperone.* Biochemistry, 2016. **55**(3): p. 501-509.

139. Howarth, M., et al., *A monovalent streptavidin with a single femtomolar biotin binding site.* Nature Methods, 2006. **3**(4): p. 267-273.

140. Chaiet, L. and F.J. Wolf, *The Properties of Streptavidin, a Biotin-Binding Protein Produced by Streptomycetes.* Arch Biochem Biophys, 1964. **106**: p. 1-5.

141. Weber, P.C., et al., *Crystallographic and Thermodynamic Comparison of Natural and Synthetic Ligands Bound to Streptavidin.* Journal of the American Chemical Society, 1992. **114**(9): p. 3197-3200.

142. Le Trong, I., et al., *Structural consequences of cutting a binding loop: two circularly permuted variants of streptavidin.* Acta Crystallographica Section D-Biological Crystallography, 2013. **69**: p. 968-977.

143. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nature Reviews Molecular Cell Biology, 2005. **6**(3): p. 197-208.

144. van der Lee, R., et al., *Classification of Intrinsically Disordered Regions and Proteins.* Chemical Reviews, 2014. **114**(13): p. 6589-6631.

145. Wright, P.E. and H.J. Dyson, *Intrinsically disordered proteins in cellular signalling and regulation.* Nature Reviews Molecular Cell Biology, 2015. **16**(1): p. 18-29.

146. D. A. Case, V.B., J. T. Berryman, et al., *Amber 14.* 2014.

147. Darden, T., D. York, and L. Pedersen, *Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems.* Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.

148. Essmann, U., et al., *A Smooth Particle Mesh Ewald Method.* Journal of Chemical Physics, 1995. **103**(19): p. 8577-8593.

149. Toukmaji, A., et al., *Efficient particle-mesh Ewald based approach to fixed and induced dipolar interactions.* Journal of Chemical Physics, 2000. **113**(24): p. 10913-10927.

150. Jorgensen, W.L., et al., *Comparison of Simple Potential Functions for Simulating Liquid Water.* Journal of Chemical Physics, 1983. **79**(2): p. 926-935.

151. Li, P.F., L.F. Song, and K.M. Merz, *Parameterization of Highly Charged Metal Ions Using the 12-6-4 LJ-Type Nonbonded Model in Explicit Water.* Journal of Physical Chemistry B, 2015. **119**(3): p. 883-895.

152. Wang, J.M., et al., *Development and testing of a general amber force field.* Journal of Computational Chemistry, 2004. **25**(9): p. 1157-1174.

153. Bour, P. and T.A. Keiderling, *Ab-Initio Simulations of the Vibrational Circular-Dichroism of Coupled Peptides.* Journal of the American Chemical Society, 1993. **115**(21): p. 9602-9607.

CHAPTER 5

Conclusions and Future Outlook

The objective of this dissertation is to contribute to the field of structure based drug design by introduction of a novel conformational sampling method for receptors in the binding pocket region. Proteins are inherently plastic and undergo a variety of motions that range from ultrafast vibrations to long-range backbone motions. In order to better understand the protein-ligand binding interactions, it is crucial to capture the flexibility of proteins in our current existing computational models. However, incorporating it into even the smallest region, for example, the binding site of a protein, is computationally demanding. Even if this task can be accomplished there is a risk of running into false positives due to the enormous conformational space involved. Crystal structure represents only one of the conformations from the ensemble and can be hardly considered as a good substitute for an entire conformational space available to proteins.

The first chapter of this thesis addresses computational approaches that have been implemented to incorporate receptor flexibility into the ligand-binding domain of a protein. The introduction is not limited to receptor flexibility methods used in molecular docking studies. Prospective validation of several docking and other comprehensive tools has also been discussed largely within the scope of blind challenges conducted by the D3R and CSAR organizations. Based on our analysis to date, we conclude that the community is moving forward by fine-tuning several computational approaches but that the statistical uncertainties in the sampling and scoring accuracy still remain significant.

In this regard, we have developed a method that can introduce receptor flexibility in the binding pocket region of the protein in a novel fashion. Chapter 2 entails the algorithm and the method development section. The algorithm is quite simple and can be succinctly described here. We treat a molecule at an atom pair level and use a distance based coordinate system. Each

selected distance is associated with pair potential look up tables for each atom pairwise interaction. These look up tables help us in generating structures on an energy scale in a very efficient manner. There are two main motions that occur in the protein-ligand binding motions-side chains and backbone motions. We included receptor flexibility in terms of both. Once the conformational ensemble is generated on an energy scale, the free energies are calculated by using Movable Type free energy method[1].

We initially applied this strategy to treat receptor flexibility in terms of side chains, which we are calling $MT_{flex[2]}$. MT_{flex} was successfully applied on a set of 159 protein-ligand systems derived from core PDBBINDv2014 dataset after excluding proteins with metal ions in the binding pocket area[3, 4]. This application is discussed in details in Chapter 3 of this dissertation. The conformational states generated by MT_{flex} were employed in subsequent docking and scoring exercise. We performed Glide docking on the generated conformational ensemble and also on the crystal structure. It was observed that Glide docking generated better binding modes in terms of both structural RMSDs (Å) and scores when docking was performed by using MT_{flex} generated conformations as compared to docking in crystal structure. Apart from Glide scoring, we also used our in-house Movable type free energy method and validated that the binding affinity showed better correlation with respect to experimental results when we employed MT_{flex} generated ensemble as compared to the score of using crystal docked structures. Overall, we showed that by including side chain flexible multiple receptor structures, docking and binding affinity measures improve as compared to including only single structure.

After including side chain receptor flexibility, we next applied our method to include backbone flexibility as well. The extension to include backbone flexibility was quite straightforward, yet computationally expensive. We are calling this strategy, $MT_{Flex\text{-}b}$. To take

care of the computational expense, we used Jacobsen's strategy of generating loops from both the terminals and meeting in the middle[5]. We successfully applied our strategy to study the $loop_{3\text{-}4}$ of the Streptavidin-Biotin system. This highly mobile, eight residue long loop undergoes upon to close transition in going from Biotin un-bound (*apo*) to bound (*holo*) state[6-9]. It has been discussed in Chapter 4 of this thesis. We generated a huge conformational ensemble of the $loop_{3\text{-}4}$ region using $MT_{Flex\text{-}b}$ with over ~11 million conformations. The ensemble contained both closed-like and open-like loop conformations with a best structural RMSD of ~1.6 Å with respect to experiment. We obtained a free energy surface (MT-FES) for both the *apo* and *holo* states and observed that the open loop is ~10.5 kcal/mol more stable in the *apo* state. In the *holo* state, the trend was reversed and it was observed that the closed loop is ~12 kcal/mol more stable than the open loop. We observed that the stability of the open loop in the *apo* state is attributed to solvation free energy, while in the *holo* state, the protein-ligand contacts are much more favorable towards the closed state, thereby compensating for the solvation free energy. We validated our MT-FES by performing MD-PMF simulations using FF99SBILDN and FF14SB force fields. A good correlation between MT-FES and MD-PMF using FF14SB simulations was observed. We also generated a thermodynamic free energy cycle and obtained a binding free energy change of –16.2 kcal/mol, which is in reasonable agreement with the experimental binding affinity of –18.3 kcal/mol. With the help of fast and efficient thermodynamic cycle and free energy surfaces, we hope that we are introducing a tool that will hugely impact the field of structure based drug design.

The ultimate goal of our study is to not only generate a conformational ensemble but to employ it in understanding the deeper knowledge of protein-ligand interactions. We hope to expand our ability to accurately predict free energies, which will hugely impact the growth of

structure based drug design[10-12]. This understanding of side chain and flap motions will help us in understanding its affects on substrate or ligand binding, which will help in the advancement of structure-based drug design efforts. This strategy can be extended to systems like Kinases, ureases, HSP90, HIV protease, *etc*. This understanding will help the community in designing better inhibitors that have desired functions.

REFERENCES

REFERENCES

1. Zheng, Z., M.N. Ucisik, and K.M. Merz, *The Movable Type Method Applied to Protein-Ligand Binding.* Journal of Chemical Theory and Computation, 2013. **9**(12): p. 5526-5538.

2. Bansal, N., Z. Zheng, and K.M. Merz, *Incorporation of side chain flexibility into protein binding pockets using MTflex.* Bioorganic & Medicinal Chemistry, 2016. **24**(20): p. 4978-4987.

3. Wang, R., et al., *The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures.* J Med Chem, 2004. **47**(12): p. 2977-80.

4. Wang, R., et al., *The PDBbind database: methodologies and updates.* J Med Chem, 2005. **48**(12): p. 4111-9.

5. Jacobson, M.P., et al., *A hierarchical approach to all-atom protein loop prediction.* Proteins-Structure Function and Bioinformatics, 2004. **55**(2): p. 351-367.

6. Chu, V., et al., *Thermodynamic and structural consequences of flexible loop deletion by circular permutation in the streptavidin-biotin system.* Protein Science, 1998. **7**(4): p. 848-859.

7. Freitag, S., et al., *Structural studies of the streptavidin binding loop.* Protein Science, 1997. **6**(6): p. 1157-1166.

8. Le Trong, I., et al., *Streptavidin and its biotin complex at atomic resolution.* Acta Crystallographica Section D-Biological Crystallography, 2011. **67**: p. 813-821.

9. Stayton, P.S., et al., *Streptavidin-biotin binding energetics.* Biomolecular Engineering, 1999. **16**(1-4): p. 39-44.

10. Gilson, M.K. and H.X. Zhou, *Calculation of protein-ligand binding affinities.* Annual Review of Biophysics and Biomolecular Structure, 2007. **36**: p. 21-42.

11. Michel, J. and J.W. Essex, *Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations.* Journal of Computer-Aided Molecular Design, 2010. **24**(8): p. 639-658.

12. Shirts, M.R., D.L. Mobley, and J.D. Chodera, *Alchemical Free Energy Calculations: Ready for Prime Time?* Annual Reports in Computational Chemistry, Vol 3, 2007. **3**: p. 41-59.