Multi-scale Modeling of Bacterial Proteasome Core Particle Assembly

By

Pushpa Itagi

Submitted to the graduate degree program in the Center for Computational Biology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chair: Dr. Yinglong Miao

_____

Co-Chair: Dr. Eric J. Deeds

_____

Dr. Ilya Vakser

_____

Dr. Joanna Slusky

_____

Dr. Jeroen Roelofs

Date Defended: 07/08/2021

The dissertation committee for Pushpa Itagi certifies that this is the approved version of the following dissertation:

Multi-scale Modeling of Bacterial Proteasome Core Particle Assembly

Chair: Dr. Yinglong Miao

Co-Chair: Dr. Eric J. Deeds

Date Approved: 07/08/2021

Abstract

Molecular machines play a central role in cellular processes like signal transduction, motility, genome duplication, transport, protein synthesis, protein degradation, and many more. These machines cannot be synthesized directly by the cell but are assembled from individual subunits through non-covalent interactions. Most of the molecular machines studied to date have evolved from ordered and hierarchical assembly pathways. In this dissertation, we focused on the assembly of a specific molecular machine – the proteasome. The proteasome is a critical component of intracellular protein degradation and is involved in numerous processes such as cell growth, maintenance, and cell division. The proteasome consists of a proteolytically-active 20S Core Particle (CP) and a 19S Regulatory Particle (RP) that binds the CP and recognizes proteins tagged for degradation. The architecture of the CP is conserved across archaea, bacteria, and eukaryotes. The CP is formed from 14 α and 14 β subunits, arranged in a barrel-shaped complex of four stacked rings in an $\alpha_7\beta_7\beta_7\alpha_7$ arrangement. The β proteins are the catalytically active subunits, whereas the α subunits serve to bind the RP and ensure that only proteins targeted for degradation enter the CP. The proteasome CP is active only when fully assembled, which begins with the formation of Half Proteasomes (HP- $\alpha_7\beta_7$), and then two HPs dimerize into a CP. All the β subunits are synthesized with an N-terminal propeptide, autocatalytically cleaved off when two HPs dimerize, and only then the CP becomes active. In this dissertation, I elucidate different steps of CP assembly in the bacterium *Rhodococcus erythroplis* (*Re*). First, we are investigating the molecular mechanism behind the separation of time scales observed in CP assembly. Experimental work on bacterial proteasomes has shown that the HP forms completely within minutes, while HP dimerization to form CPs takes hours to complete. These studies also suggested that the β propeptide plays a role in regulating the dimerization rate. Using all-atom Molecular Dynamics (MD), we investigated the role of the β propeptide in *Re* CP assembly and showed that the length and polarity in the propeptide impact dimerization rate. We further validated these findings experimentally. In all species, CP assembly always occurs with two HPs dimerizing, and we never observe CP with one or two missing subunits. We hypothesized that there exists some allosteric communication among the subunits to prevent

dimerization of near-HP structures ($\alpha_6\beta_7$, $\alpha_7\beta_6$, and $\alpha_6\beta_6$) with each other or an HP. Our Molecular simulations revealed a global conformational shift in the β subunits that causes significant conformational transitions making the near-HP structures in a non-dimerizable state. Next, we investigated the formation of HPs in three assembly CP pathways, using coarse-grained ODE mathematical models based on chemical reaction kinetics theory. Our results discuss kinetic trapping and CP assembly dynamics in different pathways. Our mathematical models reveal that there is a tradeoff between speed and robustness in these assembly pathways. Ultimately, our simulation findings have helped us address long-standing questions in the proteasome assembly field and obtain structural and molecular insights. Further, our results will lay a promising foundation for structure-based drug design for designing specific, efficient, and less toxic small-molecule proteasome assembly inhibitors to treat tuberculosis and other diseases. Lastly, our findings on self-assembling bacterial CPs will serve as a proof of concept for designing nanomachines and other nanotechnology applications.

Acknowledgments

There are several people whom I would like to thank and appreciate for supporting and encouraging me throughout my graduate school, without whom this journey would not be possible.

I sincerely express my gratitude and the highest level of appreciation to my advisor Dr. Eric J. Deeds; he has guided me with immense support, encouragement, and patience to pursue my research and has given me the freedom to explore and learn from several computational tools and techniques. I am incredibly grateful and fortunate to have worked with him and gained knowledge on macromolecular assemblies and computational models. He has been very kind to take me in the group and has been very patient in guiding me to communicate, present and write research findings effectively. He is one of the best and engaging professors I have had in my academic journey.

I am also highly grateful to committee members and mentors, Dr. Ilya Vakser, Dr. Joanna Slusky, Dr. Yinglong Miao, and Dr. Jeroen Roelofs. They have supported and guided my research and have provided their time, insightful discussions, and the scientific knowledge which has helped me grow as a researcher. In addition, I am very thankful to Dr. Wonpil Im, who guided and inspired me to learn Molecular Dynamics Simulations. Finally, I would like to thank Dr. Christian Ray, Dr. Roberto de Guzman, Dr. John Karanicolas, and the late Dr. Mark Richter for their guidance and insightful discussions.

The work in my dissertation would not have been possible without my lab members, Anupama Kante, for being my fellow lab member and friend, from providing experimental data, discussing research findings, to the good memories we have had in all these years, Leo Lagunes, for her positive words and insights in research and writing. Maulik Nariya for being a kind and helping a friend and our lab alumni, Breanne Sparta, Shamus Cooley, Tim Hamilton, Serena Hughes for their helpful discussions. I am also thankful to the Im lab members, Jumin Lee and Seonghoon Kim, for being patient and assisting me on MD simulations. Finally, I would like to thank all the faculty, colleagues, and friends at the Center for Computational Biology-KU and Quantitative Biosciences-UCLA for their discussions, and support. I am also grateful to my master's research advisor Dr. Abhijit Mitra-IIITH, and my Undergraduate research advisor Zabin Bagewadi-BVBCET and all the other faculty who guided me at both institutions.

I am very grateful to have a special group of people who have been very loving and supporting. Ashwini Bennale, my best friend since childhood, has always shown immense positivity, love, and encouragement towards me while pursuing my ambitions. Nootan Pandey has been strong support during my challenging graduate school days. She has always been caring and my go-to person in Lawrence and beyond. Soumyaroop Nandi, for being a caring and patient friend and to explore California, national parks, and many cities in the last few years. Pallavi Biswas has been very supportive and for being there since

beginning of graduate school. I would also like to thank Hara Madhav, Huijing Wang, Shristi Pawnikar, and other friends who have been there and provided motivation.

My journey would have been impossible without my wonderful and loving family, my parents, Dr. Sunanda Itagi and Dr. Kotrabasappa Itagi, my sister Sumati Itagi, my brother-in-law Swarup Hiremath, and the newest member Nihit Hiremath. I am here only because of you all, and I am fortunate to be with you all. My mother, Sunanda, has been my inspiration to pursue my doctoral degree and to be resilient and is my unwavering support system. My father is always encouraging me to be persistent and have a positive attitude.

My sister and my best friend for her kindness and encouragement and for being my constant in life. They have provided me with unconditional love, support, and nurturing. Lastly, I would like to thank all my cousins and extended family members who have supported me while I chased my dreams. Finally, I am highly grateful to God for giving me the opportunity to pursue research, survive a global pandemic, and being blessed with amazing mentors, a loving family, and a precious bunch of friends.

This thesis is dedicated to my loving parents and sister; without you, my journey was impossible.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Molecular machines are complexes that play an essential role in numerous biological processes. These machines are highly complex and efficient. Molecular machines are assembled from a set of subunits into a fully functional form; forces of evolution very likely guide the assembly of these complexes. A few examples include the proteasome, ribosome, spliceosome, myosin, motor proteins, ATP synthase, and many more. This thesis presents a detailed study of a highly critical macromolecular stacked ring complex - the proteasome, a central component of intracellular protein degradation. The proteasome complex is predominantly involved in the degradation of unwanted or damaged cellular proteins, regulates various cell processes, and plays a crucial role in maintaining homeostasis. Proteasome dysregulation has been associated with autoimmune diseases, cancer, neurodegenerative disorders, cardiomyopathies, and several other conditions.

Degradation of cellular proteins is essential in signal transduction, proteostasis, signaling, and cell cycle regulation [1-3]. Proteasomes are ubiquitous to life and are found in prokaryotes and eukaryotes. Not all prokaryotes have proteasomes; only archaea and a subset of bacteria (actinomycetes) have these complexes. The actinomycetes have most likely acquired the proteasome through a horizontal gene transfer. The other bacteria like *Escherichia coli* have simpler proteases like HslV/ClpQ, which share the same catalytic mechanism as proteasomes [3, 4]. In eukaryotes, proteasomes are found in the nucleus, attached to the endoplasmic reticulum, and in the cytosol and are essential for the viability of eukaryotic cells [3, 4]. Also, in eukaryotes, proteasomes are a part of the ubiquitin-proteasome system (UPS), which plays a central role in numerous regulatory pathways, protein quality control, antigen presentation, and other functions.

Prokaryotes have a simpler form of proteasome formed by four stacked rings, and eukaryotes have the most. complicated form, i.e., 26S proteasome. [3-6].

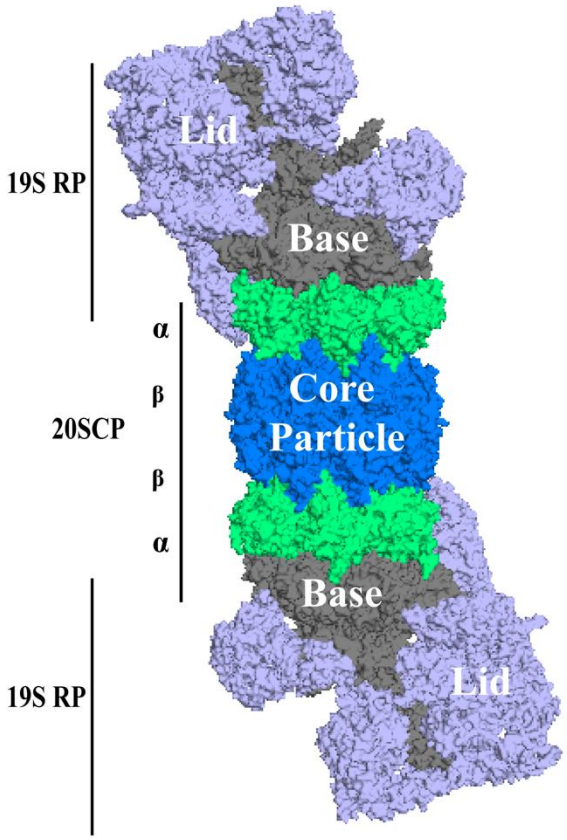


Figure 1.1: Structure of the human 26S proteasome (PDB ID:5GJR) comprising of 19S Regulatory Particle (RP) and 20S Core Particle (CP). The CP consists of four α (green) and β (blue) stacked rings. Figure rendered in Pymol.

This complex 26S proteasome (2.5 MDa) is a central protease involved in intracellular protein degradation. It consists of a 20S catalytic Core Particle (CP) of approximately 700KDa and capped by one or two 19S Regulatory Particles (RP) (**Fig. 1.1**). The 19S RP can be divided into the base and lid components, and structural studies on RP are currently relatively limited [7]. The 19S RP is found only in eukaryotes and thus is not conserved in the evolution. RP is involved in recognizing the proteins targeted for degradation and then translocating them to the CP [3, 8]. The CP complex is very well studied and

characterized and is about 15 nm in length and 11nm in diameter [9]. The active sites reside in the CP, which is a barrel-shaped complex stacked of four heptameric rings.

Prokaryotes do not have the 19S RP, but instead have AAA+ ATPases (like Pan, Arc, and Mpa), which help in protein substrate unfolding and translocation to the CP [10]. Much research is still undergoing to understand the assembly and functions of the RP. Our work focuses on studying the assembly of the proteasome CP, and more on its quaternary structure and assembly in bacterial cells is described below. The 20S CP is found in all three kingdoms archaea, bacteria, and eukaryotes, and its overall structure is highly conserved [2, 3]. It consists of four stacked rings, each ring composed of seven subunits. The two outermost rings are made up of seven α subunits each. In contrast, the two inner rings are made up of seven β subunits, each, and together these are arranged in the stoichiometric form $\alpha_7\beta_7\beta_7\alpha_7$ (**Fig. 1.2**).

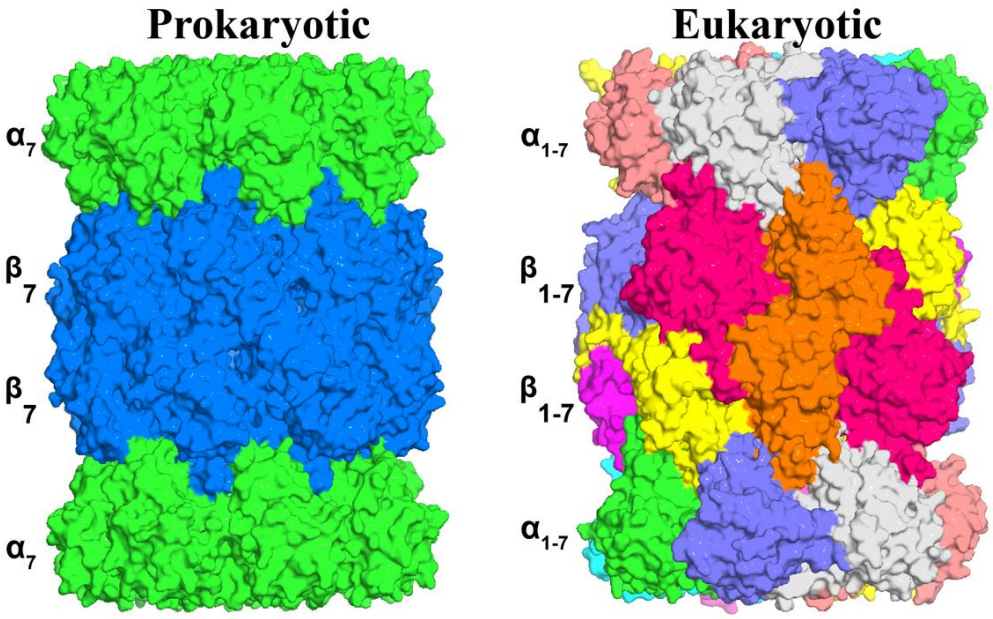


Figure 1.2: Surface representation of proteasome Core Particle structures from prokaryote (*Rhodococcus erythropolis*; PDB:1Q5Q) and eukaryote (*Saccharomyces cerevisiae*; PDB:5L52). The schematic shows that eukaryotic CP has seven different alpha and beta subunits, whereas the prokaryotic CP is simpler and made of one or two alpha and beta subunits. Figure rendered in Pymol.

Prokaryotic proteasomes contain one or two types of α and β subunits, while eukaryotic proteasomes contain seven different but related α and seven different but related β subunits [2, 3, 6]. Both α and β subunits have the same protein fold of an α+β tertiary structure [11, 12]. The α subunit has a structural

significance and is known to be involved in the gating process for many archaea and eukaryotic CP's. In contrast, the β subunits are proteolytically active and contain the active site threonine.

The subunit composition difference between prokaryotic and eukaryotic is seen in (**Fig. 1.2**) as differences in overall shape, symmetry, and interfaces. The differences in subunit composition are likely due to evolutionary pressures and selections for more complex organisms. Additionally, all the14 β subunits of prokaryotic CP are catalytically active (generally having chymotryptic peptidase activity), whereas, in eukaryotic CP's only three subunits (β_1, β_2, and β_5) are catalytically active [4]. To be specific in most organisms, β_1 catalyzes caspase-like activity, β_2 tryptic, and β_5 chymotryptic activity [4].

It is unknown what drove the origin of four inactive and three active β subunits in eukaryotic cells [4, 13]. In higher eukaryotes, there are also immunoproteasome and thymoproteasomes, which use alternative β subunits and display altered proteolytic activity. Despite the difference in subunit complexity, the CP's overall structure is highly conserved in all known prokaryotic and eukaryotic proteasomes (**Fig. 1.2**). Another striking feature of eukaryotic proteasomes is that they require dedicated assembly factors and chaperones to guide CP assembly. The eukaryotic proteasome is made of seven distinct α and β subunits; it is more complicated and requires several dedicated chaperones like PAC1-PAC2, PAC3-PAC4, and UMP1 [3]. So far, no evidence exists that prokaryotes require such chaperones for proteasome assembly; the α and β subunits have been repeatedly shown to spontaneously self-assemble into a prokaryotic CP without any additional factors [14].

In all 20S proteasomes, the monomers first form Half Proteasomes (**HP: $\alpha_7\beta_7$**), which then dimerize to form an active CP (**Fig. 1.3**). To protect the catalytically active sites, β subunits are expressed in an inactive precursor form with a propeptide sequence at the N terminus. These N-terminal β propeptides are d auto-catalytically cleaved when two Half Proteasomes (HP) associate to form a fully active CP [1] (**Fig. 1.3**). This dimerization of half proteasomes thus triggers the autocatalytic processing of propeptides.

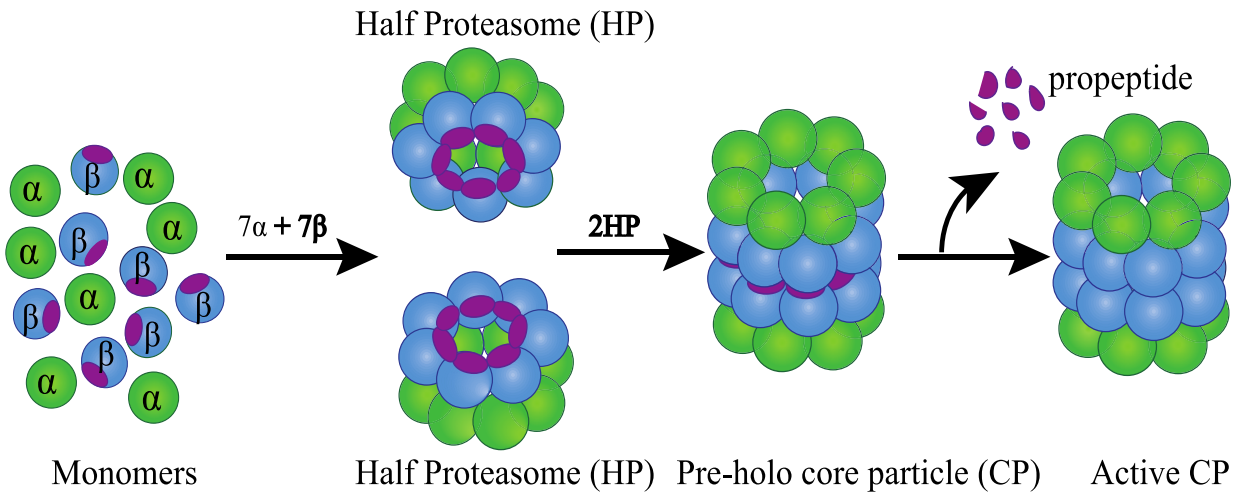


Figure 1.3: Generalized scheme of proteasome Core Particle assembly. The CP is active only after assembly. The α subunits are in green, β subunits in blue, and beta propeptides are in purple.

The CP is not active after until its fully assembled and the propeptide is autocatalytically cleaved off. Thus, the propeptide sequesters the active site from the environment until CP assembly is complete. These β subunit propeptides are of varied length and sequence, depending on the species. Some are only a few residues, while others are more than 60 residues long [9]. Previous studies have revealed that CP assembly is universally hierarchical, and that the HP is an obligate intermediate in all assembly pathways. However, the, β propeptides play different roles in the assembly and activation of CP depending on the organism [1]. Since the CP is only active when fully assembled, we expect that nature has evolved an efficient assembly pathway. Assembly of the CP in a regulated manner is highly critical because incorrect assembly might give rise to uncontrolled proteolysis (exposing active sites before assembly), non-specific cleavage, and accumulation of proteins, which destroys the cell's ability to maintain homeostasis and could eventually cause stress or even cell death. The proteasomes play a critical role in pathways fundamental to cell survival and proliferation, numerous specific inhibitors of the proteasome have been developed. Proteasome inhibition has become a novel drug target to treat multiple myeloma, malaria, neurodegenerative disorders, and other diseases. [15-19]. Over the last two decades, the FDA has approved three proteasome inhibitors: Bortezomib, Carfilzomib, and Ixazomib, all to treat multiple myeloma [20-24]. Along with the increasing crisis of antibiotic resistance, bacterial proteasomes are a promising drug target.

The bacterial pathogen *Mycobacterium tuberculosis* (*Mtb*) is known to have caused around 1 billion deaths in the past two centuries, underscoring the need to treat tuberculosis [25]. *Mtb* is a challenging bacterium to target since it has developed multidrug resistance [26]. Gene knock-out studies in mice have suggested that the proteasome is essential for virulence and survival to stress in *Mtb* [25]. These findings led to the identification of the *Mtb* proteasome as a novel drug target. Proteasome active sites are structurally conserved, making it challenging to develop the proteasome inhibitors targeting the pathogen and not human proteasomes [16].

Additionally, proteasome active site inhibitors can also non-specifically block the function of other cellular proteases that share the Ser/Thr active site architecture. To circumvent the non-specificity of active site inhibitors, we can exploit targeting proteasome assembly by designing assembly inhibitors instead of activity inhibitors. Targeting proteasome assembly has just begun to develop actively. It is an attractive approach because the subunit interfaces are different in human and *Mtb* proteasome, and the assembly pathways are likely different [25]. We thus anticipate that targeting proteasome assembly could have high specificity [27]. Since proteasomes are not active until fully assembled, small molecules that disrupt assembly represent an alternative approach to block proteasome function. To our knowledge no proteasome inhibitors are currently being investigated for tuberculosis in the clinic, further suggesting that assembly may be an attractive drug target. So far, *Mtb* CP assembly has not been well characterized, and thus understanding the assembly pathway and intermediates formed will help us design efficient inhibitors.

Besides the clinical relevance, hierarchical structures like proteasomes, the apoptosome, virus capsids, and other macromolecular complexes have likely experienced significant evolutionary pressure to assemble efficiently and quickly into a functional form. It is quite intriguing to imagine how the subunits of such complexes assemble in cells with so many possible combinations. Additionally, the cell must ensure that these complexes can find the interacting subunits despite macromolecular crowding. Another exciting application of studying self-assembly is material science, nanomachines, and protein cages. Prokaryotic proteasome assembly occurs spontaneously without any extrinsic factors; understanding the principles of self-assembly will be highly beneficial for designing emerging protein nanomachines, biomaterials, and for

synthetic biology in general.

CP assembly kinetics have been perhaps best characterized experimentally in the bacterium *Rhodococcus erythropolis* (*Re*). The *Mtb* proteasome shares about 64% sequence similarity with *Re*. Experimental studies in *R. erythropolis* have demonstrated that α and β subunits stay monomeric when expressed separately and spontaneously assemble *in vitro* into active CPs on being incubated together. Experimental findings also reveal that the HPs are formed almost immediately (seconds) after the reaction starts. The CP appears after a certain time lag and 100% CP assembly takes about 3 hours [28, 29]. This evidence suggests that dimerization of HP's is a rate-limiting step and indicates that HP is an obligatory intermediate in the hierarchical assembly pathway. The reason for this separation of timescales is currently unknown, but existing experimental evidence supports a key role for the propeptide in regulating dimerization. HP formation becomes slower if the propeptide is deleted, but HP dimerization is very fast if the propeptide is added *in trans* to an assembly reaction with the α and propeptide deletion β subunits. In that case CP assembly occurs nearly as fast as HP assembly [28]. While it is thus clear that the propeptide inhibits HP dimerization, the molecular mechanism is currently unknown. So far, the assembly pathways have been studied experimentally in the model systems of our interest primarily *R. erythropolis* and *T. acidophilum* [2, 30, 31]. So far, there has been no atomic level computational study or any molecular modeling of the prokaryotic proteasome. Thus, this approach is first of its kind to obtain atomic insights using molecular models and offers opportunities to gain in-depth understanding at molecular level for the CP assembly in the model systems of our interest.

One popular method to understand biological molecules and their molecular mechanisms is Molecular Dynamics (MD). In MD, atoms are considered spheres, and the bonds are treated as springs, and then for a collection of atoms, say proteins, Newton's equations of motion are numerically solved. It has three main components: the force field, which describes potential energy based on the laws defining the mutual interactions in the system of interest. Secondly, one must define an algorithm to integrate the equations of motion numerically, and several types of integrators like Velocity Verlet, and Leapfrog are available. Lastly, it needs a set of initial positions and velocities for all-atom systems. MD is fundamentally a

statistical mechanic method and helps explore a system's dynamics, structural and thermodynamic properties. The structural and thermodynamic properties at a level of resolution that is difficult to obtain experimentally. Hence, MD has emerged as an extremely valuable tool to understand conformational changes occurring in biological molecules like proteins, lipids, etc. All-atom simulations typically have a biological molecule of interest such as protein, lipids, DNA, or RNA and have water and ions to mimic the experimental conditions.

In the second chapter, we have employed all-atom MD simulations to elucidate the separation of time scales observed in *Re* CP assembly. From these simulations, we found that a region of the propeptide near the HP dimerization interface in *Re* is highly disordered and interacts with key residues in the HP interface important for dimerization. We also validated the MD simulation predictions with experimental findings. Understanding the molecular mechanism in *R. erythropolis* slow HP dimerization elucidated a critical step in the CP assembly pathway in detail and provides insights into the dimerization interface, an attractive target for assembly inhibitors. In the long term, this work will provide a framework to use structure-based approaches to design inhibitors of CP assembly. Particularly, studying the proteasome assembly pathway in species like pathogenic *M. tuberculosis* and the archaea, *Thermoplasma acidophilum* will aid in developing efficient and novel assembly inhibitors for tuberculosis and understand the evolutionary aspects of how ring-like complexes such as proteasome have evolved.

As discussed above, in all the organisms, CP assembly is hierarchical in that the HP is an obligate intermediate in all assembly pathways [32, 33]. To date, we have never seen evidence for a CP with missing subunits, i.e., an HP never dimerizes with another near-HP intermediate like $\alpha_6\beta_7$ (i.e., the HP missing a single α subunit), $\alpha_7\beta_6$, and $\alpha_6\beta_6$, etc. However, it is currently unknown how the subunits allosterically communicate to achieve hierarchical assembly pathways. In other words, how do the subunits "know" to first assemble the HP structure and only then dimerize? Allosteric communication likely occurs among the CP subunits, which allows the system to differentiate between obligatory and non-obligatory intermediates.

The third chapter focuses on a study of conformational allostery and how this allostery is communicated among intermediates to prevent incompatible intermediates from assembly into incorrect

CP-like structures. Determining the structure of intermediates involved in CP formation will be vital for further understanding the assembly pathway and molecular events that lead to CP formation. Crystal structure analysis, mass spectrometry, microscopy studies, and electron microscopy have provided evidence that the near-HP intermediates like $\alpha_6\beta_7$, $\alpha_7\beta_6$, and $\alpha_6\beta_6$ occur during assembly and are present for a significant time. However, these near-HP intermediates never dimerize with an HP or with one another to form atypical and incomplete structures of CP-like complexes. ($\alpha_6\beta_7\beta_7\alpha_7$, $\alpha_6\beta_6\beta_7\alpha_7$ or $\alpha_6\beta_7\beta_7\alpha_7$). The formation of such incomplete CP's could have drastic effects because of unregulated protein degradation and perturb homeostasis in cells. Understanding the allosteric regulation involved in preventing the dimerization of near-HP intermediates will help us elucidate the molecular events leading to CP assembly and understand the stability of these near-HP intermediates. A better understanding of the bacterial CP assembly pathway will provide a framework for designing new inhibitors that target CP biogenesis, impacting specific steps in the assembly pathway. The third chapter addresses why the near-HP intermediates do not dimerize with a true HP and assemble into a CP and how this information is communicated in the intermediates or other similar hierarchical assemblies.

While our findings provide insight into the separation of time scales in CP assembly and the role of allostery in CP assembly, many interesting questions are still unanswered. The proteasome CP and other proteins like AAA+ ATPases, GroEL, DNA binding proteins (e.x. RAD 52) all have a ring-like structure. These proteins would have evolved a strategy and method to overcome or avoid kinetic challenges to assemble efficiently [34]. A pioneering and elaborate study on ring-like structures provides an evolutionary rationale and understanding of self-assembly dynamics in biological molecules like the proteasome [34]. For ring-like structures, a form of kinetic trapping known as "deadlock" and evolutionary pressures have carved assembly pathways for such structures to avoid deadlock. Coarse-grained computational approaches are highly suitable to investigate the kinetics of assembly pathways for complicated structures. The coarse-graining approach is powerful and fast and allow us to understand long time-scale assembly phenomena in complex machines like the proteasome.

In the fourth chapter, we have used coarse-grained ODE models to understand the assembly kinetics of

CP assembly in *Re*. Specifically, we address the pathway adopted by the cell to form an HP and then dimerize it into a CP. Previous work based on limited evidence has led to the speculation that archaea and bacteria follow different pathways to assemble a CP [1]. Interestingly, our findings report a novel pathway that operates in *Re* in-vitro CP assembly. Our results are based on the ODE models, which follow a chemical-kinetics-based approach. In the future, these models can be explored with different parameters like association rates, binding affinity, and rate constants to understand the CP assembly for other species like *Mycobacterium tuberculosis*, *Thermoplasma acidophilum*, and even the highly complex eukaryotic CP's.

Lastly, the final Chapter 5 takes a different direction and discusses the comparative characterization of Croefelmer drug mixtures using Machine Learning and Data Mining approaches. Drugs that are manufactured using naturally occurring raw materials are highly complex and heterogeneous [35]. Croefelmer is a botanical complex mixture drug extracted from the sap of South American tree, *Croton lechleri*. It is an FDA-approved drug used to treat noninfectious diarrhea in HIV patients [36]. Such drugs derived from natural resources exhibit batch-to-batch variation and are extremely sensitive to manufacturing processes. Post-approval, these drugs can be subjected to changes in raw materials, manufacturing process, or other parameters. To ensure that the drug quality post-approval remains sufficient, extensive analytical characterization is required. Using machine learning classifiers, PCA, and mutual information approaches, we have developed a mathematical tool that can identify critical quality attributes (CQAs) that help in distinguishing different batches of Croefelmer, especially to identify if the batches are expired or not. This kind of comparative characterization can be extended to other biologic drugs to distinguish between batches and decide if two preparations of drugs are highly similar.

Chapters 2-4 utilize atomistic Molecular Dynamics, and coarse-grained (ODE) simulation approaches to understand the CP assembly and its dynamics in *Rhodococcus erythropolis*. Using all-atom Molecular Dynamics (MD) simulations, we have gathered fascinating insights on the role(s) of the propeptide in CP assembly. These atomistic simulations are explicit models and use the same salt concentration as used in proteasome *in-vitro* experiments. Additionally, we have studies that elucidate the allosteric regulation at

the atomic level among proteasome subunits, which helps prevent aberrant structures. This work further forms a promising foundation to implement similar studies in *Mtb* and higher-order organisms CP assembly. Moreover, our work potentially contributes to identifying novel inhibitors of proteasome CP assembly and can be applied to identify conformations of various intermediates in *Mtb*. In addition, our work can be applied to propose simple concepts for self-assembly in nanomaterials and protein nanomachines.

1.1 References

1. Sharon M, Witt S, Glasmacher E, Baumeister W, Robinson CV: Mass spectrometry reveals the missing links in the assembly pathway of the bacterial 20 S proteasome. *J Biol Chem* 2007, 282(25):18448-18457.
2. Kwon YD, Nagy I, Adams PD, Baumeister W, Jap BK: Crystal structures of the Rhodococcus proteasome with and without its pro-peptides: implications for the role of the pro-peptide in proteasome assembly. *J Mol Biol* 2004, 335(1):233-245.
3. António J. Marques RP, Ana C. Matias, Paula C. Ramos, and R. Jürgen Dohmen: Catalytic mechanism and assembly of the proteasome. *Chem Rev* 2009, 109(4):1509-1536.
4. Peter Zwickl WB: The Proteasome — Ubiquitin Protein Degradation Pathway, vol. 268; 2002.
5. Saeki Y, Tanaka K: Assembly and Function of the Proteasome. In: *Ubiquitin Family Modifiers and the Proteasome: Reviews and Protocols.* Edited by Dohmen RJ, Scheffner M. Totowa, NJ: Humana Press; 2012: 315-337.
6. Seemuller E, Lupas A, Baumeister W: Autocatalytic processing of the 20S proteasome. *Nature* 1996, 382(6590):468-471.
7. Kim HM, Yu Y, Cheng Y: Structure characterization of the 26S proteasome. *Biochim Biophys Acta* 2011, 1809(2):67-79.
8. Saeki Y, Toh-e A, Kudo T, Kawamura H, Tanaka K: Multiple Proteasome-Interacting Proteins Assist the Assembly of the Yeast 19S Regulatory Particle. *Cell* 2009, 137(5):900-913.
9. Seemüller E, Zwickl P, Baumeister W: 12. Self-Processing of Subunits of the Proteasome. *Enzymes* 2002, 22.
10. Maupin-Furlow J: Proteasomes and protein conjugation across domains of life. *Nature Reviews Microbiology* 2012, 10(2):100-111.
11. Lowe J, Stock D, Jap B, Zwickl P, Baumeister W, Huber R: Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 A resolution. *Science* 1995, 268(5210):533-539.
12. Groll M, Ditzel L, Löwe J, Stock D, Bochtler M, Bartunik HD, Huber R: Structure of 20S proteasome from yeast at 2.4Å resolution. *Nature* 1997, 386(6624):463-471.
13. Zwickl P: The 20S proteasome. *Curr Top Microbiol Immunol* 2002, 268:23-41.
14. Becker SH, Darwin KH: Bacterial Proteasomes: Mechanistic and Functional Insights. *Microbiol Mol Biol Rev* 2017, 81(1).
15. Basse N, Montes M, Marechal X, Qin L, Bouvier-Durand M, Genin E, Vidal J, Villoutreix BO, Reboud-Ravaux M: Novel organic proteasome inhibitors identified by virtual and in vitro screening. *J Med Chem* 2010, 53(1):509-513.
16. Cheng Y, Pieters J: Novel proteasome inhibitors as potential drugs to combat tuberculosis. *J Mol Cell Biol* 2010, 2(4):173-175.
17. Durrant JD, McCammon JA: Molecular dynamics simulations and drug discovery. *BMC Biology* 2011, 9(1):71.

18. Joazeiro CAP, Anderson KC, Hunter T: Proteasome Inhibitor Drugs on the Rise. *Cancer Research* 2006, 66(16):7840.
19. Kingwell K: Proteasome target to tackle non-replicating TB. *Nat Rev Drug Discovery* 2009, 8.
20. Bonvini P, Zorzi E, Basso G, Rosolen A: Bortezomib-mediated 26S proteasome inhibition causes cell-cycle arrest and induces apoptosis in CD-30+ anaplastic large cell lymphoma. *Leukemia* 2007, 21(4):838-842.
21. Manasanch EE, Orlowski RZ: Proteasome inhibitors in cancer therapy. *Nature Reviews Clinical Oncology* 2017, 14(7):417-433.
22. R. C. Kane ATF, R. Sridhara, and R. Pazdur: United States Food and Drug Administration approval summary: bortezomib for the treatment of progressive multiple myeloma after one prior therapy. *Clinical Cancer Research* 2006, 12:2955-2960.
23. Kuhn DJ, Chen Q, Voorhees PM, Strader JS, Shenk KD, Sun CM, Demo SD, Bennett MK, van Leeuwen FWB, Chanan-Khan AA *et al*: Potent activity of carfilzomib, a novel, irreversible inhibitor of the ubiquitin-proteasome pathway, against preclinical models of multiple myeloma. *Blood* 2007, 110(9):3281-3290.
24. Raedler LA: Ninlaro (Ixazomib): First Oral Proteasome Inhibitor Approved for the Treatment of Patients with Relapsed or Refractory Multiple Myeloma. *Am Health Drug Benefits* 2016, 9(Spec Feature):102-105.
25. Gandotra S, Schnappinger D, Monteleone M, Hillen W, Ehrt S: In vivo gene silencing identifies the Mycobacterium tuberculosis proteasome as essential for the bacteria to persist in mice. *Nat Med* 2007, 13(12):1515-1520.
26. Li D, Li H, Wang T, Pan H, Lin G, Li H: Structural basis for the assembly and gate closure mechanisms of the Mycobacterium tuberculosis 20S proteasome. *EMBO J* 2010, 29(12):2037-2047.
27. Zwickl P, Voges D, Baumeister W: The proteasome: a macromolecular assembly designed for controlled proteolysis. *Philos Trans R Soc Lond B Biol Sci* 1999, 354(1389):1501-1511.
28. Zuhl F, Seemuller E, Golbik R, Baumeister W: Dissecting the assembly pathway of the 20S proteasome. *FEBS Lett* 1997, 418(1-2):189-194.
29. Suppahia A IP, Burris A, Kim FMG, Vontz A, Kante A, Kim S, Im W, Deeds EJ, Roelofs J.: Cooperativity in Proteasome Core Particle Maturation. *iScience* 2020, 23(5).
30. Witt S, Kwon YD, Sharon M, Felderer K, Beuttler M, Robinson CV, Baumeister W, Jap BK: Proteasome assembly triggers a switch required for active-site maturation. *Structure* 2006, 14(7):1179-1188.
31. Campbell MG, Veesler D, Cheng A, Potter CS, Carragher B: 2.8 A resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy. *Elife* 2015, 4.
32. Panfair D, Ramamurthy A, Kusmierczyk AR: Alpha-ring Independent Assembly of the 20S Proteasome. *Sci Rep* 2015, 5:13130.
33. Shigeo Murata HYaKT: Molecular mechanisms of proteasome assembly. *Nature reviews* 2009, 10.
34. Deeds EJ, Bachman JA, Fontana W: Optimizing ring assembly reveals the strength of weak interactions. *Proc Natl Acad Sci U S A* 2012, 109(7):2348-2353.
35. Nariya MK, Kim JH, Xiong J, Kleindl PA, Hewarathna A, Fisher AC, Joshi SB, Schöneich C, Forrest ML, Middaugh CR *et al*: Comparative Characterization of Crofelemer Samples Using Data Mining and Machine Learning Approaches With Analytical Stability Data Sets. *J Pharm Sci* 2017, 106(11):3270-3279.
36. Frampton JE: Crofelemer: a review of its use in the management of non-infectious diarrhoea in adult patients with HIV/AIDS on antiretroviral therapy. *Drugs* 2013, 73(10):1121-1129.

Some of the results in this chapter are published in Suppahia, A., **Itagi, P**., Deeds, E. J., & Roelofs, J, et al. (2020). Cooperativity in Proteasome Core Particle Maturation. *iScience*, *23*(5).

Chapter 2

Understanding the Separation of Timescales in *Rhodococcus erythropolis* Proteasome Core Particle Assembly

2.1 Introduction

The degradation of proteins is an essential step in signal transduction, proteostasis, and the regulation of biochemical pathways [1-3]. The 26S proteasome, a massive 2.5 MDa molecular machine, is a central protease involved in intracellular protein degradation. In eukaryotes, the catalytically active 20S Core Particle (CP) is capped by two 19S Regulatory Particles (RP's), forming the 26S proteasome. The CP consists of four heptameric rings, which are stacked coaxially in a barrel-shaped structure. The α and β subunits form the outer and inner rings, respectively, with an $\alpha_7\beta_7\beta_7\alpha_7$ stoichiometry [1, 4]. The α subunits interact with regulatory particles and help control the target substrate's entry into the barrel, while the β subunits are catalytically active and carry out proteolysis. The CP is found in all three kingdoms of life, archaea, bacteria, and eukaryotes, and its overall quaternary structure is highly conserved [5, 6].

Like many molecular machines, the proteasome cannot be synthesized by the cell in an active form. Instead, it is assembled from a set of subunits into a functional quaternary structure. The proteasome CP is active only when it is fully assembled, since the β subunits are initially expressed in an inactive precursor form with a propeptide sequence at the N terminus (**Fig. 2.1A**). As a result, understanding CP assembly and

biogenesis is critical to our overall understanding of the proteasome's function and regulation *in vivo*. In particular, the proteasome is well established as a drug target for treating a variety of diseases, including cancer and tuberculosis [7-9]. Traditional approaches to targeting the proteasome have focused on small molecules like Bortezomib that directly bind to the active site and disrupt proteolysis [7, 10-12]. However, it has been suggested that inhibiting assembly could offer an alternative and relatively unexplored approach to pharmacologically disrupting proteasome function. This is particularly important for *Mycobacterium tuberculosis* (*Mtb*); it has been shown that disrupting proteasome function can ameliorate chronic *Mtb* infections, but therapies targeting proteasome assembly have yet to be developed for clinical applications [7]. A better understanding of the assembly pathways and mechanisms underlying bacterial CP biogenesis could eventually lead to a new class of therapeutics for diseases like tuberculosis.

The assembly of the CP has been studied experimentally in a wide variety of organisms [3, 13-16]. In all cases, the CP assembly pathway involves the formation of Half Proteasomes (HP: $\alpha_7\beta_7$) from α and β subunits. During CP assembly, two HPs dimerize to form the pre-holo-CP, and then the propeptide is auto-catalytically cleaved off to form the active CP [2, 17] (**Fig. 2.1A**). While assembly pathways that form the HP differ between organisms [2, 13, 18], the HP is an obligatory intermediate in all organisms studied to date [2, 13, 18-20]. Given the proteasome's intricate quaternary structure, we expect that CP assembly has evolved to be efficient, accurate, and robust. Particularly when considering that incorrect assembly could give rise to uncontrolled proteolysis, non-specific cleavage, and accumulation of proteins if active sites remain exposed [3, 15]. Exactly how the assembly pathway achieves accurate CP assembly, however, is currently unclear.

CP assembly kinetics have been perhaps best characterized experimentally in the bacterium *Rhodococcus erythropolis* (*Re*). Over twenty years ago, Baumeister and colleagues demonstrated that the *Re* α and β subunits remain monomeric when expressed and purified independently [19, 21]. This property allows us to monitor both HP CP formation *in vitro* [15, 19, 20]. Baumeister et al. showed that the HPs are fully formed almost immediately after the subunits are mixed, with complete assembly of the HP observed

at 30 seconds. However, fully formed CPs are visible after a considerable time lag of about 30 minutes, and 100% CP assembly takes up to 3 hours [2, 19]. This evidence suggests that dimerization of HP's is a rate-

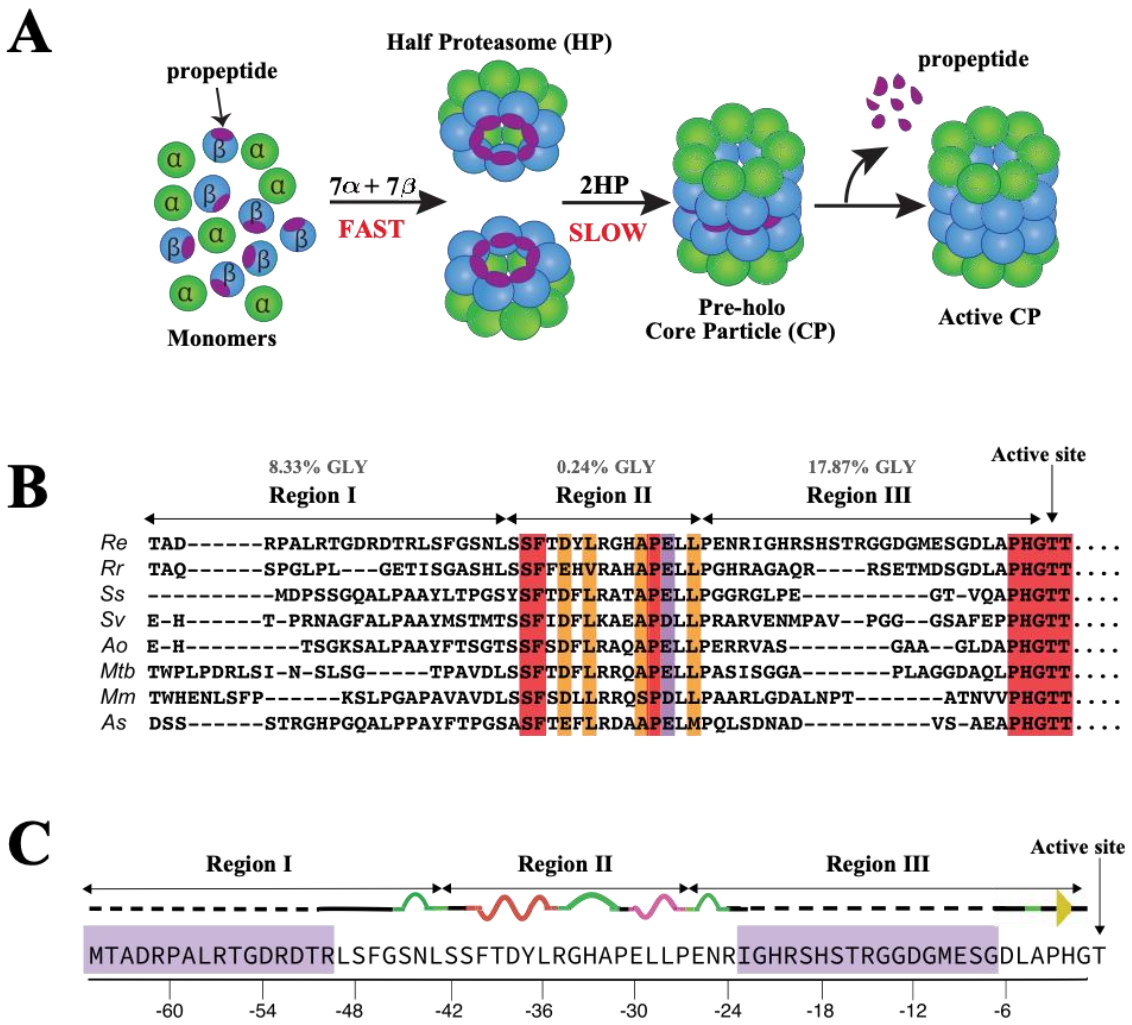


Figure 2.1: Bacterial 20S proteasome assembly and propeptide conservation. (A) Schematic of the proteasome Core Particle (CP) assembly. The α subunits are shown in green and β subunits in blue with the propeptide in purple. Arrows demonstrate progression from subunits to active CP and separation of time scales between Half Proteasome (HP) and CP assembly. (B) Note: This alignment is a representative subset of the 256 species MSA used for analysis. Eight amino acid sequences of the N-terminal β subunit propeptide from *Rhodococcus erythropolis (Re), Rhodococcus rhodnii (Rr), Saccharopolyspora shandongensis (Ss), Saccharomonospora viridis (Sv), Amycolatopsis orientalis (Ao), Mycobacterium tuberculosis (Mtb), Mycobacterium mageritense (Mm) and Actinopolyspora saharensis (As)*. The conserved residues are highlighted in red, residues in orange have conservation between amino acid groups of similar properties and the residues in purple have conservation between amino acids of weakly similar properties. The active site is shown the vertical arrow above the Threonine (T). (C) β propeptide sequence in *Re.* Region I is made up of residues -65th to -43rd, Region II is from the -42nd to the -27th residues, and Region III is from the -26th to -1st residues. The residues without electron density are highlighted in purple and are modeled for simulations.

limiting step and demonstrates a significant separation in time scales between HP formation and dimerization to form the CP (**Fig. 2.1A**).

The reason for this separation of these timescales is currently unknown, although existing experimental evidence supports a key role for the propeptide in regulating dimerization [19]. Baumeister and colleagues monitored *Re* CP assembly in three different scenarios. First, wild type (WT) α and β subunits form HPs within seconds, but there is a time lag of approximately 30 minutes in CP assembly. Second, WT α subunits and a mutant variant of the β-subunits with no propeptide (βΔpro) assembled the HP at a significantly slower rate. Subsequent crystallographic studies on the *Re* CP with a β mutant where the propeptide cannot be cleaved demonstrated that the propeptide mediates critical interactions between the α and β subunits, a possible explanation as to why assembly is attenuated in βΔpro mutants [1, 19]. Finally, they added the propeptide in trans with α and βΔpro subunits. Surprisingly, active CP was formed significantly faster, within ~30 seconds. The HPs were not observed, suggesting that they dimerized so quickly that they could not be captured in native gels [19]. Collectively, these findings provide evidence that propeptide regulates the dimerization rate by inhibiting HP dimerization. The molecular mechanisms through which the propeptide achieves this regulation, however, are not currently understood.

We recently performed a combined computational and experimental study that began to elucidate the propeptide's role in the dimerization step (19). As part of this work, we performed a Multiple Sequence Alignment (MSA) of the *Re* β subunit, including the propeptide, with sequences from various bacterial species (**Fig. 2.1B**). We divided the bacterial propeptide into three distinct regions based on this alignment and the available crystal structure (**Fig. 2.1B**) (1, 3, 5, 19). Region I is the most N-terminal segment of the propeptide (residues -65 to -43 in the *Re* sequence), is more flexible than Region II and interacts with α subunits. Region II (residues -42 to -27 in *Re*) is much more conserved than Regions I and III and forms the crucial central region (1); these residues form contacts with both the α and β subunits. Region III (residues -26 to -4) is highly flexible and is immediately N-terminal to the active site threonine of the β subunits. This region is also highly enriched in glycine residues (~17%) across all bacterial species (19) (**Fig. 2.1B**). Region III is notably near the HP dimerization interface, suggesting that it could play a crucial

role in CP assembly. Determining the role of Region III of the propeptide in HP dimerization requires detailed analysis. Molecular Dynamics (MD) simulations provide details that current biochemical approaches cannot. Further, MD simulations provide a detailed platform to gather atomistic insights into the function of the propeptide during CP assembly [22, 23]. Our preliminary MD results from previous work show that the *Re* propeptide has a high root mean square fluctuation (RMSF) value for Region III, a metric that measures a residue's flexibility averaged over time. We observed that this flexibility enables the propeptide to move near the HP dimerization interface.

Furthermore, at the HP dimerization interface, there are a set of key residues in *Re* β subunits which are involved in critical interactions (hydrogen bonds, salt bridges, or other noncovalent interactions) with the β' ring of the other HP (**Fig. 2.2A and 2.2B**), these interactions drive CP assembly [5]. Since these key

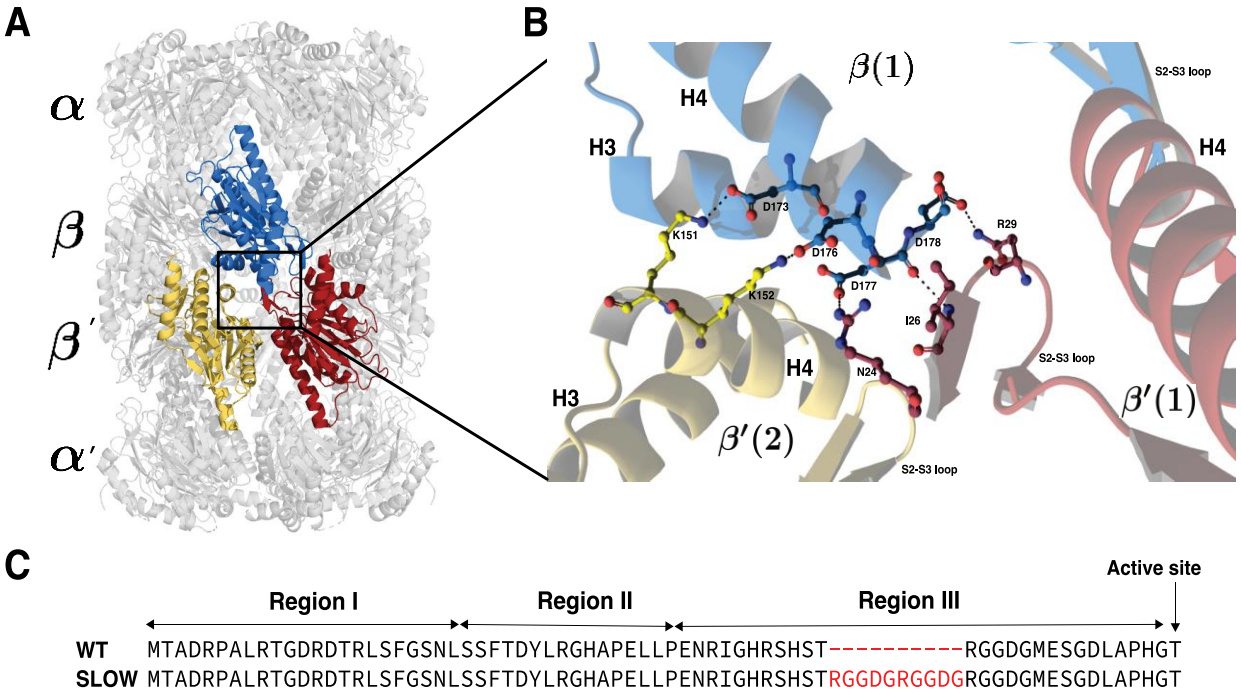


Figure 2.2 *Re* CP structure and key residues (A) Side view of the WT *Rhodococcus erythropolis* (PDB entry:1Q5R) Core Particle (CP). The colored β subunits (blue, yellow and red) highlight interactions between β and β' subunits which occur at the Half Proteasome (HP) dimerization interface. (B) This inset shows the zoom in view of the key residues associated with β-β' interactions that are part of S2-S3 loop and H3-H4 helices (REF) (C) *Re* β propeptide sequence for Wild Type (WT) and mutant version SLOW, which forms CP at slower rate.

residues occur at the HP dimerization interface, we hypothesized that the propeptide interacts with the key

residues, physically preventing them from associating with another HP and thus slowing dimerization. To further investigate flexibility and glycine enrichment shown in the MSA, we designed mutants targeting Region III of *Re* propeptide (**Fig. 2.1C**). *In vitro* experiments revealed that propeptide mutations lead to slower HP dimerization relative to WT and little to no active CP [24]. These experiments provided further evidence that *Re* propeptide regulates HP dimerization rate (9).

While our earlier experiments and simulations on *Re* CP provided significant information regarding CP assembly, to date, there have been limited attempts to determine the role of the propeptide in HP dimerization quantitatively [24]. In this work, we began investigated the role of the propeptide in *Re* CP assembly by performing extensive all-atom MD simulations of the *Re* WT HP and an extremely slowly dimerizing mutant from our previously published work which we call as the "SLOW" mutant [24]. Here, we used the Anton supercomputer to complete these simulations, allowing us to achieve microsecond sampling for very large HP structures [25]. Our simulations show specific hydrogen bonding interactions between the propeptide and the key residues at the dimerization interface. These interactions are considerably more frequent in the SLOW mutant than in WT. Analysis of these simulations suggested that mutating charged residues in Region III would reduce those hydrogen bonds and make dimerization faster. We thus computationally designed a FAST mutant by mutating two charged residues in WT β subunits to alanine. As anticipated, the FAST mutant HP simulations revealed that the propeptide makes fewer interactions with key residues than WT. We then tested these predictions experimentally using *in vitro* assembly assays and found that the FAST mutant does dimerize considerably faster than the WT. This study thus proposes a model where the *Re* HP exists in two conformational states: D+ (dimerizable) or D- (non-dimerizable). If the WT HP is in a D+ state, its propeptide residues do not interact with the dimerization interface key residues. Hence, they can bind with the β' ring of another HP and assemble into a CP. In other words, D- state refers to the conformation of HPs where the propeptide of any β subunit interacts with key residues at the dimerization interface. The propeptide interactions thus prevents the β subunit from associating with another HP. In solution, both D+ and D- states exist in equilibrium, and mutations or other perturbations can influence this equilibrium. Our results show that, in WT, the D+ and D- are both present,

but in SLOW mutant, D- conformation dominates. As anticipated, the FAST mutant more frequently resides in the D+ conformation. This model explains our previous experimental results and provides unprecedented insights into how HP dimerization is regulated.

This study thus presents, a deeper understanding of the propeptide's functional role in *Re* CP assembly. Our findings demonstrate that the propeptide regulates the HP dimerization step and is tightly coupled with the propeptide's length and amino acid composition. In the future, this work will provide support for ongoing efforts to use structure-based approaches to design inhibitors of CP assembly.

2.2 Materials and Methods

2.2.1 Half Proteasome structure for Molecular Dynamics simulations

The starting structure for the *Rhodococcus erythropolis* Core Particle (CP) was taken from PDB 1Q5R [1]. The Half Proteasome (HP) starting structure was obtained by taking the top-half of CP coordinates, with all 14 chains is used as a starting structure for MD simulations.

2.2.2 Modeling missing propeptide residues

For the SLOW mutant in this study, the missing regions of the propeptide were modeled using Rosetta Comparative Modeling [26, 27] hosted at the Robetta server. For this, we first modeled a single α-β dimer and then repeated the same model for seven-fold symmetry to obtain the HP structure.

2.2.3 Molecular Dynamics simulations setup

The simulation inputs were generated using the CHARMM-GUI solution builder module [28-30]. All the systems were neutralized with 100mM NaCl (same used for experiments) and 15A° water on each side of the protein for a rectangular box. The detailed components are given in the supplementary material. We performed a short minimization of 5000 cycles, switching from steepest descent to conjugate gradient after

2500 cycles, the system is held at constant volume, no restraints are held on atoms, and the nonbonded cutoff is 10 Å. Next it is followed by 5ns of NVT and 60ns NPT equilibration with a 2fs time step using the GPU version of AMBER18 [31, 32] using CHARMM forcefield. Each simulation took ~ 48 hours to finish 65ns of equilibration on NVIDIA RTX 2080Ti GPU. The detailed parameters for the equilibration are given in the supplementary material. After finishing AMBER equilibrations, the coordinates and restart files were used to initiate 2.5 μs simulations on Anton2 for long time scale Molecular Dynamics simulations [25]. The NPT ensemble and CHARMM 36m force fields, and TIP3P water models were used for Anton simulation runs. Additionally, the pressure and temperature are constant at 1bar and 303.15 K using Multigrator integrator and default NPT parameters [33]. The RESPA (Reference systems propagator algorithms) integration method was employed with a timestep of 2.5 fs, and the coordinates were saved every 0.24ns. All systems were simulated under periodic boundaries in the NPT ensemble. Additional details of the equilibration and MD simulations are provided in A.10.

All the simulations for WT and mutants have a backbone RMSD around 3.5 Å – 6.5 Å (Fig. S4) and they require about 500 nanoseconds to converge (Fig. S1). So as a result, we did not include the first 500 nanoseconds of the runs in our analysis. All the details and methods for *in vitro* reconstitution experiments for the WT and FAST variants are in appendix A.7.

2.2.4 Statistical analysis

To estimate the differences between the WT and mutant simulations, we used a categorical regression in R (Tables A.9). This test is used to determine if the average behavior of WT is different from SLOW and FAST or if they were drawn from different distributions. MD simulations data is time-dependent and autocorrelated, and as such it is challenging to find a statistical test to estimate the significance of MD simulations. Therefore, to account for these factors, we have used the Newey West estimators of standard errors and Ordinary Least Squares Regression with categorical variables. The use of categorical variables as separate binary variables is done for WT replicate and each mutant replicate. All the linear regression,

estimates and tests are done using the lm function in R [34] and scripts are available in the Supplementary material.

2.3 Results

2.3.1 Key residues regulate CP assembly

Shorter simulations from our previous work revealed that in the WT *Re* propeptide Region III, which is near the HP dimerization interface, is highly flexible and is generally found outside the HP barrel [24]. In this work, we performed 2.5 μs Molecular Dynamics (MD) simulations of the Half-Proteasome (HP). Closer inspection of Region III revealed a possible steric challenge for another HP to bind (**Fig. A.1**). Indeed, MD simulations show the propeptide physically blocking the regions on which a second HP could theoretically bind. In fact, a stretch of about 7-10 residues of Region III is more mobile than the Regions I or II (**Fig. A2**). Further, we found that this stretch forms hydrogen bonds with a group of residues at the dimerization interface of the HP. Interestingly, this group of residues have been previously identified as key residues for CP assembly [5]. Experimental evidence from Witt et. al showed that these key residues form critical interactions with the key residues of the opposing HP. In *Re* Half Proteasome (HP) assembly, a set of key residues on the β subunits interact with the opposing β' of another HP to drive CP formation (**Fig. 2.2A and 2.2B**). These interactions include hydrogen bonds, salt bridges, and hydrophobic interactions between the β rings [5]. Perturbing any of these critical interactions by mutating the key residues has a destabilizing effect and, in most cases, completely prevents CP formation (5). Experimental work demonstrated that alterations to the propeptide length in *Re* increased CP assembly time or prevented dimerization altogether [24]. However, from *in vitro* assembly assays alone, it remains unclear how these propeptide alterations slowed CP assembly. Thus, since the propeptide residues are within 2-3Å from the key residues, it is possible that the propeptide interacts with the key residues and regulates dimerization rates.

2.3.2 Mutant with extended propeptide Region III dimerizes very slowly

Propeptide Region III is Glycine rich and contains a $^{-13}$GDGMESG^{-7} motif that frequently interacts with the key residues (data not shown) [24]. To determine this part of the propeptide influence CP assembly and kinetics, we previously designed β mutants with a longer Region III loop. *In vitro* assembly assays showed HP formation in altered Region III mutants but no CP formation even after prolonged incubation times [24]. All the mutants considered in this previous study had a portion of Region III deleted or added near the HP dimerizing interface. One mutant, the Extended Loop SLOW (EL1 in [24]), has an additional sequence of charged residues (**Fig. 2.2C**). We introduce a repeated glycine-rich motif twice to the WT propeptide, hypothesizing that this addition can delay HP dimerization. To elucidate the mechanisms behind the slow dimerization, we simulated the WT and SLOW HP using MD simulations. We speculated that the additional charged residues result in more interactions between the key residues and the propeptide, slowing down HP dimerization. Additionally, the SLOW mutant's extra residues would cause steric hindrance for the opposing HP to collide and form CP (**Fig. A1**).

We used the Anton2 supercomputer [25] to simulate both WT and SLOW HPs for 2.5 μs. If there are hydrogen bonds between propeptide and key residues in any of the β subunits, we hypothesized the HP would be in an interacting state that cannot dimerize with another HP, thus termed non-dimerizable (D-). If there are no hydrogen bonds formed between the propeptide and the key residues, then the HP is non-interacting, free to dimerize with another HP, thus dimerizable (D+). **Fig. 2.3A** illustrates the non-dimerizable (interacting) and dimerizable (non-interacting) states. We quantified interactions by calculating the number of simulation frames in which a hydrogen bond (2.4 Å distance) between a key residue and a propeptide residue is formed. For each timestep (0.24ns), we determine the number of hydrogen bonds formed and which state the HP is dimerizable or non-dimerizable. Consistent with our hypothesis, simulation results revealed that the SLOW mutant propeptide makes many more interactions with the key residues (**Fig 2.3B)**. The ratio of simulation frames in each state shows that WT HP exists in either a D+ (25%) or D- (75%) state in the 2.5 us sampled (**Fig. 2.3B**). The HP of the SLOW mutant nearly

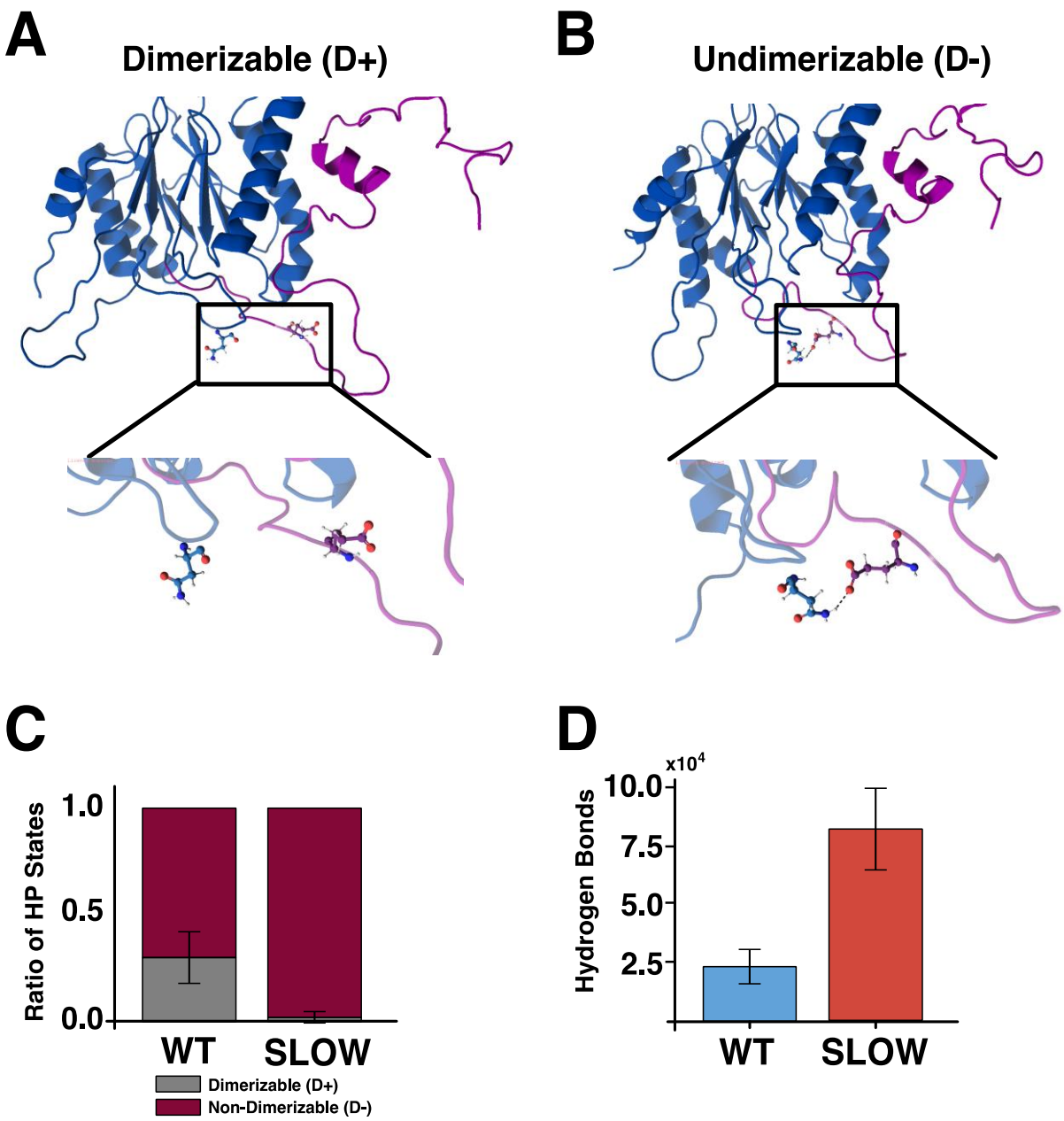


Figure 2.3: Conformational states definition and WT, and SLOW hydrogen bonds profile. (A) Ribbon diagrams of the β subunit (blue) with a full length propeptide (purple). Key residues are shown as colored atoms. (Left) β subunit of a HP in a non-dimerizable state with inset highlighting interactions. (Right) β subunit of HP in a dimerizable state for comparison. (B) Bar graph of percent of simulated time frame in D+ and D- states for both WT and SLOW β subunit mutants. Error bars signify SEM for 3 MD simulations of 2.5 μs. (C) Bar graph showing the total number of hydrogen bonds formed between the propeptide and the key residues as an average over 3 MD simulations for both WT and SLOW Half-Proteasomes (HPs). Error bars show SEM for 3 MD simulations of 2.5 μs.

always (99% of simulated time) resides in a D- state (**Fig. 2.3B**). Additionally, our simulations show many

hydrogen bonds between the propeptide, and key residues of the SLOW mutant and fewer hydrogen bonds formed in the WT HP (**Fig. 2.3C).**

Our simulation results suggest that adding the extra charged residues to Region III of propeptide makes it further extensible and interacts with key residues by making more noncovalent interactions (**Fig. 2.3**). The hydrogen bonds could be perturbing the native interactions required for CP formation. Hence, the HP with a longer propeptide Region III is not in the required confirmation to dock with the opposing HP. This may be why the SLOW mutant has no CP assembly observed even after 24 hours of incubation time (9).

2.3.3 Charged residues in Region III yields a HP in a mostly dimerizable state

As observed in the SLOW mutant, the addition of charged residues delays CP assembly; we further investigated the role of charged amino acids and hypothesized that mutating charged residues to non-polar amino acids would potentially reduce these critical interactions. To determine the role of Region III charged residues on assembly, we selected two charged residues in WT Region III, E-9, and D-12, and computationally mutated them to Alanine using CHARMM GUI [28] to generate what we term as FAST mutant (**Fig. 2.4A**). In the WT sequence, both residues make many interactions, specifically by forming side-chain hydrogen bonds with several amino acids that form salt bridges across the HP dimerization interface, particularly R29 and N24 residues. We predicted that mutating the Glutamic Acid and Aspartic Acid to Alanine would potentially contribute to faster dimerization by making fewer non-covalent interactions. We ran similar MD simulations for 2.5 μs with this FAST HP mutant to test our hypothesis and quantified the ratio of simulated time the HP spent in the D+ vs. D- state. Analysis of the FAST mutant simulations revealed that interactions between the propeptide and the key dimerization residues were much less frequent than WT, i.e., about 11% instead of 25% in WT (**Fig. 2.4B**). Further, the FAST HP formed fewer hydrogen bonds compared to the WT HP (**Fig. 2.4C**). As expected, interactions between the -9 and -12 positions of the propeptide and the key residues were much less frequent (**Fig 2.4B and 2.4C**). This suggests that the FAST mutant could dimerize faster than WT.

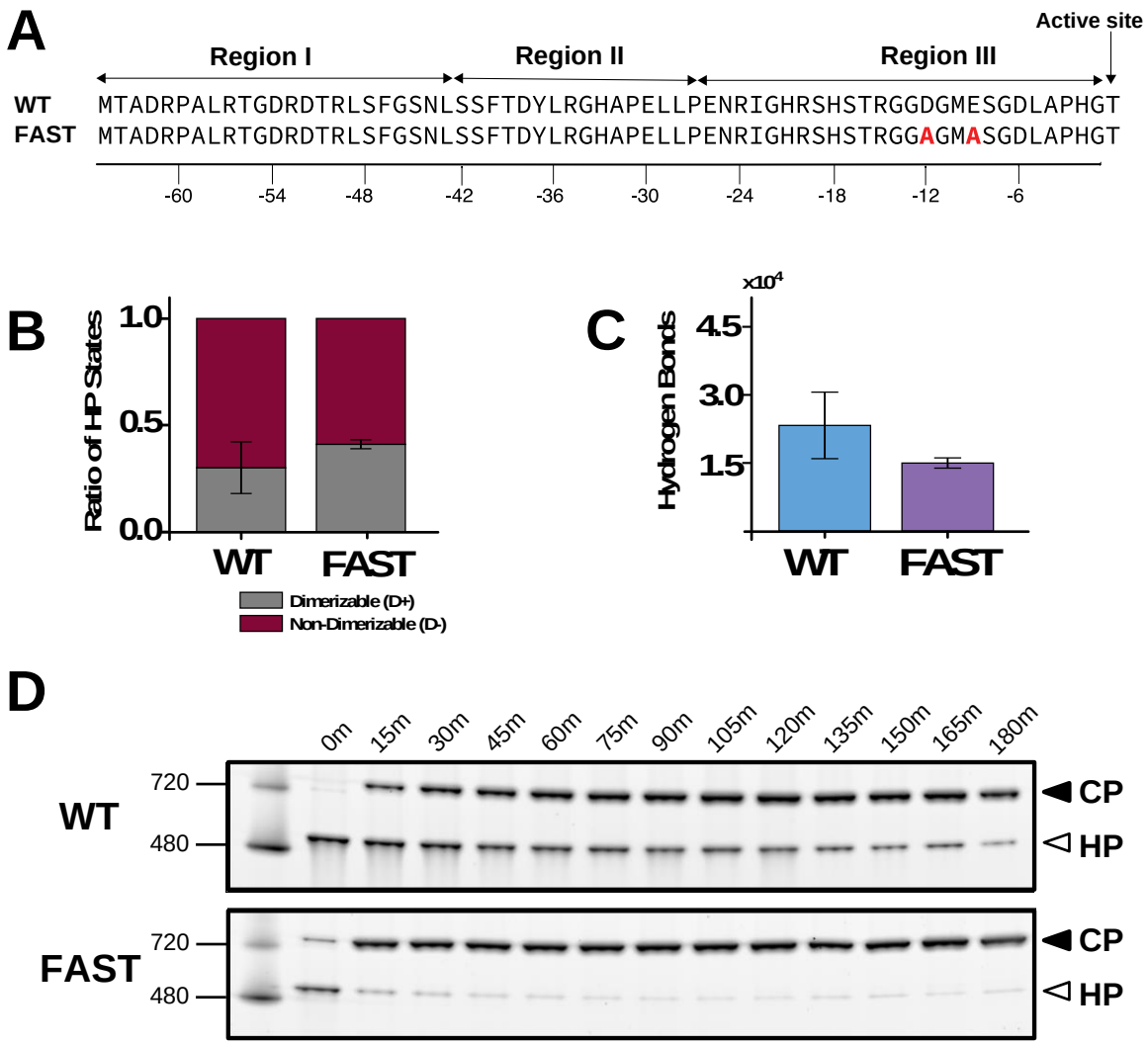


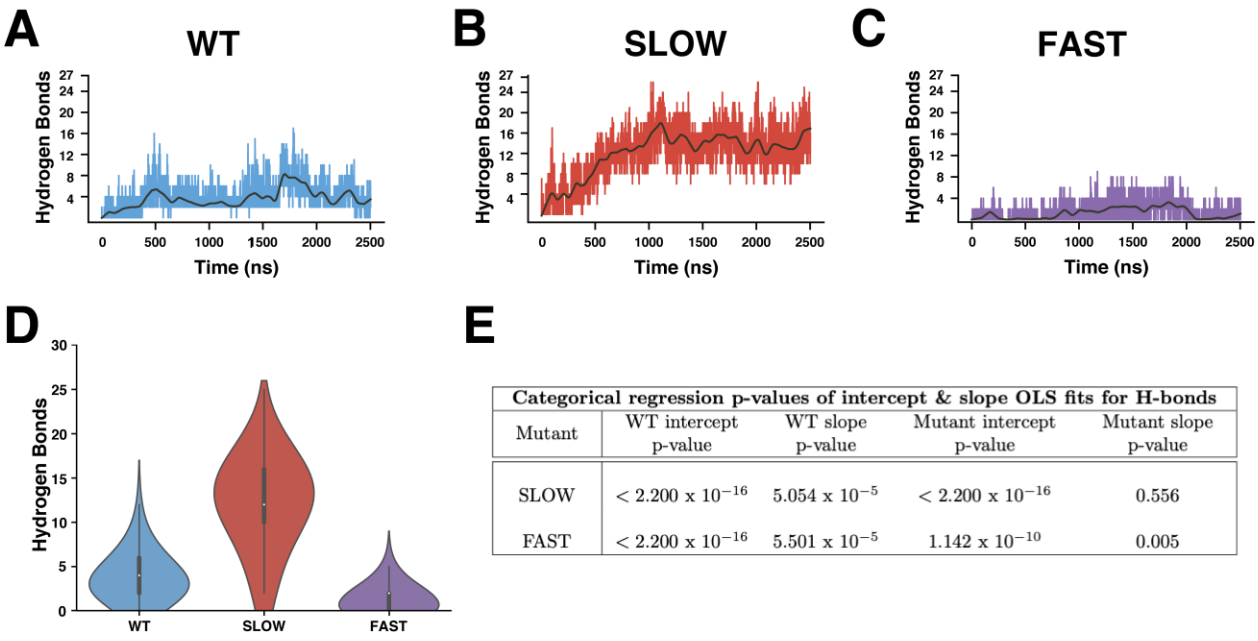
Figure 2.4: WT and FAST hydrogen bonds profile. (A) Sequence of β subunit propeptide regions in WT (top) and FAST (bottom) mutants. Red bolded residues are the altered D, E to A. (B) Bar graph of percent of simulated time frame in D- and D+ states for both WT and FAST β subunit mutants Half-Proteasomes (HPs). Error bars signify SEM for 3 MD simulations of 2.5 us. (C) Bar graph showing the total number of hydrogen bonds formed between the propeptide and the key residues as an average over 3 MD simulations for both WT and FAST HPs. Error bars show SEM for 3 MD simulations of 2.5 us. (D) 4-20% Tris-Glycine native gels from in vitro assembly assays at increasing time points (time points labeled above each lane in minutes) for WT (top) and FAST (bottom) β subunit mutants. Gels were stained with Spyro Ruby protein and visualized with a BioRad Imager.

Based on our computational predictions, we generated a double point mutation D-12A&E-9A variant (**Fig. 2.4A**), using ligation independent cloning [24]. To compare the assembly rates of WT and FAST mutants, α and β mutant substrates were mixed *in vitro* and incubated at 30^0C at increasing time points from 0 seconds to 180 minutes. This assay provides a readout of the ability of α and β subunits to assemble core particles. At each time point, we can visualize the CP assembled and the HP fractions remaining to be converted to CP (**Fig. 2.4D**). We performed those experiments to obtain time course for WT and FAST proteasome CP assembly (**Fig. 2.4D**). Native gels show that, indeed, the FAST mutant forms CP faster than WT. In fact, after 30 mins, the FAST mutant has formed CP, and there is little to no detection of HP, suggesting all the HP's have dimerized by this point. In contrast, at 30 minutes the WT subunits have formed some CP and HPs (**Fig. 2.4D**). Further, the SLOW has formed no CP at all by 30 minutes incubation [24]. The experimental results thus are consistent with the MD simulation predictions.

2.3.4 Hydrogen bond dynamics shows state transitions

To look closer at the dynamics of hydrogen bonds formed over the simulation time for every β subunit, we characterized the hydrogen bonds formed between the subunits and the key residues as a function of time. The SLOW mutant has more hydrogen bonds forming between the key residues and all seven propeptides than WT (**Fig. 2.5A, B**); FAST had fewer hydrogen bonds than the WT (**Fig. 2.5 A, C**). Additionally, WT has a lower ratio of non-dimerizable states than the FAST mutant over the simulated 2.5 us. Thus, the dynamics of hydrogen bonds forming could provide insights into the mechanisms regulating the dimerizable and non-dimerizable states. Here, we used the Locally Weighted Scatterplot Smoothing (LOWESS) method [35] to capture the hydrogen bonds taking place between propeptides and key residues over time. LOWESS plots indicate that the HP needs about 500ns to reach a converged state after equilibration (**Fig. A.3**); this is primarily due to not having the solved HP crystal structure, the large complex size, and limited sampling using all-atom unconstrained MD simulations.

LOWESS plots for our MD simulations show that during the MD simulation, WT and SLOW form more hydrogen bonds than the FAST mutant (**Fig. 2.5A**). The LOWESS plot for the FAST mutant shows frequent transitions from zero hydrogen bonds to multiple (**Fig. 2.5A**). This suggests that the FAST HP transitions from a dimerizable to a non-dimerizable state frequently in the span of a 2.5us simulation. These rapid transitions never occur in the SLOW HP, (**Fig. 2.5A**). After 500ns, the SLOW HP usually forms more than 20 hydrogen bonds and thus in a non-dimerizable conformational state (**Fig. 2.5A**). Further, the WT HP transitions between dimerizable and non-dimerizable states based on the hydrogen bond count, but less frequently than the FAST HP (**Fig. 2.5A)**



Categorical regression p-values of intercept & slope OLS fits for H-bonds

Mutant	WT intercept p-value	WT slope p-value	Mutant intercept p-value	Mutant slope p-value
SLOW	$< 2.200 \times 10^{-16}$	5.054×10^{-5}	$< 2.200 \times 10^{-16}$	0.556
FAST	$< 2.200 \times 10^{-16}$	5.501×10^{-5}	1.142×10^{-10}	0.005

Figure 2.5: Hydrogen bond dynamics in MD simulations, shown for one replicate of each type. (A)(B)(C) LOWESS plots of the number of hydrogen bonds formed by the HP over simulated time (in ns) for a single MD simulation of WT (left), SLOW (middle) and FAST (right) β subunit mutants for the first replicate. Bold line represents the non-parametric LOWESS fit. LOWESS plots for remaining replicates are included in Supplementary Material. (D) Violin plot of distribution of hydrogen bonds formed over each of three independent MD simulations of WT (blue), SLOW (red) and FAST (purple) HPs. White dot in the violin plots represents average number of hydrogen bonds for that computational replicate. E) Categorical regression *p-values* to compare the WT and the SLOW, FAST mutants shown in D).

We used violin plots to compare the overall distribution of hydrogen bonds across the simulations. Violin plots use kernel density estimates (KDE) [36] to depict distribution of hydrogen bonds formed for WT, FAST, and SLOW (**Fig 2.5B**). MD simulations are time-dependent, and to incorporate that in our statistical

analysis for comparing simulations, we used heteroscedasticity and autocorrelation consistent Newey-west estimators to assess the statistical variability and determine if the overall behavior of WT differs from SLOW and FAST. For this, we did a detailed analysis by categorial regression of the number of hydrogen bonds formed in WT vs mutants as a function of time (**Tables in A.9**). We observed that all the WT differ significantly from SLOW and FAST and display a significant *p-value* with for intercepts or slopes (**Fig. 2. 5D**). We have only shown these tests for one replicate in Fig. 2.5D the complete details are described in **Tables in A.9**.

The multimeric structure of proteasome relies on the thermodynamics principle of cooperativity to assemble subunits into a CP [37]. We looked at kymographs for each simulation to investigate if the WT, FAST, and SLOW β subunit mutants show cooperativity. Kymographs are graphical representations of hydrogen bonds made by the propeptide of each β subunit over time. These graphs allow us to determine if β subunits displayed cooperativity among themselves. In other words, if one β subunit propeptide interacts with the key residues, does that influence the other β subunits to interact as well? In **Fig. 2.6**, we show the number of hydrogen bonds formed in each β subunit over time for WT (**Fig. 2.6A**), FAST (**Fig. 2.6B**), and SLOW (**Fig. 2.6C**) HPs. We observed no cooperativity between the subunits in any of the mutants.

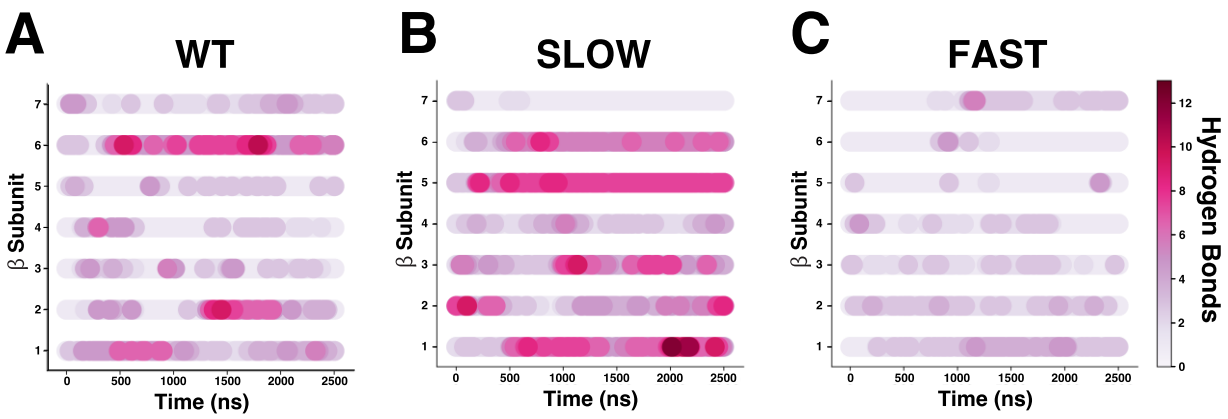


Figure 2.6: Hydrogen bond dynamics in each β subunit. Kymograph of the number of hydrogen bonds formed over simulated time for each β subunit in the HP of WT (A), SLOW (B), and FAST (C) mutants. Colors correspond to the number of hydrogen bonds formed based on colormap with the brighter red as a higher count.

Interestingly, we observed a heterogeneous distribution of hydrogen bonds among the seven β subunits. As

expected, the SLOW mutant had a higher number of hydrogen bonds across the subunits than WT and FAST. We also observed heterogeneity among the subunits, i.e., once a β subunit reaches a non-dimerizable state, it likely stays there for longer timescales (much longer than 2.5 μs). For WT we observed that the individual subunits flip between D+ and D- states. In the SLOW mutant we observed that if a subunit is in D- state (like β subunit-7) it stays for a longer time (> 2.5 μs) in the non-dimerizable state. For the FAST mutant, we saw even more transitions from D+ to D- than the WT or SLOW.

2.4 Discussion

The assembly of the active proteasome Core Particle (CP) has been previously studied in several species [18, 19, 38]. Since the proteasome is a crucial complex for regulating the cell cycle, immune response and maintaining homeostasis, the order of events taking place during CP assembly requires temporal control. Our understanding of the detailed structural aspects of the molecular mechanisms regulating these hierarchal assembly pathways is largely unknown. To our knowledge, this work is one of the first studies on the complex dynamics of CP assembly at an atomic level.

A critical step in CP assembly, which is conserved across species, is the dimerization of two Half Proteasomes (HPs). In the bacterial species *Rhodococcus erythropolis* (*Re*), experimental assembly assays show that the HP is quickly assembled while the CP assembly takes considerably longer. Our previous work identified the propeptide on the β subunits as crucial for dimerization (22). The evolutionary motivation for this separation of time scales remains unknown. Here we describe a novel mechanism to explain this separation of time scales. We observe a structural transition that we hypothesize physically blocks the dimerization of two HPs, which may explain how the propeptide regulates dimerization rates.

Here, we performed all-atom Molecular Dynamics (MD) simulations with *Re* HP, focusing on WT and β propeptide mutants. A significant result from our MD simulations suggests that the length and residue composition of the propeptide regulate the conformations for an HP to dimerize based on the interactions the propeptide makes with key residues – those associated with binding the opposing β ring. Our work

indicates that the HP exists in one of the two states that drive CP assembly, a dimerizable state or D+ that leads to two HPs association and a non-dimerizable or D- state that does not. Experimental evidence shows that with WT subunits, the HP forms quickly, while dimerization of HPs to form the CP takes relatively longer [15, 19]. With β subunit mutants with a longer propeptide, i.e., SLOW mutant, experiments show that the CP is assembled even slower, and no CP formation is seen after 24 hours [24]. Our MD simulations show that the SLOW HP propeptides form more hydrogen bonds with key residues at the dimerization interface (β ring). This binding implies that these interactions prevent the HP from associating to the opposing HP, therefore maintaining the HP in a D- state. The WT HP spends less time in this state and is thus more likely to dimerize and form a CP. Further, MD simulations show that the polarity of the propeptide plays a role in the ability of the HP to dimerize. Our simulations with altered propeptide residues, in the FAST mutant, show an HP that forms fewer hydrogen bonds between the propeptide and the key residues than a WT HP. This suggests that the FAST β mutant HP is more likely to be in a dimerizable state than the WT. Our *in vitro* assembly assays show this mutant indeed forms CPs much faster. From our MD simulations, we believe that since the FAST HP is forming fewer hydrogen bonds than the SLOW mutant, and the FAST mutant is more likely to be in a dimerizable state and able to bind with another HP.

When the propeptide from one β subunit forms hydrogen bonds with the key residues of the same or neighboring β subunit (non-dimerizable subunit), the remaining β subunits propeptides do not necessarily form such hydrogen bond. However, our understanding is that even if a single β subunit is in the non-dimerizable state, then likely, the HP cannot associate with another HP. It would be interesting to examine the β subunits involve in cooperative transitions to make other β subunits non-dimerizable. From the kymographs (time evolution of hydrogen bonds for β subunit), we observed that subunits behave independently in all simulations (Fig. 2.6 and Fig. A.8). Further, in our kymographs, we observe heterogeneity among these subunits, wherein the WT and the FAST the β subunits flip a lot between D- and D+. However, in the SLOW mutant kymographs, we do not observe such transitions between the states; instead, we see that the SLOW β subunits remain in the same state for more than 2.5 μs. For instance, in Fig. 2.6 B the β_1, β_5, and β_6 are in D- states for most of the time. Furthermore, we did not observe any

cooperative transitions among the β subunits. In other words, if one β subunit flips to D- state it does not influence its neighboring β subunits to transit into D- states. Thus, every β subunit behaves independently and contributes to the overall HP conformation and hence dimerization state. Probably, a different metric, such as a Mutual Information measurement (26, 27), could quantify cooperativity; however, our kymographs did not suggest any such correlation.

The proteasome is conserved across all kingdoms of life, but the subunit sequences are different. The β propeptide is present on all β subunits with very less sequence conservation [39]. Our previous work shows the β propeptide in bacteria can be categorized into three regions and with different amounts of glycine's in each region, and Region II being the most conserved among all bacterial species. However, the assembly pathway of different bacterial proteasomes remains unexplored, with the most common model system is that of *Re*. Future studies are critical to uncovering if the propeptide similarly regulates dimerization in other bacterial proteasomes. For example, in *Mycobacterium tuberculosis* (*MT*), the proteasome is implicated in the bacterium's ability to resist macrophages [40-42]. Upon closer inspection, *mob* proteasome shares 64% sequence similarity with *Re*, suggesting that the flexibility in Region III might be a similar feature between the two bacterial species [24]. We hypothesize that this flexibility in Region III may also play a critical role in dimerization rates in other bacterial proteasome assemblies, like what we have observed here with *R. erythropolis*.

Our results also indicate a need for further structural studies to identify and capture the *Re* HP transitioning from a non-dimerizable to dimerizable state. Electron Microscopy (EM) studies have captured HP and CP with WT subunits [16, 43]; however, the structures formed with β mutant variants remain unknown. Further, the CP with β mutant variants could be inactive, thus obtaining high-resolution structures of these HPs and CP's will be immensely valuable for understanding *Re* CP function and assembly. Notably, there may be a structural mechanism regulating dimerization states in the mutant HP.

In some eukaryotic proteasomes, β subunits have longer propeptides than the bacterial species [13, 39]. Eukaryotic proteasomes also require chaperone's assistance during assembly [38, 44, 45] to control the addition of subunits and conformational states during HP formation. It is, therefore, possible that the *Re*

bacterial propeptide acts like a chaperone to regulate dimerization rates by controlling the conformational states. However, there is a significant need to verify this experimentally. To our knowledge, there are no experimental studies that measure the assembly rates of proteasomes other than in *Re*. However, if subunits can evolve to have a faster assembly, why has the proteasome evolved to display a separation of time scales regulated by the propeptide? We speculate that the slow HP dimerization step serves as a checkpoint in cells to ensure the correct stoichiometry of the intermediates before dimerizing to form a CP and thus prevent formation of aberrant CP-like (less than 28 subunits) structures. Further, there is a possibility that mutations to the propeptide, particularly those altering length and polarity, render the CP inactive and thus are not beneficial. The *in vitro* assembly assays discussed in this study cannot demonstrate if the propeptide is autocatalytically cleaved off and CP renders active after assembly. Mutants can often have slower activity rates than WT [5, 24]. To further validate assembled CP, activity assays and FAST mutant are essential to confirm if it has altered CP activity.

Developing a complete understanding of how propeptide affects CP assembly will require comprehensive computational and experimental efforts in other organisms [13, 39, 46, 47]. However, several open questions remain unanswered; for example, would mutating the other charged propeptide residues in Region III of the FAST mutant (ARG -15 in **Fig. 2.4A**) make CP assembly even faster? Has the pathogen *mob* like *Re* have a similar mechanism for the separation of time scales in its CP? Moreover, why has this slow dimerization step evolved if there was a possibility of making assembly faster? There are still numerous long-standing questions to be answered in understanding proteasome assembly. These questions will require additional atomistic simulations and experiments for seeking out answers.

Our result on molecular simulations emphasizes the importance of the HP's ability to dimerize as the propeptide length and composition regulate it. This work will also lay the foundations for structure-based drug design by utilizing the HP structures from simulations as a template, where small-molecular assembly inhibitors can lock the HPs in a non-dimerizable state. The development of effective and potent CP assembly inhibitors will also lead to a framework where this approach of using MD simulation structures as a template, to target diseases like tuberculosis [48, 49], cancer [9, 50-52], and other diseases [53, 54].

Ultimately, our work highlights the critical need of studies combining biophysical models and experiments to elucidate hierarchical assemblies.

2.5 References

1. Kwon YD, Nagy I, Adams PD, Baumeister W, Jap BK: Crystal structures of the Rhodococcus proteasome with and without its pro-peptides: implications for the role of the pro-peptide in proteasome assembly. *J Mol Biol* 2004, 335(1):233-245.
2. Sharon M, Witt S, Glasmacher E, Baumeister W, Robinson CV: Mass spectrometry reveals the missing links in the assembly pathway of the bacterial 20 S proteasome. *J Biol Chem* 2007, 282(25):18448-18457.
3. Zwickl P, Voges D, Baumeister W: The proteasome: a macromolecular assembly designed for controlled proteolysis. *Philos Trans R Soc Lond B Biol Sci* 1999, 354(1389):1501-1511.
4. Campbell MG, Veesler D, Cheng A, Potter CS, Carragher B: 2.8 A resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy. *Elife* 2015, 4.
5. Witt S, Kwon YD, Sharon M, Felderer K, Beuttler M, Robinson CV, Baumeister W, Jap BK: Proteasome assembly triggers a switch required for active-site maturation. *Structure* 2006, 14(7):1179-1188.
6. Lupas A, Zühl F, Tamura T, Wolf S, Nagy I, De Mot R, Baumeister W: Eubacterial proteasomes. *Mol Biol Rep* 1997, 24(1-2):125-131.
7. Lin G, Li D, Chidawanyika T, Nathan C, Li H: Fellutamide B is a potent inhibitor of the Mycobacterium tuberculosis proteasome. *Arch Biochem Biophys* 2010, 501(2):214-220.
8. Cheng Y, Pieters J: Novel proteasome inhibitors as potential drugs to combat tuberculosis. *J Mol Cell Biol* 2010, 2(4):173-175.
9. Almond JB, Cohen GM: The proteasome: a novel target for cancer chemotherapy. *Leukemia* 2002, 16(4):433-443.
10. R. C. Kane PFB, A. T. Farrell, and R. Pazdur: Velcade: U.S. FDA approval for the treatment of multiple myeloma progressing on prior therapy. *Oncologist* 2003, 8.
11. Paramore A, Frantz S: Bortezomib. *Nature Reviews Drug Discovery* 2003, 2(8):611-612.
12. R. C. Kane ATF, R. Sridhara, and R. Pazdur: United States Food and Drug Administration approval summary: bortezomib for the treatment of progressive multiple myeloma after one prior therapy. *Clinical Cancer Research* 2006, 12:2955-2960.
13. António J. Marques RP, Ana C. Matias, Paula C. Ramos, and R. Jürgen Dohmen: Catalytic mechanism and assembly of the proteasome. *Chem Rev* 2009, 109(4):1509-1536.
14. Panfair D, Ramamurthy A, Kusmierczyk AR: Alpha-ring Independent Assembly of the 20S Proteasome. *Sci Rep* 2015, 5:13130.
15. Baumeister W, Walz J, Zuhl F, Seemuller E: The proteasome: paradigm of a self-compartmentalizing protease. *Cell* 1998, 92(3):367-380.
16. Tamura T, Nagy I, Lupas A, Lottspeich F, Cejka Z, Schoofs G, Tanaka K, De Mot R, Baumeister W: The first characterization of a eubacterial proteasome: the 20S complex of Rhodococcus. *Curr Biol* 1995, 5(7):766-774.
17. Seemuller E, Lupas A, Baumeister W: Autocatalytic processing of the 20S proteasome. *Nature* 1996, 382(6590):468-471.
18. Zwickl P: The 20S proteasome. *Curr Top Microbiol Immunol* 2002, 268:23-41.
19. Zuhl F, Seemuller E, Golbik R, Baumeister W: Dissecting the assembly pathway of the 20S proteasome. *FEBS Lett* 1997, 418(1-2):189-194.
20. Mayr J, Seemuller E, Muller SA, Engel A, Baumeister W: Late events in the assembly of 20S proteasomes. *J Struct Biol* 1998, 124(2-3):179-188.

21. Zwickl P, Kleinz J, Baumeister W: Critical elements in proteasome assembly. *Nat Struct Biol* 1994, 1(11):765-770.
22. Li PC, Miyashita N, Im W, Ishido S, Sugita Y: Multidimensional umbrella sampling and replica-exchange molecular dynamics simulations for structure prediction of transmembrane helix dimers. *J Comput Chem* 2014, 35(4):300-308.
23. Gandotra S, Lebron MB, Ehrt S: The Mycobacterium tuberculosis proteasome active site threonine is essential for persistence yet dispensable for replication and resistance to nitric oxide. *PLoS Pathog* 2010, 6(8):e1001040.
24. Suppahia A IP, Burris A, Kim FMG, Vontz A, Kante A, Kim S, Im W, Deeds EJ, Roelofs J.: Cooperativity in Proteasome Core Particle Maturation. *iScience* 2020, 23(5).
25. al DESe: Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *IEEE* 2014:41–53.
26. Song Y DF, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D.: High-resolution comparative modeling with RosettaCM. *tructure* 2013 Oct8, 21,10:1735-1742.
27. Raman S VR, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, Baker D.: Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 2009, 77:89-99

.
28. Jo S, Kim T, Iyer VG, Im W: CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem* 2008, 29(11):1859-1865.
29. Lee J, Cheng X, Swails JM, Yeom MS, Eastman PK, Lemkul JA, Wei S, Buckner J, Jeong JC, Qi Y *et al*: CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *Journal of Chemical Theory and Computation* 2016, 12(1):405-413.
30. Lee J, Hitzenberger M, Rieger M, Kern NR, Zacharias M, Im W: CHARMM-GUI supports the Amber force fields. *The Journal of Chemical Physics* 2020, 153(3):035103.
31. Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, Merz KM, Jr., Onufriev A, Simmerling C, Wang B, Woods RJ: The Amber biomolecular simulation programs. *Journal of computational chemistry* 2005, 26(16):1668-1688.
32. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC: Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* 2013, 9(9):3878-3888.
33. Lippert RA, Predescu C, Ierardi DJ, Mackenzie KM, Eastwood MP, Dror RO, Shaw DE: Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure. *The Journal of Chemical Physics* 2013, 139(16):164106.
34. Team RDC: R: A language and environment for statistical computing. In. Vienna, Austria; 2010.
35. Cleveland WS: Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* 1979, 74(368):829-836.
36. Hintze JL, Nelson RD: Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician* 1998, 52(2):181-184.
37. Williamson JR: Cooperativity in macromolecular assembly. *Nature Chemical Biology* 2008, 4(8):458-465.
38. Shigeo Murata HYaKT: Molecular mechanisms of proteasome assembly. *Nature reviews* 2009, 10.
39. Seemüller E, Zwickl P, Baumeister W: 12. Self-Processing of Subunits of the Proteasome. *Enzymes* 2002, 22.
40. Cheng Y, Pieters J: Novel Proteasome Inhibitors as Potential Drugs to Combat Tuberculosis. *Journal of Molecular Cell Biology* 2010, 2(4):173-175.
41. Samanovic MI, Li H, Darwin KH: The pup-proteasome system of Mycobacterium tuberculosis. *Subcell Biochem* 2013, 66:267-295.

42. Gandotra S, Schnappinger D, Monteleone M, Hillen W, Ehrt S: In vivo gene silencing identifies the Mycobacterium tuberculosis proteasome as essential for the bacteria to persist in mice. *Nat Med* 2007, 13(12):1515-1520.
43. Zühl F, Tamura T, Dolenc I, Cejka Z, Nagy I, De Mot R, Baumeister W: Subunit topology of the Rhodococcus proteasome. *FEBS Letters* 1997, 400(1):83-90.
44. Tomko RJ, Hochstrasser M: Order of the Proteasomal ATPases and Eukaryotic Proteasome Assembly. *Cell Biochemistry and Biophysics* 2011, 60(1):13-20.
45. Robert J. Tomko J, Hochstrasser M: Molecular Architecture and Assembly of the Eukaryotic Proteasome. *Annual Review of Biochemistry* 2013, 82(1):415-445.
46. Marques AJ, Palanimurugan R, Matias AC, Ramos PC, Dohmen RJ: Catalytic mechanism and assembly of the proteasome. *Chem Rev* 2009, 109(4):1509-1536.
47. Saeki Y, Tanaka K: Assembly and Function of the Proteasome. In: *Ubiquitin Family Modifiers and the Proteasome: Reviews and Protocols.* Edited by Dohmen RJ, Scheffner M. Totowa, NJ: Humana Press; 2012: 315-337.
48. Durrant JD, McCammon JA: Molecular dynamics simulations and drug discovery. *BMC Biology* 2011, 9(1):71.
49. Johnson DK, Karanicolas J: Ultra-High-Throughput Structure-Based Virtual Screening for Small-Molecule Inhibitors of Protein–Protein Interactions. *Journal of Chemical Information and Modeling* 2016, 56(2):399-411.
50. Joazeiro CAP, Anderson KC, Hunter T: Proteasome Inhibitor Drugs on the Rise. *Cancer Research* 2006, 66(16):7840.
51. Manasanch EE, Orlowski RZ: Proteasome inhibitors in cancer therapy. *Nature Reviews Clinical Oncology* 2017, 14(7):417-433.
52. Voorhees PM, Dees EC, O'Neil B, Orlowski RZ: The Proteasome as a Target for Cancer Therapy. *Clinical Cancer Research* 2003, 9(17):6316.
53. Layfield R, Lowe J, Bedford L: The ubiquitin-proteasome system and neurodegenerative disorders. *Essays in biochemistry* 2005, 41:157-171.
54. Ross CA, Pickart CM: The ubiquitin–proteasome pathway in Parkinson's disease and other neurodegenerative diseases. *Trends in Cell Biology* 2004, 14(12):703-711.

Chapter 3

Global conformational shifts act as a checkpoint in bacterial proteasome Core Particle assembly

3.1 Introduction

Proteins bind each other through noncovalent bonds and form macromolecular complexes. Macromolecular machines, like the proteasome, ribosome, spliceosome, AAA ATPases, GroEL, and virus capsids, entail many noncovalent interactions so that their subunit interfaces combine and form the final structure [1]. In general, macromolecular machines are assembled from a set of individual subunits or monomers. The assembly pathways for macromolecular machines are almost always hierarchical, with a sequence of specific steps necessary to obtain the final complex. A classic example is the proteasome, which also exhibits a hierarchal pathway to assemble into the final complex.

Proteasomes are critical proteases involved in the degradation of damaged and unwanted proteins. Since they are involved in regulating the turnover of numerous proteins in cells, their activity is tightly controlled [2]. The proteasome 20S Core Particle (CP) forms the central component of the complex and is found in all three kingdoms of life. The 20S CP quaternary structure is highly conserved and consists of four stacked rings. Each ring comprises seven subunits, the outer two rings are made up of α subunits, and the inner rings are made up of β subunits (which are catalytically active) [2]. In prokaryotes, the four CP rings are homo-heptameric (made up of one type of α and β subunits), whereas, in eukaryotes, the rings are hetero-heptameric (made up of seven distinct types of α and seven distinct β subunits).

Assembly of proteasome CP has been widely studied in organisms ranging from humans to archaea and bacteria. CP assembly begins with forming a Half Proteasome (HP- $\alpha_7\beta_7$) for all species (**Fig. 3.1**). All the

β subunits are synthesized in an inactive zymogen form with a propeptide at its N-Terminal. When two HPs dimerize, the propeptide is auto-catalytically cleaved, and an active CP is formed [3, 4]. Thus, a CP assembly does not occur unless two HPs dimerize and form the necessary interactions (**Fig. 3.1**). The pathway leading to the formation of the HP can differ depending on the organism in question. Still, CP formation always occurs from the association of two HPs in all organisms. Interestingly there is no evidence of proteasome CPs that have less than 28 subunits arranged as four stacked rings. One of the critical questions about CP assembly currently unanswered is how does the HP "know" that it must associate with another HP?

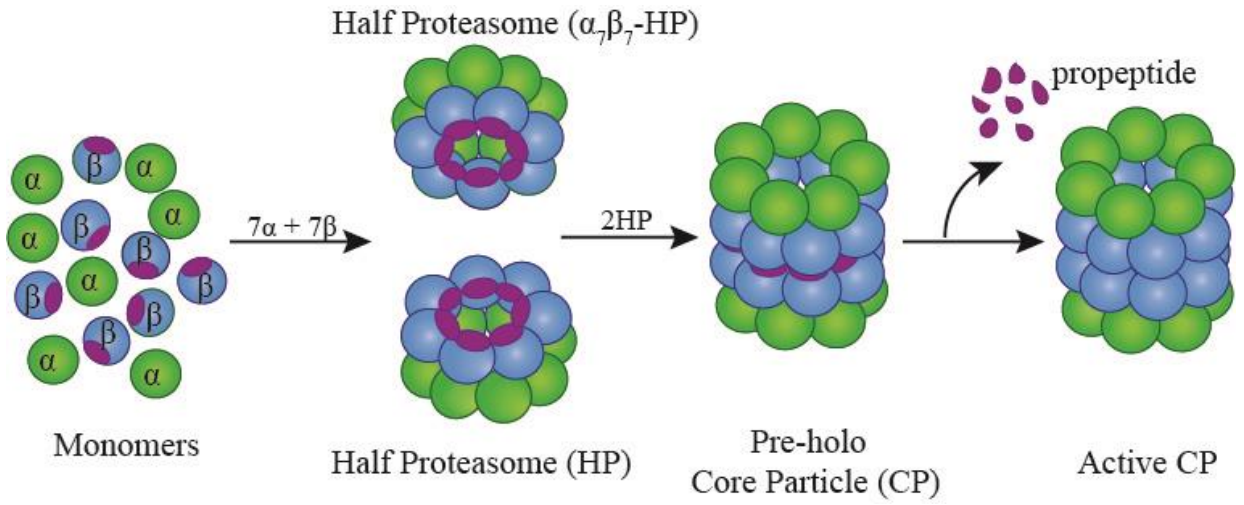


Figure 3.1: Schematic of the 20S proteasome assembly. The α subunits are shown in green and β subunits are shown in blue, and the propeptide in purple. Two Half Proteasomes (HP) associate to form a Pre-holo Core Particle and then the propeptide is autocatalytically cleaved off, assembling into the active CP.

In this work, we chose the bacterial *Rhodococcus erythropolis* proteasome CP as our model system. In prokaryotes, CP assembly is spontaneous and occurs without using any chaperones or assisting assembly factors. But, in all organisms, CP assembly is universally hierarchical (Fig. 3.1), and HP is an obligatory intermediate in the CP assembly of any organism. In Chapter 2, we addressed a critical question in CP assembly of bacterium *Rhodococcus erythropolis* (*Re*), a close relative of *Mycobacterium tuberculosis* species. Specifically, we investigated the dimerization of two HPs into a CP in *Re*. While the findings discussed in Chapter 2 provide insights into the separation of time scales in CP assembly, many questions

about CP assembly remain unanswered. As mentioned above, HP dimerization occurs only when both reactants are true HP ($\alpha_7\beta_7$) structures. Experimental work on *Re* CP assembly has shown that intermediates which are near-HP, for example, $\alpha_6\beta_7$ (i.e., the HP missing a single α subunit), $\alpha_7\beta_6$, and $\alpha_6\beta_6$ occur during assembly and that they persist for non-trivial amounts of time [3, 5, 6]. However, these intermediates never dimerize, either with one another or with "true" $\alpha_7\beta_7$ HPs, to form aberrant CP-like structures (e.g., $\alpha_6\beta_7\beta_7\alpha_7$) (**Fig. 3.2A, B and C**). For instance, in the near-HP intermediate $\alpha_6\beta$, the entire β ring is formed. Yet, these intermediates never dimerize with the $\alpha_7\beta_7$ HP (**Fig. 3. 2A**). In some way, the information about the missing subunit is transmitted over ~ 30 Å to the HP interface to prevent dimerization. There can be many such intermediates during CP assembly, but we consider three important near-HP intermediates in this work, as shown in **Fig. 3.2A, B, and C**. It is currently unknown how the proteasome subunits allosterically communicate to achieve hierarchical assembly pathways.

Incorrect assembly and incomplete structures like $\alpha_6\beta_7\beta_7\alpha_7$ or $\alpha_6\beta_6\beta_7\alpha_7$ would likely allow proteins to be degraded in an unregulated manner and disturb the cell's homeostasis. Additionally, such structures might lead to kinetic trapping, reducing the assembly yield and further influencing assembly dynamics. We chose these three intermediates $\alpha_6\beta_7$, $\alpha_7\beta_6$, and $\alpha_6\beta_6$ as they are on-pathway intermediates for the various proposed CP assembly pathways, which are discussed in detail in Chapter 4. We initially hypothesized that the β propeptide might cause conformational transitions that make the near-HP states non-dimerizable. To date, there have been no molecular simulation studies for understanding the type of allosteric communication that characterizes the intermediates.

To investigate our hypothesis, we employed Molecular Dynamics (MD) simulation on our model system - *Rhodococcus erythropolis* (*Re*). MD is an appealing method to study processes at the molecular and atomistic level, where traditional experimental techniques cannot provide detailed insights. A major limitation for assembly intermediates in *Re* is that we have no HP or any other intermediate crystal structures available. We decided to delete the subunits in the crystal structure (PDB:1Q5R) for the *Re* CP to overcome this. Since these modified starting structures are not obtained from the solution, we need extended MD simulation runs to achieve equilibrium sampling.

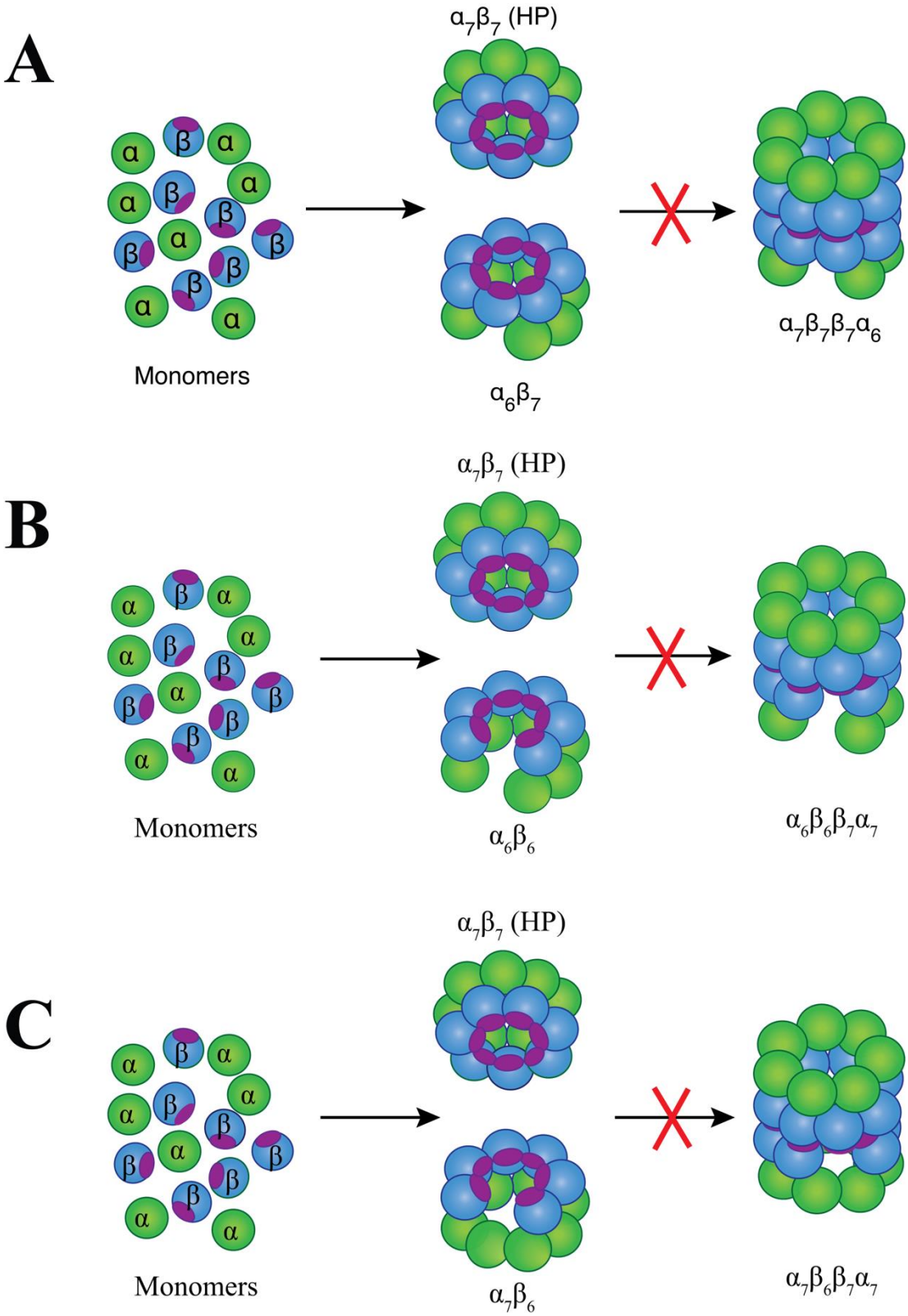


Figure 3.2: Schematic showing three examples of reactions that do not occur in CP assembly. A) The near-HP intermediate $\alpha_6\beta_7$ does not dimerize with a true HP ($\alpha_7\beta_7$). B) The near-HP intermediate $\alpha_6\beta_6$ does not dimerize with an HP C) The near-HP intermediate $\alpha_7\beta_6$ does not dimerize with an HP.

For this, we have used the special purpose supercomputer Anton and obtained 2.5 μs of simulation for three replicates of each $\alpha_6\beta_7$, $\alpha_7\beta_6$, and $\alpha_6\beta_6$ intermediates. We compared these results to the WT *Re* HP simulations described in detail in Chapter 2 of this dissertation.

Our MD simulations reveal global conformational shifts in these near-HP structures, particularly distortion in β ring symmetry, tilt, and geometry. These distortions change the entire confirmation and the structure of the intermediate. For the $\alpha_7\beta_6$ and $\alpha_6\beta_6$ simulations, we observed changes in the β ring geometry; the angles made by the β subunits relative to one another changed significantly when compared with the HP simulations. For $\alpha_6\beta_7$, we observed that the β subunit with a missing alpha (β_7) has a different tilt angle changing the interfaces with its neighboring β subunits. It thus appears that our initial hypothesis regarding β propeptide is not the primary contributor for the non-dimerization of near-HP intermediates; instead, a global conformational shift occurs in the intermediate's ensembles. Taken together, all the conformational changes and distortions in symmetry change the dimerization interface and prevent critical interactions that are needed in the near-HP structures; thus, they are likely prevented from associating with themselves or another HP.

The insights from simulations elucidated the mechanism of why near-HP intermediates cannot dimerize with HPs or themselves. Our work strongly suggests that the interfaces between the protein subunits have evolved to be frustrated [7, 8] and therefore are "intrinsically frustrated." For instance, we observed from the simulations that the β subunit angles in the $\alpha_6\beta_6$ simulations are not the same as we see in HP simulations. Therefore, the addition of the last αβ dimer generates a non-optimal set of β subunit angles, suggesting the observed bond angles in the CP structure are frustrated. These findings suggest that such global confirmational shifts may occur in many hierarchical assembly pathways and not just proteasomes. If this true, then the allosteric mechanism among subunits of macromolecular assemblies acts as a checkpoint factor to prevent aberrant structure formation and as an evolutionary approach to maximize assembly yields of complexes. Additionally, understanding the molecular mechanism of the allosteric communication, which prevents dimerization of near-HP intermediates, will give insights into specific interactions in α and β subunits. These interactions can be modulated to develop specific and less toxic drugs in the future.

3.2 Materials and Methods

3.2.1 Modeling missing propeptide residues

The starting structure for the *Rhodococcus erythropolis* CP was obtained by taking the top-half of CP (Point mutant CP with the propeptide) coordinates, i.e., 14 chains from PDB 1Q5R [4]. We mutate the A33 to K33 in the β subunits (as the crystal structure is a point mutant). The missing regions of the propeptide were modeled using Rosetta Comparative Modeling [9, 10] hosted at the Robetta server. For this, we first modeled a single αβ dimer and then repeated the same dimer model to obtain the HP structure. We used the Wild Type (WT) modeled structure used for simulations in Chapter 2.2 as our starting structure. The three intermediates were generated from this structure by deleting the relevant chains in CHARMM-GUI [11].

3.2.2 Molecular Dynamics simulations setup

The simulation inputs were generated using the CHARMM-GUI solution builder module [11-13]. The systems were neutralized with 100mM NaCl (the same salt concentration used for experiments), and 15A° water was added on each side of the protein in a rectangular box. The detailed components of the simulations are described in **Table B.8**. We performed a short minimization followed by 5ns of NVT and 60ns NPT equilibration with a 2fs time step using the GPU version of AMBER18 [14, 15] using the CHARMM36m forcefield [16]. Each simulation took ~ 48 hours to finish 65ns of equilibration on NVIDIA RTX 2080Ti GPU. After finishing AMBER equilibrations, the coordinates and restart files were used to initiate 2.5 microseconds simulations on Anton2 for long time scale Molecular Dynamics simulations [17]. The NPT ensemble and CHARMM version [18] of the TIP3P water model [16] were used for the Anton simulations. Additionally, the pressure and temperature are constant at 1bar and 303.15 K using Multigrator integrator and default NPT parameters [19]. The coordinates are saved every 0.24ns. For the RESPA (Reference systems propagator algorithms) scheme, every second time step was used for evaluating long-range interactions. The pressure was controlled using the Martyna-Tobias-Klein (MTK) barostat [20] and an interval length of 480 picoseconds. The temperature was maintained by the Nose-Hoover thermostat

[21] with an interval length of 24 picoseconds. A relaxation time of τ=0.041667 picoseconds was used for the barostat and thermostat. All systems were simulated under periodic boundaries in the NPT ensemble.

3.2.3 Estimation of the angles (βθ) of the β subunits

To estimate the βθ, which is the angle made by every β subunit with reference to the center of the β ring, we used the center of mass and fundamental vector algebra operations. For this, we select all atoms in the β ring (without propeptide) and translate its center of mass to origin. Next, we select every β subunit (β_n) in a clockwise direction (where n refers to the number of β subunits in the simulation), and its left neighboring β subunit (β_{n+1}) and obtain the center of mass for both these selections (**Fig. B.3.1**). This results in two vectors for each set of β subunits. Next, we use the dot product to calculate the angle made between these two β subunits (βθ) (**Fig. B.3.1**). This process is repeated for seven β subunits in $\alpha_6\beta_7$ and HP simulations, and six β subunits for $\alpha_7\beta_6$, and $\alpha_6\beta_6$. To estimate the distortion in β ring geometry of the intermediates for intermediates $\alpha_7\beta_6$, and $\alpha_6\beta_6$, we calculate angle β_6-β_1 to which is then compared with the β_6-β_1 of HP and represents the extent to which the $\alpha_7\beta_6$ and $\alpha_6\beta_6$ structures are collapsing.

3.2.4 Estimation of the angles (βØtilt) of the β subunits

To estimate the βØ tilt, which is to estimate the rotation of β subunits in the intermediate's simulations. For this, we select the β ring (without propeptide) and translate its center of mass to origin. As we are estimating the torsional angles, we need to have four points in two planes. One set of points are the center of mass of two selected β subunits, and the other set of points are the center of mass of a helix (H1: residues 42 to 62) in the β subunits. Next, we select every β subunit in an anti-clockwise direction and its right neighboring β subunit and obtain the center of mass for both these selections (**Fig. B.3.2**). So, we get two vectors for every four points for each set of β subunits. Next, we use these vectors to calculate the torsional angle between the β subunits, which we refer to as "βθtilt" (**Fig. B.3.2**). This process is repeated for seven β subunits in $\alpha_6\beta_7$ and HP simulations, and six β subunits for $\alpha_7\beta_6$, and $\alpha_6\beta_6$.

3.2.5 Statistical analysis for βØ and βØtilt values

To estimate the significance of changes in βØ and βØtilt for the WT and intermediates, we have used categorical regression using the Newey-West estimator using R packages [22]. This test is used to determine if the time-dependent behavior of $\alpha_7\beta_7$ is different from $\alpha_6\beta_7$, $\alpha_7\beta_6$, and $\alpha_6\beta_6$, in other words, to estimate the probability that they are from the same distributions. Unfortunately, MD simulation data is time-dependent and hence autocorrelated; it is challenging to find a statistical test to estimate the significance of our simulations. To account for these limitations, we used Newey-West estimators [22]. The *p-values* from the categorical regression statistical tests for every replicate and each β subunit for βθ and βØtilt calculations are reported in Appendix (**Figs. B.5 and B.7**).

3.3. Results

3.3.1 Region I have higher RMSF in all the near-HP's than HP simulations

Electron microscopy, Mass spectrometry, and native-gel experiments provide evidence of near-HP intermediates like $\alpha_6\beta_7$, $\alpha_7\beta_6$, $\alpha_6\beta_6$, and $\alpha_6\beta_5$, etc. during Core Particle (CP) assembly [3, 5, 6]. These near-HP intermediates are non-obligatory and assembly-incompetent as they do not dimerize with each other or another "true Half Proteasome" (HP- $\alpha_7\beta_7$) to form a CP (**Fig. 3.1**). In Chapter 2, we focused on a specific step in *Rhodococcus erythropolis* CP assembly, where we demonstrated that the β propeptide is involved in regulating the dimerization rates. Changes to the Region III length or amino acid composition impact the time required for CP assembly. Here, we focus on a different aspect of assembly, i.e., why an HP and near-HP intermediates do not dimerize. We hypothesized that similar to the separation of timescales, the propeptide induces transitions in near-HP intermediates to a non-dimerizable state that cannot associate with an HP, and dimerization is highly unlikely to occur. $\alpha_7\beta_7$

The *Re* propeptide is 65 residues long and is categorized into three regions based on position in the *Re* HP, conserved residues, secondary structure, and electron density of propeptide residues (**Fig. 3.3.1A**). The three propeptide regions (sequence numbering refers to *Re* crystal structure) are shown in (**Fig. 3.3.1A, B)**. Region I is N terminal Residues from -65th to -43rd, which are near the α subunits. Next, Region II: Residues from -42nd to -27th have several conserved residues and forms the crucial central box region [3]; this region residues form contacts with both α and β subunits. Lastly, Region III comprises from Residues from -26th to -1st, which are near the β subunits and located around the HP dimerization interface.

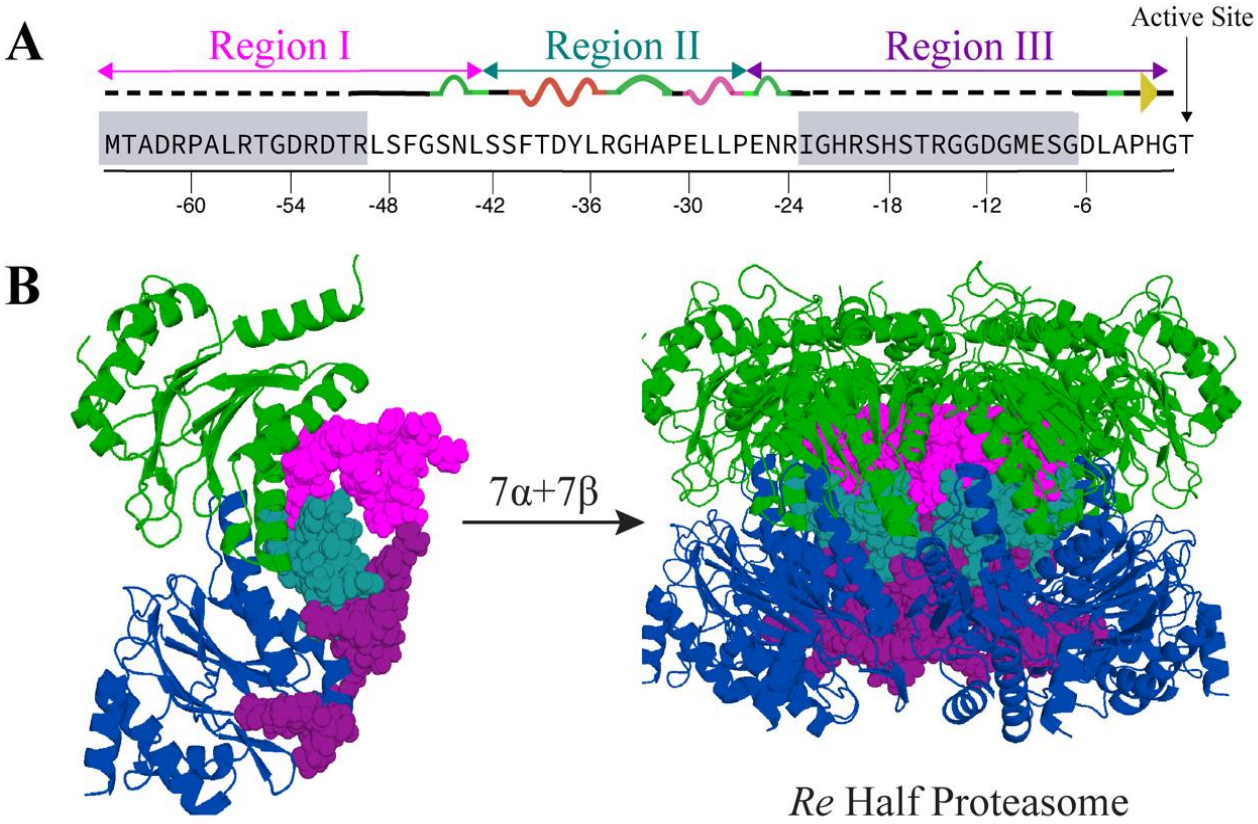


Figure 3.3.1: A) *Re* propeptide sequence showing the three regions. The residues shaded with a grey background do not have electron density in the crystal structure (1Q5R) and are modeled. Region I (magenta) is near to α subunits, Region II (teal) is at the interface of α and β subunits, and Region III (purple) is near β subunits and the HP dimerization interface. B) A α (green) and β (blue) dimer and the *Re* HP are shown in cartoon representation.

To investigate our hypothesis that the propeptide induces transitions that make the near-HP conformations non-dimerizable, we performed all-atom unbiased MD simulations of three intermediates $\alpha_6\beta_7$, $\alpha_7\beta_6$, and $\alpha_6\beta_6$. These intermediates are missing one or two subunits from the true HP. Our simulations

for the HP are extracted from a mutant CP (PDB: 1Q5R) crystal structure that cannot cleave the propeptide. One significant challenge faced in this work is that there are no available structures of the intermediates we wish to simulate. All available structural data is for the entire CP and not for any of the CP assembly intermediates. So, we begin all the intermediate simulations using coordinates extracted from the top half of a CP structure and then deleting the subunits (**Fig. B.1**) as needed in CHARMM-GUI [11].

3.3.2 Propeptide Region I is highly flexible in near-HP intermediates

All-atom MD simulations were run for 2.5 μs on the Anton2 supercomputing resource [17]. After visualizing the trajectories, we observed that Region I (-65^{th} to $-43^{rd)}$ of the propeptide was highly flexible and fluctuated more than Region II (-42^{nd} to -27^{th}) and Region III (-26^{th} to -1^{st}). Interestingly, we also saw that Region I of the propeptide comes near the space where the missing subunits (either α or β) would occupy if it were a complete HP. In other words, the propeptides of the neighbors of missing the β subunits or diagonally opposite of the missing β subunits essentially move to partially fill in the space of the missing α or β subunits.

We performed a Root Mean Square Fluctuation (RMSF) analysis to calculate the overall fluctuations of the entire 65 residues long propeptide for HP-$\alpha_7\beta_7$ and the three intermediate simulations. As seen in **Fig. 3.3.2** the Region, I have a 3- 4Å higher RMSF than the *Re* HP. We also observed that Region III does not fluctuate as much for near-HP intermediates as for WT and the mutants used in the earlier study the Chapter 2 (**Fig. B2).** The Region I RMSD of the intermediate simulations further confirmed that all the three intermediates have a notably higher Region I RMSD than the HP, and the missing subunits likely give the compact propeptides more space and mobility inside the rings.

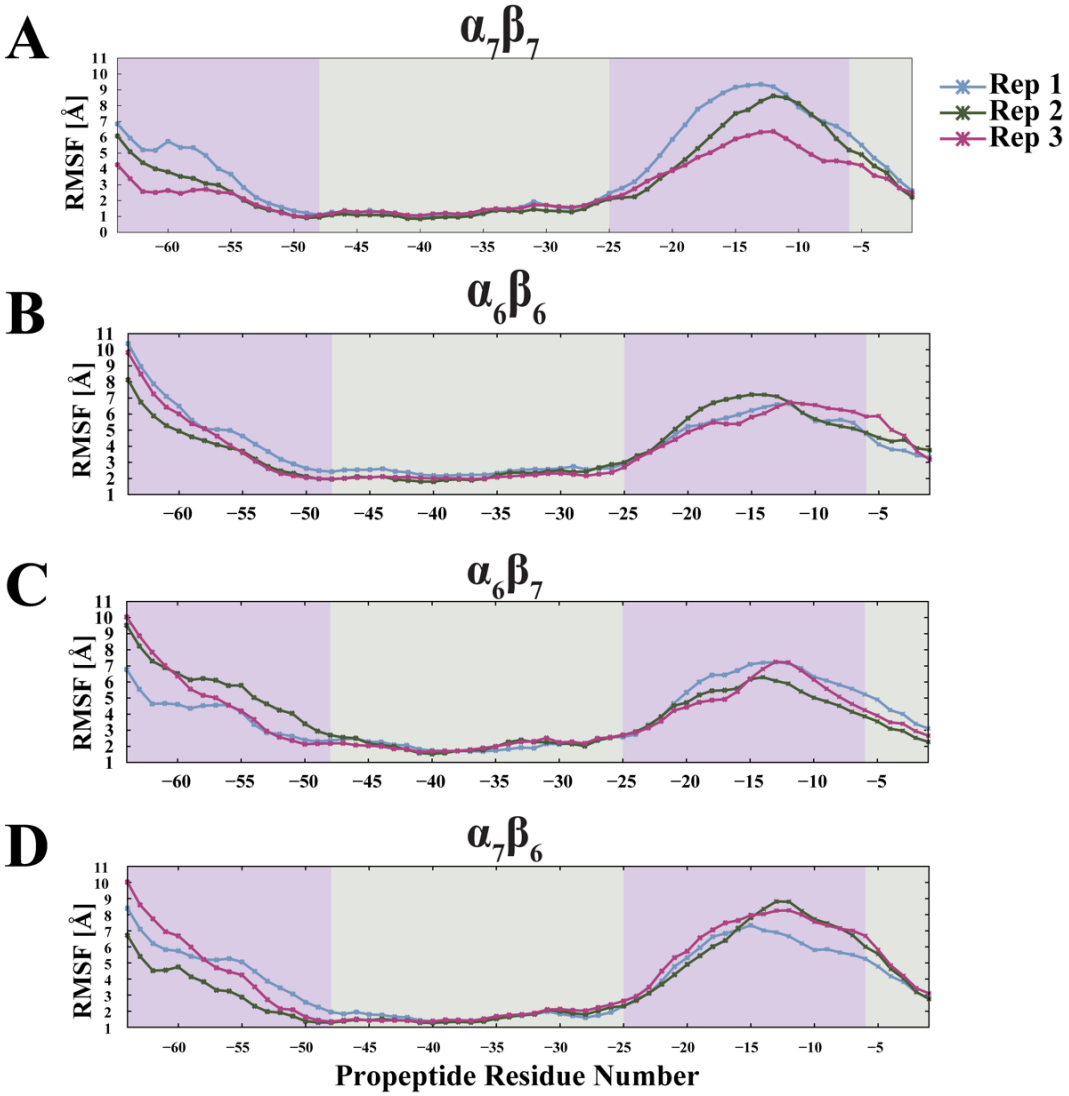


Figure 3.3.2: RMSF plots of propeptide backbone atoms in WT and intermediates are shown for three replicates and averages over 2.5 microseconds. The pink shaded regions indicate the residues with missing electron density, and grey regions have electron density. The panel's A) B) C) and D) are for the four different intermediates simulations as indicates. The three lines correspond to the three replicates for each system.

3.3.3 The $\alpha_6\beta_6$ intermediate collapses into a more compact information

The $\alpha_6\beta_6$ like the other simulations began with the crystal structure and were equilibrated for 65 ns before running on the Anton resource. Nearly after ~ 500-600ns, the subunits without a neighbor start to come closer to one another. At the beginning of the simulation, we see a void for the missing β in $\alpha_6\beta_6$ (**Fig. 3.4.1 and 3.4.2**). With time the β_1 and β_6 subunits come very close to each other and fill up that void (**Fig. 3.4.1**).

This conformational shift is clearly seen in the bottom view of β subunits; the closed conformation is seen at the end of 2μs (**Fig. 3.4.2**) and highlights the structural rearrangement. This interesting structural rearrangement could be the structure of $\alpha_6\beta_6$ in solution, but we do not have any crystal structures of $\alpha_6\beta_6$ or any of the intermediates. It is also interesting to see that these changes are for the entire $\alpha_6\beta_6$ and not just one or two subunits.

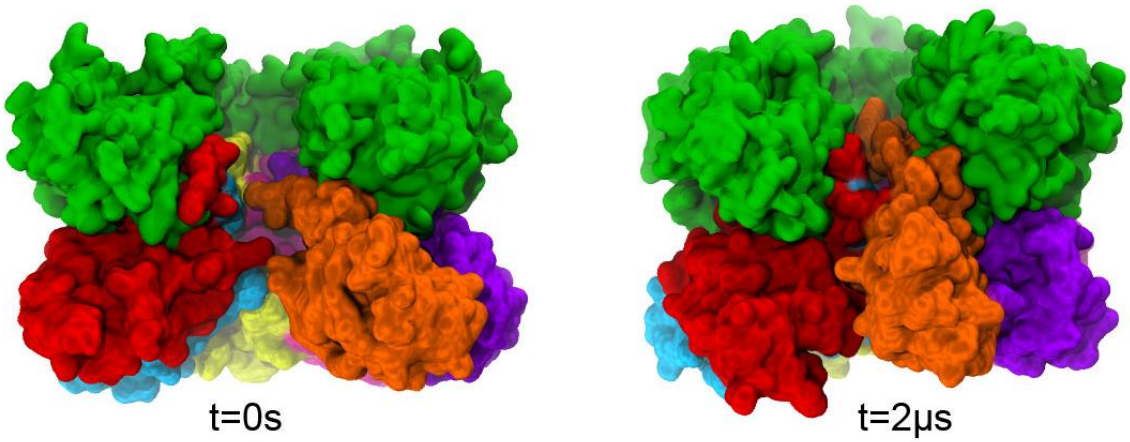


Figure 3.4.1: Side view of $\alpha_6\beta_6$ MD simulations at the beginning (0 seconds) and the end of 2μs shown in surface representations. The α subunits are in green, and the β subunits are in six different colors.

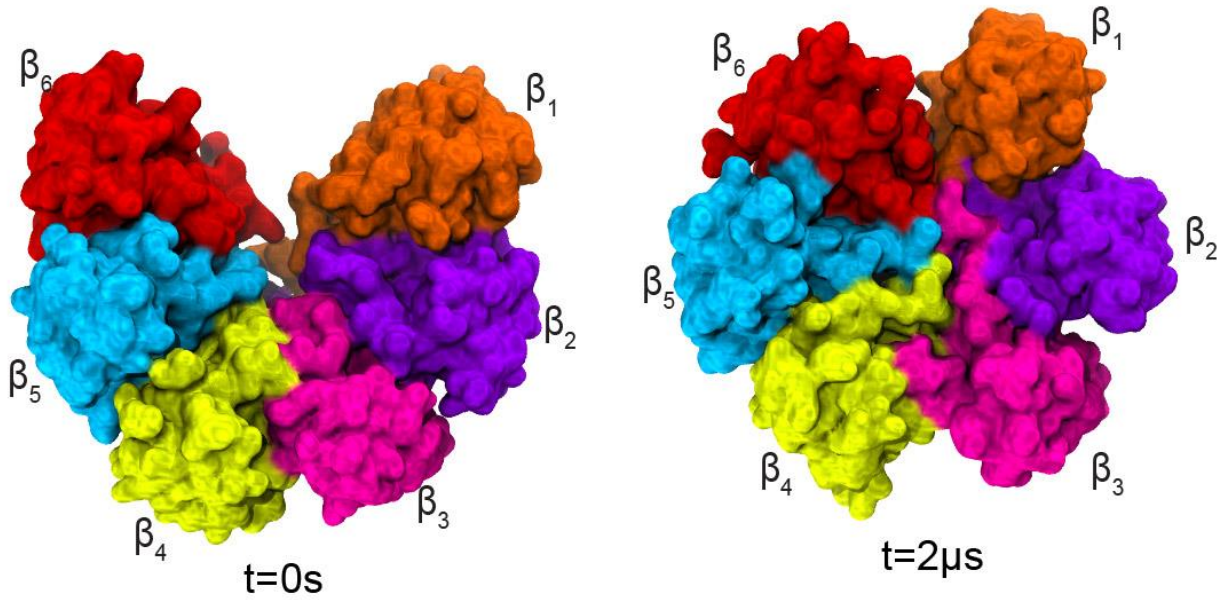


Figure 3.4.2: Bottom view of β subunits in $\alpha_6\beta_6$ MD simulations at the beginning (0 seconds) and the end of 2μs shown in surface representations. The α subunits are hidden, and the β subunits are in six different colors as indicated.

3.3.4 The $\alpha_7\beta_6$ intermediate shows β subunits getting closer

As observed in $\alpha_6\beta_6$ we also noticed structural rearrangements in the β ring for $\alpha_7\beta_6$.

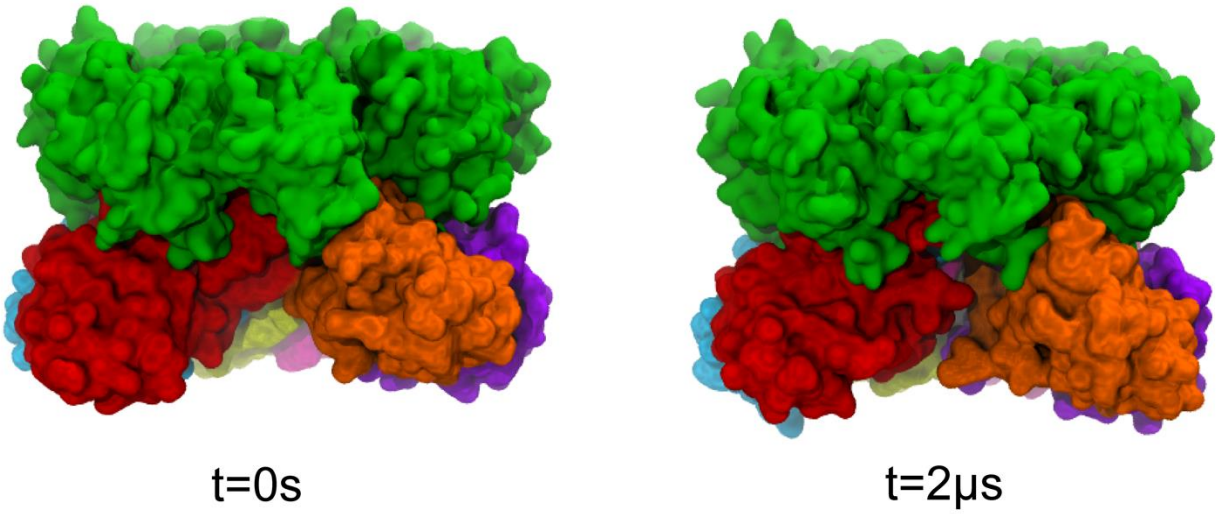


Figure 3.5.1: Side view of $\alpha_7\beta_6$ MD simulations at the beginning (0 seconds) and the end of 2μs shown in surface representations. The α subunits are in green, and the β subunits are in six different colors.

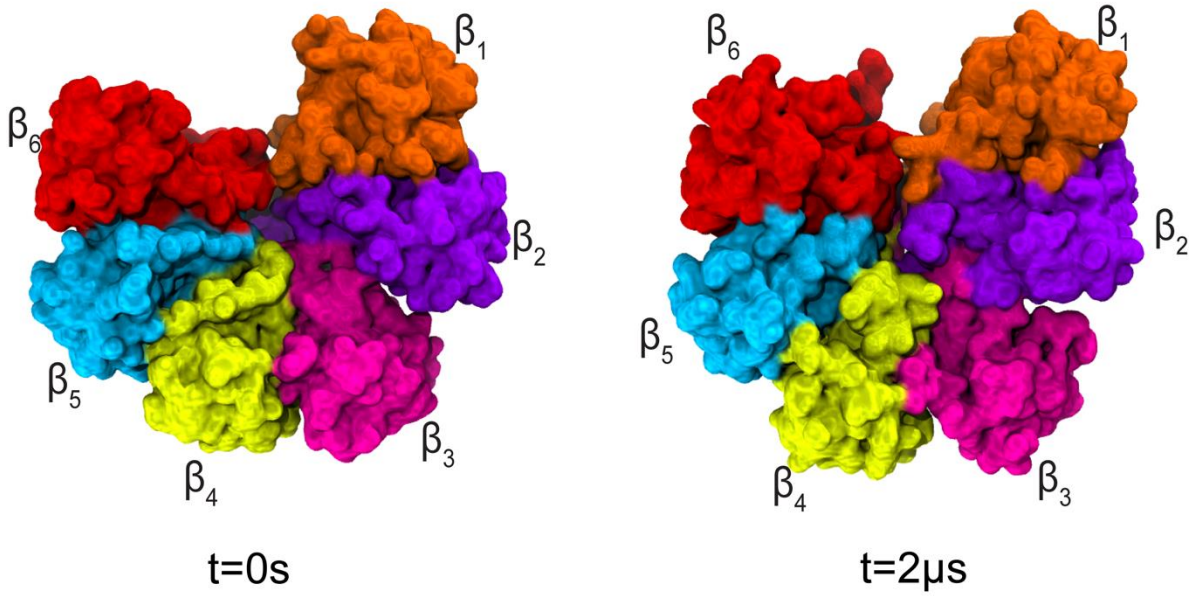


Figure 3.5.2: Bottom view of β subunits in $\alpha_7\beta_6$ MD simulations at the beginning (0 seconds) and the end of 2μs shown in surface representations. The α subunits are hidden, and the β subunits are in six different colors as indicated.

3.3.5 $\alpha_7\beta_7$ simulations shows no similar transitions as seen in $\alpha_6\beta_6$ and $\alpha_7\beta_6$

The conformational transitions that we had observed in the intermediate simulations are likely due to the lesser stability of the interfaces, which arises from missing subunits. As a control simulation, we looked at the HP ($\alpha_7\beta_7$) simulations (Chapter 2) we ran under the same conditions and simulation length. We did not observe any drastic transitions in any of the β subunits (**Fig. 3.6.1 and Fig. 3.6.2**). The HP side view and the bottom view of β subunits show transitions that reflect protein dynamics and are likely how the HP structures look in the solution.

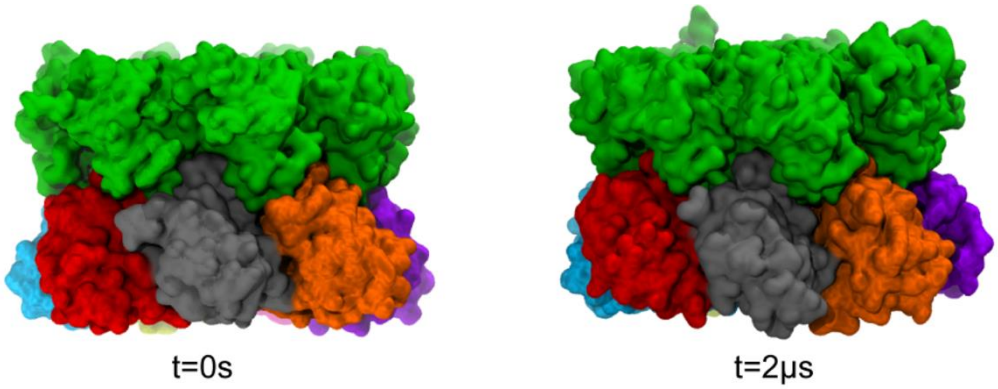


Figure 3.6.1: Side view of $\alpha_7\beta_7$ – HP MD simulations at the beginning (0 seconds) and the end of 2μs shown in surface representations. The α subunits are in green, and the β subunits are in seven different colors.

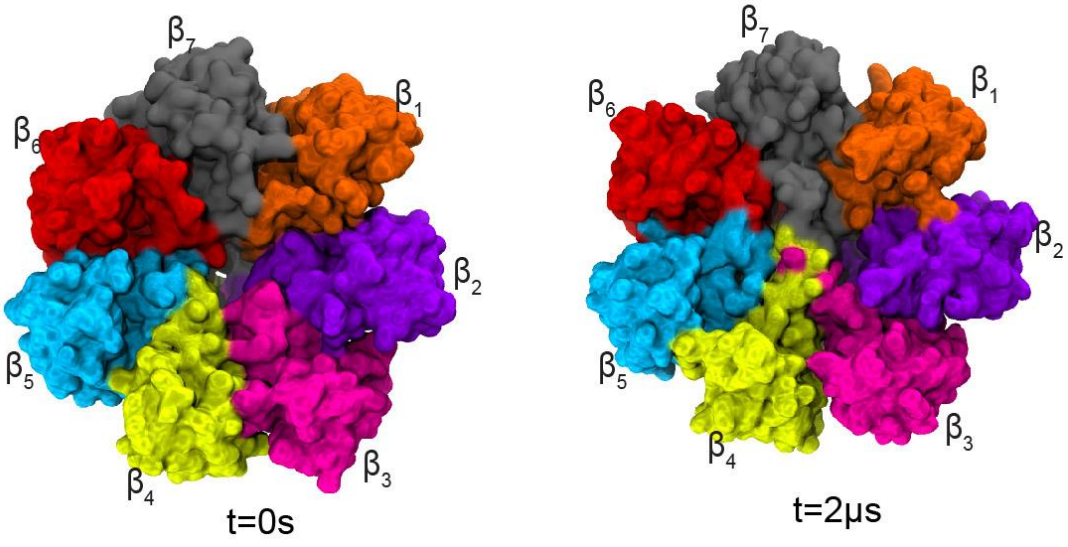


Figure 3.6.2: Bottom view of β subunits in $\alpha_7\beta_7$ - HP MD simulations at the beginning (0 seconds) and the end of 2μs shown in surface representations. The α subunits are hidden, and the β subunits are in seven different colors as indicated.

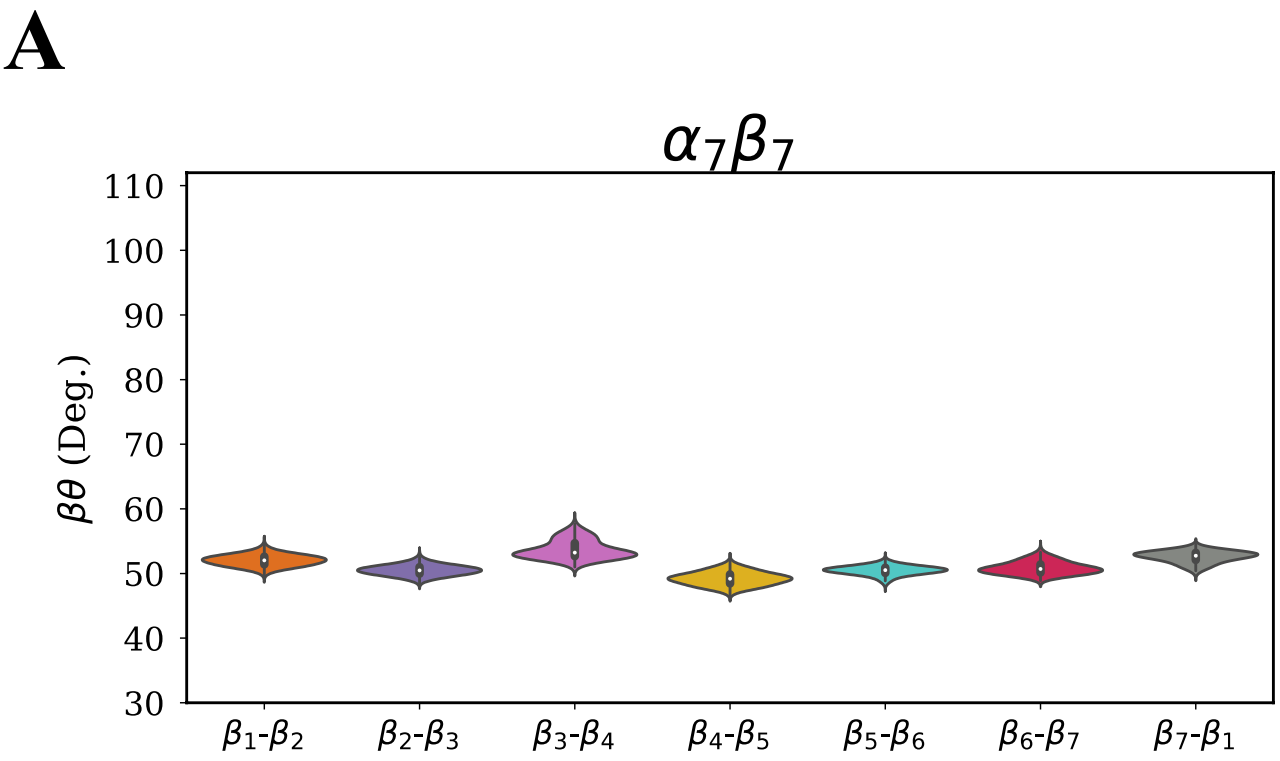
3.3.6 The angle between β subunits (βθ) in intermediate is significantly different from HP simulations

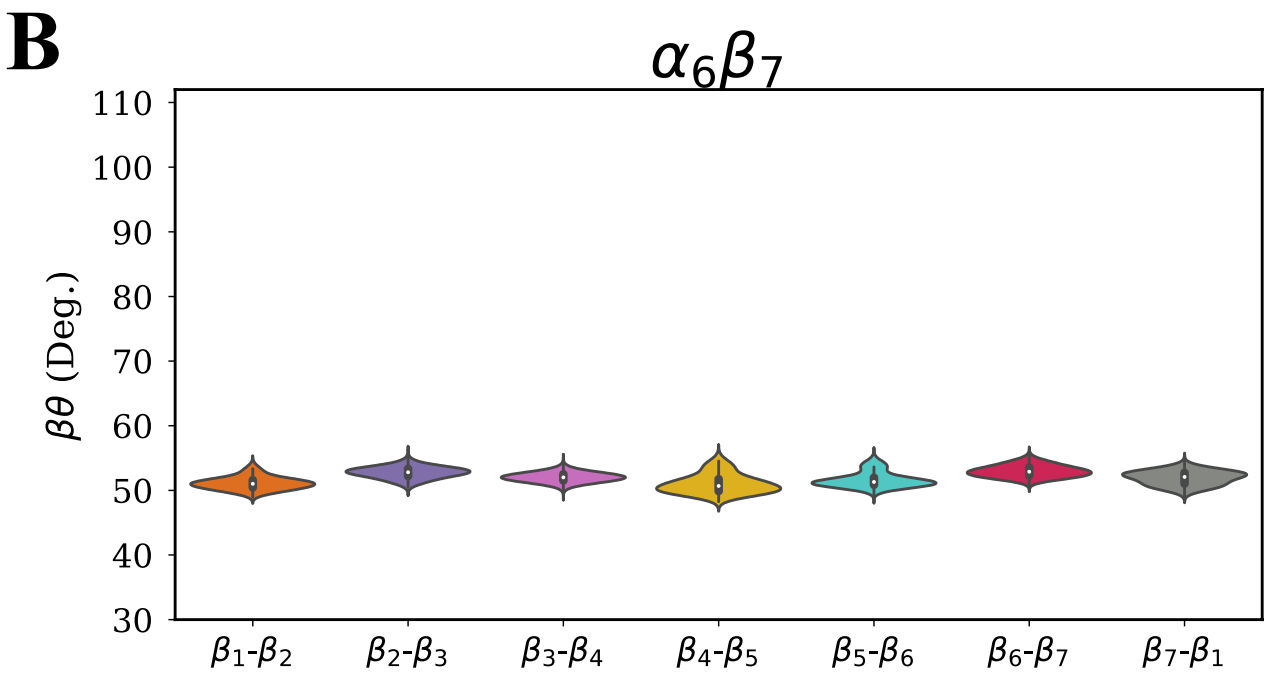
All 20S proteasomes have a C2 symmetry [23]; having such symmetry is advantageous for protein complexes because they offer increased stability, cooperativity and lead to lower energy complexes. We designed a metric βθ to quantify the changes and distortion seen in the simulation trajectories (βθ). Here, we calculate the angle made between every β subunit and its neighboring β on the right-side and the center of mass of the seven membered β ring (**Fig. B.3**). The *Re* HP crystal structure has βθ around 51^0 for all β subunits; we hypothesized that the intermediates would have a considerably different number and distribution from a βθ of 51^0.

The violin plots in Fig. 3.7 describe the distribution of βθ values for each β subunit in the simulation from 500ns to 2.5μs (500 ns simulations are considered equilibration). The median values of βθ (for all seven β subunits like the initial β_1-β_2 etc.,) in $\alpha_7\beta_7$ and $\alpha_6\beta_7$ are around 50^0- 60^0 (**Fig. 3.7A, B**). As seen in **Fig. 3.7 A, B**, the distributions for all β subunits do not fluctuate much above the βθ that is seen in the *Re* crystal structure (51^0). The $\alpha_6\beta_7$ looks very similar to $\alpha_7\beta_7$ as it has all seven β subunits, and this helps to maintain the total of about ~360^0 for the entire β ring. We see by visualization (**Fig. 3.7C, D**) that the other two intermediates, $\alpha_6\beta_6$ and $\alpha_7\beta_6$, have a very different distribution when compared to $\alpha_7\beta_7$ and thus are in a distinctly different conformation and symmetry. This distortion in ring geometry and shape indicates a global conformational transition in the intermediates. Very likely, the interfaces in $\alpha_6\beta_6$ and $\alpha_7\beta_6$ are likely not in the correct orientation and conformation to associate with another HP

For $\alpha_6\beta_6$ simulations, as seen in **Fig. 3.7C**, the subunits β_1-β_4 have the βθ much higher than $\alpha_7\beta_7$. Interestingly, the β_6-β_1 angle has a broad distribution, indicates the position where the missing β occupies becomes collapsed. Moreover, when we compare the same β_6-β_1 angle in $\alpha_7\beta_7$ (HP), the value should be 102^0. Therefore, the global conformational shifts bring β_1 and β_6 very close, and thus another β subunit cannot easily fit into the ring due to the steric hindrances. The βθ for subunits $\beta_{6\text{-}1}$ (**Fig. 3.7C** and **Fig. B.4C**) rapidly decreased after 500ns, indicating that the symmetry of the intermediate has changed. Similarly, for

$\alpha_7\beta_6$ in **Fig. 3.7 D**, the incomplete β ring symmetry is distorted, and the structure is also more closed and





collapsed. We also see that the β subunits which have neighboring β and α subunits, specifically β_2-β_3 and β_3-β_4, are very distorted and different from the HP structures shown in **Fig. 3.7A**. This is because the β_6 is

getting closer to β_1 as shown in red violin (Fig. 3.3 D), and this forces the distant β subunits to have a higher βθ to retain the ~360^0 total angles (sum of all seven β subunits).

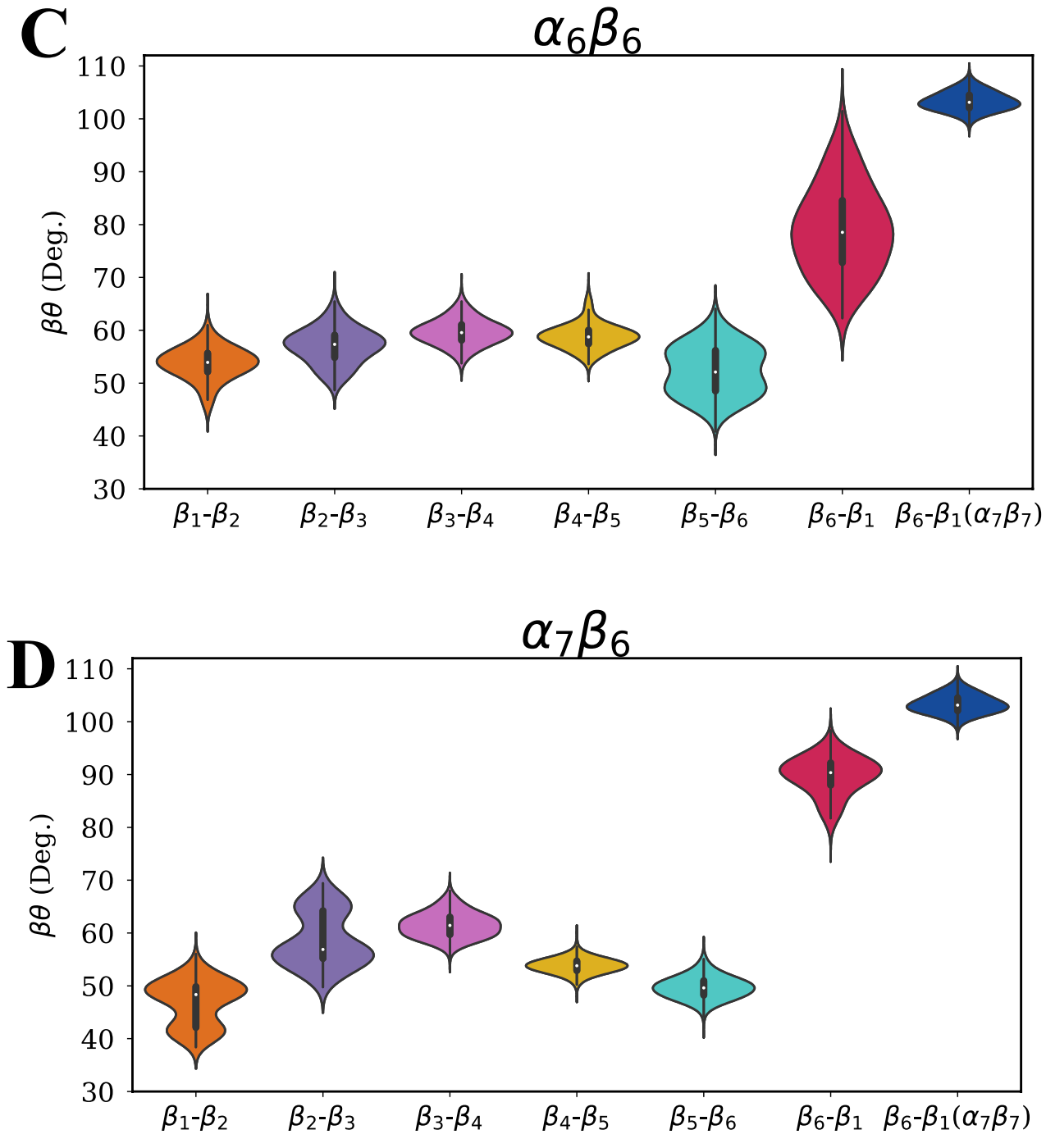


Figure 3.7: Violin distributions of βθ for all β subunits in different colors of the four sets of simulations. A) $\alpha_7\beta_7$ (HP) simulations and their distributions. B) $\alpha_6\beta_7$ C) $\alpha_6\beta_6$ including the $\beta_{6\text{-}1}$ of $\alpha_7\beta_7$. D) $\alpha_7\beta_6$ including the $\beta_{6\text{-}1}$ of $\alpha_7\beta_7$ in blue.

The set of key residues at the HP dimerization interface described in Chapter 2 (**Fig. 2.2**) are a part of highly critical interactions that drive CP assembly; any perturbations to interactions has a drastic impact on the CP formation [24]. After looking at the βθ distributions and the snapshots from simulations, it is very unlikely for the key residues of another HP to bind with near-HP intermediates and participate in essential interactions like salt bridges, hydrophobic interactions, and hydrogen. Hence, we conclude that the $\alpha_6\beta_6$ and $\alpha_7\beta_6$ intermediates are in a non-dimerizable state and incorrect conformation to associate with an HP. Change in βθ as a function of simulation time is shown in **Fig. B.4**. We can see the heterogeneity among the β subunits and clear outliers that distinguish the intermediates, especially $\alpha_6\beta_6$ and $\alpha_7\beta_6$, from the HP. Though we did not see anything notably distinct in βθ distributions for $\alpha_6\beta_7$, we did observe (**Fig. 3.5**) that a few of the β subunits which are around the missing α attain a different conformation and are distinct from the other subunits throughout the trajectory. Very likely, $\alpha_6\beta_7$ follows a different mechanism than $\alpha_6\beta_6$ and $\alpha_7\beta_6$ for not associating with an HP in the CP assembly step.

3.3.7 Statistical Analysis for βθ

A common feature of Molecular Dynamics simulations involving multiple replicates is that independent simulations can evolve differently and lead to different results, despite starting from the same starting structure and parameters. A significant challenge with MD simulations is that they are time-dependent and autocorrelated, and this needs to be considered in any statistical analysis. We thus have used the Newey West estimators [22], which are robust to autocorrelators and heteroscedasticities, with categorical variables to account for these factors. Categorical variables are used as separate binary variables for every $\alpha_7\beta_7$ replicate and every intermediate replicate. The linear regression, estimates, and tests are done using the linear model (lm) function in R 3.5.2 [25].

To consider that multiple simulation replicates can lead to different results, we have done a pairwise categorical regression, i.e., every HP replicate ($\alpha_7\beta_7$) is compared with every intermediate. Since we are looking for changes in β subunits, all the β subunits comparisons are tested.

	HP Intercept	HP Slope	Rep1-$\alpha_6\beta_6$ Intercept	Rep1-$\alpha_6\beta_6$ Slope	HP Intercept	HP Slope	Rep2-$\alpha_6\beta_6$ Intercept	Rep2-$\alpha_6\beta_6$ Slope	HP Intercept	HP Slope	Rep3-$\alpha_6\beta_6$ Intercept	Rep3-$\alpha_6\beta_6$ Slope
Rep1-HP β_6-β_1	2.00E-16	0.02085	2.00E-16	2.00E-16	2.00E-16	0.03349	2.00E-16	2.00E-16	<2e-16	0.0409	<2e-16	<2e-16
Rep2-HP β_6-β_1	2.20E-16	1.44E-11	2.20E-16	2.20E-16	2.20E-16	6.40E-10	2.20E-16	2.20E-16	2.20E-16	1.01E-09	2.20E-16	2.20E-16
Rep3-HP β_6-β_1	2.00E-16	0.01766	2.00E-16	2.00E-16	2.00E-16	0.02978	2.00E-16	2.00E-16	2.00E-16	0.03618	2.00E-16	2.00E-16

Table 3.8.1: *p-values* for the intercepts and slope from Newey-West estimators for the βθ as a function of time (500ns to 2.5μs) in HP ($\alpha_7\beta_7$) and the intermediate ($\alpha_6\beta_6$) simulations. The *p-values* of slopes or intercepts of HP that are insignificant are highlighted in blue.

Our null hypothesis is that there is no statistical difference in βθ values between the HP ($\alpha_7\beta_7$) and intermediates. The regression model is of the form $y = \beta 0 + \beta 1.\ X + B2.C + B3.C.X + \varepsilon$ where y is the βθ, β0 is the intercept for $\alpha_7\beta_7$, β1 is the slope of $\alpha_7\beta_7$, β2 is the intercept of intermediate, β3 is the slope of intermediate, ε is the error term, and C is the categorical variable. We have shown only a subset of statistical results for i.e., βθ for β_6-β_1 in $\alpha_6\beta_6$ and $\alpha_7\beta_6$ simulations in **Tables 3.8.1 and 3.8.2** and the entire set of statistical results are in **Fig. B.5**. The *p-values* with insignificant slopes are highlighted in blue.

	HP Intercept	HP Slope	Rep1-$\alpha_7\beta_6$ Intercept	Rep1-$\alpha_7\beta_6$ Slope	HP Intercept	HP Slope	Rep2-$\alpha_7\beta_6$ Intercept	Rep2-$\alpha_7\beta_6$ Slope	HP Intercept	HP Slope	Rep3-$\alpha_7\beta_6$ Intercept	Rep3-$\alpha_7\beta_6$ Slope
Rep1-HP β_6-β_1	2.00E-16	3.96E-02	2.00E-16	2.00E-16	2.00E-16	2.83E-02	2.00E-16	2.00E-16	2.20E-16	0.03832	2.20E-16	2.30E-10
Rep2-HP β_6-β_1	2.20E-16	1.94E-09	2.20E-16	2.20E-16	2.20E-16	1.86E-10	2.20E-16	2.20E-16	2.20E-16	1.46E-09	2.20E-16	5.47E-07
Rep3-HP β_6-β_1	2.20E-16	0.03534	2.20E-16	2.45E-11	2.20E-16	0.02487	2.20E-16	6.35E-10	2.00E-16	0.03403	2.00E-16	2.00E-16

Table 3.8.2: *p-values* for the intercepts and slope from Newey-West estimators for the βθ as a function of time (500ns to 2.5μs) in HP ($\alpha_7\beta_7$) and the intermediate ($\alpha_7\beta_6$) simulations. The *p-values* of slopes or intercepts of HP that are insignificant are highlighted in blue.

Initially, a significance of α=0.05 was used and later, we corrected it for multiple hypothesis tests, using the Bonferroni adjustment, which changes the form α from 0.05 to $1.95X10^{-4}$ ($\alpha_6\beta_6$ and $\alpha_7\beta_6$) where n (216) refers to the total number of statistical tests done. The *p-values* in **Table 3.8.1**. confirm that βθ distributions of β_6-$\beta_1$in $\alpha_6\beta_6$ simulations are significantly different from the βθ distributions of β_6-$\beta_1$in HP simulations. In **Tables 3.8.1**. and **Table 3.8.2** the *p-values* with the HP slope being insignificant (blue highlighted) indicate that the βθ in HPs will not change with time, and the simulation has reached its convergence. But

the βθ slope in $\alpha_6\beta_6$ and $\alpha_7\beta_6$ is significant in all replicates and will change with time, and hence we likely will observe more distortions to βθ. Overall, the statistical results conclude that besides a few exceptions (red highlighted *p-values* in **Tables B.5**), all the replicates of the intermediate simulations are significantly different in their βθ from the HPs βθ, which confirms that the βθ conformational shifts are significant.

3.3.8 $\alpha_6\beta_7$ shows subunits near missing alpha in a different conformation which destabilizes its structure.

In the $\alpha_6\beta_7$ MD simulations, we observed that the β_7 unit adopts a very different conformation, and it appeared as it shifted to a different plane. We have quantified the angle made by every β subunit with reference to the center of mass of the β ring (**Fig. B.3.2**). In the side view, we mainly saw β_6 and β_7 appear distorted. These distortions reflect the changes in the quaternary structure, in HP the interfaces between α and β subunits participate in non-covalent interactions and thus keep them in the right symmetry. Such noncovalent bonds are not possible if β_6 and β_7 are tilted. Though the bottom views of $\alpha_6\beta_7$ (**Fig. 3.9.2**) do not look strikingly different, the side views (**Fig. 3.9.1**) show the flip in β_7.

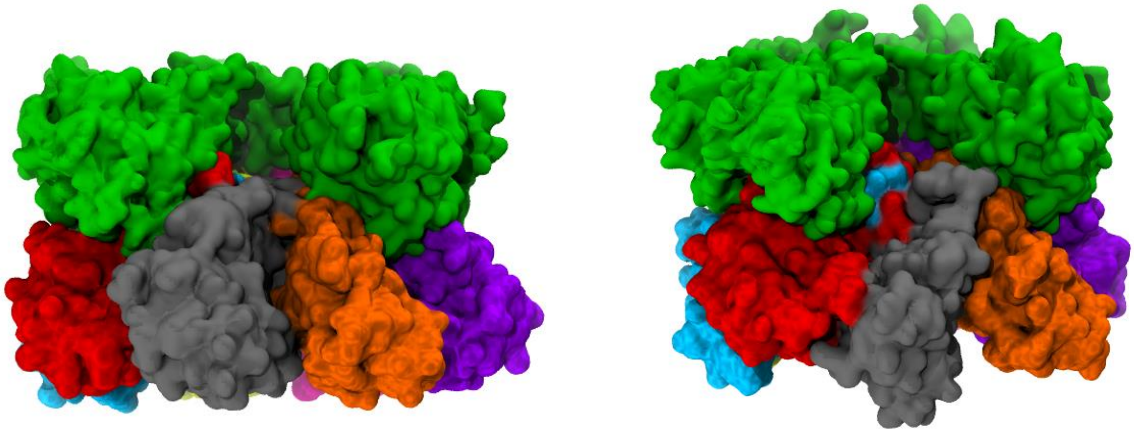


Figure 3.9.1: Side view of $\alpha_6\beta_7$ MD simulations at the beginning (0 seconds) and the end of 2μs shown in surface representations. The α subunits are in green, and the β subunits are in seven different colors.

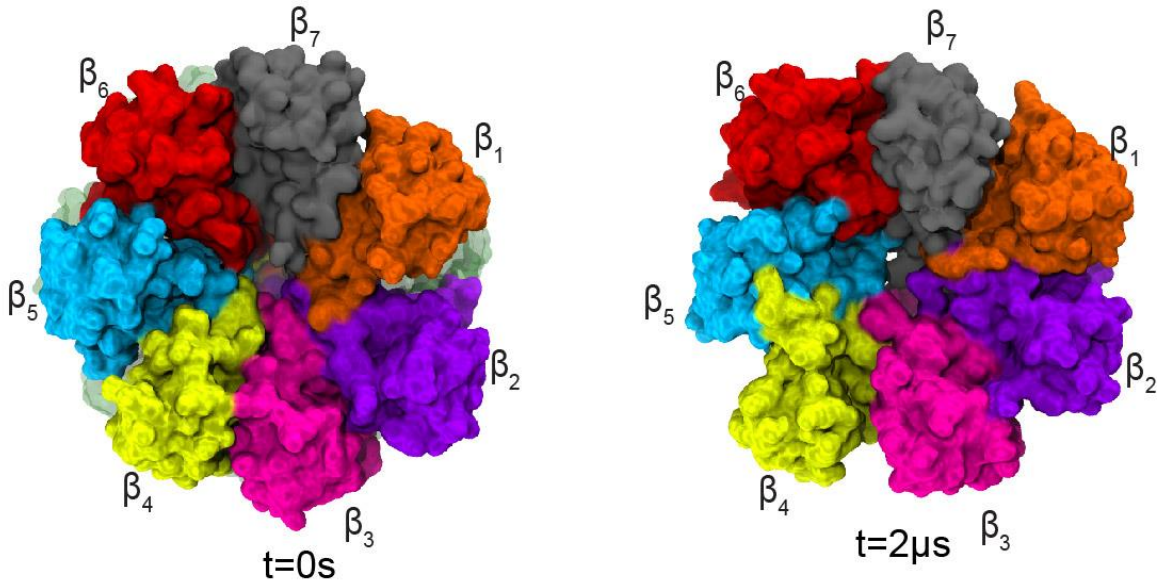
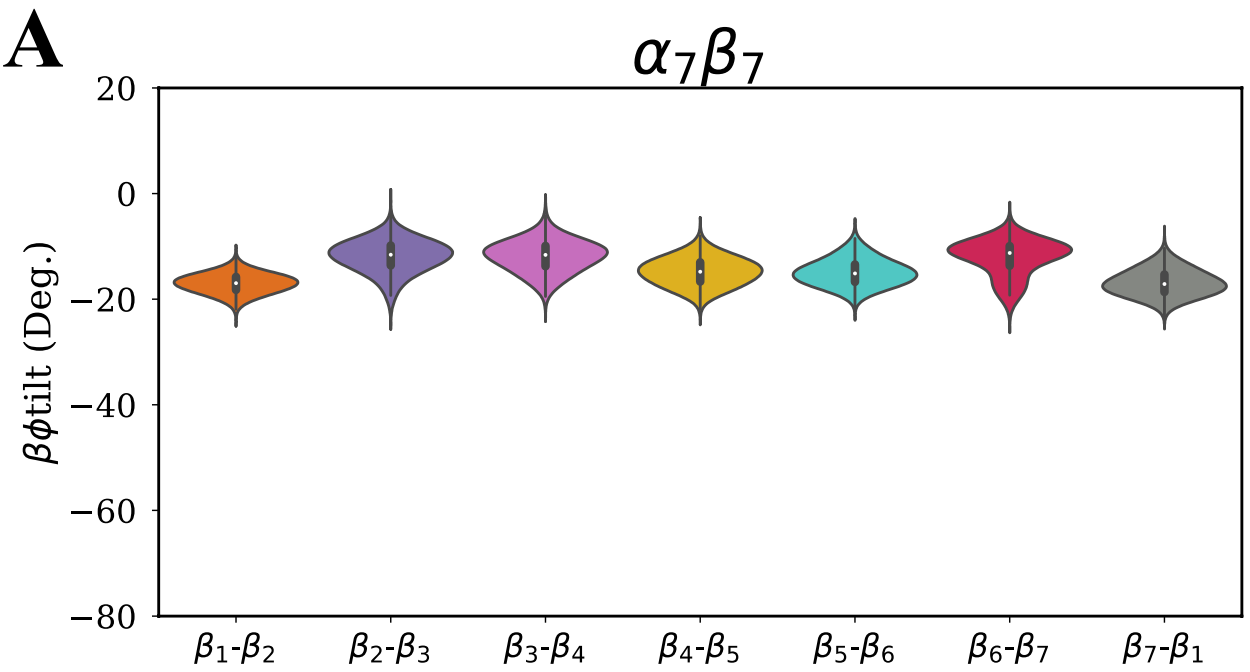
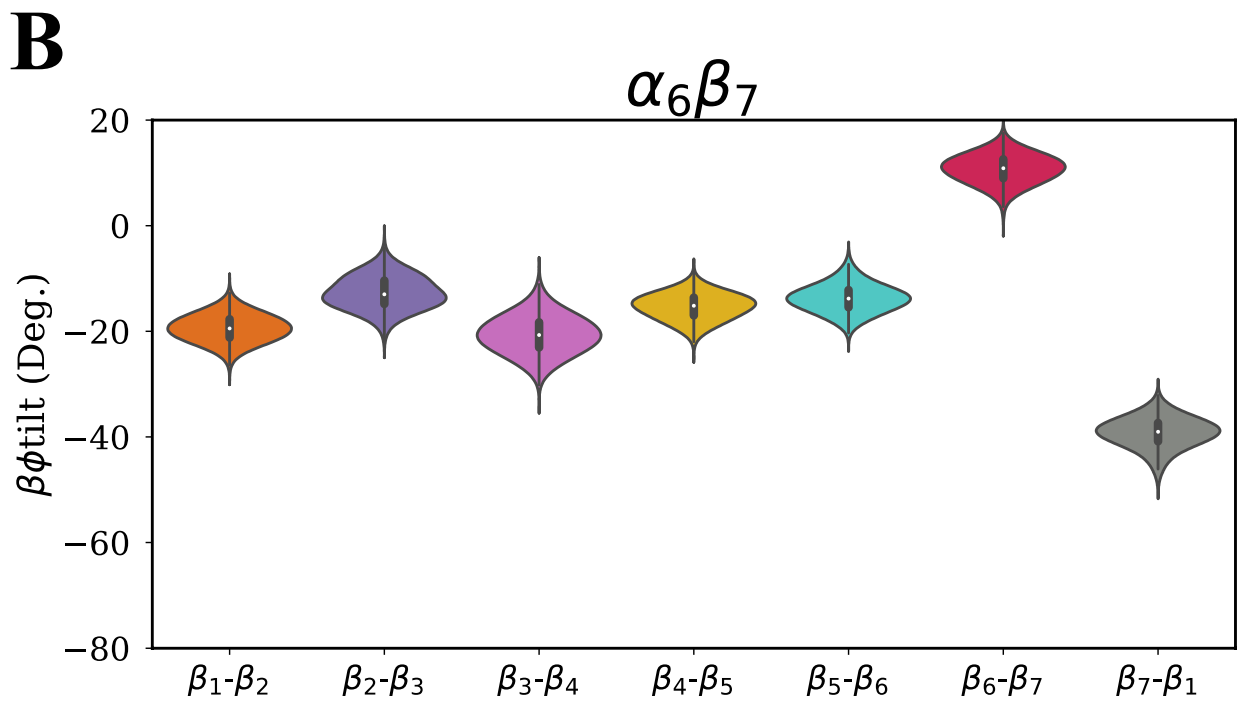


Figure 3.9.2: Bottom view of β subunits in $\alpha_6\beta_7$ MD simulations at the beginning (0 seconds) and the end of 2μs shown in surface representations. The α subunits are transparent (green), and the β subunits are in seven different colors as indicated.

3.3.9 $\alpha_6\beta_7$ shows a distorted structure due to change in the rotation of the β subunits (βØtilt)

Though in 3.3.4, we observed structural transitions in $\alpha_6\beta_6$ and $\alpha_7\beta_6$, which explains why they are unable to dimerize with an HP. By visualization, it appeared that $\alpha_6\beta_7$ has a different mechanism describing why it cannot dimerize. To investigate this underlying molecular mechanism, we used the torsional or dihedral angles (βθtilt) to calculate the rotation of the β subunits, as seen in **Fig. B.3.2**. The βØtilt angle in the crystal structure of *Re* is about $-19.1^{0,}$ and hence we expect the $\alpha_7\beta_7$ simulations to be closer to this number. As expected, all the β subunits in HP are approximately around 19.1^0 (**Fig. 3.10A**). In the $\alpha_6\beta_7$, we observed that the β_1, β_6, and β_7 are displaced from their positions compare to the WT HP (**Fig. 3.10B**) as the α ring is incomplete, causing several interfaces in the CP to have a weak affinity. As seen in **Figs. 3.9.1 and 3.9.2**, the β_6 and β_7 are tilted and rotated with reference to the β ring. Interestingly, we also observed that in $\alpha_6\beta_6$ and $\alpha_7\beta_6$, the distributions of βØtilt are noticeably different from the HP (**Fig. 3.10**). We expected that $\alpha_6\beta_7$ could not dimerize because its β ring is distorted, but $\alpha_6\beta_6$ and $\alpha_7\beta_6$ are also having rotational distortions along with the angular changes in the incomplete β ring. Therefore, we summarize that, on average, all the three intermediates differ from HP-$\alpha_7\beta_7$ conformationally and are in a non-dimerizable state.





To observe the βØtilt evolution as a function of time, we have time-series plots for the torsion angle calculated in (**Fig. B.6 A, B, C, and D**). Therefore, we summarize that, on average, all the three intermediates differ from HP-$\alpha_7\beta_7$ conformationally and are in a non-dimerizable state. To observe the

βØtilt evolution as a function of time, we have time-series plots for the torsion angle calculated in (**Fig. B.6 A, B, C, and D**)

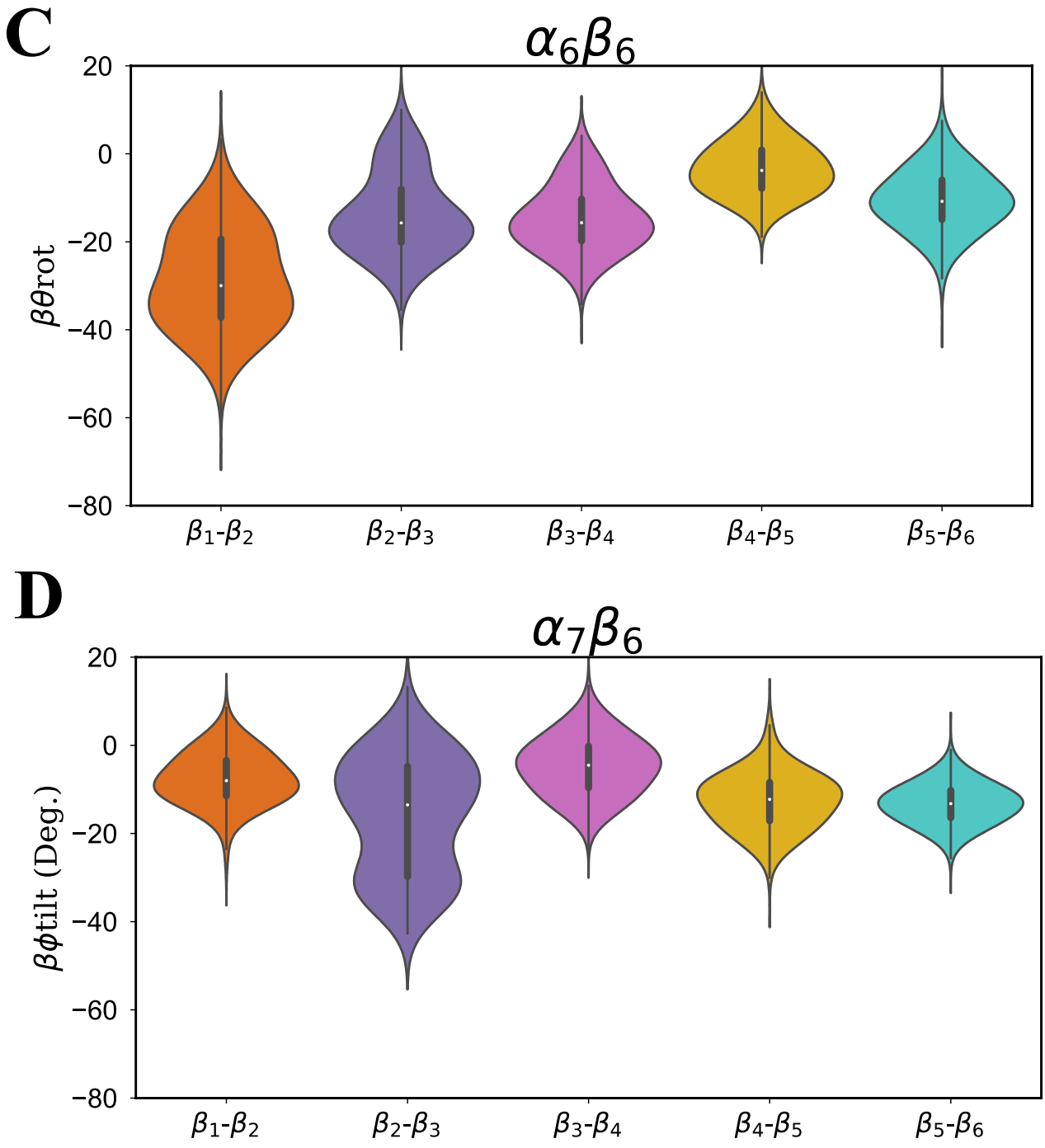


Figure 3.10: Violin distributions of βØtilt for all β subunits in different colors of the four sets of simulations. A) $\alpha_7\beta_7$ (HP) simulations and their distributions. B) $\alpha_6\beta_7$ C) $\alpha_6\beta_6$ D) $\alpha_7\beta_6$.

3.3.10 Statistical Analysis for βθtilt

On similar lines with βθ statistical analysis in section 3.3.7, we have performed Linear Regression using categorical variables for every subunit and replicate for βθtilt. Our null hypothesis states no statistical difference in βØtilt values between the HP ($\alpha_7\beta_7$) and $\alpha_6\beta_7$. The model is of the form $y = \beta 0 + \beta 1.\, X + B2.\, C + B3.\, C.\, X + \varepsilon$ where y is the βØtilt, $\beta 0$ is the $\alpha_7\beta_7$ intercept, $\beta 1$ is the slope of $\alpha_7\beta_7$, $\beta 2$ is the intercept of $\alpha_6\beta_7$, $\beta 3$ is the slope of intermediate, ε is the error term, and C is the categorical variable. We have shown only a subset of statistical results (βØtilt for subunits β_6 and β_7) in (**Table 3.11**), and the entire set of results are in **Tables B.7.**

	HP Intercept	HP Slope	Rep1-$\alpha_6\beta_7$ Intercept	Rep1-$\alpha_6\beta_7$ Slope	HP Intercept	HP Slope	Rep2-$\alpha_6\beta_7$ Intercept	Rep2-$\alpha_6\beta_7$ Slope	HP Intercept	HP Slope	Rep3-$\alpha_6\beta_7$ Intercept	Rep3-$\alpha_6\beta_7$ Slope
Rep1-HP β6	2.00E-16	2.00E-16	2.00E-16	0.01212	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
Rep1-HP β7	2.20E-16	2.20E-16	2.20E-16	5.44E-09	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
Rep2-HP β6	2.20E-16	2.20E-16	2.20E-16	1.36E-09	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
Rep2-HP β7	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.23E-14	2.20E-16	0.08289	2.20E-16	8.16E-14	2.20E-16	2.20E-16
Rep3-HP β6	4.26E-07	0.005332	2.20E-16	1.68E-12	3.95E-06	0.01099	2.20E-16	0.31489	6.45E-06	0.0129	2.20E-16	1.80E-08
Rep3-HP β7	0.3542	<2e-16	<2e-16	<2e-16	0.3783	2.20E-16	2.20E-16	3.49E-12	0.3853	<2e-16	<2e-16	<2e-16

Table 3.11: *p-values* for the intercepts and slope from Newey-West estimators for the βØtilt as a function of time (500ns to 2.5μs) in HP ($\alpha_7\beta_7$) and the $\alpha_6\beta_7$ intermediate simulations. This table has only a subset of *p-values* for β_6 and β_7. The *p-values* of insignificant slopes are in blue and insignificant intercepts are insignificant are in blue.

As seen in **Table 3.11**, we did not find any cases with insignificant *p*-values for both intercept and the slope. Specifically, we were looking for cases where both the intercept and slope are above a threshold of $> 2.31X10^{-4}$ (Bonferroni corrected for 256 tests from initial α=0.05). The *p-values* shaded in yellow indicates that the intercept is insignificant. We found only three cases where the HP simulations (replicate 3) has an insignificant intercept. The *p-values* shaded in blue indicates that the slope is insignificant. Here, we found three cases where the HP slopes are insignificant and three for $\alpha_6\beta_7$ βØtilt values. If the slope is insignificant then, it only indicates that there is no change in slope as a function of time indicates because these βØtilt values are converged. We did not find in this subset any β subunits in $\alpha_6\beta_7$ with an insignificant intercept, which depicts that these are having a significantly βØtilt when compared to the HP simulations.

Overall, for subunits β_6 and β_7 the βØtilt values in $\alpha_6\beta_7$ are significantly different in reference to the HP, and thus have a distinct rotation in β subunits.

3.4 Discussion

This study addresses a specific step of CP assembly in which two HP's associate to form a CP. All CP assembly studies to date have shown that two HPs assemble into a CP. However, we never observe a near-HP intermediate (HP with one or two missing subunits) associating with a "true HP" ($\alpha_7\beta_7$) and forming CP-like complexes ($\alpha_7\beta_7\beta_6\alpha_6$). It is completely unclear, however, how these intermediates are prevented from dimerizing. We used all-atom MD simulations to understand the molecular mechanism in *Re* CP assembly, which prevents such near-HP intermediates association. These near-HP intermediates do exist in solution and are likely to be present for non-trivial amounts of time. Their persistence can lead to kinetic traps reducing the CP assembly yields. If such CP-like structures become active, it can allow for unregulated protein degradation and disrupts homeostasis in cells. Therefore, likely an allosteric mechanism operates in cells that prevent such destabilizing intermediates association. Since the *Re* CP occurs without the use of any external factors or chaperones, the information for the allosteric mechanism must already be incorporated in the sequence and shape of α and β subunits.

We initially hypothesized that the β propeptide blocks association of near-HP intermediates with a true HP. Basically, the propeptide perturbs the critical interactions required for CP assembly (**Fig. 2.2**) and causes transitions making the intermediates non-dimerizable. Alternatively, we hypothesized that an alternative allosteric mechanism is possible, which causes conformational changes. To test our hypotheses, we have simulated three intermediates $\alpha_6\beta_7$, $\alpha_6\beta_6$, and $\alpha_7\beta_6$. Our results did not show any direct correlation between the β propeptide and non-dimerizable conformational states. Instead, we saw interesting global conformational shifts that happen in the subunits, which causes changes to the entire quaternary structure and rings geometry. In all the intermediate simulations, we observed distortion in the β ring symmetry and global conformational shifts. In the intermediates with incomplete β ring, we observed that the subunits

beside missing α or β collapse and became compact. Specifically, in $\alpha_6\beta_6$, its overall geometry changes significantly, as seen in **Fig. 3. 7C**-red violin, the angle-β in subunits $\beta_6 - \beta_1$, drastically shifts and has a distinct distribution compared to the $\beta_6 - \beta_1$ in HP simulations (**Fig. 3. 7C** - blue violin). Similarly, the $\alpha_7\beta_6$ intermediate, had several structural changes when visualized (Fig. 3.5). The angle in $\beta_6 - \beta_1$ of $\alpha_7\beta_6$ has a very different distribution from $\beta_6 - \beta_1$ of HP (**Fig.3.7D**- red and blue violins). The intermediate $\alpha_6\beta_7$ likely follows a different mechanism than $\alpha_6\beta_6$ or $\alpha_7\beta_6$. As seen in **Fig. 3.7B**, the βθ distributions are very similar to that of HP-$\alpha_7\beta_7$. This could also be because $\alpha_6\beta_7$ has the complete β ring and maintains a total angle of about 360_0. We observed that the β subunits near the missing α have a tilt and thus changing the overall dimerization interface shape. As seen in **Fig. 3.10 A,** and **B**, the $\alpha_6\beta_7$ has totally different distributions of subunits tilt.

Another consequence of the global confirmational shifts in the near-HPs, are the distortions in dimerization interface. These changes make dimerization interface incompatible and non-complementary. Thus, its highly probable, that the near- cannot make the crucial interactions with the key residues at the dimerization interface of opposing HP. In Chapter 2, I discussed in detail about these crucial interactions [24], and how small perturbations of these interactions effects CP assembly (**Fig. 2.2**). All the near-HP intermediates have many confirmational shifts, making them highly unlikely to be in the correct orientation for participating in the non-covalent interactions. Specifically, the key secondary structural elements at the HP dimerization interface, which include the S2-S3 loops and H3-H4 helices, would be in incorrect orientation to form the critical interactions that drive CP assembly [24]. Therefore, the shifts in angles-βθ, and subunit tilt-βØtilt acts as a cause the near-HP intermediates to be in a non-dimerizable states and such conformational shifts act as a checkpoint factor to prevent incorrect assemblies.

Our results have several implications. Firstly, the intrinsic β subunit angles in the intermediates, are highly different from the angles in HP. This difference likely stems from the evolution of frustrated interfaces in unbound protein structures (subunits). Upon binding, these protein subunits become less frustrated [7, 8]. For instance, as seen in the simulation snapshots, the structure of HP is very different from

$\alpha_6\beta_6$ (**Fig. 3.4**). Because of the two missing subunits, the entire $\alpha_6\beta_6$ structure has collapsed, and relaxes to a differed conformation. Currently our results are preliminary, but our findings of global conformational shifts suggests that the subunits are intrinsically frustrated.

These frustrations also have impacts in completing the assembly, because they can have kinetic and thermodynamic challenges to overcome. As seen in $\alpha_6\beta_6$ or $\alpha_6\beta_7$, the frustrated interfaces of the subunits have made the structures highly compact and distorted, such that the missing subunits cannot readily associate with the intermediates. From our findings, it's hard to explain how these intermediates can accommodate missing subunits to form HPs. The $\alpha_6\beta_6$ simulations show clearly, the collapse in structure and do not have any void for the missing α or β subunit to integrate into the structure.

With these simulations going forward we can use these intermediates structures to screen small-molecule inhibitors for assembly. As seen the interfaces are very different in intermediates when compared to the HPs or assembled structures. More importantly our work lays foundation to investigate if similar mechanisms are conserved in other hierarchical assemblies. Would such global confirmational shifts also occur in other non-proteasome assemblies?[26] Do other multi-subunit complexes have conformational shifts acting as assembly check points? Is this a conserved mechanism for preventing incorrect assemblies?

Ultimately, our work is the first molecular simulation study that shows evidence for allosteric communication among CP subunits. There have been no such molecular simulation studies on near-HP intermediates for any 20S CP. In the future, our results will aid in investigating specific interfaces for contacts and interactions that are unique in these intermediates. This specificity from different interfaces can be exploited to design effective and novel CP small-molecule assembly inhibitors for targeting specific assembly steps in the bacterial proteasomes.

[5, 23, 27-32]

3.5 References

1. Chapter 5 - Macromolecular Assembly. In: *Cell Biology (Third Edition)*. Edited by Pollard TD, Earnshaw WC, Lippincott-Schwartz J, Johnson GT: Elsevier; 2017: 63-74.
2. Livneh I, Cohen-Kaplan V, Cohen-Rosenzweig C, Avni N, Ciechanover A: The life cycle of the 26S proteasome: from birth, through regulation and function, and onto its death. *Cell Research* 2016, 26(8):869-885.
3. Sharon M, Witt S, Glasmacher E, Baumeister W, Robinson CV: Mass spectrometry reveals the missing links in the assembly pathway of the bacterial 20 S proteasome. *J Biol Chem* 2007, 282(25):18448-18457.
4. Kwon YD, Nagy I, Adams PD, Baumeister W, Jap BK: Crystal structures of the Rhodococcus proteasome with and without its pro-peptides: implications for the role of the pro-peptide in proteasome assembly. *J Mol Biol* 2004, 335(1):233-245.
5. Suppahia A IP, Burris A, Kim FMG, Vontz A, Kante A, Kim S, Im W, Deeds EJ, Roelofs J.: Cooperativity in Proteasome Core Particle Maturation. *iScience* 2020, 23(5).
6. Zuhl F, Seemuller E, Golbik R, Baumeister W: Dissecting the assembly pathway of the 20S proteasome. *FEBS Lett* 1997, 418(1-2):189-194.
7. Ferreiro DU, Hegler JA, Komives EA, Wolynes PG: Localizing frustration in native proteins and protein assemblies. *Proceedings of the National Academy of Sciences* 2007, 104(50):19819.
8. Ferreiro DU, Komives EA, Wolynes PG: Frustration in biomolecules. *Quarterly Reviews of Biophysics* 2014, 47(4):285-363.
9. Song Y DF, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D.: High-resolution comparative modeling with RosettaCM. *tructure* 2013 Oct8, 21,10:1735-1742.
10. Raman S VR, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, Baker D.: Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 2009, 77:89-99.
11. Jo S, Kim T, Iyer VG, Im W: CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem* 2008, 29(11):1859-1865.
12. Lee J, Cheng X, Swails JM, Yeom MS, Eastman PK, Lemkul JA, Wei S, Buckner J, Jeong JC, Qi Y *et al*: CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *Journal of Chemical Theory and Computation* 2016, 12(1):405-413.
13. Lee J, Hitzenberger M, Rieger M, Kern NR, Zacharias M, Im W: CHARMM-GUI supports the Amber force fields. *The Journal of Chemical Physics* 2020, 153(3):035103.
14. Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, Merz KM, Jr., Onufriev A, Simmerling C, Wang B, Woods RJ: The Amber biomolecular simulation programs. *Journal of computational chemistry* 2005, 26(16):1668-1688.
15. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC: Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* 2013, 9(9):3878-3888.
16. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S *et al*: All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 1998, 102(18):3586-3616.
17. al DESe: Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *IEEE* 2014:41–53.
18. Best RB, Zhu X, Shim J, Lopes PE, Mittal J, Feig M, Mackerell AD, Jr.: Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ(1) and χ(2) dihedral angles. *J Chem Theory Comput* 2012, 8(9):3257-3273.
19. Lippert RA, Predescu C, Ierardi DJ, Mackenzie KM, Eastwood MP, Dror RO, Shaw DE: Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure. *The Journal of Chemical Physics* 2013, 139(16):164106.

20. Martyna GJ, Tobias DJ, Klein ML: Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* 1994, 101(5):4177-4189.
21. Nosé S: A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics* 1984, 52(2):255-268.
22. Newey WK, West KD: A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 1987, 55(3):703-708.
23. António J. Marques RP, Ana C. Matias, Paula C. Ramos, and R. Jürgen Dohmen: Catalytic mechanism and assembly of the proteasome. *Chem Rev* 2009, 109(4):1509-1536.
24. Witt S, Kwon YD, Sharon M, Felderer K, Beuttler M, Robinson CV, Baumeister W, Jap BK: Proteasome assembly triggers a switch required for active-site maturation. *Structure* 2006, 14(7):1179-1188.
25. Team RDC: R: A language and environment for statistical computing. In. Vienna, Austria; 2010.
26. Levy ED, Erba EB, Robinson CV, Teichmann SA: Assembly reflects evolution of protein complexes. *Nature* 2008, 453(7199):1262-1265.
27. Campos LA, Sharma R, Alvira S, Ruiz FM, Ibarra-Molero B, Sadqi M, Alfonso C, Rivas G, Sanchez-Ruiz JM, Romero Garrido A *et al*: Engineering protein assemblies with allosteric control via monomer fold-switching. *Nature Communications* 2019, 10(1):5703.
28. McLeish T, Schaefer C, von der Heydt AC: The 'allosteron' model for entropic allostery of self-assembly. *Philos Trans R Soc Lond B Biol Sci* 2018, 373(1749).
29. Wodak SJ, Paci E, Dokholyan NV, Berezovsky IN, Horovitz A, Li J, Hilser VJ, Bahar I, Karanicolas J, Stock G *et al*: Allostery in Its Many Disguises: From Theory to Applications. *Structure* 2019, 27(4):566-578.
30. Zlotnick A, Mukhopadhyay S: Virus assembly, allostery and antivirals. *Trends Microbiol* 2011, 19(1):14-23.
31. Greco TM, Cristea IM: The Biochemical Evolution of Protein Complexes. *Trends in biochemical sciences* 2016, 41(1):4-6.
32. Deeds EJ, Bachman JA, Fontana W: Optimizing ring assembly reveals the strength of weak interactions. *Proc Natl Acad Sci U S A* 2012, 109(7):2348-2353.

Chapter 4

Kinetic Trapping and Robustness in Bacterial Core Particle Assembly

4.1 Introduction

Assemblies of proteins are involved in every major function of a cell. Cells require molecular machines like ATP synthase, the proteasome, ribosome, the apoptosome, nucleosome, and several other macromolecular complexes to maintain cellular homeostasis and carry out protein synthesis and protein breakdown. These large molecular machines are composed of many subunits, are highly coordinated, and have evolved to function in an efficient and organized manner. Cells cannot synthesize such molecular machines directly, but instead they are assembled from a set of subunits into a fully functional structure. A comprehensive understanding of the function of the molecular machines requires knowing the structure, kinetics, and energetics of its reaction intermediates [1]. To fully decipher the role and understand the importance of macromolecular complexes requires an understanding of these complexes' kinetics, thermodynamics, and assembly.

Assembly of molecular machines is often thought to be hierarchical and ordered. Though several theoretical studies on assembly assume that the assembly pathways need not be hierarchical. Understanding assembly pathways gives us insights into the evolutionary process that have favored some intermediates and prevented other off-pathway or kinetically trapped intermediates. Macromolecular assembly dynamics and kinetics often encounter situations where the subunits required for the final structure are exhausted and exist as incompatible intermediates that cannot interact to form the final complex. These incompatible intermediates are kinetically trapped and reduce assembly speed and waste energy. Studying assembly of complexes like CP and others can us give more insights into how these machines are built in cells. The

principles used in constructing these machines, especially self-assembling complexes like bacterial CP can serve as a proof of concept in building nanomachines, and nanoscale structures. Additionally, assembly studies can be applied to design highly specific, potent assembly inhibitors.

Protein degradation is tightly regulated as the misfolded or damaged proteins must be removed selectively from cells. Cells carry out intracellular protein degradation through molecular machines called proteasomes. These massive molecular machines are critical for a variety of cellular functions Proteasomes are found in archaea, bacteria, and eukaryotes. The 20S Core Particle (CP) forms the catalytic core, whereas the 19S Regulatory Particles (RP) caps the CP and is involved in regulatory functions [2]. Structural studies have shown that the 20S CP quaternary structure is highly conserved in all kingdoms of life [2]. The CP's show a characteristic pattern of four coaxially stacked heptameric rings of either seven α or seven β subunits in an $\alpha_7\beta_7\beta_7\alpha_7$ arrangement. The α subunits form the outer rings, and the β subunits are catalytically active and form the inner rings. There are substantial differences in the subunit complexity in prokaryotic and eukaryotic CPs. Prokaryotes CP assembly takes place spontaneously, but eukaryotic CP assembly is more complicated and requires dedicated chaperones and assembly factors to mediate CP formation [3]. In this study, we have demonstrated kinetic trapping computationally and experimentally in the macromolecular complex proteasome Core Particle.

Several studies have proposed two distinct pathways for prokaryotic CP assembly [4-6]. Assembly of CPs in archaea is thought to begin with the formation of the α ring first (ARF), and then the β subunits append to it (**Fig. 4.1**). On the other hand, in bacteria, seven αβ dimers (ABD) are thought to assemble to form a CP (**Fig. 4.1**). Both pathways are hierarchal, and therefore we propose another model of CP assembly pathway, which we call the unordered model (UOM). Where, the initial nucleation step involves the formation of a αβ dimer, but then all possible association are possible up to the formation of a HP (**Fig. 4.1**). This UOM is not investigated by any other studies on CP assembly. To further understand these pathway assembly dynamics, we need to develop models so that we can have defined parameters that govern the assemblies. Previous work has demonstrated that ring-like structures assemblies are susceptible

to “deadlock” during the assembly process [7]. Such studies will provide insights into the evolutionary process have guided the assembly of ring-like structures such as the CP.

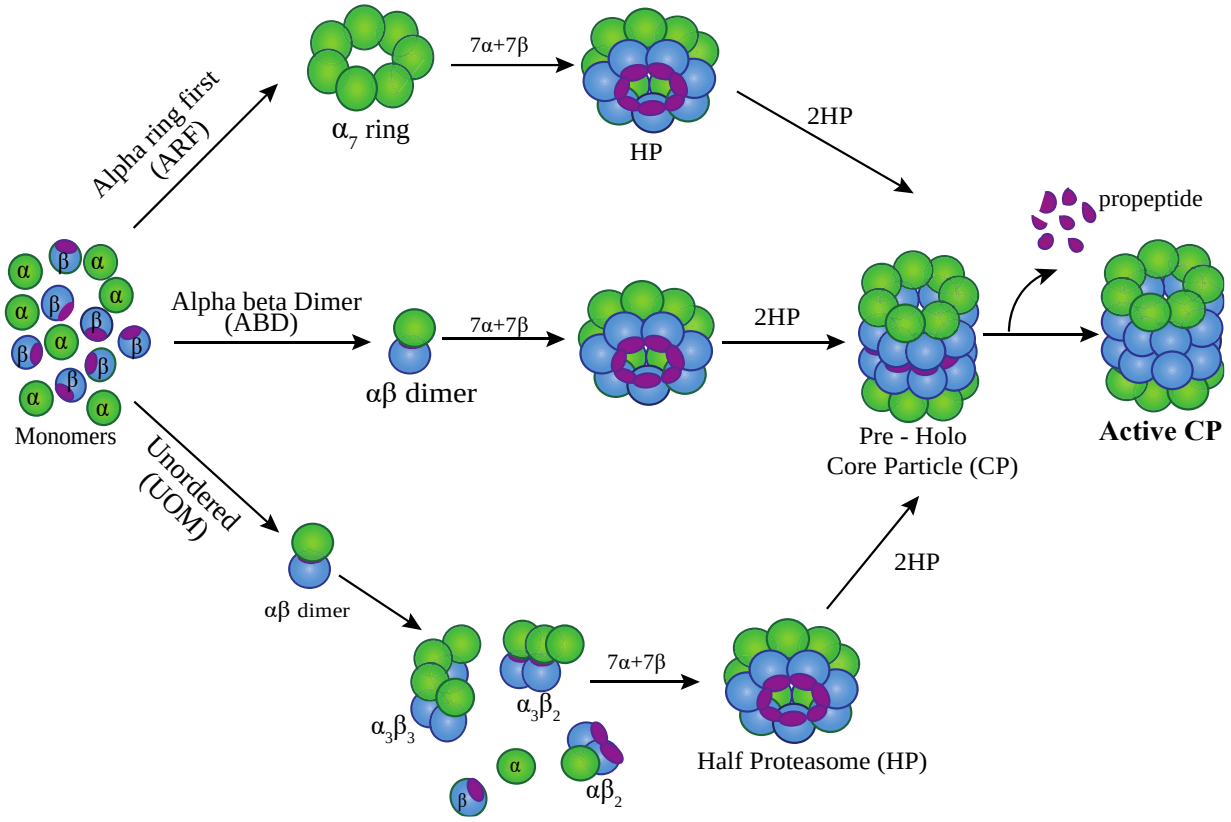


Figure 4.1: Schematic of the three CP assembly pathways. Alpha Ring First (ARF) is known to occur in archaea, Alpha Beta Dimer (ABD) is known to occur in bacteria, and the Unordered Model (UOM) of assembly is the new pathway we propose in this study. All the α subunits are in green, β is in blue and the propeptide is in purple. Irrespective of the pathways involved in formation of HP, all CPs are formed only by association of two HPs.

In this work, we have chosen the actinomycete bacterium *Rhodococcus erythropolis* (*Re*) as our model system to study CP assembly. The *Re* CP assembly is well studied, and its subunits can be purified separately as monomers, and they spontaneously self-assemble into active CPs *in vitro* [4, 8]. We developed mathematical models to understand the assembly dynamics of homomeric stacked rings like proteasome CP. We also validated our model findings by *in vitro* experiments in *Re*. Our results indicate that bacterial assembly is affected by kinetic trapping, and there is a tradeoff between speed and robustness in CP assembly pathways. We also observed that kinetic trapping is dependent on serval parameters like

interaction affinities, rate constants. Also, our models indicate that in bacterial CP assembly, speed is more crucial than robustness for both *in vivo* and *in vitro* scenarios. Additionally, the studies present the assembly dynamics and kinetics of the newly proposed CP assembly pathway – UOM.

4.2 Materials and Methods

4.2.1 Mathematical models

The mathematical framework used for developing the ODE models is largely based on the ring assembly dynamics and is described in the main and the Supplemental text of a previous study on ring assembly [7]. For CP, the modeling approach begins with generating different species required for CP assembly. The molecular species can be generated from size 1 to size 28 (four stacked rings). The simulation methodology is based on "Chemical Reaction Network" (CRN) approaches and inspired by rule-based approaches based on molecular graphs used by several groups [7, 9, 10]. The species in CP assembly are presented used a bitwise representation, of 0's (absence) and 1's (presence) as a formal notation to represent all possible intermediates in the assembly process as described by other works [7, 11, 12].

We have three CP assembly pathways -ARF, ABD, and UOM, constructed differently due to their hierarchy. The number of species for each pathway differs, in ARF 27 different species can be formed, which includes the two monomers and seven different ways of forming a seven-membered alpha ring. Then the remaining species include the different ways seven beta rings can append to this alpha ring. For the ABD model, 7 specie are possible which reflect the seven dimers assembling into an HP. Lastly, the UOM allows many different on-pathway intermediates, and hence we have 875 species in total. Next, the reactions are enumerated for each model; there are few rules, including ring-sidedness and no clashes. Considering these, we will have 14795 reactions for ARF, 13 reactions for ABD, and 18149 reactions for UOM. Next, a system of ODE is derived, which describes the ODE's time evolution of the concentration of any intermediate and for the four stacked rings. The Odes are integrated using CVODE libraries, and the Backward Differential Methods [13, 14].

Assembly of CPs is majority carried by cells that constantly produces monomers to balance the decrease in the concentration of intermediates and fully assembled complexes that happened due to cell growth, dilution, and protein degradation. To explore these *in vivo* conditions, we incorporated the synthesis and degradation processes in our models using two terms – rate of synthesis (Q) and rate of degradation (δ). Thus, the concentration of subunits (C) depends on both these rates as $C=Q/\delta$. These models are described in detail in a different study [7].

4.2.2 *In vitro* native gel assembly experiments

The α and β subunits were expressed in *E.coli* and purified using affinity chromatography and ion-exchange chromatography (Details of all experimental protocols may be found in supporting information). Dynamic light scattering was used to ensure the subunits were in a monomeric state. The α and β subunits were concentrated to roughly 35uM using Amicon Ultra 10K centrifugal filters (Millipore) and diluted to 0.5uM, 1uM, 2uM, 4uM, 8uM, 16uM, and 32uM in an assembly buffer (HNE). Both subunits were then mixed in equal volumes to obtain a final subunit concentration of 0.25uM, 0.5uM, 1uM, 2uM, 4uM, 8uM, and 16uM. For each concentration, the subunits were mixed in equimolar ratios. The assembly reaction was allowed to proceed for 24hours at 30°C. An equal volume of the assembly reactions was mixed with an equal volume of loading dye (0.8M HEPES,0.1% Bromophenol Blue, 20% Glycerol). These samples were then loaded on a 4-20% native gel (Invitrogen). The gels were run at four °C and 120 V for 12 hours. Gels were stained with Sypro Ruby protein stain, visualized using Licor Odyssey Fc imager, and quantified using ImageStudio Lite software.

The α and β subunits were transferred into a working buffer (HNE 20mM HEPES, 100mM NaCl,1mM EDTA, 5mM DTT PH 7.0) and concentrated using Amicon Ultra 10K centrifugal filters (Millipore). Protein concentrations were estimated to be roughly 35uM by measuring absorbance at 280nm and using the molar extinction coefficient of 16390 for α and 17880 for β. These stocks were then diluted to 0.5uM, 1uM, 2uM, 4uM, 8uM, 16uM, and 32uM. Both subunits and mixed in equal volumes to obtain a final subunit

concentration of 0.25uM, 0.5uM, 1uM, 2uM, 4uM, 8uM, and 16uM. The assembly reaction was allowed to proceed for 24hours at 30°C. An equal volume of the assembly reactions was mixed with an equal volume of loading dye (0.8M HEPES,0.1% Bromophenol Blue, 20% Glycerol). These samples were then loaded on a 4-20% native gel (Invitrogen). The gels were run at four °C and 120 V for 12 hours (**Fig. C.5**).

4.3 Results

4.3.1 Homomeric trimeric and trimeric stacked rings assembly

In three-membered homomeric (trimeric) rings, when all the interaction affinities (binding affinities/ strengths) between the subunits are equal, and at a fixed concentration of the subunits, we observe the existence of the plateau or "deadlock" for the trimeric rings after a certain amount of time [7] (**Fig. 4.2**). The depletion of monomers and dimers contributes to the formation of trimeric rings. But, after a certain time, as the formation of trimers reaches the plateau, and the system reaches a kinetically trapped state or “deadlock”. This happens because all the intermediates remaining are not compatible with forming a

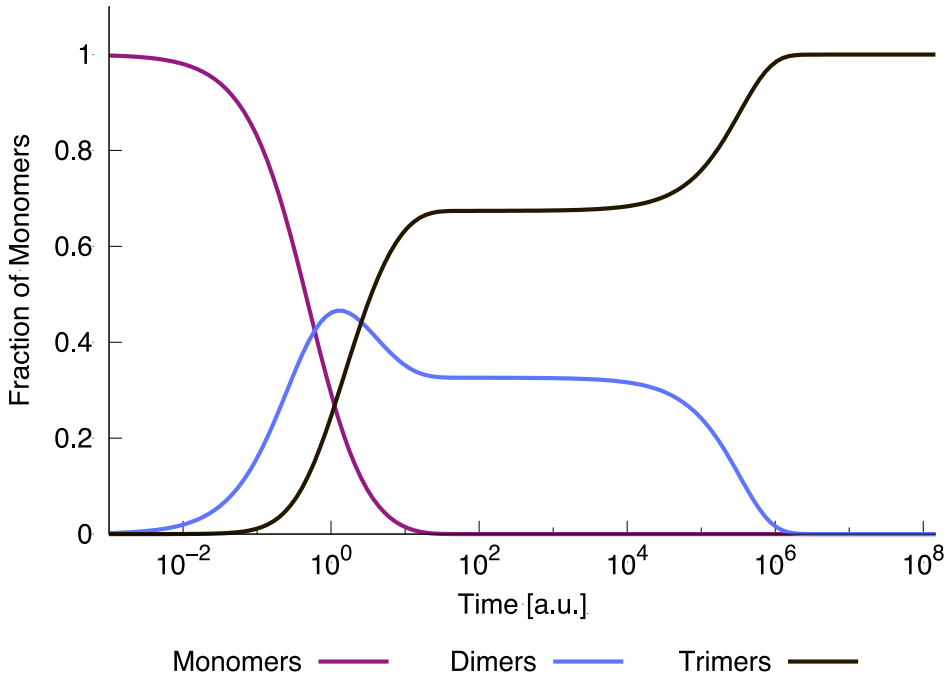


Figure 4.2: The assembly dynamics information of a trimeric homomeric ring. Depletion of monomers (purple) results in the formation of dimers (blue) and trimers (black). The plateau seen in trimers is the deadlock phase, where the intermediates are kinetically trapped.

trimeric ring and are kinetically trapped. So, for instance, we can have intermediates like two dimers that cannot interact to form a homomeric trimeric ring. This deadlock is alleviated only after a certain time when the dimers begin to dissociate. The compatible intermediates (monomers and dimers) are again available, which can form the trimeric ring. The existence and duration of the plateau depends on several parameters of the model [7].

This model is based on the chemical kinetics approach, and here we first enumerate the species (intermediates) involved in the assembly of a trimeric ring. Next, we enumerate all the reactions involved in the assembly, in other words two monomers reacting with each other to form a dimer, a monomer and a dimer reacting to form a trimer. In the model, we also have few rules which dictate the dynamics of ring assembly. For instance, we consider a "sidedness," which incorporates the asymmetric aspects of the subunits [7]. As observed in nature, all known subunits or monomers in ring-like protein structures are internally asymmetric. To account for asymmetry in our model (see C.1), binding is permitted only if interactions can occur from the right side of one subunit with the left side of the other subunit (**Fig. C.1**). Also, we do not allow steric clashes in the model, such as two dimers interacting to form a tetramer; such reactions cannot occur. After considering these rules, we define a system of Ordinary Differential Equations (ODEs) used to determine the species and reactions and further solve each species time evolution based on the law of mass action [15].

For heteromeric rings, the assembly situation differs because we can vary the interaction affinities independently for every subunit. As discussed in pioneering study [7], Deeds. et al. examined the influence of assembly dynamics by changing the interaction affinities of a heteromeric trimeric ring. They found that at least one weak interaction (“single weak interaction” strategy) provides robust and efficient ring assembly with maximal yield. Ring-like structures are very thermodynamically stable and thus dominate at equilibrium, we can expect that for stacked ring structures like our model – the proteasome CP- the assembly dynamics and deadlock formation can be quite dramatic [7]. We began by investigating the assembly dynamics of stacked rings by considering a simpler system, i.e., a trimeric stacked ring; as

described earlier in this model, we enumerate all the species and reactions and followed by numerically integrating the ODE's using CVODE libraries.

We found key differences in the assembly of trimeric rings and trimeric stacked rings. Firstly, the deadlock plateau for trimeric stacked rings reduced assembly efficiency more dramatically, and the deadlock plateau exists for much longer times (**Fig. 4.3**). This is mainly because the intermediates, which are kinetically trapped in the formation of a trimeric stacked ring, are also ring-like structures. For example, we can have tetrameric stacked rings highly stable, and these intermediates do not dissociate on biologically relevant time scales. Therefore, the deadlock in stacked rings exists for an indefinite period.

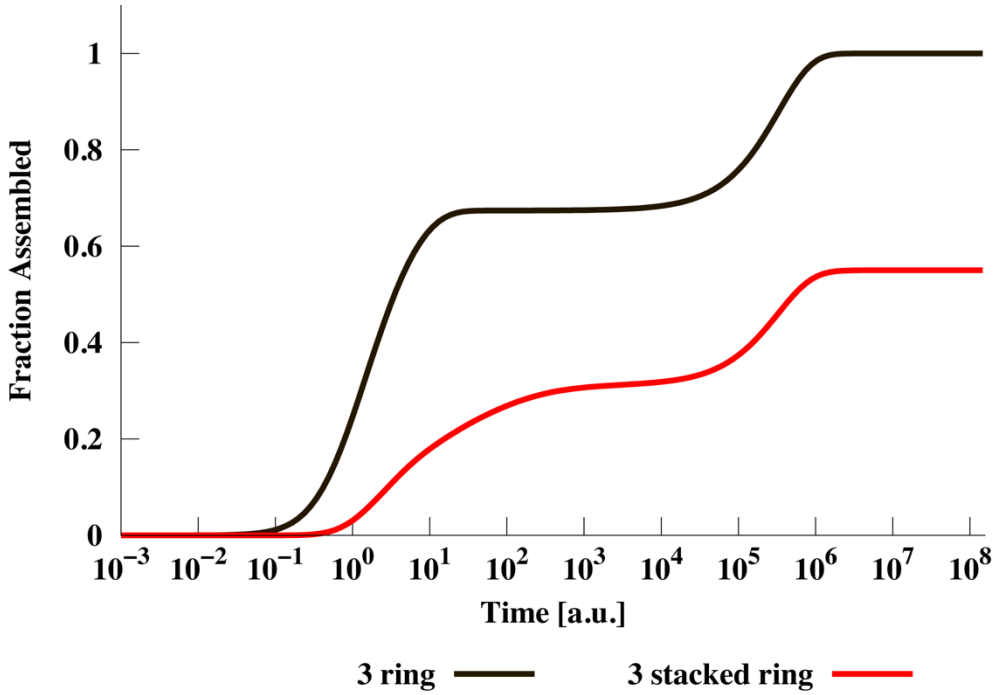


Figure 4.3: Assembly dynamics for a three-membered ring and three-membered stacked ring as a function of time. The red curve depicts the formation of the three-membered ring, and the black curve for three-membered stacked rings. The plateau phase (middle) in both the curves is the "deadlock."

The deadlock plateau observed in these rings is also dependent on the concentration of the subunits, and the assembly dynamics are different for rings and stacked rings. When we compare the fraction assembled in a single ring and stacked ring as a function of the concentration of subunits, we observe the impact of longer deadlocks (> 24hrs) in stacked rings (**Fig. 4.4**). With increasing subunit concentration, the assembly

efficiency increases up to a certain concentration (1 μM in this example). After the maximum assembly efficiency is reached, increasing concentration; causes increase in kinetically trapped intermediates and thus deadlock to decrease the assembly efficiency. The concentration needed to achieve maximum yield and the reduction in assembly efficiency depend on several parameters like interaction affinities and rate constants. But increasing the subunit concentration after a certain limit does not affect assembly efficiency because the system is now dependent on the dissociation of the kinetically trapped intermediates. These intermediates are very stable and hence do not dissociate readily on biologically relevant timescales.

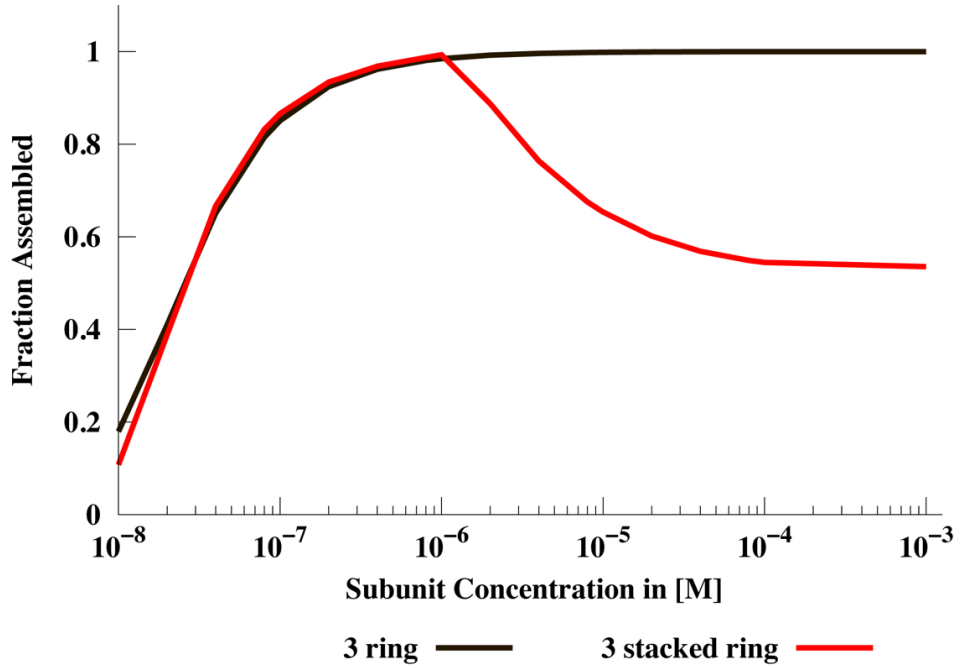


Figure 4.4: Fraction of assembled three-membered single and three-membered stacked ring after 24 hours as a function of subunit concentration. This model shows the appearance of deadlock after the concentration increases 1uM in the three membered stacked rings.

This highlights that in stacked rings, the highly stable kinetically trapped intermediates can cause a detrimental impact, especially if they are part of biological molecules. For ring-like protein complexes, the deadlock effect can disturb the cells homeostasis and unfruitful investment of ATP towards the synthesis of subunits. In this work, we have bacterial CP as our model system, and the ODE models for these are more complex than those for simple rings; hence we incorporated additional parameters based on other macromolecular simulation studies and experimental evidence.

4.3.2 ODE models for investigating bacterial CP assembly dynamics

To further understand the impact of deadlock in larger rings and protein complexes, we employed in-house ODE models as several studies have proposed that the CP assembly pathway is Alpha Ring First (ARF) or Alpha Beta Dimer (ABD) (**Fig. 4.1**). It is thought that the archaea follow ARF, and bacteria follow the ABD pathway. So far, no concrete study demonstrates that bacteria follow only the ABD pathway for CP assembly. Both pathways are hierarchical; the ARF requires the formation of a seven-membered α ring and then proceeds to form HP's. The ABD pathway requires seven α-β dimers and then forms the HP's and CP. In this work, we propose a new pathway- the "Unordered Model" (UOM), which is non-hierarchical. Thus, the intermediates can be any of the possible species up to the formation of an HP (**Fig. 4.1**). As previously described the, HP is an obligatory intermediate, and irrespective of the pathway, CP assembly occurs from the dimerization of two HP's and not by other intermediate species. Experimental evidence suggests that dimerization of HP takes hours to finish, but HP formation occurs almost immediately after the reaction begins (Chapters 2 and 3). The scope of this work is focused on understanding the assembly pathway of HP, which eventually assembles into a CP.

We have several ODE assembly models based on simulations for complex macromolecular assemblies and experimental studies on bacterial CP's. For all the three models of CP assembly, we have three primary parameters, interaction affinity (K_D) (binding affinity or binding strength between two interfaces, **see C.2**) and association rates for two HP's (k_{+HP}), and association rates for subunits (k_+). The other parameters include subunit concentration, simulation time, and error tolerance. The interaction affinities (K_D), which represent how strong can two interfaces associate, we obtain this information approximately from the surface contact area between interfaces. As a convention, we have defined six unique protein-protein interfaces in a CP. We primarily refer to the *Rhodococcus erythroplis* (*Re*) CP interfaces in this study (**Fig. C.2**). Out of these six interfaces, IN5 and IN6 are involved in HP dimerization, and the other four interfaces are a part of HP formation. Thus, interfaces IN5 and IN6 are not included in our models since they do not contribute to the assembly dynamics but are a part of HP dimerization.

The interfaces (1- 4) are our regions of interest; IN1 is the interface between two alpha subunits, IN2 is the interface between two beta subunits, IN3 and IN4 are the interfaces between alpha and beta subunits (**Fig. C.2**). These binding affinities are connected to the contact surface area buried between the subunits involved in the interface. *Re* Crystal structure studies suggest that IN4 is the strongest in binding affinity and IN1 for *Thermoplasma. acidophilum* (*Ta*), which is thought to follow the ARF pathway [16]. As seen in the assembly models (**Fig. 4.1**), ARF pathways needs first the alpha ring to be formed, and this requires a strong affinity to begin the assembly reactions, thus IN1 interaction affinity i.e., K_{D1} is critical for ARF. In ABD and UOM, we focus on IN4 (largest interface for alpha beta dimer) as the formation of αβ dimers is a nucleating step, so thus IN4 interaction affinity i.e., K_{D4} is critical. The other three remaining interfaces K_D's are set to a lower affinity (10^{-2} M). The other parameter k_+ which represents the association rate of subunits in most cells is about 10^5 $M^{-1}s^{-1}$to 10^7 $M^{-1}s^{-1}$ [17].

4.3.3 Deadlock formation is varying in the three models

We selected a range of commonly used concentrations in CP assembly experiments and similarly chose the K_D's of interfaces based on modeling approaches performed by several groups [7, 12, 18, 19]. The HP dimerization rate (k_{+HP}) was approximated based on the time course assembly experiments (Chapter 2). Next, we simulated the three assembly models with this set of parameters and observed interesting differences in assembly dynamics.

For all the three models (**Fig. 4.5**), similar what we see in the assembly of stacked rings (**Fig. 4.2**), increasing subunit concentrations, gives higher assembly efficiency, and then the yield gets reduced due to deadlock. Interestingly, the extent of deadlock varies among the three models, with ARF showing no deadlock even at higher concentrations (Fig. 4.4). Hence, the assembly in ARF is very robust and non-susceptible to kinetic trapping. In ABD, deadlock reduces the assembly efficiency by 20-25%, as the kinetically trapped intermediates are stable and do not dissociate on these timescales. The effect of kinetic trapping is seen by the decrease in assembly efficiency. In the UOM a similar trend occurs, but this model is most susceptible to deadlock, where the assembly efficiency after 24 hours is reduced by ~ 50%. The

kinetically tapped intermediates are abundant in UOM since it allows any many possible combinations of on-pathways intermediates, hence the proportion of kinetically trapped intermediates is higher.

The strength of K_D used in **Fig. 4.5** is independently varied for each model. It is tuned such that the maximum assembly efficiency occurs at 1 μM (used for assembly experiments), and the parameter used to show the assembly dynamics as seen in Fig. 4.4 are described in C.3. The interaction affinity plays a critical role in achieving maximum yield at different concentrations. Depending on the assembly model, the K_D modulates the maximum yield, since a stronger K_D will yield maximum assembly efficiency at lower concentrations and weaker K_D will yield maximum yield at a higher concentration of subunits (**Fig. 4.5 and C.3**).

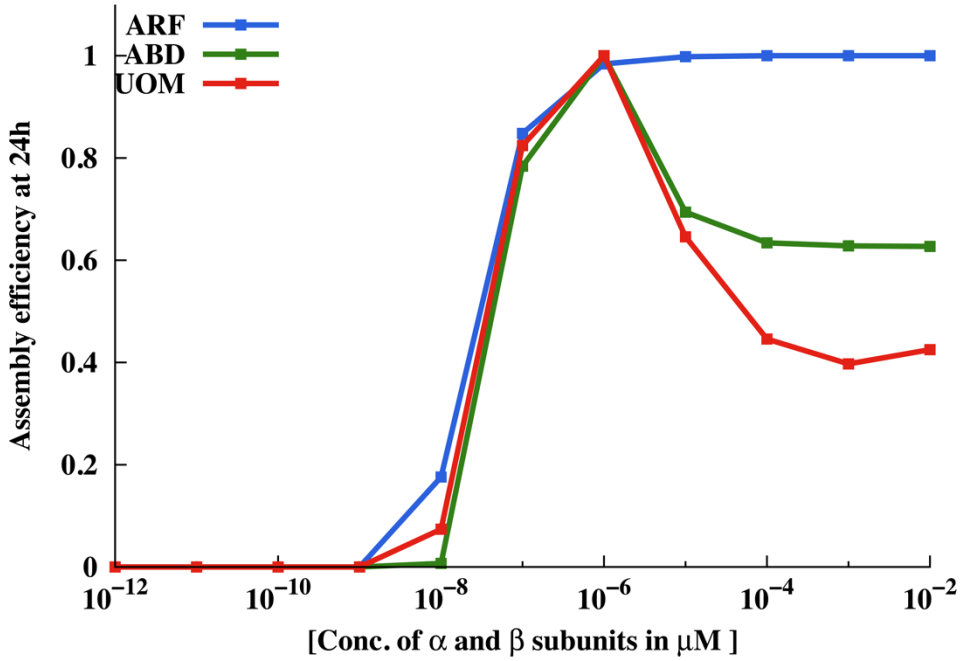


Figure 4.5: CP assembly at varying concentrations with [α] = [β]. The three models exhibit different fraction CP assembled (or assembly efficiency) and show a varying deadlocked plateau. The affinities and association parameters were chosen such that they have a maximum assembly at 10^{-6} M. The interaction affinities and the association rates for each model are in (see C.3) The x-axis is in log scale.

4.3.4 *In vitro* CP assembly dynamics in *Rhodococcus erythropolis* (*Re*).

The above discussion focuses on the ODE models, which suggest kinetic trapping in CP assembly. We further wanted to experimentally investigate kinetic trapping, and the type of CP assembly pathway which

operates in our model system *Re*. This bacterium has been well characterized experimentally and has a simple subunit composition compared to the complex eukaryotic CPs [3]. In the bacteria *Re*, subunits stay monomeric until mixed. This finding led to the hypothesis that in bacteria, α and β dimers are formed, which further assemble into an HP. As we had seen in the CP assembly models described in the previous section, we hypothesized that CP assembly is also susceptible to kinetic trapping.

To investigate if kinetic trapping occurs *in vitro*, a set of assembly experiments was performed. As described (**Fig. C.5**), we conducted a set of native gel CP assembly experiments. Interestingly, we observed a decrease in assembly efficiency as the α and β subunit concentrations were increased. After a particular concentration (1μM), we found that the assembly efficiency starts decreasing. This finding confirmed our hypothesis of kinetic trapping in proteasome CP *in-vitro* assembly (**Fig. 4.5**). This result suggests that, like the ODE models, *in vitro* assembly of CP is initially slower because the time it takes to for subunits to associate at low concentration, followed by the formation of kinetically trapped intermediates at higher subunit concentration.

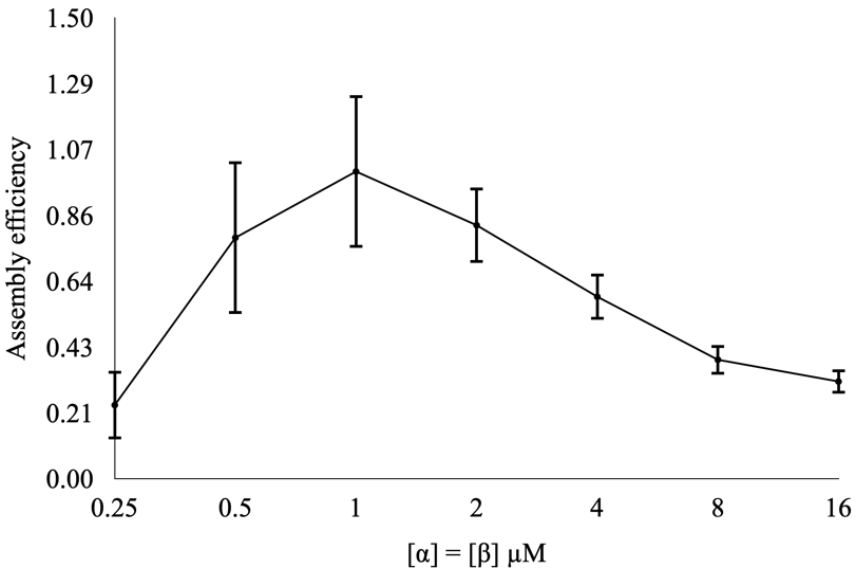


Figure 4.6: Quantification of Native-PAGE gels.
Gels were stained with Sypro Ruby protein stain and imaged using a Licor Odyssey Fc imager. Bands were quantified using the ImageStudio Lite software. Band intensities are considered proportional to CP concentration. Assembly efficiency was determined by dividing the band intensity with individual subunit concentration and normalized to the highest value.

4.3.5 Investigating *Re* CP assembly pathway from *in-vitro* experiments

Based on the visual comparison of the *in vitro* results of CP assembly in *Re*, we speculated that it follows the UOM of assembly. This is surprising, as it was previously thought that CP assembly occurs through the ABD pathway in bacteria. To further confirm our hypothesis, we performed fits of the assembly models with the experimental results. For this, we varied three parameters, $K_{D,}$ the strongest interaction affinity between protein subunits, k_{+}, and k_{+HP}. The three parameters were varied over a range of physiologically relevant values (C.6). We used the Root Mean Square Error (RMSE) statistic to fit the assembly efficiencies over a range of concentrations (0.5 μM to 16 μM) to that of the three assembly models.

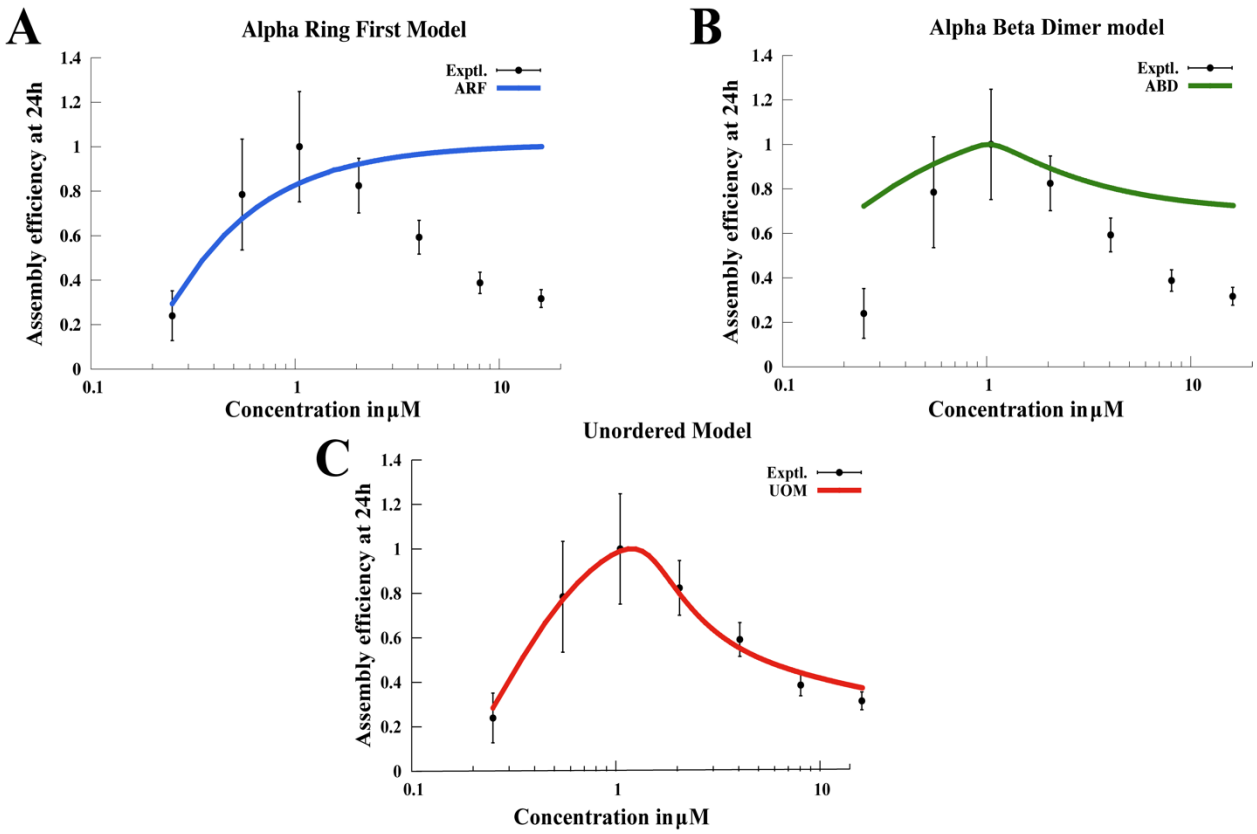


Figure 4.7: The ODE models and the *in vitro* experimental fits for the three CP assembly pathways. A) Alpha Ring First, the model is shown in blue curve, B) Alpha Beta Dimer the model is shown in green C) Unordered model the model is shown in red. UOM model has the best fit to the experimental data indicating that *Rhodococcus erythropolis* likely follows Unordered Model of CP assembly *in vitro*. All the parameters for the fits are given in C.6.

CP assembly in *Re* appears significantly different from the ARF model predictions, as we could not find a good set of parameters that yielded a perfect fit (**Fig. 4.7A**). While the ABD model provided a better fit

to the data, the fit was far from perfect, particularly at lower and higher subunit concentrations (**Fig. 4.7B**). Though the ABD pathway is thought to operate in bacteria, there is no concrete evidence for this finding. Our fits indicate that ABD pathway of CP assembly likely does not occur in *Re*. Despite the ABD pathways being widely accepted in field for bacterial system, it does not fit well with experimental data (**Fig. 4.7B**). For the UOM model, we got excellent fits with the experiments **(Fig. 4.7C)**. This finding is interesting since the UOM is susceptible to kinetic trapping and the deadlock reduces the assembly efficiency and it is the least robust. This leads to the question if a fast and non-hierarchical assembly pathway has some evolutionary advantages for the evolutionary advantages for the cells.

4.3.6 CP assembly models display a tradeoff between speed and robustness

The influence of deadlock on different assembly pathways (**Fig. 4.5**) indicated that more hierarchical pathways are less susceptible to deadlock. But we know from our previous observations on stacked rings that kinetically trapped intermediates do not resolve biologically relevant time scales. To further elucidate the assembly dynamics and understand how the assembly yields change over time, we examined the above simulation parameters differently. All the above simulations in **section 4.3.2** have the same concentration of subunits but independent interaction affinities (K_D) for each pathway. Each pathway has only one KD stronger and the remaining three interfaces K_D relatively weaker. To understand the differences in assembly kinetics of the CP in each model, we simulated having the same K_D's and association rate for all pathways and at the same subunit concentration. We observed that UOM was the fastest (**Fig. 4.8**), then ABD and ARF.

In the ARF pathway for CP assembly, a seven-membered α ring must be formed first, and only then can β subunits bind; since the α ring takes time to fully to form, CP assembly is slower. In the ABD pathway, αβ dimers must form first, and then other dimers can associate to form stable structures and finally assemble into the CP (**Fig. 4.1**). In the UOM, only the first nucleation step requires an αβ dimer, but after that any possible intermediate can associate with the dimer and quickly form an HP and then dimerize into CP. This non-hierarchy also makes this model prone to deadlock because there are more possible combinations to

generate incompatible intermediates. Therefore, though UOM is less robust, it is very fast in forming CP's, and it's followed by ABD and ARF (**Fig. 4.8**), indicating a tradeoff between speed and robustness in the CP assembly dynamics. Also, if we increase subunit concentration > 10^{-6} M (**Fig. C.4A, B, and C**), we notice the appearance of deadlock which reduces the assembly yield in UOM and ABD, as previously seen in **Fig. 4.3**. This indicates that UOM is fastest, but not robust, therefore there is a tradeoff between speed and robustness in these assembly pathways.

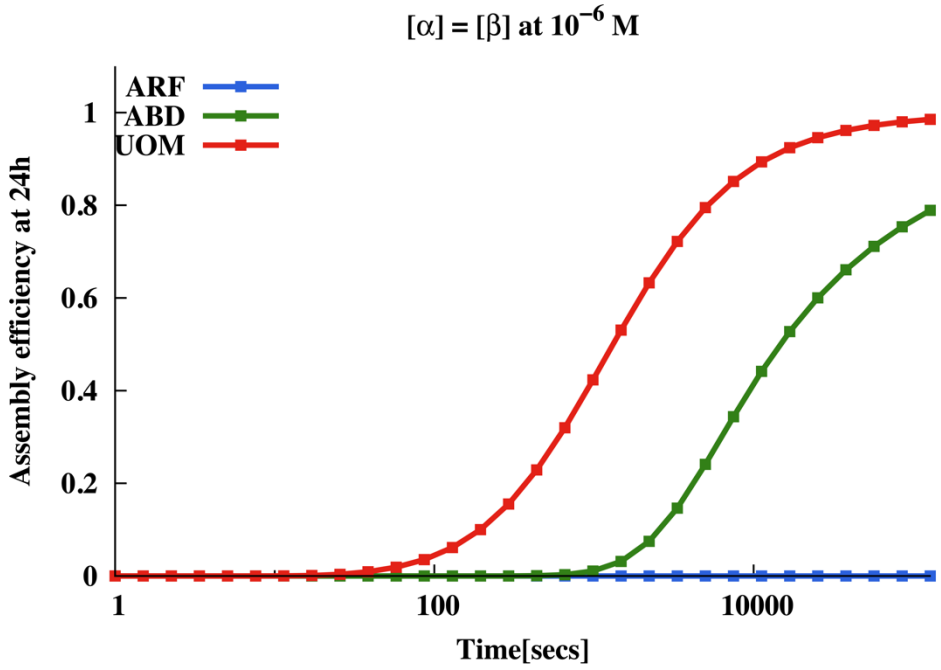


Figure 4.8: CP assembly time courses of the three models ARF, ABD, and UOM. At concentration [α] = [β] = 10^{-6} M. The time evolution clearly shows that UOM is the fastest. The detailed parameters are described in section C.4.

4.3.7 CP assembly dynamics *in vivo*

The above findings describe a situation that begins with a certain fixed concentration of subunits. The cells scenario is entirely different from this, where ring-like structures are constantly synthesized and lost. Complexes like proteasomes are actively being lost due to dilution, degradation, cell growth, and division [7]. This represents typical "*in vivo*" conditions, where the subunits are constantly synthesized and

degraded. To explore the "*in vivo*" CP assembly dynamics, we included synthesis and degradation rates for the three CP assembly models.

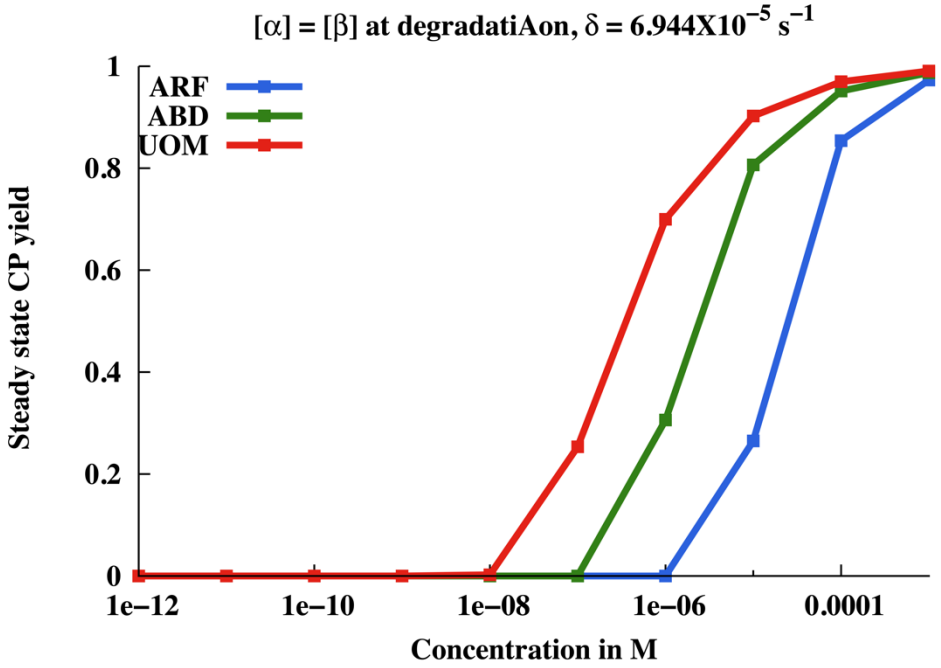


Figure 4.9: Steady state Core Particle assembly yield in the presence of synthesis and degradation for *Re*. The steady state yield for the three models with varying synthesis and degradation rates are shown as a function of concentration of monomer. The degradation rates (δ) correspond to the doubling time of the bacteria and the α, β monomers synthesis rate (Q) corresponds to the concentration of α and β monomers. The synthesis rate at various concentrations is calculated by CT = Q / δ. The parameters in this case are degradation rate δ = 5.55X10-4s^{-1} and monomers synthesis rate Q and CT varies from 10^{-2}M to 10^{-8}M. This degradation rate corresponds to the average doubling time of *R. erythropolis* bacteria. The steady state yield plots for slower degradation rates are discussed in C.7.

To incorporate degradation rate (δ), we considered the value of 6.944 X10^{-5} $s^{-1,}$ which are chosen to represent the half-life of proteins in *Re* (doubling time is ~ 4hours) cells. The slower the doubling time of the species, the slower is the effective degradation rate. In the above simulations for three assembly models, we observed that in the UOM, the steady-state CP yield (fraction of monomers in CP) reaches a maximum at lower concentrations (**Fig. 4.9**). In comparison, ABD and ARF need a much higher concentration of subunits than UOM. Like what we observed in the "*in vitro*" CP assembly models (**Fig. 4.9**), we found similar assembly dynamics in the "*in vivo*" models, where UOM is fastest and is followed by ABD, and ARF. These findings suggest that for hierarchical pathways like ABD and ARF, CP assembly takes a longer

time as the intermediate species require more time to associate and dissociate to form an HP. This happens because the longer the intermediate persists, the greater the probability it will be degraded, and the more intermediates are degraded the less they contribute to CP formation, and thus lower is the assembly efficiency. Our findings for the *in vivo* scenario suggest that CP assembly in cells has evolved to be fast and optimize resource utilization faster than hierarchy and robustness to deadlock.

4.4 Discussion

Macromolecular machines are large complexes involved in integral biological functions like signal transduction, cell motility, protein synthesis and degradation and hence critically influence many cell processes. These machines are assembled from a set of subunits or monomers into the final functional form. The assembly pathways of these complexes depend on the order and rates at which the subunits bind. All self-assembly pathways depend on diffusion-driven random collisions among the subunits in cells [15]. The widespread notion in the field is that all macromolecular complexes have evolved to assemble in an ordered and hierarchical manner. This requires the assembly to occur in a series of steps, for example in archaeal CP assembly, in the Alpha Ring First (ARF) assembly operates where, the seven-membered α ring must be formed before the β subunits can bind to it, and only then is an HP is formed, and only then can two HPs assemble into a CP. Simulations of such hierarchal pathways, have shown the existence of kinetically trapped intermediates. Theoretical studies have shown that incompatible intermediates can dominate the assembly dynamics. These kinetically trapped structures reduce the assembly yield (number of functional structures formed given a fixed concentration of subunits). This reduced assembly yield is also known as a deadlock. Naturally we can expect that, deadlock can be highly disadvantageous to the cell, and evolution would have designed some ways to overcome this kinetic challenge. In this work, we used a combination of theoretical models and in vitro experiments to understand the kinetic trapping in bacterial proteasomes.

For the first time, this study demonstrates that kinetic trapping exists *in vitro*, and the impact of kinetic trapping depends on the assembly pathways. The ARF and ABD assembly pathways are hypothesized to

exist in archaea and bacteria, respectively. Here, we have proposed another new pathway which is non-hierarchical, called the Unordered Model. Assembly kinetics from the ODE models for these three CP assembly pathways indicate presence of deadlock at higher subunit concentrations. As the α and β subunit concentration increases, the intermediates in UOM and ABD get kinetically trapped, and hence further increase in subunit concentration does not increase CP yield. To further understand the assembly dynamics, we have chosen *Re* CP as the model system. *In vitro* studies on the *Re* CP, the subunits stay monomeric when expressed and purified separately in *E. coli*, and on mixing, they spontaneously assemble into full CP. Interestingly, in experiments, we saw that as the concentration increased, the assembly yield started to reduce (**Fig. 4.3**). This decline is likely due to the kinetic trapping, as demonstrated in our ODE assembly models.

We compared the *in vitro* assembly kinetics with the ODE models and performed an RMSE fit to examine which pathways of CP assembly likely operate in *Re*. This analysis suggests that *Re*, most likely follows the UOM of CP assembly. This indicates that speed is more crucial than hierarchy, and the cells require faster assembly than being more robust. Most cells have a continuous synthesis and degradation of proteins; to explore this, we modified our models to include synthesis and degradation rates. The degradation rate δ is approximate to the inverse of the doubling time of the *Re* bacteria, and thus we have a $\delta=6.944X10^{-5}$ s^{-1} for our models. These models or "*in vivo*" scenario also indicates that UOM has higher steady state yield than ABD and ARF. Also, with decreasing degradation rate (C.7), we obtain higher yields at lower concentrations, showing that the cells have enough time to alleviate deadlock and get maximum steady-state CP yield. In both *in vitro* and *in vivo* scenarios, the UOM gives the maximum CP yield and is likely an evolutionary selection for cells, since the UOM pathway is shown to be the fastest as demonstrated in our models and experiments. Our results suggest that the bacterial CP assembly process selects speed of assembly over robustness to kinetic trapping.

This work employs a coarse-grained approach using ODE models to help us understand CP assembly pathways and assembly kinetics for bacterial proteasomes. Further, we can use such models to study CP

assembly in other bacteria and archaeal CPs to examine if there is a similar tradeoff between speed and robustness. Experiments like mass-spectrometry and Cryo-EM can help us further characterize the kinetically trapped intermediates that we observe *in-vitro* and the simulations. Additionally, we can also examine if any specific intermediates dominate in the solution. For the ODE simulations we used a constant association rate ($k_+=10^6\,M^{-1}s^{-1}$), that is often found in most protein-protein association reactions [15]. We would need further detailed mathematical models to change these rates as a function of the intermediates size or stoichiometry and then observe the changes to assembly dynamics and kinetics. These types of coarse-grained ODE models also can be developed for other self-assembly complexes like virus capsids, AAA ATPases, etc., Our self-assembly studies can inform for future designing of nanomachines and nanotechnology applications in pharmaceuticals (drug delivery) and semiconductor industry (photolithography). Additionally, identifying stable and transient intermediates will also help us to obtain new drug targets for CP assembly.

4.5 References

1. Bruce A: Molecular biology of the cell: Second edition. New York : Garland Pub., [1989] ©1989; 1989.
2. Lowe J, Stock D, Jap B, Zwickl P, Baumeister W, Huber R: Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 A resolution. *Science* 1995, 268(5210):533-539.
3. António J. Marques RP, Ana C. Matias, Paula C. Ramos, and R. Jürgen Dohmen: Catalytic mechanism and assembly of the proteasome. *Chem Rev* 2009, 109(4):1509-1536.
4. Sharon M, Witt S, Glasmacher E, Baumeister W, Robinson CV: Mass spectrometry reveals the missing links in the assembly pathway of the bacterial 20 S proteasome. *J Biol Chem* 2007, 282(25):18448-18457.
5. Panfair D, Ramamurthy A, Kusmierczyk AR: Alpha-ring Independent Assembly of the 20S Proteasome. *Sci Rep* 2015, 5:13130.
6. Tamura T, Nagy I, Lupas A, Lottspeich F, Cejka Z, Schoofs G, Tanaka K, De Mot R, Baumeister W: The first characterization of a eubacterial proteasome: the 20S complex of Rhodococcus. *Curr Biol* 1995, 5(7):766-774.
7. Deeds EJ, Bachman JA, Fontana W: Optimizing ring assembly reveals the strength of weak interactions. *Proc Natl Acad Sci U S A* 2012, 109(7):2348-2353.
8. Suppahia A IP, Burris A, Kim FMG, Vontz A, Kante A, Kim S, Im W, Deeds EJ, Roelofs J.: Cooperativity in Proteasome Core Particle Maturation. *iScience* 2020, 23(5).
9. Danos V, Laneve C: Formal molecular biology. *Theoretical Computer Science* 2004, 325(1):69-110.
10. Feret J, Danos V, Krivine J, Harmer R, Fontana W: Internal coarse-graining of molecular systems. *Proc Natl Acad Sci U S A* 2009, 106(16):6453-6458.
11. Suderman R, Deeds EJ: Machines vs. ensembles: effective MAPK signaling through heterogeneous sets of protein complexes. *PLoS Comput Biol* 2013, 9(10):e1003278.

12. Saiz L, Vilar JM: Stochastic dynamics of macromolecular-assembly networks. *Mol Syst Biol* 2006, 2:2006.0024.
13. Hindmarsh AC, Brown PN, Grant KE, Lee SL, Serban R, Shumaker DE, Woodward CS: SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. *ACM Trans Math Softw* 2005, 31(3):363–396.
14. Cohen SD, Hindmarsh AC, Dubois PF: CVODE, A Stiff/Nonstiff ODE Solver in C. *Computers in Physics* 1996, 10(2):138-143.
15. Voit EO, Martens HA, Omholt SW: 150 years of the mass action law. *PLoS computational biology* 2015, 11(1):e1004012-e1004012.
16. Kwon YD, Nagy I, Adams PD, Baumeister W, Jap BK: Crystal structures of the Rhodococcus proteasome with and without its pro-peptides: implications for the role of the pro-peptide in proteasome assembly. *J Mol Biol* 2004, 335(1):233-245.
17. Chapter 5 - Macromolecular Assembly. In: *Cell Biology (Third Edition)*. Edited by Pollard TD, Earnshaw WC, Lippincott-Schwartz J, Johnson GT: Elsevier; 2017: 63-74.
18. Bray D, Lay S: Computer-based analysis of the binding steps in protein complex formation. *Proceedings of the National Academy of Sciences* 1997, 94(25):13493.
19. Sweeney B, Zhang T, Schwartz R: Exploring the parameter space of complex self-assembly through virus capsid models. *Biophys J* 2008, 94(3):772-783.

Chapter 5

Comparative Characterization of Crofelemer Drug Mixtures Using Machine Learning and Data Mining Approaches

5.1 Introduction

Natural products and their derivatives have played a significant role in drug discovery. Over the past decades, drug discovery mainly relied on synthetic compounds as the source, which are comparingly easy to produce, supply, and test. However, there has been a lot of scientific interest in obtaining drugs from natural products as they provide a lot of structural diversity compared to standard combinatorial chemistry. Archaeological evidence suggests that the use of medicinal plants for therapeutic purposes began thousands of years ago [1]. In 1827, the plant alkaloid morphine was the first commercial natural product drug, introduced by Merck [1]. Over the years, the term "biologics" has been used to describe drugs whose source material comes from microorganisms or living systems. Biologics have revolutionized the treatment of severe diseases and illnesses. Drugs like penicillin, Humira, Taxol, Avastin, and monoclonal antibodies are a few of the drugs that have been developed from natural resources.

The FDA has shown interest in the past few decades in developing botanical drugs. Currently, there are two FDA-approved botanical drugs Veregen (sine catechins) and Mytesi (Crofelemer) [2]. Botanical drugs are mixtures of several chemical entities and are highly heterogeneous due to their natural origin and often lack distinct active ingredients [3]. Due to the heterogeneous nature of raw material, there is a critical need

to ensure that these drugs are consistent across batches in effectiveness, safety, and potency. This work presents a Comparative Characterization study of the botanical drug Crofelemer (Mytesi™, Fulyzaq™).

Crofelemer was approved in 2006 by the FDA to treat noninfectious diarrhea in HIV patients undergoing antiretroviral therapy [4-6]. More than a quarter of HIV-infected patients or these therapies reported suffering from diarrhea, impacting their health [2]. This drug is administered orally and works by inhibiting chloride ion channels (the CFTR, the CaCC, and the CLC-2 channels) in the gastrointestinal tract [7, 8]. This drug is extracted from the sap of the South American tree *Croton lechleri*, commonly known as "dragon's blood", due to its deep red color. The Crofelemer is a complex mixture of procyanidins and prodelphinidins forming oligomers of 5-11 linearly covalently linked monomers in varying ratios [2]. This heterogeneous mix makes Croefelmer a complex drug product and complicates chemical and physical characterizations [4]. There can also be changes to the manufacturing process during the approval process or post-commercialization in a drug's lifecycle. This highlights the company's need for a critical assessment, which can show that the changes in the manufacturing process or raw material do not affect the safety, purity, and efficacy of the approved drug. Thus, any process changes need to be supported by substantial data that show any manufacturing process changes do not have any adverse effects and result in highly similar quality attributes. Many post-drug approval changes like the manufacturing process can lead to detectable analytical differences compared to the original (reference) product. From a regulatory perspective, the reference and consequently produced biologic drugs must be highly similar in efficacy and safety. These kinds of comparison studies are called comparability assessments and biosimilarity studies [9, 10].

In this work, we performed a comparability assessment of different batches and mixtures, which assess the ability of analytical techniques to capture batch-to-batch variability in Crofelemer (CF). The mathematical model approach used is similar to a previous study, on Comparative Characterization of Crofelemer stability studies and is published in three companion articles [4, 11, 12]. For ease, this CF stabilities study will be referred to as "Year 1 CF samples". For the stability studies, a single CF lot [4] was obtained from the manufacturer, and then "virtual" lots were created through a combination of techniques

like dialysis, fractionation by, centrifugation, and forced degradation studies [4, 12]. In total, the Year 1 CF study had 35 distinct samples, for two different temperatures. These samples were subjected to physical and chemical assays. The resulting data was analyzed using data mining and machine learning approaches to identify the analytical differences in treated CF materials [11].

Using a similar mathematical model and data mining approaches, we are characterizing CF mixtures in this Chapter. Many analyses and techniques are adapted from previous studies done in the group [11, 13]. The CF mixtures used in this study will be referred to as "Year2 CF samples", which focuses more on differentiating mixtures and identifying batch-to-batch variability. For this, three different batches or "pure lots" were obtained, and then six different mixtures were generated by combining different proportions of the three pure lots (**Fig. 5.1A**). To further prepare the samples, these CF lots were subjected to filtration via centrifugation to obtain three fractions (10kDTop, 10kD bottom, and unfractionated) (**Fig. 5.1B**). In total, we had 27 distinct samples, each with 3-4 replicates.

A

Mixture / Crofelemer	P1	P2	P3	M1	M2	M3	M4	M5	M6
	Percentage (%) of Pills								
Lot 1	100			10	10	33.3			
Lot 2		100		90		33.3	50	10	90
Lot 3			100		90	33.3	50	90	10

B

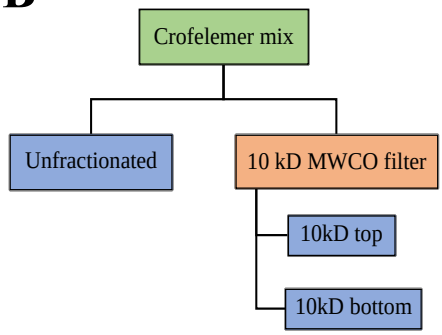


Figure 5.1: The mixtures and fractions used for data collection. A) Table showing the percentages of three pure lots (batches) and six different mixtures (generated from different combinations of pure lots). B) The centrifugal filters used for collection of fractions.

Traditional analysis of results from physicochemical techniques can easily miss identifying important qualities, called “critical quality attributes” (CQAs). Computational methods like machine learning, information theory, statistics, and data processing are essential to identify CQAs that can differentiate between the batches or lots. Aside from identifying CQAs, the computational approach used here can also be applied to the study of biologics and to address the issue of biosimilarity. These tools are highly useful

for evaluating large data sets like CF and similar biologics. The results obtained can help identify batches with contaminants or other variants that might not have the same efficacy as the original approved drug or reference drug. Hence, the combination of computational analysis and physiochemical characterization can help maintain regulatory standards and provide a more structured way to evaluate changes in the drug manufacturing process.

This work aims to distinguish CF mixtures from one another, and we achieved this by using a combination of mutual information (MI), principal component analysis (PCA), similarity analysis, and machine learning classifiers from the scikit learn package in Python [14]. Our study distinguished expired CF lots from valid lots and mixtures, and we also simulated a biosimilarity experiment, where our mathematical model successfully identified active vs. inactive Croefelmer groups. Ultimately, this approach lays a foundation for future works on assessing biologics and addressing issues of biosimilarity. Additionally, it highlights the need for more robust machine learning-based approaches to analyze the results of physical, chemical, and biological assays in pharmaceutical characterization.

5.2 Materials and Methods

5.2.1 Sample preparation

For this study, the Croefelmer mixtures were prepared similarly as described in the articles [4, 12]. The further details of mixtures and sample preparation are not described in this chapter and are out of scope. The data was provided by our collaborators using similar methods as published in previous studies [4, 11, 12]. As described in a similar study, the Fulzaq tablets were obtained in three batches [11]. The filtrate after centrifugation was categorized into two fractions, the 10-kDa top fraction, which contains molecules with Molecular Weight greater than 10 kDa, and the 10kDa bottom fractions, which contains molecules less than 10k Da. The third fraction was the unfractionated CF drug. The mixtures were prepared by mixing different proportions of three pure lots (batches) in various combinations or rations (Fig. 5.1). The table shown in

Fig. 5.1 has the three pure lots (first three mixtures) and then mixtures (4 - 9) with different combinations. In total, we have nine mixtures, each with three fractions and each fraction having four replicates; precisely, we have = 9(mixtures) X 3 (fractions) = 27 distinct samples.

5.2.2 Physical assays datasets preprocessing

A wide range of techniques like Ultraviolet-visible absorption spectroscopy (UV-Vis), Fourier transform infrared spectroscopy (FTIR), circular dichroism (CD), Nuclear magnetic spectroscopy (NMR), normal phase high-pressure liquid chromatography (HPLC), and other HPLC techniques like size-exclusion chromatography (SEC) and hydrophilic interaction chromatography (HILIC) were employed to characterize the CF mixtures and generate the datasets used for our analysis. (All of the experimental data for this study are done by our collaborators).

The data for UV-Vis's spectroscopy is recorded from 190 nm to 1100 nm. The FTIR data measures absorbances from 900 cm^{-1} to 4000 cm^{-1}. The CD measures the absorbances from 200 to 250nm. For the normal phase HPLC, we have absorbances at 280 nm and retention times 0 to 65 minutes (at 640-millisecond intervals). For SEC we had absorbances collected from 190 nm to 800 nm with the retention times from 0 to 30 minutes (at 640 milliseconds interval). Similarly, for HILIC, we have absorbances collected from 190 nm to 800 nm with retention times from 0 to 50 minutes (at 640 milliseconds interval). For CD, HILIC, and SEC, the experiment was also run with buffer (without any sample treatment). We subtracted these background correction values from the raw data before normalizing them by concentration. To summarize we had 5.502,379 total features (150:CD, 912: UV-Vis,3215: FTIR, NMR:1,769,526, 6096: HPLC, 1,396,240:SEC and 2326240: HILIC for each replicate of each sample).

Some of these techniques had artifactual data for certain ranges of wavelengths, thus we did not include the data from these regions in our analysis. We only have included the wavelengths between 240 and 600 nm for UV and wavenumbers between 1100 and 1700 cm^{-1} for FTIR. For SEC and HILIC we used the retention times from 10 to 18 mins and 3.5 to 46.5 mins, respectively. For ^{13}CNMR, we had 0 to 170 ppm, and $^{for\ 1}$HNMR, we had 2.5 to 7.5 ppm. We have not included HPLC and both NMR techniques for

classification in the later analysis as they had low information regions. After excluding the artifactual data our total number of features reduced from 3,725,757 to 2,419,134 features (150:CD, 351: UV-Vis,622: FTIR, 371,999:SEC and 2,046,000: HILIC).

5.2.3 Mutual Information

Mutual information (MI) measures the relatedness of two random variables [15, 16]. We referred to the method employed in related studies [11, 12, 17]. All the calculations are done using the mutual_info_score method from the scikit-learn package [14] in Python to calculate the mutual information score (MIS). Mutual information is a data mining technique that we used to identify regions of the data set that are rich in information content. The mutual information between 2 discrete random variables is defined as follows [17].

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

where X and Y are sets of possible x and y bins, p(x,y) is the joint probability distribution function of X and Y, and p(x) and p(y) are the marginal probability of distribution functions of X and Y, respectively. All the figures with MI have units of bits as we have used base two logarithms. As done in previous study [11], for calculating MI, y-bins are 27 categorical variables (nine mixtures with three fractions each), and x -bins are 6 (arbitrary) estimate the probability of range of feature values.

Feature reduction is often made to reduce dimensionalities and remove the redundant and low variance regions that add computational complexity to the problem. For our analysis feature selection was accomplished by ranking every feature based on their MI scores. So, a high MI score indicates that the data point is more dependent on the labels (each label is a distinct sample). We selected the top hundred MI scores after combining all the physical assay datasets for part of our analyses.

5.2.4 Principal Component Analysis (PCA)

To visualize many features in two-dimensional space for a more intuitive understanding, we utilized the popular dimensionality reduction technique of Principal Component Analysis (PCA). After preprocessing data and performing the background correction when needed, we performed PCA on the resulting matrix. The rows represent the different CF pure lots and mixtures, and each column represents the features. We used the function scale () from scikit-learn for standardization and *PCA ().fit().transform()* from scikit-learn in Python to calculate the principal components for the data [14].

5.2.5 Similarity analysis

A similarity measure is used to determine how similar two variables and here we quantified the similarity between two objects by measuring the Euclidean distance between them. We have six mixtures and three pure lots for our case, each having three different fractions; for consistency, we chose only three replicates from all fractions. Thus, we have an 81X81 matrix that contains the distances between samples by combining the data and representing it as a heatmap to visualize the high dimensional space in 2D. Since we are using data from different experimental techniques, it is crucial to perform standardization of the input feature matrix. We used the function euclidean_distances() from sklearn.metrics.pairwise in Python scikit-learn [14].

5.2.6 Machine learning Classifiers and Cross-Validation

For this study, we used six classifiers which include k-nearest neighbor (kNN) [18], linear discriminant analysis (LDA) [19, 20], support vector machine (SVM) [21, 22], Decision tree (DT) [23], Random Forest (RF) [24] and Ada-Boosted Decision Tree (AdaBoost) [25]. The parameters were optimized for each classifier by calculating the accuracies as done in [11].

Cross-validation is a statistical technique used to assess the effectiveness of a machine learning model on unseen data. This technique typically involves splitting the data into training and test datasets. The model

is trained using the training data set and later validated/evaluated over the test data set. There are several methods of cross-validation; in this work, we have used two methods:

1. Test-train split, or Monte Carlo test-train split: In this method, we randomly split data into subsets, like 70:30, 80:20, or 90:10. The larger number represents the training dataset subset, and the smaller represents the test data subset. For our work, we have split the data into 90:10 and repeated the test-train split 100 times to achieve statistical accuracy. The results reported are percentage accuracy, on the test data, averaged over these 100 individual cross validation experiments.
2. Leave – one – out: This a type of exhaustive cross-validation, where N samples are randomly chosen as a test set, and the remaining form the training set. For our analysis, N=1, so we had one (out of 108) sample as the test set and the remaining data for training our classifiers. Other popular variations include, leave-11-out or leave-8-out and so on. The results reported are average percentage accuracy.

5.2.7. Biosimilarity assessment

To validate if the classifiers distinguish between active and inactive samples, we performed a biosimilarity assessment combining the Croefelmer mixtures datasets used in this study ("Year2 CF lot") and the Croefelmer stability datasets ("Year1 CF lot") which is used in previous studies as described in three companion articles [4, 11, 12]. The "Year1 CF lot" were the stability studies conducted in 2016, which had a single batch of surface scraped Fulyzaq tablets were dissolved in water and centrifuged, the resulting filtrate was fractionated with 10kDa and 3kDa filers. This yielded 10kDa top, 10kDa bottom, 3kDa top, and 3kDa bottom fractions. Also, an unfractionated set of samples was used for this study. All these five fractions were then maintained at two temperatures 25^0C and 40^0C for 0 days (had only25^0C), 2 days, 1 week, and 1 month. These totally produced 35 distinct samples (4 days, two temperatures (not for day 0), and 5 fractions). For the biosimilarity experiment, the active group comprised of all unfractionated samples for 9 mixtures of "Year2 CF lot" and all unfractionated samples maintained at 0 days. This leads to 10

distinct samples for the active group. The inactive group, consisted of only "Year2 CF lot" 10kDa top and 3kDa Top maintained at 2days, 1week, and 1 month (25^0C and 40^0C). Hence, we have 12 samples for the inactive group.

5.3 Results

5.3.1 Analysis of UV-Vis, FTIR, CD physical assays data

Each of the Croefelmer pure lots and mixtures was subjected to UV-Vis's absorption, FTIR, and CD physical techniques. After, we did background corrections and normalization (see 5.2.2), we found that all these three biophysical techniques have high levels of mutual information (MI) (**Fig. 5.2**). MI is an information theoretic measure of statistical correlation between variables. We used MI scores to quantify

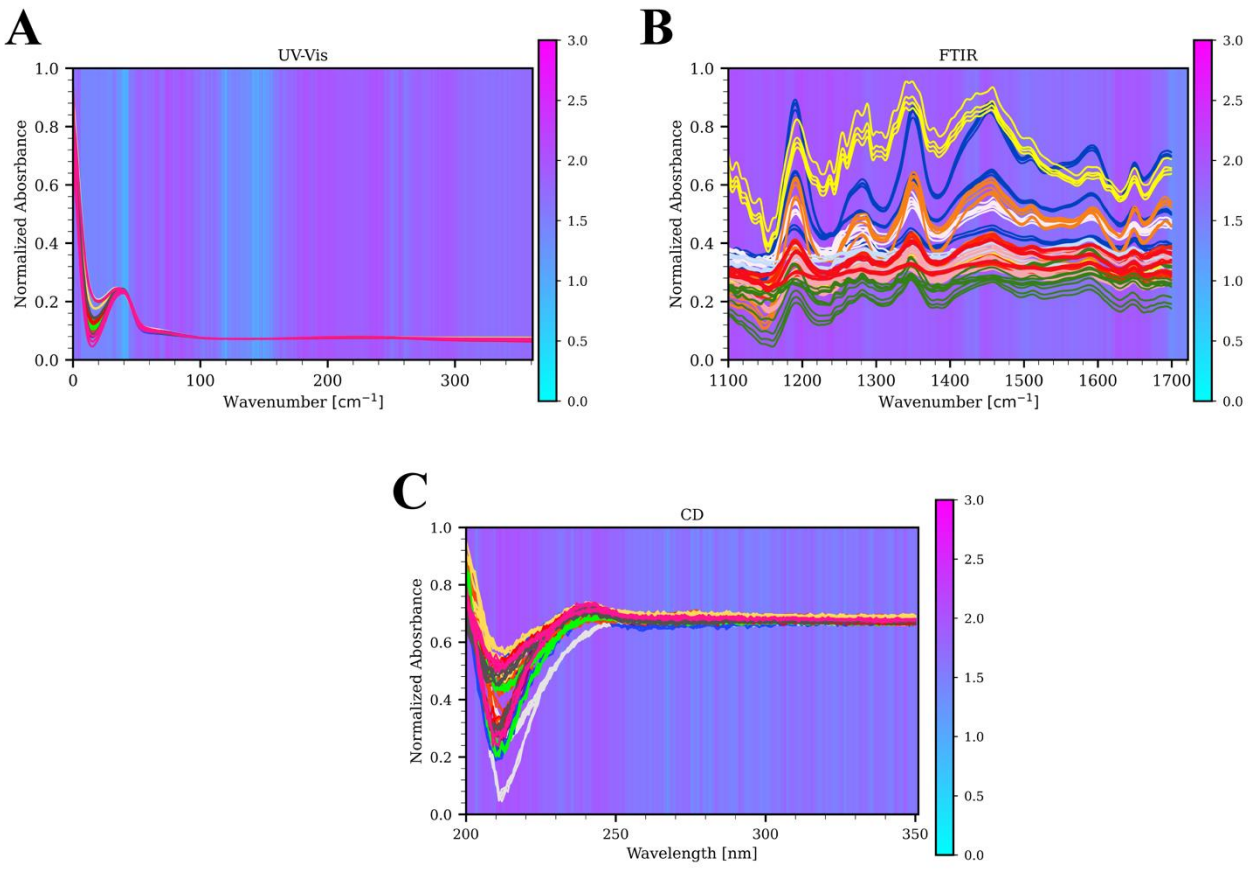


Figure 5.2: Raw data and mutual information score (in bits) for (A) UV-Vis absorption, (B) CD, and (C) FTIR data from Crofelemer pure lots and mixtures. In each plot, the colored lines show the normalized data for the corresponding technique and the background shows the mutual information score. In each case, the raw data were divided by the concentration of the samples, and the maximum intensity in each case was normalized to 1. The lines represent the different replicates from the pure lots and mixtures.

the data for each data point for all techniques and then used heat map to represent along with the actual data to visualize differences among mixtures. The MI scores are independently calculated between each signal and wavelength for each sample type (see 5.2.3). This analysis showed that UV, FTIR and CD have relatively high MI scores (1.7-2 bits) and they can discriminate between the Croefelmer mixture samples to some extent.

5.3.2 Analysis of NMR and HPLC data

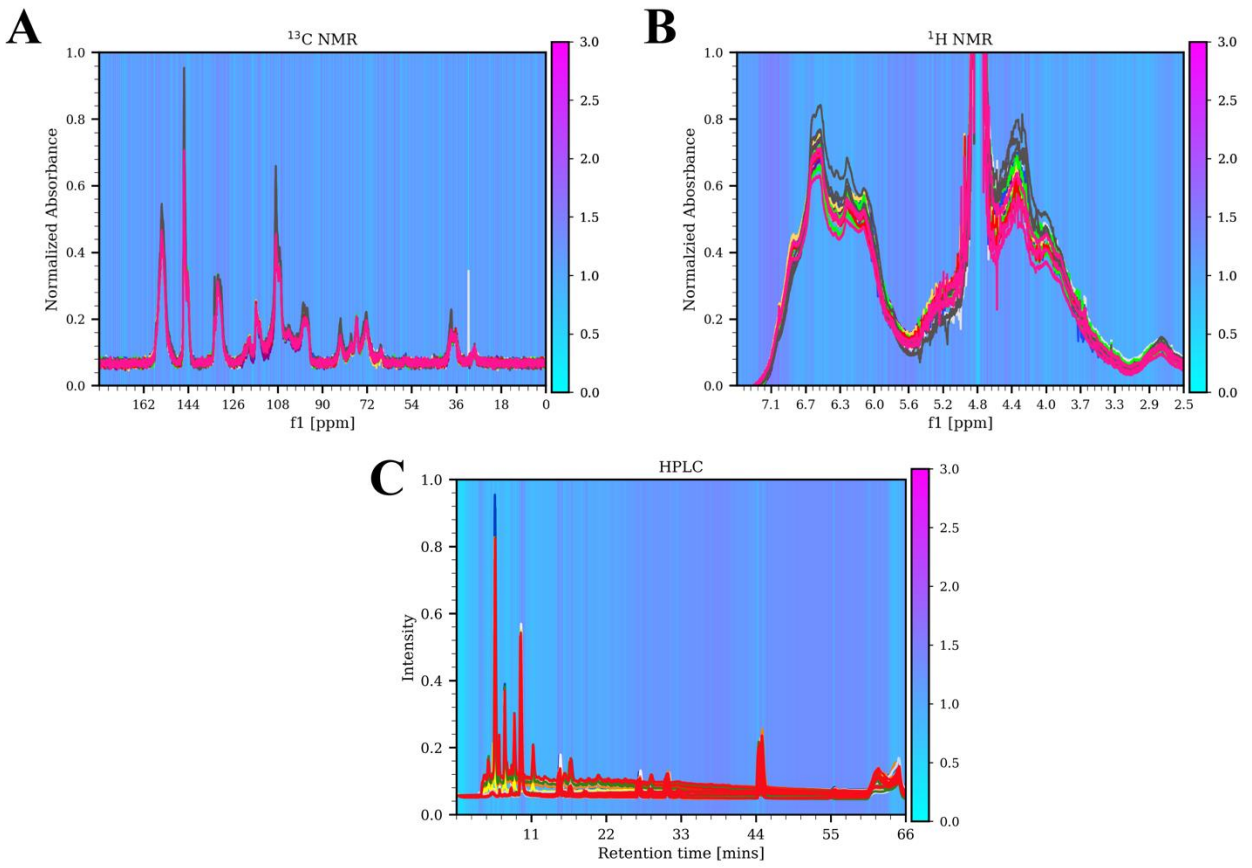


Figure 5.3: Raw data and mutual information score (in bits) for (A) ^{13}C NMR, (B) ^{1}H NMR, and (C) HPLC data from Crofelemer pure lots and mixtures. In each plot, the colored lines show the normalized data for the corresponding technique and the background shows the mutual information score. In each case, the raw data were divided by the concentration of the samples, and the maximum intensity in each case was normalized to 1. The lines represent the different replicates from the pure lots and mixtures.

Along with UV, FTIR, and CD we had three additional data sets from NMR and HPLC techniques. After computing MI scores as done in section 5.3.1, we observed that overall, for NMR and HPLC

techniques the MI scores ^{13}C NMR and HPLC have moderate levels of information and ^{1}H NMR has very low levels of information around 1.1 bits (**Fig. 5.3**). For further analysis and classification, we did not include NMR and HPLC data. Visualization analysis like PCA also showed relatively less clustering for NMR and HPLC (data not shown).

5.3.3 Analysis of SEC and HILIC physical assays data

Chromatography techniques like SEC, and HILIC HPLC were also applied to the Croefelmer mixtures, and these two techniques provide large amounts of data. For SEC, the data was collected from retention times of 0 to 30 minutes (at 640 millisecond interval) and the absorption spectra was collected from 190nm to 800nm for each retention time (Not matching the methods). Similarly, for HILIC the data was collected from retention times of 0 to 50 minutes (at 640 millisecond interval) and the absorption spectra was collected from 190nm to 800nm for each retention time. In total we have 1,396,240 features for SEC and 2,326,240 features for HILIC. This enormous amount of data is tedious and impractical to use for further characterization. So, we employed the MI score approach to examine the find useful subsets of the chromatography data for further analysis of Croefelmer mixtures.

For SEC we observed high MI scores for broader wavelength but smaller spans of retention times (~10-19mins) (**Fig. 5.4A**). To further quantify this observation, we averaged all MI scores over all wavelengths for a given retention time (**Fig. 5.4 B** blue curve) and separately averaged all retention times for a given wavelength (**Fig. 5.4 B** red curve). The average MI was highest for retention times between 10 to 12 mins. As earlier observed (**Fig. 5.4 A**) the curve showing MI scores averaged over time (**Fig. 5.4 B** red curve) did not show a prominent peak for all wavelengths.

HILIC data showed a different behavior, where we observed high MI scores for all retention times (0-50mins) for wavelength from 190nm to 255nm (**Fig. 5.4C**). We observed that retention times from 12mins to 50 mins for wavelength of 257nm to 460nm had very low MI scores (~ 0 bits) (**Fig. 5.4C**). Like the SEC case, we averaged the MI scores over all wavelengths and over all the times for HILIC data (**Fig. 5.4D**).

Here, the absorbance values for the wavelength near 220 nm had the maximum average MI score across all retention times. Additionally, we observed a maximum average MI score at 2 minutes averaged over all wavelengths (**Fig. 5.4D**). These results emphasize the utility of using MI scores as a tool for identifying regions with extremely rich sets of information.

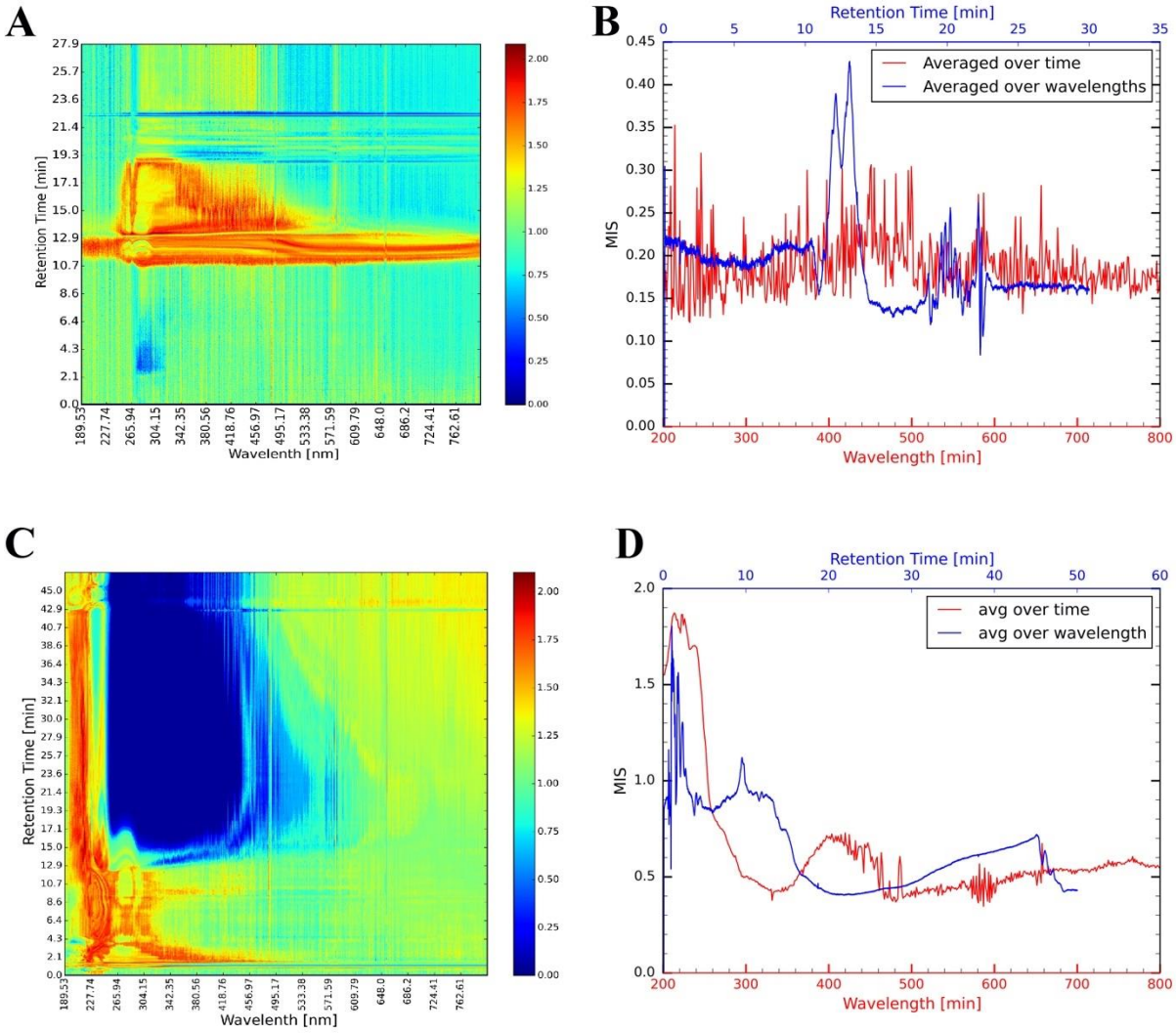


Figure 5.4: Plots for mutual information score (MIS in bits) and for SEC and HILIC of Croefelmer mixtures. (A) Heat map of mutual information score for SEC. (B) The red curve represents the mutual information score averaged over all retention times, and the blue curve represents the mutual information score averaged over all wavelengths for SEC data. (C) and (D) same as in (A) and (B) for HILIC data.

5.3.4 Visualization of data

From the MI scores data, we selected several subsets from all the techniques that would be useful for classifying the different pure lots and mixtures. From UV-vis we took data from (240 – 300 nm), CD from (200 – 350 nm), FTIR from (1100 cm^{-1}-1700cm^{-1}) (**Fig. 5.3**), SEC all wavelengths from retention times of 10-18 mins (**Fig. 5.4A, B**), and for HILIC we selected all wavelengths from retention times of 3.52 to 47.5 mins (**Fig. 5.4 C, D**). All these five techniques combined (NMR and HPLC were not included due to low MI scores) gave a total of 2,419,134 features per replicate/sample (total 108 reps). To visualize the data better we performed a PCA with the 2,419,134 features per sample (**Fig. 5.5A**). We can still see that the data does not show clear separations among pure lots and mixtures, even though the first two principal components capture almost 80% variance (**Fig. 5.5A**), where each fraction and mixture are shown with a different legend. Additionally, we also used a similarity analysis visualization which represents data based on Euclidean distances (**Fig. 5.5B**). The similarity analysis and PCA helps us visualize the Croefelmer mixtures among large feature space. The similarity analysis heat-map (**Fig. 5.5B**) shows that all the three pure lots and six mixtures cannot be distinguished in the combined data set. As seen, the pure lots (P1, P2 and P3) with fractions (Top-T, Bottom-B, and Unfractionated-U) are all equidistant and not distinct.

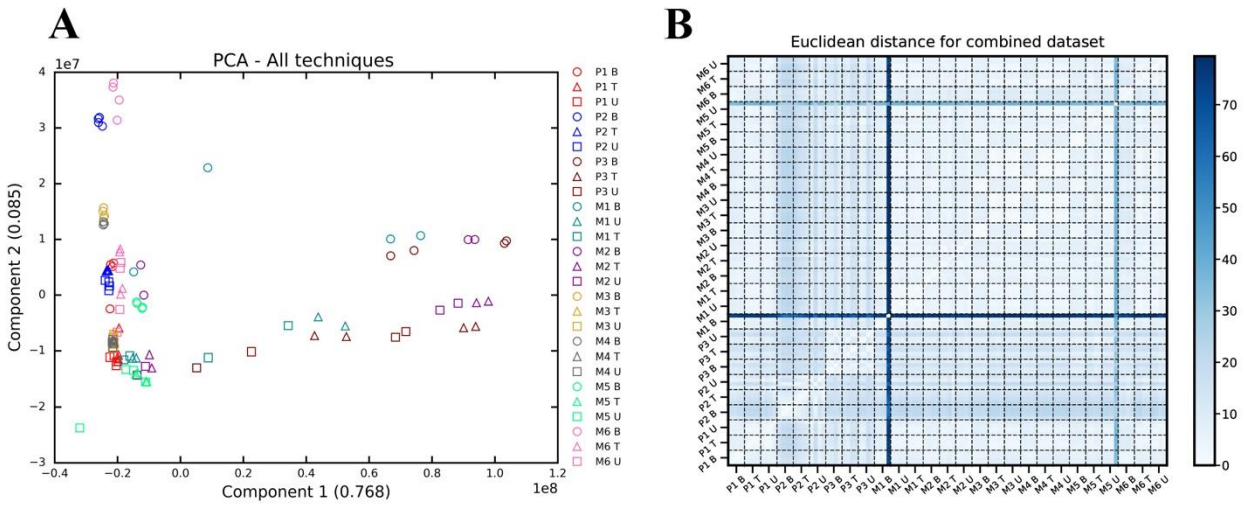


Figure 5.5: A) Principal Component Analysis and B) Similarity Analysis for five techniques combined.

5.3.5 Visualization of the physical techniques of top 100 mutual information scores from datasets

For feature selection we decided to select the top 100 features with the highest Mutual Information Scores (MIS) among all the data. After selecting the top 100 MIS, we observed that the 1st and 2nd principal components capture more than 82% of the variance in the data (**Fig. 5.6A**) and the heat-map from similarity analysis shows a clear distinction between bottom and other fractions (**Fig. 5.6B**).

Selecting the features with top 100 MI scores, thus does a better job in capturing differences between samples. We also saw that the three pure lots (P1, P2 and P3), which are three different batches, shows some clustering in the PCA representation (**Fig. 5.6A**). Overall, selection of the top 100 features leads to somewhat better clustering and much better similarity analysis (**Fig. 5.6A, B**).

We investigated further to see which techniques contribute the most to top 100 MI scores, and we observed that the high dimensional techniques i.e., SEC and HILIC, contribute the most. As seen in the pie chart **Fig. 5.7** HILIC contributes about 64%, followed by SEC with a 22%, FTIR about 9% and UV-Vis with 5%. Interestingly, we did not see any features from CD in the top 100 MI scores (**Fig. 5.7**).

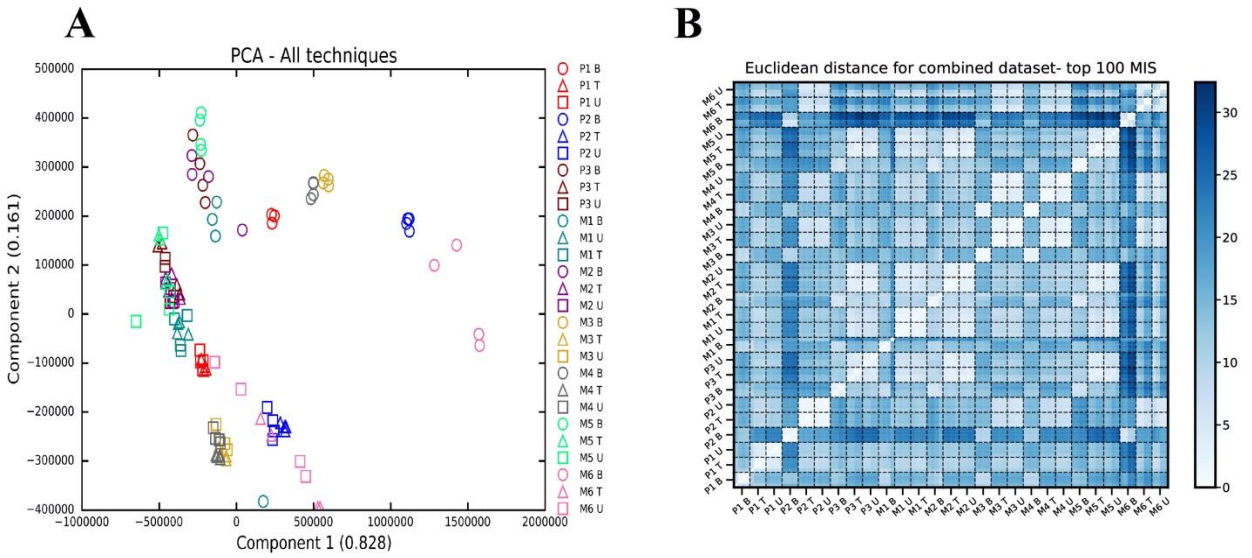


Figure 5.6: A) Principal Component Analysis for top 100 MIS and B) Similarity Analysis for five techniques combined for top 100 MIS.

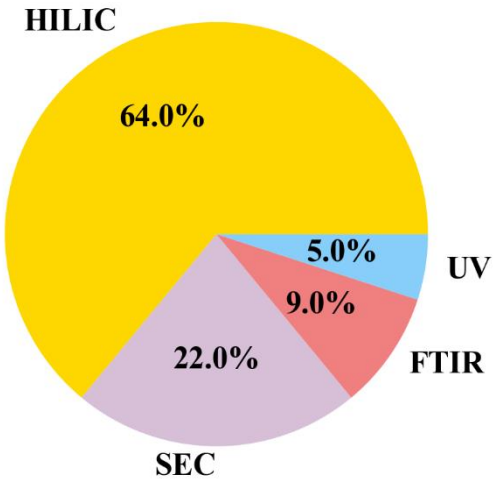


Figure 5.7: Pie chart shows the percentage of features that are part of the top 100 mutual information scores.

5.3.6. Classification Analysis

The MI score is a promising information theoretic measure that can be used to extract useful features. As seen in the MI scores analysis, most techniques have high information about score (2.5 bits) on their own. But no single feature can identify the different mixtures, to have a more robust classification, we have used the machine learning classifiers on every technique. We performed supervised learning classification using the seven classifiers (see 5.2.6) on individual techniques and combing all the techniques. Similar to the approach used in a previous study [11] we used Monte Carlo test train split with 91% of data as training set and 9% as test set and averaged the accuracies over 100 iterations to determine statistical significance. From the seven classifiers used LDA performed the best and kNN, SVM (linear and radial), and random forest also performed well (**Table 5.8**). Clearly, as MI scores indicated NMR and HPLC techniques had very low classification accuracy (**Table 5.8**).

We also used leave-one-out cross validation. The same classifiers which performed well for Test-train split had higher accuracies under leave-one-out cross validation technique (**Table 5.9**).

Technique	kNN	LDA	SVM (Linear)	SVM (Radial)	ADA Boost	Decision Tree	Random Forest
CD	0.83	0.87	0.88	0.87	0.19	0.50	0.78
UV	0.88	0.90	0.92	0.91	0.24	0.64	0.85
FTIR	0.84	0.99	0.91	0.88	0.34	0.77	0.92
HPLC	0.12	0.42	0.18	0.15	0.11	0.17	0.22
^{13}C NMR	0.24	0.15	0.16	0.09	0.07	0.15	0.16
^{1}H NMR	0.32	0.10	0.21	0.17	0.08	0.14	0.18
SEC	0.83	0.85	0.67	0.55	0.13	0.61	0.51
HILIC	0.52	0.54	0.68	0.45	0.13	0.63	0.61

Table 5.8: Classification accuracies under Test-train split for individual technique averaged over 100 iterations of cross-validation. Accuracies greater than 75% are in red.

Technique	kNN	LDA	SVM (Linear)	SVM (Radial)	ADA Boost	Decision Tree	Random Forest
CD	0.84	0.93	0.88	0.12	0.31	0.50	0.81
UV	0.89	0.94	0.95	0.18	0.27	0.57	0.72
FTIR	0.96	1.0	0.98	0.16	0.17	0.79	0.94
HPLC	0.08	0.48	0.20	0.12	0.0	0.12	0.17
^{13}C NMR	0.33	0.14	0.25	0.11	0.0	0.03	0.11
^{1}H NMR	0.11	0.18	0.14	0.03	0.07	0.03	0.22
SEC	0.66	0.79	0.66	0.16	0.04	0.45	0.55
HILIC	0.57	0.60	0.58	0.14	0.13	0.37	0.52

Table 5.9: Classification accuracies under leave -one -out cross validation for individual technique averaged over 100 iterations o cross-validation. Accuracies greater than 75% are in red.

For leave-one-out classification accuracies, we noticed that LDA has 100% accuracy for the FTIR technique. Interestingly, similar studies on biosimilarity assessment using leave-one-out cross-validation also had the LDA classifier showing as high as 100% accuracy [11, 13].

Classifier	Test-train split	Leave-one-out cross-validation
kNN	0.65	0.67
LDA	0.51	0.59
SVM(Linear)	0.67	0.66
ADB	0.13	0.24
SVM(Radial)	0.33	0.17
Decision Tree	0.55	0.48
Random Forest	0.50	0.51

Table 5.10: Classification accuracies under test-train split and leave -one -out cross validation for combined datasets averaged over 100 iterations. None of the accuracies were above 75%.

Similar to test-train split NMR and HPLC had very low classification accuracies for the leave-one out scheme as well (**Table 5.10**). Next, we used data combined from five techniques (UV, CD, FTIR, SEC and HILIC) to classify using the test-train split and leave-one out cross-validation schemes. Combination of all data leads to low classification accuracies, even LDA which had 100% accuracy under leave-one out cross validation for FTIR shows low accuracy.

Technique	Test-train split	Leave-one-out cross-validation
kNN	0.95	0.79
LDA	0.97	0.99
SVM(Linear)	0.97	0.97
ADB	0.5	0.29
SVM(Radial)	0.97	0.22
Decision Tree	0.71	0.76
Random Forest	1.0	0.84

Table 5.11: Classification accuracies under test-train split and leave-one-out cross validation for combined datasets averaged over 100 iterations. Classification accuracies which are above 75% are highlighted in red.

We did not find any classifiers greater than 75% accuracy for the combined data set (Table. 5.11). We had previously seen (**Fig. 5.5**) that combining these techniques shows poor clustering and cannot distinguish between Croefelmer mixtures. This is very likely due to the low variance and low information data in the physical techniques; these data points lower the classification accuracy and makes the samples indistinguishable. Next, we performed feature selection as previously described in section 5.2.3 by selecting top 100 features with highest MI scores. We see improvement in classification accuracy with this MI feature selection (**Table 5.11**). All the techniques except Adaptive Naïve Bayes (ADB) had greater than 75% accuracy for at least one cross-validation scheme. kNN, LDA, SVM (linear and radial), and random forest all have very high classification accuracies for both the cross-validation schemes. These accuracies are consistent with our previous clustering and visualization results described in section

5.3.7. Biosimilarity Experiment

Biosimilar drugs are expected to be highly similar to the FDA-Approved drugs (reference drugs). The biosimilar drug should have no have no clinically meaningful differences in safety, purity, or potency (safety and effectiveness) compared to the reference product. The biosimilar and reference drugs should exhibit similar biochemistry in structure and function. A previous stability study on Croefelmer drug by our group has used similar mathematical approaches to identify signatures to distinguish Croefelmer drug samples [11].

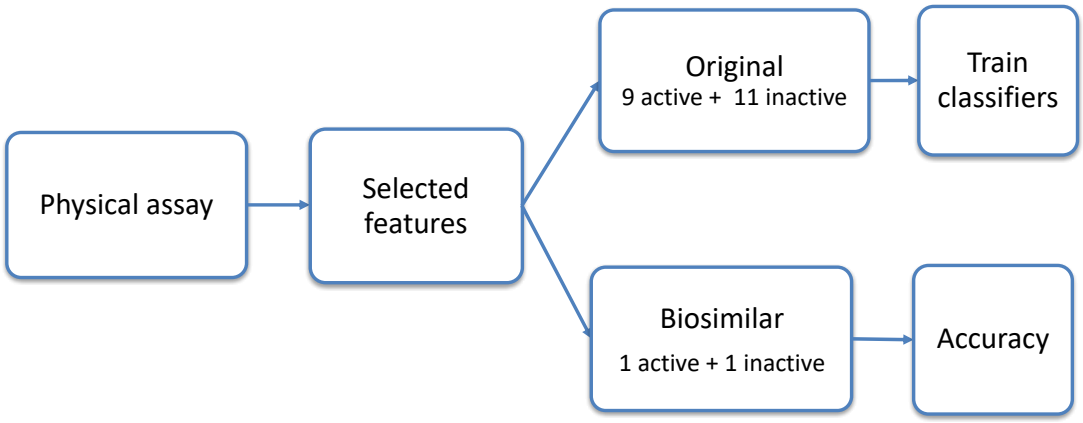


Figure 5.12: Model of the Comparative Characterization experiment for CF mixtures and stability data.

The data from this study also labeled as "Year 1 CF lot" Croefelmer data was utilized for simulating a biosimilarity experiment (See 5.2.7). The Croefelmer mixtures datasets used for this Chapter are labeled as "Year 2 CF lot". The Croefelmer datasets were divided into two categories: active and inactive. The active category has all unfractionated samples from the three pure lots and six mixture datasets (**Fig. 5.1**) and additionally it has all the unfractionated samples of Day 0 from "Year 1 CF lot". These fractions have the active ingredients in them. So, in total the active dataset has ten distinct samples (nine mixtures from "Year 2 CF lot" and one mixture from "Year 1 CF lot"). The inactive dataset has all unfractionated and 3kD top fractions for the days 2,7, and 30 (for two temperatures i.e., 25^0C and 4^0C). Thus, it has twelve

distinct samples which are not having the active ingredient and hence the CF from these mixtures are not effective.

From the active and inactive datasets, we randomly chose one sample from each group and left it out as biosimilar data i.e., two samples (1 sample from 10 active and 1 from 12 inactive) (**Fig. 5.12**). The remaining twenty samples from the training or original dataset. Next, we performed supervised learning classification using Test-train split (91% training and 1% test) and leave-one out cross validation schemes. The results of this classification are summarized in **Table 5.13**. Note that, there are 120 combinations ($^{(Active)10}C_1$ X $^{(inactive)12}C_1$ = 120) of leaving out the data for this active and inactive groups. The classifiers are optimized and calculated for all the possible combinations. We see that on an average we obtain greater than 75% accuracies for most classifiers.

Technique	Test-train split	Leave-one-out cross-validation
kNN	0.99	0.95
LDA	0.99	0.99
SVM(Linear)	0.99	1.00
ADB	0.97	0.98
SVM(Radial)	0.39	0.97
Decision Tree	0.97	0.98
Random Forest	0.97	0.98

Table 5.13: Classification accuracies under test-train split and leave-one-out cross validation for biosimilarity experiment averaged over 100 iterations. Classification accuracies which are above 75% are highlighted in red.

We observed very high classification accuracies for the experiment, besides SVM (radial) under test-train split. SVM (linear) shows a 100% accuracy for the biosimilarity assessment (**Table. 5.13**). These high accuracies implied a very high rate of success to model biosimilarity assessment using physical techniques data. In future, this model can be also applied to other biosimilar drugs and understand their biological activity and lot-to-lot variability during manufacturing.

5.4 Discussion

Drugs made using natural sources as raw materials are highly heterogeneous and are subjected to batch-to-batch variability [4, 10, 11, 13]. In this work, we have different mixtures of the botanical drug Crofelemer (CF), which forms our model system. This characterization study aimed to identify subtle differences between these samples, which are missed in the traditional data analysis and analytical characterization methods. The mathematical model helps for a better understanding of data and select the regions of high information. Further, the use of classifiers can be used to identify differences in physical assays. This study is similar and uses similar methods and data mining approaches in our research group [11].

Physiochemical assays like UV-vis, CD, FTIR, NMR, SEC, and HILIC were then performed on these mixtures to obtain the analytical datasets. The UV, CD, and FTIR were used to obtain spectroscopy characters, NMR, SEC, and HILIC to get the information on the degree of polymerization, composition, and molecular weight distribution [4]. In this work, we have applied a mathematical model for the comparative characterization of CF mixtures datasets. Modern data mining and machine learning methods make a robust and well-suited technique to analyze large datasets. We demonstrated that using techniques like mutual information, PCA, similarity analysis, and machine learning methods, it is possible to differentiate and classify the different mixtures and address batch-to-batch variability issues.

Our analysis dataset consists of the CF mixtures from three pure lots (different batches) and six mixtures. Further, each of these fractions has multiple replicates; the bottom fractions, the top, and unfractionated fractions have four replicates. Therefore, for nine mixtures and three fractions with multiple replicates, we have a total of 108 samples. Our analysis using MI scores showed that most techniques have regions of high information, except HPLC and the NMR techniques. Even PCA showed good clustering and separation and separation for UV-vis, CD, and FTIR but had poor separation and less structuring for HPLC and NMR (data not shown). Most of the PCA results on individual techniques showed a clear separation of the bottom fractions, which are inactive. The PCA of CD also showed a clear separation of the pure lots

and mixtures (data not shown). Overall, the five techniques - UV-vis, CD, FTIR, SEC, and HILIC contained rich and essential information, and HPLC and the NMR techniques did not have enough information to classify the CF mixtures.

The chromatography techniques like SEC and HILIC generate vast data sets as these techniques collect data over a range of absorption spectra (200 nm-800 nm) for intervals in retention times (30-60 minutes). Analyzing such large datasets with traditional methods is not feasible; our approach of using information-theoretic metric - mutual information (MI) helped identify information-rich regions. Feature selection based on MI scores enabled us to reduce the features from 1,396,240 to 371,999 in SEC and from 2,326,240 to 2,046,000 in HILIC datasets for each sample. To see if a single feature can differentiate between all mixtures, we combined the data from five techniques. The combined datasets PCA and similarity analysis (**Fig. 5.6A, B**) successfully determine the bottom fractions of very low concentrations (less than 10kDa) from the unfractionated and top. Moreover, we also saw the separation between the three different batches: pure lot 1, 2, and 3 (PL in **Fig. 5.6 A**). The similarity analysis heatmap also shows that bottom fractions are very distinct because of low concentration and are inactive, shown in **Fig. 5.6 B**.

We used the supervised machine learning classifiers to classify the samples. When used for a single technique, we observed that for two cross-validation schemes, we obtained high accuracy for UV, CD, FTIR, and SEC (**Fig. 5.9 and 5.10**). We excluded the NMR and HPLC datasets with significantly less information and then combined the remaining five UV, CD, FTIR, SEC, and HPLC techniques. Using classifiers on this combined dataset gave low accuracy, and none of the classifiers used had greater than 75% accuracy (**Fig. 5.11**). Interestingly, the top 100 MI scores and select those features for classification accuracies allowed several classifiers to achieve near 100% accuracy for the random forest, and other classifiers performed very well. This is one of the critical results of our work, suggesting that it is possible to improve classification accuracies by using top-ranked features for the physical assays (**Table 5.13**).

Another interesting application of using mathematical models for such data sets is to simulate biosimilarity experiments. For our work, we collected the data from previous stability studies on CF [4, 11, 12], which had 35 distinct samples (Year 1 CF lot). This stability study started with a single batch of CF

drug, and then it was subjected to thermal degradation at two temperatures -25^{0}C and 40^{0}C and incubated for different lengths of time (See 5.2.7). We know that for day 0 and temperature of 25^{0}C, the unfractionated samples are active. This sample from Year 1 datasets and the nine samples of the study of the current mixture (only unfractionated samples from three pure lots and six mixtures) together formed ten samples of the active group. The inactive group had 12 samples from the "Year 1 CF lot" (which were incubated for more than a day) [11]. Then, we randomly chose one sample from the active and the inactive group as our test set and train the machine learning classifiers on the remaining data. We observed very high classification accuracies. The classifiers successfully distinguished among the active and inactive samples.

This simulated experiment is a great example of applying machine learning methods to differentiate active and inactive samples with high statistical significance. Such methods can be applied to biosimilars. The new product is an identical copy of the reference (original) product and is expected to show similar biochemistry, safety, and efficacy as the reference product. The mathematical model in our study can have also been applied to monoclonal antibodies (IgG) datasets from physical, chemical, and biological assays. In the future, similar methods can be applied for cases where there has been a change in the manufacturing process, raw material, or contaminated drugs manufactured, and these models help evaluate the new product before it is available for public use.

Overall, our results in Chapter 5 suggest that using machine learning methods is a promising approach to differentiate between mixtures or batches of biological origin drugs. They can classify mixtures exhibiting different biological activity. Also, the mathematical model used can discriminate between active and inactive drugs (expired or not having potency). Perhaps, additional chemical and biological assays for the CF mixtures like fluorescence assays, single-cell patch-clamp assays, and oxidation relates studies can give us richer information and help further evaluate our model to determine if it was robust. Applying our model to physical, chemical, and biological assays data can help to reduce the number of assays and experiments in analytical characterization, and further highlight subtle CQAs missed in the routine analysis. More studies for characterizing biopharmaceuticals are essential and are upcoming with the recent advances in machine learning and artificial intelligence. These techniques have also found application in studies on

microRNA datasets [26], nanoparticles of medical interest [27], air pollution epidemiology [28], gene expression data [29], and many more studies [30, 31]. These studies describe the enormous potential and challenges faced in using machine learning and data mining approaches systematically to obtain proper classification and biomedical information to advance scientific discoveries in the new computational era.

5.5 References

1. Patridge E, Gareiss P, Kinch MS, Hoyer D: An analysis of FDA-approved drugs: natural products and their derivatives. *Drug Discov Today* 2016, 21(2):204-207.
2. Frampton JE: Crofelemer: a review of its use in the management of non-infectious diarrhoea in adult patients with HIV/AIDS on antiretroviral therapy. *Drugs* 2013, 73(10):1121-1129.
3. Are Botanical Drugs, Herbal Medicinal Supplements, and Natural Product Drugs 505(b)(2)s, Too? [https://camargopharma.com/resources/blog/are-botanical-drugs-herbal-medicinal-supplements-and-natural-product-drugs-505b2s-too/]
4. Kleindl PA, Xiong J, Hewarathna A, Mozziconacci O, Nariya MK, Fisher AC, Deeds EJ, Joshi SB, Middaugh CR, Schöneich C *et al*: The Botanical Drug Substance Crofelemer as a Model System for Comparative Characterization of Complex Mixture Drugs. *Journal of pharmaceutical sciences* 2017, 106(11):3242-3256.
5. Cottreau J, Tucker A, Crutchley R, Garey KW: Crofelemer for the treatment of secretory diarrhea. *Expert Rev Gastroenterol Hepatol* 2012, 6(1):17-23.
6. Crutchley RD, Miller J, Garey KW: Crofelemer, a novel agent for treatment of secretory diarrhea. *Ann Pharmacother* 2010, 44(5):878-884.
7. Chordia P, MacArthur RD: Crofelemer, a novel agent for treatment of non-infectious diarrhea in HIV-infected persons. *Expert Review of Gastroenterology & Hepatology* 2013, 7(7):591-600.
8. Frampton JE: Crofelemer: A Review of its Use in the Management of Non-Infectious Diarrhoea in Adult Patients with HIV/AIDS on Antiretroviral Therapy. *Drugs* 2013, 73(10):1121-1129.
9. Federici M, Lubiniecki A, Manikwar P, Volkin DB: Analytical lessons learned from selected therapeutic protein drug comparability studies. *Biologicals* 2013, 41(3):131-147.
10. Lubiniecki A, Volkin DB, Federici M, Bond MD, Nedved ML, Hendricks L, Mehndiratta P, Bruner M, Burman S, DalMonte P *et al*: Comparability assessments of process and product changes made during development of two different monoclonal antibodies. *Biologicals* 2011, 39(1):9-22.
11. Nariya MK, Kim JH, Xiong J, Kleindl PA, Hewarathna A, Fisher AC, Joshi SB, Schöneich C, Forrest ML, Middaugh CR *et al*: Comparative Characterization of Crofelemer Samples Using Data Mining and Machine Learning Approaches With Analytical Stability Data Sets. *J Pharm Sci* 2017, 106(11):3270-3279.
12. Hewarathna A, Mozziconacci O, Nariya MK, Kleindl PA, Xiong J, Fisher AC, Joshi SB, Middaugh CR, Forrest ML, Volkin DB *et al*: Chemical Stability of the Botanical Drug Substance Crofelemer: A Model System for Comparative Characterization of Complex Mixture Drugs. *J Pharm Sci* 2017, 106(11):3257-3269.
13. Kim JH, Joshi SB, Tolbert TJ, Middaugh CR, Volkin DB, Smalter Hall A: Biosimilarity Assessments of Model IgG1-Fc Glycoforms Using a Machine Learning Approach. *Journal of Pharmaceutical Sciences* 2016, 105(2):602-612.
14. Fabian Pedregosa GV, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. : Scikit-learn: Machine Learning in Python. *JMLR* 2011 2(85).

15. Duncan TE: On the Calculation of Mutual Information. *SIAM Journal on Applied Mathematics* 1970, 19(1):215-220.
16. Kraskov A, Stögbauer H, Grassberger P: Estimating mutual information. *Physical Review E* 2004, 69(6):066138.
17. Thomas M. Cover. J, A. Thomas: Elements of Information Theory: Wiley-Interscience; 2006.
18. Altman NS: An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* 1992, 46(3):175-185.
19. Huberty CJ: Applied Discriminant Analysis (Wiley Series in Probability and Statistics) Wiley Series in Probability and Statistics); 1994.
20. McLachlan GJ: Discriminant Analysis and Statistical Pattern Recognition. *Wiley Series in Probability and Statistics* 2004.
21. Vapnik V: The Nature of Statistical Learning Theory; 1995.
22. Yin-Wen Chang C-JH, Kai-Wei Chang, Michael Ringgaard, Chih-Jen Lin: Training and Testing Low-degree Polynomial Data Mappings via Linear SVM. *JMLR* 2010, 11(48).
23. Quinlan R: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning. San Francisco:

Morgan Kaufmann Publishers Inc; 1993.
24. Breiman L: Random Forests. *Machine Learning* 2001, 45(1):5-32.
25. Schapire YFaRE: Experiments with a New Boosting Algorithm *Machine Learning: Proceedings of the Thirteenth International Conference* 1996.
26. Piątek Ł, Grzymała-Busse JW: LEMRG: Decision Rule Generation Algorithm for Mining MicroRNA Expression Data. *Adv Exp Med Biol* 2017, 1028:105-137.
27. Jones DE, Ghandehari H, Facelli JC: A review of the applications of data mining and machine learning for the prediction of biomedical properties of nanoparticles. *Comput Methods Programs Biomed* 2016, 132:93-103.
28. Bellinger C, Mohomed Jabbar MS, Zaïane O, Osornio-Vargas A: A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health* 2017, 17(1):907.
29. Tuana G, Volpato V, Ricciardi-Castagnoli P, Zolezzi F, Stella F, Foti M: Classification of dendritic cell phenotypes from gene expression data. *BMC Immunol* 2011, 12:50.
30. Zlotogorski-Hurvitz A, Dekel BZ, Malonek D, Yahalom R, Vered M: FTIR-based spectrum of salivary exosomes coupled with computational-aided discriminating analysis in the diagnosis of oral cancer. *J Cancer Res Clin Oncol* 2019, 145(3):685-694.
31. Sampson DL, Parker TJ, Upton Z, Hurst CP: A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches. *PLoS One* 2011, 6(9):e24973.

Chapter 6

Conclusion and Future Directions

Macromolecular machines are crucial components of living systems and are assembled from multiple subunits. These machines are evolutionarily designed to be effective, robust and are involved in nearly all cellular processes. Ribosomes, nucleosomes, spliceosomes, GroEL, Clp proteases, proteasomes, virus capsids, and motor proteins are some examples of such machines, and they are present in all life forms, from tiny viruses, ancient archaea, and bacteria to the complex eukaryotes. The prokaryotic proteasome 20S Core Particle (CP) is the molecular machine of interest in this thesis. The proteasomal CP consists of four heptameric rings arranged in a cylindrical pattern $\alpha_7\beta_7\beta_7\alpha_7$. Here, we address multiple questions on CP assembly using *Rhodococcus erythropolis* (*Re*) bacterium as our model system. The CP has gained significant interest recently as it is a novel drug target for the treatment of tuberculosis [1-4], multiple myeloma [5-9], and other diseases [6, 10, 11]. Chapters 2-4 address specific questions in bacterial CP assembly, which have long been unanswered in biology. Chapter 2 addresses the slow dimerization of HPs into CP. Chapter 3 provides evidence of why near-HP intermediates do not associate with HP or each other. Chapter 4 investigates the formation of HPs in assembly and assembly dynamics of different CP assembly pathways.

Experimental studies indicate that the *Re* β propeptide can regulate the slow dimerization in CP assembly. In Chapter 2, we used Molecular Dynamics (MD) simulations of *Re* pre-holo CP- $\alpha_7\beta_7\beta_7\alpha_7$ to investigate the role of β propeptide in dimerization of two Half Proteasomes (HP- $\alpha_7\beta_7$) into a fully assembled CP. We categorize the *Re* β propeptide into three regions (I, II, and III). Our findings suggest that the length and polarity of Region III impact dimerization rate. We observed that extending Region III by including charged residues yields more interactions between the propeptide and a set of key residues at

the dimerization interface. These non-covalent interactions delay HP dimerization and thus CP assembly. The mutation of charged residues to Alanine (FAST mutant) in Region III reduces the interactions between propeptide and key residues, thus allowing the key interactions to form between two opposing HPs and making CP assembly faster than in WT. Statistical analysis showed that the distributions of hydrogen bond interactions between propeptide and key residues in the WT simulations differed significantly from the mutants. Both the FAST and SLOW mutants are validated experimentally. Our MD simulations analysis revealed other charged residues in Region III (like ARG-15 in **Fig. 2.1C**), potentially mutated, and studied to observe if it is faster than the FAST mutant.

The *Mycobacterium tuberculosis* pathogen is a close relative of *Re* and a similar length β propeptide. In the future, we can investigate if similar separation of time scales also exists in *Mtb* and if β propeptide can impact its HP dimerization rate and hence CP assembly. Additionally, structure-based drug design approaches can utilize the HP structures from our MD simulations to screen novel and potent small molecule CP assembly inhibitors. As there are no crystal structures available for any CP assembly intermediates, our MD simulations can serve as ideal templates for drug design approaches.

The HP is an obligatory intermediate in every species CP assembly pathway; we have no evidence of CPs with less than 28 subunits (CP-like). Near-HP structures ($\alpha_6\beta_7$, $\alpha_6\beta_6$, and $\alpha_7\beta_6$) never dimerize with a true HP($\alpha_7\beta_7$) or each other. We hypothesized that likely an allosteric mechanism operates in these intermediates to prevent the formation of CP-like structures. The near-HP intermediates MD simulations discussed in Chapter 3 reveal a global conformational shift and distortion in the α and β rings geometry. We saw a significant change in the β subunits angle in the intermediates with the incomplete β ring. In $\alpha_6\beta_7$, we noticed that β subunits near the missing α subunit change their tilt angle and are distorted in shape and geometry. These significant shifts are allosterically communicated through the propeptide and the subunits and conformational changes to the complete quaternary structure. As a result, these intermediates have incompatible interfaces and cannot associate stably with HPs or each other. Such global conformational shifts may indicate that the subunits of the CP are intrinsically frustrated in true HPs ($\alpha_7\beta_7$). This would

need further analysis to investigate the interactions between interfaces and the thermodynamic contributions to global conformational shifts.

. We also do not understand why an β or an α monomer cannot bind with an HP; the molecular mechanism which prevents their association that forms incompatible intermediates like $\alpha_7\beta_7$-β remains unexplored. It would also be intriguing to develop theoretical biophysical models and design more simulations for hierarchical assemblies other than the proteasome. We speculate that the subunits in assembled multi-subunit non-proteasome complexes are likely intrinsically frustrated. Additional analysis from the *Re* near-HP simulations, for instance, investigating different interface contacts, the surface contact area, will also provide further evidence for our speculation that these subunits are intrinsically frustrated.

In the final study on CP assembly in Chapter 4, we utilized Ordinary Differential Equations to develop mathematical models for investigating kinetic trapping and the assembly dynamics in bacterial CP. Previous studies on ring-like structures [12] demonstrated that in conditions where the initial subunit concentrations are fixed, like in the case of *in vitro* experiments, increasing the subunit concentration gives rise to a plateau phase called "deadlock" because the incompatible intermediates dominate. Deadlock is a type of kinetic trapping phenomenon, which occurs when monomers are exhausted, and the stable incompatible intermediates accumulate and dominate the system. These intermediates are very stable (ring-like); hence do not dissociate readily and are kinetically trapped. Our deterministic simulations were run for three CP assembly models, two of which are widely accepted – Alpha Ring First (ARF) and Alpha Beta Dimer (ABD) in the proteasome field. Here, we proposed a new pathway of CP assembly- Unordered Model (UOM), which is partially hierarchal and has not been examined before in any CP assembly study. Our simulations indicate that kinetic trapping occurs in UOM and ABD but not in the ARF. The kinetic trapping is also observed for *in vitro* experiments on *Re* CP. These assembly models demonstrated a tradeoff between speed and robustness, where UOM is the fastest and least robust, whereas ARF is the lowest but very robust. Interestingly, the UOM is also fast *in vivo*, implying that evolution has favored faster assembly over robustness. We can develop more models to investigate the effect of association rates on assembly dynamics

and kinetics in the future. Also, more comprehensive models can be developed to investigate how evolution has overcome the challenge of kinetic trapping in other molecular machines.

In the miscellaneous Chapter 5, we discuss the application of machine learning and data mining methods for biopharmaceuticals characterization. Botanical drug Croefelmer displays heterogeneity and batch-to-batch variation due to its natural source of raw material. We used a mathematical model for comparative characterization of complex mixture drugs using Croefelmer (CF) as a model system. Our collaborators had generated CF drug mixtures and then subjected them to physicochemical assays like UV-Vis, CD, NMR, HPLC, SEC and HILIC, for data collection. We used PCA and similarity analysis techniques to visualize the data, and mutual information (MI) score to identify "high information" regions within the datasets. Finally, we performed supervised machine learning classification to detect differences in different CF mixtures. Feature selection (selecting a subset of data) was made by ranking every data point by their MI scores and selecting the top 100 features with the highest MI scores. One of the key results of this work is that it was possible to significantly improve classification accuracy and discriminate CF mixtures by using the top-ranked features for the physiochemical assays based on their MI scores. This combination of mathematical models can also be applied to characterize other biologics, identify batch-to-batch variability, or even analyze large volumes of physiochemical assays data. As previous studies by our group [13, 14], MI scores offer an effective way to identify species that differ due to different storage temperatures or times and provide further insights.

Molecular machine assembly is hierarchal and can assemble via different pathways to regulate assembly efficiency. As shown in this dissertation, we have investigated three steps in bacterial CP assembly using computational and theoretical models. Specifically, we have elucidated molecular mechanisms which regulate HP dimerization and thus CP assembly. Also, we observe global conformational shifts, which likely act as a checkpoint factor in CP assembly, to prevent aberrant CP-like structures which can cause uncontrolled proteolysis. Lastly, our findings show a tradeoff between speed and robustness in hierarchal pathways, suggesting evolution favors speed over robustness in bacterial CP assembly. Ultimately, this work has helped advance our knowledge on bacterial CP assembly and

understand the role of allostery in hierarchal assemblies. In the future, our results can be applied to study how similar molecular mechanisms operate in other species of CP as well and test if these assembly mechanisms are conserved across different organisms.

6.1 References:

1. Burns KE, Pearce MJ, Darwin KH: Prokaryotic ubiquitin-like protein provides a two-part degron to Mycobacterium proteasome substrates. *J Bacteriol* 2010, 192(11):2933-2935.
2. Cheng Y, Pieters J: Novel proteasome inhibitors as potential drugs to combat tuberculosis. *J Mol Cell Biol* 2010, 2(4):173-175.
3. Lin G, Li D, Chidawanyika T, Nathan C, Li H: Fellutamide B is a potent inhibitor of the Mycobacterium tuberculosis proteasome. *Arch Biochem Biophys* 2010, 501(2):214-220.
4. Martins M, Viveiros M, Couto I, Amaral L: Targeting human macrophages for enhanced killing of intracellular XDR-TB and MDR-TB. *Int J Tuberc Lung Dis* 2009, 13(5):569-573.
5. Goldberg AL: Development of proteasome inhibitors as research tools and cancer drugs. *Journal of Cell Biology* 2012, 199(4):583-588.
6. Joazeiro CAP, Anderson KC, Hunter T: Proteasome Inhibitor Drugs on the Rise. *Cancer Research* 2006, 66(16):7840.
7. Kuhn DJ, Chen Q, Voorhees PM, Strader JS, Shenk KD, Sun CM, Demo SD, Bennett MK, van Leeuwen FWB, Chanan-Khan AA *et al*: Potent activity of carfilzomib, a novel, irreversible inhibitor of the ubiquitin-proteasome pathway, against preclinical models of multiple myeloma. *Blood* 2007, 110(9):3281-3290.
8. Manasanch EE, Orlowski RZ: Proteasome inhibitors in cancer therapy. *Nature Reviews Clinical Oncology* 2017, 14(7):417-433.
9. R. C. Kane ATF, R. Sridhara, and R. Pazdur: United States Food and Drug Administration approval summary: bortezomib for the treatment of progressive multiple myeloma after one prior therapy. *Clinical Cancer Research* 2006, 12:2955-2960.
10. Voorhees PM, Dees EC, O'Neil B, Orlowski RZ: The Proteasome as a Target for Cancer Therapy. *Clinical Cancer Research* 2003, 9(17):6316.
11. Zhang X LS, Bazzaro M: Drug Development Targeting the Ubiquitin–Proteasome System (UPS) for the Treatment of Human Cancers. *Cancers* 2020, 12(4).
12. Deeds EJ, Bachman JA, Fontana W: Optimizing ring assembly reveals the strength of weak interactions. *Proc Natl Acad Sci U S A* 2012, 109(7):2348-2353.
13. Hewarathna A, Mozziconacci O, Nariya MK, Kleindl PA, Xiong J, Fisher AC, Joshi SB, Middaugh CR, Forrest ML, Volkin DB *et al*: Chemical Stability of the Botanical Drug Substance Crofelemer: A Model System for Comparative Characterization of Complex Mixture Drugs. *J Pharm Sci* 2017, 106(11):3257-3269.
14. Nariya MK, Kim JH, Xiong J, Kleindl PA, Hewarathna A, Fisher AC, Joshi SB, Schöneich C, Forrest ML, Middaugh CR *et al*: Comparative Characterization of Crofelemer Samples Using Data Mining and Machine Learning Approaches With Analytical Stability Data Sets. *J Pharm Sci* 2017, 106(11):3270-3279.

Appendix A

Appendix for Chapter 2

Understanding the Separation of Timescales in *Rhodococcus erythropolis* Proteasome Core Particle Assembly

A.1 Propeptide comes out of the HP barrel in WT simulations

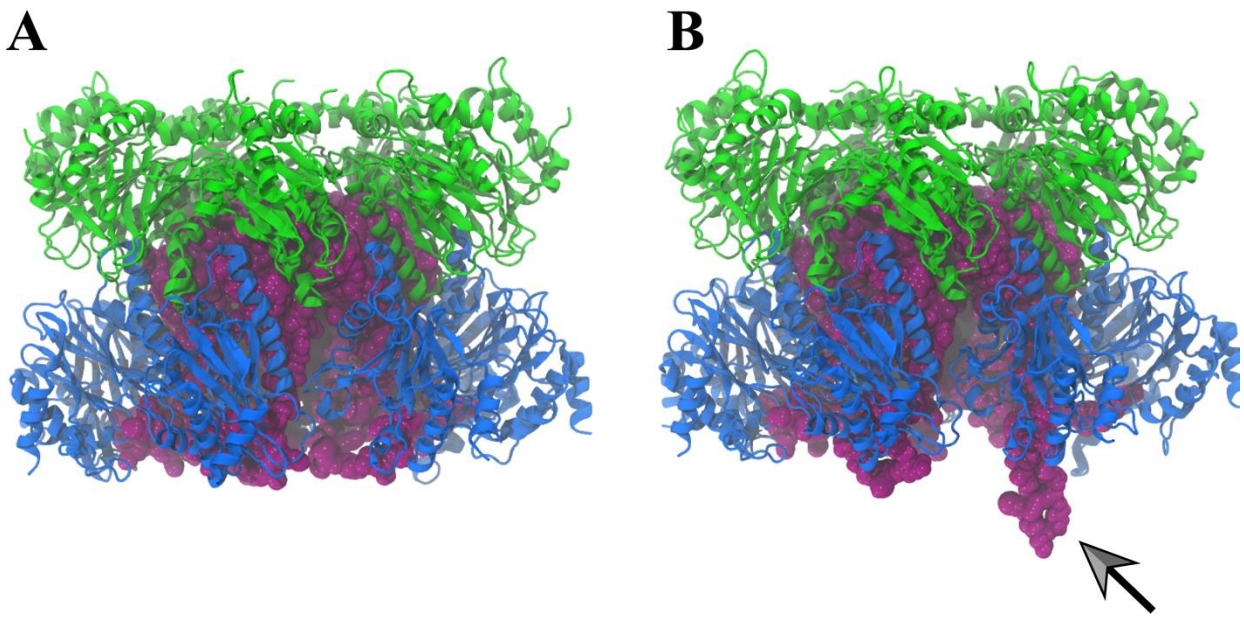


Figure A.1. Cartoon representation of the WT Half Proteasome from Molecular Dynamics simulations of *Re* bacterium. In both images, the α subunits are shown in green, β subunits in blue and the β propeptides in the purple spheres. (A) The WT HP at 996 ns. (B) WT HP at 1240.8 ns. The arrow serves to highlight the protruding propeptide. Both images were rendered using VMD.

A.2 Root Mean Square Fluctuation (Å) for the propeptide in WT, SLOW and FAST simulations

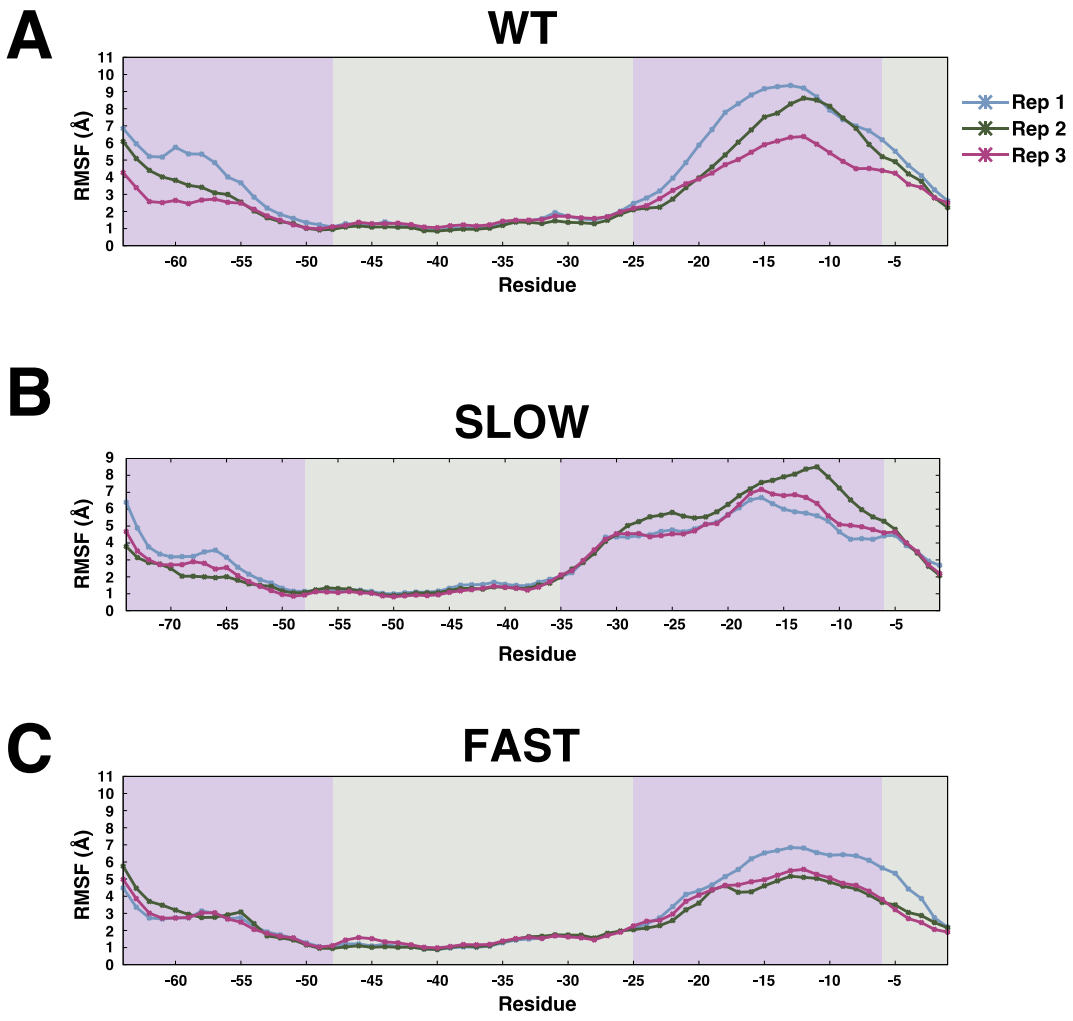


Figure A.2. Root Mean-Square Fluctuations (RMSF) of the β propeptide in WT, SLOW and FAST HPs. The panels show the average RMSF after 2.5 μs for all the backbone atoms of each residue of the β propeptides in (A) WT, (B) SLOW and (C) FAST HP for three independent Molecular Dynamics simulations. For all panels, electron density of the residues in crystal structure is depicted by the colored bars; missing electron density residues (purple) and residues with electron density (grey). The missing electron density residues were modeled by Rosetta. The RMSF values measured for Replicate Rep. 1 is shown by the blue curve, Rep. 2 by the green and Rep. 3 by the pink colors.

A.3 Potential energy profiles of the Anton simulations

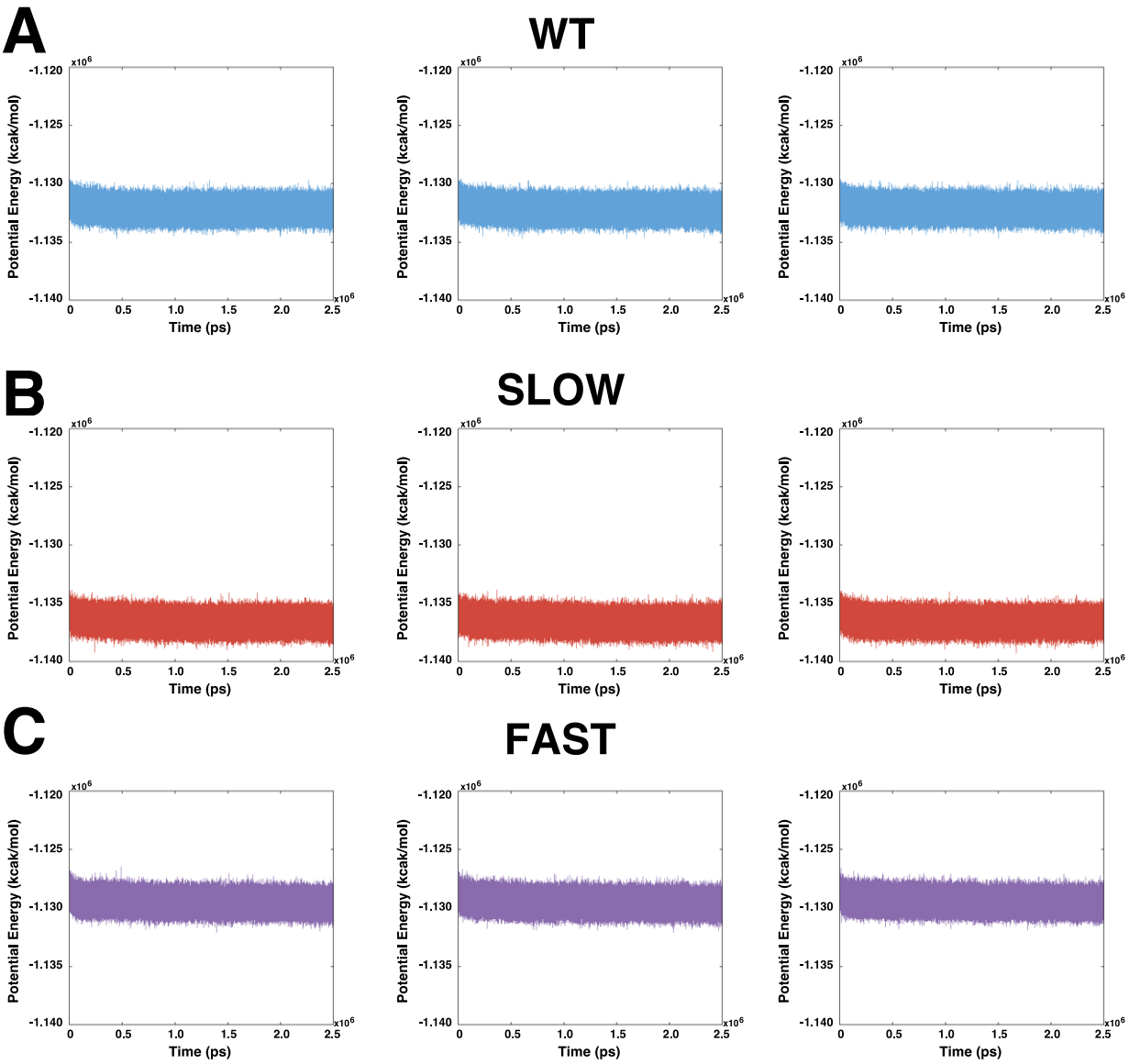


Figure A.3. Potential energy of each simulation. The potential energy of each 2.5 μs simulation on Anton with (A) WT, (B) SLOW and (C) FAST HPs. Plots show the potential energy for Rep. 1 (left), Rep. 2 (middle) and Rep. 3 (right) separately for each HP type. Every simulation took about 500 ns to converge.

A.4 RMSD (without propeptide) of all backbone atoms as a function of time

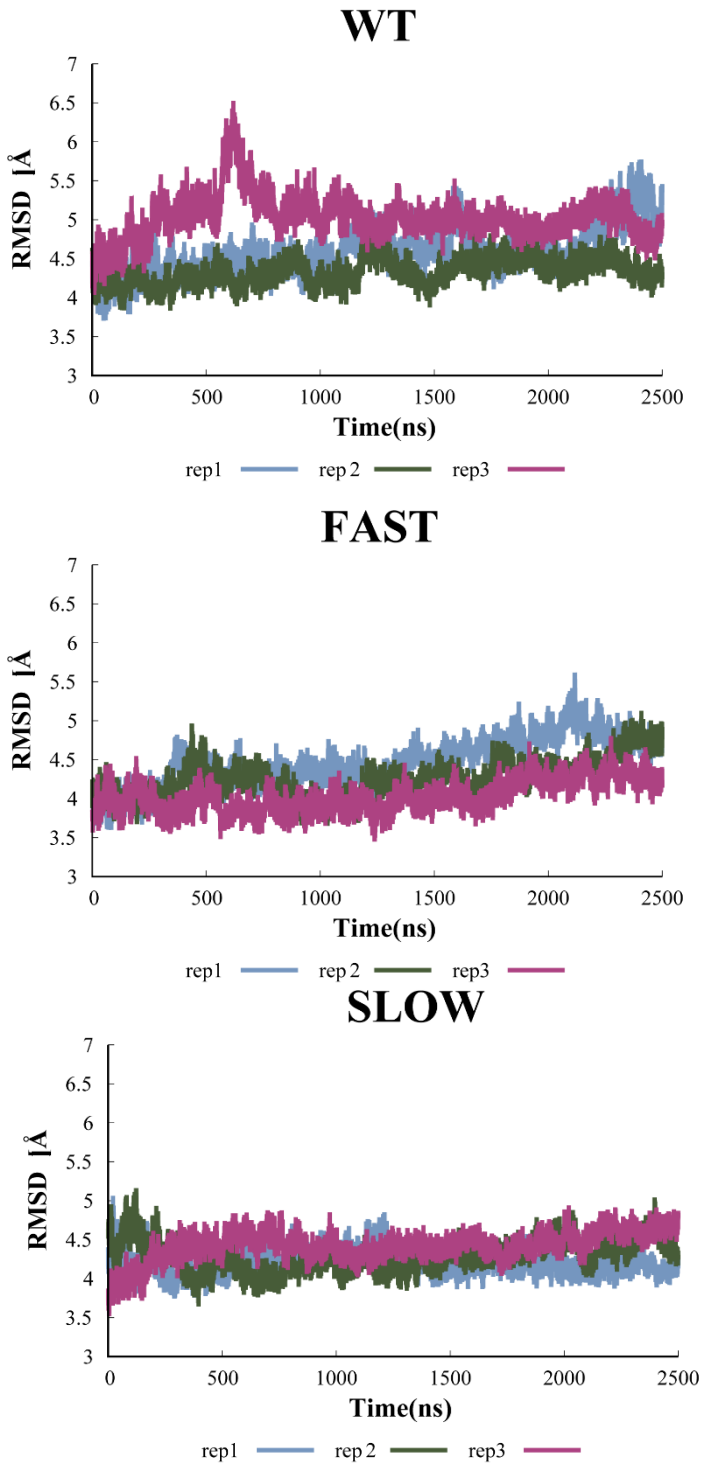


Figure A.4. RMSD for HP without the propeptide. Panels show the average RMSD for all seven β subunits of the HP without including the propeptide residues in the RMSD calculations at each time point for (A) WT, (B) SLOW and (C) FAST HPs. For comparison, all three replicates (blue, green, and pink) are shown on the same axes.

A.5 LOWESS plots of hydrogen bonds between propeptide and key residues

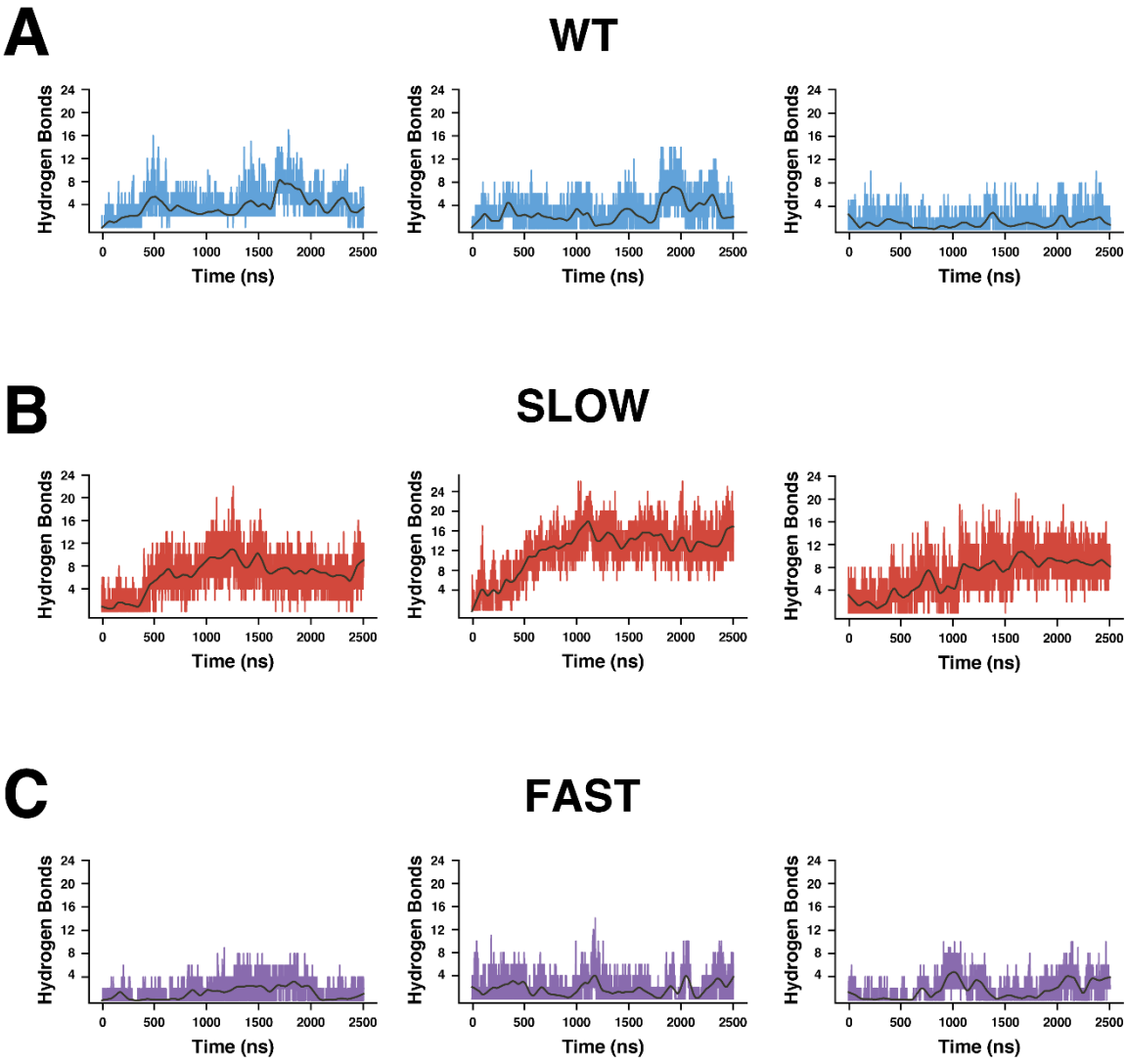


Figure A.5. LOWESS plots for HPs. Plots show the number of hydrogen bonds formed by the propeptide of (A) WT, (B) SLOW and (C) FAST HPs at each time point. Plots show the hydrogen bonds for Rep. 1 (left), Rep. 2 (middle) and Rep. 3 (right) separately for each HP type. Black line indicates the non-parametric LOWESS fit.

A.6 Violin distributions for all replicates of WT, SLOW and FAST

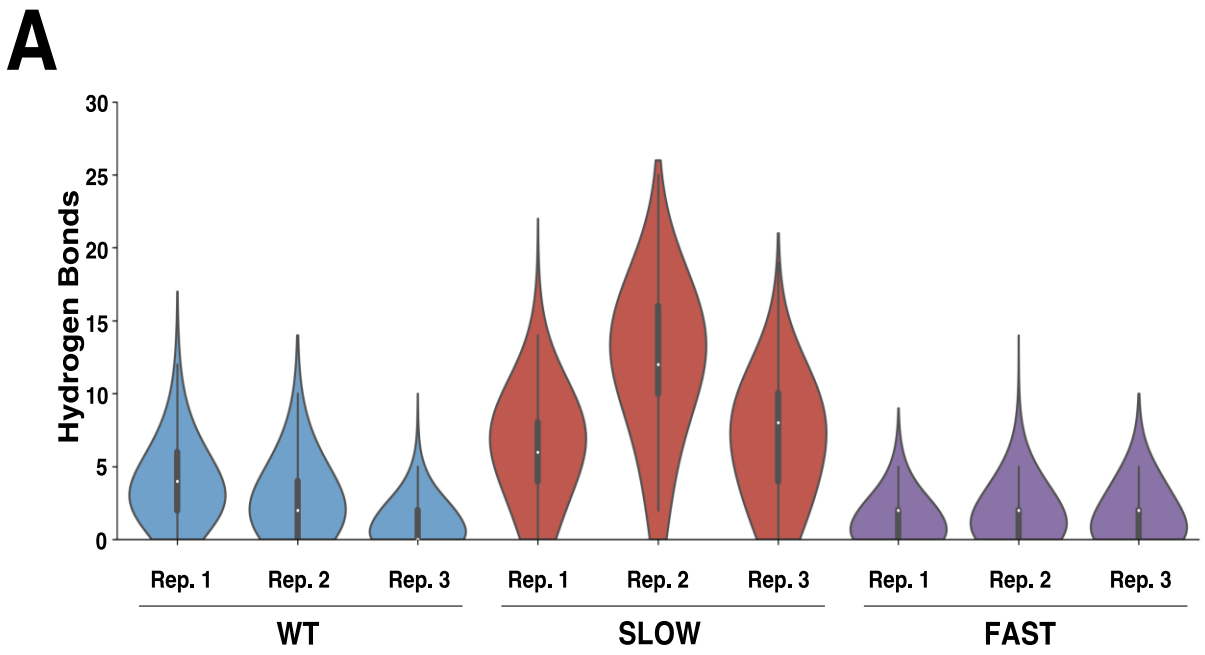


Figure A.6. Violin plots of hydrogen bonds formed by propeptide. Violin plots showing the total number of hydrogen bonds formed between the propeptide and key residues at the HP dimerization interface for each replicate of WT (blue), SLOW (red) and FAST (purple) HP.

A.7 *In vitro* reconstitution experiments / Experimental Methods

The α, Wild Type β and FAST β proteins were expressed and purified. All proteins were concentrated to 8uM using Amicon Ultra 10K centrifugal filters (Millipore) in an assembly buffer HNE (20mM HEPES, 100mM NaCl, 1mM EDTA, 5mM DTT, pH7.0). The α subunit was mixed with wild type β and FAST mutant β separately in equi-molar ratio to obtain a final subunit concentration of 4uM. Assembly reactions were allowed to proceed for 3hrs at 30°C. At the end of the 3hrs time course, equal volume of loading dye (0.8M HEPES, 0.1% Bromophenol Blue, 20% Glycerol) was added to the reactions. Samples were loaded on a 4-20% native gel (Invitrogen). Gels were run at 4°C, 120V for 12 hours, stained with Sypro Ruby (Thermofisher Scientific, catalog number S12000) protein stain as described by Thermofisher Scientific [1] in the manual, visualized using Biorad ChemiDoc imager and quantified using ImageLab software.

A.8 Kymographs for all replicates of WT, SLOW and FAST

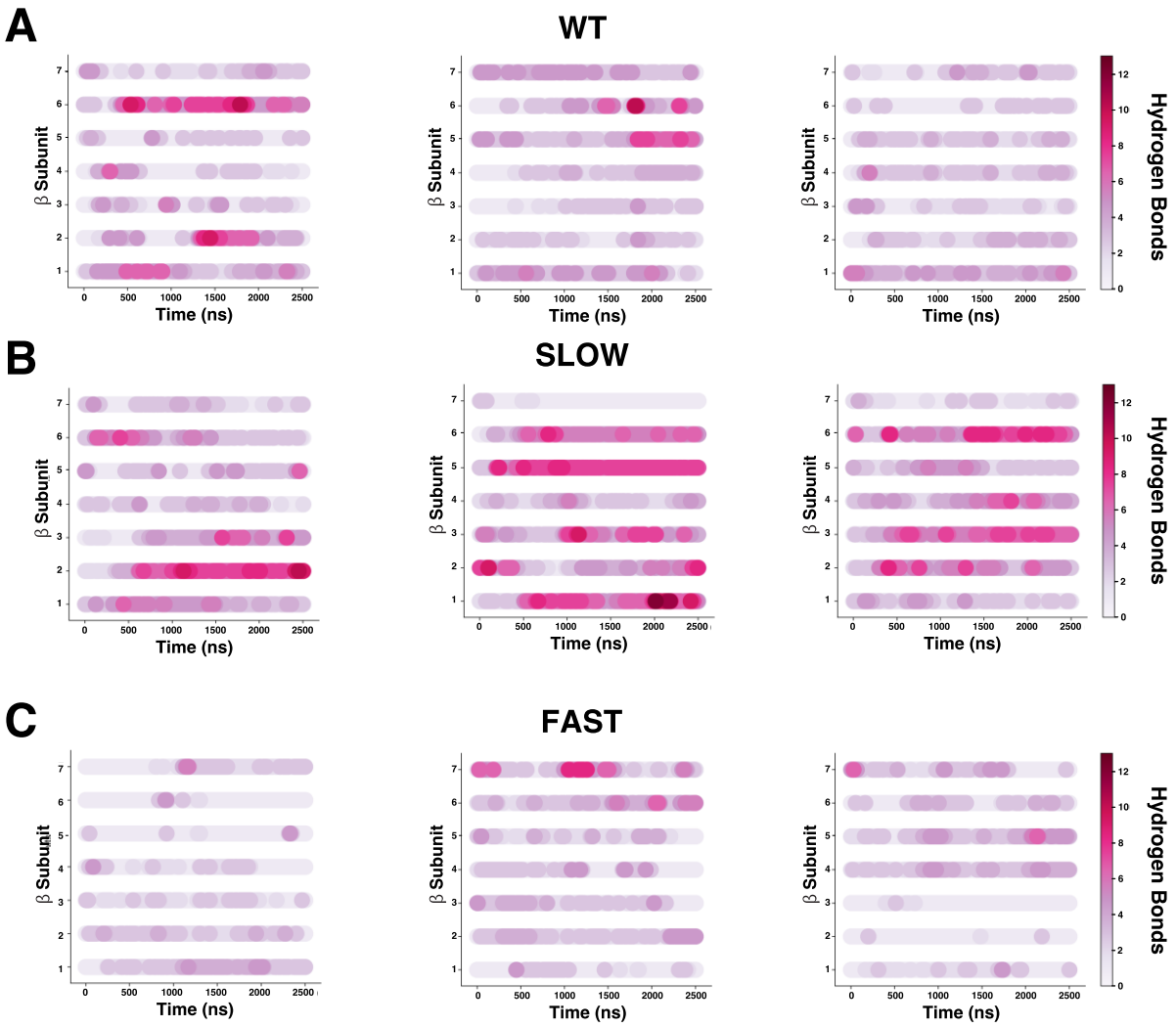


Figure A.7. Kymographs for WT, and mutants. The kymographs show the total number of hydrogen bonds formed between each β subunits and the key resides at each time point for (A) WT, (B) SLOW and (C) FAST HPs. Plots show the hydrogen bonds for Rep. 1 (left), Rep. 2 (middle) and Rep. 3 (right) separately for each simulation type.

A.9 Statistical Tables for categorical regression analysis

A.9.1. Model 1 statistical *p-values*: one intercept and one slope

Simulation	*Intercept p-value*	*Slope p-value*
WT-rep1	< 2.2e-16	1.65e-05
WT-rep2	2.0e-4	< 2.2e-16
WT-rep3	2.7e-4	1.73e-12
SLOW-rep1	< 2.2e-16	7.6e-4
SLOW-rep2	< 2.2e-16	2.05e-05
SLOW-rep3	< 2.2e-16	< 2.2e-16
FAST-rep1	< 2e-16	0.2072
FAST-rep2	6.77e-09	3.1e-04
FAST-rep3	1.37e-06	6.94e-09

Table A.9.1: Table for categorical regression p-values for the intercept and slope from Newey-West estimator fits for number of hydrogen bonds as a function of time (500ns to 2.5 μs).The model is of the for $y = \beta 0 + \beta 1.\ X + \varepsilon$ where y is the number of hydrogen bonds between propeptide residues and the set of key residues at HP dimerization interface, $\beta 0$ is the coefficient, $\beta 1\ \ is\ slope,$ and ε is the error term. The insignificant p-values are highlighted in grey.

The null hypothesis would be that the average number of hydrogen bonds that assist or delay CP assembly will not follow the same trend in longer simulations (seconds) as shown in our current μs simulations. The α threshold is 0.05 and only one replicate i.e., FAST mutant replicate 1 has an insignificant slope i.e., the num. of hydrogen bonds being formed will not increase with longer simulations and this results the slope being equal to 0. Also, all the SLOW mutant replicates are having very high significance with p-value less than 2.2 e-16. This indicates, that on an average the number of interactions between propeptide and key residues which delay dimerization are significantly in more numbers in SLOW than the FAST or WT simulations. Additionally, all the replicates of WT specially, rep2 and rep3 do not look very different from

the FAST mutant simulations, but if we see the *p-values* for WT slope they all are significant, and hence with more longer simulations all the WT replicates will look significantly different than FAST.

A.9.2 Model 2 statistical *p-values*: two intercepts and one slope

Wild-Type (WT)	Mutant	WT-Intercept *p-value*	WT & Mutant-Slope *p-value*	Mutant-Intercept *p-value*
WT-rep1	SLOW-rep1	3.92e-16	0.2527	< 2.2e-16
	SLOWrep2	< 2.2e-16	1.98e-05	< 2.2e-16
	SLOW-rep3	3.92e-16	< 2.2e-16	< 2.2e-16
	FAST-rep1	< 2.2e-16	3.86e-05	9.70e-08
	FAST-rep2	< 2.2e-16	9.70e-08	< 2.2e-16
	FAST-rep3	< 2.2e-16	5.66e-11	< 2.2e-16
WT-rep2	SLOW-rep1	< 2.2e-16	1.23e-06	< 2.2e-16
	SLOW-rep2	1.15e-03	< 2.2e-16	< 2.2e-16
	SLOW-rep3	0.3865	< 2.2e-16	< 2.2e-16
	FAST-rep1	< 2.2e-16	< 2.2e-16	< 2.2e-16
	FAST-rep2	< 2.2e-16	< 2.2e-16	< 2.2e-16
	FAST-rep3	< 2.2e-16	< 2.2e-16	6.14e-13
WT-rep3	SLOW-rep1	2.87e-14	0.3975	< 2.2e-16
	SLOW-rep2	0.3425	6.04e-10	< 2.2e-16
	SLOW-rep3	7.44e-14	< 2.2e-16	< 2.2e-16
	FAST-rep1	1.67e-10	3.79e-07	7.75e-07
	FAST-rep2	3.48e-04	8.59e-11	2.59e-10
	FAST-rep3	0.0479	3.614e-16	< 2.2e-16

Table A.9.2: Table for categorical regression p-values for the intercepts and slope from Newey-West estimator fits for the number of hydrogen bonds as a function of time (500ns to 2.5 μs). The model is of the for $y = \beta 0 + \beta 1.\ X + B2.C + \varepsilon$ where y is the number of hydrogen bonds between propeptide residues and the set of key residues at HP dimerization interface, $\beta 0$ is the coefficient for WT, $\beta 1$ is the mutant coefficient of intercept, $\beta 2$ is the combined slope, ε is the error term and C is the categorical variable. The simulations whose p-value slope are not significant is highlighted in grey background.

The null hypothesis would be that on average the WT simulations are not different from the SLOW and FAST mutants. The α threshold is 0.05 and in Table A.9.2 three cases the intercept is insignificant and two cases the slope. Thus, for cases where intercept is insignificant, we observe that the slope is significant and

thus with longer simulations we have enough evidence that they will be different. If the slope is insignificant i.e., the simulation has already reached equilibrium and the number of hydrogen bonds as shown in here would not change with longer simulations. Overall, we found no replicates in **Table A9.2** where both intercept and slope are insignificant, hence we have sufficient evidence that the WT replicates are different from SLOW and FAST.

A.9.3 Model 3 statistical *p-values:* two intercepts and two slopes

Wild-Type (WT)	Mutant	WT- Intercept *p-value*	WT-Slope *p-value*	Mutant-Intercept *p-value*	Mutant-Slope *p-value*
WT-rep1	SLOW-rep1	< 2.2e-16	5.49e-05	< 2.2e-16	3.65e-06
	SLOW-rep2	< 2.2e-16	5.05e-05	< 2.2e-16	0.5556
	SLOW-rep3	< 2.2e-16	5.47e-05	0.1767	< 2.2e-16
	FAST-rep1	< 2.2e-16	5.50e-05	1.142e-10	0.0053
	FAST-rep2	< 2.2e-16	6.50e-05	1.410e-11	0.3192
	FAST-rep3	< 2.2e-16	< 5.09e-05	4.706e-12	0.8151
WT-rep2	SLOW-rep1	0.0004	< 2.2e-16	< 2.2e-16	< 2.2e-16
	SLOW-rep2	0.0003	< 2.2e-16	< 2.2e-16	0.0072
	SLOW-rep3	0.0004	< 2.2e-16	4.71e-16	6.28e-07
	FAST-rep1	0.0004	< 2.2e-16	0.0104	9.26e-13
	FAST-rep2	0.0004	< 2.2e-16	0.2020	5.27e-07
	FAST-rep3	0.0003	< 2.2e-16	0.3905	0.0001
WT-rep3	SLOW-rep1	0.0043	1.86e-11	< 2.2e-16	1.12e-07
	SLOW-rep2	0.0041	1.57e-11	< 2.2e-16	0.2489
	SLOW-rep3	0.0043	1.85e-11	< 2.2e-16	< 2.2e-16
	FAST-rep1	0.0043	1.90e-11	9.07e-08	0.0009
	FAST-rep2	0.0046	2.68e-11	0.0013	0.4784
	FAST-rep3	0.0042	1.63e-11	0.0099	0.3358

Table A.9.3: Table for categorical regression p-values for the intercepts and slope from Newey-West estimators for the number of hydrogen bonds as a function of time (500ns to 2.5 μs). The model is of the form $y = \beta 0 + \beta 1.\ X + B2.C + B3.C.X + \varepsilon$ where y is the number of hydrogen bonds between propeptide residues and the set of key residues at HP dimerization interface, $\beta 0$ is the coefficient for WT, $\beta 1$ is the slope of WT, $\beta 2$ is the intercept of mutant, $\beta 3$ is the slope of mutant, ε is the error term and C is the categorical variable. The simulations whose p-values are not significant are highlighted in grey background.

The null hypothesis would be that on average the WT simulations are not different from the SLOW and FAST mutants. The α threshold is 0.05 and we observed in three cases the intercept is insignificant and five cases the slope. Thus, for cases where intercept is insignificant, has the slope very significant (Ex: WT-rep1 and SLOW-rep3) and hence with longer simulations we have enough evidence that WT and mutants will remain different. If the slope is insignificant (Ex: WT-rep1 and SLOW-rep2) i.e., the simulation (mutant) has already reached equilibrium and the number of hydrogen bonds in our μs's simulations would not change with time. Overall, we found no replicates where both intercept and slope are insignificant, hence we have sufficient evidence that the WT replicates are different from SLOW and FAST, and the statistics show significant evidence and are drawn from different distributions.

A.10 MD simulations systems details

Simulation	Box size (Å)	Number of atoms	Number of water molecules	Ions
WT	155 X 155 X 155	352979	298725	Na^+ = 275 Cl^- = 184
SLOW	155 X 155 X 155	353567	298515	Na^+ = 275 Cl^- = 184
FAST	155 X 155 X 155	345648	291465	Na^+ = 257 Cl^- = 180

Table A.10: System properties and details of the WT, FAST, and SLOW simulations. All the simulations are run in a rectangular water box with 15 Å water on each side of the protein.

A.11 Additional MD simulations details.

Proteins are described by CHARMM36 force field and explicit water was modeled with the CHARMM version of the TIP3P water model [2]. All the simulations are explicit solvent and TIP3P water model with CHARMM version [3]. In the Anton2 simulations, integration was carried out using the Multigrator algorithm [4] with a 2.5 fs time step. For the RESPA scheme every second time step was used for long range interactions. Pressure was controlled using the Martyna-Tobias-Klein (MTK) barostat [5] and with an interval length of 480 ps. The temperature was maintained by the Nose-Hoover thermostat [6] with an interval length of 24 ps. A relaxation time of τ=0.041667 ps was used for barostat and thermostat.

A.12 References

1. SYPRO Ruby Protein Gel Stain [https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2Fmp12000.pdf&title=U1lQUk8gUnVieSBQcm90ZWluIEdlbCBTdGFpbg==]
2. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S *et al*: All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 1998, 102(18):3586-3616.
3. Best RB, Zhu X, Shim J, Lopes PE, Mittal J, Feig M, Mackerell AD, Jr.: Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J Chem Theory Comput* 2012, 8(9):3257-3273.
4. Lippert RA, Predescu C, Ierardi DJ, Mackenzie KM, Eastwood MP, Dror RO, Shaw DE: Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure. *The Journal of Chemical Physics* 2013, 139(16):164106.
5. Martyna GJ, Tobias DJ, Klein ML: Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* 1994, 101(5):4177-4189.
6. Nosé S: A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics* 1984, 52(2):255-268.

APPENDIX B

Appendix for Chapter3

Global conformational shifts act as a checkpoint in bacterial Core Particle assembly

B.1 Cartoon representation of near-HP intermediates structures

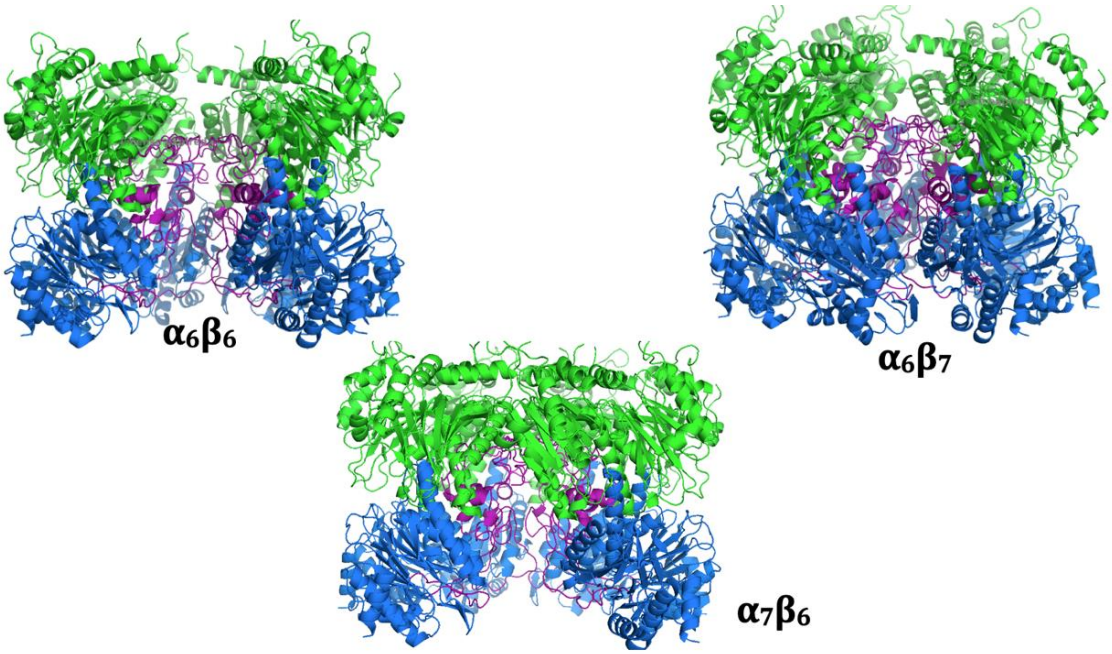


Figure B.1: Cartoon representations of the three near-HP intermediates, which are simulated for 2.5 μs. The three models were built using the crystal structure of the WT proteasome, with its propeptide present, as a starting point. These were developed by starting from our previous WT HP simulations and removing the relevant subunits. These are $\alpha_6\beta_7$ and $\alpha_7\beta_6$ (i.e., HPs missing just one α or β subunit), $\alpha_6\beta_6$ (an HP missing an entire α/β dimer). All α subunits are shown in green and β subunits are blue, and the propeptide is purple.

B.2 Propeptide Region I RMSD shown as violin distributions

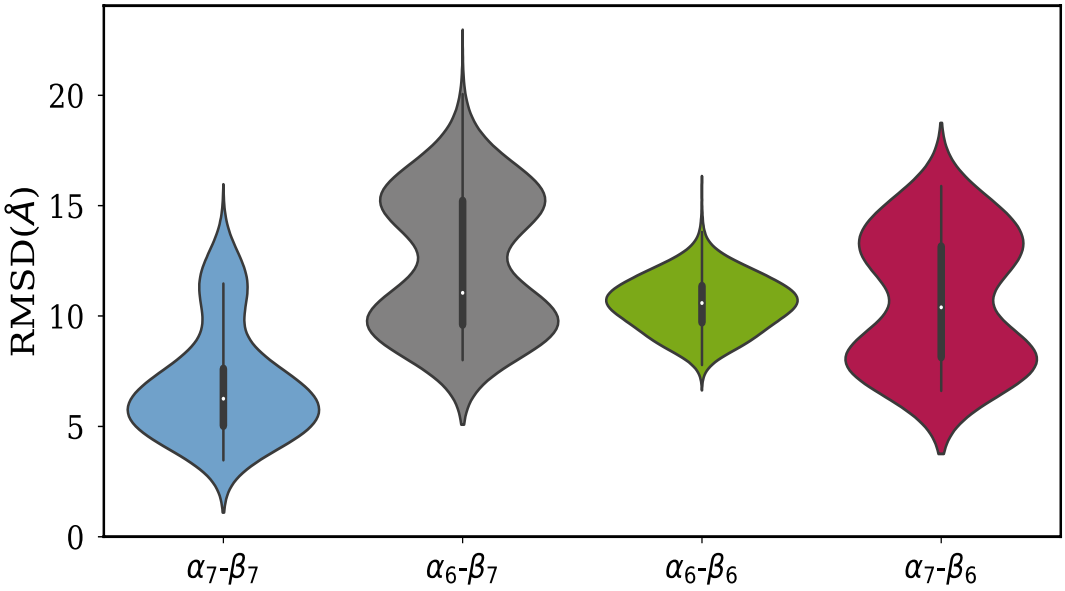


Figure B.2: Region I RMSD shown as violin plots. Each violin represents each intermediate and has observations for simulations from all the three replicates (from 500 ns to 2.5 μs) combined into one violin.

B.3.1 Method for calculating the βθ

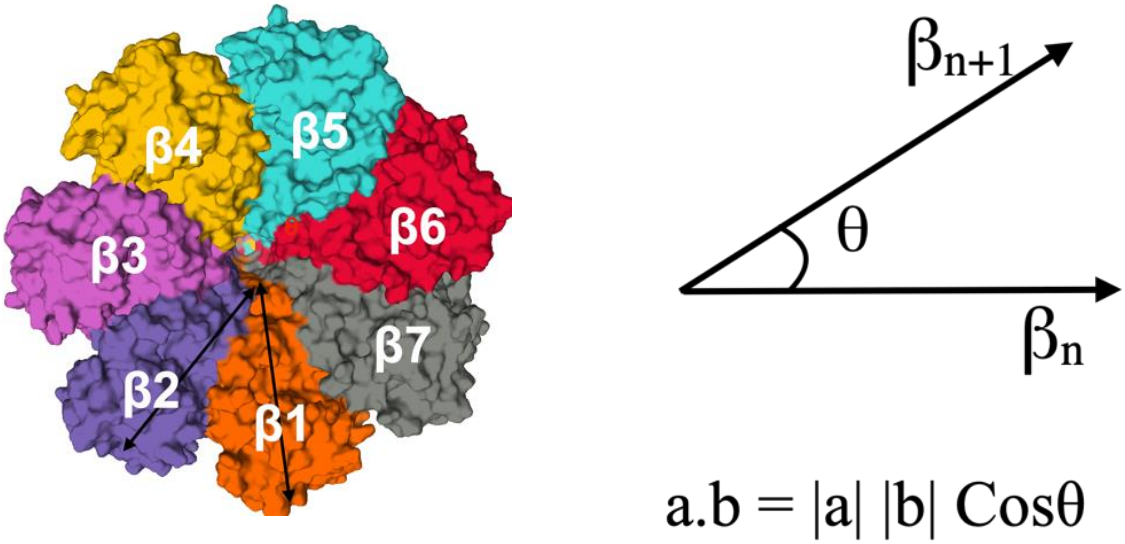


Figure B.3.1: The method for calculating βθ of the β subunits. All the seven β subunits of the CP are colored differently. Every angle is calculated between two β subunits and the Center of Mass of the β ring.

B.3.2 Method for calculating the βØtilt

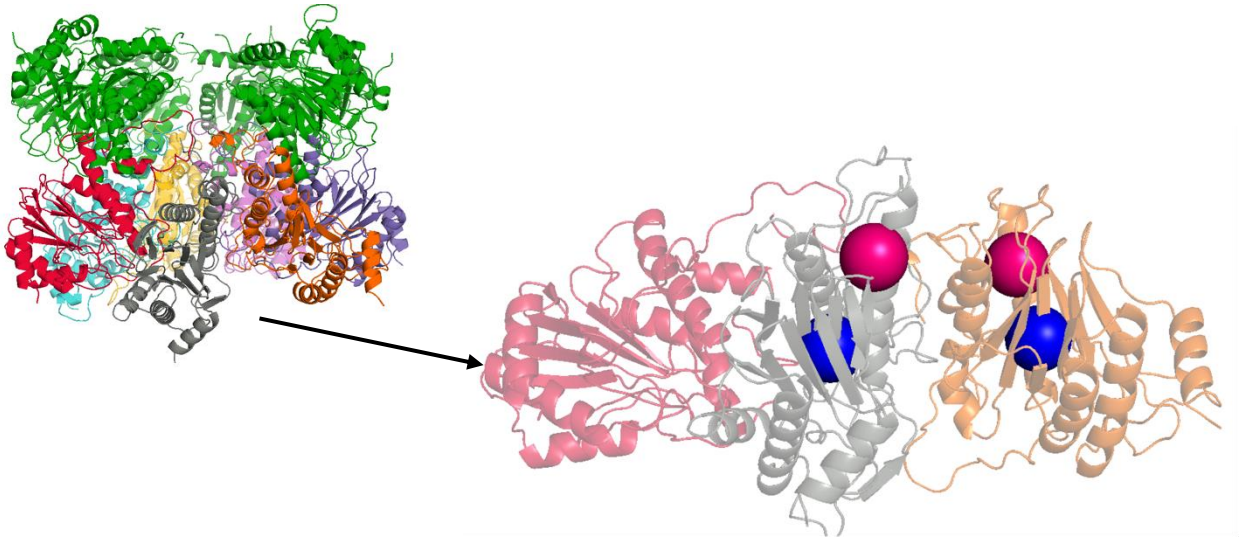


Figure B.3.2: Method to calculate the dihedral angle βØtilt made by every β subunit in the ring. The $\alpha_6\beta_7$ intermediate is shown in cartoon representation. With alpha subunits in green, and β subunits in different colors. The zoom in picture shows only three β subunits β_6(red), β_7(grey) and β_1(orange). The blue spheres represent the center of mass of the β subunits, and the magenta spheres represent the center of mass of the H1 helix of β subunits.

B.4 βθ as a function of time for the HP and intermediate simulations

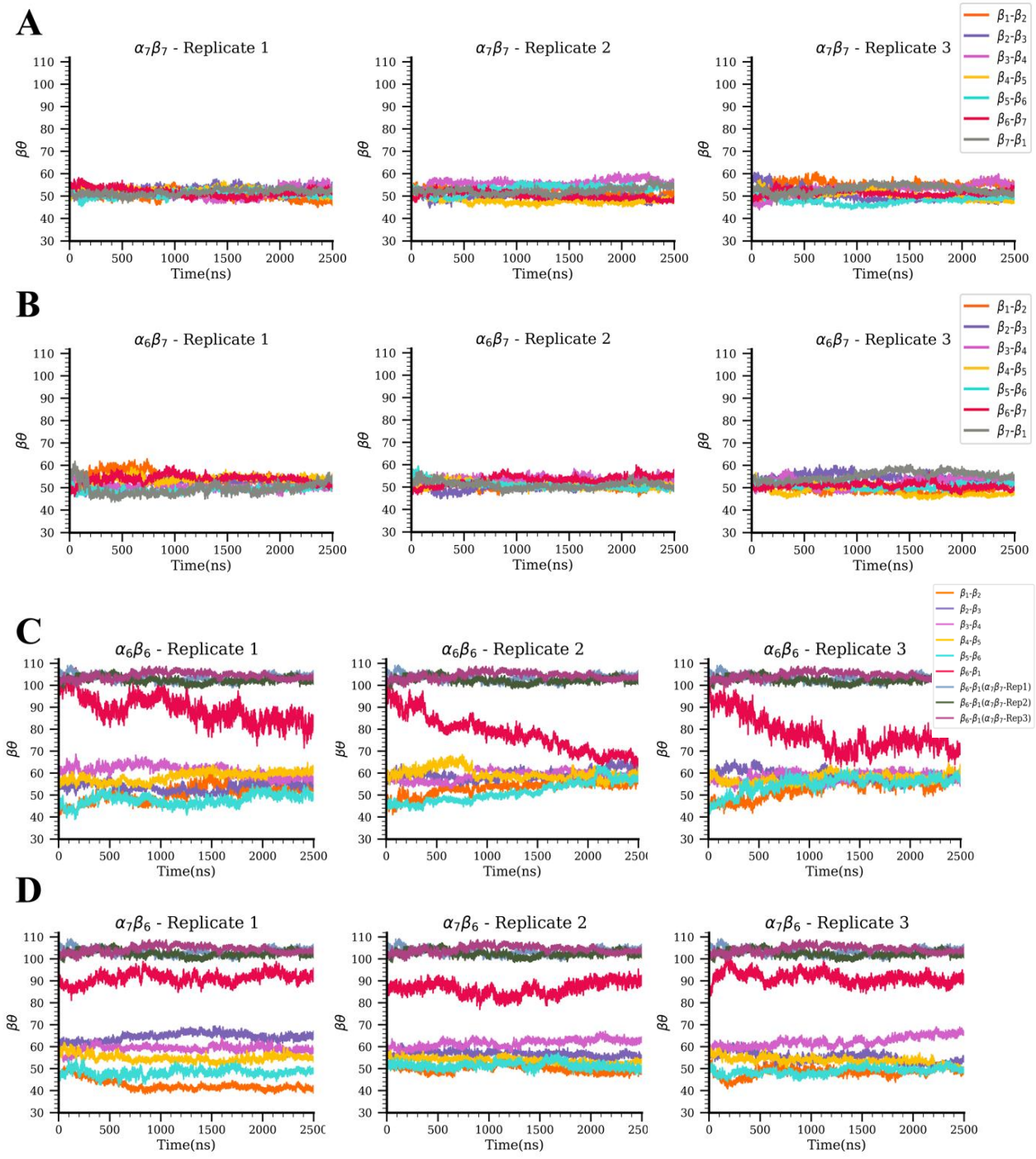


Figure B.4: The values of βθ as a function of time for all subunits of the β ring. A) HP-$\alpha_7\beta_7$ B) $\alpha_6\beta_7$ C) $\alpha_6\beta_6$ and D) $\alpha_7\beta6$. Each angle for the β subunits is in a different color, and the $\alpha_6\beta$, and $\alpha_7\beta6$ simulations have the βθ values from HP simulations for comparison.

B.5 Statistics Tables for βθ categorical regression

B.5.1 $\alpha_7\beta_7$ and $\alpha_6\beta_7$

		Replicate 1				Replicate 2				Replicate 3			
		$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_7$ Intercept	$\alpha_6\beta_7$ Slope	$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_7$ Intercept	$\alpha_6\beta_7$ Slope	$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_7$ Intercept	$\alpha_6\beta_7$ Slope
$\alpha_7\beta_7$ Replicate 1	$\alpha_6\beta_7$ -β_1	2.20E-16	3.50E-13	0.05809	0.01182	2.20E-16	1.15E-13	0.0003748	4.23E-06	2.20E-16	3.36E-14	3.67E-16	0.7206
	$\alpha_6\beta_7$ -β_2	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_3	1.05E-09	2.20E-16	2.20E-16	2.20E-16	2.36E-09	2.20E-16	2.20E-16	2.20E-16	9.97E-12	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_4	2.20E-16	0.08509	1.06E-13	2.20E-16	2.20E-16	0.07316	0.34101	1.02E-08	2.20E-16	0.0911237	2.20E-16	0.0004222
	$\alpha_6\beta_7$ -β_5	2.20E-16	0.000105	2.20E-16	2.20E-16	2.20E-16	7.31E-05	2.20E-16	0.0658	2.20E-16	0.000108	2.20E-16	2.30E-10
	$\alpha_6\beta_7$ -β_6	2.00E-16	2.00E-16	2.00E-16	0.01212	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_7	2.20E-16	2.20E-16	1.51E-07	0.4039	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.47E-09
$\alpha_7\beta_7$ Replicate 2	$\alpha_6\beta_7$ -β_1	2.20E-16	2.20E-16	2.20E-16	5.44E-09	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_2	2.20E-16	7.60E-05	6.72E-15	0.9335	2.20E-16	5.42E-05	2.20E-16	0.0189	2.20E-16	3.69E-05	8.48E-05	0.1063
	$\alpha_6\beta_7$ -β_3	2.20E-16	4.07E-07	2.20E-16	3.50E-08	2.20E-16	2.07E-07	0.01872	2.20E-16	2.20E-16	1.02E-07	2.20E-16	6.29E-14
	$\alpha_6\beta_7$ -β_4	0.5295	<2e-16	<2e-16	<2e-16	0.538	<2e-16	<2e-16	<2e-16	0.4835	<2e-16	<2e-16	<2e-16
	$\alpha_6\beta_7$ -β_5	2.00E-16	0.09364	2.00E-16	2.00E-16	2.20E-16	0.08266	2.20E-16	1.61E-14	2.20E-16	0.09947	6.29E-11	0.03952
	$\alpha_6\beta_7$ -β_6	2.20E-16	0.001564	2.20E-16	2.20E-16	2.20E-16	0.001279	9.36E-13	0.576102	2.20E-16	0.001587	1.46E-07	1.37E-10
	$\alpha_6\beta_7$ -β_7	2.20E-16	2.20E-16	9.20E-13	0.617	2.20E-16	2.20E-16	1.06E-13	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.33E-10
$\alpha_7\beta_7$ Replicate 3	$\alpha_6\beta_7$ -β_1	2.20E-16	2.20E-16	2.20E-16	1.36E-09	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_2	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.23E-14	2.20E-16	0.08289	2.20E-16	8.16E-14	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_3	2.20E-16	4.07E-13	0.554169	0.0001723	2.20E-16	1.18E-13	1.66E-07	3.86E-08	2.20E-16	2.68E-14	6.90E-12	0.5251
	$\alpha_6\beta_7$ -β_4	6.76E-11	0.009707	7.48E-14	1.34E-05	2.15E-11	0.007969	0.00112	1.53E-14	6.28E-12	0.00645	2.20E-16	2.58E-09
	$\alpha_6\beta_7$ -β_5	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_6	2.00E-16	0.03878	2.00E-16	2.00E-16	2.20E-16	0.03083	2.20E-16	1.53E-14	2.00E-16	0.04294	2.00E-16	0.71694
	$\alpha_6\beta_7$ -β_7	2.20E-16	1.86E-07	2.20E-16	0.1274	2.20E-16	3.59E-07	0.7486	2.20E-16	2.20E-16	5.78E-11	6.47E-12	2.20E-16

All are *p-values*, with a threshold **α = 1.98 e-4,** Bonferroni corrected.

Table B.5.1: *p-values* of 256 categorical regression tests for βθ as a function of time all seven β subunits in $\alpha_7\beta_7$ and $\alpha_6\beta_7$. The *p-values* with insignificant intercept are marked in yellow, *p-values* with insignificant slope are marked in blue, and in red if both intercept and slope are insignificant. The null hypothesis is rejected for tests whose *p-values* are marked in red.

We observed only two cases where βθ values of both the intercept and slope of $\alpha_6\beta_7$ were not significantly different from the $\alpha_7\beta_7$ HP simulations (red highlighted). We also see many *p-values* with an insignificant slope (blue), indicating that the βθ values will not change with time and have the simulations converged. The cases with insignificant *p-values* of slope (yellow) all have the corresponding slope *p-value* significant, thus for these cases with extended simulations, the βθ values of intercept will be significant.

B.5.2 $\alpha_7\beta_7$ and $\alpha_6\beta_6$

				Replicate 1				*Replicate 2*				*Replicate 3*	
		$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_6$ Intercept	$\alpha_6\beta_6$ Slope	$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_6$ Intercept	$\alpha_6\beta_6$ Slope	$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_6$ Intercept	$\alpha_6\beta_6$ Slope
$\alpha_7\beta_7$ Replicate 1	$\alpha_6\beta_6$ -β_1	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_2	2.20E-16	6.36E-09	2.82E-10	0.2941	2.20E-16	2.80E-08	2.20E-16	2.20E-16	2.20E-16	2.06E-08	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_3	2.20E-16	8.29E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	4.65E-06	<2e-16	<2e-16	<2e-16	0.7471
	$\alpha_6\beta_6$ -β_4	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.05E-13	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_5	2.20E-16	2.20E-16	2.20E-16	9.48E-10	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.00E-16	2.00E-16	0.01769	2.00E-16
	$\alpha_6\beta_6$ -β_6	2.00E-16	0.02085	2.00E-16	2.00E-16	2.00E-16	0.03349	2.00E-16	2.00E-16	<2e-16	0.0409	<2e-16	<2e-16
$\alpha_7\beta_7$ Replicate 2	$\alpha_6\beta_6$ -β_1	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_2	2.20E-16	2.62E-16	2.20E-16	2.20E-16	2.20E-16	6.75E-15	2.20E-16	2.20E-16	2.20E-16	3.33E-15	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_3	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.00E-16	2.00E-16	2.00E-16	0.01396	2.00E-16	2.00E-16	2.00E-16	0.05132
	$\alpha_6\beta_6$ -β_4	2.20E-16	9.39E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	5.09E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_5	<2e-16	<2e-16	<2e-16	0.5565	2.20E-16	2.20E-16	2.20E-16	2.20E-16	<2e-16	<2e-16	0.3261	<2e-16
	$\alpha_6\beta_6$ -β_6	2.20E-16	1.44E-11	2.20E-16	2.20E-16	2.20E-16	6.40E-10	2.20E-16	2.20E-16	2.20E-16	1.01E-09	2.20E-16	2.20E-16
$\alpha_7\beta_7$ Replicate 3	$\alpha_6\beta_6$ -β_1	2.20E-16	2.30E-07	2.20E-16	2.20E-16	2.20E-16	7.54E-06	2.20E-16	2.20E-16	2.20E-16	3.03E-08	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_2	<2e-16	<2e-16	0.1728	<2e-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.48E-05
	$\alpha_6\beta_6$ -β_3	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_4	2.20E-16	2.20E-16	2.20E-16	2.20E-16	<2e-16	<2e-16	<2e-16	0.724	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_5	2.20E-16	0.0007137	5.71E-10	4.38E-14	2.20E-16	0.000167	2.20E-16	2.20E-16	2.20E-16	0.000105	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_6	2.00E-16	0.01766	2.00E-16	2.00E-16	2.00E-16	0.02978	2.00E-16	2.00E-16	2.00E-16	0.03618	2.00E-16	2.00E-16

All are p-values, with a threshold **α = 2.3 e-4,** Bonferroni corrected.

Table B.5.2: *p-values* of 216 categorical regression tests for βθ as a function of time all seven β subunits in $\alpha_7\beta_7$ and $\alpha_6\beta_6$. The *p-values* with insignificant intercept are marked in yellow, *p-values* with insignificant slope are marked in blue, and in red if both intercept and slope are insignificant. The null hypothesis is not rejected for any tests in this table.

We observed no cases where *p-values* of both intercept and the slope are above our threshold of $2.3e^{-4}$. Thus, all the βθ values in all the β subunits of $\alpha_6\beta_6$ are significantly different from $\alpha_7\beta_7$. We also see many *p-values* with an insignificant slope which indicates that the βθ values will not change with time and have converged. The cases with insignificant *p-values* of slope (yellow) all have the corresponding slope value significant, thus for these cases with extended simulations, the βθ values of intercept will be significant.

B.5.3 $\alpha_7\beta_7$ and $\alpha_7\beta_6$

We observed only two cases where βθ values of both the intercept and slope of $\alpha_7\beta_6$ were not significantly different from the $\alpha_7\beta_7$-HP simulations (red highlighted). We also see many *p-values* with an insignificant slope(blue), indicating that the βθ values will not change with time and have converged. The cases with insignificant *p-values* of slope (yellow) all have the corresponding slope value significant, thus for these cases with extended simulations, the βθ values of intercept will be significant. Compared to the βθ values of $\alpha_6\beta_7$, the $\alpha_7\beta_6$ intermediate has only four cases of insignificant slope, which depicts that for our simulations βθ distributions fluctuate more in $\alpha_7\beta_6$ than $\alpha_6\beta_7$ compared to the HP.

		Replicate 1				*Replicate 2*				*Replicate 3*			
		$\alpha_7\beta_7$ Intercept	**$\alpha_7\beta_7$ Slope**	**$\alpha_7\beta_6$ Intercept**	**$\alpha_7\beta_6$ Slope**	**$\alpha_7\beta_7$ Intercept**	**$\alpha_7\beta_7$ Slope**	**$\alpha_7\beta_6$ Intercept**	**$\alpha_7\beta_6$ Slope**	**$\alpha_7\beta_7$ Intercept**	**$\alpha_7\beta_7$ Slope**	**$\alpha_7\beta_6$ Intercept**	**$\alpha_7\beta_6$ Slope**
$\alpha_7\beta_7$ Replicate 1	**$\alpha_7\beta_6$ -β_1**	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.00E-16	2.00E-16	1.11E-02	1.33E-03	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_2	2.20E-16	5.58E-07	2.20E-16	3.00E-11	2.20E-16	7.84E-07	2.20E-16	2.20E-16	2.20E-16	5.09E-07	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_3	2.20E-16	4.53E-15	2.20E-16	1.58E-05	2.20E-16	2.27E-15	2.20E-16	3.86E-09	2.20E-16	4.27E-16	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_4	<2e-16	<2e-16	<2e-16	5.31E-01	2.20E-16	2.20E-16	2.20E-16	9.10E-08	2.20E-16	2.20E-16	2.20E-16	4.97E-14
	$\alpha_7\beta_6$ -β_5	2.20E-16	2.20E-16	2.20E-16	0.0002738	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	5.04E-06
	$\alpha_7\beta_6$ -β_6	2.00E-16	3.96E-02	2.00E-16	2.00E-16	2.00E-16	2.83E-02	2.00E-16	2.00E-16	2.20E-16	0.03832	2.20E-16	2.30E-10
$\alpha_7\beta_7$ Replicate 2	**$\alpha_7\beta_6$ -β_1**	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.00E-16	2.00E-16	2.00E-16	0.00237	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_2	2.20E-16	2.92E-12	2.20E-16	2.20E-16	2.20E-16	5.73E-12	2.20E-16	0.01229	2.20E-16	2.51E-12	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_3	2.20E-16	2.20E-16	2.20E-16	3.12E-09	2.20E-16	2.20E-16	2.20E-16	3.71E-05	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_4	<2e-16	<2e-16	<2e-16	0.8068	2.20E-16	5.46E-15	2.20E-16	2.28E-10	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_5	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	0.0007524	2.20E-16	2.20E-16	2.20E-16	2.20E-16	5.70E-06
	$\alpha_7\beta_6$ -β_6	2.20E-16	1.94E-09	2.20E-16	2.20E-16	2.20E-16	1.86E-10	2.20E-16	2.20E-16	2.20E-16	1.46E-09	2.20E-16	5.47E-07
$\alpha_7\beta_7$ Replicate 3	**$\alpha_7\beta_6$ -β_1**	2.20E-16	1.01E-05	2.20E-16	2.20E-16	2.20E-16	1.59E-05	2.20E-16	2.20E-16	2.20E-16	1.24E-05	2.20E-16	0.04389
	$\alpha_7\beta_6$ -β_2	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	0.0007092
	$\alpha_7\beta_6$ -β_3	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.03E-07
	$\alpha_7\beta_6$ -β_4	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.56E-10	2.20E-16	2.20E-16	2.20E-16	0.006186
	$\alpha_7\beta_6$ -β_5	2.20E-16	0.001181	0.074614	0.645465	2.20E-16	0.0015858	2.20E-16	0.0003044	2.20E-16	0.001146	0.822381	2.75E-10
	$\alpha_7\beta_6$ -β_6	2.20E-16	0.03534	2.20E-16	2.45E-11	2.20E-16	0.02487	2.20E-16	6.35E-10	2.00E-16	0.03403	2.00E-16	2.00E-16

All are p-values, with a threshold **α = 2.3 e-4,** Bonferroni corrected.

Table B.5.3: *p-values* of 216 categorical regression tests for βθ as a function of time all seven β subunits in $\alpha_7\beta_7$ and $\alpha_7\beta_6$. The *p-values* with insignificant intercept are marked in yellow, *p-values* with insignificant slope are marked in blue, and in red if both intercept and slope are insignificant. The null hypothesis is rejected for tests whose *p-values* are marked in red.

B.6 βØtilt as a function of time for the HP($\alpha_7\beta_7$) and intermediate simulations

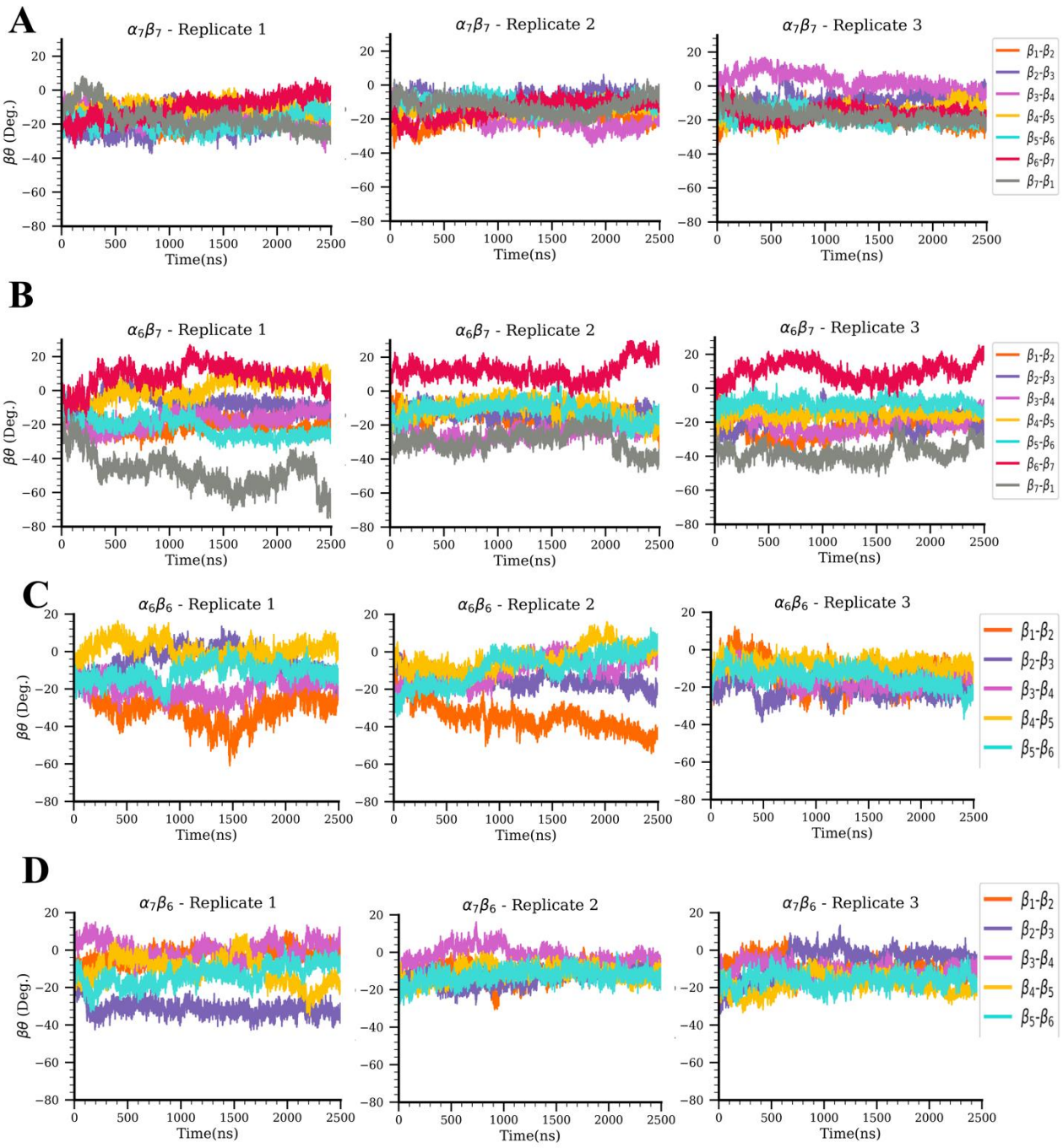


Figure B.6: The values of βØtilt as a function of time for all subunits of the β ring. A) HP-$\alpha_7\beta_7$ B) $\alpha_6\beta_7$ C) $\alpha_6\beta_7$ and D) $\alpha_7\beta_6$. Each β subunit tilt is in a different color.

B.7 Statistics Tables for βØtilt categorical regression

B.7.1 $\alpha_7\beta_7$ statistical results and $\alpha_6\beta_7$

		Replicate 1				Replicate 2				Replicate 3			
		$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_7$ Intercept	$\alpha_6\beta_7$ Slope	$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_7$ Intercept	$\alpha_6\beta_7$ Slope	$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_7$ Intercept	$\alpha_6\beta_7$ Slope
$\alpha_7\beta_7$ Replicate 1	$\alpha_6\beta_7$ -β_1	2.20E-16	3.50E-13	0.05809	0.01182	2.20E-16	1.15E-13	0.0003748	4.23E-06	2.20E-16	3.36E-14	3.67E-16	0.7206
	$\alpha_6\beta_7$ -β_2	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_3	1.05E-09	2.20E-16	2.20E-16	2.20E-16	2.36E-09	2.20E-16	2.20E-16	2.20E-16	9.97E-12	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_4	2.20E-16	0.08509	1.06E-13	2.20E-16	2.20E-16	0.07316	0.34101	1.02E-08	2.20E-16	0.0911237	2.20E-16	0.0004222
	$\alpha_6\beta_7$ -β_5	2.20E-16	0.000105	2.20E-16	2.20E-16	2.20E-16	7.31E-05	2.20E-16	0.0658	2.20E-16	0.000108	2.20E-16	2.30E-10
	$\alpha_6\beta_7$ -β_6	2.00E-16	2.00E-16	2.00E-16	0.01212	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_7	2.20E-16	2.20E-16	2.20E-16	5.44E-09	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
$\alpha_7\beta_7$ Replicate 2	$\alpha_6\beta_7$ -β_1	2.20E-16	7.60E-05	6.72E-15	0.9335	2.20E-16	5.42E-05	2.20E-16	0.0189	2.20E-16	3.69E-05	8.48E-05	0.1063
	$\alpha_6\beta_7$ -β_2	2.20E-16	4.07E-07	2.20E-16	3.50E-08	2.20E-16	2.07E-07	0.01872	2.20E-16	2.20E-16	1.02E-07	2.20E-16	6.29E-14
	$\alpha_6\beta_7$ -β_3	0.5295	<2e-16	<2e-16	<2e-16	0.538	<2e-16	<2e-16	<2e-16	0.4835	<2e-16	<2e-16	<2e-16
	$\alpha_6\beta_7$ -β_4	2.00E-16	0.09364	2.00E-16	2.00E-16	2.20E-16	0.08266	2.20E-16	1.61E-14	2.20E-16	0.09947	6.29E-11	0.03952
	$\alpha_6\beta_7$ -β_5	2.20E-16	0.001564	2.20E-16	2.20E-16	2.20E-16	0.001279	9.36E-13	0.576102	2.20E-16	0.001587	1.46E-07	1.37E-10
	$\alpha_6\beta_7$ -β_6	2.20E-16	2.20E-16	2.20E-16	1.36E-09	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_7	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.23E-14	2.20E-16	0.08289	2.20E-16	8.16E-14	2.20E-16	2.20E-16
$\alpha_7\beta_7$ Replicate 1	$\alpha_6\beta_7$ -β_1	2.20E-16	4.07E-13	0.554169	0.0001723	2.20E-16	1.18E-13	1.66E-07	3.86E-08	2.20E-16	2.68E-14	6.90E-12	0.5251
	$\alpha_6\beta_7$ -β_2	6.76E-11	0.009707	7.48E-14	1.34E-05	2.15E-11	0.007969	0.00112	1.53E-14	6.28E-12	0.00645	2.20E-16	2.58E-09
	$\alpha_6\beta_7$ -β_3	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_7$ -β_4	2.00E-16	0.03878	2.00E-16	2.00E-16	2.20E-16	0.03083	2.20E-16	1.53E-14	2.00E-16	0.04294	2.00E-16	0.71694
	$\alpha_6\beta_7$ -β_5	8.44E-09	2.20E-16	0.003193	0.121028	4.21E-09	2.20E-16	0.0006836	3.79E-12	8.87E-09	2.20E-16	2.20E-16	8.10E-05
	$\alpha_6\beta_7$ -β_6	4.26E-07	0.005332	2.20E-16	1.68E-12	3.95E-06	0.01099	2.20E-16	0.31489	6.45E-06	0.0129	2.20E-16	1.80E-08
	$\alpha_6\beta_7$ -β_7	0.3542	<2e-16	<2e-16	<2e-16	0.3783	2.20E-16	2.20E-16	3.49E-12	0.3853	<2e-16	<2e-16	<2e-16

All are *p-values*, with a threshold **α = 1.98 e-4,** Bonferroni corrected.

Table B.7.1: *p-values* of 256 categorical regression tests for βØtilt as a function of time all seven β subunits in $\alpha_7\beta_7$ and $\alpha_6\beta_7$. The *p-values* with insignificant intercept are marked in yellow, *p-values* with insignificant slope are marked in blue, and in red if both intercept and slope are insignificant. The null hypothesis is rejected for tests whose *p-values* are marked in red.

We observed three cases where βØtilt values of both the intercept and slope of $\alpha_6\beta_7$ were not significantly different from the $\alpha_7\beta_7$-HP simulations (red highlighted). We also see many *p-values* with an insignificant slope (blue), indicating that the βØtilt *values* will not change with time and have converged. The cases with insignificant *p-values* of slope (yellow) all have the corresponding slope value significant, thus for these cases with extended simulations, the βØtilt values of intercept will be significant. But in all the three replicates of $\alpha_6\beta_7$ we see that their βØtilt values for β_6, and β_7 are significantly different from the HP simulations.

B.7.2 βØtilt statistical results for $\alpha_7\beta_7$ and $\alpha_6\beta_6$

		Replicate 1				Replicate 2				Replicate 3			
		$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_6$ Intercept	$\alpha_6\beta_6$ Slope	$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_6$ Intercept	$\alpha_6\beta_6$ Slope	$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_6\beta_6$ Intercept	$\alpha_6\beta_6$ Slope
$\alpha_7\beta_7$ Replicate 1	$\alpha_6\beta_6$ -β_1	2.20E-16	4.08E-14	2.20E-16	0.0023	2.20E-16	3.26E-13	0.000417	2.20E-16	2.20E-16	9.99E-14	2.20E-16	7.13E-06
	$\alpha_6\beta_6$ -β_2	2.20E-16	2.20E-16	2.20E-16	1.28E-15	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_3	1.44E-09	2.20E-16	2.20E-16	2.20E-16	6.46E-10	2.20E-16	2.20E-16	2.20E-16	4.61E-10	2.20E-16	2.20E-16	2.77E-14
	$\alpha_6\beta_6$ -β_4	2.00E-16	0.08224	2.00E-16	0.50359	2.00E-16	0.08411	2.00E-16	2.00E-16	2.20E-16	0.0886	2.13E-09	3.86E-05
	$\alpha_6\beta_6$ -β_5	2.20E-16	6.38E-05	2.20E-16	7.23E-08	2.20E-16	8.55E-05	0.2006	2.20E-16	2.20E-16	9.22E-05	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_6	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.00E-16	2.00E-16	2.00E-16	0.08575
$\alpha_7\beta_7$ Replicate 2	$\alpha_6\beta_6$ -β_1	2.20E-16	3.95E-05	4.35E-05	0.09119	2.20E-16	7.43E-05	0.001122	2.20E-16	2.20E-16	5.20E-05	2.20E-16	0.001384
	$\alpha_6\beta_6$ -β_2	2.20E-16	6.00E-08	4.78E-13	0.1201	2.20E-16	3.06E-07	2.20E-16	1.63E-13	2.20E-16	2.75E-07	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_3	0.5327	<2e-16	<2e-16	<2e-16	0.5245	<2e-16	<2e-16	<2e-16	0.5208	2.20E-16	1.30E-06	2.20E-16
	$\alpha_6\beta_6$ -β_4	2.00E-16	0.09092	2.00E-16	0.2277	2.00E-16	0.09302	2.00E-16	2.00E-16	2.00E-16	0.09677	2.00E-16	2.00E-16
	$\alpha_6\beta_6$ -β_5	2.20E-16	0.001154	2.09E-10	2.20E-16	2.20E-16	0.001376	2.20E-16	2.20E-16	2.20E-16	0.001428	0.913006	2.20E-16
	$\alpha_6\beta_6$ -β_6	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.96E-14	2.20E-16	2.00E-16	2.00E-16	2.00E-16	0.01008
$\alpha_7\beta_7$ Replicate 3	$\alpha_6\beta_6$ -β_1	2.20E-16	3.75E-14	1.53E-14	7.98E-05	2.20E-16	3.75E-13	0.1126	2.20E-16	2.20E-16	1.05E-13	2.20E-16	1.52E-07
	$\alpha_6\beta_6$ -β_2	2.54E-12	0.005524	2.58E-06	0.336477	4.13E-11	0.008915	1.06E-15	4.52E-08	3.45E-11	0.008645	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_3	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_6\beta_6$ -β_4	2.00E-16	0.03687	2.00E-16	0.07199	2.00E-16	0.03817	2.00E-16	2.00E-16	2.20E-16	0.04113	2.20E-16	7.27E-13
	$\alpha_6\beta_6$ -β_5	2.99E-09	2.20E-16	5.92E-08	2.20E-16	5.46E-09	2.20E-16	2.20E-16	2.20E-16	6.20E-09	2.20E-16	2.20E-16	0.03247
	$\alpha_6\beta_6$ -β_6	3.83E-06	0.01088	2.20E-16	4.68E-12	2.68E-06	0.009685	2.20E-16	2.20E-16	7.53E-06	0.01358	7.71E-07	2.20E-16

All are p-values, with a threshold **α = 2.3 e-4,** Bonferroni corrected.

Table B.7.2: *p-values* of 216 categorical regression tests for βØtilt as a function of time all seven β subunits in $\alpha_7\beta_7$ and $\alpha_6\beta_6$. The *p-values* with insignificant intercept are marked in yellow, *p-values* with insignificant slope are marked in blue, and in red if both intercept and slope are insignificant. The null hypothesis is not rejected for any tests in this table.

Similar to Table B.5.2, we observed no cases where *p-values* of both intercept and the slope are above our threshold of $2.3e^{-4}$ for the intermediate $\alpha_6\beta_6$. Thus, all the βØtilt values in all the β subunits of $\alpha_6\beta_6$ are significantly different from $\alpha_7\beta_7$.

B.7.3 βØtilt statistical results for $\alpha_7\beta_7$ and $\alpha_7\beta_6$

We observed only two cases where βØtilt *values* of both the intercept and slope of $\alpha_7\beta_6$ were not significantly different from the $\alpha_7\beta_7$-HP simulations (red highlighted). We also see many *p-values* with an insignificant slope (blue), indicating that the βØtilt values will not change with time and have converged. The cases with insignificant *p-values* of slope (yellow) all have the corresponding slope value significant, thus for these cases with extended simulations, the βθ values of intercept will be significant. Interestingly, we see only two cases where the intercepts (yellow) are insignificant, compared to the four insignificant slopes in βθ values in **Table B.5.3**.

		Replicate 1				Replicate 2				Replicate 3			
		$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_7\beta_6$ Intercept	$\alpha_7\beta_6$ Slope	$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_7\beta_6$ Intercept	$\alpha_7\beta_6$ Slope	$\alpha_7\beta_7$ Intercept	$\alpha_7\beta_7$ Slope	$\alpha_7\beta_6$ Intercept	$\alpha_7\beta_6$ Slope
$\alpha_7\beta_7$ Replicate 1	$\alpha_7\beta_6$ -β_1	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.00E-16	2.00E-16	1.11E-02	1.33E-03	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_2	2.20E-16	5.58E-07	2.20E-16	3.00E-11	2.20E-16	7.84E-07	2.20E-16	2.20E-16	2.20E-16	5.09E-07	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_3	2.20E-16	4.53E-15	2.20E-16	1.58E-05	2.20E-16	2.27E-15	2.20E-16	3.86E-09	2.20E-16	4.27E-16	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_4	<2e-16	<2e-16	<2e-16	5.31E-01	2.20E-16	2.20E-16	2.20E-16	9.10E-08	2.20E-16	2.20E-16	2.20E-16	4.97E-14
	$\alpha_7\beta_6$ -β_5	2.20E-16	2.20E-16	2.20E-16	0.0002738	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	5.04E-06
	$\alpha_7\beta_6$ -β_6	2.00E-16	3.96E-02	2.00E-16	2.00E-16	2.00E-16	2.83E-02	2.00E-16	2.00E-16	2.20E-16	0.03832	2.20E-16	2.30E-10
$\alpha_7\beta_7$ Replicate 2	$\alpha_7\beta_6$ -β_1	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.00E-16	2.00E-16	2.00E-16	0.00237	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_2	2.20E-16	2.92E-12	2.20E-16	2.20E-16	2.20E-16	5.73E-12	2.20E-16	0.01229	2.20E-16	2.51E-12	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_3	2.20E-16	2.20E-16	2.20E-16	3.12E-09	2.20E-16	2.20E-16	2.20E-16	3.71E-05	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_4	<2e-16	<2e-16	<2e-16	0.8068	2.20E-16	5.46E-15	2.20E-16	2.28E-10	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	$\alpha_7\beta_6$ -β_5	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	0.0007524	2.20E-16	2.20E-16	2.20E-16	2.20E-16	5.70E-06
	$\alpha_7\beta_6$ -β_6	2.20E-16	1.94E-09	2.20E-16	2.20E-16	2.20E-16	1.86E-10	2.20E-16	2.20E-16	2.20E-16	1.46E-09	2.20E-16	5.47E-07
$\alpha_7\beta_7$ Replicate 3	$\alpha_7\beta_6$ -β_1	2.20E-16	1.01E-05	2.20E-16	2.20E-16	2.20E-16	1.59E-05	2.20E-16	2.20E-16	2.20E-16	1.24E-05	2.20E-16	0.04389
	$\alpha_7\beta_6$ -β_2	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	0.0007092
	$\alpha_7\beta_6$ -β_3	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.03E-07
	$\alpha_7\beta_6$ -β_4	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.56E-10	2.20E-16	2.20E-16	2.20E-16	0.006186
	$\alpha_7\beta_6$ -β_5	2.20E-16	0.001181	0.074614	0.645465	2.20E-16	0.0015858	2.20E-16	0.0003044	2.20E-16	0.001146	0.822381	2.75E-10
	$\alpha_7\beta_6$ -β_6	2.20E-16	0.03534	2.20E-16	2.45E-11	2.20E-16	0.02487	2.20E-16	6.35E-10	2.00E-16	0.03403	2.00E-16	2.00E-16

All are *p-values*, with a threshold **α = 2.3 e-4,** Bonferroni corrected, from 216 tests.

Table B.7.3: *p-values* of 216 categorical regression tests for βØtilt as a function of time all seven β subunits in $\alpha_7\beta_7$ and $\alpha_7\beta_6$. The *p-values* with insignificant intercept are marked in yellow, *p-values* with insignificant slope are marked in blue, and in red if both intercept and slope are insignificant. The null hypothesis is rejected for tests whose *p-values* are marked in red.

B.8 Simulations System Information

Simulation	Box size (Å)	Number of atoms	Number of water molecules	Ions
HP-$\alpha_7\beta_7$	155 X 155 X 155	352979	298725	Na^+ = 275 Cl^- = 184
$\alpha_6\beta_7$	155 X 155 X 155	386163	335430	Na^+ = 285 Cl^- = 209
$\alpha_6\beta_6$	155 X 155 X 155	353249	306735	Na^+ = 256 Cl^- = 190
$\alpha_7\beta_6$	155 X 155 X 155	379857	329751	Na^+ = 284 Cl^- = 205

Table B.8: System properties and details of the HP and near-HP simulations. All the simulations are run in a rectangular water box with 15Å water on each side of the protein.

APPENDIX C

Appendix for Chapter 4
Robustness and Kinetic Trapping in Bacterial Core Particle Assembly

C.1 Sidedness of subunits in ODE models

C.2 Definition of the CP interfaces for interaction affinity (K_D)

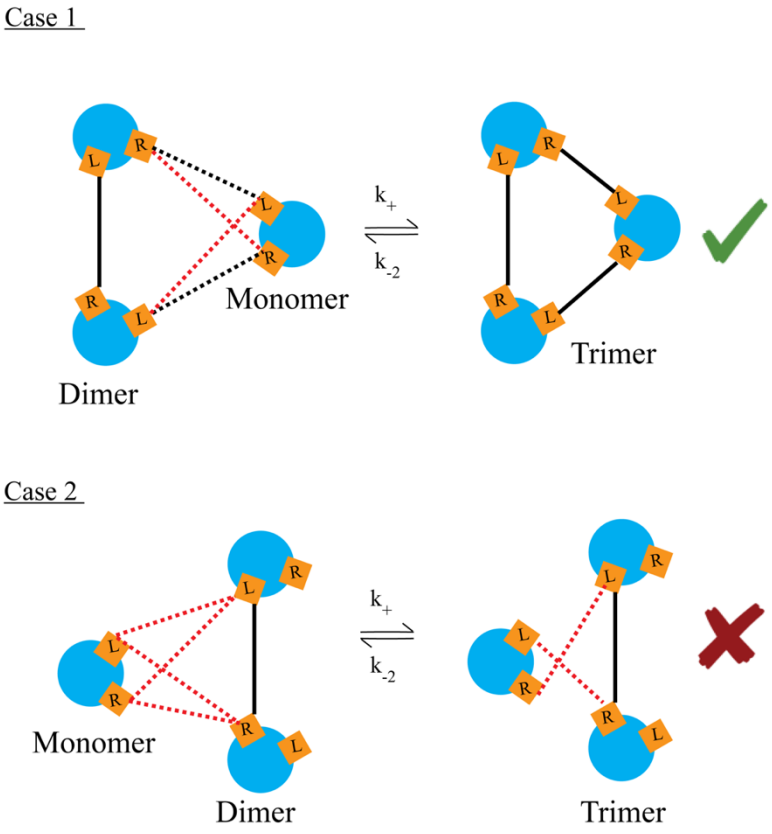


Figure C.1**:** Schematic of a trimer formation displaying the sidedness used for developing ODE models. Each subunit has a distinct left (L) and a right (R) side, and interactions can occur only between the right side of one subunit and the left side of the other subunit. Case 1 shows the allowed reactions and Case2 shows the reactions which are not allowed due to incorrect sides interacting.

All proteins are internally asymmetric, and there is a sidedness to the protein structure [1]. We have defined six unique interfaces as a convention for our proteasome CP studies. Since the 20S CP quaternary structure is conserved, these conventions can be applied to CPs from different species. For our studies, we are referring to the *Re* bacterium 20S CP (PDB:1Q5R).

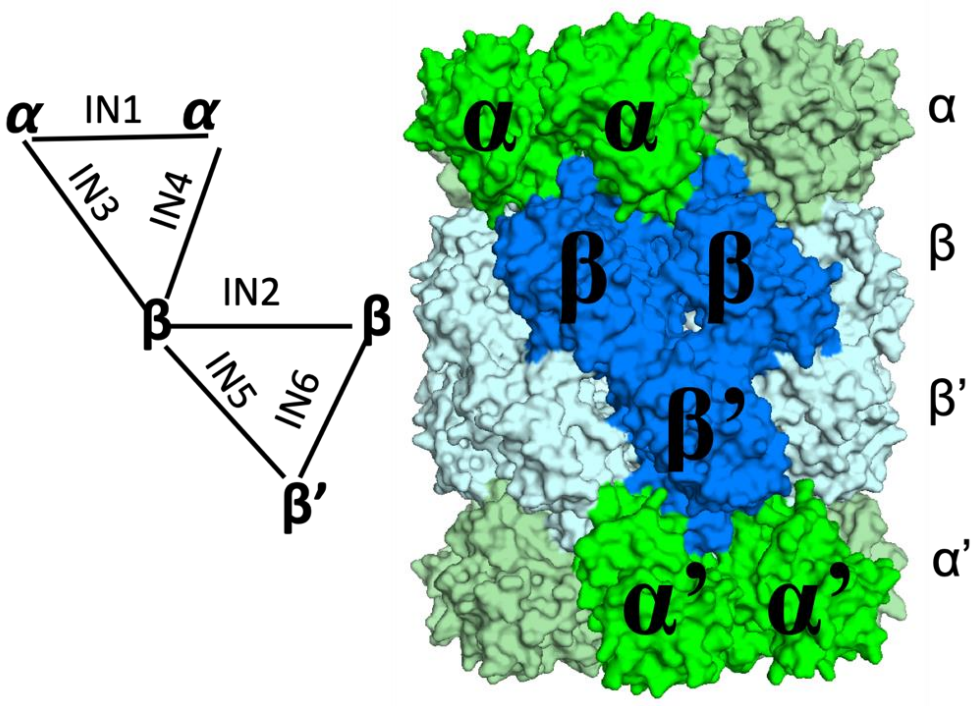


Figure C.2: Schematic of the *Re* CP crystal structure (PDB ID: 1Q5R) to show the different interfaces (IN). There are six unique interfaces for every CP.

Interface 1(IN1): Is the interface between two α subunits. Interface 2(IN2): is the interface between two β subunits. Interface 3(IN3) and Interface 4(IN4): are the two interfaces between α and β subunits. The last two are interfaces that form the HP dimerization surface, i.e., Interface 5 (IN5) and Interface 6 (IN6): where two β rings of HP's interact and assemble into the CP. For the ODE models, we associate each interface with an interaction affinity or binding strength (KD) in M. Stronger interfaces have higher; for example, the avidin-biotin complex is the strongest known non-covalent interaction with KD = 10^{-15} M. We are looking into only the four interfaces (IN1, IN2, IN3, and IN4) for our models because we are explicitly looking to form HP's. For ARF, we vary KD for IN1 (KD1) and maintain the remaining KD values at a

constant number. Similarly, we vary KD for IN4 (KD4) for ABD and UOM while retaining the other interface's KD values at a constant number.

C.3 Proteasome assembly dynamics and deadlock

As discussed in the main text, CP assembly dynamics are susceptible to deadlock, and this reduces assembly efficiency. For the three models of CP assembly described-ARF, ABD and UOM, the partly hierarchical and less robust UOM shows are highly susceptible to deadlock. As described in detail in Supplementary Section 4 [1], every ring-like structure has an optimal affinity based on its ring size and subunit concentration. To generate the results in Fig. C.3 we adjusted the interaction affinities so that the maximum CP efficiency at 1 μM. The subunit concentration ([α]=[β]) was from 10^{-12} to 10^{-2} M, with the interaction affinities (K_D) as follows: UOM and ABD, K_D at IN4 is the strongest (UOM=1X10^{-4}M; ABD=1X10^{-6}M), but ARF K_D at IN1 is the strongest (1X10^{-8}M). All the remaining interfaces have a weak affinity (K_D = 10^{-2} M) for all models. The association rates k_+=10^6 $M^{-1}s^{-1}$, which is the protein association rates observed in most reactions, and the k_{+HP} = 10^3 $M^{-1}s^{-1}$ which is approximated from the experimental evidence that we observe for dimerization of HP's (Chapter 2).

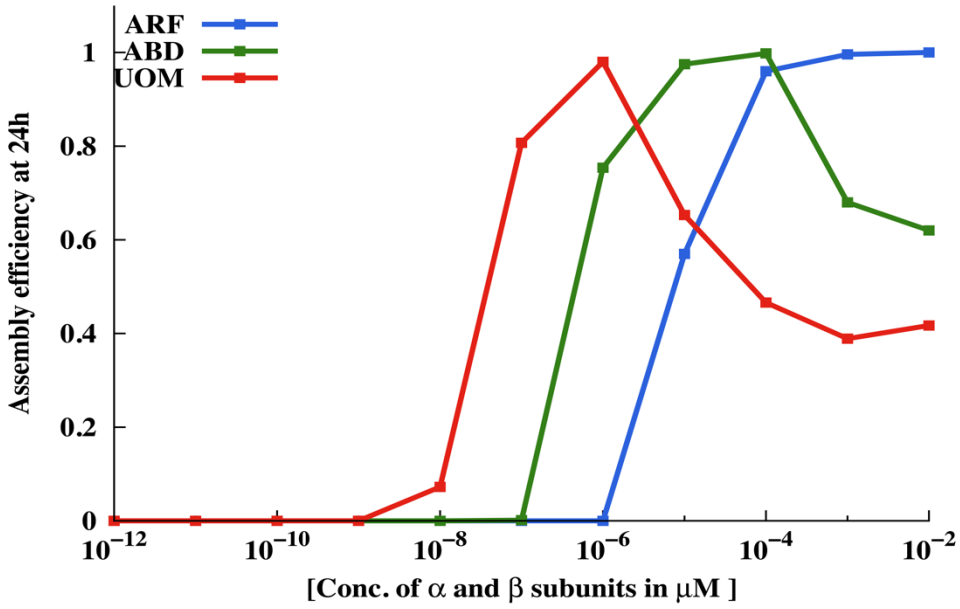


Figure C.3: CP assembly at varying concentrations with [α] = [β]. The three models exhibit different fraction CP assembled (or assembly efficiency) and show a varying deadlocked plateau. The x-axis is in log scale to clearly illustrate the deadlock plateau. The interaction affinities for each model are equivalent in this plot.

C.4 Time profiles of the CP assembly kinetics

If we have the same K_D for all the three model's we observe that their maximum assembly efficiency peaks at different concentrations. This implies that based on the assembly pathway an optimal affinity exists, and if the K_D is too weak then the intermediates fall apart even before they are assembled and if K_D is very strong then there exists a longer deadlock phase. In (Fig. C.3) we have all the parameters same across the three models. The K_{D4} =1X10^{-4}M for UOM and ABD, but for ARF K_{D1} =1X10^{-4}M, and the remaining three affinities in either case are set to a lower strength of K_D =1X10^{-2}M.

The initial concentration of subunits also affects the CP assembly dynamics, as shown in Fig. C.4. We observe that at lower concentration UOM (Fig. C.4.A) is fastest and then as subunit concentration increases (Fig. C.4.B, and 4C) there is deadlock induced in UOM, which reduces its assembly efficiency. The ABD model has maximum efficiency at 10^{-5}M and then with further increase in concentration it started declining.

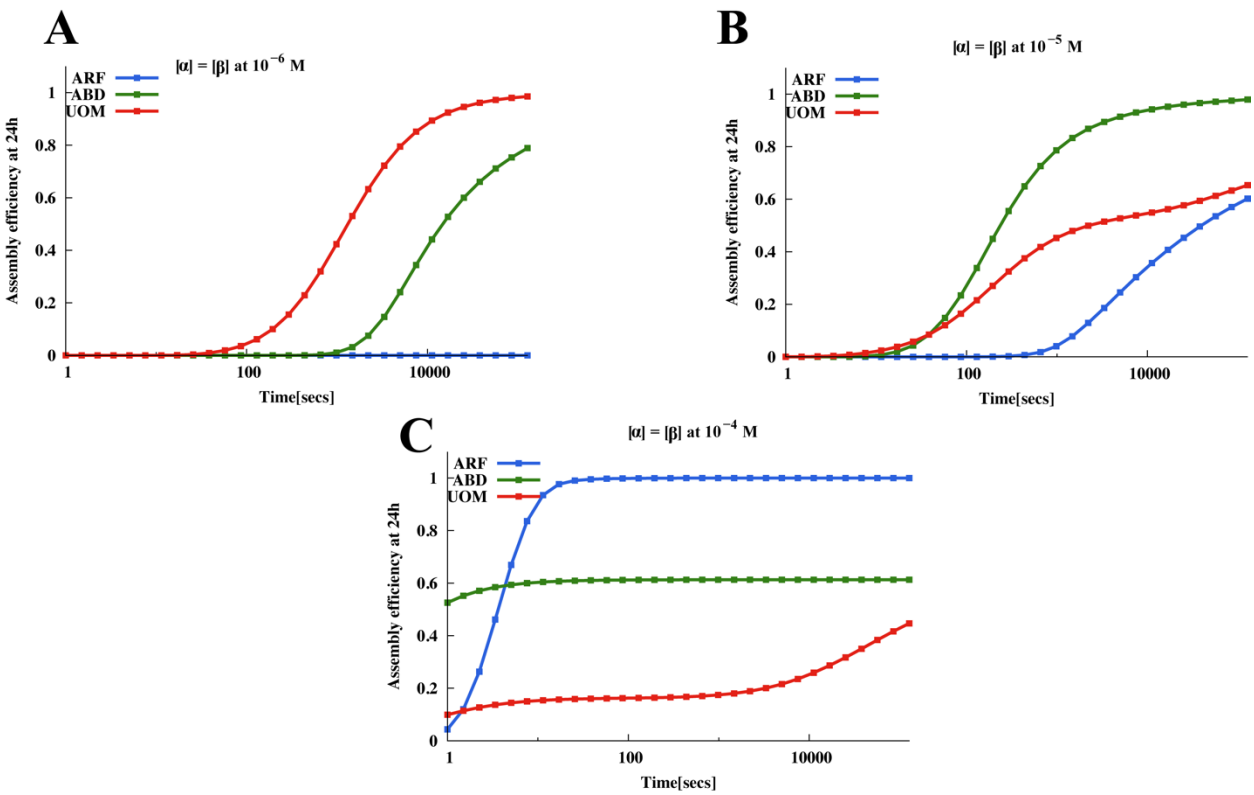


Figure C.4: Effect of subunits concentration on CP assembly dynamics A) At concentration [α] = [β] = 10^{-6} M. B) At concentration [α] = [β] = 10^{-5} M. C) At concentration [α] = [β] = 10^{-4} M. The k_{+HP} = 10^{3} M^{-1}s^{-1}, k_{+}=10^{6} M^{-1}s^{-1}, K_{D4} =1X10^{-4}M (UOM, ABD), K_{D1} =1X10^{-4}M (ARF), and the other K_D are at a lower affinity of 1X10^{-2}M.

The ARF (**Fig. C.4.C**) has higher assembly efficiency at higher concentrations than UOM, and ABD. The ARF is slower (**Fig. C.4C**) because the K_{D1} =1X10^{-4}M is a weak interaction affinity and thus delays the CP assembly because the intermediates are not stable and fall apart quickly. As seen in **Fig. C.3** the optimal affinity for ARF is K_{D1} =1X10^{-8}M.

C.5 *In vitro* native gel CP assembly experiments

For the assembly experiments, which are done by a graduate student in Dr. Eric J. Deeds lab-Anupama Kante, began with the cloning the α1 and β1 genes of *Re* in separate expression plasmid vectors. The α and β subunits were expressed in *E. coli* and purified separately using metal affinity chromatography (see methods). As seen in several other studies [2-4], she was able to obtain fully assembled CPs on mixing the two subunits together. To determine the effect of subunit concentration on the assembly efficiency, we incubated the α and β subunits together in an equimolar ratio at seven concentrations ranging from 0.25uM to 16uM. The assembly process was allowed to proceed for 24hrs at 30^{0}C and formation of CP was determined by native PAGE **(Fig. C.5)**. Gels were stained using Sypro Ruby [5] and imaged using a Licor imager. The sensitivity of the Sypro Ruby stain [5] was established by the linearity of the band intensities corresponding to concentrations of pre-assembled CPs. Assembly reactions at all concentrations show two distinct bands – a 740kD proteasome CP and a smaller 480kD HP band. As defined earlier, assembly efficiency is the percent of monomers present in the CP. In this assay, band intensity is proportional to the concentration of CP. The CP assembly efficiency was thus calculated by dividing the band intensity with respective subunit concentration and normalized to the highest value in the assay. In **Fig. 4.5**, we see an increase in assembly efficiency with initial increase in the subunit concentration but, as concentration increased above 1uM, assembly efficiency started to decline. The *in vitro* assembly yield (**Fig. C.5)** has a similar profile as that of the simulation showing the effect of concentration on the assembly yield of three membered stacked ring in **Fig. 4.1**. This result concretely suggests that like the model, *in vitro* assembly of

CP is initially limited by diffusion of subunits at low concentrations and by dissociation of kinetically trapped intermediates at high concentrations.

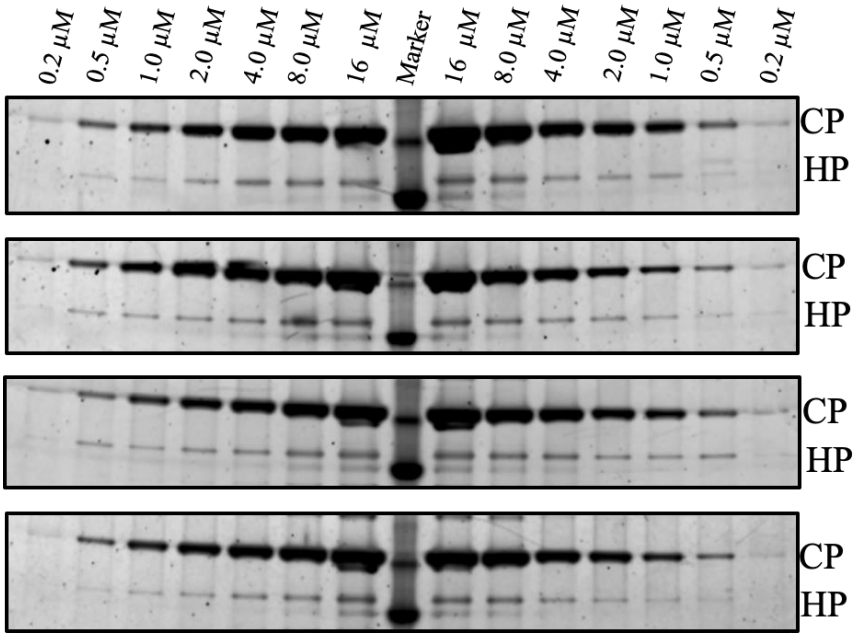


Figure C.5: Native-PAGE Analysis of *in vitro* assembly of 20S proteasome Core Particle *from Rhodococcus erythropolis.*

C.6 Parameter selection for comparing models and experimental data

To fit the assembly models with *in vitro* CP assembly dynamics, we did a maximum likelihood estimation using Root Mean Square Error (RMSE) as a measure statistic to identify the absolute fit of the model to the experimental data. Lower values of RMSE indicate a better fit. We initially started with varying three parameters — k_{+}, k_{+HP}, and K_D. The k_{+} and k_{+HP} was varied from 10^2 to 10^8 $M^{-1}s^{-1}$ in increments of 10 and K_D was varied to make the affinities range from 10^{-2}M to 10^{-8}M. For all these combinations a RMSE was calculated to search for the lowest value as close to 0. The **Table C.6** shows the parameters which were selected for the final fit results that are shown in **Fig. 4.6**.

Model	k_+	k_{+HP}	K_D
ARF	10^8	10^8	$5X10^{-6}$
ABD	10^6	10^0	$8x10^{-7}$
UOM	1100	270	$3x10^{-5}$

Table C.6: Table showing the list of parameters optimized for final fits of the models and experimental data which are shown in Chapter 4 (Fig. 4.6).

C.7 *In vivo* assembly dynamics with synthesis and degradation rates

We included synthesis (Q) and degradation rates (δ), for various concentrations as used in Chapter 4 (**Fig. 4.7**) to explore if deadlock can still impact assembly efficiency in "in *vivo*" scenarios. Here, the subunits are constantly supplied and degraded as it happens in living systems. In **Fig. C.7** we demonstrate the impact of degradation rates on assembly efficiency. The synthesis rate is calculated by $C_T = Q / \delta$. The parameters for **Fig. C.7** are three degradation rates and monomers synthesis rate Q and C_T varies from 10^{-2}M to 10^{-8}M. Another feature of these models is that the synthesis rate is only for the subunits or monomers and no other intermediates.

We observe that for slower degradation rates $\delta = 5.55X10^{-6}\ s^{-1}$ (**Fig. C.7C**) the subunits have enough time to alleviate deadlock and get higher yields at lower concentrations and (10^{-7}M), whereas for higher degradation rates $\delta = 5.55X10^{-4}\ s^{-1}$ (**Fig. C.7A**) we obtain higher yields at higher subunit concentrations because the intermediates are degraded faster than they are bring used for assembly.

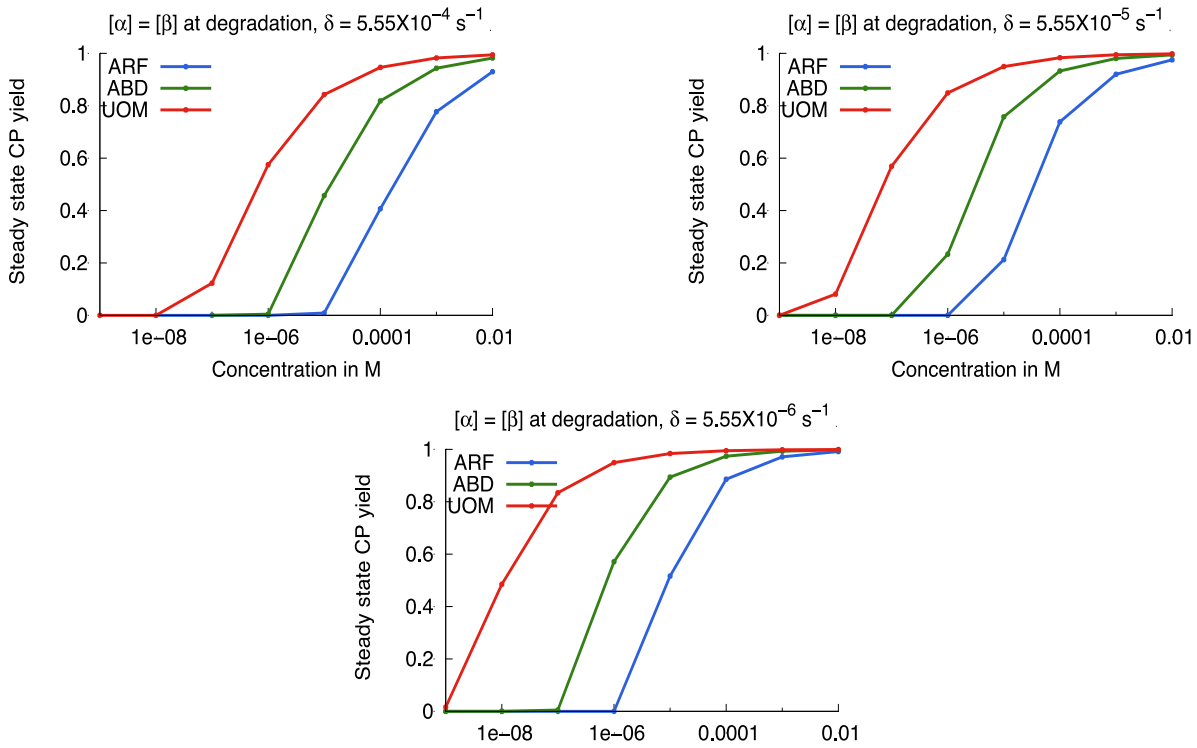


Figure C.7: Assembly dynamics for models with synthesis and degradation rates. A) $\delta = 5.55X10^{-4}s^{-1}$ B) δ = 5.55X10-5 s^{-1} C) $\delta = 5.55X10^{-6}s^{-1}$.

C.8 References

1. Deeds EJ, Bachman JA, Fontana W: Optimizing ring assembly reveals the strength of weak interactions. *Proc Natl Acad Sci U S A* 2012, 109(7):2348-2353.
2. Mayr J, Seemuller E, Muller SA, Engel A, Baumeister W: Late events in the assembly of 20S proteasomes. *J Struct Biol* 1998, 124(2-3):179-188.
3. Zuhl F, Seemuller E, Golbik R, Baumeister W: Dissecting the assembly pathway of the 20S proteasome. *FEBS Lett* 1997, 418(1-2):189-194.
4. Zühl F, Tamura T, Dolenc I, Cejka Z, Nagy I, De Mot R, Baumeister W: Subunit topology of the Rhodococcus proteasome. *FEBS Letters* 1997, 400(1):83-90.
5. SYPRO Ruby Protein Gel Stain [https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2Fmp12000.pdf&title=U1lQUk8gUnVieSBQcm90ZWluIEdlbCBTdGFpbg==]