Freie Universität Berlin

Conformational Dynamics of the Multivalent Targets

Inaugural-Dissertation to obtain the academic degree *Doctor rerum naturalium* (Dr. rer. nat.) submitted to the Department of Biology, Chemistry and Pharmacy of Freie Universität Berlin

by

M.Sc. Stevan Aleksić

from Novi Sad (Serbia)

YEAR OF SUBMISSION 2018

1st Reviewer:
Prof. Dr. Bettina Keller
Department of Biology, Chemistry and Pharmacy
Freie Universität Berlin

2nd Reviewer:
Prof. Dr. Beate Paulus
Department of Biology, Chemistry and Pharmacy
Freie Universität Berlin

Date of defence: 16.10.2018.

Acknowledgements

Foremost, I would like to express my sincere gratitude to Prof. Bettina Keller for an opportunity to conduct doctoral research, her generous support and mentorship. It was a great pleasure to be one among the first Ph.D. students of Prof. Keller, and I am convinced that many exciting scientific contributions will emerge from her growing working group.

Moreover, I would like to acknowledge Prof. Beate Paulus as my second supervisor, for the time reviewing this thesis, and proving a stimulating scientific and working environment for the theory enthusiasts at Freie Universität.

Furthermore, I would like to thank the team of Sonderforschungsbereich 765 "Multivalenz als chemisches Organisations- und Wirkprinzip" for providing an excellent platform for scientific collaborations. As a member of SFB765 I was fortune to collaborate with numerous scientists on three intriguing projects. First, I would like to thank to Dr. Jonas Hanske and Dr. Christoph Rademacher for all the experimental work and stimulating discussions on Langerin allostery project, which yielded a JACS publication. Next, I would like to thank to a big team still preparing a publication on the sialosides project, consisting of Pallavi Kiran, Dr. Sumati Bhatia, Prof. Rainer Haag, Dr. Susanne Liese, Prof. Dr. Roland Netz, Dr. Daniel Lauster, and Prof. Andreas Hermann. Finally, I would like to acknowledge Miriam Bertazzon, Dr. Jana Sticht, and Prof. Christian Freund for the experimental insights, ongoing discussion and a publication preparation regarding the WW domains project.

It was exciting four years of sharing the ups and downs of Ph.D. life with other Ph.D. fellows of theory groups at Freie Universität. My special thanks go to all Biomolecular Dynamics coworkers for an astonishing team spirit, not necessarily only science related.

Finally, I would like to express my greatest gratitude to my father and all my closest friends for the enormous support along this "turbulent" journey deeper in the world of Computational Chemistry.

List of Publications:

- Jonas Hanske*, Stevan Aleksić*, Martin Ballaschk, Marcel Jurk, Elena Shanina, Monika Beerbaum, Peter Schmieder, Bettina G. Keller, and Christoph Rademacher. "Intradomain Allosteric Network Modulates Calcium Affinity of the C-Type Lectin Receptor Langerin". In: Journal of the American Chemical Society 138.37 (2016), pp. 12176–12186. doi:10.1021/jacs.6b05458.

*These authors contributed equally to the publication.

Contribution to the publication:

A detailed description of my contribution to this publication is provided in a preface to **Chapter 3**.

Summary

Multivalency is a ubiquitous mechanism in nature involved in numerous biological processes such as recognition, adhesion, self-organization of matter and signal transduction. Multivalency can be defined as the binding of an n-valent ligand to an m-valent receptor through non-covalent, strong, but reversible interactions. Recently, multivalency has been applied as a principle to design novel molecules with a potential to fight the different microbes. Therefore, the majority of the experimental and theoretical framework has been developed in the areas facilitating the ligand design, while its respective receptor is usually assumed to be somewhat rigid. However, this assumption does not always hold. Thus the primary focus of this thesis was to investigate the conformational behavior of three multivalent systems with computational methods and to correlate these findings with the outcomes of the complementary experiments.

First, we investigated the carbohydrate uptake and release by a trivalent C-type lectin receptor Langerin, since the little was known about this mechanism. The carbohydrate recognition is dependent on the Ca^{2+} cofactor. We demonstrated that the Ca^{2+} binding to Langerin is pH-sensitive and under control of a robust allosteric network. Additionally, we showed that the conformational dynamics of Langerin comprised several events occurring at different timescales.

Then, I focused on elucidating the conformational dynamics and its influence on the design and the potency of the trivalent sialosides to inhibit a viral protein Hemagglutinin.

Last, I explored a bivalent recognition process on the example of a proline-rich peptide SmB_2 and tandem-WW domains of a spliceosomal Formin Binding Protein 21. In this study, I reported a highly complex conformational dynamics of the apo receptor, shed light on its respective free energy landscape, proposed a scheme for determining a binding-competent structure and modeled a binding complex.

Zusammenfassung

Multivalenz ist ein allgegenwärtiger Mechanismus in der Natur, der in eine Vielzahl biologischer Prozesse wie zur Erkennung, Adhäsion, Selbstorganisation von Materie und Signalübertragung involviert ist. Multivalenz kann als Bindung eines n-valenten Liganden an einen m-valenten Rezeptor durch nichtkovalente, starke, aber reversible Wechselwirkungen definiert werden. Unlängst wurde Multivalenz als Prinzip für das Design neuartiger Moleküle mit Potential zur Bekämpfung verschiedener Mikroben angewandt. Daher wurde ein Großteil des experimentellen und theoretischen Grundgerüsts in Bereichen entwickelt, die das Ligandendesign unterstützen, während der entsprechende Rezeptor gewöhnlich als eher unbeweglich betrachtet wird. Allerdings ist diese Annahme nicht immer berechtigt. Deshalb liegt der primäre Fokus dieser Arbeit auf der Untersuchung des konformationellen Verhaltens dreier multivalenter Systeme mit computerunterstützten Methoden und der Korrelation der Resultate mit den Ergebnissen ergänzender Experimente.

Zunächst untersuchten wir die Kohlenhydrataufnahme und -freisetzung des trivalenten C-typ Lectinrezeptors Langerin, da nur wenig über diesen Mechanismus bekannt ist. Die Kohlenhydraterkennung ist auf einen Ca^{2+}-Kofaktor angewiesen. Wir wiesen nach, dass die Ca^{2+}-Bindung an Langerin pH-sensitiv ist und unter Kontrolle eines robusten allosterischen Netzwerks steht. Außerdem zeigten wir, dass die Konformationsdynamik von Langerin mehrere Vorgänge auf unterschiedlichen Zeitskalen umfasst.

Weiterhin konzentrierte ich mich auf die Aufklärung der Konformationsdynamik und deren Einfluss auf Design und Wirksamkeit trivalenter Sialoside das virale Protein Hämagglutinin zu hemmen.

Zuletzt untersuchte ich einen bivalenten Erkennungsprozess am Beispiel des prolinreichen Peptids SmB_2 und Tandem-WW-Domänen des spliceosomalen Formin-bindenden Proteins 21. In dieser Studie berichtete ich von der hochkomplexen Konformationsdynamik des Aporezeptors, klärte die entsprechende Freie Energie-Landschaft auf, schlug ein Schema zur Bestimmung einer bindungskompetenten Struktur vor und modellierte einen bindenden Komplex.

Contents

Chapter 1

Introduction

1.1 Multivalency

Multivalency is an important mechanism in nature that comprises the non-covalent, strong and reversible binding between m-valent ligands and n-valent receptors (where m,n >1) [1]. Examples of multivalency include attachment of pathogens to the host cells through carbohydrate-protein interactions [2], cell-cell adhesion, multivalent DNA-protein interactions essential for the gene expression, and multivalent protein-protein interactions that are the key element in the antibody-mediated immunological responses [3].

Recently an idea of fighting the nature with its own weapons arose through a concept of a multivalent ligand design. In such ligands, the carbohydrate moieties are assembled into the supramolecular structures by the spacers or interfaces of the diverse origins. To quantify the increase in the binding affinity Mammen and coworkers introduced the enhancement factor

$$\beta = \frac{K_{d,multi}}{K_{d,mono}} \tag{1.1}$$

where $K_{d,multi}$ and $K_{d,mono}$ are the dissociation constants of the multivalent ligand and its monovalent counterpart respectively [4]. An advantage of this enhancement factor β is that it can be used when the exact number of the ligand-receptor entities is unknown a priori, which is often the case in biological systems. However, this affinity measure cannot distinguish multivalency from the cooperativity and symmetry effects [5].

The interaction of the n monovalent ligands with an m-valent receptor is not considered as multivalent. Nonetheless, the multivalent architecture of the receptor itself increases the probability of the monovalent ligand to bind more often. Similarly, this symmetry effect is incorporated in the multivalency. Cooperativity

effect describes how the binding of one ligand changes the affinity of the multivalent receptor to support further binding. Cooperativity is said to be neutral when all ligands have the same affinity for the given multivalent receptor (noncooperative effect). Then, positive cooperativity implies that affinity for the first ligand is the lowest compared to all other ligands of the binding series. Finally, when the binding of the first ligand decreases the probability of the subsequent binding events, one assumes negative cooperativity. Multivalency is often confused with the positive (synergetic) cooperativity. Furthermore, the negative cooperativity effect was reported for the most multivalent carbohydrate-lectin interactions and multivalent molecular machines. [6, 7].

Affinity enhancement can be explained regarding the favorable thermodynamics or kinetics of the multivalent binding event when compared to the monovalent case. Sometimes, the affinity enhancement can be attributed to a rebinding effect. The increased concentration of the ligand in the proximity of the receptor causes the rebinding events, which in turn govern the system to a state where all ligands are bound.

Thermodynamics model of a monovalent ligand binding to a multivalent receptor still can be represented by the binding free energy (ΔG) as following:

$$\Delta G = \Delta H - T\Delta S. \tag{1.2}$$

The enthalpic component of binding ΔH accounts for all the non-covalent interactions formed between the ligand and the binding site, whereas entropic component represents the number of degrees of freedom lost upon binding (considering ligand and receptor combined). Further, the entropic component ΔS can be realized as the sum of the translational (ΔS_{trans}), rotational (ΔS_{rot}) and conformational entropy (ΔS_{conf}) of the ligand, and the desolvation effects occurring at the receptor (ΔS_{solv}):

$$\Delta S = \Delta S_{trans} + \Delta S_{rot} + \Delta S_{conf} + \Delta S_{solv}. \tag{1.3}$$

The translation entropy represents the freedom of a molecule to independently move through space. It is directly dependent on the logarithm of the mass of a molecule ($\Delta S_{trans} \propto \log(\mathrm{M})$), and inversely proportional to the logarithm of its concentration ($\Delta S_{trans} \propto \log([L])^{-1}$). The rotational entropy represents the possibility of a molecule to rotate around all three Cartesian axes and it proportional to the product of three momenta of inertia ($\Delta S_{rot} \propto \log(I_x I_y I_z)$). Therefore, those two entropy contributions are weakly dependent on the mass and dimensions of a molecule. If it is assumed that $\Delta \mathrm{S}_{trans}$ and $\Delta \mathrm{S}_{rot}$ are equal for the ligand,

the receptor and the ligand-receptor species, then, upon ligand-receptor complex formation, three translational and three rotational degrees of freedom are lost. If the difference in the masses of the ligand and receptor is neglected then the total translational and rotational entropic cost of the ligand-receptor complex formation (independent of the mono-/multivalent nature) is approximately the same given that both molecules are at the same concentration in solution.

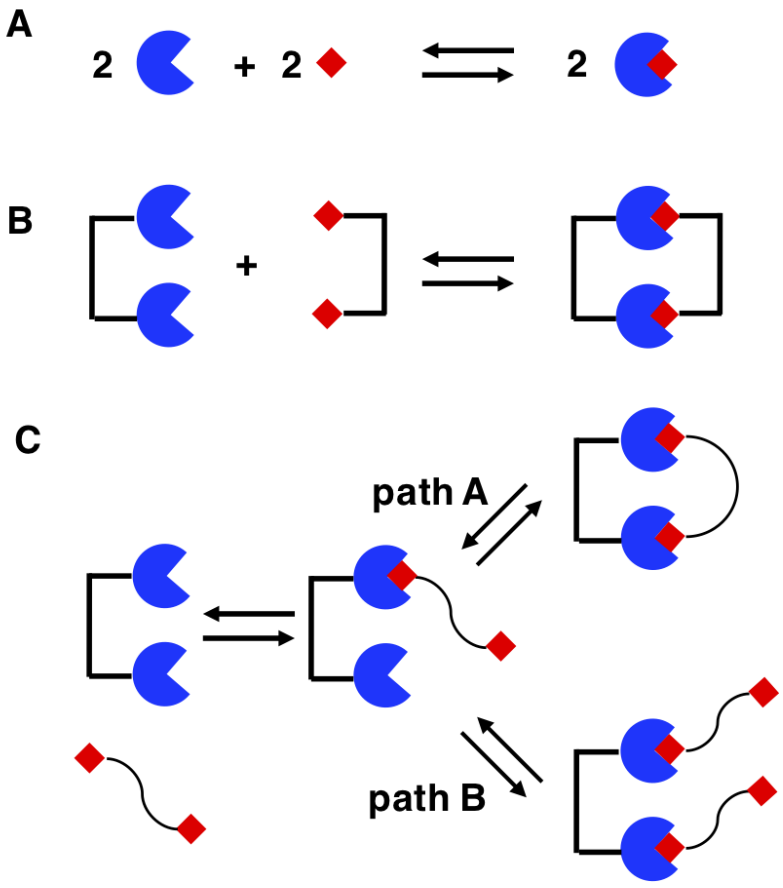


FIGURE 1.1: Thermodynamic models of mono/bivalent binding: (A) Binding of a rigid monovalent ligand to a rigid monovalent receptor; (B) Binding of a rigid bivalent ligand to a rigid bivalent receptor; (C) Binding of a flexible bivalent ligand to a somewhat flexible bivalent receptor.

Consider binding of two monovalent ligands to two monovalent receptors. If both binding partner are absolutely rigid resulting in the conformational entropy equals to zero and the total entropic cost is (Figure 1.1A):

$$\Delta S = 2\Delta S_{trans}^{mono} + 2\Delta S_{rot}^{mono}. \tag{1.4}$$

When a rigid bivalent ligand associates with a rigid bivalent receptor (ΔS_{conf}=*0*, once the first ligand-complex is formed, there is no further translational and

rotational cost involved in the subsequent ligand-receptor pair formation (Figure 1.1B):

$$\Delta S = \Delta S_{trans}^{mono} + \Delta S_{rot}^{mono}. \tag{1.5}$$

Hence, this type of multivalent binding is thought to be entropically enhanced. However, this binding scenario is rather unrealistic, as multivalent ligands and receptors pose the certain degree of flexibility encoded in their nature. For the realistic systems, the conformational entropy ΔS_{conf} has almost always negative value due to the reduction of conformational spaces of both binding partners. Now assume that the binding event occurs between a pair of somewhat flexible bivalent ligand and bivalent receptor (Figure 1.1C), and the flexibility of a bivalent receptor is restricted to two binding sites. If binding event follows a path A and:

$$\Delta S_{conf}^{mono} < \Delta S_{trans}^{mono} + \Delta S_{rot}^{mono} \tag{1.6}$$

then the binding in such a bivalent system is still entropically enhanced compared to the monovalent case. This is the favourable scenario in the design of the multivalent ligands. When there is an equal probability of the bivalent system taking both binding paths, then, the binding of the bivalent ligand to the bivalent receptor is entropically neutral and that is given as follows:

$$\Delta S_{conf}^{mono} = \Delta S_{trans}^{mono} + \Delta S_{rot}^{mono}. \tag{1.7}$$

Improper design of the bivalent ligand can result in no affinity gain compared to its monovalent counterpart, as the binding is said to be entropically diminished, since:

$$\Delta S_{conf}^{mono} > \Delta S_{trans}^{mono} + \Delta S_{rot}^{mono} \tag{1.8}$$

and only a binding path B is possible. From a thermodynamical perspective, designing a multivalent ligand with spacer of modest flexibility proved to be the best approach in achieving affinity enhancement [1, 4].

Kinetic studies of the multivalent binding revealed that the association constant k_{on} is not significantly dependent on the multivalent nature of the system. However, the dissociation rate constant k_{off} of a multivalent ligand is reduced compared to its monovalent counterpart. To achieve the full dissociation, N ligand-receptor interactions should be broken. Thus, the slower k_{off} rate contributes to the affinity enhancement. [5, 8].

Rao *et al.* reported that high affinity of trivalent a vancomycin-based ligand had a kinetic origin due to the strong rebinding effect [9]. It was proved that

thermodynamics representation for such systems failed to explain the rebinding effect, and consequently, this phenomenon should be addressed kinetically. In a typical kinetic model, a ligand can assume three states: an unbound, a transition, and a bound state. The ligand should overcome a transition energy barrier to form an initial complex with the receptor. From the transition complex, the system progresses fast to the fully bound state. Nonetheless, this kinetic model is not applicable for studying the rebinding effect due to the low activation barrier of the transition complex, which increases the probability of binding. The local concentration of the ligand also contributes to the improved k_{on} rates. Weber *et. al* [10] proposed an additional almost bound state to account for the rebinding effect and discussed the following cases:

- **strong ligand with activation barrier** independently of the spacer presence is always totally bound or totally unbound. The almost bound state is the kinetically unstable entity and immediately transits to the totally bound state (Figure 1.2A).
- **weak ligand without activation barrier** also lacks the almost bound state, and the singly bound state is absent when ligands are connected by a spacer (Figure 1.2B).
- **very weak ligand without activation barrier** is a case where the presence of a spacer strongly influences the existence of the respective kinetic entities. Spacer-free ligands lack the totally bound state due to the low affinity. In contrast, the almost bound state is kinetically stable. The presence of a spacer shifts the equilibrium to the totally bound state, as the dissociation occurs at the slower rate allowing the rebinding effect to happen (Figure 1.2C).

Principles of multivalency have been used to design new antibacterial and antiviral agents. A potent multivalent drug should shield the pathogen surface and lead to pathogen removal from the body. Alternatively, a multivalent ligand can interact with the receptor on the cell membrane and prevent the pathogen from entering the cytoplasm causing infection. Since multivalent agent binds to the multiple binding sites simultaneously, the possibility for the resistance development is highly reduced. Multivalent glycoconjugates have been employed in the development of vaccines to fight the various pathogen. Broader advancement in this area is held by the obstacles of detecting the potent carbohydrate antigens or poor antibody responses to them. Efficacy, bioavailability, and toxicity of the multivalent ligands are still marginally addressed in the literature [1, 5, 11, 12].

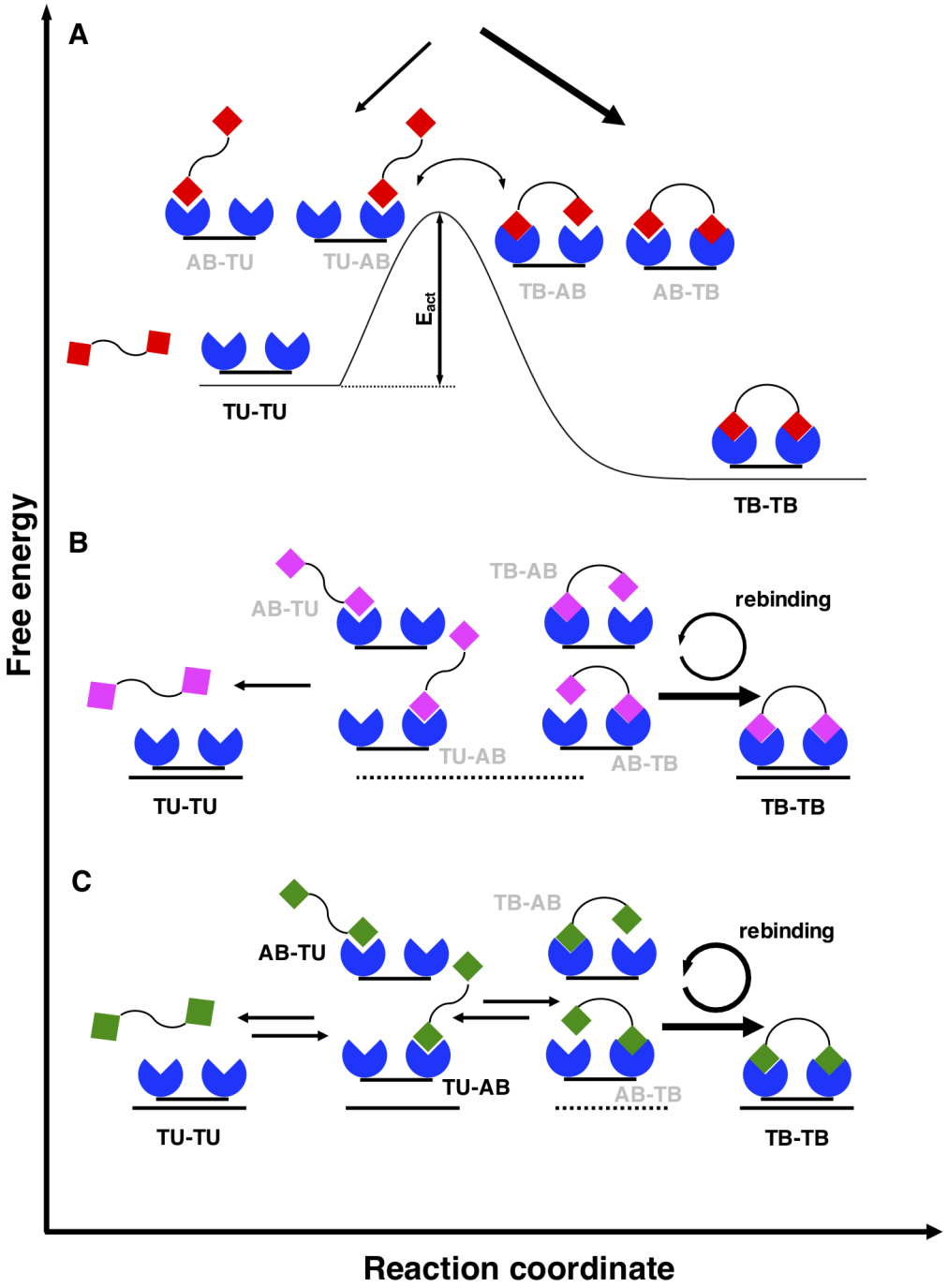


FIGURE 1.2: Kinetic models of a multivalent binding: (A) Strong ligand with activation barrier (ligand in red); (B) Weak ligand without activation barrier (ligand in magenta); (C) Very weak ligand without activation barrier (ligand in green). Unstable kinetic species are indicated in gray letters.

1.2 Allostery

Much progress has been reported recently on fine-tuning the affinity of multivalent ligands. On the contrary, dynamics of multivalent receptors have rarely

been studied, assuming rigid body representation of such a receptor. However, the effect of the ligand binding on the population shifts in the receptor's conformational ensemble cannot be neglected. In a classic example of the oxygen transport by the oligomeric protein hemoglobin, the binding of the first oxygen molecule introduces the conformational changes of the receptor, which are then associated with the positive cooperativity effect allowing further three oxygen molecules to bind with increased affinity [13]. Although this type of binding is not multivalent per se, it triggers the question if the allostery may play a role in actual multivalent systems.

The term allostery was coined in 1961 by Monod and Jacob to describe the inhibition of the enzyme L-threonine deaminase by its end-product L-isoleucine. The authors argued that L-isoleucine binds to the different (allosteric) site then the substrate L-threonine causing inhibition [14]. As early as in 1935, Pauling proposed a model to account for the positive cooperativity in hemoglobin. In the mid-sixties, the hemoglobin case was revisited and the first two models of allostery emerged. Monod *et al.* proposed the "concerted" or MWC model based on at least two known conformational states of the deoxy- and oxyhemoglobin. According to MWC model, the allosteric proteins are symmetric multimers, and each monomer is either in tensed (binding incompetent) or relaxed state (binding competent state). The interconversion between two conformational entities is then concerted. Tensed and relaxed states differ in their affinities for the substrate. The authors excluded that mixed tensed-relaxed (TR) state existed [15]. In a sequential (KNF) model, Koshland *et al.* extended the previous work of Pauling and stated that subunits of the oligomer change their conformation one at the time [16]. Hence, a hybrid TR state may exist in the conformational ensemble. Taken together, MWC and KNF model can be combined in a general model of allostery.

In the decades to follow the numerous proves of the allosteric mechanism in the monomeric proteins emerged. Even for a rather rigid protein lacking the huge conformational changes, an allosteric mechanism could not be excluded. Those findings led to the shift in the paradigm of understanding allostery. All but the fibrous proteins exist as a mixture of the different conformational states in a solution. Weber *et al.* proposed that ligand binding only shift the conformational ensemble towards the elevated populations of the binding-competent structures (Figure 1.3) [17]. Starting from this powerful concept Gansekaren *et al.* postulated that the allostery is likely to be an intrinsic property of all dynamical proteins. Pieces of evidence for such hypothesis are numerous. Initially, the

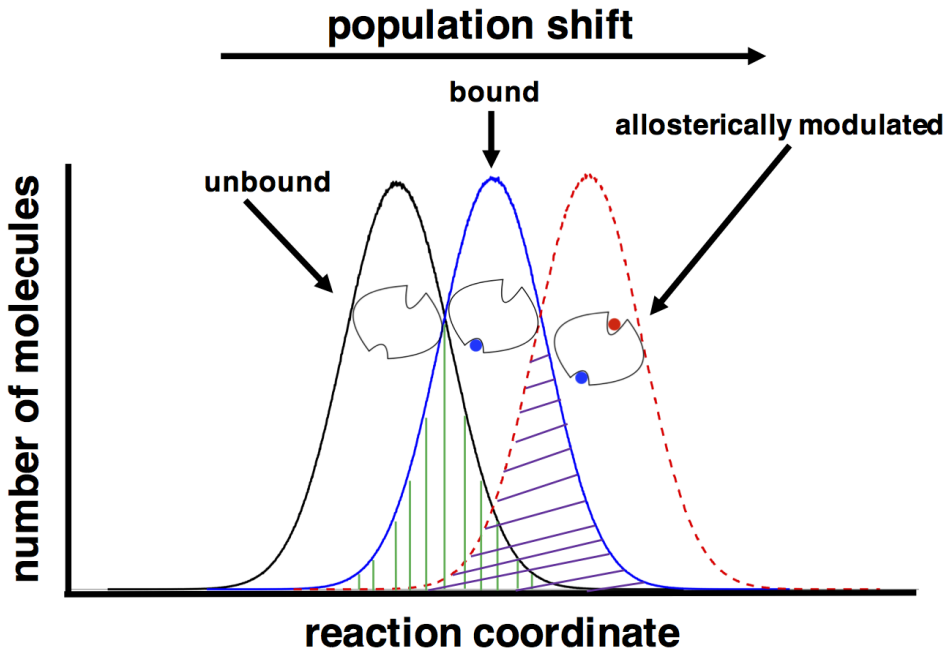


FIGURE 1.3: An allosteric modulator causes a shift in the conformational ensemble.

proteins thought to be non-allosteric were later allosterically modulated by the site-directed mutagenesis, chemical modifications (i.e., phosphorylation of the key residues), or by the detection of the potent allosteric agents [18].

Current view on the allostery can be summarized in terms of "domino" and "violin" models. The domino model is in line with the traditional view on the allostery. The allosteric modulator binds to the allosteric site distant from the orthosteric site and provokes strong signal transduction followed by significant conformational changes. On the other hand, in violin model, the allosteric signal is spread through protein along several relatively weak pathways mainly as vibrational fluctuations of the protein structure [19].

While the MWC and KNF models could not explain the allostery at the atomistic level, advances in the experimental and computational techniques broaden our perception of allostery [20–22]. Consequently, allosteric modulators gained the momentum in drug design. Considering the rationale that almost all proteins can be allosterically triggered then there is a possibility to design drugs hitting the "cryptic" sites. As a proof of concept, Bowman *et al.* utilized the combination of molecular dynamics and cysteine-labeling experiments to detect cryptic sites in three known allosteric drug targets [23]. Binding to the allosteric sites has a considerable advantage, as those sites are protein specific unlike the orthosteric site conserved within the protein family and among different species.

Then, an allosteric drug can be applied in lower doses compared to an orthosteric counterpart directly competing with an endogenous substrate. Finally, allosteric drugs binding at the protein-protein interfaces has also been reported [20, 24].

1.3 Protein-Protein Interactions

Antigen-antibody interactions [25] and spliceosomal proteins are well studied examples of multivalent protein-protein interplays [26]. 20000 coding genes give rise to approximately 200000 proteins in the humans. This huge variety is a consequence of alternative splicing and post-translational modifications [27]. Proteins work in concert with other proteins through protein-protein interactions (PPIs). PPIs are physical contacts established between two or more proteins through van der Waals and electrostatic interactions. Due to temporal and spatial dimensions of the PPIs, it is hard to estimate the exact number of unique PPIs. In general, PPIs are highly dynamical complexes dependent on the condition and state of the cell [28].

PPIs can be classified based on the composition, stability, lifetime and affinity of the protein complex. By the composition, a complex can be homo- or heterooligomer. A homooligomer consists of the identical subunits, while the formation of heterooligomers occurs between nonidentical protein chains. Regarding stability, PPIs are said to be obligate, when monomers cannot exist independently in vivo. Such proteins request a binding partner or partners to maintain their fold and stability. On the contrary, protein participating in nonobligate PPIs, first fold and then find the partner(s) to form the complex. Further, classification based on the lifetime of the protein complex only applies to the nonobligate PPIs. Transient protein complexes associate and dissociate temporarily, and they are often found in signaling and regulatory pathways. Permanent interactions are typically very stable. An example of the permanent PPI is an antigen-antibody complex. Finally, affinity of the PPI implies if the interaction is weak or strong [20].

Proteins interact through interfaces, which are usually flat, large and exposed to solvent (Figure 1.4A). The area of a typical protein interface is between 1500 and 3000 Å. Additionally, PPIs can be classified based on the architecture of the interface [32]:

- **linear epitopes** comprise the short linear peptides or turn that fit into the interface of the binding partner (Figure 1.4B);

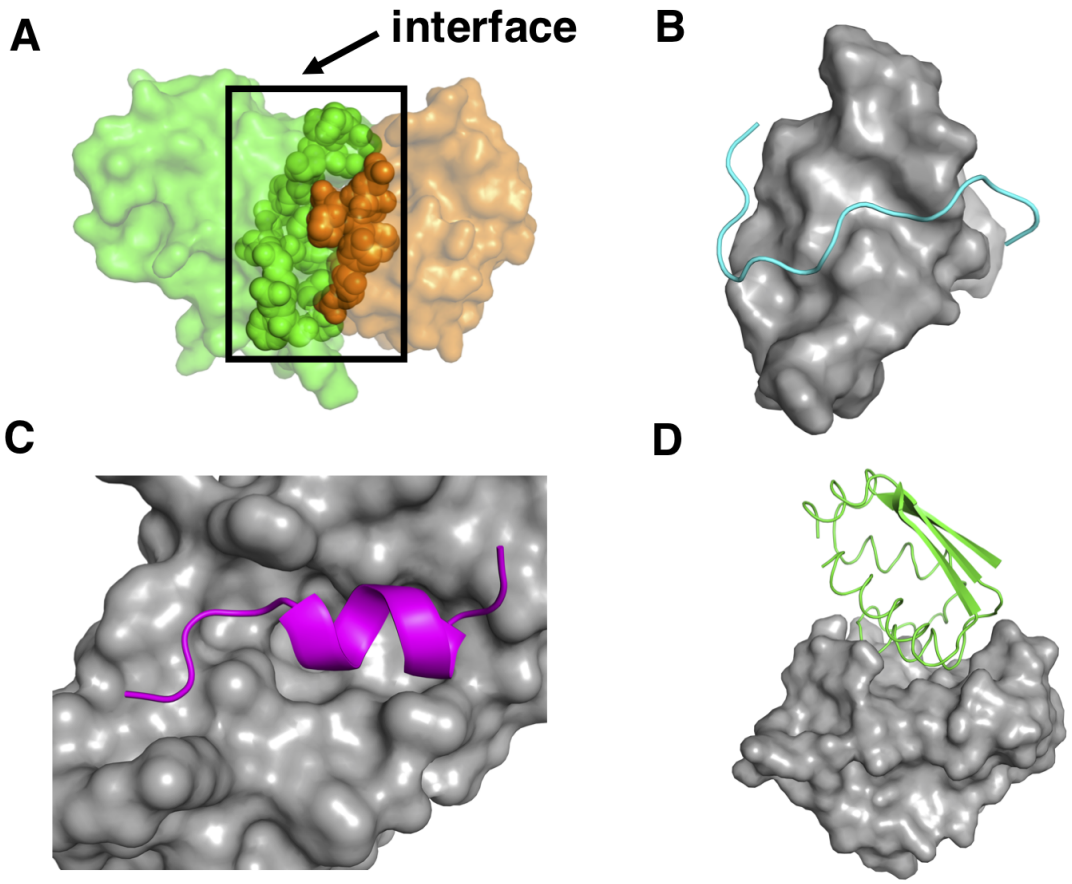


FIGURE 1.4: Protein-Protein Interactions: (A) Two proteins form an interface comprised of non-convalent interactions; (B) Linear epitopes - Smad 7 peptide (shown in cyan) bound to Smurf 1 WW_2 domain (shown in gray) (pdb: 2ltx [29]); (C) Secondary epitopes - MDM2 (in magenta) bound to the transactivation domain of p53 (pdb: 1ycr [30])) ; (D) Tertiary epitopes - Barnase-Barstar complex (barnase shown in gray, barstar in green) (pdb: 1brs [31]).

- **secondary epitopes** represent the most common type of PPIs. A single secondary structure element (α-helix, β-sheet or longer peptides) binds to the groove of its counterpart (Figure 1.4C);
- **tertiary epitopes** comprise the multiple sequences of both binding partners engaged in the interface formation (Figure 1.4D).

PPI interfaces consist of the core and rim regions. The core region is mainly hydrophobic and resembles the protein interior. In contrast, the rim region is exhibited to the solvent and mimics the surface of a protein. Albeit PPI contain a large number of residues, only a small subset of them contributes significantly to the binding affinity. They are named "hot spots" residues [33] and located in

the core region. Tryptophans, arginines, and tyrosines are the most abundant "hot spots" residues [34].

Lately, the "hot spot" residues have been tested as the potential drug targets since they can provide a variety of the intermolecular non-covalent bonds ranging from π-π stacking interactions to strong salt bridges. The linear and secondary epitopes have been employed as the lead compounds of peptidomimetics approach. Furthermore, even tertiary epitopes initially thought to be undruggable, have been targeted successfully with macrocyclic compounds [32, 35, 36].

1.4 Thesis Objectives

I aimed at understanding the dynamics of the multivalent receptors and its implications on the binding of the multivalent ligands with the atomistic precision. To achieve this goal, I applied various computational methods, in particular, molecular dynamics simulations (MD). In **Chapter 2**, I presented a theoretical overview of the methods facilitating the research conducted in this dissertation.

Chapter 3 is dedicated to the study on elucidating the structural determinants of Ca^{2+} binding to the trimeric C-type lectin receptor Langerin. I employed MD simulations to investigate the Langerin conformational dynamics mimicking the conditions in extracellular and intracellular compartments. Then, by implementing the Mutual Information, I constructed a network of dynamically coupled residues. This network guided mutagenesis experiments (conducted by Dr. Jonas Hanske from Structural Glycobiology Group led by Dr. Christoph Rademacher), which then shed light on the role of an allosteric network triggered by Ca^{2+} binding.

In **Chapter 4**, I studied the conformational dynamics of the trivalent sialic acid based constructs targeting Haemagglutinin. MD simulations revealed that PEG spacer collapsed into an ensemble of coiled structures further stabilized by the presence of the intramolecular hydrogen bonds. Those findings can guide further efforts in optimizing such multivalent ligands. In this project I collaborated with Pallavi Kiran, and Dr. Sumati Bhatia (Macromolecular Chemistry Group led by Prof. Dr. Rainer Haag), Dr. Susanne Liese (Bio-soft Matter Theory Group led by Prof. Dr. Roland Netz) and Dr. Daniel Lauster (Molecular Biophysics Group led by Prof. Dr. Andreas Hermann).

In **Chapter 5**, I addressed multivalent protein-protein interactions. I was interested in understanding how the highly complex conformational space of tandem-WW domains influenced the binding of a bivalent proline-rich sequence. To answer this question, I constructed a Markov state model (MSM) of the apo-receptor dynamics. MSM analysis revealed that the slowest kinetic modes entailed for the structures with the formed interdomain interface. In turn, interface formation blocked "hot spot" residues from simultaneous recognition of both valences. I found a binding-competent structure in the ensemble of the fast interconverting structures lacking the defined interface. Finally, I modeled the binding event by employing HADDOCK protein-protein protocol. Experimental insights for this project were provided by Miriam Bertazzon and Dr. Jana Sticht of Protein Biochemistry Group led by Prof. Dr. Christian Freund.

The outlook for the further research in all three areas covered in this thesis was summarized in **Chapter 6**.

Chapter 2

Methods and Theory

2.1 Molecular Dynamics

With the advent of X-ray crystallography and nuclear magnetic resonance (NMR) in the middle of last century, scientists were able to determine the structures of biomolecules with the atomic resolution for the first time. However, those biophysical methods are limited to the rather rigid representations of biomolecules. The internal motions of atoms result in the conformational changes, which are in turn necessary for a biomolecule to fulfill its biological role. Molecular simulations emerged as a computational technique complementing experiments in the efforts of answering complex biological questions [37, 38].

Molecular simulations found applications in the sampling of the conformational space of a target molecule. Then, they can recover the thermodynamical properties of an investigated system. Finally, they examine the actual dynamics of a biomolecule. For the first two application, one can choose between Monte Carlo and Molecular Dynamics simulations (MD), yet only MD simulations describe the time evolution of the system and can account for the dynamics [39].

The all-atom (classical) MD simulations are based on the laws of classical mechanics. Each atom in the investigated system is treated as a point mass (m_i). The movement of a point particle i can be represented with a position vector $r_i(t)$ with an arbitrary origin in a 3D-Cartesian space:

$$r_i(t) = r(x, y, z) = x(t)\vec{e_x} + y(t)\vec{e_y} + z(t)\vec{e_z} \tag{2.1}$$

where $x(t)$, $y(t)$, and $z(t)$ are the time-dependent displacements of the point particle i along the respective Cartesian coordinate, while $\vec{e_x}$, $\vec{e_y}$, and $\vec{e_z}$ are the orthogonal basis vectors. The current velocity (Eq. 2.2) and acceleration (Eq. 2.3) of the point particle i calculated for a small time-step $\Delta t = t_2 - t_1 \approx 0$,

represent the first and the second derivatives of $r_i(t)$ respectively:

$$v_i = \lim_{\Delta t \to 0} \frac{r_i(t_2) - r_i(t_1)}{\Delta t} = \frac{d}{dt} r_i(t) = \dot{r}_i(t) \tag{2.2}$$

$$a_i = \lim_{\Delta t \to 0} \frac{v_i(t_2) - v_i(t_1)}{\Delta t} = \frac{d}{dt} \dot{v}_i(t) = \frac{d}{dt} \frac{d}{dt} \dot{r}_i(t) = \ddot{r}_i(t) \tag{2.3}$$

The force vector $F_i(t)$ acting on the point particle i and causing the movement along the free energy surface is denoted according to the Newton's second law of motion:

$$F_i(t) = m_i a_i = m_i \ddot{r}_i(t) \tag{2.4}$$

Since a biomolecule consists of a large number of atoms integrating the set of equations of motions (Eq. 2.2, 2.3, 2.4) analytically becomes unfeasible. To circumvent this problem and to assure the low computational demand, yet still, to maintain the high accuracy of an MD algorithm, a number of numerical integration methods have been developed. Among them, Verlet-type algorithms are the most common. The general equation of a Verlet algorithm for the point particle i can be derived from the forward and the backward Taylor expansion of the $r_i(t)$:

$$\begin{aligned}
r_i(t+\Delta) &= r_i(t) + \dot{r}_i(t)\Delta + \frac{1}{2}\ddot{r}_i(t)\Delta^2 + \frac{1}{3}\dddot{r}_i(t)\Delta^3 + \mathcal{O}(\Delta^4) \\
r_i(t-\Delta) &= r_i(t) - \dot{r}_i(t)\Delta + \frac{1}{2}\ddot{r}_i(t)\Delta^2 - \frac{1}{3}\dddot{r}_i(t)\Delta^3 + \mathcal{O}(\Delta^4) \\
&\Downarrow \\
r_i(t+\Delta) &= 2r_i(t) - r_i(t-\Delta) + \ddot{r}_i(t)\Delta^2 + \mathcal{O}(\Delta^4) \\
&\Downarrow \\
r_i(t+\Delta) &= 2r_i(t) - r_i(t-\Delta) + \frac{F_i(t)}{2m_i}\Delta^2 + \mathcal{O}(\Delta^4).
\end{aligned} \tag{2.5}$$

suggesting that the position of the particle i $r_i(t+\Delta)$ in the next simulation step is determined by the current position $r_i(t)$, the position at the previous time step $r_i(t-\Delta)$, and the force vector acting on the the particle i at the current time step $F_i(t)$. Thus, the Verlet algorithm is time-reversible by definition. The velocity of the particle i at the current time step t is calculated according:

$$v_i(t) = \frac{r_i(t+\Delta) - r_i(t-\Delta)}{2\Delta} \tag{2.6}$$

Verlet type algorithms (Verlet, velocity-Verlet, leap-frog) are fourth-order algorithms, as the error of calculating the position of the particle i is approximated

with the fourth power of the integration time step Δ [40]. The value of the integration step Δ is determined by the fastest motions of the system. The bonds connecting the hydrogen and heavy atoms vibrate with periods of 10 fs. Therefore, the integration time step should be at least one order of magnitude smaller. If such vibrations are constrained by means of SHAKE [41] or LINCS [42] algorithms, integrations time step can be increased up to 4 fs to further speed up calculations. Verlet type algorithms preserve quantities like the total (mechanical) energy, momentum, and angular momentum of the system sampled in an MD simulation. For a system comprising N atoms (point particles), with the conserved mechanical energy, its classical Hamiltonian (H) is denoted as follows:

$$\begin{aligned} H &= \sum_{i=1}^{N} V(r_i(t)) + \sum_{i=1}^{N} K(p_i) \\ &= \sum_{i=1}^{N} V(r_i(t)) + \frac{1}{2}\dot{r}_i(t) \sum_{i=1}^{N} m_i \dot{r}_i(t) = const. \end{aligned} \tag{2.7}$$

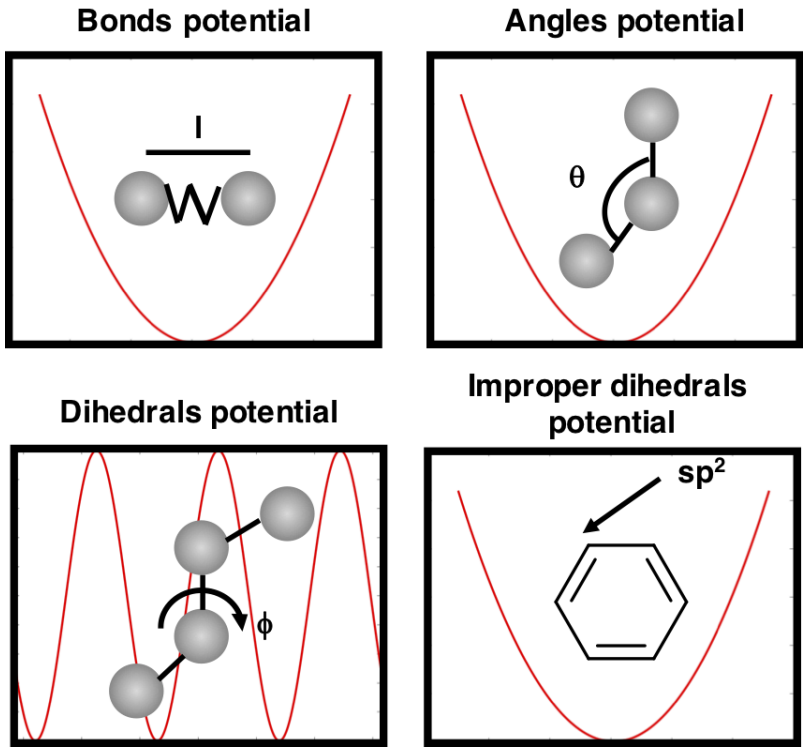


FIGURE 2.1: Schematic representation of the bonded potentials of a force field.

The sum of the all momenta is constant according to the principle of action and reaction. To keep the potential energy term constant, one assumes that only a set of conservative forces (F_i vector) causes the displacement of the molecule. A

conservative force is a force only dependent on the coordinates of an object on which it is acting, but not on the speed of that object. Hence, the work associated with a molecule displacement is independent of the path taken. In the context of the of the MD simulations, the set of conservative forces acting on the investigated molecule is often referred as a force field. The force field is given as a sum of the interatomic potentials accounting for the bonded and non-bonded interactions. Equation 2.4 can be further expanded, since $F_i(t)$ equals to the negative gradient of the potential energy:

$$F_i(t) = m_i \ddot{r}_i(t) = -\nabla V(R(t)) \tag{2.8}$$

A mathematical expression for a typical force field is given as:

$$\begin{aligned} V(R(t)) = & \overbrace{\sum_{bonds} \frac{1}{2} k_l (l - l_{eq})^2 + \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_{eq})^2}^{\text{bonded terms}} \\ & + \overbrace{\sum_{dihedrals} k_\phi [1 + \cos(n\phi - \gamma)] + \sum_{\substack{\text{improper} \\ \text{dihedrals}}} \frac{1}{2} k_\zeta (\zeta - \zeta_{eq})^2}^{\text{bonded terms}} \\ & + \overbrace{\sum_{\substack{\text{non-bonded} \\ \text{atom pairs}}} 4\eta((\frac{\sigma}{d})^{12} - (\frac{\sigma}{d})^6)) + \sum_{\substack{\text{non-bonded} \\ \text{atom pairs}}} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1} \frac{1}{d_{char}}}^{\text{non-bonded terms}} \end{aligned} \tag{2.9}$$

where the position vector of the investigated molecule is realized as $R(t) = \{ r_1(t), r_2(t), ..., r_N(t) \}$. The bonds and angles potentials in the force field equation treat the stretching and bending motions around a chemical bond respectively. They ensure the correct chemical structure of the investigated molecule and prevent bond breaking. Both potentials are approximated as harmonic oscillators (Figure 2.1) and denoted according to a Hooke's law formula. In the bond potential, k_l is a force constant, l is a bond length at the simulation time step t, and l_{eq} is an equilibrium bond length for the same atom pair determined experimentally (i.e., X-ray crystallography) or obtained from the high-level quantum mechanics calculations. Analogously to the bond potential, the angle potential (three atoms are needed to define a bond angle θ) is defined with k_θ (force constant), θ (angle value at the current simulation time step t), and θ_{eq} (angle value at equilibrium) terms. The dihedral angles potential describes the rotation around a chemical

bond (four atoms are needed to define a dihedral angle), and it is given as a periodic cosine function. In the dihedral potential term of Eq. 2.9 k_ϕ stands for a force constant, n is a number of minima in a dihedral potential energy profile, and γ is a phase offset (Figure 2.1). The improper dihedrals term is needed to ensure the planarity of some particular groups, such as sp^2 hybridized carbons in carbonyl groups or in aromatic rings. Similarly to bonds and angles potentials, the improper dihedral potential is approximated with a harmonic oscillator (oscillation from the equilibrium value of the respective improper dihedral ζ_{eq} with force constant k_ζ (Figure 2.1)). The last two terms of the force field equation account for the long-range non-bonded (non-covalent) interactions. The repulsive and attracting van der Waals interatomic forces are modeled with the Lennard-Jones (LJ) 12-6 potential. The η term of the LJ potential represents the depth of the potential well (Figure 2.2 left-hand side panel), while σ term is an interatomic distance at which V_{LJ} =0. The electrostatic interactions between non-bonded (partially) charged atoms are modeled according to the Coulomb's law. ϵ_0 and ϵ_i are the dielectric permeabilities of the vacuum and a solvent in which the system is simulated respectively, while q_i and q_j are (partial) charges of ij atoms, and d_{char} the distance between charged species. [43].

Molecular Dynamics is a statistical mechanics method. The macroscopic properties of the system can be calculated by averaging over the sets of microstates (configurations) according to the Boltzmann distribution. To better mimic the actual macroscopic behavior of the system, it is possible to couple the simulation box to external thermostatic bath (canonical (NVT) ensemble in which the number of particles N, the volume of the system V and temperature T are constant). Thermostats (Andersen [44], Noose-Hover [45], Berendsen [46]) are algorithms that keep the temperature of the system around the desired macroscopic value through a proper alternation of the equations of motion. Similarly, the pressure can be maintained at the constant value (isothermal-isobaric ensemble, NPT = const) by the employment of the barostat algorithms.

To address the finite size effects and surface properties periodic boundary conditions (PBC) are commonly applied in an MD simulation. The system of interest is placed in a unit cell, which is in turn replicated in all directions. The coordinates and velocities are propagated for the unit cell exclusively, while the non-bonded terms of a force field are determined over all periodic images. As Lennard Jones potential decays with d^{-12} and d^{-6} a spherical cut off 1 nm is used when computing van der Waals forces [40]. Due to the long-range nature of the electrostatic interactions, they should be evaluated over the whole periodic

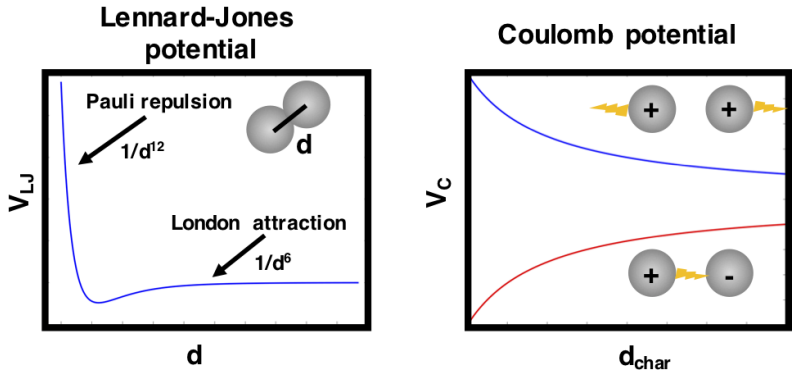


FIGURE 2.2: Schematic representation of the non-bonded potentials of a force field.

lattice. As shown in Figure 2.2 (right-hand side panel), the electrostatic potential decays very slowly with the factor $\frac{1}{d_{char}}$ resulting in an increase of the computation cost for calculating all electrostatic interactions. Ewald sum methods have been utilized in solving this computationally demanding problem [47].

2.2 Computational Approaches for Prediction of NMR Observables

In recent decades NMR became a widely used technique for the structural determination of the biomolecules in particular proteins. ^{1}H, ^{13}C, ^{15}N chemical shifts are utilized to determine the presence of the secondary structure elements by Chemical Shift Index method [48]. The ensemble average of the backbone and sidechain dihedrals are obtained from the scalar coupling measurements (COSY spectrum). Then, the interatomic distances are deduced from NOESY spectrum (spatial magnetization transfer - Nuclear Overhauser effect). Finally, the relative orientation of the tertiary structure elements can be concluded by Residual Dipolar Coupling measurements. One can combine these NMR experiments to elucidate an ensemble of structures for a protein of interest.

Within this thesis, I computed the chemical shifts, and the ^{3}J-coupling constants based on the performed MD simulations of the molecules investigated in **Chapters 3** and **5** (only chemical shifts were computed in this chapter). Then, I compared computed data with the experimentally determined values, which is a common approach for the validation of the quality of the MD simulations.

To introduce the concepts of chemical shift and ^{3}J-coupling constant consider a proton (^{1}H) NMR experiment. A nucleus of interest in such experiment is

a spinning proton. By applying an external magnetic field B_0, the spin of the proton may be aligned with B_0 (spin number $+\frac{1}{2}$, α spin state), or against B_0 (spin number $-\frac{1}{2}$, β spin state). If an imaginary sample for an NMR experiment would consist only of protons, then the magnetic field at a H^+ would equal to B_0, as no additional magnetic field was induced in the absence of an electron. Hence, the energy difference ΔE between the α and β spin states would be given as:

$$\Delta E = \frac{\gamma h B_0}{2\pi} = h\nu_{sample} \tag{2.10}$$

where γ is a magnetogyric constant, h is a Planck constant, and ν_{sample} is a resonance frequency of this imaginary sample. Now assume that a sample for an NMR experiment consists only of hydrogen atoms, then a spinning electron would induce an additional magnetic field B_{ind}, which would be aligned against the external magnetic field B_0. Thus, the proton is thought to be shielded by the electron, as the effective magnetic field at the proton B_{eff} is lower compared to B_0. Then, the equation 2.10 can be reformulated as follows:

$$\Delta E = \frac{\gamma h(B_0 - B_{ind})}{2\pi} = \frac{\gamma h B_{eff}}{2\pi} = h\nu_{sample}. \tag{2.11}$$

This shielding effect yields a lower value for ΔE between the α and β spin states of a hydrogen atom compared to an imaginary sample only containing protons (H^+). Since ΔE and ν_{sample} are directly proportional, a signal for a hydrogen

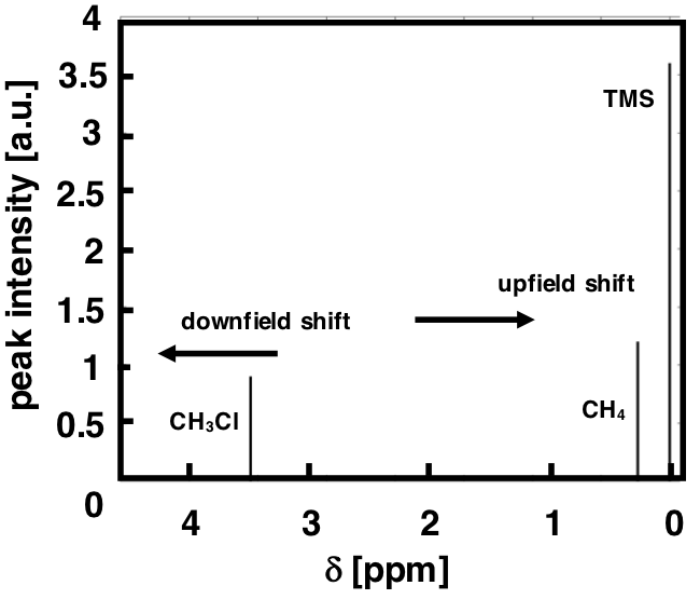


FIGURE 2.3: Schematic representation of a combined ^{1}H-NMR spectrum of CH_3Cl and CH_4.

atom would be shifted to the lower frequency range compared to a proton signal (upfield shift). If one would compare signals of protons in spectra of methane (CH_4) and chloromethane (CH_3Cl), a signal for chloromethane protons would be shifted to the higher frequency range (downfield shift). An electronegative Cl atom attracts the electron density towards itself resulting in deshielding effect on three protons of CH_3Cl (Figure 2.3). Therefore, the effective magnetic field would be higher compared to the effective magnetic field felt by methane protons. As different NMR spectrometers operate at different working frequencies (different strength of an external magnetic field), signals in a 1D-NMR spectrum are reported as chemical shifts [ppm] or shifts from the signal of a standard compound. The standard compound like tetramethylsilane $(CH_3)_4Si$ (TMS) consists of almost completely shielded nuclei. Chemical shift δ of the respective nucleus in a sample represents a measure of the nuclear (de)shielding effect, which is in turn in direct correlation to the chemical environment surrounding that specific nucleus. Chemical shift δ is calculated as:

$$\delta = \frac{\nu_{samp} - \nu_{stand}}{\nu_{stand}} \quad \frac{[Hz]}{[MHz]} = [ppm] \tag{2.12}$$

Because of the total nuclear shielding effect the absolute resonance frequency of TMS ν_{stand} equals to the operating frequency of a NMR spectrometer. Hence, the chemical shift of TMS is set to 0 ppm. The deshielding effect at any other sample that is not a standard compound causes the downfield shift of its resonance frequency (ν_{sample}-ν_{stand}).

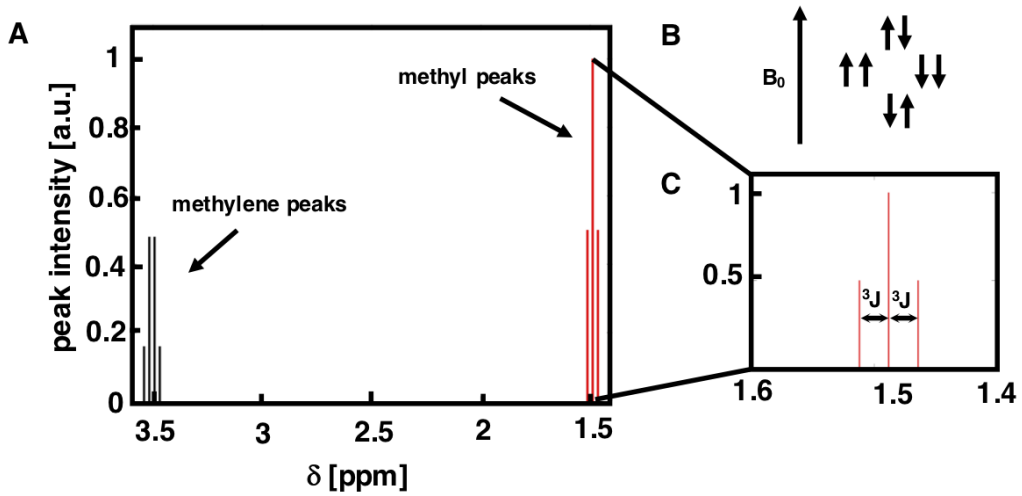


FIGURE 2.4: (A) Schematic representation of ^{1}H-NMR spectrum of CH_3CH_2Cl according to SDBS No. 3347HPM-02-151 spectrum [49]; (B) Spin states of the methylene protons (spin-spin coupling); (C) Enlarged methyl peaks with denoted ^{3}-J coupling constants.

In Figure 2.4A a schematic representation of a ^{1}H-NMR spectrum of chloroethane is given. One would expect to observe just two peaks, a peak representing the methyl group protons, and a peak for the methylene protons. However, the methyl peak has been split into three peaks, while the methylene peak into four peaks. This splitting pattern can be explained with spin-spin interaction (coupling) between two group of protons bound to the vicinal carbons. To explain why the methyl peak is split into a triplet, one should look at the possible orientations of the spins of two methylene protons. Their spins can be aligned to B_0 or against it resulting in four combinations of their spins states (Fig 2.4B). Two combinations of the spin states are equal ($\uparrow\downarrow$, $\downarrow\uparrow$ and the induced magnetic fields cancel each other) and yield a middle peak. When both protons have spins paired in the direction of B_0, then a downfield peak is produced. On the contrary, impaired spins produce an upfield peak. The multiplicity of the peaks is determined by the number of chemically equivalent protons at the vicinal carbon(s) increased by 1 (n+1 rule). The chemically equivalent nuclei do not couple to each other. By applying the same n+1 rule, then the methylene peak is split into a quartet. The peaks in both triplet (Figure 2.4C) and quartet are split with the same spacing independent of B_0 called ^3J-coupling constant [Hz] (spin-spin coupling occurs between two protons separated by three chemical bonds) [50].

The magnitude of ^3J-coupling constant provides a wealth of information on the spatial orientation of two considered protons. For a pair of protons separated by three chemical bonds, ^3J-coupling constant is given by Karplus equation:

$$^3J_{H,H'} = Acos^2\theta + Bcos\theta + C \tag{2.13}$$

where θ is a dihedral angle defined by four covalently bound atoms and A, B, C are empirically determined parameters [51]. Those parameters are obtained by fitting measured ^3J-coupling constants for which respective torsion angle has been elucidated previously by X-ray crystallography. Determining $^3J_{H^\alpha,H^N}$ and $^3J_{H^\alpha,H^\beta}$ provides the structural insight on the conformational space sampled by the backbone and sidechain ϕ, χ_1 dihedrals respectively. As shown in Figure 2.5, the $^3J_{H^\alpha,H^N}$-coupling is the highest inside the β-sheets. The spatial orientation of the same atoms when located in a right α-helix correlates with the lowest values of $^3J_{H^\alpha,H^N}$-constants. For the highly dynamical residues located in the loops, ^3J constant cannot be used to unambiguous resolve value of the dihedral ϕ, as it corresponds to fully sampled ϕ-space. Nonetheless, Karplus equation still holds as a powerful tool for the validation of the simulation data. The time series of the dihedral ϕ is easily accessible allowing comparison of the computed and

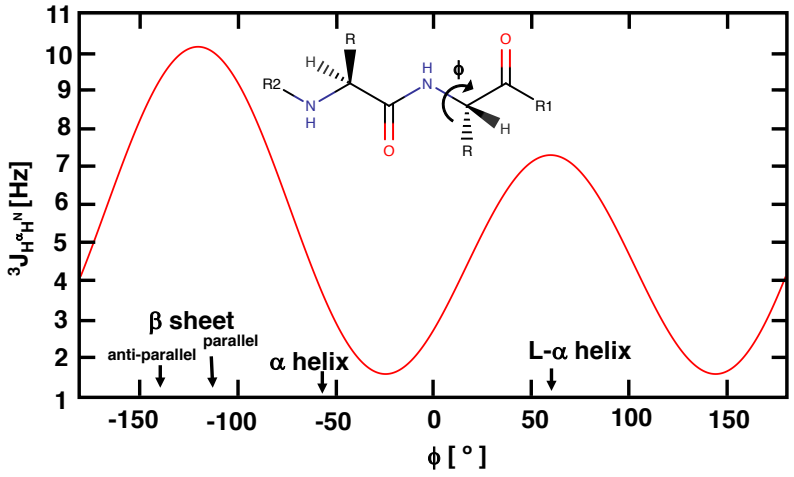


FIGURE 2.5: ^{3}J-coupling constant as a function of ϕ-dihedral of a canonical residue.

experimentally determined ^3J-coupling constants [52].

Broadening our understanding of the critical physico-chemical factors responsible for the chemical shift changes paved the way for the computational methods for the chemical shift prediction. These methods can be classified into three groups. The first group comprises the methods based on the use of sequence/structure alignment against protein chemical shifts databases. The second group is based on calculating chemical shifts directly from atomic coordinates by employing empirical equations derived from classical physics and experimental data or by applying density functional theory. Lately, the former two methods were combined into several hybrid methods. They have become popular as they yield more accurate results compared the older approaches.

One can use a chemical shifts prediction software to validate the quality of MD data. SPARTA$^+$ is software with built-in artificial neural network trained on the database of 580 proteins for which relations between tertiary structure and chemical shifts are established. SPARTA$^+$ protocol incorporates both structural (backbone and sidechain dihedrals) and dynamical inputs (S^2-order parameter) with the information on the local interactions (i.e., hydrogen bonds) to elucidate the chemical shifts for a query structure [53]. Beauchamp *et al.* critically assessed the quality of SPARTA$^+$ protocol for prediction of backbone atoms chemical shifts based on the MD derived ensembles. The reported errors for the AMBER99-ILDN

family of force fields were in line with the experimental errors for nonproline canonical residues [54].

2.3 Markov State Models of Protein Dynamics

The classical molecular dynamics simulations generate the ample of data regarding the conformational behavior, the thermodynamical and kinetic properties of a system of interest. Only recently the Markov State Models (MSMs) have emerged as a powerful framework capable of resolving a free energy landscape of a simulated biomolecule. Hence, MSM provides insights into long-lived conformations of the system and the kinetic rates at which the system interconverts between the minima of the free energy landscape. The method has been applied in studying biologically relevant phenomena like the conformational dynamics [55], ligand binding [56], protein folding [57], and allostery [23].

Consider a state space Ω comprising all the coordinates and velocities of the particles in a system of interest. To construct the MSM of a molecule of interest one assumes:

- Markovian property of the system, implying that the time evolution of the system is only determined by the current state, but not dependent on the history of the system;
- that system is ergodic, hence all parts of the state space Ω are dynamically connected;
- that detailed balance condition is satisfied meaning that the portion of the system transitioning from the region i of Ω to the region j per time unit equals the portion of the system transitioning from j to i.

Now assume that several MD trajectories of the system of interest run in parallel and that the starting points of the respective simulations are distributed according to some initial probability density function $u_{t=0}$. Continuous treatment of the state space Ω relies on the transfer operator formalism, which can be summarized as follows:

$$u_{t+\tau}(y) = \mathcal{T}(\tau)u_t(y) = \frac{1}{\pi(y)}\int_{\Omega} p(x, y; \tau)\pi(x)u_t(x)\mathrm{d}x. \tag{2.14}$$

If the system is initialized in the y region of Ω (at the time step t), by applying the transfer operator $\mathcal{T}(\tau)$ to the probability density function $\mathrm{u}_t(\mathrm{y})$ over the time discretization τ, $\mathrm{u}_t(\mathrm{y})$ will evolve to a modified probability density function

$u_{t+\tau}(y)$. If this procedure is repeated long enough, the system will transition from the region y to the region x of Ω, which is given by the transition probability density $p(x,y;\tau)$. For infinite amount of time the system will visit the whole state space Ω, while $u_t(y)$ will converged to a constant function independent of the stationary distribution $\pi(y)$. If the system is simulated in an NVT or an NPT thermodynamical ensemble, the stationary distribution $\pi(y)$ is equivalent to the Boltzmann distribution at a temperature of simulation:

$$\pi(x) = \frac{\exp(-\frac{1}{k_B T} H(x))}{\int_\Omega \mathrm{d}x \exp(-\frac{1}{k_B T} H(x))} \tag{2.15}$$

where $H(x)$ is Hamiltonian of the system, and k_B is the Boltzmann constant. [58]. As discussed in [59, 60], the transfer operator $\mathcal{T}(\tau)$ is self-adjoint implying that its eigenvalues $\lambda_k(\tau)$ and eigenfunctions $r_k(y)$ are real-valued. Since the system is assumed to be ergodic and reversible, solving Eq. 2.16:

$$\mathcal{T}(\tau) r_k(y) = \lambda_k(\tau) r_k(y) \tag{2.16}$$

would yield information on the dynamical modes of the system. In the full state space Ω molecular dynamics is Markovian by construction, however, finding the analytical solution to the eigenvalues and the eigenfunctions of the transfer operator is not feasible for any biomolecular system. Hence, within MSM framework one usually works in a discretized state space $\Omega_{,disc}$, partitioned in a set of N non-overlapping states ($S_i \cap S_j = 0$ and $\cup_{i=1}^{N} S_i = \Omega$). Those discrete states are called microstates and can be thought as a small portion (volume) of the state space. Ideally, discretization algorithm should be purely kinetic, so the barrier between microstates would overlap with the barriers of the free energy surface. In the discrete state space the transfer operator is approximated with the transition matrix $T(\tau)$, in which the transition probabilities of jumping from the microstate S_i to S_j within the observation interval or lag time τ are stored. The probability density $p_t(y)$ is given as:

$$p_t(y) = \pi(y) u_t(y) \tag{2.17}$$

where $\pi(y)$ and $u_t(y)$ are the stationary distribution and the probability density function respectively. In the discrete state space $\Omega_{,disc}$ the probability density is given as a vector containing the same number of elements as the number of microstates. Each entry represents the probability of finding the system in that specific microstate at the time t. Since the state space is fully partitioned, the sum of all entries of the $p_t(y)$ vector always equals to 1. As the system evolves

over the time, the $p_t(y)$ vector is updated according to:

$$p_{t+\tau}(y) = T(\tau)p_t(y). \tag{2.18}$$

The transition matrix $T(\tau)$ is computed based on a count matrix. The count matrix is a N×N matrix, where N is the number of microstates. In a sliding window approach, the transitions are counted as follows:

- check in which microstate is the projected trajectory at the time step t=0 (S_i) and t=0+τ (S_j), and then increase the count for the element C_{ij} by one;
- move to the next time step (t=1), and compare it with the time step t=1+τ, while increasing the count of the respective element C_{ij} by one;
- repeat the counting procedure until the time step t=T-τ (where T is the length of a trajectory).

Finally, to enforce the detailed balance condition the total number of the transitions between ij and ji microstate pairs are averaged:

$$C_{ij} = C_{ji} = \frac{C_{ij} + C_{ji}}{2}. \tag{2.19}$$

Then the transition matrix is realized as:

$$T_{ij} = \frac{C_{ij}}{\sum_k C_{ik}}. \tag{2.20}$$

The transition matrix is a row stochastic matrix, meaning that each entry in the k^{th} row of C is normalized with respect to the sum of all entries of the k^{th} row.

For an ergodic and reversible system, analogously to the equation 2.16, the transition matrix $T(\tau)$ is subject of the eigendecomposition as given by:

$$l_i^T T(\tau) = \lambda_i(\tau) l_i^T \tag{2.21}$$

where $\lambda_i(\tau)$ and l_i are the real-valued eigenvalues and left eigenvectors of $T(\tau)$ respectively. The eigenvalues $\lambda_i(\tau)$ are bound to the interval [-1,1]. Consequently, the eigenspectrum of $T(\tau)$ is bound from above:

$$|\lambda_i(\tau)| \leq \lambda_1 = 1. \tag{2.22}$$

Since the eigenvalues $\lambda_i(\tau)$ (i>1) decay exponentially, the conformational dynamics of the system is approximated by the linear combination of the N dominant

eigenvectors (determined by the initial slow decay of the corresponding $\lambda_i(\tau)$). Therefore, the probability density p_t can be expressed in the terms of the expansion coefficients c_i determined by the initial probability density $p_{t=0}$:

$$p(t) = \sum_{i=1}^{\infty} c_i \lambda_i^{\frac{t}{\tau}}(\tau) l_i \approx \sum_{i=1}^{N} c_i \lambda_i^{\frac{t}{\tau}}(\tau) l_i. \tag{2.23}$$

All the entries of the first left eigenvector l_1 are positive, and the i^{th} entry of l_1 represents the probability of the system to be found in the S_i microstate of $\Omega_{,disc}$. All other left eigenvectors l_i (i>1) contain the positive, negative, and entries close or equals to zeros (Figure 2.6-right panel). This change in the sign structure of

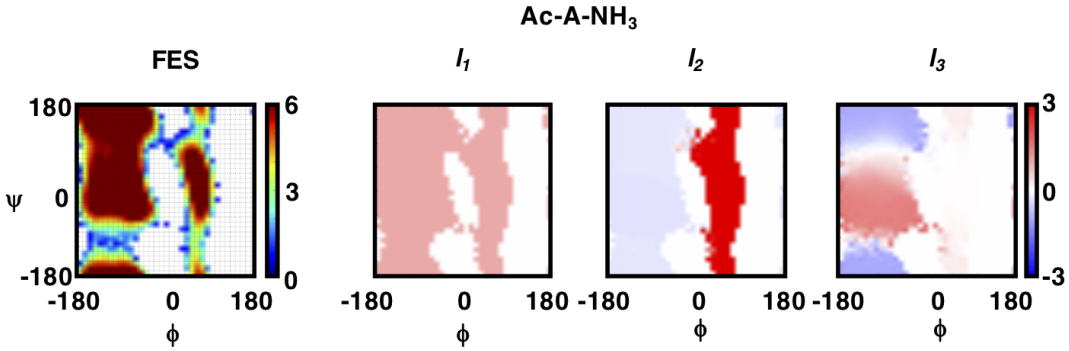


FIGURE 2.6: Free energy landscape of the alanine dipeptide projected onto ϕ-ψ plane discretized into 36×36 grid cells; Left eigenvectors (l_{1-3}) reshaped as 36×36 matrices (right panel).

the eigenvectors entries (for l_i (i>1)) is a basis for the Perron Cluster Cluster Analysis (PCCA). PCCA is an iterative algorithm considering a single eigenvector of $T(\tau)$ at a time. For the simplicity, now consider alanine dipeptide, since its free energy surface is well represented by a projection onto a Ramachandran plane spanned over two backbone dihedrals. It comprises three distinguishable minima, namely α, β, and L-α regions. Additionaly assume, that an MSM for this system has been contstructed by discretizing the free energy surface into 36×36=1296 grid cells (microstates). As shown in Figure 2.6, the microstates corresponding to the L-α region have positive values assigned in l_2, while microstates of the α and β regions have negative values. One interprets the sign structure of l_2 as the kinetic exchange across an energy barrier centered at ϕ=0°. This is the slowest dynamic mode of the alanine dipeptide conformational dynamics. PCCA algorithm has split alanine dipeptide free energy surface into two metastable regions (macrostates). Next look at the sign structure of l_3. Interestingly, the microstates corresponding to the L-α region have all assigned values close to 0. The kinetic exchange occurs between the α region (positive sign structure), and the β region

(negative sign structure), while L-α microstates are excluded from this dynamic mode. Hence, the free energy surface is further split into two macrostates, yielding in total three macrostates, which are analogous to three minima of the Ramachandran plane [61]. To treat more complex molecules (free energy surfaces), a recent algorithm called PCCA$^+$ has been reported [62]. PCCA$^+$ algorithm considers several eigenvectors at the same time, thus the degree of membership to a macrostate is reported for each microstate. An implied time scale of the nth dynamic mode can be determined by rewriting equation 2.23:

$$c_i \lambda_i^{\frac{t}{\tau}}(\tau) l_i = c_i \exp(-\frac{t}{t_i}), \tag{2.24}$$

which yields:

$$t_i = -\frac{\tau}{ln\lambda_i(\tau)} \tag{2.25}$$

For increasing values of the lag time τ the implied time scales converge to the constant values, and are independent of the lag time τ for which the transition matrix of MSM is estimated (Champman-Kolmogorov test).

Several schemes for the discretization of the state space have been proposed in the literature [63]. Often schemes that fully partition Ω yield a poor statistics in the transition regions of the free energy surface, leading to the instability of the model. To circumvent such problem and to minimize the discretization error, the core set approach for the discretization of the state space has been introduced. The idea behind this approach is that a core set represents a metastable region or a minimum of the free energy landscape. Thus, the union of the disjoint core sets B_i does not fully discretize the state space ($\cup_{i=1}^{N}$ $B_i \subset \Omega$) unlike the union of the disjoint microstates. As a result of this discretization scheme some of the regions of Ω do not belong to any of the core sets. This is so-called the intervening space I= Ω / $\cup_{i=1}^{N}$ B_i. It is assumed when the system occasionally leaves a core set and visit the intervening space, a time scale associated with that conformational event is faster then the fastest dynamic mode estimated by the MSM. Equation 2.23 is analogous to the approximation of the transfer operator by the basis functions ψ(x), which form a complete basis of the state space Ω. One can rewrite it as follows:

$$r(x) \approx \sum_{i}^{N} \tilde{c}_i \psi_i(x) \tag{2.26}$$

where r(x) are the eigenfunctions of the transfer operator (Eq. 2.16). In the core set approach the basis functions are modeled with the committor functions q_i. Let assume that the free energy surface of the system has only two minima

represented by the core sets B_0 and B_1. Only one committor function is needed to approximate the transfer operator. If the system is in B_0, then the committor function assumes the value 0, while in B_1 the value of the committor function is 1. As system interconverts from B_0 to B_1, in the intervening space the committor function interpolates between 0 and 1. In the systems with more complex free energy surface, additional committor functions are needed to account for other possible B_i to B_j transitions. Typically, the committor functions are not known analytically, and as shown by Schütte *et al.* [64], they can be approximated by the trajectory projections dependent on the history and the future of the system. These projections are called milestoning processes. The forward milestoning (m^+) process assigns a value 1 if the trajectory is in the core set B_i. Additionally, the value 1 is assigned if the trajectory hit the intervening space, but will next visit the core set B_i. In all other cases, the trajectory is projected to a value 0. The backward milestoning (m^-) again assigns the value 1 when the trajectory is in the core set B_i, or visit the intervening space by coming from B_i. Otherwise, the trajectory is projected to the value 0. As shown by Lemke *et al.* [65], the equation 2.26 can be further expanded into:

$$\tilde{c}^T P(\tau) M^{-1} = \lambda \tilde{c}^T. \tag{2.27}$$

The matrix elements of matrices M and $P(\tau)$ are given as the (time-lagged) correlation functions between a ij pair of the forward and backward milestoning processes according to:

$$\begin{aligned} M_{ij} &= \frac{1}{T} \sum_{t=0}^{T} m_i^-(t) m_j^+(t), \\ P_{ij}(\tau) &= \frac{1}{T-\tau} \sum_{t=0}^{T-\tau} m_i^-(t) m_j^+(t+\tau). \end{aligned} \tag{2.28}$$

The probability of the system visiting the core set B_j next after the system resided in the core set B_i is stored as a matrix element M_{ij}. The matrix $P(\tau)$ is time-lagged analog of the mass matrix M. The product $P(\tau)\mathrm{M}^{-1}$ (Eq. 2.27) is equivalent to the transition matrix $T(\tau)$ of the conventional (fully partitioned) MSM.

Instead of working in a multidimensional state space sampled by an MD trajectory, one can define a low-dimensional subspace spanned over several "reaction coordinates" capable of capturing the most relevant conformational events of the system. However, finding those "reaction coordinates" is not always a trivial

task. In the context of **Chapter 5** studying a highly complex conformational dynamics of the Formin Binding Protein 21 tandem-WW domains, I proceeded with a two-step protocol for the dimensionality reduction. First, I applied the Principal Component Analysis (PCA) on the set of the backbone atoms coordinates to detect the internal motions relevant for this system. Next, to define the slow subspace of the PCA space capturing the most of the kinetic variance of the system, I performed the Time-lagged Independent Component Analysis (tICA). Finally, this projected representation of the tandem-WW domains free energy landscape was subject to a density-based clustering algorithm, namely Common-Nearest Neighbor (CNN) algorithm to determine the core sets used for the MSM construction. The theoretical framework of the PCA, tICA, and CNN is summarized in the **subsections 2.3.1, 2.3.2** and **2.3.3** respectively.

2.3.1 Principal Component Analysis

Principle component analysis (PCA) introduced by Karl Pearson in 1901 has been successfully applied to analyze MD simulations of the numerous biomolecules. To detect the internal motion of the system, rotational and translational degrees of freedom should be removed from the MD trajectory. Kabsch's method [66] for the rotational fit is implemented in GROMACS package and minimize the least square distance Δ between instantaneous *r(t)* and reference structure $\tilde{r}$:

$$\Delta = \sum_{i=1}^{N} m(r(t) - \tilde{r})^2. \tag{2.29}$$

Now consider a case when the internal motions of the system are represented by a set of mass weighted Cartesian coordinates X=$\{x_1,\ x_2,...,\ x_p\}$. PCA linearly transforms the dataset X of p correlated coordinates into a set of p uncorrelated variables PC=$\left\{z_1,\ z_2,...,\ z_p\right\}$ called principal components, where p equals to 3N (number of the considered atoms of the system). The first principle component is a linear combination (dot product) of the original dataset X with a set of weights α:

$$z_1 = \alpha_1^T X = \alpha_{11}x_1 + \alpha_{12}x_2 + ... + \alpha_{1p}x_p \tag{2.30}$$

The weight are constrained as follows:

$$\alpha_{11}^2 + \alpha_{12}^2 + ... + \alpha_{1p}^2 = 1 \tag{2.31}$$

and determined in such a way that the variance of the original dataset X is maximized. The second principle component z_2=${\alpha_2}^T X$ is estimated to capture

the maximum variance in the dataset X uncorrelated with z_1. Again, it is subject to the same normalization constraint applied on the set of α_2 weights. All other principle components z_3, z_4,...,z_p are determined in the same fashion.

The principal components are calculated following the eigendecomposition of the $p \times p$ covariance matrix σ. The matrix elements of σ are calculated as

$$\sigma_{ij} = \frac{1}{N} \sum_{k=1}^{N} \big(x_i - \langle x_i \rangle\big)\big(x_j - \langle x_j \rangle\big) \qquad i, j = 1, 2, ..., p, \tag{2.32}$$

where $\langle x_i \rangle$,$\langle x_j \rangle$ are averaged i^{th} and j^{th} coordinates respectively. The eigenvectors of the covariance matrix are sorted according the decreasing eigenvalues $\zeta_1 \geq \zeta_2 \geq ... \geq \zeta_p$. Eigenvalues of the covariance matrix describe the degree of the variance of the dataset in the direction of the corresponding eigenvector. The principal components z_i are realized as projections of the original data onto the eigenvectors of σ [67]. By calculating the cumulative variance over all eigenvalues, one can show that first several principal components account for the most important internal motions, while the rest of the principal components entails for the local fluctuations. [68].

2.3.2 Time-lagged Independent Component Analysis

By performing PCA the state space Ω was linearly transformed into a set of n most dominant principal components $PC' = \{z_1, z_2, ..., z_n\}$ capturing most of the conformational variance of the system. However, some of the projected conformational events contained in the PCA subspace, might have huge magnitude, although happening at a fast time scale. As previously discussed, the ultimate goal of the MSM framework is to find the slowest dynamic modes of the investigated system. The PC' set was once again linearly transformed with the Time-lagged Independent Component Analysis into a set of independent components yielding $IC = \{z'_1, z'_2, ..., z'_m\}$ (where m<n). Similarly to PCA, one computes the matrix elements of a time-lagged correlation matrix $\sigma^z_{ij}(\tau)$:

$$\sigma^z_{ij}(\tau) = \langle z_i(t) z_j(t+\tau) \rangle_t \tag{2.33}$$

which can be given in the following form as well:

$$\sigma^z_{ij}(\tau) = \frac{1}{T - \tau - 1} \sum_{t=1}^{T-\tau} z_i(t) z_j(t+\tau) \tag{2.34}$$

where T is the length of the MD trajectory, τ the lag time, and z_i, z_j the principle components for which the correlation function is computed. The independent components should be uncorrelated, while their autocovariance (at the lag-time τ) should be maximized with the constrain $\sigma^z_{ij}(0)=\delta_{ij}$. Analogously to PCA, the eigendecomposition of $\sigma^z_{ij}(\tau)$ and the projection of PC' onto its dominant eigenvectors yields $IC=\{z'_1, z'_2,..., z'_m\}$ capturing the system's degrees of freedom with the greatest kinetic variance . Now, one can work in a slow subspace of the full state space Ω [69].

2.3.3 Common Nearest Neighbors Clustering Algorithm

If an MD trajectory is projected on the slow subspace spanned over the m dominant independent components, to determine the core sets, a density-based clustering algorithm can be applied. Common Nearest Neighbors (CNN) algorithm like any other density-based algorithm assigns a data point x_i of the multivariate data set to a cluster CS_k, if x_i is density-reachable to any of the previously determined members of the cluster CS_k. When applying CNN algorithm one should define a neighborhood parameter R and the number of shared neighbors N. In the CNN algorithm, the neighborhood parameter R is a radius of the cluster CS_k. If a cluster CS_k is initialized at the data point x_j then the data point x_i will be assigned to the same cluster when x_j and x_i share at least N neighbors, while being in each other's neighbors list. This implies that the maximum distance between x_j and x_i points equals to the radius R. Data points not assigned to any cluster are considered as noise points, and in the core set approach, they correspond to the intervening space [65].

2.4 Mutual Information - Computational Approach to Study Allostery

The recent developments in the understanding of the allostery can be significantly attributed to the numerous emerging computational methods which aim to detect the potential allosteric sites or to elucidate the possible pathway(s) of the allosteric signal transduction. Additionally, the increase of the computer power and improved simulations protocols enabled the studies on the dynamical aspects of the allostery. Early computational methods considered the analysis of the protein sequence concerning the evolutionarily conserved residues. The rationale behind such methods is that a set of highly conserved residues (single site methods), or the clusters of residues that coevolved together ("coupled sites

methods"), play a significant role in allostery. However, they cannot distinguish between the multiple roles that conserved residues may assume within a protein structure [70].

Protein allostery is tightly connected with the intrinsic dynamics. An allosteric site communicates with an orthosteric site through substantial conformational changes induced by the binding of the allosteric modulator, or through subtle correlated internal motions. While the time scales of the large conformational changes are still rarely reachable by unbiased MD simulations, the fast internal motions especially the oscillation of the sidechains can be well sampled by the all-atom MD simulations. The protein dynamics can be visualized as a network utilized in subsequent allosteric network analysis. Residues are represented by the nodes in such a network, whereas the edge thickness connecting a specific residue pair is proportional to the degree of interdependence in their motion [21].

A common statistical metric of the correlation implemented in various sequence-based and dynamical network analysis methods is the Mutual Information (MI). Mutual Information of two continuous random variables X, Y or the information content conveyed about X by Y is given as follows:

$$MI(X,Y) = \int_X \int_Y p(X,Y) \log(\frac{p(X,Y)}{p(X)p(Y)}) \mathrm{d}X \mathrm{d}Y \tag{2.35}$$

where *p(X,Y)* is the joint distribution, and *p(X)*, *p(Y)* are marginal distributions. By definition, *MI* is 0 only and only if the product of marginals distributions equals to the joint distribution. As *MI* is not restricted to a certain interval, one should normalize *MI* with the respect to the Shannon information entropy:

$$NMI(X,Y) = \frac{MI(X,Y)}{min(H(X), H(Y))} \tag{2.36}$$

Shanon entropy of a random continuous variable X is defined as :

$$H(X) = -\int_X p(X) \log(p(X)) \mathrm{d}X \tag{2.37}$$

If two random variables X, Y assume *NMI* value of 0, they are said to be mutually independent. In contrast, if *NMI* equals 1, then the random variable X is fully determined by variable Y [71].

Several implementations of the MI have been reported elsewhere [72, 73]. The main advantage of MI application in the context of protein allostery is that this correlation measure can account for the non-linear correlations, which is often

the case when MI is applied to the set of internal coordinates such as dihedral angles [74].

2.5 Protein-Protein Docking

Molecular docking is a computational method that aims at predicting the ligand-receptor complex structure. Traditionally, a protein target has been treated as a rigid body in docking calculations. Molecular docking comprises two steps. First, the conformational space of the ligand in the active site is sampled by a search algorithm (molecular mechanics, Monte Carlo sampling). In a later step, binding conformations (poses) are ranked according to an implemented scoring function. Ideally, the best-scored poses should resemble the experimentally determined binding poses [75].

The computational methods for studying PPI can be classified into two categories. The first category comprises the data-driven methods utilizing a machine learning algorithm to mine the interactome databases (databases of known protein-protein interactions). Such methods are complementary to protein-protein docking methods, as they provide the input on the interacting residues for the docking algorithm [76]. A good protein-protein docking software should address the dynamical nature of the binding partners, at least at the level of the binding interface [77].

HADDOCK protein-protein docking protocol employs the NMR observables such as the chemical shift perturbation (CSP) or the NOE distances to guide docking calculations aiming at producing the high quality (near-native) complexes of the interacting partners. For instance, the residues exhibiting significant CSP upon binding are set as active and treated as flexible during the docking run. The rest of the surrounding residues are set to be passive, hence rigid. Information on the interacting residues is supplied in the form of ambiguous interacting restrains (AIR). AIR is an ambiguous intermolecular distance $d_{i,AB}$ calculated between any atom m of an active residue i of protein A (m_{iA}) and any atom n of both active and passive residues k (N_{res} in the protein B (n_{kB})). The effective distance is given according to:

$$d_{iAB}^{eff} = \left(\sum_{m_{i,A}=1}^{N_{atoms}} \sum_{k=1}^{N_{res}B} \sum_{n_{kB}=1}^{N_{atoms}} \frac{1}{d_{m_{iA}n_{kB}}^{6}} \right)^{(-1/6)} \tag{2.38}$$

,where N_{atoms} denotes all atoms of an active residue i in either of the binding partners, while N_{res} represents th sum of active and passive residues. $d_{i,AB}^{eff}$ is

limited to a maximum value of 3 Å, and is scaled with -$\frac{1}{6}$ power to mimic the London term of the Lenard-Jones potential.

The docking protocol follows a three-step procedure. First, two proteins are separated by the 150 Å distance and randomly rotate around the respective center of mass. During the rigid body energy minimization the translational degrees of freedom are also allowed resulting typically in 1000 different docking complexes which are evaluated in terms of the HADDOCK scoring function. The HADDOCK scoring function (HSF) is realized as a sum of:

$$HSF = E_{vdw} + E_{elec} + E_{desolv} + E_{BSA} + E_{AIR} \tag{2.39}$$

where E_{vdW}, and E_{elec} describe the non-bonded interactions, E_{desolv} is an empirical desolvation constant, E_{BSA} is the buried surface area upon complex formation in Å and E_{AIR} is the restraint violation energy [78]. Then, the best 200 conformations are further optimized through a simulated annealing MD run. Lastly, the docking complex is solvated with TIP3P water molecules, and subjected to a short MD run. The pair-wise backbone RMSD at the interface (iRMSD) is used as a similarity measure to cluster docking solutions. All structures, for which an iRMSD < 1.0 Å is reported, are assigned to the same cluster. The resulting clusters are again evaluated in terms of HADDOCK scoring function [79, 80].

Chapter 3

Intradomain Allosteric Network Modulates Calcium Affinity of the C-Type Lectin Receptor Langerin

Lectins are ubiquitous proteins capable of binding a soluble carbohydrate or a carbohydrate component of a glycoprotein or a glycolipid [81]. Members of the C-type lectin (CTL) family requires a Ca^{2+} cofactor to recognize a sugar moiety. CTL are oligomeric receptors that mediate cell to cell adhesion and apoptosis. They are also an integral part of the innate immune system [82].

Langerin is a trimeric C-type lectin receptor (CTLR) expressed on the surface of the Langerhans cells. It triggers the immunological responses towards invading pathogens such as HIV, influenza virus, Mycobacterium spp., or different types of fungi [83]. Each of three identical monomers comprises a carbohydrate recognition domain (CRD), an α-helical tail domain, and a transmembrane domain. The CRDs and α-helical tails of all three monomers form a so-called extracellular domain (ECD) [84]. The CRD bears a single Ca^{2+}-site, which comprises E285 and N287 of a highly conserved EPN motif (comprising residues E285, P286, and N287) of the long loop, and adjacent residues E293, N307 and D308. The EPN motif determines Langerin specificity for mannose-based oligosaccharides [85]. Recently, Munoz-Garcia *et al.* reported that Langerin trimer binds glycosaminoglycans through an interdomain interface pointing at the importance of Langerin in the cell to cell adhesion [86].

Prior to this study, a little was known about how the sugar cargo is processed once it is attached to Langerin and internalized to a dendritic cell. Therefore, we were interested in elucidating the molecular mechanism(s) facilitating the extracellular sugar uptake and intracellular release in the acid environment of an early endosome. To gain insights into these processes, we combined complementary

techniques, nuclear magnetic resonance (NMR), isothermal titration calorimetry (ITC), and molecular dynamics (MD) simulations.

I set up and analyzed the molecular dynamics simulations. Then, I implemented Mutual Information as referred in **section 2.4**. I generated Figure 3 (panels A-D), Figure 4 and Figure 6 (panels B, C, E, F) in addition to the Supporting Information Figures (S7, S9, S10, S15, S16). Last, I participated in the manuscript preparation together with Dr. Jonas Hanske, Dr. Christoph Rademacher, and Prof. Dr. Bettina Keller.

We proposed a novel mechanism of fine-tuning Ca^{2+} achieved through a large and robust allosteric network that coupled the dynamics of the adjacent structural elements, namely the long and short loops. Ca^{2+} binding is a pH-sensitive process. The drop in Ca^{2+} affinity is paired with the drop in pH occurring in the early endosome. We demonstrated that pH sensitivity is due to two pH sensors: (i) a central hub residue H294 and (ii) an unknown pH sensor probably located in the Ca^{2+}-binding site itself. Then, we excluded a possibility that allosteric signal was transmitted downstream the receptor and the allosteric network was a part of a broader signaling pathway. Finally, our NMR relaxation experiments and MD simulations revealed that Langerin was rather rigid. The loops coupling takes place at the nanosecond timescales. The loss of Ca^{2+} cofactor triggers the structural rearrangements of the highly conserved proline residue P286 of the EPN motif. The latter conformational event occurs in the micro- to millisecond regime. It is an intrinsic feature of the apo Langerin structure independent of the pH changes and not under allosteric control. However, at the steady Ca^{2+} concentration of the extracellular matrix Langerin is completely shifted to the holo cis conformer.

https://doi.org/10.1021/jacs.6b05458

Chapter 4

Exploring Rigid Core and Flexible Core Trivalent Sialosides for Influenza Virus Inhibition

4.1 Introduction

Hemagglutinin (HA) is a glycoprotein expressed on a coat of the influenza virus. It comprises of three identical monomers, each having a globular domain and α-helical tail that docks HA to the viral coat. HA interacts with the sialic acid residues (SA) on the membrane of respiratory cells and erythrocytes. An SA-binding site resides in the globular domain of a HA monomer, pointing outwards and being solvent exposed. Two SA sites of a HA trimer (HA_3) are separated by approximately 4.3 nm, while two adjacent copies of HA_3 are 10-12 nm distant from each other [87]. One can apply these spatial constraints to design an SA-based trivalent ligand and achieve inhibition of the influenza virus with high affinity. Several approaches for the SA-based multivalent ligand design have been proposed in the literature [88, 89].

The aim of this study was to design and synthesize a trivalent sialoside that could bind with the HA_3 and influenza virus with high binding affinity. I collaborated with Pallavi Kiran, and Dr. Sumati Bhatia (Macromolecular Chemistry Group led by Prof. Dr. Rainer Haag), Dr. Susanne Liese (Bio-soft Matter Theory Group led by Prof. Dr. Roland Netz) and Dr. Daniel Lauster (Molecular Biophysics Group led by Prof. Dr. Andreas Hermann).

Pallavi Kiran and Dr. Sumati Bhatia selected polyethylene glycol (PEG) as a spacer due to its high water solubility and biocompatibility. Considering the geometry of HA_3, Dr. Susanne Liese proposed two paths that connect the core of HA_3 and respective sialic acid binding sites. The first path has a length of 3.06 nm and goes through a "valley" between two adjacent HA monomers. The second

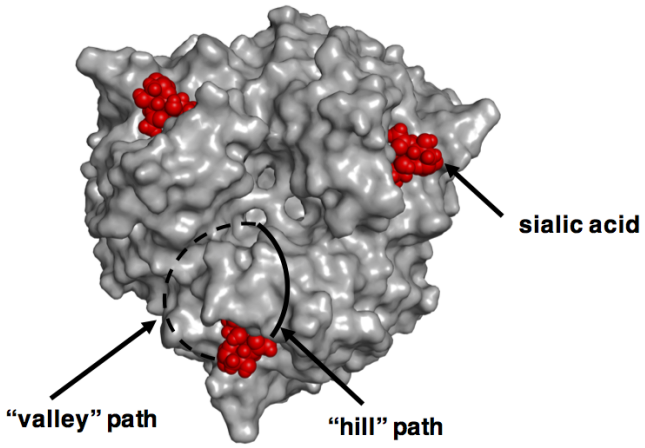


FIGURE 4.1: Three HA monomers are shown in grey, while the respective SA-sites are shown in red. Possible paths that connect the SA-binding site with the center of the trimer are indicated with black lines (the "valley" path with a dashed line, the "hill" path with a bold line.

path goes over a "hill" and is 3.35 nm long (Figure 4.1). The rigid segment of a trivalent ligand has a length of approximately 1.2 nm, thus leaving the PEG spacer to bridge the distance of 1.86 nm or 2.15 nm, depending on the assumed path. Based on the findings of Liese *et al.*, the PEG spacer should comprise at least six to 14 ethylene-glycol units [90]. Hence, Pallavi Kiran prepared several trivalent ligands containing either a rigid adamantane core or a flexible TRIS core. Dr. Daniel Lauster measured the affinity of those tripods to bind to a HA trimer in a hemagglutination inhibition assay and to inhibit the influenza virus in a microscale thermophoresis assay. As indicated in Table 4.1 only compound 10 had the inhibition constant in the micromolar range.

$(EG)_n$, n=	core	HAI (K_i)	MST (K_d (μM))
	α-methylsialoside	2 mM	2800 ± 300
	2,6-sialyl lactose	100 mM	ND
1	Ada	>100 mM	ND
6	Ada (compound 10)	97 μM	113 (1mM)
14	Ada (compound 11)	>100 mM	ND
1	Tris	>100 mM	ND
4	Tris	20 mM	ND
7	Tris	>100 mM	ND

TABLE 4.1: Tripods affinity data.

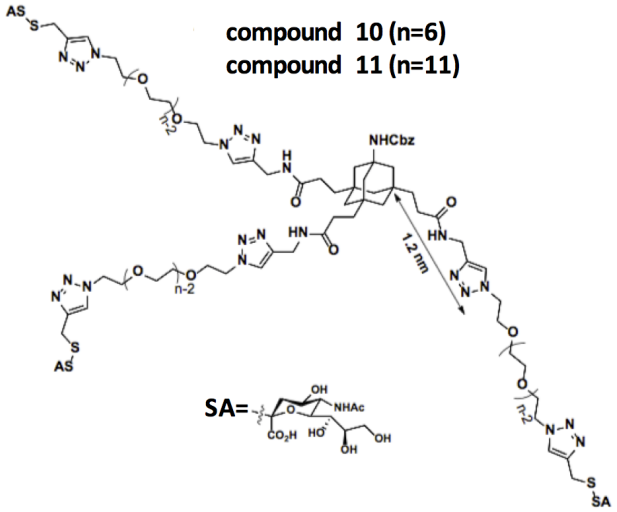


FIGURE 4.2: Schematic representation of the tripod constructs.

To answer the question why only compound 10 exhibited the desirable potency, I set up molecular dynamics simulations of the compounds **10** and **11** with the PEG spacer and the adamantane core (Figure 4.2).

4.2 Methods

4.2.1 Molecular Dynamics Simulations Set-up

All-atom molecular dynamics (MD) simulations were performed for the monovalent counterparts of compounds **10** and **11** (later referred as ligands **10** and **11** respectively) in explicit water (TIP3P water model [91]), using the GROMACS 5.0.2 simulation package [92]. The initial structures of the ligands were drawn in Marwin Sketch [93]. Both ligands were parametrized in Acpype [94]. The topologies of the ligands **10** and **11** were generated according to the General Amber Force Field (GAFF) [95]. The semi-empirical quantum chemistry programme SQM was used to assign the partial charges with AM1-BC level of theory [96]. The systems were minimised in vacuum with the steepest decent algorithm [97] (emtol = 1000.0 (kJ/mol)/nm, nsteps = 50000). 4069 and 14441 water molecules were added to solvate the ligands 10 and 11 respectively in a dodecahedron box (volume of the ligand **10** box 130.51 nm3, volume of the ligand **11** box 443.76 nm3). Then solvated ligands were brought to another round of minimization with the same parameters, followed by two equilibration runs, first in the NVT ensemble at 300 K (V-rescale thermostat [98], time constant = 0.1 ps), and then in NPT ensemble (Parrinello-Rahman barostat [99], reference

pressure = 1 bar, time constant = 2 ps) for 100 ps respectively. Five starting structures per ligand were randomly taken from the NPT equilibration run, and used later in the subsequent MD runs. The production runs were simulated in the NPT ensemble (temperature 300 K, pressure 1 bar). The covalent bonds to all hydrogen atoms were constrained with the LINCS algorithm [42] (lincs iter = 1, lincs order = 4), allowing for an integration time step of 2 fs. Newton's equations of motion were integrated with the leap-frog scheme. The cut-off for Lennard-Jones interactions was set to 1 nm. The electrostatic interactions were treated with the Particle-Mesh Ewald (PME) algorithm [47] a real space cut off 1 nm, a Fourier grid spacing of 0.16 nm, and an interpolation order of 4. Periodic boundary conditions were applied in all three dimensions. The solute coordinates were written to the trajectory file every 1 ps. In total, 1 μs of the simulations were obtained for each ligand.

4.2.2 Molecular Dynamics Data Analysis

The respective distances used to characterize ligands **10** and **11** were obtained with the GROMACS command **g_mindist**. Then, they were visualized with an in-house MATLAB [100] script. The intramolecular hydrogen bond networks were computed with GROMACS command **g_hond** and the hydrogen bond occupancy was calculated with an in-house Python [101] script.

4.3 Results and Discussion

In a preliminary study conducted on the similar set of trivalent ligands (containing the adamantane core and the PEG linker), I found that all three PEG arms collapsed shortly after the start of MD simulations. Each arm formed an ensemble of the coil-like structures with the same average end-to-end distance measured between the outermost oxygen atoms. Hence, I simulated only monovalent counterparts of ligands **10** and **11** solvated in water. To trace the conformational behavior of simulated ligands, I measured the distance distributions for several atom pairs as indicated in Figure 4.3. First, the distance distribution of the C_8, N_5 atom pair lays in the range of 0.7 nm to 0.9 nm for the most of simulation time (1 μs) in both ligands. Nonetheless, the overlap of both distributions is expected, as the planarity of the peptide bond and adjacent triazole ring provide the limited flexibility of this building block. Thus, the composition in the rest of the molecule does not influence the dynamics of the core. Second, the distance distributions of the C_8, S_1 atom pair correspond to the formation of PEG coils.

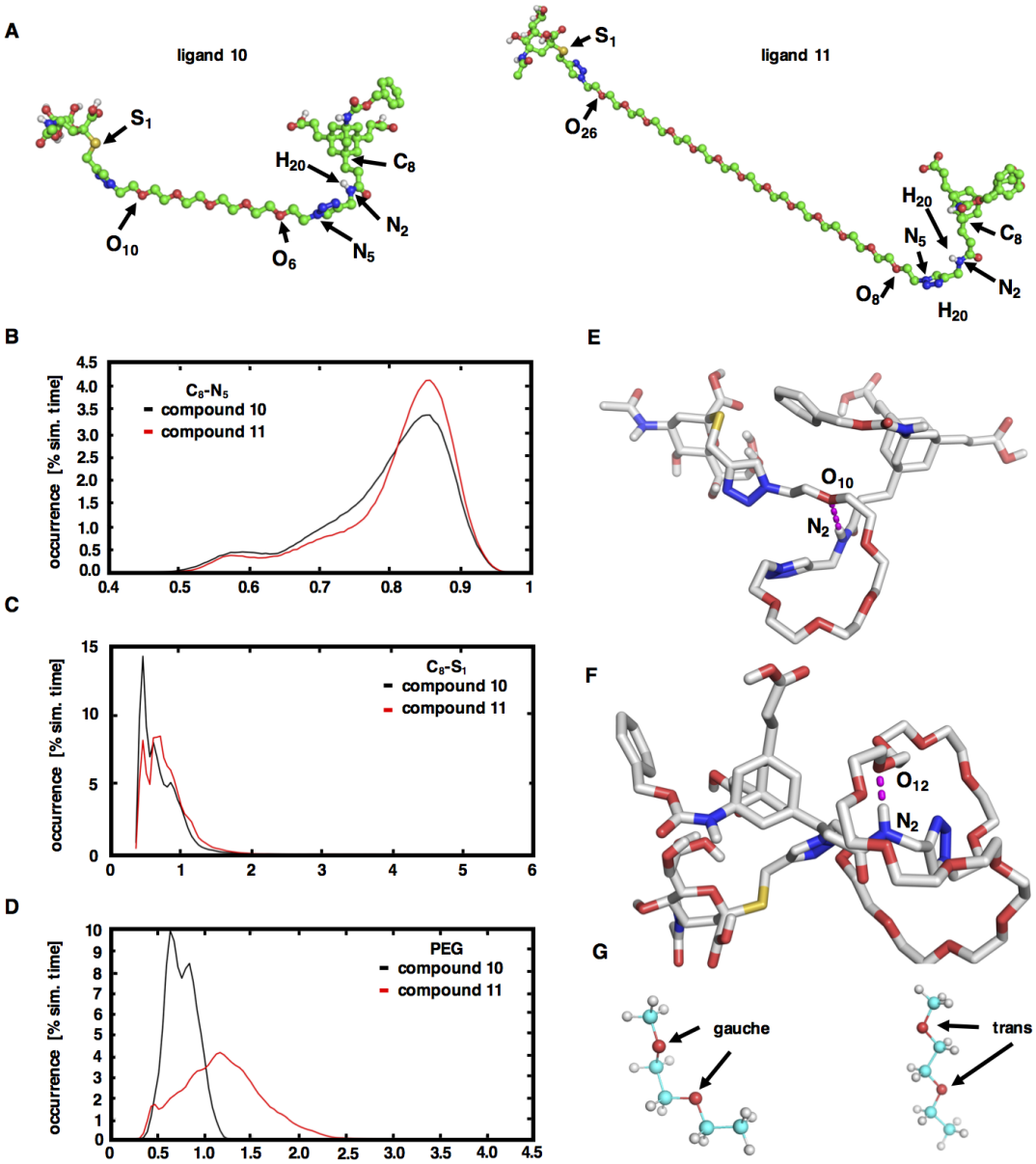


FIGURE 4.3: (A) Pymol representations of the monovalent counterparts of ligand **10** (left panel) and ligand **11** (right panel). The atoms used to characterize the conformational dynamics of the ligands are marked according to the Acpype parametrization. (B) The distance distribution of the C_8-N_5 atom pair capturing the dynamics of the rigid part of the ligands. (C) The distance distribution of the C_8-S_1 atom pair capturing the distance between the rigid core and sialic acid. (D) The distance distribution of the outermost oxygen atoms of PEG spacers. (E) Pymol representation of the most stable intramolecular hydrogen bond of ligand **10**. (F) Pymol representation of the most stable intramolecular hydrogen bond of ligand **11**. (G) As indicated in [90] a PEG linker can adopt two conformations similar to alkanes: gauche (left panel) and trans (right panel).

Both distributions are centered below 1 nm suggesting that PEG arms in the absence of receptor do not stretch enough to reach the range estimated to facilitate binding. Surprisingly, they cover also distances lower than the C_8-N_5 distance (<0.5 nm). In such conformations, I found the sialic acid moiety underneath the adamantane core, which is sterically impossible for the adjacent triazole ring. Last, the distance distribution of the outermost oxygen atoms of PEG spacers shows that on average PEG arms sample the longer distances than the distances between the adamantane core and sialic acid. This observation was somehow confusing and triggered the question if there was an attractive interaction between the rigid core and the sialic acid moiety.

To test this hypothesis, I analyzed the formation of the intramolecular hydrogen bonds between all the pairs of hydrogen bond donors and acceptors. My analysis revealed that sialic acid remained free to potentially interact with the respective binding site on the surface of HA. In both simulated systems, the bonded atom pair N_2-H_{20} was the most dominant hydrogen bond donor. It brings the PEG spacer and triazole ring adjacent to the sialic acid in the vicinity of the adamantane core through the formation of several transient hydrogen bonds. In case of compound **10**, N_2-H_{20} donor participates in the intramolecular hydrogen bonding for 34.5% of the simulation time, whereas in case of compound **11** it acts as the donor for the half of the simulation time. As shown in Figures 4.3E and 4.3F, the formation of the most stable hydrogen bond between N_2-H_{20} and a PEG oxygen atom resulted in all PEG oxygens of the linker to assume the gauche conformation (Figure 4G). It is in line with findings in the reference [90]. As Liese *et al.* showed, a gauche conformation of PEG oxygens allows for the surrounding water molecules to be doubly coordinated with the PEG backbone. Taken together the intramolecular hydrogen bond network and the water shell around PEG linker discussed by Liese *et al.*, one should expect increased conformational and desolvation entropic penalties for ligand **11** compared to ligand **10**. Consequently, ligand **10** unfolds from the coil-like structure easier than ligand **11** leading to μM inhibitory constant.

Chapter 5

Structural Basis for Recognition of a Bivalent Proline-rich Sequence (SmB_2) by FBP21 tandem-WW Domains

5.1 Introduction

Proteins are complex machineries that usually requires binding partners to fulfill their biological functions. Many cellular pathways such as cell to cell communication, signal transduction, or transcription are executed through protein-protein interactions. Protein-protein interactions (PPI) are the physical contacts formed between two or more proteins through van der Waals and electrostatic interactions and maintained through a big (1000-6000 Å), and solvent exposed surface [20]. A particular subset of residues found in the interaction surface contributes significantly to the binding affinity. Those residues are referred as "hot spots" [33]. There is a tendency for such residues to be organized in local clusters, more than to be evenly distributed across the binding surface. Several classifications of the PPI have been proposed in the literature and are discussed in detail in **section 1.3**.

Eukaryotic genes are usually transcribed in the form of a precursor mRNAs (pre-mRNAs), which are processed to mRNAs during splicing. Noncoding sequences (introns) are removed from the pre-mRNA, and coding sequences (exons) are ligated to form mRNA later transcribed to a functional protein. Splicing takes place in a ribonucleoprotein (RNP) complex called spliceosome. The spliceosome is a highly dynamical assembly. The final composition of the spliceosome is determined by the type of the introns to be spliced and comprised of the small nuclear ribonucleoproteins (snRNPs) and numerous non-snRNP proteins. The

building blocks of a snRNP are a molecule of RNA (for U4/U6 two molecules), a set of seven Sm proteins (B/B', D3, D2, D1, E, F, and G) and a variable number of spliceosome-specific proteins [102]. The splicing factors or the protein components of the spliceosome maintain the structure of the snRNPs and fine-tune the splicing activity by interacting with other spliceosomal components [103].

A prominent example of PPI in the eukaryotic proteome is the recognition of the proline-rich sequences (PRS) by the spliceosomal proteins. PRS are typically 5-10 residues long peptide sequences. There are 3-6 proline residues in the so-called "core motif," hence, the name proline-rich sequence. PRS play an active role in the spliceosome formation through interaction with the proline-rich sequence recognition domains (PRD) such as GYF, and WW domains [104]. The architecture of a PRS recognition domain is characterized by a shallow binding groove containing a cluster of solvent-exposed aromatic residues, which form hydrophobic but unspecific contacts to the conserved prolines in the PRS core motif [105]. Fine-tuning in both affinity, and specificity is achieved through additional interactions between the flanking residues of the core motif and a WW domain. In a bound state, PRS assume almost an ideal left-handed polyproline II (PPII) helix. Known PRS have a C_3 rotational pseudo-symmetry about the helical axis, with three residues per turn leading to PxxPxxP or PxPPxPP consensus sequence of the core motif. Additionally, PPII helix has C_2 rotational pseudo-symmetry about the axis perpendicular to the helical axis. These features results in the ability of a PRS to bind in both N- to C-, and C- to N- directions, as the backbone, and the side chains of the core motif have the same position upon rotation of 180° [106].

Formin binding protein 21 (FBP21) is a spliceosomal protein, which is a binding partner of the various splicing factors. It recognizes the PRS of the core splicing factor SmB/SmB' with its tandem-WW (t-WW) domains. The two WW domains are bridged with a highly flexible linker (Figure 5.1). Whereas two domains themselves are rather rigid, the overall flexibility of the t-WW domains is solely determined by the conformational behavior of the unstructured linker [107].

The basis for this study was the previous report on the affinity and specificity of the FBP21 t-WW domains for several PRS with varying number of proline-rich motifs by Klippel and coworkers [108]. They demonstrated that a monovalent peptide (SmB_1) derived from the SmB splicing factor binds with a relatively low affinity ($K_d \approx 300\ \mu M$) to the t-WW domains of FBP21. By the addition of the second proline-rich motif (SmB_2 peptide), the binding affinity of such peptide for

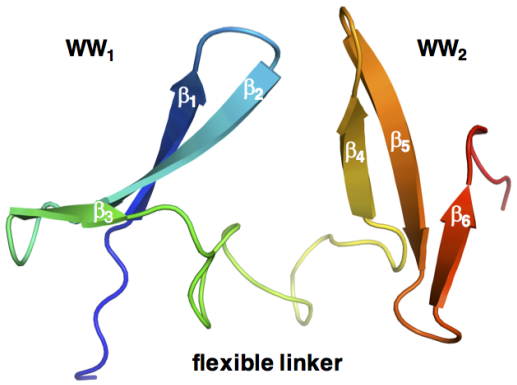


FIGURE 5.1: Three-dimensional architecture of FBP21 t-WW domains (pdb: 2jxw).

t-WW domains increased by more than ten folds instead of the mainly expected factor 2.

We aimed at understanding how the conformational dynamics of the apo t-WW domains influenced the improved binding affinity of the bivalent SmB_2 peptide. To elucidate this process at the atomistic resolution, we combined molecular dynamics (MD) simulations, protein-protein docking, Markov state model analysis, nuclear magnetic resonance (NMR), isothermal titration calorimetry (ITC), and site-directed mutagenesis.

In this project, I collaborated with Miriam Bertazzon and Dr. Jana Sticht from Protein Biochemistry Group led by Prof. Dr. Christian Freund. I set-up and analyzed all MD simulations and protein-protein docking calculations. Miriam Bertazzon prepared all protein constructs of interest and ran NMR and ITC measurements. Together with Dr. Sticht, they analyzed the experimental results.

5.2 Methods

5.2.1 Principal Component Analysis

Principal component analysis (PCA) is a common dimensionality reduction method which captures the dimensions of the largest conformational variance. PCA linearly transforms the initial set of the atomistic coordinates by the diagonalisation of the covariance matrix to a set of linearly uncorrelated coordinates. The elements of the covariance matrix σ_{ij} are given as:

$$\sigma_{i,j} = \langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle \tag{5.1}$$

where $r_i - \langle r_i \rangle$ are protein atomic displacement vectors in the 3N dimensional conformational space, N is the number of atoms in the observed system, r_i is the atomic coordinate vector, and $\langle r_i \rangle$ is the average atomic coordinate vector over an MD trajectory. The diagonalisation of the covariance matrix produces 3N eigenvectors (v_i) describing the modes of the protein dynamics, and 3N eigenvalues (χ_i) describing their respective amplitudes. Eigenvectors are sorted according to decreasing eigenvalues [68]. The slow modes of the protein dynamics are represented by the motion of the backbone atoms. The 76 trajectories of the apo t-WW domains were reduced to the backbone atoms trajectories and fitted to the backbone atoms of the starting structure by **trjconv** GROMACS command. All the trajectories were merged into a single trajectory, prior the computation of the covariance matrix with **g_covar** GROMACS command. Then, the backbone coordinates were projected onto the eigenvectors with **g_anaeig** GROMACS command.

5.2.2 Markov State Models

In recent years, Markov state models (MSMs) have emerged as a powerful tool to understand the complexity of data sampled by MD simulations. MSMs have been successfully employed to study biological phenomena such as the conformational dynamics of biomolecules, protein folding, ligand-receptor interactions, and allostery [23, 55–57]. Prior to construction of an MSM, one assumes that a process sampled by MD simulations is Markovian, ergodic and reversible. In the next step, a proper discretization scheme is selected for the partitioning of the conformational space into a set of microstates, followed by the estimating the transition jump probabilities between (micro)states pairs. Those transitions probabilities are stored in a transition matrix, which is a row stochastic matrix (specified by denominator in Eq. 5.2) and it elements are computed according to:

$$T_{i,j}(\tau) = \frac{C_{i,j}}{\sum_k C_{ik}}. \tag{5.2}$$

where C_{ij} are the matrix elements of a count matrix C containing the counts of transitions between ij pairs of (micro)states. The eigendecomposition of the transition matrix yields the information on the time scales at which slow dynamics modes occur, the long-lived conformations (metastable states), and their hierarchy in the free energy surface [58].

The conformational space of the apo t-WW domains was projected onto the subspace spanned over the first 15 principal components accounting for the 97%

of the conformational variance. Then, this subspace was a subject of the Time-lagged Independent Component Analysis (tICA) to detect the rare events in this projected subspace [69]. Time-lag correlation matrix $\Omega_{i,j}(\tau)$ was computed with pyEMMA Python package [109] according to:

$$\Omega_{i,j}(\tau) = \langle r_i(t) r_j(t+\tau) \rangle_\tau \tag{5.3}$$

for lag time $\tau = 1$ ns. Eigenvectors of the time-lagged covariance matrix Ω represent the linear combination of the slowly decomposing degrees of freedom. The conformational space of t-WW domains further reduced to the tIC space spanned over ten slowest tICs, was then discretized with the core set approach. The metastable regions in the tIC subspace were detected with the Common Nearest Neighbour (CNN) density-based cluster algorithm. The density criterion parameters (in detailed discussed in **section 2.3.3**) like the cluster radius (R) and the minimum number of shared neighbors (N) in CNN algorithms were set to 2.4 and 2 respectively. This parametrization yielded 50 disjoint core sets, 47 of which were dynamically connected and comprised 80.85 % of the conformational space, suggesting a highly complex conformational dynamics of the t-WW domains. The remainder of the conformational space is referred to as intervening space, which can be interpreted as an ensemble of transient and unstable structures. The transitions between the 47 core sets were computed by defining the forward $m_i{}^+(t)$ and backward $m_i{}^-(t)$ milestoning processes. The milestoning processes are defined based on the history ($m_i{}^-(t)$) and future ($m_i{}^+(t)$) of the system and constructed for each core separately ((in detailed discussed in **section 2.3.3**)). The transition matrix of the MSM was given according

$$T(\tau) = M^{-1} P(\tau) \tag{5.4}$$

The matrix elemens of *P(τ)* and M were computed as (time-lagged) correlation function between the forward and backward milestoning processes [65]:

$$P_{i,j}(\tau) = \frac{1}{T-\tau} \sum_{t=0}^{T-\tau} m_i^-(t) m_j^+(t+\tau) \tag{5.5}$$

$$M_{i,j} = \sum_{t=0}^{T} m_i^-(t) m_j^+(t). \tag{5.6}$$

The MSM was validated in terms of the convergence of the implied time scales (μ):

$$\mu = -\frac{\tau}{ln(\lambda_{i,T(\tau)})} \tag{5.7}$$

where $\lambda_{i,T(\tau)}$ (i>1) are the eigenvalues of the transition matrix $T(\tau)$. By considering the eigenspectrum of $T(\tau)$, core sets were then lumped with Perron Cluster Cluster Analysis (PCCA) [61] into long-lived structures separated by the free energy barriers.

5.2.3 Molecular Dynamics Simulations

All-atom molecular dynamics simulations were carried out for the following apo systems: FBP21 t-WW domains (system #1), FBP21 WW1 domain (system #2), and FBP21 WW2 domain (system #3). To study the binding of the SmB_2 peptide (GTPMGMPPPGMRPPPPGMRGLL) to t-WW domains, four SmB_2:t-WW complexes were selected as indicated in Table 5.1 (systems #4-7). The structure of the bivalent ligand SmB_2 was prepared in Chimera [110] molecular modeling software with the **Build Structure tool**. As previously reported by Ball *et al.* [106], PRSs are the linear peptides and form a PPII helical structure due to the presence of a proline-rich motif. All the backbone dihedrals were set to ϕ=-78°, ψ= 146° to maintain a PPII helix. The terminal residues in the systems #1-3 were kept as charged. All the simulations were performed in explicit water (TIP3P water model [91]), using the GROMACS 5.0.2 simulation package [92] and the AMBER ff99SB*-ILDNP force field [111]. From the 2JXW ensemble of NMR structures [107], the third structure was selected, since it lacks inter-domain contacts. The initial structures of the respective WW domains were obtained by saving the coordinates of the residues 1 to 37 for WW_1 domain and of the residues 38 to 75 for WW_2 domain to match the sequence of the expressed singular domain constructs. The initial structures of the bound complexes were the docking poses obtained from the HADDOCK protocol [79]. The systems were minimized in the vacuum with the steepest decent algorithm [97] (emtol = 1000.0 (kJ/mol)/nm, nsteps = 50000), followed by solvation in dodecahedron boxes. Na^+ ions were added to neutralize the simulation boxes (Table 5.1.). The solvation box of the WW_2 was larger than the box of the WW_1 as the linker was in the extended conformation. Solvated systems were brought to another round of minimization with the same parameters, followed by an equilibration in the NVT ensemble at 300 K (V-rescale thermostat [98] time constant = 0.1 ps), and NPT ensemble (Parrinello-Rahman barostat [99], reference pressure = 1 bar, time constant = 2 ps) for 100 ps respectively. The covalent bonds to all hydrogen atoms were constrained with the LINCS algorithm [42] (lincs iter = 1, lincs order = 4), allowing for an integration time step of 2 fs. Newton's equations of motion were integrated with leap-frog scheme. The cut-off for Lennard-Jones interactions was set to 1 nm. The electrostatic interactions were treated with

the Particle-Mesh Ewald (PME) algorithm [47] with a real space cut off of 1 nm, a Fourier grid spacing of 0.16 nm, and an interpolation order of 4. Periodic boundary conditions were applied in all three dimensions. The first set of apo t-WW domains simulations was initialized from 12 structures attained from a short simulation at 350 K (1 ns). Prior to the start of MD runs, the temperature of the system was scaled back to 300 K. Those 12 parallel simulations yielded 8.9 μs of simulation time in total. This MD data set was a subject of core set analysis on the PCA subspace spanned over 15 principle components as described in **sections 5.2.1** and **5.2.2**. Further simulations were started from the structures corresponding to the core sets. An additional round of the core sets detection was performed when 16 μs of the simulation data was gathered. In total, 76 trajectories were produced which varied from 150 ns to 1.1 μs in length, resulting in overall 36.7 μs of the simulation time for the apo t-WW domains. The main aim of these two rounds of the adaptive sampling was to achieve the connectivity between trajectories and to ensure the sampling according to the Boltzmann distribution. The production runs for all investigated systems were simulated in the NPT ensemble (temperature 300 K, pressure 1 bar). For two singular apo constructs, ten trajectories with a total simulation length of 8.9 μs were generated. The holo simulations were directly started after the NPT minimization. For the holo systems trajectories of 100 ns were produced. The solute coordinates were written to the trajectory file every 1 ps.

Ref #	System	Box size [nm^3]	# H_20	Na^+	Simulation time
1	t-WW domains	291.64	9282	7	36.7 μs
2	WW_1 domain	122.59	3808	4	8.9 μs
3	WW_2 domain	197.26	6248	4	8.9 μs
4	$(SmB_2{:}CS_3)_{\text{canonical,1}}$	482.70	15176	5	100 ns
5	$(SmB_2{:}CS_3)_{\text{inverted,1}}$	474.61	14801	5	100 ns
6	$(SmB_2{:}CS_3)_{\text{canonical,2}}$	501.17	15874	5	100 ns
7	$(SmB_2{:}CS_3)_{\text{inverted,2}}$	489.27	15436	5	100 ns

TABLE 5.1: Volume of the simulation box, number of water molecules per simulation box, number of counter ions per simulation box, and total simulation time for each system.

5.2.4 Dihedral Angles Analysis

The flexibility of the protein backbone can be evaluated in terms of the ϕ- and ψ-backbone dihedrals, which jointly form the Ramachandran space of the corresponding residue. The backbone dihedrals of the systems #1, #2, #3 were extracted with the GROMACS command **g_rama**. An in-house developed MATLAB script [100] based on the discretization of the Ramachandran plane into 360 x 360 = 1296000 bins (the bin width of 1°), was employed to project the time series onto the grid. The Ramachandran plots were produced for all the residues of #1, #2, #3. Changes in the backbone dynamics of WW_1 domain simulated as a singular construct (system #2), and in the t-WW domains (system #1) were visualized through pair-wise Ramachandran difference plots. The same analysis was also performed for the #3, #1 system pairs.

5.2.5 Hydrogen Bond Analysis

All trajectories obtained for the respective system were concatenated and then down-sampled with GROMACS command **trjcat** resulting in 20000 equidistant frames for systems #1,#2,#3 (later referred as down-sampled trajectories). Then, they were loaded into VMD [112] for the subsequent hydrogen bond extraction with Hydrogen Bond tool. The donor-acceptor distance was set to 3.5 Å, while hydrogen bond angle was cut off at 30°. An in-house Python-based [101] script analyzed occupancy of the hydrogen bonds detected in the MD trajectories. This script reads VMD output files, filters out the hydrogen bonds with low occupancy (threshold of 10% used), detects the conserved hydrogen bonds present in all analyzed systems (i.e., #2 versus #1 analysis) and creates a list of a system specific hydrogen bonds.

5.2.6 The Chemical Shifts Prediction Based on the MD Simulations

To validate the quality of the MD the chemical shifts were predicted from the downsampled trajectories for the systems #1, #2, and #3 by with the SPARTA$^+$ software [53]. This software predicts the chemical shifts based on the neural network algorithm that considers both structural (backbone and sidechain torsions) and dynamical inputs (S^2-order parameter) with the information on the local interactions (i.e., hydrogen bonds).

5.2.7 DSSP Analysis

DSSP is a computer algorithm which considers the hydrogen bond network and geometrical features to assign every residue in a protein sequence to a specific secondary structure element [113]. To evaluate the stability of the secondary structures in the conformational ensemble, the **compute_dssp** function of the MDtraj [114] Python library was employed to the down-sampled trajectories of #1, #2, #3 systems.

5.2.8 Protein-Protein Docking

HADDOCK [79], a protein-protein docking web server (version 2.2) was employed to study the interactions between t-WW domains and SmB_2 peptide. HADDOCK is a force-field based docking approach. A user can provide experimentally determined restrains referred as ambiguous interaction restraints (AIR), such as chemical shifts, or NOE distances. Then, the user should specify a list of the active residues, usually experimentally observed to be important in the binding process. Nearby residues at the protein-protein interface are treated as passive residues. The HADDOCK docking protocol comprises three steps. First, a contact interface based on the provided AIRs is formed in a rigid body search, followed by the generation of 10000 structures of a protein-protein complex. Then, they are ranked based on a HADDOCK scoring function . This scoring function includes five terms: van der Waals, electrostatic, desolvation, restraint energy, and buried surface area term [80]. The top 200 structures are then refined in the torsion angle space by the simulated annealing. In the last step, those 200 structures are placed in TIP3P water boxes, and further refined in the Cartesian space. Finally, they are re-ranked with HADDOCK scoring function and clustered according to similarity. The decision tree for selecting a binding competent structure employed in the HADDOCK protocol is discussed in **section 5.3.4**.

5.2.9 Contact Maps

To asses if the proline-rich motifs of SmB_2 remained in close contact with the binding grooves of two WW domains, the contact distance matrices for the C-α atoms of the SmB_2:t-WW complex were computed with the **gmx mdmat** GROMACS command for the systems #6 and #7. Then, those matrices were visualized as contact maps with a MATLAB script.

5.3 Results

5.3.1 Matching Molecular Dynamics Simulations and NMR Measurements

I computed the ^{1}H-^{15}N HSQC spectra of the systems #1, #2, #3 from my simulations and compared the results to the experimentally measured chemical shifts (Figure A.1). The computed chemical shifts for the backbone ^{15}N atoms matched the experimental values within an RMSD of 2.64 ppm for the t-WW domains, 2.69 ppm for the WW_1 construct, and 2.60 ppm for the WW_2 construct. The computed chemical shifts of ^{1}H atoms matched to the experimental values with RMSD between 0.35 to 0.43 ppm. The reported accuracy of the chemical shift calculations for the current force fields is 2.45 ppm for ^{15}N and 0.43 ppm for ^{1}H, respectively [54]. Thus, my simulations are in good agreement with the NMR experiments for all three observed systems.

5.3.2 Refolding Does Not Play a Role in the PRS Recognition

A previous study on YES-associated protein 2 (YAP-2) tandem-WW domains involved in the Hippo signaling pathway [115] (control of the tissue growth) showed that the WW_1 domain is partially unfolded in the absence of the proline-rich ligand and refolds during the binding process [116]. I also examined the stability of the singular WW domains and t-WW domains in my simulations. DSSP analysis revealed that the secondary structure was stable in all three systems, in particular, the two anti-parallel β-sheets remained folded throughout the simulations (Figure 5.3). Moreover, the structurally important hydrogen bonds remained stable in the singular domains (Tables A.1, A.2), whereas pair-wise Ramachandran difference plots showed no significant changes of the backbone dynamics for the residues in the all six β-strands (Figures A.2 and A.3). The singular constructs showed a minor increase in the backbone flexibility of the loops connecting the core β-strands compared to the t-WW domains. Thus, MD simulation did not show unfolding in either of the two WW domains in the absence of the ligand. I, therefore, rule out the possibility that refolding of the WW domains plays a role in recognition of a PRS by FBP21 t-WW domains. Instead, the dynamics of the two domains relative to each other is likely to be the dominating factor in the recognition process.

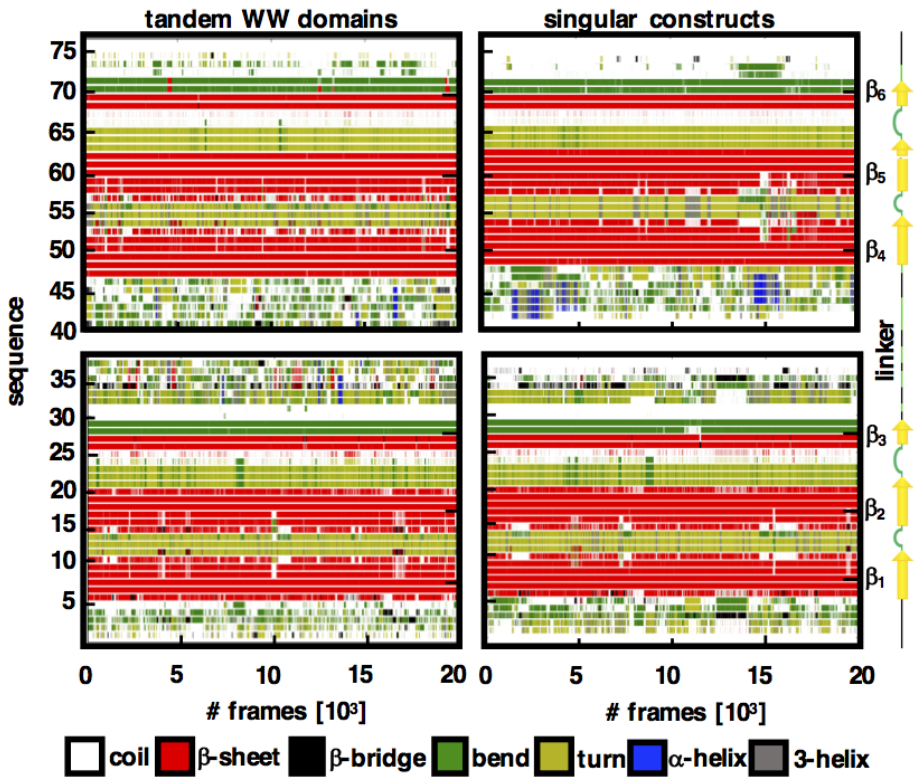


FIGURE 5.2: DSSP analysis performed on the downsampled trajectories (20000 equidistant frames) of t-WW domains (left panel) and the singular domain constructs (right panel) revealed that the β-sheet fold remained stable in three investigated systems.

5.3.3 The Conformational Ensemble of t-WW Domains is Dominated by an Inter-domain Interface Formation

Visual inspection of the MD trajectories of the t-WW domians confirmed that the dynamics of the two domains relative to each other is very complex. They adopt a wide range of relative orientations, some of which are stabilized by the formation of inter-domain contacts. Accordingly, the linker residues were highly flexible and visited all three allowed regions of the Ramachandran space (Figure 5.3.). The linker did not form any stable secondary structure element (Figure 5.2.), which is in line with a previous NMR study on the FBP21 t-WW domains [107].

To obtain a low-dimensional representation of the conformational space sampled by the t-WW construct, first I performed a principal component analysis (PCA) on the Cartesian coordinates of the backbone atoms. I considered the first 15 principal components accounting for 97% of the conformational variance. As I confirmed by visual inspection of the MD trajectories, some of the conformational changes had a large magnitude, yet occurred rather fastly. Then, to distinguish

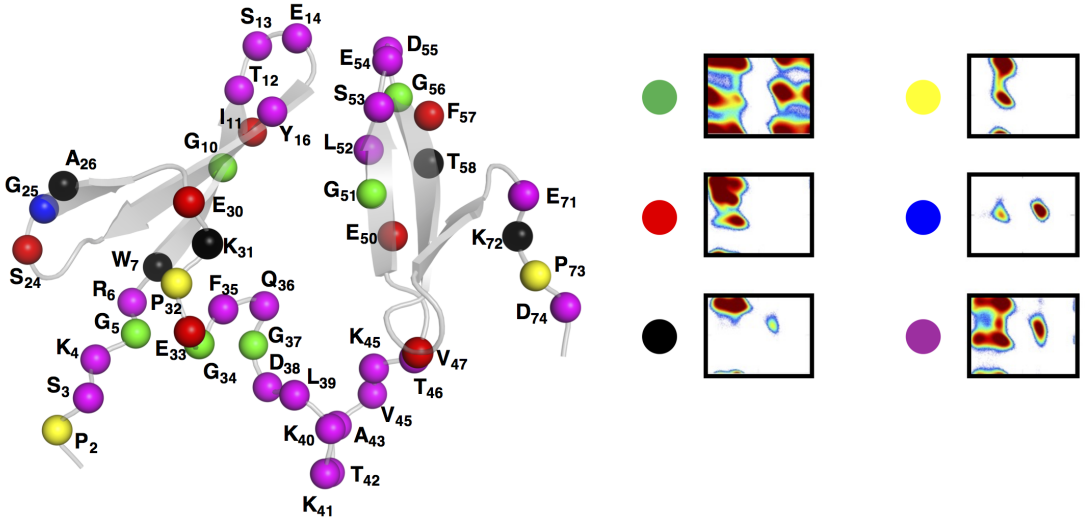


FIGURE 5.3: Backbone flexibility of the t-WW domains; Coloring is done based on the legend depicted in the right-hand side panel.

the rare conformational events from the fast fluctuations, I additionally performed a time-lagged independent component analysis (tICA) on the PCA subspace. Ten time-lagged independent components (tIC) represented 90% of the kinetic variance. Last, I analyzed this 10-dimensional tIC subspace with Common-Nearest Neighbours (CNN) density-based algorithm to detect densely populated regions of the t-WW domains conformational space.

To decide on a representative structure of each core set, I investigated their respective hydrogen bond networks. I filtered all the hydrogen bonds facilitating the β-strand fold of two WW domains and focused on the hydrogen bonds connecting two distinct structural elements (i.e., N- or C- tail and the linker, the linker and any of two WW domain, or two WW domains). Every core set exhibited several unique hydrogen bonds, which I considered when extracting the representative structure (Figure A.4). Interestingly, some of the hydrogen bonds appeared across the multiple core sets (Table A.3). The sidechain of the R_6 is a prominent donor in several semi-conserved salt bridges, which resulted in the WW_1 domain having close contacts with either linker of WW_2 domain. Strikingly, in 43 out 50 core sets, two WW domains were in close contact and formed an inter-domain interface. In the core sets CS_{16}, CS_{19}, CS_{38}, CS_{45}, CS_{48}, and CS_{50} the flexible linker was placed between two domains as dictated by the formation of the unique and the semi-conserved hydrogen bonds only between the linker and two WW domains (Figure 5.4A). In a single core set CS_{39}, the hydrogen bond network comprised contacts between the N- and C- termini and the end residues of the linker exclusively, causing two WW domains to be separated from each

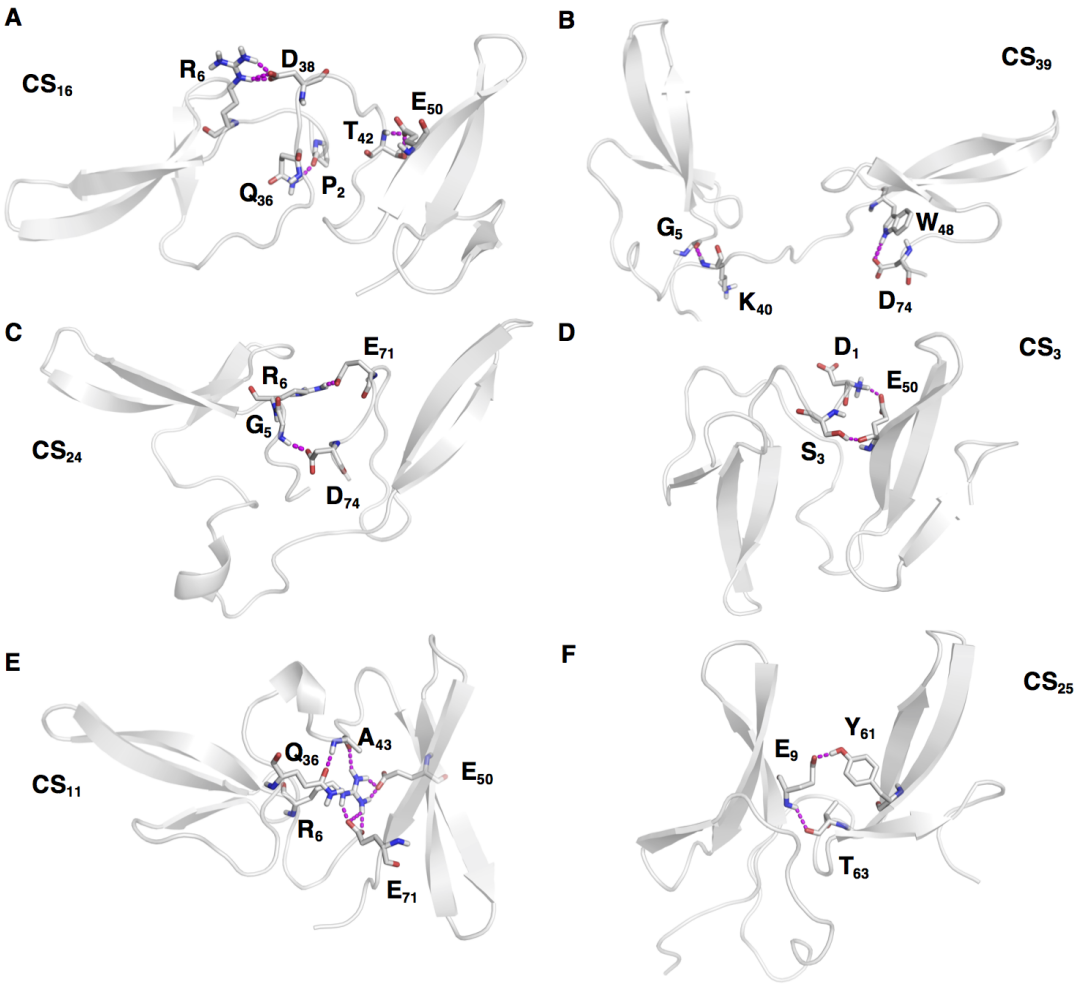


FIGURE 5.4: (A) Linker is placed between two WW domains; (B) No inter-domain interface was formed; (C) inter-domain interface comprising N- and C-terminal residues; (D) N-terminal residues hydrogen bonded with the WW_2 domains residues; (E) highly complex interface - mainly achieved through the polar linker residues and their counterparts in either of two WW domains; (F) interface stabilized by the hydrogen bonds formed between residues of the WW_1 and their counterparts in the WW_2 domains.

other (Figure 5.4B). The inter-domain interfaces varied among 43 core sets and could be grouped in the four categories:

- interface was established between the N- and C-terminal residues: core sets CS_{24}, and CS_{44} (Figure 5.4C);
- interface was formed via N-terminus being placed between two WW domains: core sets CS_3, and CS_{49} (Figure 5.4D);
- majority of the hydrogen bond participating in the interface were formed between the linker residues and their counterparts in either of two WW

domains: core sets CS_7, CS_9, CS_{11}, CS_{12}, CS_{19}, CS_{21}, CS_{28}, CS_{30}, CS_{33}, CS_{34}, CS_{36}, CS_{37}, and CS_{40} (Figure 5.4E)

- interface was achieved mainly through hydrogen bonded residues of WW_1 and WW_2 domains: remaining core sets (Figure 5.4F).

Next, Miriam Bertazzon calculated the chemical shift perturbations by comparing the ^{1}H-^{15}-N HSQC spectra recorded for the singular constructs to those of the t-WW domains. The largest chemical shifts were observed for residues S_3,Q_{36}, and E_{50}. Given that all affected residues are solvent exposed and that the secondary structure remains stable, it is likely that these chemical shifts are caused by inter-domain contacts of the affected residues, in particular by the interface formation.

To categorize the relative orientation and position of two WW domains in the different core sets, I defined four vectors between the C-α atoms of the following residues: V_8, G_{10}, K_{40}, V_{49}, G_{51} (Figure A.4). The angle α between the vectors $V_8 \rightarrow G_{10}$, and $V_{49} \rightarrow G_{51}$ described the orientation of the WW_1-domain relative to the WW_2-domain. The other two vectors, $K_{40} \rightarrow V_8$, and $K_{40} \rightarrow V_{49}$, were centered at the linker residue K_{40}. Their angle β accounts for the relative position of the two domains, where K_{40} was defined as the origin. The results of the vector analysis were summarized in Table A.4. Two WW-domains could adopt three relative orientations to each other, namely perpendicular ($\uparrow \rightarrow$), parallel ($\uparrow \uparrow$) and anti-parallel ($\uparrow \downarrow$), as depicted in Figure 5.5. By far the most dominant orientation of two WW domains observed in the conformational ensemble was perpendicular (25 core sets), then 17 core sets were found in anti-parallel orientation, while only eight core sets assumed parallel orientation. However, there was no strong correlation between four types of the inter-domain interface and the relative orientation of the WW_1, WW_2 domains. However, I assumed that the relative orientation of the two domains to each other might play an important role in determining a binding-competent structure, which is further discussed in **section 5.3.4**.

I proceeded with the investigation of the hierarchy in the free energy surface (FES) of the t-WW domains by constructing a Markov State Model (MSM) using 50 core sets discretization previously discussed. The core set CS_{26} (relative population 0.37%) was not reachable from the rest of conformational space. Next, I checked if the mass matrix M was diagonally dominant to ensure that other 49 core sets were metastable enough. After several rounds of merging the low populated core sets with the neighboring core sets in the respective trajectories, the mass matrix became diagonally dominant, but two core sets CS_{12} (relative

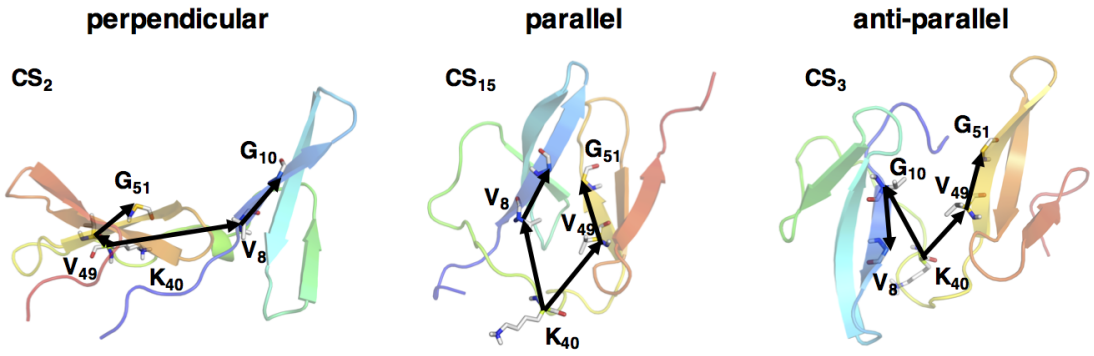


FIGURE 5.5: Two WW domains can adopt three possible orientations respect to each other: perpendicular (left panel), parallel (middle panel, and anti-parallel (right panel).

population 0.84%) and CS_{39} (relative population 0.18%) lost connectivity to the remainder of the conformational space. Then the MSM was constructed on the largest connected set comprising 47 core sets. With the adjusted mass matrix, I computed the transition matrices T for the lag times τ in the range from 100 ps to 70 ns. The implied time scales were rather well converged for the lag times 20 to 70 ns (Figure A.5). I performed the PCCA analysis on the dominant eigenvectors of the transition matrix T (computed for τ=60 ns), yielding six metastable sets (M_1-M_6) separated by the five highest barriers in the FES and corresponding to dynamical modes of the system. The FES of the apo t-WW domains is presented in Figure 5.6. Interestingly, each of the metastable sets M_1, M_2, M_3, M_4, and M_5 corresponded to a single core set CS_2, CS_8, CS_{11}, CS_{10}, and CS_5 respectively. The metastable sets M_1-M_5 shared two features: (i) the perpendicular orientation of two WW domains and (ii) the inter-domain interface. The metastable set M_6 contained remaining 42 core sets considered in the model. In the slowest dynamical process the M_1 set (CS_2, relative population 9.84%) kinetically exchanged with the rest of the conformational ensemble with an implied time scale of 7.4 μs. This minimum of FES was determined by the following hydrogen bonds R_6 $\rightarrow$ G_{56}, R_6 $\rightarrow$ E_{50}, and L_{39} $\rightarrow$ E_{50}. The M_2 (CS_8, relative population 1.83%) set was stabilized by the hydrogen bonds mainly established between N-terminal residues (S_3, K_4) and WW_2 residues (Y_{61}, N_{62}, T_{63}).

It kinetically exchanged with the sets M_3-M_6 at an implied time scale of 5.8 μs. In the set M_3 (CS_{11}, relative population 1.13%), the N- and C-terminal, linker, and WW_2 domain residues participated in the hydrogen bond network stabilizing a highly complex inter-domain interface. The t-WW domains visited any other part of the conformational space covered by the M_4-M_6 metastable sets every 5.3 μs. Next in the hierarchy of the FES was the metastable set M_4 (CS_{10}, relative

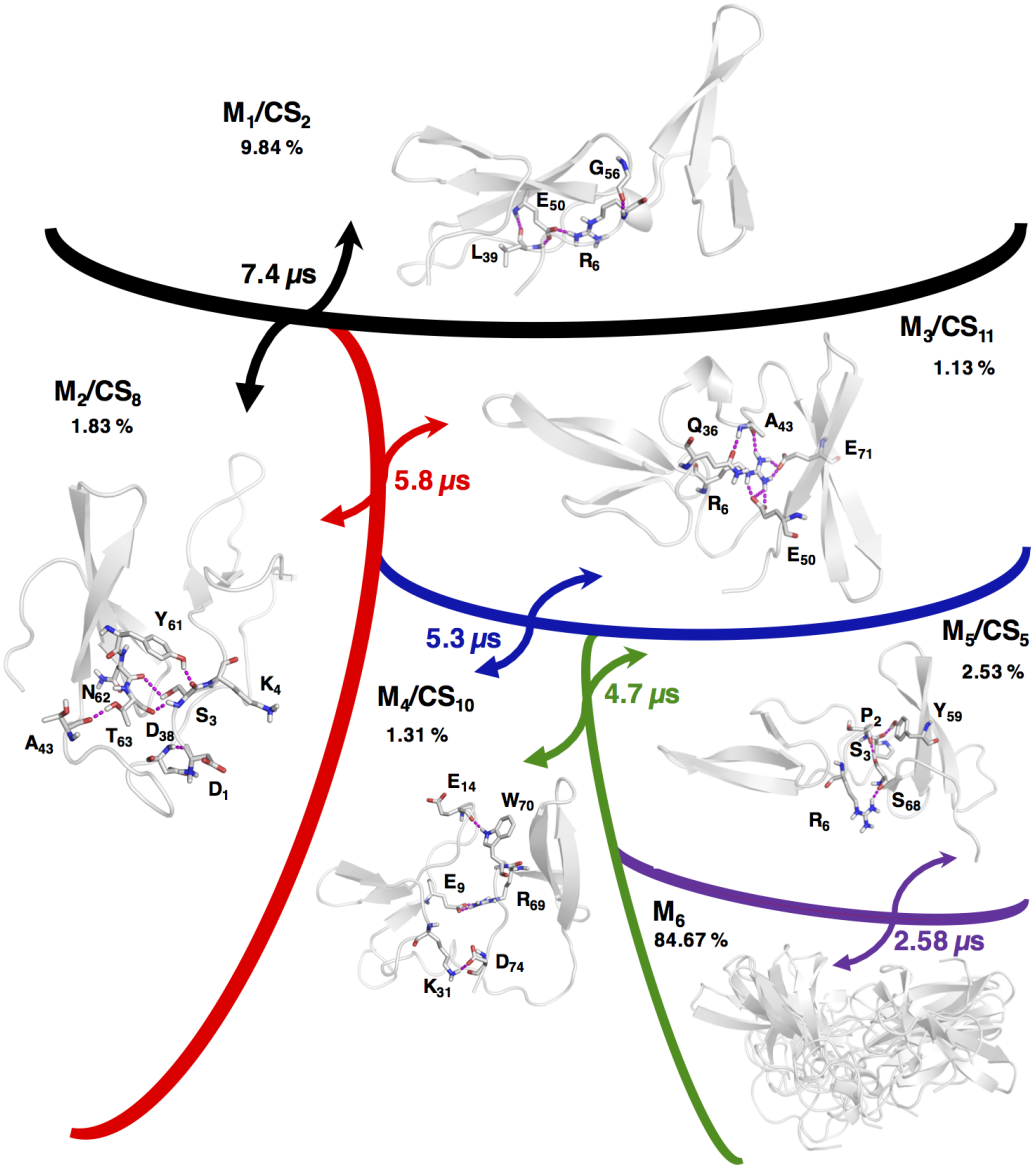


FIGURE 5.6: Kinetic model of the apo t-WW domains.

population 1.31%). In the representative structure of this metastable set M_4 two WW domains were brought in an inter-domain interface through the side chain contacts of R_{69} and W_{70} of the WW_2 with E_9, and E_{14} of the WW_1 domain respectively. The next kinetic process occurred at an implied time scale 4.7 μs and comprised interconversion of the t-WW domains from the metastable set M_4 to either set M_5 or M_6. The last distinguishable metastable set M_5 (CS_5, relative population 2.53%) was determined by an interface formed by the N-terminal residues P_2, S_3 and R_6 bridged with the counterparts of the WW_2 domain Y_{59} and S_{68}. The core sets belonging to the M_6 metastable set were interconverting at faster time scales than 2.58 μs, and the timescales of such conformational changes could not be determined by the current model. Those core sets can be assumed

to belong to a rather shallow, huge and rugged minimum of the free energy surface. Taken together a visual inspection and the hydrogen bond networks of the respective core sets, in addition to the kinetic model, I concluded that the conformational dynamics of the t-WW domains was dominated by the formation of the various inter-domain interfaces.

5.3.4 Binding-competent Structure

I previously mentioned in **chapter 5.1** that WW domains bind a PRS with their shallow and solvent exposed binding grooves (Figure 5.7). The most important recognition element of a WW domain binding groove is a cluster of aromatic residues, which packs against a proline-rich motif of a PRS [106]. In the previous study by Huang *et al.* [107], it was shown that by mutating the tryptophans W_{29} (WW_1 domain), and W_{70} (WW_2 domain), the affinity of the respective WW domain for a PRS was completely abolished. Hence, W_{29} and W_{70} are essential for the binding of a PRS. Additionally, tyrosines adjacent to W_{29} (Y_{18}, Y_{20}), and to W_{70} (Y_{59}, Y_{61}) exhibited relatively strong chemical shift perturbations upon PRS binding [108], suggesting their importance as the recognition elements. As discussed by Ball *et al.* [106], the lack of the intra-molecular hydrogen bonds in a PPII helix, leaves backbone carbonyl oxygen of a proline free to form a hydrogen bond with the side chain of the essential tryptophan of the respective WW domain. This binding pattern was observed in available X-ray and NMR structures [29, 105, 117]. In the absence of the X-ray or NMR structure of a bivalent PRS bound to the t-WW domains, I assumed three conditions that both t-WW domains and SmB_2 should meet to facilitate SmB_2:t-WW complex formation:

(i) all residues of SmB_2 have the fixed backbone dihedrals ϕ=-78°, ψ=146° to maintain the PPII helixal conformation of the peptide (Figure 5.7A). Thus the distance between the backbone carbonyl oxygens of the outermost prolines of two proline-rich motif stood at ∼30 Å, consequently this condition implies that the imidazole nitrogens of W_{29} and W_{70} should be also ∼30 Å apart from each other in a binding competent structure;

(ii) the relative orientation of two WW domains should allow for the simultaneous recognition of both proline-rich motifs of the SmB_2 peptide;

(iii) two aromatic clusters or binding grooves (Y_{18},Y_{20}, W_{29} in the WW_1 domain, and Y_{59},Y_{61}, W_{70} in the WW_2 domain) should be solvent exposed, therefore, not trapped in an inter-domain interface.

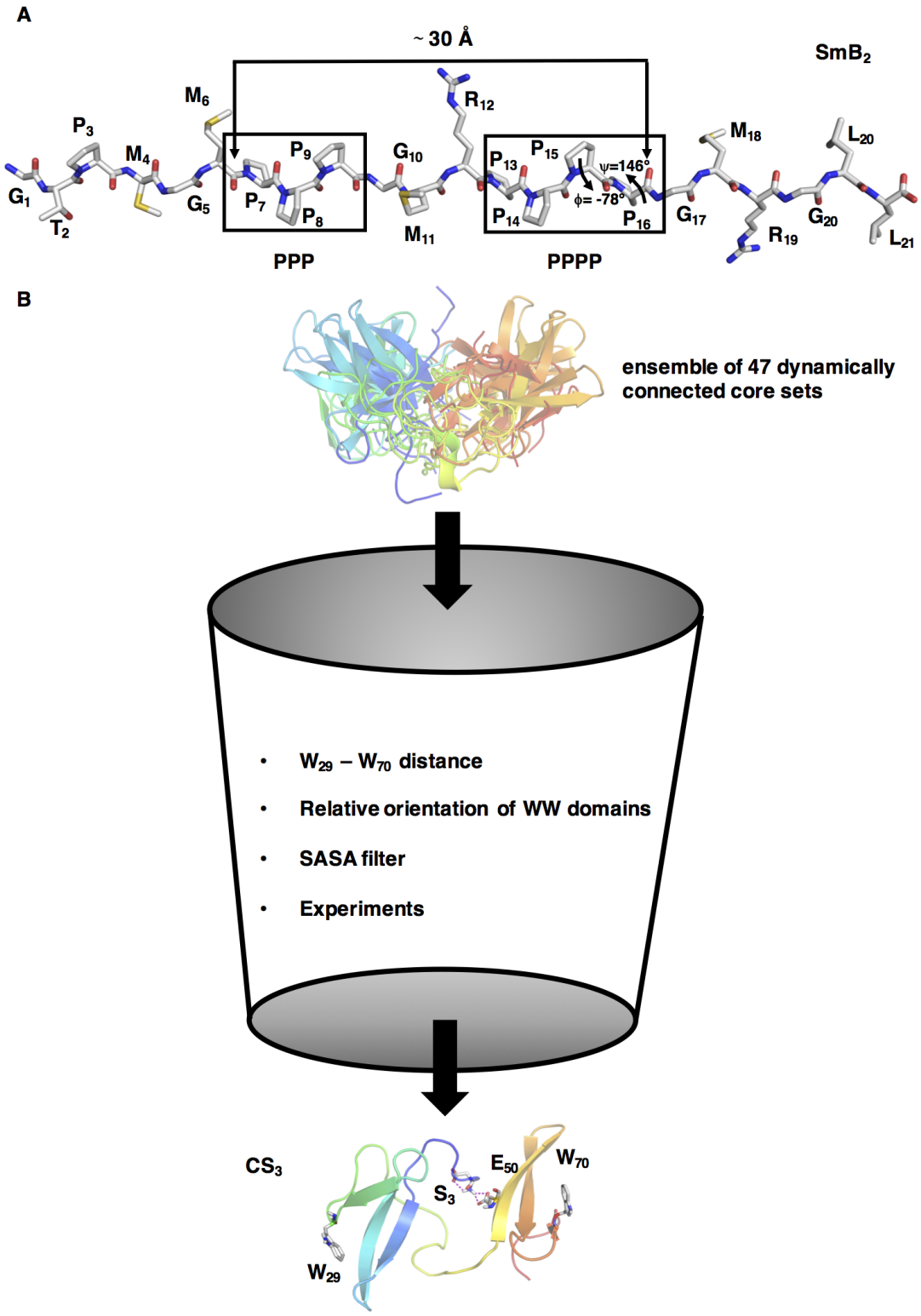


FIGURE 5.7: (A) Pymol representation of the SmB_2 in a PPII helix conformation; (B) Binding-competent structure decision scheme.

With the notion that the conformation which the t-WW domains adopt in the bound state was already present in the simulated conformational ensemble of the unbound state, I next identified core set(s) which would be capable of forming

a complex with the SmB_2. 47 core sets used in the construction of the MSM were considered for the further analysis. By applying the relaxed condition (i) and using the distance between the imidazole nitrogens of W_{29} and W_{70} in the range of ∼25-35 Å, I marked 29 core sets as not binding competent. In the remaining 18 core set, I visually inspected the relative orientations of two WW domains (condition (ii)), resulting in only core sets CS_3, CS_4, CS_{33}, and CS_{41} to be further considered. Last, I excluded the core set CS_{33} as the side chain of Y_{61} was blocked in a hydrogen bond with the side chain of T_{42}, while the core set CS_{41} was filtered out due to Y_{59}, and Y_{61} having low solvent exposure. The reported SASA of Y_{59}, and Y_{61} stood at 19Å, and 12Å respectively (condition (iii)). Furthermore, none of three dynamically disconnected core sets were a potential binding-competent structure according to conditions (i-iii).

system	$K_{D,NMR}$ [μM]	$K_{D,ITC}$ [μM]
wt t-WW	84.17 ± 22.42	41.84 ± 3.45
R6A t-WW	111.55 ± 35.83	45.45 ± 8.79

TABLE 5.2: Affinity measurements

To further support the decision on the binding-competent structure Miriam Bertazzon expressed R6A mutant of the t-WW domains as the side chain of R_6 participated in a salt bridge with the side chain of E_{50} stabilizing the core set CS_4. Then she employed isothermal titration calorimetry and NMR titration experiments to measure the affinity of the wild-type (wt) t-WW, and R6A t-WW domains for the SmB_2 peptide (Table 5.2). Additionally, Miriam recorded a NOESY spectrum of the SmB_2:t-WW complex. This experiment revealed that only the backbone nitrogen atoms of the residues adjacent in the primary sequence were at the NOE distances lower then 5 Å (upper limit of the method), suggesting that two domains were not in a close contact upon binding of the SmB_2 peptide to the wild-type t-WW domains. No NOEs were detected between the SmB_2 peptide and the t-WW domains. This was somehow expected, since Miriam used of SmB_2 peptide construct without the ^{15}-N labeled backbone nitrogens. However, the Smb_2:t-WW complexed was formed as indicated in Table 5.2. Based on the experimental findings, I concluded that CS_4 was not a binding-competent structure. Hence, I used the representative structure of the core sets CS_3 (relative population 5.39%) in the subsequent docking calculations (Figure 5.7B).

5.3.5 Docking and Molecular Dynamics Yields Candidates for the Structure of the SmB_2:t-WW domains Complex

Next, I checked whether a docking protocol would yield suitable complexes of SmB_2 with the selected t-WW domains structure CS_3. I employed NMR-guided protein-protein docking with the HADDOCK protocol, in which one can specify a set of active residues, which side chains conformations are optimized together with the ligand conformation during docking. Residues E_9, G_{10}, Y_{18}, A_{26}, Q_{28}, W_{29}, S_{53}, E_{54}, T_{58}, Y_{60}, R_{69}, W_{70}, and E_{71} were set as active based on the chemical shift perturbation calculations (Miriam Bertazzon). Additionally, three tyrosines Y_{20}, Y_{59}, and Y_{61} were marked as active as they were expected to facilitate packing of the proline-rich motifs of SmB_2. The HADDOCK protocol yielded several docking poses of the SmB_2:t-WW complex. However, visual inspection showed that the HADDOCK score was a poorly suited to compare the various poses. In particular, in some cases, a relatively high HADDOCK score was assigned to poses in which SmB_2 was docked to regions outside of the binding groove. Therefore, I selected interesting docking poses based on the following criteria:

(i) SmB_2 can bind to the t-WW domains in the N- to C- direction (canonical binding mode), and C- to N- direction (inverted binding mode) as indicated by NMR spin labeling experiments [108];

(ii) the core motif of SmB_2 has to be in the proximity of W_{29} and W_{70}. Ideally, it formed an intermolecular hydrogen bond to the side chain of the two residues;

(iii) as HADDOCK yielded docking poses in which the central arginine R_{12} of the SmB_2 was pointing towards either of two WW domains. Hence I considered two distinct R_{12} conformations.

Thus, I obtained four candidates of the potential SmB_2:t-WW complex:

Ref #	SmB_2:t-WW complex	binding direction	R_{12} conformation
4	SmB_2:CS_3	canonical	towards WW_1
5	SmB_2:CS_3	inverted	towards WW_1
6	SmB_2:CS_3	canonical	towards WW_2
7	SmB_2:CS_3	inverted	towards WW_2

TABLE 5.3: Docking poses probed by the MD simulations.

In the system #4 SmB_2 was bound to the t-WW domains in a canonical direction (Figure A.6). The sidechain of W_{29} formed a hydrogen bond with the backbone carbonyl oxygen of M_6 of the Smb_2 preceding the P_{7-9} proline-rich motif. Both proline-rich motifs were packed against the respective aromatic cluster, albeit a hydrogen bond with the sidechain of W_{70} was missing. The SmB_2:t-WW complex was additionally stabilized by a salt bridge between E_{71} and R_{19}. Again in the system #5, when SmB_2 was bound to the t-WW domains in the inverted mode, W_{29} established a hydrogen bond with the SmB_2 (carbonyl oxygen of G_{17}), while the side chain of W_{70} did not participate in any hydrogen bond formation (Figure A.6). Similarly, in the systems #6 and #7 the side chain of W_{70} did not establish any hydrogen bond with the proline-rich motifs of the SmB_2, although it was placed in their proximity. The P_{7-9} proline-rich motif was attached to the WW_1 domain through hydrogen bonds with the side chains of S_{27} and W_{29}, while the side chain of the central R_{12} established contacts with β_5-β_6 loop residues T_{63} and E_{64} in the system #6 (Figure 5.8A). The P_{7-9} proline-rich motif in the system #7 formed two hydrogen bonds with the side chain of S_{68} of the WW_2 domain, while the backbone of M_{18} following the P_{13-16} proline-rich motif made a hydrogen bond with the side chain of W_{29} (Figure 5.8A).

To investigate the role of the R_{12} in the formation of the SmB2:t-WW complex I initialized the MD simulations from the selected docking poses (Table 5.3). Furthermore, I wondered if the #4, #6 and #5, #7 respectively would converge to a single complex per binding direction. Interestingly, in the systems #4 and #5 (Figure A.6), the initial orientation of the R_{12} pointing towards WW_1 domain led to overall instability of the SmB2:t-WW complex. In both system, the proline-rich motif of SmB_2 in the vicinity of the WW_2 domain, actually detached from the WW_2 domain, leaving systems #6 and #7 as the potential candidates for the bound complex. The lack of charged residues in the β_2-β_3 loop resulted in no salt bridges formed with the side chain of R_{12} in system #4. On the contrary, in systems #6 and #7 glutamate E_{64} (β_5-β_6 loop) formed a stable salt bridge with SmB_2 arginine R_{12}, suggesting its importance for the stability of the SmB2:t-WW complex (Figure 5.8B). Remarkably, during the MD simulations, the stable hydrogen bonds with the side chain of W_{70} were formed between the backbone carbonyl oxygens of G_5 and G_{17} in the #6 and #7 systems respectively. Additionally, in both systems, the proline-rich motif bound to the WW_1 domain was further stabilized by the side chain of S_{27}. In the inverted binding mode (system #7), the P_{7-9} proline-rich motif was kept in the WW_2 domain through an additional hydrogen bond with the side chain of S_{68}. The flanking arginine R_{19} of SmB_2 further facilitated bound complexes through salt bridges with E_{71}

(system #6), and E_{30} (system #7). The computed contact maps for the #6 and #7 systems revealed that both proline-rich motifs of the SmB_2 remained in close contact with two aromatic clusters (Figure 5.8C, Table A.5). Taken together the hydrogen bond networks and the good packing of the proline-rich motifs against the cluster of aromatic residues, I concluded that systems #6 and #7 were meaningful candidates for the SmB2:t-WW complex representing two experimentally determined binding directions of the SmB_2.

5.4 Discussion and Conclusions

The presented work shed light on the recognition of bivalent proline-rich sequences by the tandem WW domains of the spliceosomal forming binding protein 21 (FBP21). Taken together the findings of the NMR titration experiments and the MD simulations of the systems #1, #2 and #3, I concluded that the β-strand fold of two WW domains remained stable in the absence of a PRS in all three investigated systems. This finding led to a conclusion that the refolding of any two domains was essential during the binding event (Figure 5.2). Extensive molecular dynamics simulations revealed a highly complex dynamics of the apo t-WW domains. The conformational ensemble of this system was characterized by 50 metastable structures. Interestingly, in 43 of those structures, two WW domains remained in close contact due to the formation of an inter-domain interface. The interface formations can be attributed to the abundance of the charged residues encoded in the primary structure of the t-WW domains. Consequently, the side chains of such residues bridged two WW domains with each other, then WW domains with the flexible linker and both termini. The inter-domain interfaces differed among the affected metastable structures, and I classified them into four major types based on the participating secondary structure elements.

In the kinetic model of the conformational dynamics of the apo t-WW domains, five slow dynamics modes were identified. They described the kinetic exchanged between the six minima of the free energy surface occurring at the time scales between 2.58 and 7.4 μs. Each of the minima M_1-M_5 corresponded to a single metastable structure exhibiting the inter-domain interface, albeit the type of the interface varied among different minima (Figure 5.6). Then in the minima, M_1-M_5, two WW domains did not assume the relative orientation to each other which would promote simultaneous recognition of both proline-rich motifs of the SmB_2 peptide. Also, the essential aromatic residues of the WW_2 domains participated in the inter-domain interface of the minima M_2, M_4 and M_5. Such structural features of the five deepest minima of the free energy surface implied that (i) a

binding-competent structure might be found in the vast and shallow minimum M_6; (ii) if such structure existed then presence of the PRS would provoke its formation excluding the conformational selection on the level of protein as a mechanism of the PRS recognition. Indeed, I was able to identify a single binding-competent structure by following the criteria discussed in the **section 5.3.4** and considering the outcomes of the site-directed mutagenesis study combined with the SmB_2 affinity measurements and the NOESY experiments.

The SmB_2 peptide was modeled to be in the PPII conformation in the subsequent docking calculation, which yielded four potential candidates of the SmB_2:t-WW complex. However, it turned that the orientation of the central arginine R_{12} of SmB_2 dictated instability of two selected docking poses (systems #4 and #5), leading to the detachment of the SmB_2 from the WW_2 domain, as confirmed by the MD simulations (Figure A.6). MD simulations of the remaining two docking poses (Figure 5.8), showed that the side chain of R_{12} was hydrogen-bonded to the β_5-β_6 loop contributing to the stable SmB_2:t-WW complexes (systems #6 and #7).

The protein-protein docking and MD simulations proved the dual binding mode of the SmB_2 was possible. Since the SmB_2 peptide remained bound to the t-WW domains, no clear conclusion on the preference for the binding direction could be drawn. Given the proposed binding models and the hierarchy of the free energy surface of the apo t-WW domains, it is likely that the combination of the conformational selection and the induced fit on the level of the protein plays a role in the PRS recognition. During the MD simulations of the systems #6 and #7, the t-WW domains fluctuated from the starting structures with the RMSD of 2-4 Å, allowing for the subtle rearrangements of the residues essential for the recognition of the SmB_2.

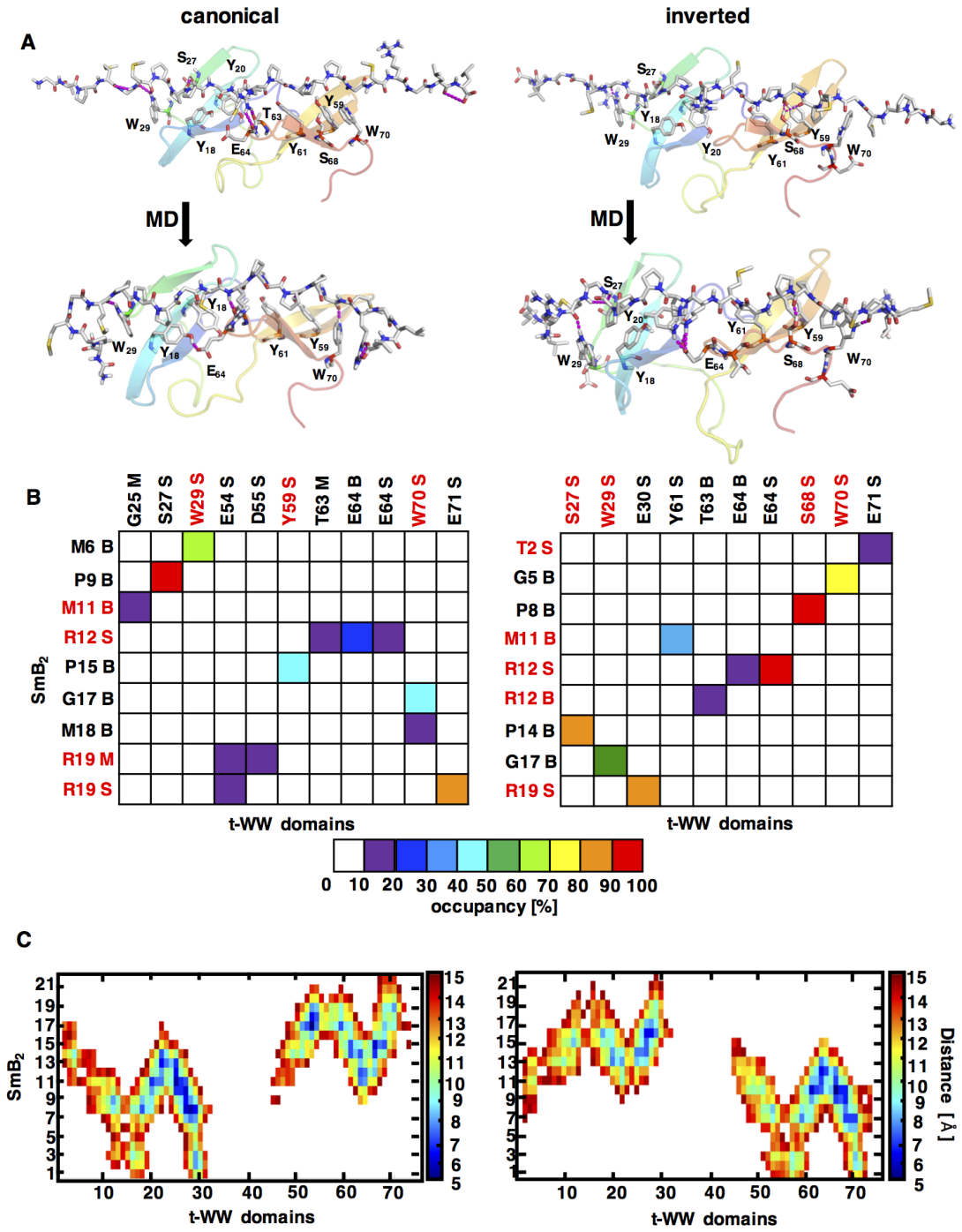


FIGURE 5.8: (A) Docking poses of the SmB_2:t-WW complex in the canonical binding mode (left upper panel) subject of the 100 ns MD run (system #6) with the last simulation frame presented (left lower panel), and an inverted binding mode (right upper panel) subject of the 100 ns MD run (system #7) with the last simulation frame presented (right lower panel); (B) Hydrogen bond networks of two SmB_2:t-WW complexes (canonical binding mode-left panel; inverted binding mode-right panel) - acceptors marked in red; (C) Contact maps of the C-α atoms of two SmB_2:t-WW complexes (canonical binding mode-left panel; inverted binding mode-right panel).

Chapter 6

Conclusions and Outlook

In recent decades molecular dynamics has evolved into a robust method to study dynamical properties of biomolecules and became a complementary method to other structural techniques. Therefore, it was a method of choice in three projects covered in this thesis, whose primary focus was the analysis of the multivalent systems.

As presented in **Chapter 3**, I contributed to a structural study on the trimeric C-type lectin receptor Langerin, an important component of the innate immune system responsible for pathogen recognition and antigen presentation. Prior to this study, a mechanism governing extracellular carbohydrate uptake and intracellular release mediated by Langerin was not well understood. Like other C-type lectins, Langerin requires the Ca^{2+} cofactor to perform its biological role. We (a collaborative project with Structural Glycobiology Group led by Dr. Christoph Rademacher) showed that no interdomain cooperativity was involved in the Ca^{2+} binding and sugar recognition. However, Langerin affinity for the Ca^{2+} ion was pH-sensitive, and it was under control of an intradomain allosteric network. This was a somewhat surprising outcome since other studied C-type lectin exhibited different mechanisms of sugar uptake and release. Further, we were interested in elucidating the role(s) of such an allosteric network. The pH-competent dictating Langerin affinity for the Ca^{2+} ion was encoded in this robust allosteric network and controlled by two pH sensors: (i) a hub residue of the allosteric network H294 (ii) and an unknown pH sensor. The protonation of H294 affected the coupling of the highly flexible short loop with the adjacent long loop bearing the Ca^{2+} binding site. We proposed that the unknown pH sensor was located in the Ca^{2+} binding site itself. This finding triggers several questions yet to be answered:

- which residue among the charged residues of the Ca^{+2} site (E285, E293, and D308) is the second pH sensor;
- do all the three charged residues form a pH triad, and then the positive charge is transferred among them;

- does the protonation of the second pH sensor lead to Ca^{2+}/sugar release in the first place;
- does the opening of the Ca^{2+} channels in the endosomal compartment and the overall decrease of the Ca^{2+} concentration cause Ca^{2+}/sugar release allowing the protonation of the second pH sensor, which in turn initializes the conformational changes in the Ca^{+2} site preventing the Ca^{2+} rebinding.

Further investigations into this questions are currently performed. Then, we excluded a possibility that the allosteric network was a part of a broader signaling pathway. It was not clear with the current data whether the allosteric network evolved as an intrinsic feature of Langerin structure compensating for the loss of the conformational entropy upon a ligand binding. Finally, we reported several conformational events taking place at different time scales. While the loops coupling was in the nanoscale range, the cis-trans isomerization of the highly conserved P286 (located in the long loop) occurred in the micro- to millisecond regime.

In **Chapter 4**, I focused on investigating the conformational dynamics of the trivalent sialosides designed to inhibit an influenza virus protein, namely Hemagglutinin. Such ligands consist of a rigid adamantane core and the flexible PEG linkers to which sialic residues were attached. I showed that PEG linkers tend to collapse rather easily to the coil-like structures, suggesting a high conformational entropic cost upon the binding to the receptor. Then, I identified several intramolecular hydrogen bonds facilitating the formation of the coil-like structures of the PEG linkers. The relative occurrence of such hydrogen bonds could partially entail for the discrepancy in K_d of the two investigated ligands. As indicated by molecular dynamics simulations, one might consider replacing the carbamoyl moiety that connects the adamantane core with PEG linkers and contains the dominant N_2-H_{20} donor with an ether group to further fine-tune such trivalent sialosides.

From investigating the conformational dynamics of a receptor and a ligand separately, I proceeded in **Chapter 5** with a project that aimed at elucidating the binding of a bivalent proline-rich peptide (SmB_2) to the tandem-WW domains of Formin Binding Protein 21 (FBP21). The molecular dynamics simulations of the apo t-WW domains revealed a highly complex conformational dynamics. In-depth analysis of the metastable regions of the conformational space showed that tandem-WW domains tend to form various interdomain interfaces characterized by a (core)set specific hydrogen bond network. Additionally, I constructed

a kinetic model of the apo t-WW domains, shedding light on the organization of the free energy landscape. There were five distinguishable minima each corresponding to a single metastable structure. The sixth minimum comprised of 42 fast interconverting metastable sets and can be assumed as broad, shallow and rugged. Remarkably, a binding-competent structure was found in this broad minimum, and it lacked an interdomain interface. This finding suggested that combination of the conformational selection and induced fit on the side of the protein plays a role in the binding of the bivalent SmB_2 peptide. The bound complex was modeled by the combination of HADDOCK docking protocol and the molecular dynamics simulations yielding the SmB_2 peptide bound to the t-WW domains in a canonical and in an inverted direction respectively. Still, there is an open question if the proposed binding models would apply to any other bivalent proline-rich peptide. The novelty of this work is reflected in the combination of molecular dynamics, Markov State Model theory, protein-protein docking and NMR experiments to study protein-protein interactions.

Appendix A

Chapter 5: Supplementary Material

Donor	**Acceptor**	% **WW**$_1$	**% t-WW**
R6 SC	D21 SC	100	33.14
S24 SC	D21 SC	95.94	100
V8 BB	Y20 BB	90.51	88.78
Y20 BB	V8 BB	90.05	89.5
Q28 BB	Y19 BB	85.72	85.29
Y19 BB	Q28 BB	82.77	84.46
D21 BB	A26 BB	78.56	78.12
S24 BB	D21 SC	78.39	83.34
Y18 BB	G10 BB	77.2	72.63
Q28 SC	E30 BB	75.62	77.75
I23 BB	D21 SC	67.64	75.9
T12 BB	Y16 BB	67.38	68.89
G10 BB	Y18 BB	65.12	66.49
L22 BB	R6 BB	58.83	61.49
G15 BB	T12 BB	52.49	52
S3 SC	D1 SC	40.65	14.73
H17 SC	E9 SC	38.53	41.29
Y16 BB	T12 BB	35.22	37.22
W7 BB	G34 BB	30.81	29.53
R6 BB	D21 SC	24.58	44.15
A26 BB	S24 SC	22.59	24.86
F35 BB	P32 BB	21.2	20.16
Q36 BB	W7 BB	16.84	16.98
T12 SC	Y16 BB	14.84	13.13
W7 SC	K4 BB	10.45	13.92

TABLE A.1: Hydrogen bond network comparison WW_1 versus t-WW domains.

Donor	**Acceptor**	% **WW_2**	**% t-WW**
Y61 BB	V49 BB	88.59	88.8
V49 BB	Y61 BB	85.07	84.77
T63 BB	V47 BB	84.52	84.96
R69 BB	Y60 BB	82.57	82.02
Y60 BB	R69 BB	78.38	77.15
T65 BB	N62 SC	76.01	74.16
Y59 BB	G51 BB	71.77	68.68
N62 BB	E67 BB	70.77	70.54
N62 SC	W48 SC	63.93	60.35
G51 BB	Y59 BB	59.27	62.7
S53 BB	F57 BB	53.64	52.14
G56 BB	S53 BB	53.02	44.85
T63 SC	V47 BB	52.84	57.39
E64 BB	N62 SC	51.98	54.35
R69 SC	E67 SC	45.94	49.32
F57 BB	S53 BB	44.43	31.99
T65 SC	E67 SC	43.63	41.82
E67 BB	T65 SC	34.58	33.84
R69 SC	E71 BB	24.86	20.97
S53 SC	F57 BB	23.84	20.26
K40 SC	D38 SC	23.64	10.14
R69 SC	D74 SC	18.23	18.01
F57 BB	E54 BB	14.8	18.57
W48 BB	K45 BB	14.65	13.16
R69 SC	E71 SC	14.45	11.2
G66 BB	N62 BB	10.77	12.65

TABLE A.2: Hydrogen bond network comparison WW_2 versus t-WW domains.

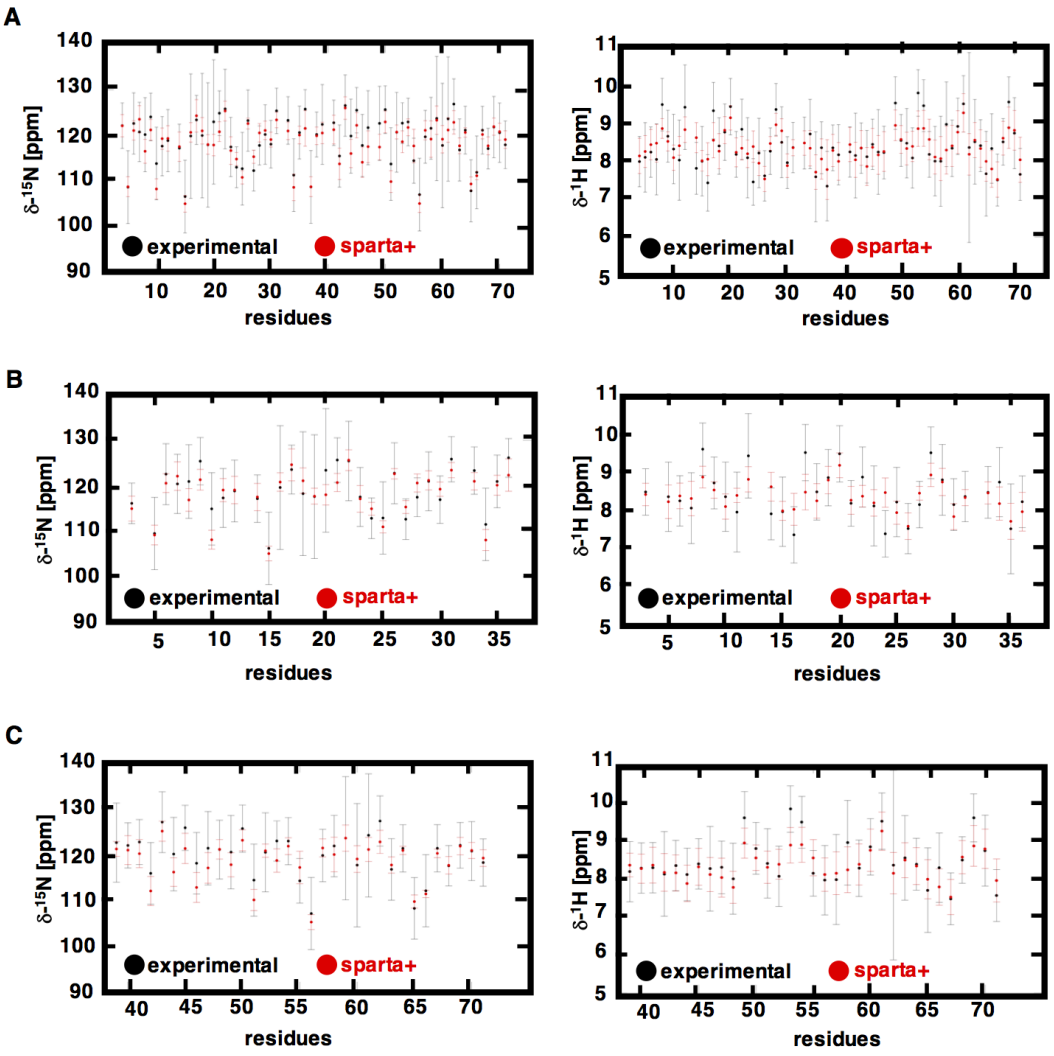
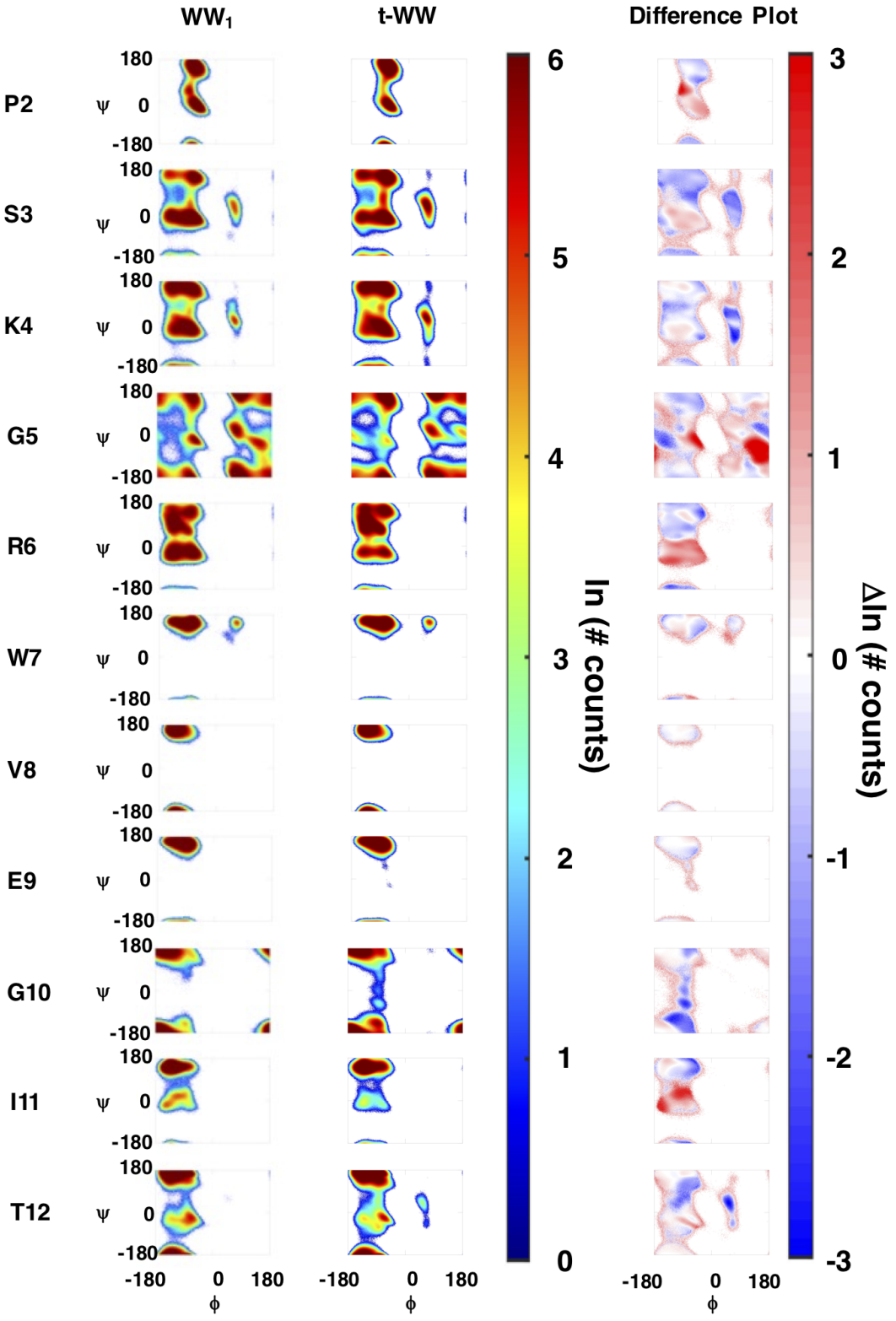
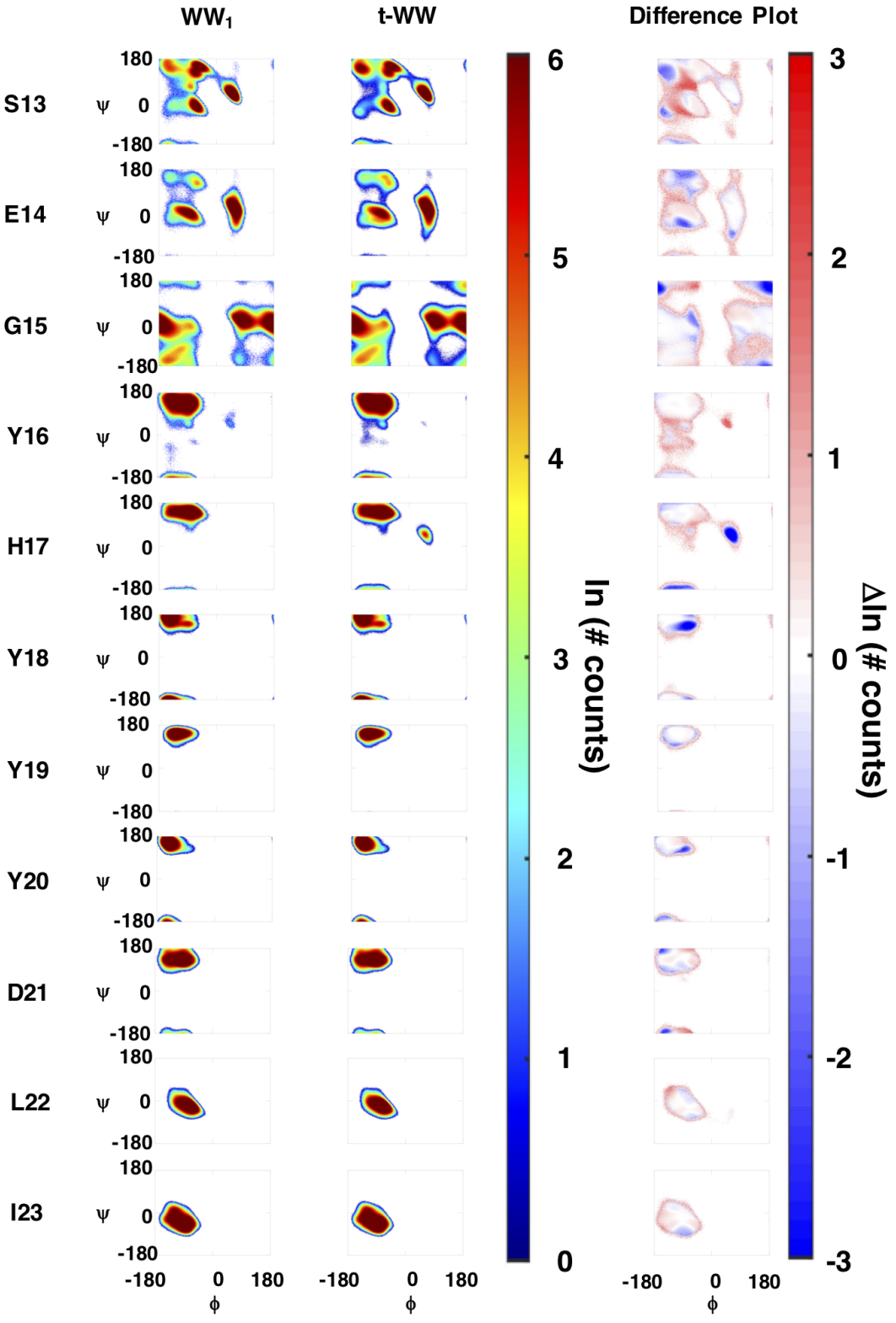


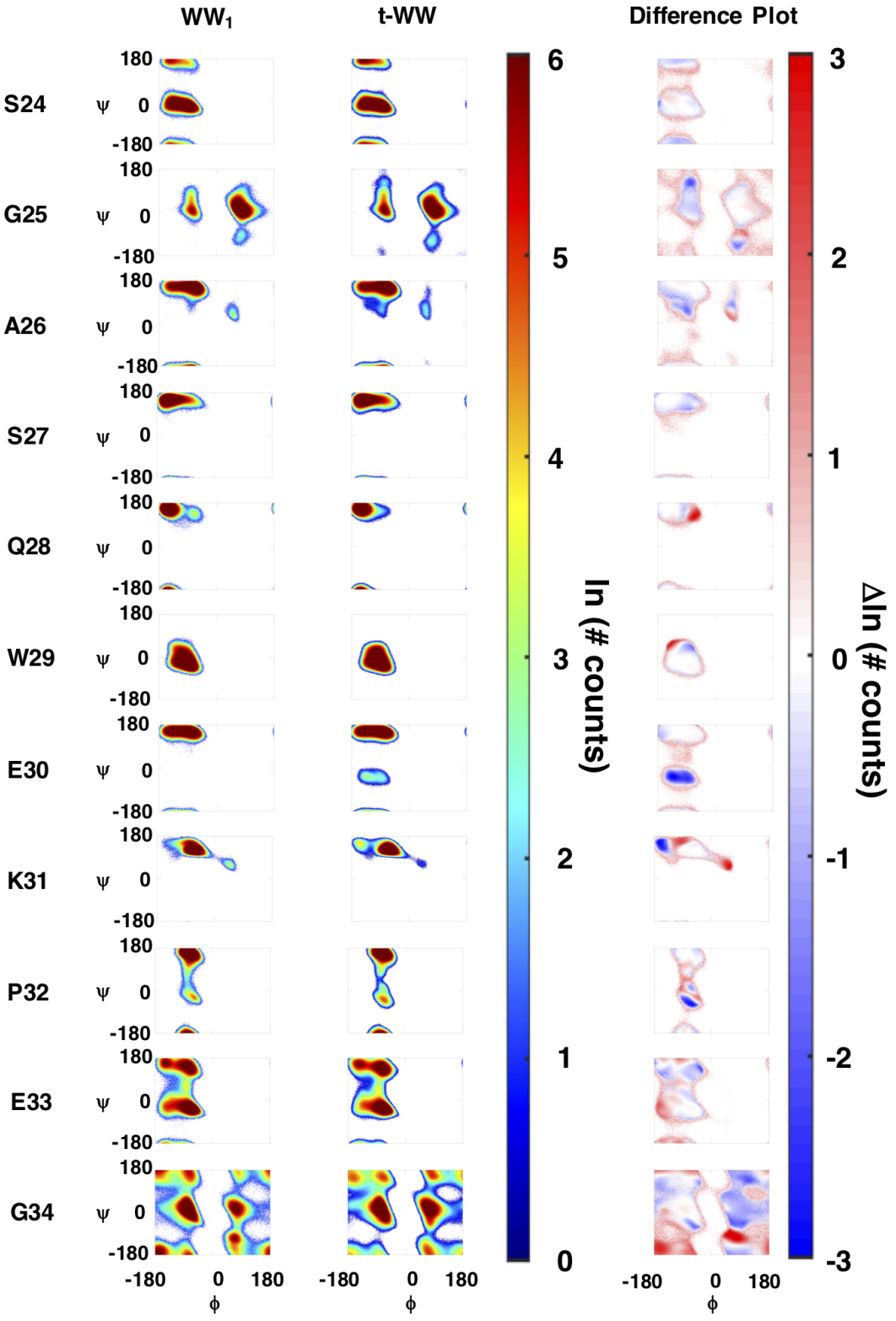
FIGURE A.1: The chemical shifts of the backbone ^{15}N atoms (left panel) and ^{1}H (right panel) are calculated with SPARTA+ software (colored in red) and mapped to the experimentally determined chemical shifts (colored in black): (A) t-WW domains; (B) WW_1 construct; (C) WW_2 construct.

Donor	**Acceptor**	[%]	**# core sets**
Q36 BB	W7 BB	12 - 81 %	13
D1 BB	E33 SC	14 - 83 %	12
T63 SC	T46 BB	10 - 85 %	12
R6 SC	D74 SC	15 - 100 %	10
K4 SC	E33 SC	12 - 65 %	10
K41 SC	E64 SC	10 - 50 %	10
K40 BB	K4 BB	14 - 80 %	8
R6 BB	K40 BB	18 - 72 %	7
R6 SC	D38 SC	14 - 100 %	7
R6 SC	E50 SC	31 - 100 %	7
K40 SC	D1 SC	11 - 46 %	7
W7 BB	Q36 SC	53 - 86 %	7
R6 SC	E64 SC	87 - 100 %	6
R6 SC	E67 SC	12 - 100 %	6
R6 SC	L39 BB	27 - 68 %	6
K40 SC	E64 SC	11 - 53 %	6
K41 SC	E67 SC	19 - 42 %	6
S3 BB	E33 BB	14 - 62 %	6
R6 BB	G34 BB	13 - 40 %	5
D1 BB	D38 SC	10 - 37 %	5
E9 BB	G37 BB	31 - 68 %	5
K4 SC	D38 SC	11 - 41 %	5
K40 SC	E9 SC	12 - 64 %	5
K72 SC	D38 SC	16 - 51 %	5
T42 BB	E64 SC	14 - 69 %	5

TABLE A.3: Semi-conserved hydrogen bonds appearing across several core sets.


WW1
t-WW
Difference Plot
P2
S3
K4
G5
R6
W7
V8
E9
G10
I11
T12
ψ
180
0
-180
ϕ
ln (# counts)
Δln (# counts)
6
5
4
3
2
1
0
-1
-2
-3


WW1
t-WW
Difference Plot
S13
E14
G15
Y16
H17
Y18
Y19
Y20
D21
L22
I23
ψ
180
0
-180
ϕ
ln (# counts)
Δln (# counts)
6
5
4
3
2
1
0
-1
-2
-3


WW1
t-WW
Difference Plot
S24
G25
A26
S27
Q28
W29
E30
K31
P32
E33
G34
ψ
ϕ
180
0
-180
ln (# counts)
Δln (# counts)
6
5
4
3
2
1
0
3
2
1
0
-1
-2
-3

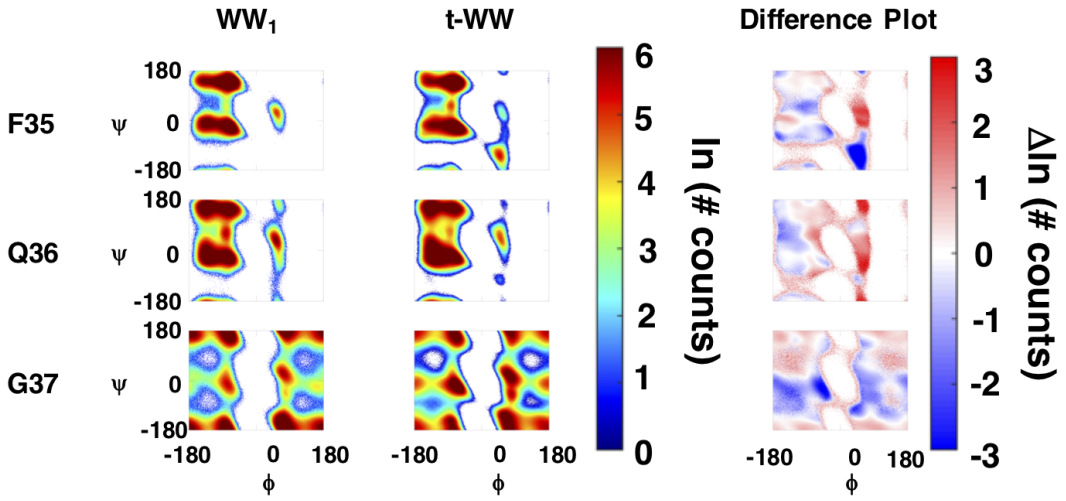
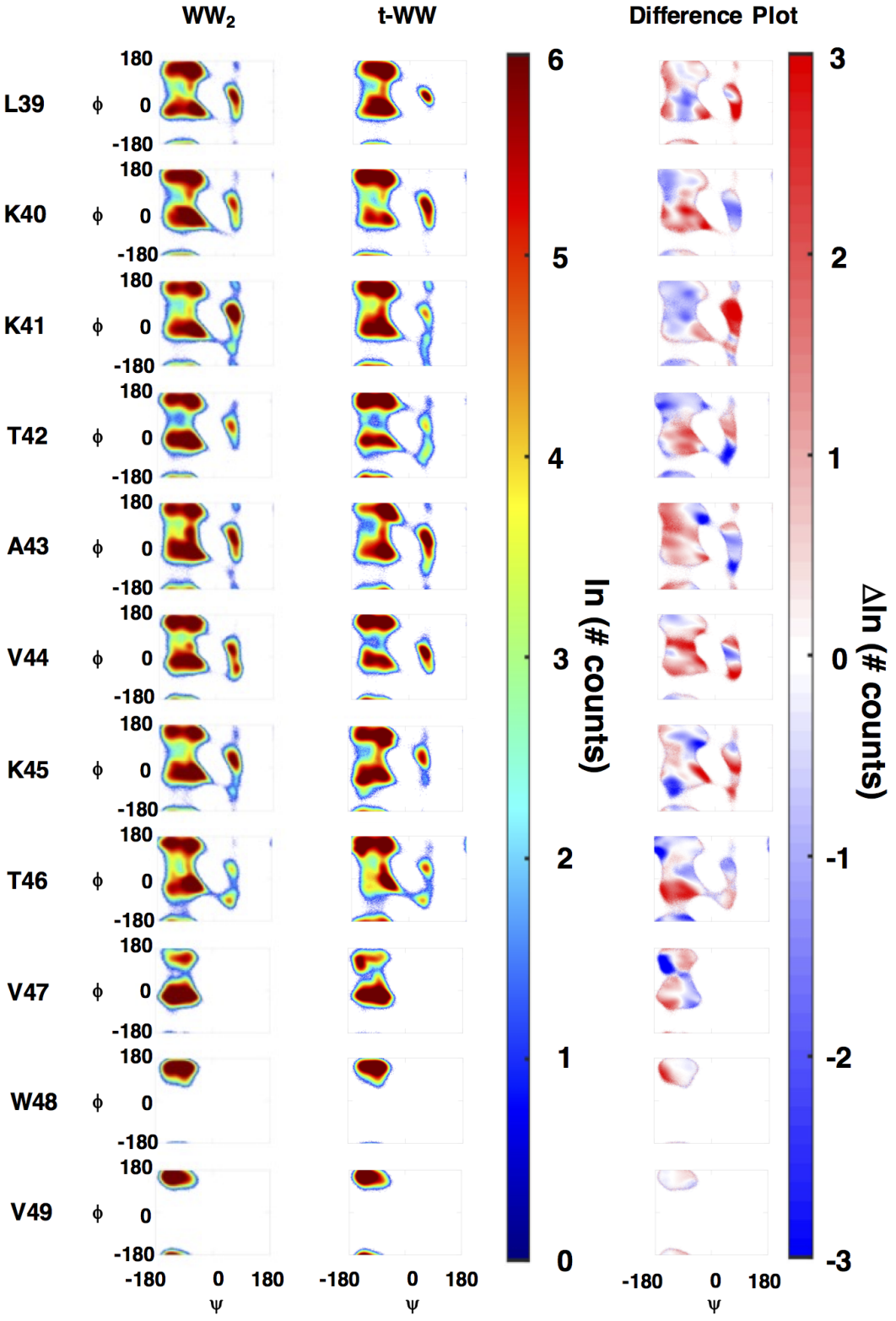
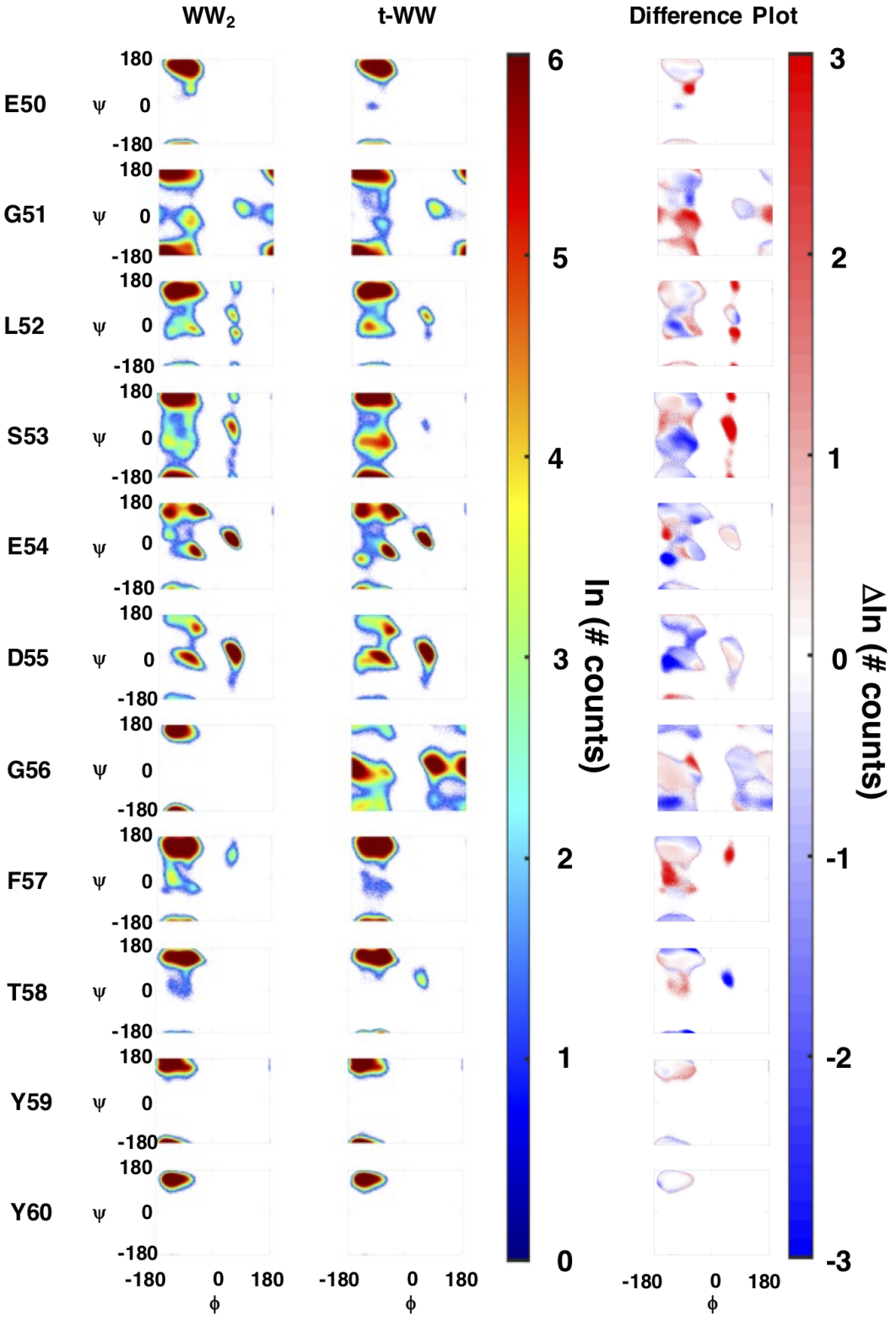
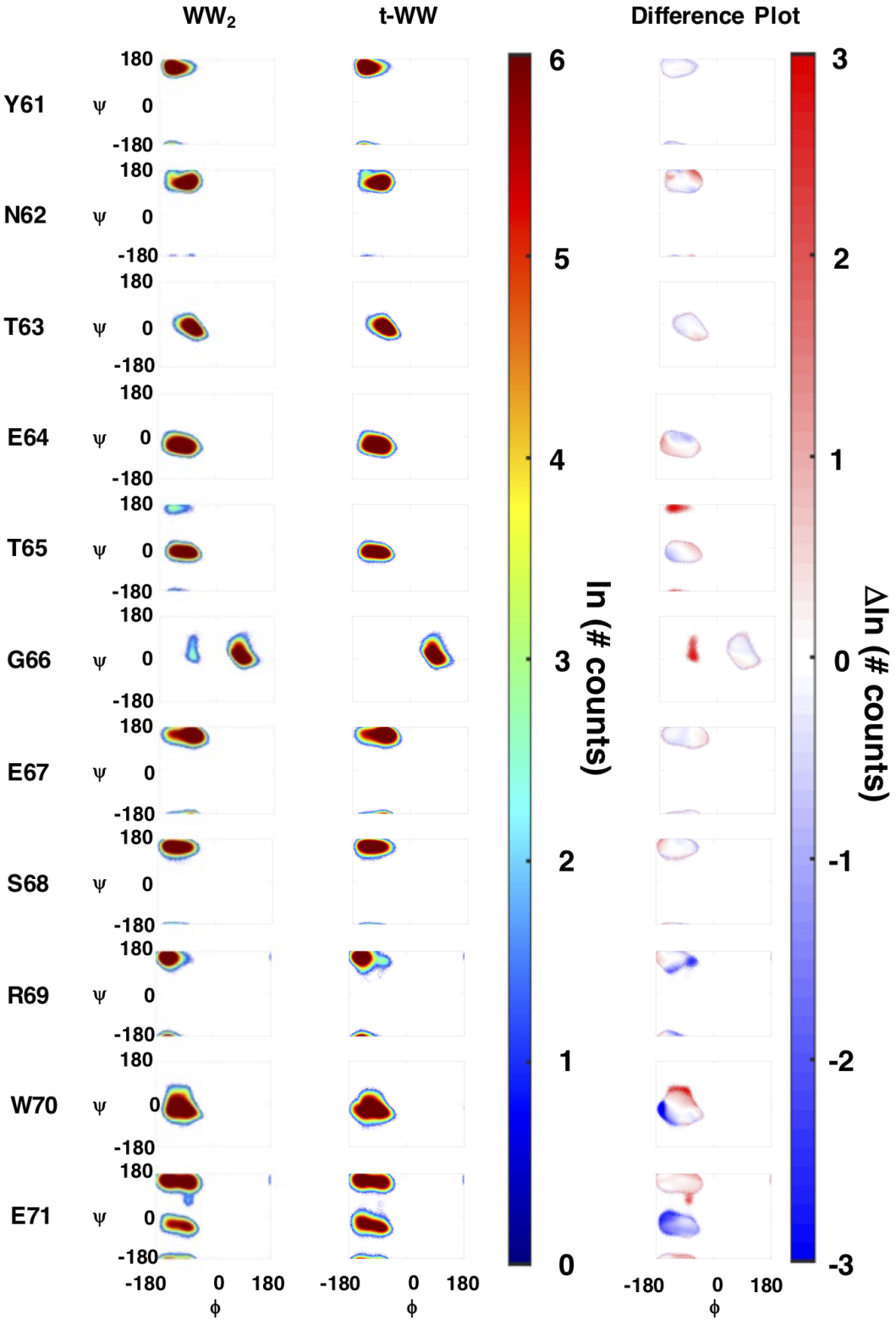


FIGURE A.2: Backbone dynamics analysis in Ramachandran space for WW_1 versus t-WW domains.


WW2
t-WW
Difference Plot
L39
K40
K41
T42
A43
V44
K45
T46
V47
W48
V49
ϕ
ψ
180
0
-180
ln (# counts)
Δln (# counts)
6
5
4
3
2
1
0
3
2
1
0
-1
-2
-3


WW2
t-WW
Difference Plot
E50
G51
L52
S53
E54
D55
G56
F57
T58
Y59
Y60
ψ
180
0
-180
ϕ
-180
0
180
ln (# counts)
6
5
4
3
2
1
0
Δln (# counts)
3
2
1
0
-1
-2
-3


WW2
t-WW
Difference Plot
Y61
N62
T63
E64
T65
G66
E67
S68
R69
W70
E71
ψ
180
0
-180
-180
0
180
ϕ
6
5
4
3
2
1
0
ln (# counts)
3
2
1
0
-1
-2
-3
Δln (# counts)

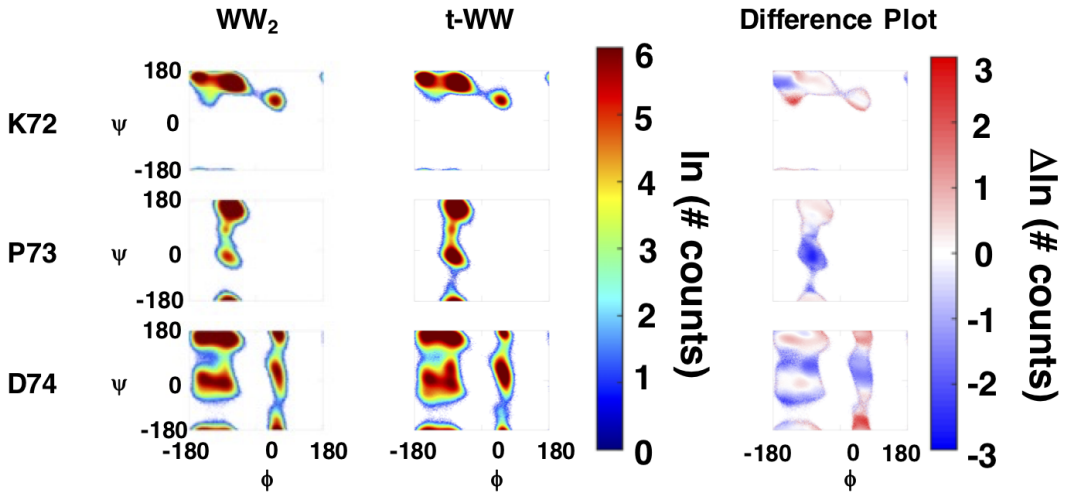
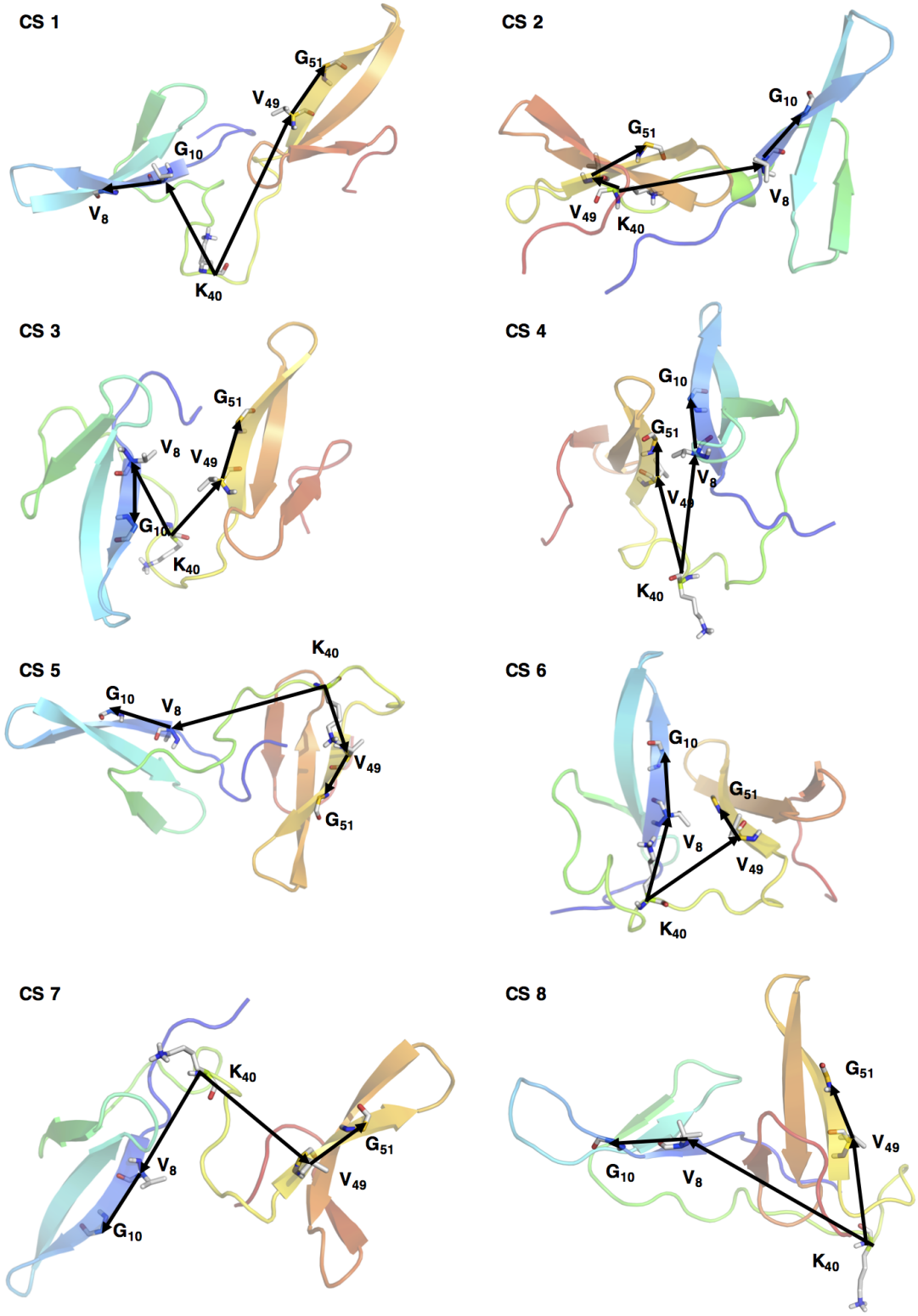
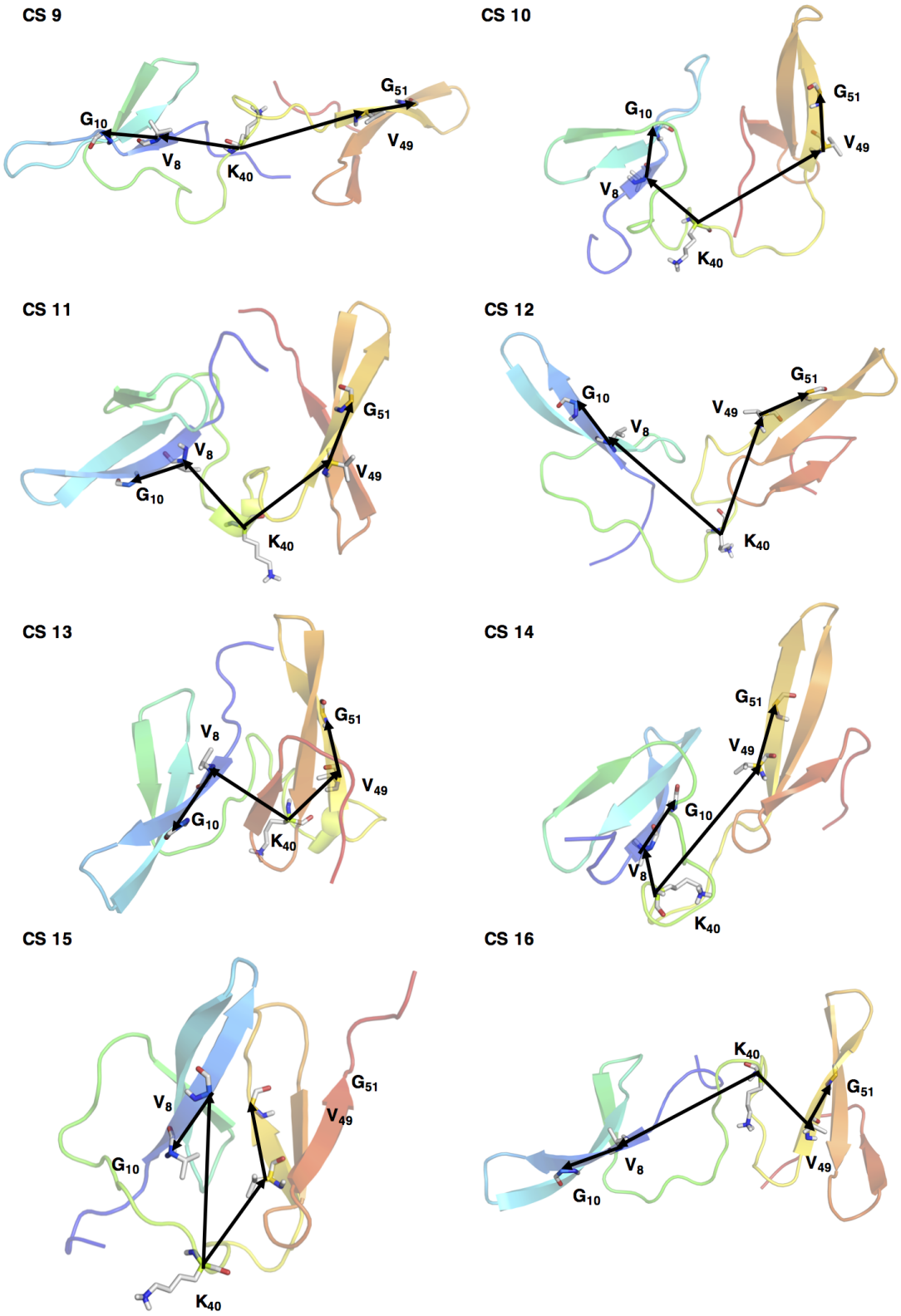


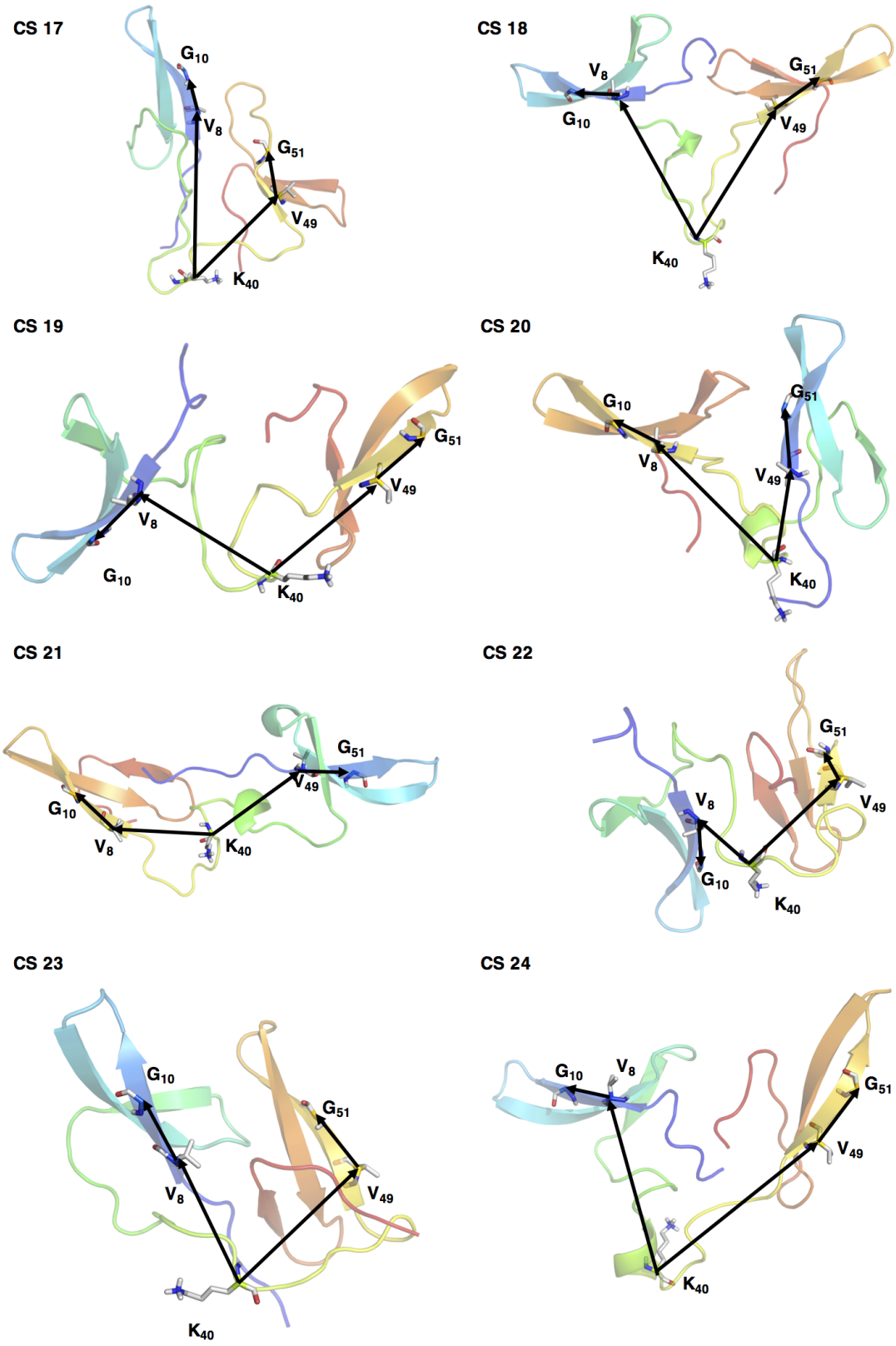
FIGURE A.3: Backbone dynamics analysis in Ramachandran space for WW_2 versus t-WW domains.

core set	angle α	angle β	orien-tation		core set	angle α	angle β	orien-tation
1	127.55	97.49	$\uparrow\rightarrow$		26	27.61	152.21	$\uparrow\uparrow$
2	20.01	144.95	$\uparrow\rightarrow$		27	106.18	72.77	$\uparrow\downarrow$
3	118.41	104.00	$\uparrow\downarrow$		28	154.36	77.00	$\uparrow\downarrow$
4	61.61	164.31	$\uparrow\rightarrow$		29	75.97	121.24	$\uparrow\rightarrow$
5	60.58	90.94	$\uparrow\rightarrow$		30	102.19	102.03	$\uparrow\downarrow$
6	53.69	139.46	$\uparrow\rightarrow$		31	112.04	66.98	$\uparrow\rightarrow$
7	136.24	74.84	$\uparrow\downarrow$		32	30.25	131.04	$\uparrow\downarrow$
8	63.98	131.82	$\uparrow\rightarrow$		33	80.99	57.02	$\uparrow\downarrow$
9	166.38	31.91	$\uparrow\downarrow$		34	144.76	70.63	$\uparrow\downarrow$
10	29.83	104.01	$\uparrow\rightarrow$		35	52.67	93.16	$\uparrow\rightarrow$
11	141.09	73.41	$\uparrow\rightarrow$		36	79.24	106.69	$\uparrow\uparrow$
12	107.14	111.98	$\uparrow\rightarrow$		37	124.65	30.84	$\uparrow\downarrow$
13	122.35	111.41	$\uparrow\downarrow$		38	99.92	62.13	$\uparrow\rightarrow$
14	72.06	143.60	$\uparrow\rightarrow$		39	50.38	96.22	$\uparrow\rightarrow$
15	56.21	139.43	$\uparrow\uparrow$		40	87.55	134.39	$\uparrow\rightarrow$
16	135.62	48.85	$\uparrow\rightarrow$		41	79.22	71.14	$\uparrow\rightarrow$
17	43.30	142.55	$\uparrow\rightarrow$		42	123.45	36.04	$\uparrow\downarrow$
18	127.08	101.58	$\uparrow\downarrow$		43	159.71	12.81	$\uparrow\downarrow$
19	113.31	56.99	$\uparrow\downarrow$		44	81.58	98.53	$\uparrow\rightarrow$
20	50.72	125.34	$\uparrow\rightarrow$		45	147.82	45.55	$\uparrow\downarrow$
21	128.09	31.40	$\uparrow\downarrow$		46	37.73	157.52	$\uparrow\uparrow$
22	52.78	94.63	$\uparrow\downarrow$		47	42.73	134.48	$\uparrow\uparrow$
23	50.45	116.13	$\uparrow\rightarrow$		48	68.52	71.06	$\uparrow\uparrow$
24	106.64	107.22	$\uparrow\rightarrow$		49	132.52	89.74	$\uparrow\downarrow$
25	39.50	148.03	$\uparrow\uparrow$		50	121.56	108.04	$\uparrow\rightarrow$

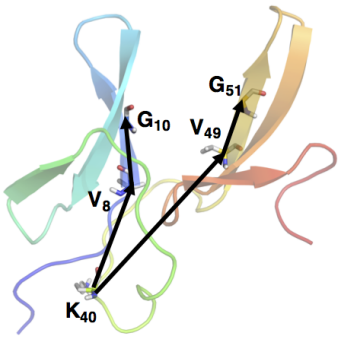
TABLE A.4: Vector Analysis.


CS 1
G51
V49
G10
V8
K40
CS 2
G10
G51
V8
V49
K40
CS 3
G51
V8
V49
G10
K40
CS 4
G10
G51
V8
V49
K40
CS 5
K40
G10
V8
V49
G51
CS 6
G10
G51
V8
V49
K40
CS 7
K40
V8
G51
V49
G10
CS 8
G51
V49
G10
V8
K40

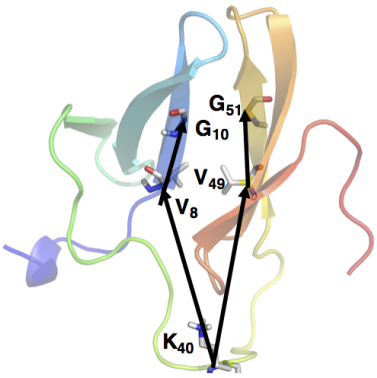

CS 9
G10
V8
K40
G51
V49
CS 10
G10
V8
K40
G51
V49
CS 11
V8
G10
K40
G51
V49
CS 12
G10
V8
V49
G51
K40
CS 13
V8
G10
K40
G51
V49
CS 14
G51
V49
G10
V8
K40
CS 15
V8
G10
K40
G51
V49
CS 16
K40
V8
G10
G51
V49


CS 17
G10
V8
G51
V49
K40
CS 18
V8
G10
G51
V49
K40
CS 19
G51
V49
V8
G10
K40
CS 20
G10
G51
V8
V49
K40
CS 21
G51
V49
G10
V8
K40
CS 22
G51
V8
V49
G10
K40
CS 23
G10
G51
V8
V49
K40
CS 24
G10
V8
G51
V49
K40

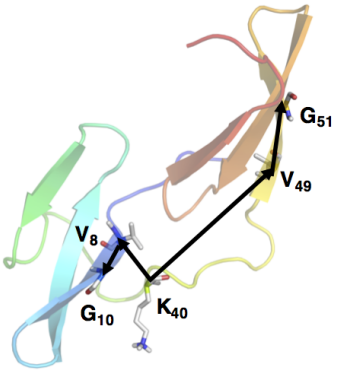
CS 25



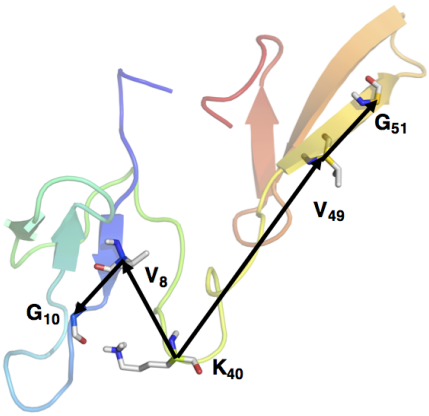
CS 26



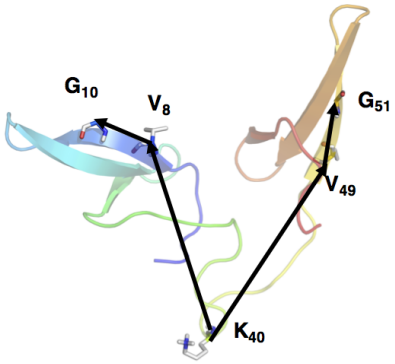
CS 27



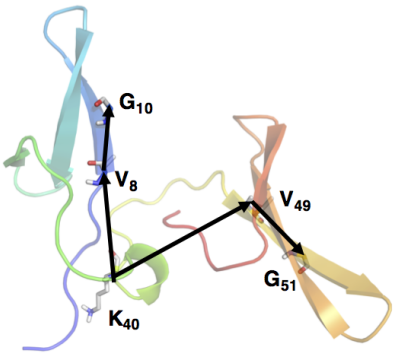
CS 28



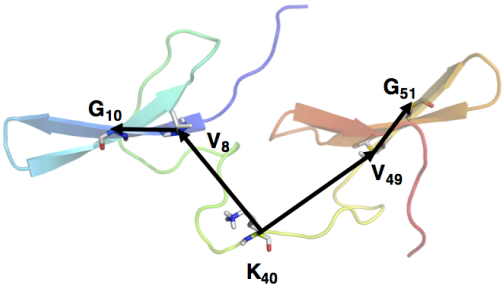
CS 29



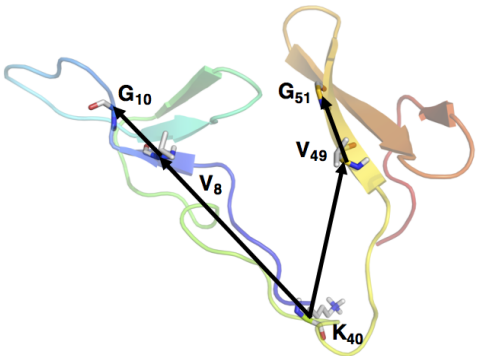
CS 30



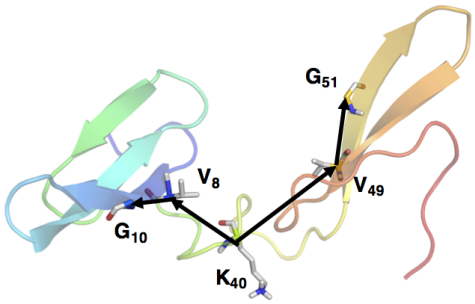
CS 31



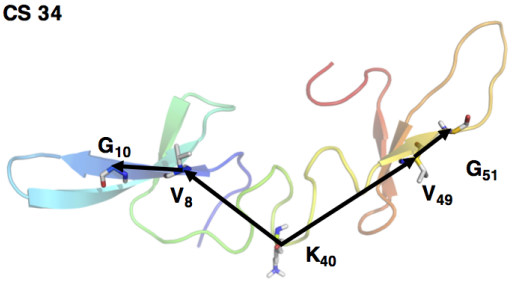
CS 32



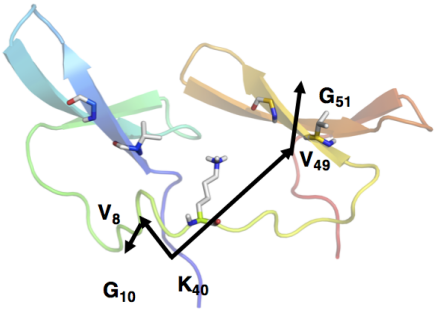
CS 33



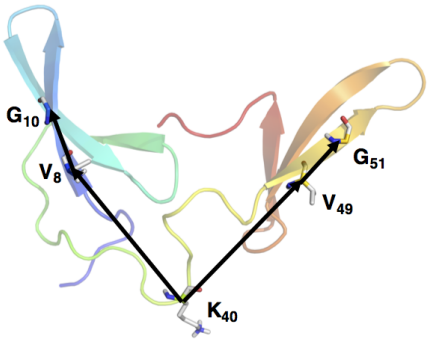
CS 34



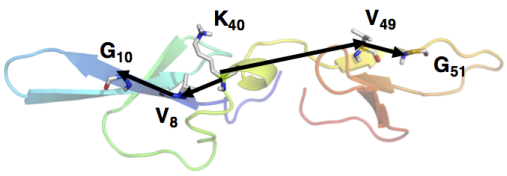
CS 35



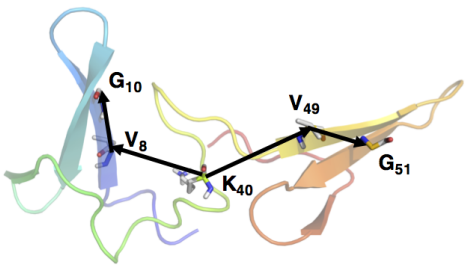
CS 36



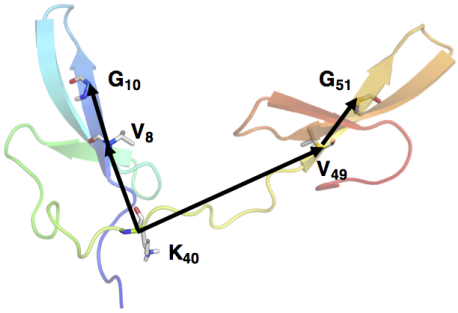
CS 37



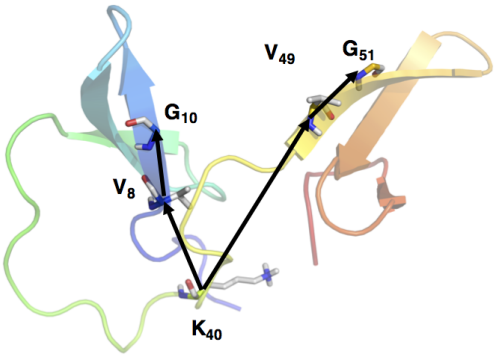
CS 38



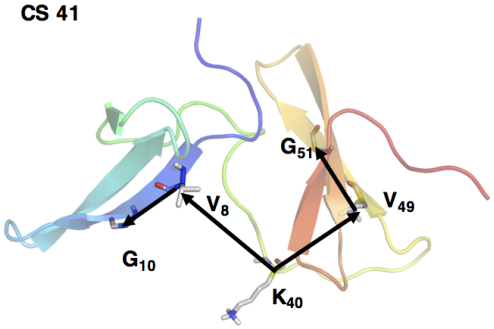
CS 39



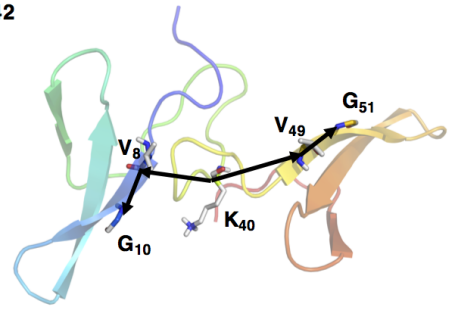
CS 40



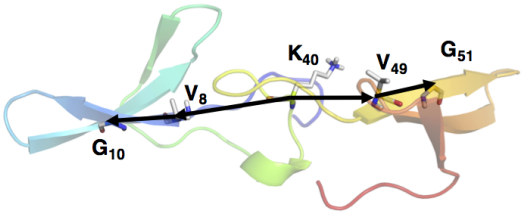
CS 41



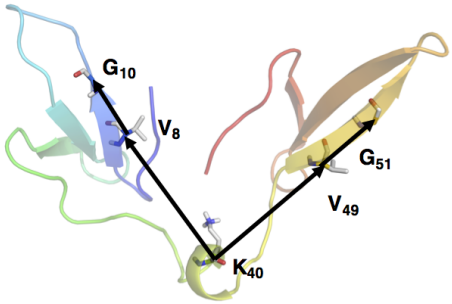
CS 42



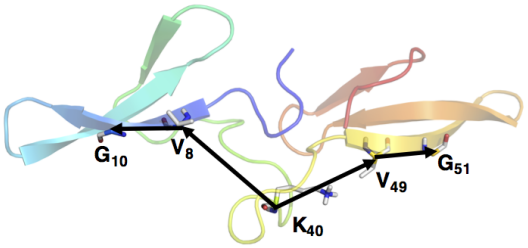
CS 43



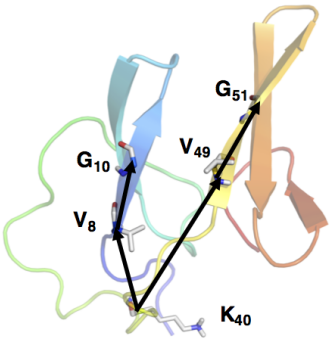
CS 44



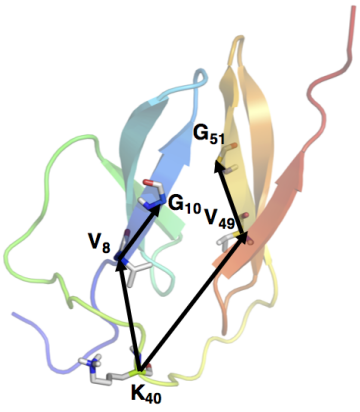
CS 45



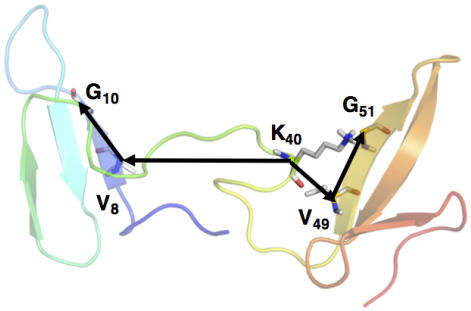
CS 46



CS 47



CS 48



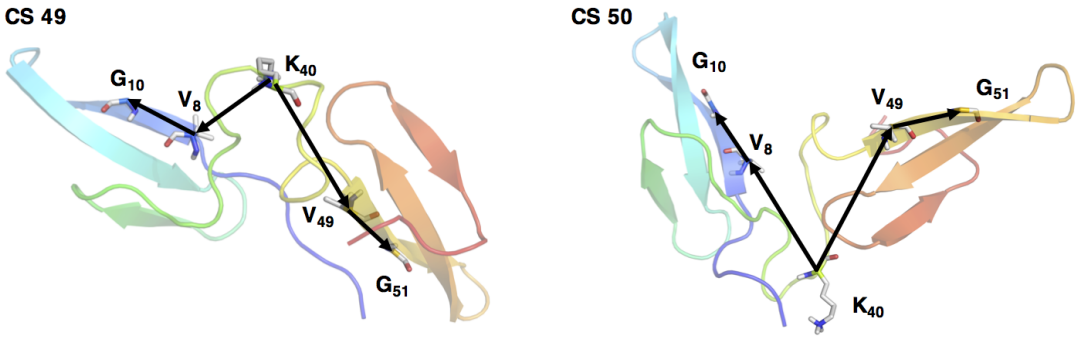


FIGURE A.4: Vector analysis.

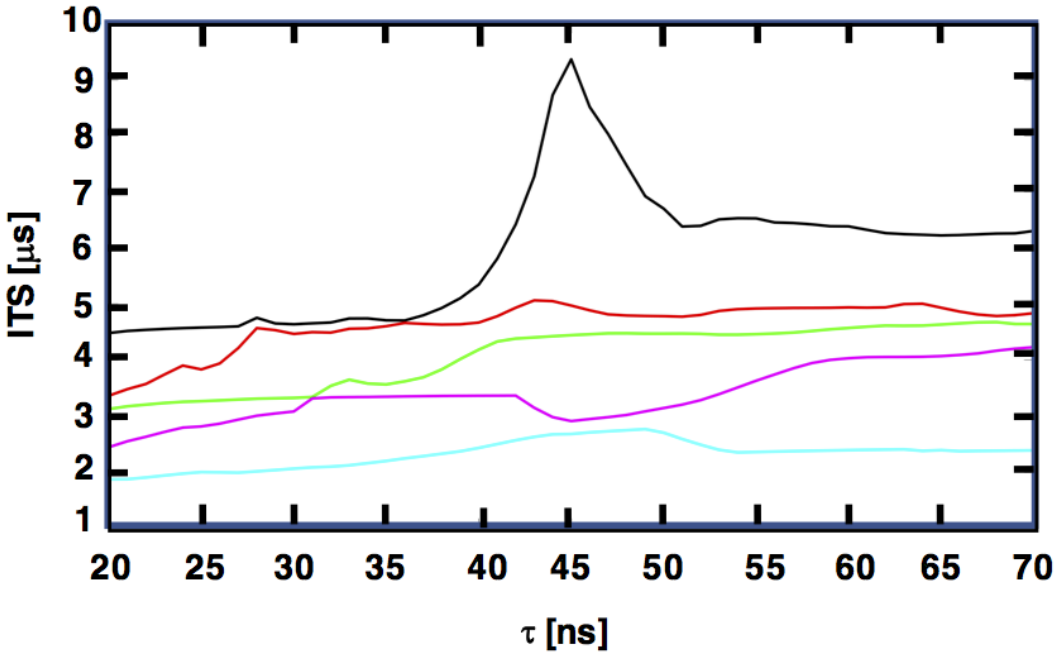


FIGURE A.5: Implied time scales of the kinetic model.

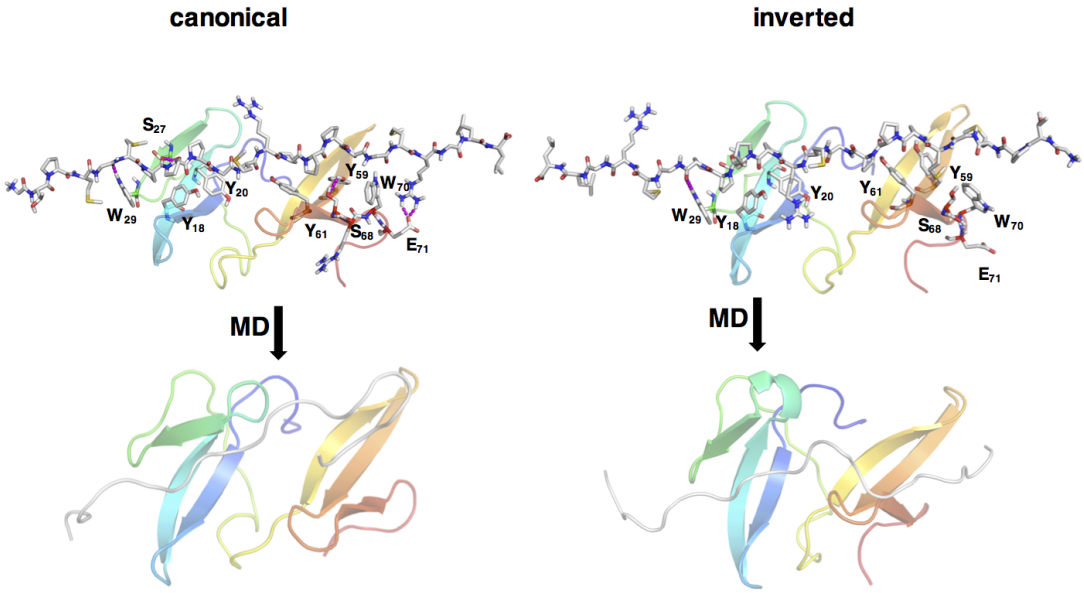


FIGURE A.6: Docking poses of the SmB_2:t-WW complex in the canonical binding mode (left upper panel) subject of the 100 ns MD run (system #4) with the last simulation frame presented (left lower panel), and an inverted binding mode (right upper panel) subject of the 100 ns MD run (system #5) with the last simulation frame presented (right lower panel).

$SmB_{2,canon}$	t-WW	Distance [Å]		$SmB_{2,invert}$	t-WW	Distance [Å]
P_7	W_{29}	8.01 ± 0.68		P_7	Y_{59}	6.80 ± 0.41
				P_7	Y_{61}	8.81 ± 0.29
				P_7	W_{70}	8.81 ± 0.29
P_8	Y_{18}	7.29 ± 0.42		P_8	Y_{59}	8.14 ± 0.57
P_8	W_{29}	6.38 ± 0.44		P_8	Y_{61}	8.49 ± 0.33
P_9	Y_{18}	8.54 ± 0.81		P_9	W_{70}	7.30 ± 0.35
P_9	Y_{20}	8.89 ± 0.42				
				P_{13}	Y_{18}	7.03 ± 0.5
P_{14}	Y_{61}	7.49 ± 0.46		P_{14}	Y_{20}	7.47 ± 0.69
P_{15}	Y_{61}	7.26 ± 0.38		P_{15}	Y_{18}	8.66 ± 0.81
				P_{15}	Y_{20}	8.43 ± 0.46
				P_{16}	Y_{18}	7.47 ± 0.43
				P_{16}	Y_{20}	8.33 ± 0.36
				P_{16}	W_{29}	6.94 ± 0.57

TABLE A.5: The average distance between the Cα-atoms of the proline-rich motifs of SmB_2 and the aromatic clusters-only atoms pairs at the average distance lower then 9 Å reported.

Bibliography

[1] Sumati Bhatia, Luis Cuellar Camacho, and Rainer Haag. "Pathogen Inhibition by Multivalent Ligand Architectures". In: *Journal of the American Chemical Society* 138.28 (2016), 8654–8666.

[2] D. Bhella. "The role of cellular adhesion molecules in virus attachment and entry". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1661 (2014), pp. 20140035–20140035.

[3] Chad T. Varner et al. "Recent Advances in Engineering Polyvalent Biological Interactions". In: *Biomacromolecules* 16.1 (2014), pp. 43–55.

[4] Choi Seok-Ki Whitesides George M. Mammen Mathai. "Polyvalent Interactions in Biological Systems: Implications for Design and Use of Multivalent Ligands and Inhibitors". In: *Angewandte Chemie* 37 (1998), 2754–2794.

[5] Carlo Fasting et al. "Multivalency as a Chemical Organization and Action Principle". In: *Angewandte Chemie International Edition* 51.42 (2012), pp. 10472–10498.

[6] Jovica D. Badjić et al. "Multivalency and Cooperativity in Supramolecular Chemistry". In: *Accounts of Chemical Research* 38.9 (2005), pp. 723–732.

[7] Hans-Jörg Schneider. "Binding Mechanisms in Supramolecular Complexes". In: *Angewandte Chemie International Edition* 48.22 (2009), pp. 3924–3977.

[8] C. Rankl et al. "Multiple receptors involved in human rhinovirus attachment to live cells". In: *Proceedings of the National Academy of Sciences* 105.46 (2008), pp. 17778–17783.

[9] Jianghong Rao et al. "Design, Synthesis, and Characterization of a High-Affinity Trivalent System Derived from Vancomycin andl-Lys-d-Ala-d-Ala". In: *Journal of the American Chemical Society* 122.12 (2000), pp. 2698–2710.

[10] Marcus Weber, Alexander Bujotzek, and Rainer Haag. "Quantifying the rebinding effect in multivalent chemical ligand-receptor systems". In: *The Journal of Chemical Physics* 137.5 (2012), p. 054111.

[11] Jeffrey J. Landers et al. "Prevention of Influenza Pneumonitis by Sialic Acid–Conjugated Dendritic Polymers". In: *The Journal of Infectious Diseases* 186.9 (2002), pp. 1222–1230.

[12] Jon Jin Kim, Keir, and Marks. "Shigatoxin-associated hemolytic uremic syndrome: current molecular mechanisms and future therapies". In: *Drug Design, Development and Therapy* (2012), p. 195.

[13] Maria del Mar Hernandez and Marco V Jose. "Positive cooperativity induces multimodal site and thermodynamic affinity distributions in multivalent proteins". In: *Analytical Biochemistry* 313.2 (2003), pp. 226–233.

[14] François Jacob and Jacques Monod. "Genetic regulatory mechanisms in the synthesis of proteins". In: *Journal of Molecular Biology* 3.3 (1961), pp. 318–356.

[15] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. "On the Nature of Allosteric Transitions: A Plausible Model". In: *J. Mol. Biol.* 12 (1965), pp. 88–118.

[16] D. E. Koshland, G. Namethy, and D. Filmer. "Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits". In: *ACS Biochem.* 5.1 (1966), pp. 365–385.

[17] Gregorio Weber. "Ligand binding and internal equilibiums in proteins". In: *Biochemistry* 11.5 (1972), pp. 864–878.

[18] K. Gunasekaran, Buyong Ma, and Ruth Nussinov. "Is allostery an intrinsic property of all dynamic proteins?" In: *Proteins: Structure, Function, and Bioinformatics* 57.3 (2004), pp. 433–443.

[19] Alexandr P. Kornev and Susan S. Taylor. "Dynamics-Driven Allostery in Protein Kinases". In: *Trends in Biochemical Sciences* 40.11 (2015), pp. 628–647.

[20] Ozlem Keskin, Nurcan Tuncbag, and Attila Gursoy. "Predicting Protein–Protein Interactions from the Molecular to the Proteome Level". In: *Chemical Reviews* 116.8 (2016), pp. 4884–4909.

[21] Victoria A Feher et al. "Computational approaches to mapping allosteric pathways". In: *Current Opinion in Structural Biology* 25 (2014), pp. 98–103.

[22] Giuseppina La Sala et al. "Allosteric Communication Networks in Proteins Revealed through Pocket Crosstalk Analysis". In: *ACS Central Science* 3.9 (2017), pp. 949–960.

[23] Gregory R Bowman et al. "Discovery of multiple hidden allosteric sites by combining Markov state models and experiments." In: *Proc. Natl. Acad. Sci. U.S.A.* 112.9 (2015), pp. 2734–2739.

[24] Ashok Kumar Grover. "Use of Allosteric Targets in the Discovery of Safer Drugs". In: *Medical Principles and Practice* 22.5 (2013), pp. 418–426.

[25] Inbal Sela-Culang, Vered Kunik, and Yanay Ofran. "The Structural Basis of Antibody-Antigen Recognition". In: *Frontiers in Immunology* 4 (2013).

[26] C. L. Will and R. Luhrmann. "Spliceosome Structure and Function". In: *Cold Spring Harbor Perspectives in Biology* 3.7 (2010), a003707–a003707.

[27] Paul Flicek et al. "Ensembl 2014". In: *Nucleic Acids Research* 42.D1 (2013), pp. D749–D755.

[28] M. P. H. Stumpf et al. "Estimating the size of the human interactome". In: *Proceedings of the National Academy of Sciences* 105.19 (2008), pp. 6959–6964.

[29] Eric Aragón et al. "Structural Basis for the Versatile Interactions of Smad7 with Regulator WW Domains in TGF- Pathways". In: *Structure* 20.10 (2012), pp. 1726–1736.

[30] Paul H. Kussie et al. "Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain". In: *Science* 274.5289 (1996), pp. 948–953.

[31] Ashley M. Buckle, Gideon Schreiber, and Alan R. Fersht. "Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-.ANG. resolution". In: *Biochemistry* 33.30 (1994), pp. 8878–8889.

[32] Michelle R. Arkin, Yinyan Tang, and James A. Wells. "Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing toward the Reality". In: *Chemistry & Biology* 21.9 (2014), pp. 1102–1114.

[33] T. Clackson and J. Wells. "A hot spot of binding energy in a hormone-receptor interface". In: *Science* 267.5196 (1995), pp. 383–386.

[34] T. Kortemme and D. Baker. "A simple physical model for binding energy hot spots in protein-protein complexes". In: *Proceedings of the National Academy of Sciences* 99.22 (2002), pp. 14116–14121.

[35] Nir London, Barak Raveh, and Ora Schueler-Furman. "Druggable protein–protein interactions – from hot spots to hot segments". In: *Current Opinion in Chemical Biology* 17.6 (2013), pp. 952–959.

[36] C. G. M. Wilson and M. R. Arkin. "Small-Molecule Inhibitors of IL-2/IL-2R: Lessons Learned and Applied". In: *Current Topics in Microbiology and Immunology*. Springer Berlin Heidelberg, 2010, pp. 25–59.

[37] Martin Karplus and J. Andrew McCammon. "Molecular dynamics simulations of biomolecules". In: *Nature Structural Biology* 9.9 (2002), pp. 646–652.

[38] Josep Gelpi et al. "Molecular dynamics simulations: advances and applications". In: *Advances and Applications in Bioinformatics and Chemistry* (2015), p. 37.

[39] Harold A. Scheraga, Mey Khalili, and Adam Liwo. "Protein-Folding Dynamics: Overview of Molecular Simulation Techniques". In: *Annual Review of Physical Chemistry* 58.1 (2007), pp. 57–83.

[40] *Understanding Molecular Simulation*. Vol. 1. Academic Press, 2002.

[41] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes". In: *Journal of Computational Physics* 23.3 (1977), pp. 327–341.

[42] Berk Hess et al. "LINCS: A linear constraint solver for molecular simulations". In: *Journal of Computational Chemistry* 18.12 (1997), pp. 1463–1472.

[43] Marco De Vivo et al. "Role of Molecular Dynamics and Related Methods in Drug Discovery". In: *Journal of Medicinal Chemistry* 59.9 (2016), pp. 4035–4061.

[44] Hans C. Andersen. "Molecular dynamics simulations at constant pressure and/or temperature". In: *The Journal of Chemical Physics* 72.4 (1980), pp. 2384–2393.

[45] Shūichi Nosé. "A molecular dynamics method for simulations in the canonical ensemble". In: *Molecular Physics* 52.2 (1984), pp. 255–268.

[46] H. J. C. Berendsen et al. "Molecular dynamics with coupling to an external bath". In: *The Journal of Chemical Physics* 81.8 (1984), pp. 3684–3690.

[47] Tom Darden, Darrin York, and Lee Pedersen. "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems". In: *The Journal of Chemical Physics* 98.12 (1993), pp. 10089–10092.

[48] D. S. Wishart, B. D. Sykes, and F. M. Richards. "The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy". In: *Biochemistry* 31.6 (1992), pp. 1647–1651.

[49] *Spectral Database for Organic Compounds.*

[50] Neil E. Jacobsen. *NMR Spectroscopy Explained.* John Wiley & Sons, Inc., 2007.

[51] Martin Karplus. "Contact Electron-Spin Coupling of Nuclear Magnetic Moments". In: *The Journal of Chemical Physics* 30.1 (1959), pp. 11–15.

[52] Bettina Keller et al. "On using oscillating time-dependent restraints in MD simulation". In: *Journal of Biomolecular NMR* 37.1 (2006), pp. 1–14.

[53] Yang Shen and Ad Bax. "SPARTA$^+$ modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network". In: *Journal of Biomolecular NMR* 48.1 (2010), pp. 13–22.

[54] Kyle A. Beauchamp et al. "Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements". In: *Journal of Chemical Theory and Computation* 8.4 (2012), pp. 1409–1414.

[55] Jagna Witek et al. "Kinetic Models of Cyclosporin A in Polar and Apolar Environments Reveal Multiple Congruent Conformational States". In: *Journal of Chemical Information and Modeling* 56.8 (2016), pp. 1547–1562.

[56] Nuria Plattner and Frank Noé. "Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models". In: *Nat. Commun.* 6 (2015), pp. 7653–10.

[57] Frank Noé et al. "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations." In: *Proc. Natl. Acad. Sci. U.S.A.* 106 (2009), pp. 19011–19016.

[58] Jan-Hendrik Prinz et al. "Markov models of molecular kinetics: Generation and validation". In: *The Journal of Chemical Physics* 134.17 (2011), p. 174105.

[59] Feliks Nüske et al. "Variational Approach to Molecular Kinetics". In: *Journal of Chemical Theory and Computation* 10.4 (2014), pp. 1739–1752.

[60] F. Vitalini, F. Noé, and B. G. Keller. "A Basis Set for Peptides for the Variational Approach to Conformational Kinetics". In: *Journal of Chemical Theory and Computation* 11.9 (2015), pp. 3992–4004.

[61] Peter Deuflhard and Marcus Weber. "Robust Perron cluster analysis in conformation dynamics". In: *Linear Algebra Appl.* 398 (2005), pp. 161–184.

[62] Susanna Röblitz and Marcus Weber. "Fuzzy spectral clustering by PCCA: application to Markov state models and data classification". In: *Advances in Data Analysis and Classification* 7.2 (2013), pp. 147–179.

[63] Gregory R. Bowman, Vijay S. Pande, and Frank Noé, eds. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Springer Netherlands, 2014.

[64] Christof Schütte et al. "Markov state models based on milestoning". In: *The Journal of Chemical Physics* 134.20 (2011), p. 204105.

[65] Oliver Lemke and Bettina G. Keller. "Density-based cluster algorithms for the identification of core sets". In: *The Journal of Chemical Physics* 145.16 (2016), p. 164104.

[66] W. Kabsch. "A solution for the best rotation to relate two sets of vectors". In: *Acta Crystallographica Section A* 32.5 (1976), pp. 922–923.

[67] Sarah A. Mueller Stein et al. "Chapter 13 Principal Components Analysis: A Review of its Application on Molecular Dynamics Data". In: *Annual Reports in Computational Chemistry*. Elsevier, 2006, pp. 233–261.

[68] Florian Sittel, Abhinav Jain, and Gerhard Stock. "Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates". In: *The Journal of Chemical Physics* 141.1 (2014), p. 014111.

[69] Guillermo Pérez-Hernández et al. "Identification of slow molecular order parameters for Markov model construction". In: *The Journal of Chemical Physics* 139.1 (2013), p. 015102.

[70] Jeffrey R. Wagner et al. "Emerging Computational Methods for the Rational Discovery of Allosteric Drugs". In: *Chemical Reviews* 116.11 (2016), pp. 6370–6390.

[71] S. Vinga. "Information theory applications for biological sequence analysis". In: *Briefings in Bioinformatics* 15.3 (2013), pp. 376–389.

[72] Adam T. VanWart et al. "Exploring Residue Component Contributions to Dynamical Network Models of Allostery". In: *Journal of Chemical Theory and Computation* 8.8 (2012), pp. 2949–2961.

[73] Christopher L. McClendon et al. "Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles". In: *Journal of Chemical Theory and Computation* 5.9 (2009), pp. 2486–2502.

[74] Bettina Keller, Zrinka Gattin, and Wilfred F. van Gunsteren. "What stabilizes the 314-helix in β3-peptides? A conformational analysis using molecular simulation". In: *Proteins: Structure, Function, and Bioinformatics* (2010), pp. 1677–1690.

[75] Xuan-Yu Meng et al. "Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery". In: *Current Computer Aided-Drug Design* 7.2 (2011), pp. 146–157.

[76] Li C. Xue et al. "Computational prediction of protein interfaces: A review of data driven methods". In: *FEBS Letters* 589.23 (2015), pp. 3516–3526.

[77] Ilya A. Vakser. "Protein-Protein Docking: From Interaction to Interactome". In: *Biophysical Journal* 107.8 (2014), pp. 1785–1793.

[78] A. Vangone et al. "Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1". In: *Proteins: Structure, Function, and Bioinformatics* 85.3 (2016), pp. 417–423.

[79] Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. "HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information". In: *Journal of the American Chemical Society* 125.7 (2003), pp. 1731–1737.

[80] G.C.P. van Zundert et al. "The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes". In: *Journal of Molecular Biology* 428.4 (2016), pp. 720–725.

[81] Matthew Brudner et al. "Lectin-Dependent Enhancement of Ebola Virus Infection via Soluble and Transmembrane C-type Lectin Receptors". In: *PLoS ONE* 8.4 (2013). Ed. by Bradley S. Schneider, e60838.

[82] Alex N Zelensky and Jill E Gready. "The C-type lectin-like domain superfamily". In: *FEBS Journal* 272.24 (2005), pp. 6179–6217.

[83] Michel Thépaut et al. "Structural Studies of Langerin and Birbeck Granule: A Macromolecular Organization Model†‡". In: *Biochemistry* 48.12 (2009), pp. 2684–2698.

[84] Hadar Feinberg et al. "Structural Basis for Langerin Recognition of Diverse Pathogen and Mammalian Glycans through a Single Binding Site". In: *Journal of Molecular Biology* 405.4 (2011), pp. 1027–1039.

[85] Hadar Feinberg et al. "Structural Basis for Langerin Recognition of Diverse Pathogen and Mammalian Glycans through a Single Binding Site". In: *Journal of Molecular Biology* 405.4 (2011), pp. 1027–1039.

[86] Juan C. Muñoz-García et al. "Langerin–Heparin Interaction: Two Binding Sites for Small and Large Ligands As Revealed by a Combination of NMR Spectroscopy and Cross-Linking Mapping Experiments". In: *Journal of the American Chemical Society* 137.12 (2015), pp. 4100–4110.

[87] W. Weis et al. "Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid". In: *Nature* 333.6172 (1988), pp. 426–431.

[88] Takashi Ohta et al. "Glycotentacles: Synthesis of Cyclic Glycopeptides, Toward a Tailored Blocker of Influenza Virus Hemagglutinin". In: *Angewandte Chemie International Edition* 42.42 (2003), pp. 5186–5189.

[89] Moritz Waldmann et al. "A Nanomolar Multivalent Ligand as Entry Inhibitor of the Hemagglutinin of Avian Influenza". In: *Journal of the American Chemical Society* 136.2 (2014), pp. 783–788.

[90] Susanne Liese et al. "Hydration Effects Turn a Highly Stretched Polymer from an Entropic into an Energetic Spring". In: *ACS Nano* 11.1 (2016), pp. 702–712.

[91] William L. Jorgensen et al. "Comparison of simple potential functions for simulating liquid water". In: *The Journal of Chemical Physics* 79.2 (1983), pp. 926–935.

[92] David Van Der Spoel et al. "GROMACS: Fast, flexible, and free". In: *Journal of Computational Chemistry* 26.16 (2005), pp. 1701–1718.

[93] *Marvin Sketch version 6.2.2 , calculation module developed by ChemAxon*. 2014. URL: http://www.chemaxon.com/products/marvin/marvinsketch/.

[94] Alan W Sousa da Silva and Wim F Vranken. "ACPYPE - AnteChamber PYthon Parser interfacE". In: *BMC Research Notes* 5.1 (2012), p. 367.

[95] Junmei Wang et al. "Development and testing of a general amber force field". In: *Journal of Computational Chemistry* 25.9 (2004), pp. 1157–1174.

[96] Ross C. Walker, Michael F. Crowley, and David A. Case. "The implementation of a fast and accurate QM/MM potential method in Amber". In: *Journal of Computational Chemistry* 29.7 (2008), pp. 1019–1031.

[97] Randall S. Dumont. "A metropolis Monte Carlo method for computing microcanonical statistical rate constants". In: *Journal of Computational Chemistry* 12.3 (1991), pp. 391–401.

[98] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical sampling through velocity rescaling". In: *The Journal of Chemical Physics* 126.1 (2007), p. 014101.

[99] M. Parrinello and A. Rahman. "Polymorphic transitions in single crystals: A new molecular dynamics method". In: *Journal of Applied Physics* 52.12 (1981), pp. 7182–7190.

[100] *MATLAB*. 2016b.

[101] *Python Software Foundation. Python Language Reference, version 2.7. Available at.* URL: http://www.python.org.

[102] C. L. Will and R. Luhrmann. "Spliceosome Structure and Function". In: *Cold Spring Harbor Perspectives in Biology* 3.7 (2010), a003707–a003707.

[103] Hsin-Chou Chen and Soo-Chen Cheng. "Functional roles of protein splicing factors". In: *Bioscience Reports* 32.4 (2012), pp. 345–359.

[104] Michael Kofler et al. "Proline-rich Sequence Recognition". In: *Molecular & Cellular Proteomics* 8.11 (2009), pp. 2461–2473.

[105] Xin Huang et al. In: *Nature Structural Biology* 7.8 (2000), pp. 634–638.

[106] Linda J. Ball et al. "Recognition of Proline-Rich Motifs by Protein-Protein-Interaction Domains". In: *Angewandte Chemie International Edition* 44.19 (2005), pp. 2852–2869.

[107] Xiaojuan Huang et al. "Structure and Function of the Two Tandem WW Domains of the Pre-mRNA Splicing Factor FBP21 (Formin-binding Protein 21)". In: *Journal of Biological Chemistry* 284.37 (2009), pp. 25375–25387.

[108] Stefan Klippel et al. "Multivalent Binding of Formin-binding Protein 21 (FBP21)-Tandem-WW Domains Fosters Protein Recognition in the Pre-spliceosome". In: *Journal of Biological Chemistry* 286.44 (2011), pp. 38478–38487.

[109] Martin K. Scherer et al. "PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models". In: *Journal of Chemical Theory and Computation* 11.11 (2015), pp. 5525–5542.

[110] Eric F. Pettersen et al. “UCSF Chimera?A visualization system for exploratory research and analysis”. In: *Journal of Computational Chemistry* 25.13 (2004), pp. 1605–1612.

[111] Abil E. Aliev et al. “Motional timescale predictions by molecular dynamics simulations: Case study using proline and hydroxyproline sidechain dynamics”. In: *Proteins: Structure, Function, and Bioinformatics* 82.2 (2013), pp. 195–215.

[112] William Humphrey, Andrew Dalke, and Klaus Schulten. “VMD – Visual Molecular Dynamics”. In: *Journal of Molecular Graphics* 14 (1996), pp. 33–38.

[113] Wolfgang Kabsch and Christian Sander. “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features”. In: *Biopolymers* 22.12 (1983), pp. 2577–2637.

[114] Robert T. McGibbon et al. “MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories”. In: *Biophysical Journal* 109.8 (2015), pp. 1528 –1532.

[115] B. Zhao et al. “Inactivation of YAP oncoprotein by the Hippo pathway is involved in cell contact inhibition and tissue growth control”. In: *Genes & Development* 21.21 (2007), pp. 2747–2761.

[116] Brett J. Schuchardt et al. “Ligand binding to WW tandem domains of YAP2 transcriptional regulator is under negative cooperativity”. In: *FEBS Journal* 281.24 (2014), pp. 5532–5551.

[117] Mark A. Verdecia et al. In: *Nature Structural Biology* 7.8 (2000), pp. 639–643.

Curriculum Vitae

For reasons of data protection, the curriculum vitae is not published in the electronic version.