SCHOOL OF HEALTH SCIENCES
DEPARTMENT OF MOLECULAR BIOLOGY & GENETICS

BSc Thesis

“Folding of the human Pin1 WW domain using molecular dynamics simulations”

Author’s Name: Aristea – Marina Zigkou

Supervisor’s name: Dr. Nicholas M. Glykos
Associate Professor of Structural and Computational Biology

Alexandroupolis, Greece

August, 2023

Abstract

In recent years, there has been significant progress in the field of Molecular Dynamics Simulations, providing us with the ability to understand the behavior and dynamics of molecular systems at the atomic level. The challenges and failures encountered by researchers in their efforts to create effective force fields have been insightful. In this thesis, we explore these possibilities using the well-studied β-sheet fold of the Fip mutant, which constitutes the WW domain of the Pin1 protein. Understanding the folding process of a β-sheet, which is a dominant secondary structure, is of great importance. Our objective is to analyze and confirm whether the applied force fields and parameters are capable of successfully simulating the folding process of our protein. Two separate folding attempts of the Fip protein were conducted in our laboratory using MD simulations, with the peptide chain initially unfolded. The first attempt involved using the Amber ff99SB-ILDN force field for a 10 μs simulation and the second attempt used the Amber ff99SB*-ILDN force field for a 15 μs simulation. We attempt to analyze these two trajectories. For our analyses, we extensively utilized the user-friendly (GUI) Grcarma, which is based on the Carma program. Both tools have been developed by our laboratory with the purpose of analyzing trajectories generated from MD simulations. Remarkably, the ff99SB*-ILDN force field, in contrast to the ff99SB-ILDN force field, successfully folded our protein in approximately half the total simulation time (around 7.2 μs). In conclusion, our study highlights the importance of Molecular Dynamics Simulations in understanding the folding of the β-sheet structure in protein Fip. By using advanced force fields and tools like Grcarma, we gained valuable insights into the dynamics of the Pin1's WW domain.

Περίληψη

Τα τελευταία χρόνια, έχει παρατηρηθεί σπουδαία εξέλιξη στον τομέα του Molecular Dynamics Simulations, παρέχοντάς μας την ικανότητα να κατανοούμε σε ατομικό επίπεδο τη συμπεριφορά και τη δυναμική των μοριακών συστημάτων. Οι δυσκολίες και οι αποτυχίες που συνάντησαν οι ερευνητές στην προσπάθειά τους να δημιουργήσουν ικανά force field ήταν καθοριστικές. Στην παρούσα πτυχιακή, εξερευνούμε αυτές τις δυνατότητες κάνοντας χρήση του πολύ καλά μελετημένου β-πτυχωτού φύλλου του μεταλλάγματος Fip, το οποίο αποτελεί το WW domain της πρωτεΐνης Pin1. Το να μπορέσουμε να κατανοήσουμε τη διαδικασία αναδίπλωσης ενός β-πτυχωτού φύλλου, όπου είναι μία δευτεροταγής δομή η οποία είναι παρούσα παντού στη φύση, είναι εξέχουσας σημασίας. Στόχος μας είναι να αναλύσουμε και να επιβεβαιώσουμε εάν τα force fields και κατ' επέκταση οι παράμετροι που έχουν εφαρμοστεί, είναι ικανά να προσομοιώσουν επιτυχώς τη διαδικασία αναδίπλωσης της πρωτεΐνης μας. Από το εργαστήριό μας έχουν γίνει δύο ξεχωριστές προσπάθειες αναδίπλωσης της πρωτεΐνης Fip, μέσω MD simulations, έχοντας αρχικά την πεπτιδική αλυσίδα ξεδιπλωμένη. Τη μία φορά με τη χρήση του Amber ff99SB-ILDN force field επιδιώξαν την αναδίπλωση της πρωτεινης σε συνολική διάρκεια 10μs και τη δεύτερη φορά με τη χρήση του Amber ff99SB*-ILDN force field σε συνολική διάρκεια 15 μs. Τα δύο τροχιακά που παραχθήκαν είναι αυτά που θα εξετάσουμε. Για τις αναλύσεις μας, χρησιμοποιήσαμε κατά κόρον το GUI Grcarma το οποίο έχει βασιστεί στο πρόγραμμα Carma. Και τα δύο έχουν δημιουργηθεί από το εργαστήριό μας με σκοπό τις αναλύσεις τροχιακών παραγόμενων από MD simulations. Το ff99SB*-ILDN force field, εν αντιθέση με το ff99SB-ILDN force field, κατάφερε να αναδιπλώσει επιτυχώς την πρωτεΐνη μας στον μισό περίπου χρόνο της συνολικής διάρκειας της προσομοίωσης, (περίπου στα 7.2 μs). Εν κατακλείδι, η έρευνά μας τονίζει τη σημασία των Molecular Dynamics Simulations στην κατανοήση της διαδικασίας αναδίπλωσης του β-πτυχωτού φύλλου της πρωτεΐνης Fip. Χρησιμοποιώντας προηγμένα force fields και εργαλεία όπως το Grcarma, αποκτήσαμε πολύτιμες γνώσεις για τη δυναμική του WW domain της πρωτεΐνης Pin1.

Acknowledgments

I would like to sincerely thank my mentor, Dr. Nicholas M. Glykos, for his invaluable guidance and patience throughout my thesis. His clear instructions and prompt responses greatly contributed to the success of my research. Working in his laboratory's fascinating field of study was a true privilege, providing an stimulating environment.

I also want to express my heartfelt gratitude to my family and partner for their unwavering support. Their understanding and encouragement played a vital role in making this academic journey fulfilling and possible. Their belief in me kept me motivated during the challenges I faced throughout my studies.

Table of Contents

1. Introduction

1.1 The significance of the N-terminal WW domain in Pin1 protein, structure, function and implications

Protein Pin1 is a highly conserved enzyme with a spherical structure and it is known for its important role in the cell's transition from G2 phase to M phase. It basically acts as a regulator of mitosis and it has been found to be involved in interactions with cell cycle regulatory kinase NIMA. The whole protein consists of 163 amino acids but we will not study the whole protein, instead we are turning our focus on a very specific part of it, the N-terminal 39-residue WW domain. Among the reasons for selecting this particular protein, the N-terminal 39-residue WW domain stands out as a well-studied domain. The WW domain of the Pin1 protein, in its wild-type form, typically features two highly conserved tryptophan residues, which are responsible for its distinctive terminology. Acting as an intermediate for protein-protein interactions, the WW domain binds to short proline-rich sequences present in other proteins. [1][2] Additionally, it is worth noting that homologous family members containing the WW domain can be found in the yeast and some fungi, emphasizing its significance in these organisms.[3] Notably, the WW domain is characterized by its fairly fast folding process (approximately 100 μs) [4], [5], small size, high solubility and ability to fold into a twisted three-strand antiparallel β-sheet secondary structure.[6] This domain has served as an invaluable model for investigating the folding process of β-sheet secondary structure. Furthermore, even when isolated, the WW domain maintains its expected secondary structure, folded conformation and functional properties. [3]

1.2 Characterization of mutant forms of Pin1 Protein's WW domain and their role in folding and stability

In our study, we focused on a mutant variant of the Pin1 protein called Fip, which carries significant point mutations. The folding kinetics and stability of Pin1 heavily rely on its amino acid sequence. The rate of the folding process is determined by the formation of loop 1, which connects β-strand I and β-strand II, while the interactions between the hydrophobic core of the protein and loop 2, connecting β-strand II and β-strand III, contribute to its overall stability. [7] Mutations in loop 1 can impact the folding kinetics. [7]–[9]

The wild-type Pin1's WW domain sequence, as represented in the Protein Data Bank (PDB) entry 1PIN, is as follows:
MADEEKLPPGWEKRMSRSSGRVYYFNHITNASQWERPSG
The Fip mutant's WW domain sequence, represented in the PDB entry 2F21, is:
MADEEKLPPGWEKRMSADGRVYYFNHITNASQWERPSG

In our study, we further modified the Fip sequence to create two additional mutants:

(1) GSKLPPGWEKRMSRDGRVYYFNHITNASQFERPSG (referred to as Fip35)

(2) SKLPPGWEKRMSRDGRVYYFNHITNASQFERPSG (referred to as Fip34)

The Fip35 mutant consists of 35 amino acids and served as the reference sequence for molecular dynamics simulations using the Amber ff99SB-ILDN force field. On the other hand, the Fip34 mutant consists of 34 amino acids and was employed as the reference sequence for molecular dynamics simulations using the Amber ff99SB*-ILDN force field.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
M	A	D	E	E	K	L	P	P	G	W	E	K	R	M	S	R	S	S	G	R	V	Y	Y	F	N	H	I	T	N	A	S	Q	W	E	R	P	S	G
-	-	-	G	S	K	L	P	P	G	W	E	K	R	M	S	R	D	-	G	R	V	Y	Y	F	N	H	I	T	N	A	S	Q	F	E	R	P	S	G
-	-	-	-	S	K	L	P	P	G	W	E	K	R	M	S	R	D	-	G	R	V	Y	Y	F	N	H	I	T	N	A	S	Q	F	E	R	P	S	G

Figure 1| *Comparison of Pin1 WW domain and mutant variants.*
In this figure, we present an alignment of the three sequences: the wild-type Pin1 WW domain, mutant 1 (Fip35) and mutant 2 (Fip34). The red color represents the mutations (deletions and substitutions), the blue color represents the three β-strands and all the other residues are black.

Starting from the N-terminal, we made specific modifications to the sequence. In both mutants, the first five residues (M, A, D, E, E) were deleted and in mutant 1, residues G and S were added, while in mutant 2, only residue S was added. Within loop 1, we introduced a substitution at residue 18, Ser18Asp and deleted Serine at position 19 in both mutants. Furthermore, a substitution was made at residue 34, Trp34Phe, which was observed in both mutants (**Figure 1**). These modifications have significantly reduced the protein's folding time from around 100 μs to approximately 12 μs, making it perfectly suitable for our needs. [4], [5], [10]

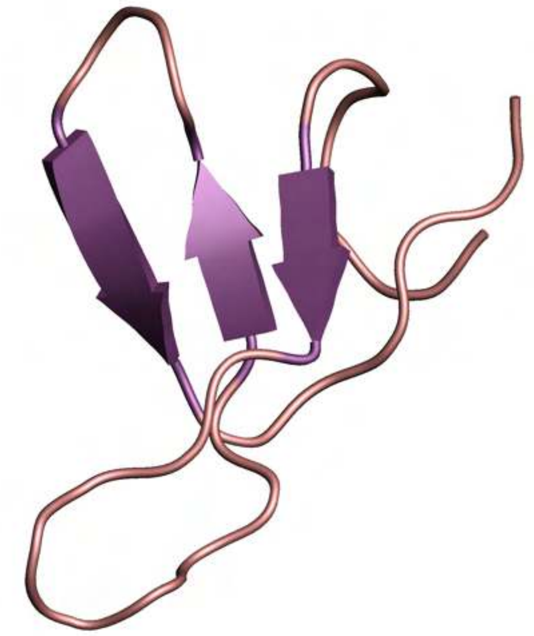


Figure 1 | *Fip35 WW domain*
Using pymol, structure is colored by secondary structure: sheet (magenta), helix (cyan) and loop (salmon)

1.3 Exploring the folding landscape of the WW domain, lessons from past molecular dynamics simulations

This section delves into the extensive efforts made over the years to achieve successful folding of the WW domain using molecular dynamics simulations. We highlight the significant insights gained through these attempts, including force field's empirical parameters and techniques that have been discovered.

In recent years, significant progress has been made in utilizing force fields with empirical parameters, improving their accuracy in reproducing the folding events and folded structures of the WW domain. This has enabled more realistic simulations and enhanced our understanding in molecular interactions.

However, the limited processing power of conventional computational resources creates a significant barrier to fully exploring the complex protein folding process. Supercomputers with massive parallel processing capabilities have been developed to overcome this challenge and achieve the necessary timescale sampling, but their availability remains limited.

1.3.1 CHARMM22 force field with CMAP corrections fails to fold the Fip35 protein

In this experiment, scientists used Fip35 mutant variant of the human Pin1 WW domain. The simulation parameters included the use of the CHARMM22 force field with CMAP corrections, explicit water model (TIP3P) with 30 mM NaCl and a custom version of NAMD 2.6 as the simulation software. The simulation was performed at a temperature of 337K on the abe cluster at the National Center for Supercomputing Applications.[10] The CMAP corrections in the CHARMM22 force field improved how the peptide backbone is represented. They used quantum mechanical calculations and crystallographic data to fix deviations in φ and ψ values, making protein folding and conformational dynamics simulations more accurate.[11]

The stability of Fip35's native state was confirmed through a 200 ns simulation, with Ca root mean-square deviation consistently below 1.4 Å, showing a stable structure. During folding, Fip35 first collapsed to a molten globule state with a hydrophobic core in 500 ns. Subsequently, specific residues formed α-helices and the trajectory switched between states with helices and sheet/disordered coil regions. Surprisingly, the simulation consistently showed α-helical conformations throughout the 10 μs trajectory, deviating from the expected folding pathway. [10]

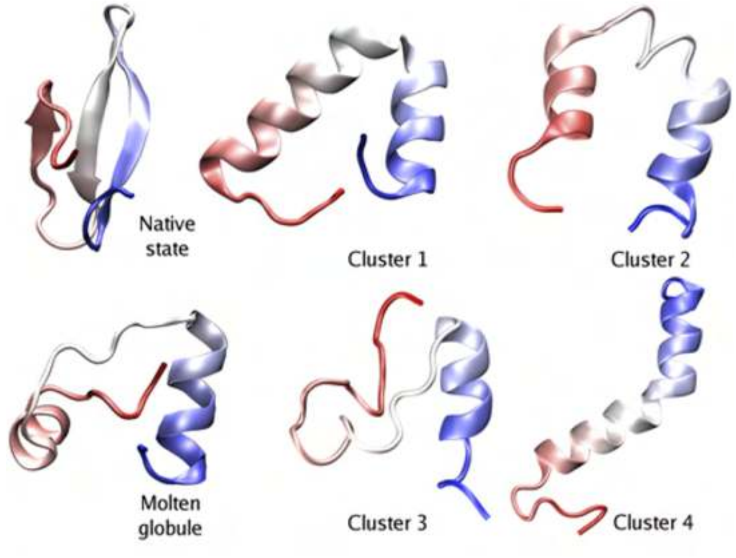


Figure 2 | *Native state, molten globule formed after 500 ns and the first four representative structures from clustering analysis. (reproduced without permission from Freddolino et al., Biophysical Journal, 2008)*

1.3.2 Investigating structural heterogeneity, inability in accurately folding Fip35 protein using AMBER96 and distributed computing, Folding@home

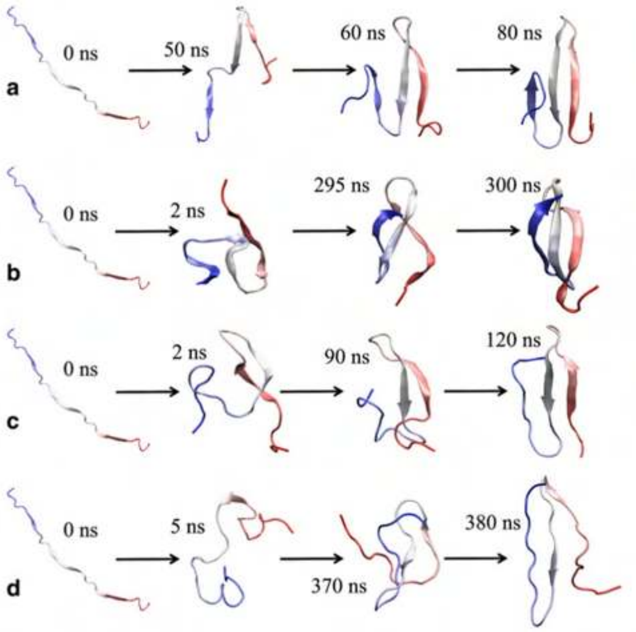
The researchers conducted an experiment to investigate the folding mechanism of the Fip35 WW domain protein. They employed molecular dynamics simulations using the AMBER96 force field and the Folding@home distributed computing platform. Thousands of folding trajectories were generated, revealing structural heterogeneity in the folding process.

By analyzing the root mean-square deviation (RMSD) and using DSSP analysis to determine the conformational state, the researchers observed the formation of three-stranded β-sheet conformations in some of the folding trajectories. However, only one trajectory closely matched the reference structure. This suggests difficulties in achieving accurate protein folding.

The simulations gave us important insights into how Fip35 folds. They showed us that considering multiple trajectories and the diverse folding pathways is crucial. The study emphasized the power of distributed computing and GPU technology for exploring the complex process of protein folding.

Figure 3 illustrates four folding trajectories with different folding patterns. In **Figure 3a**, the first hairpin forms early, followed by strand I and II adopting β-sheet conformations, leading to complete folding. **Figure 3b** shows a quick initial collapse, leading to simultaneous fulfillment of β-sheet criteria. **Figure 3c** demonstrates a fast collapse with all three strands acquiring β-sheet conformations simultaneously. **Figure 3d** exhibits the formation of the second hairpin first, followed by strand II and III adopting β-sheet conformations before strand I. [12]

Figure 3 | *Four folding trajectories from (a) T300- γ91, (b) T300-γ1, (c), and (d) T330-γ1(reproduced without permission from Ensign et al., Biophysical Journal, 2009)*



1.3.3 Revealing transferability, *β*-sheet Pin1 WW domain folding with enhanced Amber ff03* force field via simple backbone correction

In the following experiment, replica exchange molecular dynamics (REMD) was used to investigate the folding behavior of the Pin1 WW domain. Utilizing the Amber ff03* force field and the TIP3P water model, the simulations were performed with GROMACS 4.0.5. The simulations started from unfolded configurations at 800 K and used a Langevin integrator for the dynamics. Every 10 ps, exchanges between neighboring replicas were attempted among a total of 32 replicas, covering temperatures from 300 to 457 K.

During the REMD simulations, the Pin1 WW domain variant took longer to fold. However, after about 1.25 μs, it successfully folded close to the experimental structure within 2.0 Å. It's important to note that the folding simulation did not match the accuracy of the simulation started from the folded state, possibly due to a misalignment in the strand 2:3 interaction, which might be caused by limited sampling.

Despite this variation, the improvements in the force field showcased in this study are significant and represent a step forward in achieving more accurate and predictive protein folding simulations. [13]

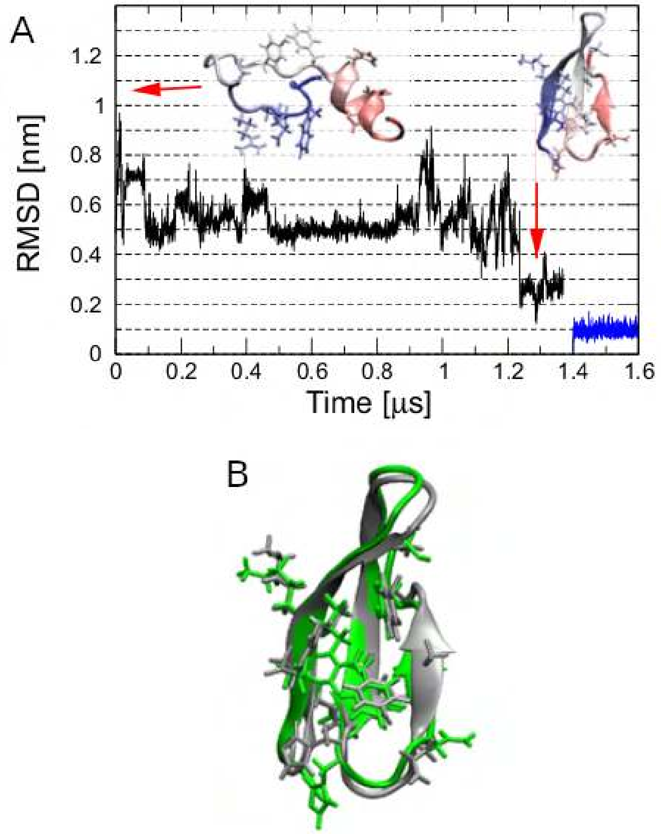


Figure 4 | *Protein folding trajectories of Pin1 WW domain (A) On the right side, the blue traces depict the backbone root mean square deviation (RMSD) of a 0.2-μs simulation initiated from the folded state of the protein. The simulation shows both the initial and folded structures, with a focus on highlighting the trajectory of each domain. The RMSD analysis considered only the structured region from residues 7 to 30 (B) The folded structures obtained from the simulations (in green) have been superimposed with the experimental structures (in silver) for the Pin1 WW domain, allowing for a comparison between the two sets of data. (reproduced without permission from Mittal et al., Biophysical Journal, 2010)*

1.3.4 Reversible folding and unfolding events of Fip35 protein and its GTT mutant, using Amber ff99SB-ILDN force field and supercomputer Anton, increased folding rate and stability in GTT mutant

Scientists conducted various experiments on Fip35 protein and on GTT protein which is a Fip35's mutant, containing three specific consecutive mutations on loop's 2 residues 26, 27 and 28. The new substitutions are glycine, threonine and threonine for 26, 27 and 28 residues respectively, hence its name. [14]

The two proteins were initially in an unfolded state and placed inside a cubic box of about 50 Å side length, surrounded by approximately 4000 TIP3P water molecules. The simulations were conducted at a temperature of 337 K, which corresponds to the predicted melting temperature for Fip35. The researchers selected the Amber ff99SB-ILDN force field, with recently improved side chain, to avoid excessive stabilization of β-sheet conformations. [15], [16] They managed to perform a total of 1197 μs of molecular dynamics (MD) simulations for both proteins. For GTT, they conducted four simulations with lengths of 83 μs, 118 μs, 124 μs and 272 μs, [14] while for Fip35, they performed four 100-μs simulations [14] and two additional 100-μs simulations from previous publication [17]. To carry out these extensive simulations, the researchers utilized the Anton supercomputer, known for generating trajectories up to 1 ms in length. [18]

Due to the modifications on GTT protein, they achieved their initial goal, i.e., to enhance and stabilize the formation of loop 2. For Fip35, the population of the folded state was 62% and for GTT was 74%. Compare to Fip35, GTT's stability and melting temperature increased by 0.5 kcal mol^{-1} and 7 K respectively. [14]

Using Amber ff99SB-ILDN force field, Fip35 was able to undergo various folding and unfolding occurrences under equilibrium conditions. The folding time for Fip35 was approximately 10 ± 3 μs, which closely matches the experimental folding time. These large-scale simulations [17], provided scientists with extensive data, leading to significant findings. They discovered that the roughness of the energy landscape plays a crucial role in determining the rates of conformational transitions in biological molecules. Moreover, the simulations confirmed the accuracy of the known force fields in representing the structure and dynamics of proteins, demonstrating their reliability. [17]

The duration of these simulations provided researchers with a unique opportunity to witness numerous occurrences of both folding and unfolding transitions taking place at the protein's melting temperature (**Figure 5**). [14]

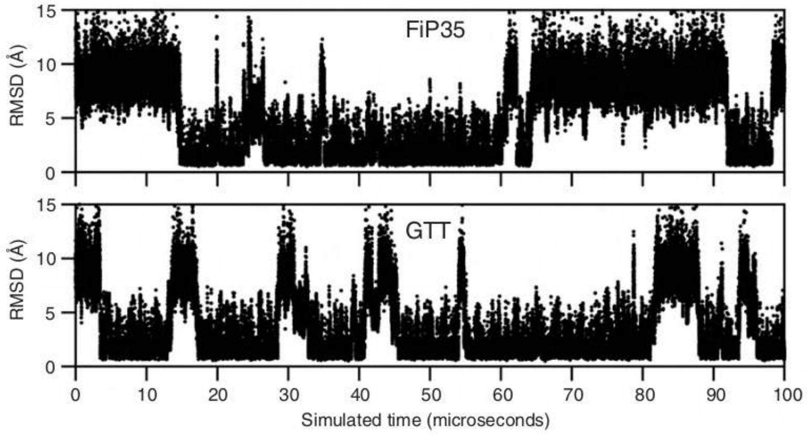


Figure 5 | *Reversible folding simulations of Fip35 and GTT. RMSD in two representative 100-µs MD simulations of FiP35 (top) and the GTT variant (bottom). (reproduced without permission from Piana et al., Journal of Molecular Biology, 2011)*

1.3.5 Evaluation of various MD force fields and development of an advanced one suited for both folded and disordered proteins, various proteins tested, including protein GTT, a Fip35 mutant of the WW domain

Scientists conducted an experiment to find a force field that can accurately describe both folded and disordered proteins. However, they discovered that none of the tested force fields were able to achieve this. Specifically, the force fields couldn't accurately represent folded protein structures, the dimensions of disordered proteins and the tendencies of disordered proteins to adopt specific secondary structures. [19]

The researchers tested six advanced force fields from CHARMM and Amber families. To assess their performance, they created a benchmark set containing 21 well-known proteins and peptides with diverse characteristics. This set included folded proteins with defined structures, fast-folding proteins, weakly structured peptides and disordered proteins with varying degrees of secondary structure. All data in the benchmark set were obtained from experimental techniques like X-ray crystallography and NMR spectroscopy. [19]

Amber force fields:

- ff99SB*-ILDN with the TIP3P water model
- ff99SB-ILDN with the TIP4P-D water model
- ff03ws force field containing empirically optimized solute-solvent dispersion interactions
- ff99SB-UCB force field with modified Lennard-Jones (LJ) parameters, dihedral modifications and TIP4P-Ew water model

CHARMM force fields:

- CHARMM22* with TIP3P water model
- CHARMM36m with TIP3P water model

Based on the simulations results, the researchers decided to make modifications to the parameters of Amber ff99SB-ILDN force field with TIP4P-D water model. These modifications aimed to improve the agreement between the simulation results and the experimental data for disordered proteins, while still maintaining the high accuracy of the force field for folded proteins. [19]
The resulting modified force field, called ff99SB-disp, contained modifications to the water model, backbone torsion corrections and the strength of a backbone O-H LJ pair, demonstrated excellent agreement with experimental observations for disordered proteins. [19]
During their simulations, among these 21 proteins and peptides, the GTT mutant [14] which is a fast-folding protein, was also tested. It was observed that the ff99SB*-ILDN force field [15], [20] exhibited the closest agreement with the experimental melting curves (**Figure 6**). For disordered proteins, ff99SB*-ILDN showed distorted dimensions and did not align well with the tendencies of residual secondary structure, although it performed well for folded proteins, small disordered peptides and fast-folding proteins. [19]

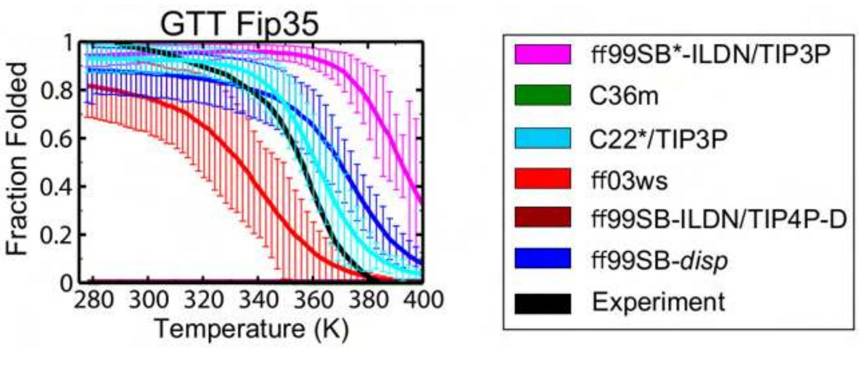


Figure 6 | *Stability of GTT Fip35 protein from simulated tempering simulations. Experimental melting curves are shown in black. No folded structures were observed with ff99SB-ILDN/TIP4P-D force field. (adapted without permission from Robustelli et al., Proceedings of the National Academy of Sciences, 2018)*

1.3.6 Exploring folding stability of Fip protein, thermodynamic contributions from 1μs-length folded/unfolded-state simulations using Amber ff99SB-ILDN force field

The scientists carried out two-state simulations, using the AMBER 18 package with the ff99SB-ILDN protein force field and the TIP3P water model. At the folded-state simulation, they used as initial structure, the structure with ID: 2F21 from PDB data base. The protein was solvated in water with chloride ions in a cubic periodic box. The simulation process involved energy minimization, equilibration in the NVT ensemble to gradually raise the temperature from 0 – 300 K and equilibration in the NPT ensemble at 300 K temperature and 1 bar pressure. Six independent trajectories with different initial velocities were generated by repeating the production MD simulation for a duration of 1 μs at T = 300 K and P = 1 bar. [21]

For the unfolded-state simulations, the initial configuration obtained from the NPT equilibration process in the folded state was used as a starting point. The system was first heated to 600 K under the NVT ensemble, subjecting the protein to heat-denaturation and then a simulated cooling process was conducted by gradually reducing the temperature by 50 K for every 1 ns of NVT ensemble simulation until reaching the final temperature of 300 K. Afterward, an NPT ensemble equilibration simulation was performed for 5 ns at 300 K and 1 bar pressure. Subsequently, they ran a 2 μs simulation and obtained eight separate trajectories with random starting velocities. The analysis focused on the last 1 μs segments of the unfolded-state simulations. [21]

Upon performing structural and thermodynamic analysis at these fourteen trajectories, the researchers gained valuable insights into the folding stability of the Pin1 WW domain. They were able to characterize the contributions of individual residues to the thermodynamic stability of the protein. The analysis considered both the backbone and side-chain interactions. Notably, β-sheet regions were found to have a favorable impact on folding, while turn and terminal regions had a destabilizing effect. Additionally, the formation of side-chain hydrophobic core regions played a crucial role in enhancing the thermal stability of the protein. This study provides a deeper understanding of the folding stability of the Pin1 WW domain and contributes to the broader knowledge of protein folding dynamics. [21]

1.4 Improved parameterizations resulted in the creation of two force fields, namely Amber ff99SB-ILDN and Amber ff99SB*-ILDN. These force fields were subsequently employed in our own MD simulations.

Prior to the development of the Amber ff99SB-ILDN force field, researchers made adjustments to the backbone dihedral parameters in the existing ff94 and ff99 force fields. These adjustments aimed to greatly improve the preferred conformations of typical secondary structures. Through the optimization of these parameters, they successfully enhanced the performance of the ff94 and ff99 force fields. Additionally, they specifically addressed the long-standing problem of glycine sampling in protein simulations. As a result, they developed a modified force field known as ff99SB. [16]

To further improve the accuracy of protein simulations, researchers conducted simulations using the ff99SB force field and encountered side chains that behaved differently than expected. Through a three-step process, they compared the behavior of χ_1 dihedrals in short helical peptides to the Protein Data Bank (PDB) to identify problematic residues. Four residues (isoleucine, leucine, aspartate and asparagine) showed significant deviations, indicating issues with the ff99SB force field. To address this, they created new χ_1 side-chain torsion potentials for these residues by adjusting force-field parameters based on advanced computational techniques. To validate the improvements, they compared the simulated results with experimental NMR data and found a closer match in the observed conformational states. This refinement resulted in the development of the enhanced Amber ff99SB-ILDN force field (ILDN is the specified code assigned to the side chains for which the potentials are adjusted). [15]

The improvements were exclusively targeted at dealing with the four specific residues (Ile, Leu, Asp and Asn) in the ff99SB force field. The adjustments made for these four residues can significantly influence the stability of protein structures and flexible regions. [15]

Some other scientists studied thoroughly the outcomes from prior MD simulations utilizing the previously revised Amber ff99SB force field. [16], [22] Subsequently, they decided to advance their research by investigating two peptides with helix-forming properties, Ala_5 and $Ac\text{-}(AAQAA)_3\text{-}NH_2$. Their intention was to make further adjustments to the ff99SB force field. Following torsional modifications (described below), a newly revised force field named Amber ff99SB* emerged. Their approach involved identifying the smallest necessary adjustment (**Equation 1**) to ensure that the force fields (ff99SB and ff99SB*) accurately replicated both the scalar couplings observed in Ala_5 and the proportion of helical structure in $Ac\text{-}(AAQAA)_3\text{-}NH_2$ at 300 K. [20]

$$\chi^2=\frac{1}{N}\sum_{i=1}^{N}\frac{\left(\langle J_i\rangle_{simu}-J_{i,expt}\right)^2}{\sigma_i^2}$$

Equation 1 | *The parameter χ^2 measures the agreement between experimental data and simulation results. It quantifies the mean-square deviation, normalized for uncertainty. For scalar couplings J, it is calculated using this equation that compares the simulated mean* $\langle J_i\rangle_{simu}$ *with the corresponding experimental value* $J_{i,expt}$.

Scientists selected the φ and ψ torsional parameters due to their direct correlation with the Ramachandran plot. They employed this concept as the foundation for modifying the ff99SB* force field. In an effort to prevent overfitting and reduce the number of independent variables, a simple cosine correction term is applied to the ψ torsion angle (**Equation 2**). This adjustment is particularly significant as ψ plays a major role in determining the propensity for helix formation. [20]

$$V_1(\psi;k_\psi,\delta_\psi)=k_\psi[1+\cos(\psi-\delta_\psi)]$$

Equation 2 | *Parameters k_ψ and δ_ψ represents the magnitude and phase offset of the corrections respectively. Corrections are applicable to all amino acids except glycine and proline. The resulting ff99SB* parameters are as follows: k_ψ = 0.1788 kcal/mol and δ_ψ = 105.4 deg*

After evaluating their findings, scientists noticed that the amount of helix present in the simulations does not significantly change with temperature. They concluded that this weak dependence on temperature is unlikely to be due to factors such as sampling, pressure, water model, simulation protocol, or how the "helix" is defined. Instead, they believed that there were deeper problems with the force field being used. Therefore, they conducted thermodynamic analysis, examining the energy-related aspects of the helix-coil equilibrium by separating it into enthalpic and entropic components. By employing a Bayesian approach, the researchers identified the optimal LR parameters *u* (enthalpic) and *w* (entropic) that describe helix formation for each force field and temperature. [20]

Despite overall improvements, the modified force fields still had issues with describing helix-coil transitions. The force fields underestimated the cooperativity of the transition and resulted in more fragmented long helices compared to experimental observations. The discrepancies can be attributed to the small magnitude of entropy loss and enthalpy gain during helix formation. Further corrections are needed to address these issues and improve the force fields. [20]

The adjustments made to the original ff99SB force field can be considered as a refinement process. However, when considering the impact on the overall conformational distribution, the correction terms have a significant effect. As a result, the scientists recommend using the modified force field, ff99SB*, for simulations involving weakly structured, unfolded peptides and proteins and also peptides that form β-hairpin structures. [20]

Over the course of time, many scientists, including those from our own laboratory, observing the results from simulations that used a force field from the Amber family, they decided to combine ff99SB-ILDN and ff99SB* in their simulations, leading to the development of a new and more advanced force field in the Amber family, referred to as ff99SB*-ILDN. [23]–[31], [32]–[34] This combination has gained significant popularity and has been widely adopted by researchers. In section (1.3.5), during the historical progression of WW domain folding using MD simulations, we already described an experiment conducted by other scientists, where they found that the ff99SB*-ILDN force field, in combination with the TIP3P water model, yielded the most stable simulations of proteins and showed the closest agreement with experimental chemical shifts and NOE violations. [19] These findings further support the effectiveness and reliability of the ff99SB*-ILDN force field.

In this thesis, we will present a comprehensive analysis of two pre-existing MD simulations. One simulation utilized the ff99SB-ILDN force field, while the other employed the ff99SB*-ILDN force field. Both simulations aimed to successfully replicate the folding dynamics of the WW domain's Fip mutant. We will discuss the findings and outcomes of these simulations in detail.

2. Analysis of Molecular Dynamics Simulations

The MD simulations were conducted using the NAMD software [35], which is a program designed for performing molecular dynamics simulations of large biological structures on powerful computer systems. The obtained simulations were mainly analyzed by using the user-friendly graphical user interface called Grcarma. [36] Grcarma provides a set of tools and pipelines built around the Carma program [37], which is used for analyzing molecular dynamics data. Carma is capable of processing different file formats such as PDB, PSF, and binary DCD files, which contain information about the simulated molecules.

We used the program Carma/Grcarma for most of our analyses, such as calculation of RMSD matrix, covariance matrix, RMS from average, radius of gyration, fraction of native contacts, dihedral and cartesian principal component analysis, determining secondary structure using the executable of the STRIDE program and pdb files portraying representative, average and superpositon structures. For the calculation of rmsf we used the plotting tool Grace, for the display of pdb files we used the molecular visualization program PyMOL and for the alignment of structures we used the protein complex structural alignment program MM-align.

My colleagues conducted two molecular dynamics (MD) simulations. A 10μs simulation attempting to fold a 35-residue amino acid sequence (see Introduction, section 1.2) using the ff99SB-ILDN force field with the Fip35 mutant as the reference structure. They also performed a 15μs simulation attempting to fold a 34-residue amino acid sequence (see Introduction, section 1.2) using the ff99SB*-ILDN force field, with the Fip34 mutant as a reference structure. Both simulations utilized the explicit TIP3P water model and the systems were equilibrated at 360K and 1 atm. [15], [20]

2.1 Comparing and highlighting differences in RMSD matrices and Secondary Structures derived from MD simulations using the ff99SB*-ILDN and ff99SB-ILDN force fields

The first step we chose to take in order to conduct an initial assessment of the two trajectories resulting from the ff99SB*-ILDN and ff99SB-ILDN force fields, is to quantify the differences and similarities of the structures adopted by the peptide chain throughout the simulation (RMSD matrix). The second step involves visualizing the secondary structures obtained throughout the simulation. By aligning these results accurately, we aim to observe how the two trajectories correspond (whether they exhibit transient structures, how many transient structures are identified, how long they persist, which secondary structures are acquired, whether they eventually attain the desired secondary structure and if they remain in it until the end).

Figure 7 is a graphical representation of the RMSD matrix and the secondary structure [28] of the trajectory derived from ff99SB*-ILDN force field. We selected the tasks RMSD matrix (specific parameters: CA atoms, first frame: 1, last frame: 15229200, step: 5076) and secondary structure (specific parameters: first frame: 1, last frame: 15229200, step: 508) from grcarma’s selection panel. The RMSD matrix (**Figure 7** top) compares all peptide conformations observed during the simulation and color-codes them based on their similarity. Conformations with low RMSD values are indicated in dark blue, representing highly similar structures, while the dissimilar structures are shown in yellow. The conformations observed along the diagonal, which starts from the origin of the axes (0,0), represent the persistent structures over time. The presence of off-diagonal blue areas indicates that a particular conformation has been visited multiple times during the simulation, appearing independently on different occasions.

The **Figure 7 (**bottom) displays the secondary structure assignments for the respective peptide conformations, aligning directly with the RMSD matrix. The α-helix is indicated in magenta, the turn in cyan, the β-sheet in yellow, and the random coil in white.

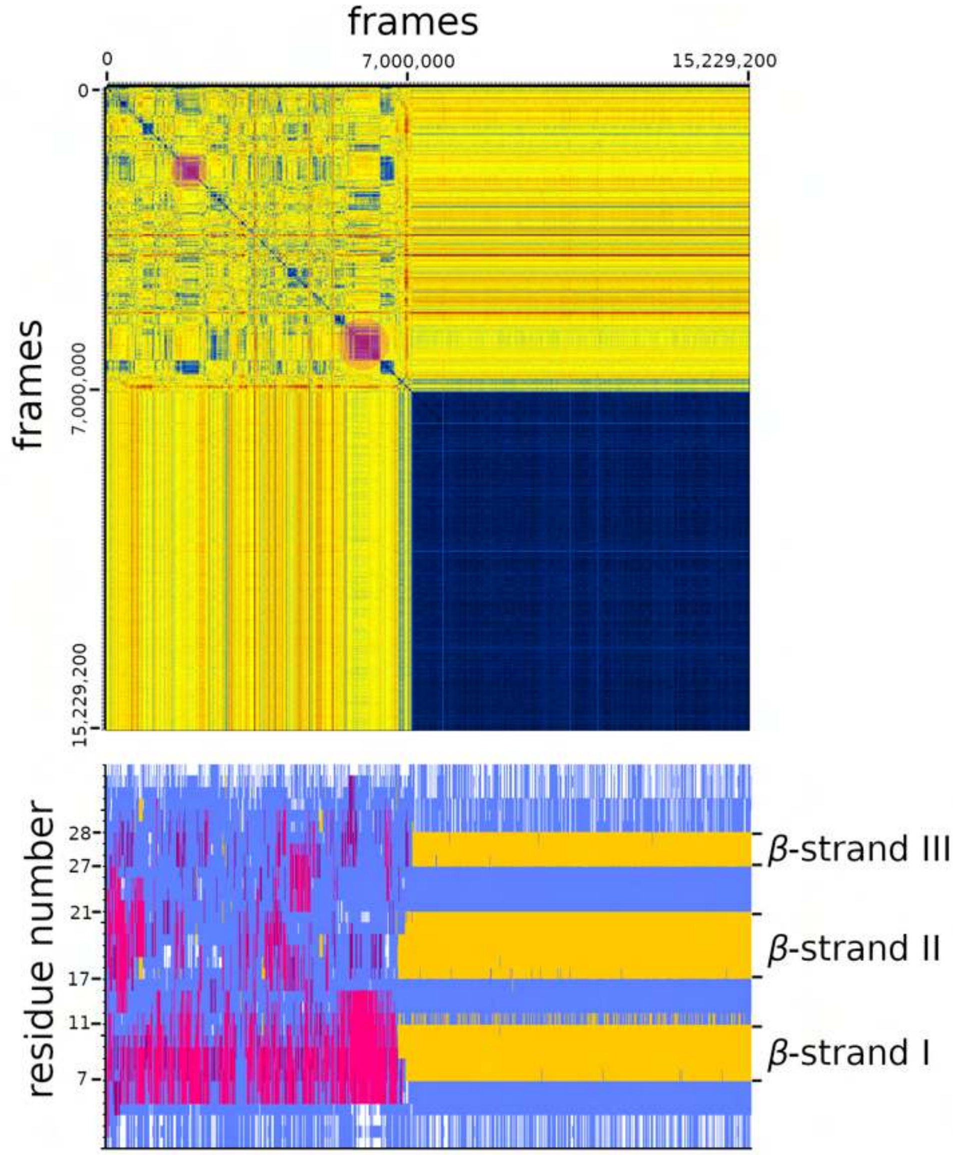


Figure 7 | *RMSD matrix (top panel), secondary structure (bottom panel) from ff99SB*-ILDN force field using the whole length of the trajectory. The RMSD matrix, top panel, is perfectly symmetric and has been computed based on the peptide's C_α atoms. The bottom panel represents the secondary structure assignments per residue.*

By examining both panels in **Figure 7**, we observe the followings: During the initial 7,000,000 frames, the chain is mostly disorder and rapidly switches between various conformations, mostly *α*-helices and turns, but only a few of them are relatively stable and can be observed during the simulation. We can detect some of them in the red circled areas on the RMSD matrix of **Figure 7**.

After undergoing various transient conformations, it appears that at ~7,200,000th frame, the peptide adopts a final conformation and remains in it for the remaining simulation time. The final conformation composed of turns and a β-sheet consisting of three β-strands. The first β-strand is formed by 5 residues (7th-11th), the second again by 5 residues (17th-21st) and the third by 2 residues (27th-28th). It is noteworthy that among the various transient conformations, there are almost no observed β-sheet structures, which are the dominant secondary structures in the simulation's last approximately 8,000,000 frames. Instead, a-helices, turns and random coils are observed.

Next, we present **Figure 8** which provides a graphical representation of the RMSD matrix and the secondary structure of the trajectory derived from ff99SB-ILDN force field. We selected the tasks RMSD matrix (specific parameters: CA atoms, first frame: 1, last frame: 10,021600, step: 3341) and secondary structure (specific parameters: first frame: 1, last frame: 10,021,600, step: 334) from grcarma's selection panel.

In **Figure 8**, the same characteristics apply as in **Figure 7** regarding the features of both graphical representations (such as the color coding) and their interpretation. Once more, the secondary structure assignments are in direct alignment with the RMSD matrix. In **Figure 8**'s top panel, the RMSD matrix reveals that during the initial 8,500,000 frames of the simulation, we observe our chain transiently adopting a few conformations, which it holds for a brief period before unfolding once more. At approximately 8,500,000th frame, it undergoes a final folding event, forming a stable structure that persists until the end of the simulation.

Upon initial observation, in **Figure 8**'s bottom panel, it is clear that throughout the simulation, there are many turns and β-sheets. Looking closer at the region "a" in the graph, we notice something interesting, the structure formed is a β-sheet structure with three strands. Surprisingly, our chain adopts this structure within the first 3,800,000 frames. Unfortunately, it doesn't last long and is only maintained for ~ 600,000 frames. We can also notice some other β-sheet structures in the first 8,500,000 frames, like the region b, which has a β-sheet with two strands. However, this is also temporary, lasting for only ~ 400,000 frames. At approximately 8,500,000th frame, region c, it folds one final time and remains in this conformation until the end. The final conformation is different from the one obtained from ff99SB*-ILDN force field, as it includes an α-helix consisting of 7 amino acids at its c-terminus, along with turns and a β-sheet composed of two β-strands. What is interesting is that, during the initial 8,500,000 frames, there are no significant α-helices observed, but in the final conformation, among other elements, there is a relatively large α-helix considering the total size of the chain.

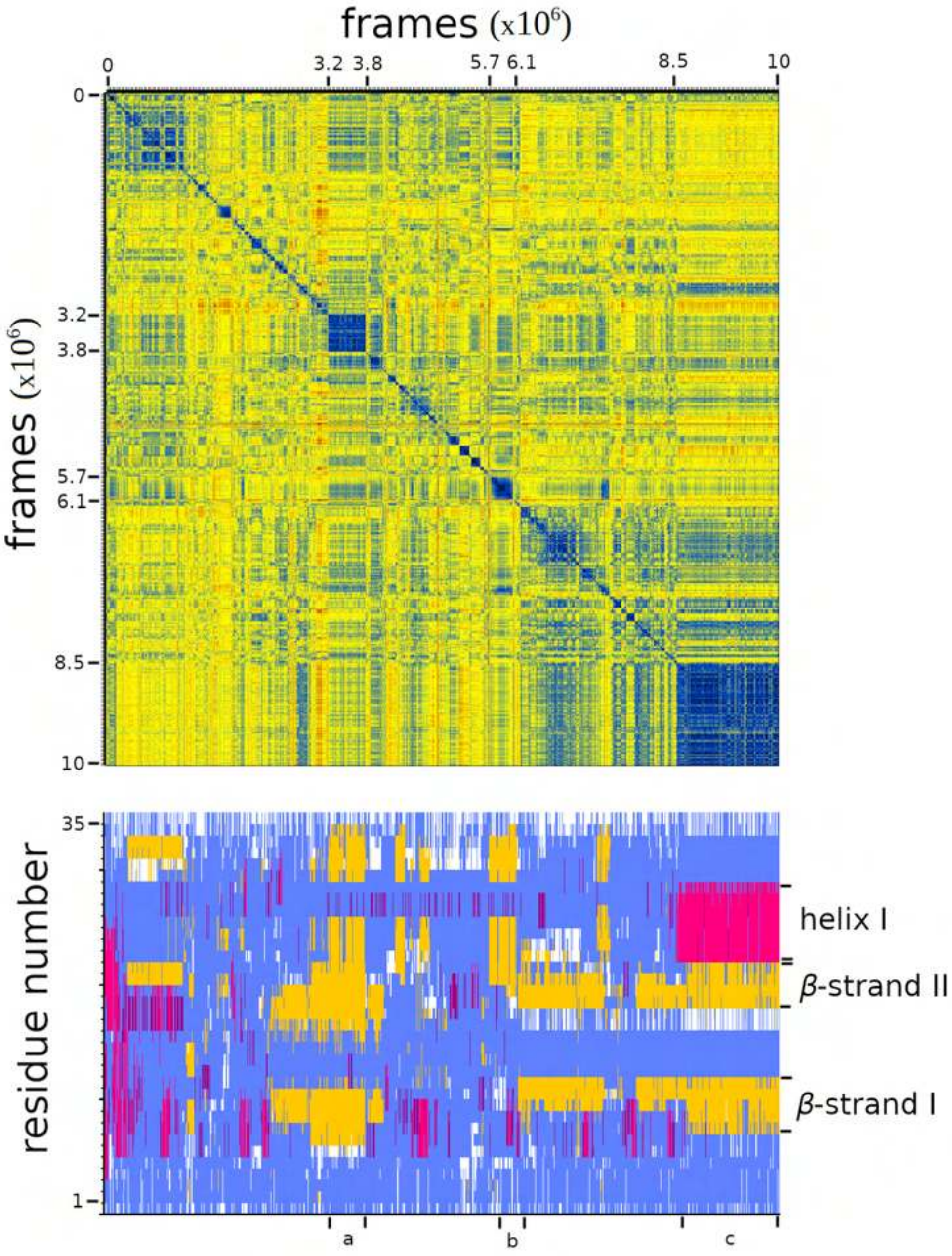


Figure 8 | *RMSD matrix (top panel), secondary structure (bottom panel) from ff99SB-ILDN force field using the whole length of the trajectory. The RMSD matrix, top panel, is perfectly symmetric and has been computed based on the peptide's C_α atoms. The bottom panel represents the secondary structure assignments per residue.*

Concluding this section, we will attempt to explore in greater depth the secondary structure that has been formed using the ff99SB*-ILDN force field. We provide **Figure 9**, which presents a graphical representation of the RMSD matrix and the secondary structure of two distinct segments extracted from the trajectory. In order to do that, we will attempt to isolate specific frames, meaning we will zoom in on our structure. By using a smaller step, we will be able to achieve greater clarity. The specific parameters below specify these two segments in frames. We selected the tasks RMSD matrix (specific parameters: CA atoms, first frame: 7,213,004, last frame: 15,229,200, step: 500) and secondary structure (specific parameters: first frame: 6,800,000, last frame: 7,300,000, step: 50) from grcarma's selection panel.

In **Figure 9** (top), our graphical representation reveals a single color (blue) as a result of this isolation, indicating low RMSD values, as we previously discussed. This was the expected outcome, as we visually identified the limits where our final conformation forms through the initial analysis of the RMSD Matrix (**Figure 7**, top). Consequently, we successfully isolated our final conformation, which was attained approximately at frame 7,213,004 and maintained throughout the entire simulation.

In **Figure 9** (bottom), through the selection of frame 6,800,000 as the starting point and frame 7,300,000 as the endpoint, we were able to zoom in on the phase in which the formation of our final conformation, the β-sheet, initiates. This detailed analysis allowed us to distinguish the order in which the individual β-strands of the β-sheet emerged. It was observed that β-strand I and β-strand II formed simultaneously, while β-strand III appeared approximately 400,000 frames later. These findings align with existing literature about β-proteins's WW domain [9], [17], [38]–[43], which proposes that, β-strands I and II undergo initial formation via loop 1, followed by the subsequent creation of β-strand III via loop 2 located between β-strands II and III.

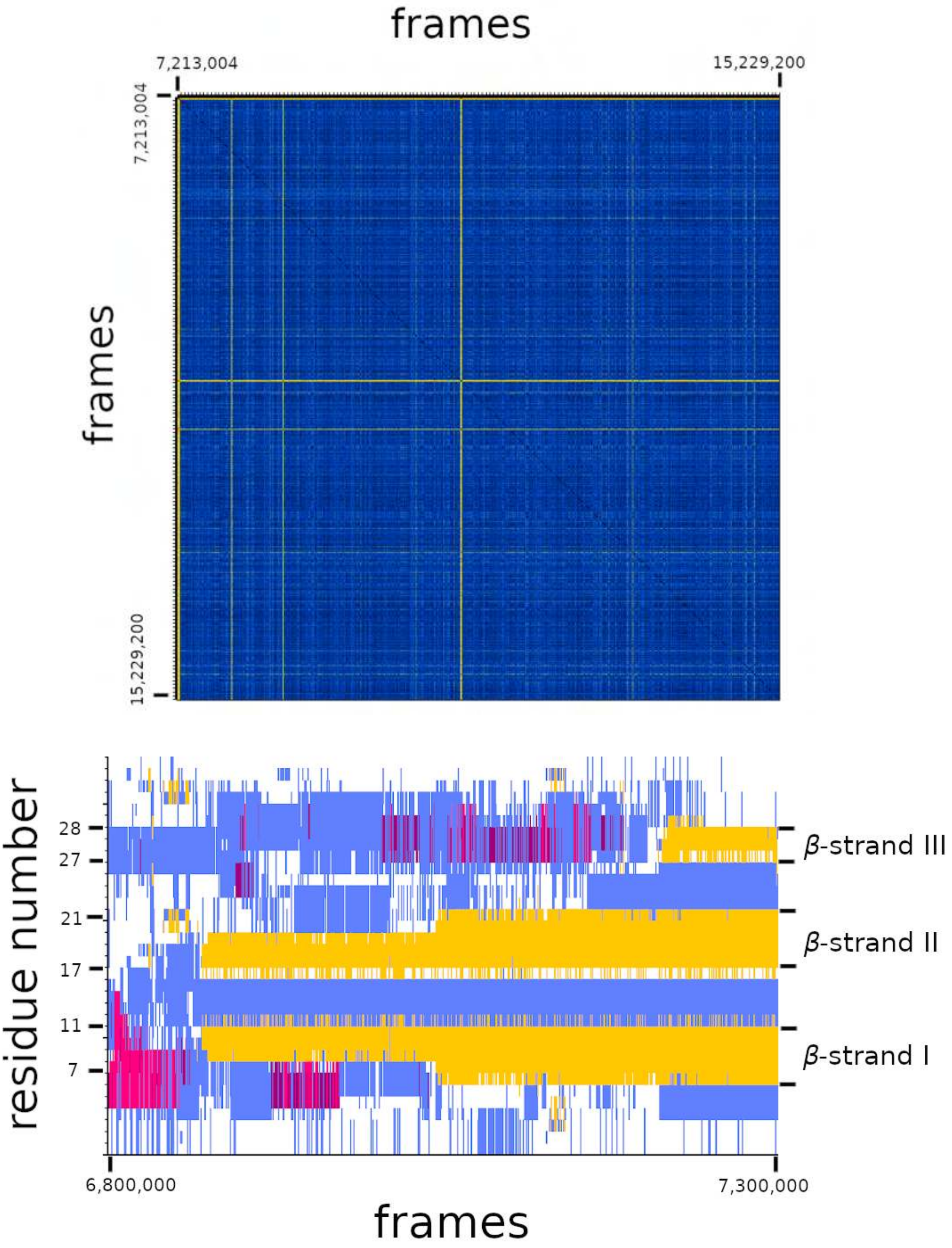


Figure 9 | *RMSD matrix (top panel), secondary structure (bottom panel) from ff99SB*-ILDN force field using two specific segments of the trajectory.*

2.2 Exploring the ff99SB*-ILDN force field, covariance analysis, RMS from average, RMS fluctuation and representative structure visualization

In the following phases of our analysis, we will basically focus on the trajectory generated by the ff99SB*-ILDN force field. Based on the results of the RMSD matrix and secondary structure, this trajectory was able to acquire the desired secondary structure, the β-sheet and maintain it until the end of the simulation. Therefore, we want to further analyze this trajectory to determine if this particular force field can indeed produce a protein that resembles the native protein.

In the current phase of the analysis, we will calculate the trajectory's covariance matrix [44], RMS from average and representative structure. To narrow down our analysis to the phases where the final conformation has already been reached, we isolated these specific frames and performed the analysis solely on them. We based the frame selection on the outcomes we observed during the analysis we already performed on the previous section (**Figure 9**). The selected limits (in frames) are 7,300,000 (lower limit) and 15,229,200 (upper limit).

By executing the tasks: (Covariance, average and representative structures) from the task selection panel, several files are generated. A graphical representation of the covariance matrix, which captures the relationships among different elements in the dataset, a graphical representation of RMS from average, pdb files to visually represent the average structure and a superposition of 500 structures generated from the trajectory and a pdb file to visually represent the trajectory's representative structure. [36]

A crucial step before executing the aforementioned tasks is to superimpose/fit all the structures within these frames, allowing them to align with one another. Additionally, in the Atom Selection field, we chose Backbone over CA, which includes the four atoms N, C, O, and CA, providing greater accuracy in our analysis. We also clicked the option: (Use dot product (needed for average structures)) and lastly, we normalized our data by clicking the option: (Calculate normalised matrices).

After several unsuccessful attempts, it was deemed necessary during the fitting phase to introduce an additional adjustment/parameter (use a subset of the residues for the fitting) and in the residue selection, amino acids from 4 to 31 were chosen. With this adjustment, we instruct the program to perform the fitting based on the motion of amino acids 4-31 while disregarding the movements of the terminal amino acids. The reason for this is that the terminal amino acids, due to their greater freedom of motion, introduced "noise" and distorted the results. This allows us to reduce the "noise" without excluding any amino acids during the analysis phase.

2.2.1 Calculation and insights from the covariance matrix, exploring the relationship among amino acids

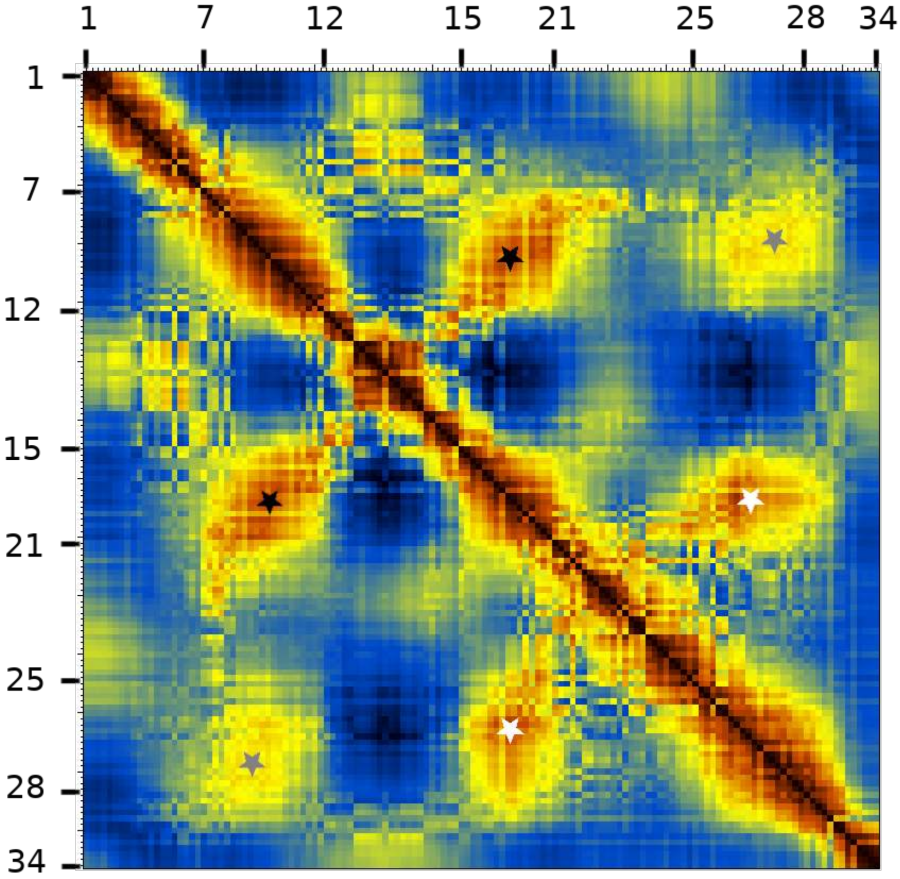


Figure 10 | *Covariance matrix*

Through the analysis of the covariance matrix in **Figure 10**, we can comprehend how the 34 amino acids of our peptide chain are correlated with each other once the chain has received its final conformation. Specifically, it shows the relationships among all amino acids, not just neighboring ones. This graphical representation utilizes a specific color code, where dark red regions represent correlated amino acids that are connected to each other, yellow regions indicate amino acids that influence each other to a lesser extent and dark blue regions represent uncorrelated amino acids that are independent from each other.

The color observed along the diagonal, which divides the square exactly in the middle, is dark red. This is expected, as it signifies the relationship of each amino acid with itself. The most significant relationships observed in this phase are related to the amino acids that compose the three β-strands. Specifically, three main relationships have been identified. It is evident that the amino acids of the first β-strand (amino acids 7-12) are correlated with the amino acids of the second β-strand (amino acids 15-21), indicating that the motion of the one β-strand follows the motion of the other. We can observe this relationship in the region where the two black stars are located. The same relationship is observed between the amino acids of the third β-strand (amino acids 25-28) and those of the second β-strand, as indicated by the region with white stars. The region with two gray stars represents the relationship between the amino acids of the first and third β-strands. The fact that we see predominantly yellow color and not red in their relationship indicates that even though both are correlated with the same strand, the second strand, they are not fully correlated with each other but they are correlated to a lesser extent.

2.2.2 Calculation of the root-mean-square from the average structure, quantifying deviations throughout the simulation

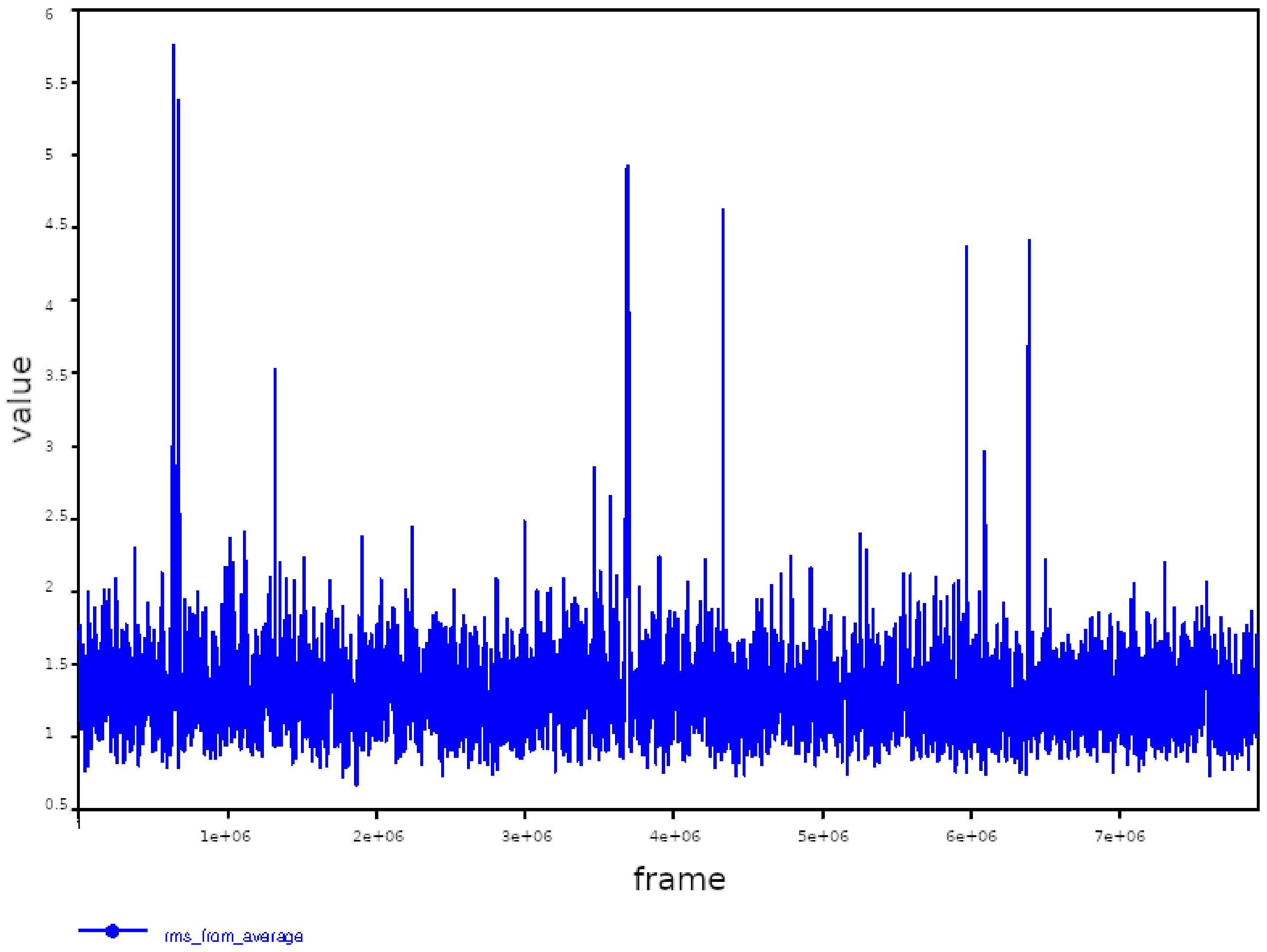


Figure 11 | *RMS from Average*

In **Figure 11**, we have created a diagram that represents the root-mean-square (rms) from the average. The information provided relates to the average structure of each coordinate/frame and compares the values obtained during the simulation to the average structure of our protein. The larger the value, indicating a higher vertical line, the greater the deviation between this average coordinate and our average structure. The frame with the minimum value is the one that most closely resembles the average and this is our representative structure.

2.2.3 Root-mean-square fluctuation analysis, exploring structural flexibility at an atomic level

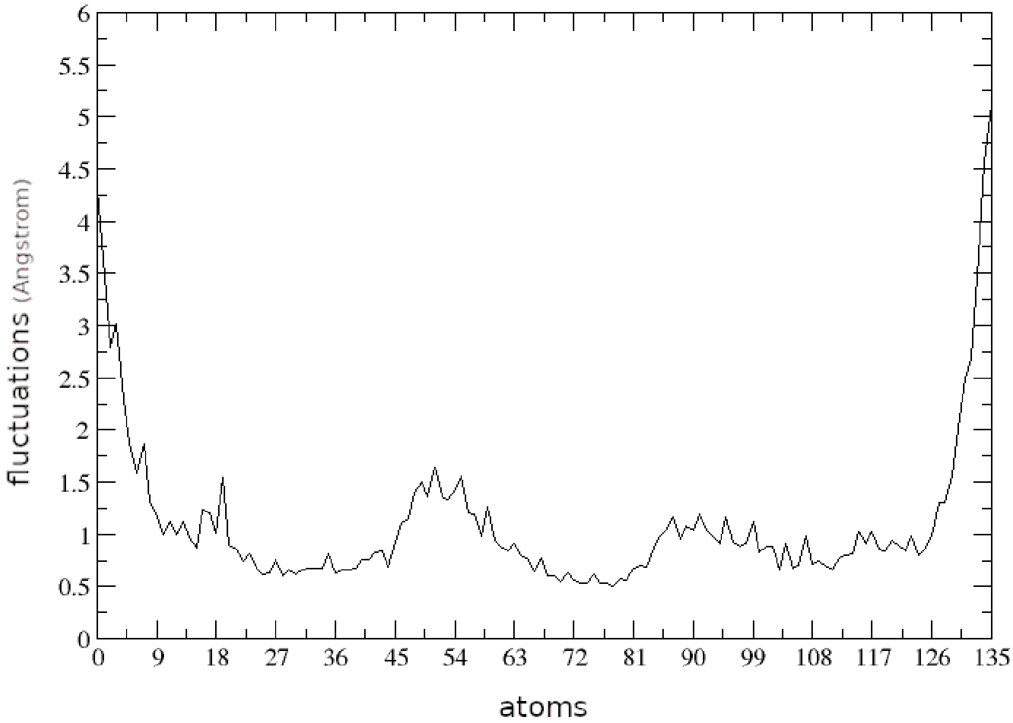


Figure 12 | *Root mean square fluctuation*

In **Figure 12**, we created a plot that represents the RMSF (root-mean-square fluctuation) of the protein's average structure. For the plot, we extracted data from the pdb file of the average structure and used the Grace tool. On the x-axis, we used the 136 atoms of our chain and on the y-axis, we have the values in Angstrom indicating the mobility/fluctuation of each atom in the chain. The smaller the value, the less movement the atom exhibits in space. By observing the shape of the diagram, we can see that the regions with lower values correspond to the atoms responsible for the formation of the three β-strands (atoms 28-45, 60-84 and 100-112). The regions between the β-strands, which correspond to the loops, have higher values, indicating greater mobility. Even higher values are observed for the atoms at the two ends of the chain. Therefore, this diagram confirms what we already know, that when our peptide chain has formed the β-sheet, the amino acids that form the β-strands are more stable in space, the intermediate regions, i.e., the loops, exhibit more movement and finally, the two ends have the highest degree of freedom. We had previously noticed that the two termini showed significant mobility when attempting to run the task: (Covariance, average and representative structures), leading to "noise" and negative impacts on our results.

2.2.4 Visualizing the representative structure, insights into protein conformation and mobility

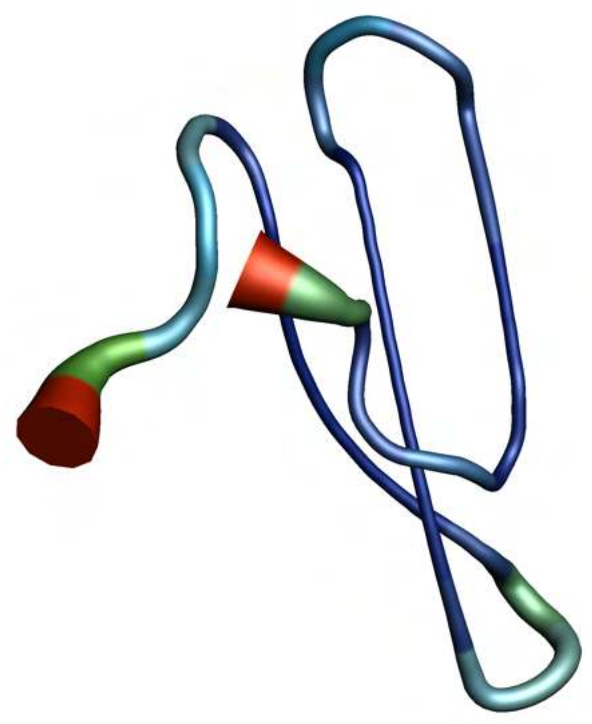


Figure 13 | *Representative Structure visualization with termini*

Using PyMOL, we created **Figure 13** by performing a visualization of our representative structure. The input for this image was the pdb file of the representative structure, generated earlier during the task: (Covariance, average and representative structures). For the visualization, we selected the "cartoon putty" style, which highlights the overall shape and folding pattern.

We selected the specific configuration (cartoon putty) as it allows us to visually discern the mobility of different segments. Essentially, this command displays the stable regions as very thin and increases their thickness as their mobility increases. Therefore, in this particular image, we observe the very thin regions that constitute the β-sheet itself and the thicker regions at the termini. However, in **Figure 12**, we observed that the β-sheet does not exhibit the same mobility throughout its length. The three β-strands appear almost immobile, while the two loops connecting the β-strands show significantly higher mobility. The reason why, in **Figure 13**, the entire β-sheet (loops and β-strands) appears uniformly thick is because the shape obtained when applying the cartoon putty style reflects the overall mobility of the structure. The much greater mobility of the structure's two termini essentially overshadows any mobility exhibited by the rest of the structure, making it nearly negligible in comparison. The information provided by the cartoon putty representation is not limited to the structure's morphology but also incorporates a specific color code. Additionally, the color code used enriches the aforementioned conclusion. Dark blue represents immobility, light blue represents moderate mobility and red represents high mobility. Thus, the stable segments of the protein are blue, the segments connecting the β-strands are light blue and the termini are red. The conclusions drawn here align with those of the diagram in **Figure 12**.

Furthermore, we generated an additional pdb file for the representative structure (**Figure 14**), excluding amino acids 1-3 and 32-34, aiming to assess the extent to which their presence affects our results. Upon comparing **Figures 13** and **14**, it becomes apparent that the considerable oscillation of these specific amino acids does indeed overshadow the oscillation of the β-sheet. In **Figure 14**, the difference in oscillation intensity between the two loops and the three strands is more distinctly highlighted, both in chromatic variation and thickness.

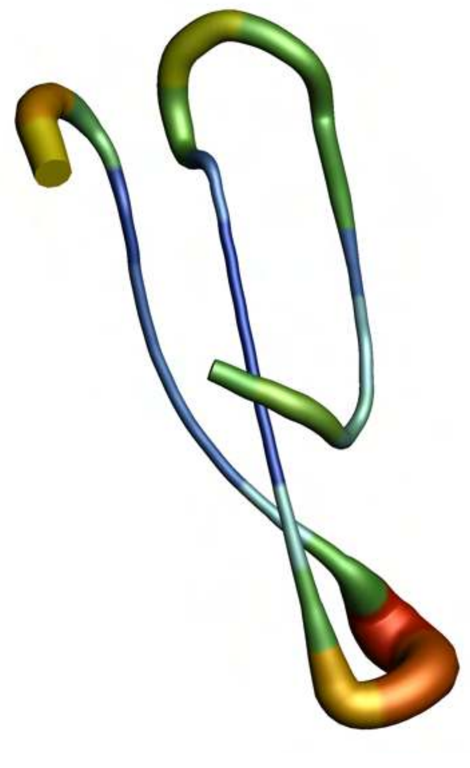


Figure 14 | *Representative Structure visualization without termini*

2.3 Radius of gyration analysis in the ff99SB*-ILDN trajectory, exploring the compactness of the structure

In the current analysis, we use the ff99SB*-ILDN trajectory without superposition/fitting. We analyze the entire simulation range, not a specific part. We selected the task Radius of gyration (specific parameters: Heavy atoms, All residues, first frame: 1, last frame: 15229200, step: 1)

In **Figure 15**, we employed the Radius of gyration [45] task. By using the radius of gyration (Rg) calculation, we can assess the distance among atoms, meaning how spread out the mass of our structure is around its center of mass. It gives us an idea of the compactness of our structure. Mathematically, Rg is defined as the root mean square distance between each atom in the molecule and the center of mass. By analyzing the graph in the figure, we can observe that from the beginning of the simulation until just before the final folding occurs, we mainly obtain large values for the radius of gyration, indicating that our structure is quite extended in space (unfolded), which is reasonable since it has not yet adopted its final conformation. However, in frames 2e+06 and 6e+06, for a certain period of time, we can observe low values in the radius of gyration, indicating the formation of transient structures that are sufficiently compact enough to appear in this graph. These two transient structures can also be observed in the two red circled areas on the RMSD matrix (**Figure 7**). Upon matching them with the secondary structure analysis (**Figure 7** bottom), it becomes evident that they are mainly composed of α-helices. After the 7.2e+06th frame, the structure once again undergoes folding and becomes even more compact, a fact that is also confirmed by the RMSD matrix in **Figure 7** (the big dark-blue square).

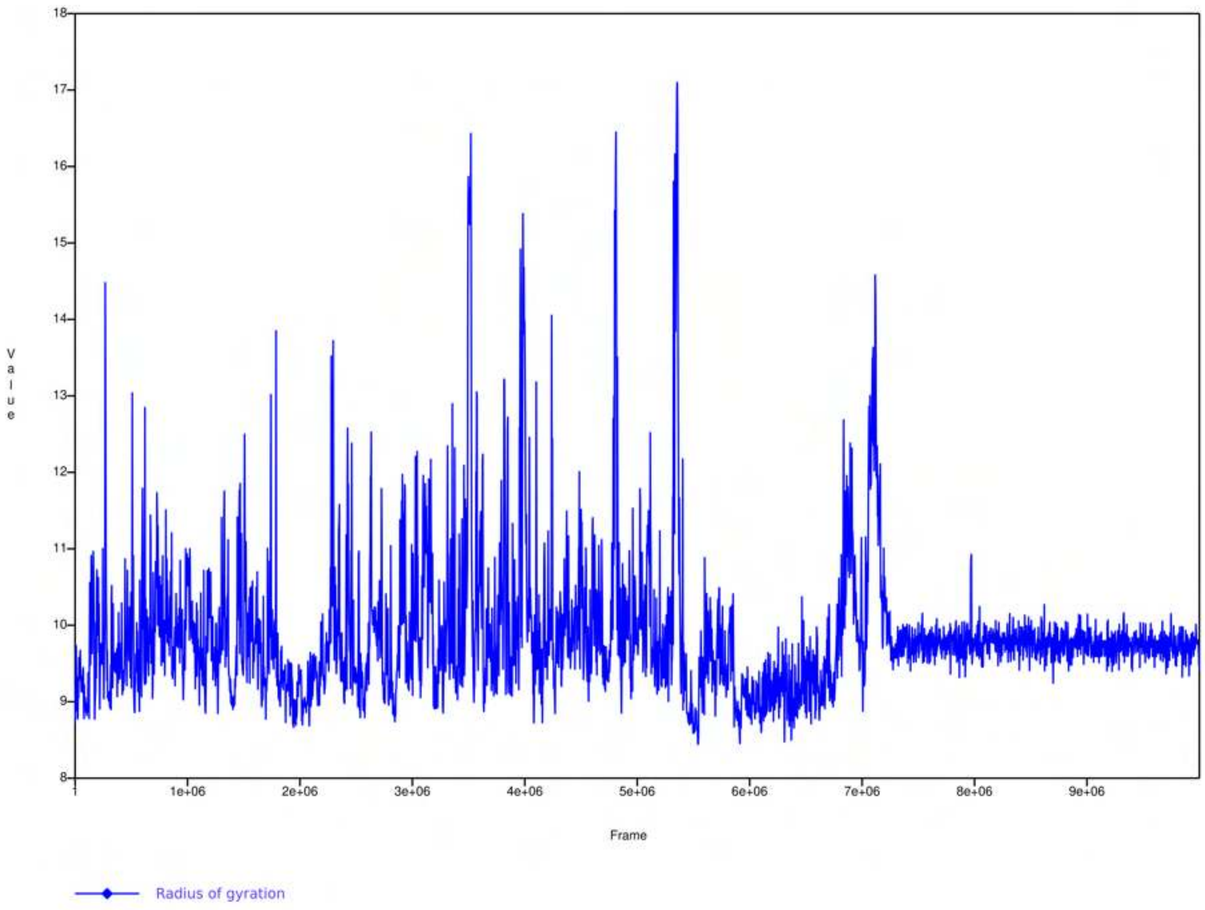


Figure 15 | *Radius of gyration diagram*

2.4 Exploring fraction of native contacts, bond distances insights in the ff99SB*-ILDN trajectory

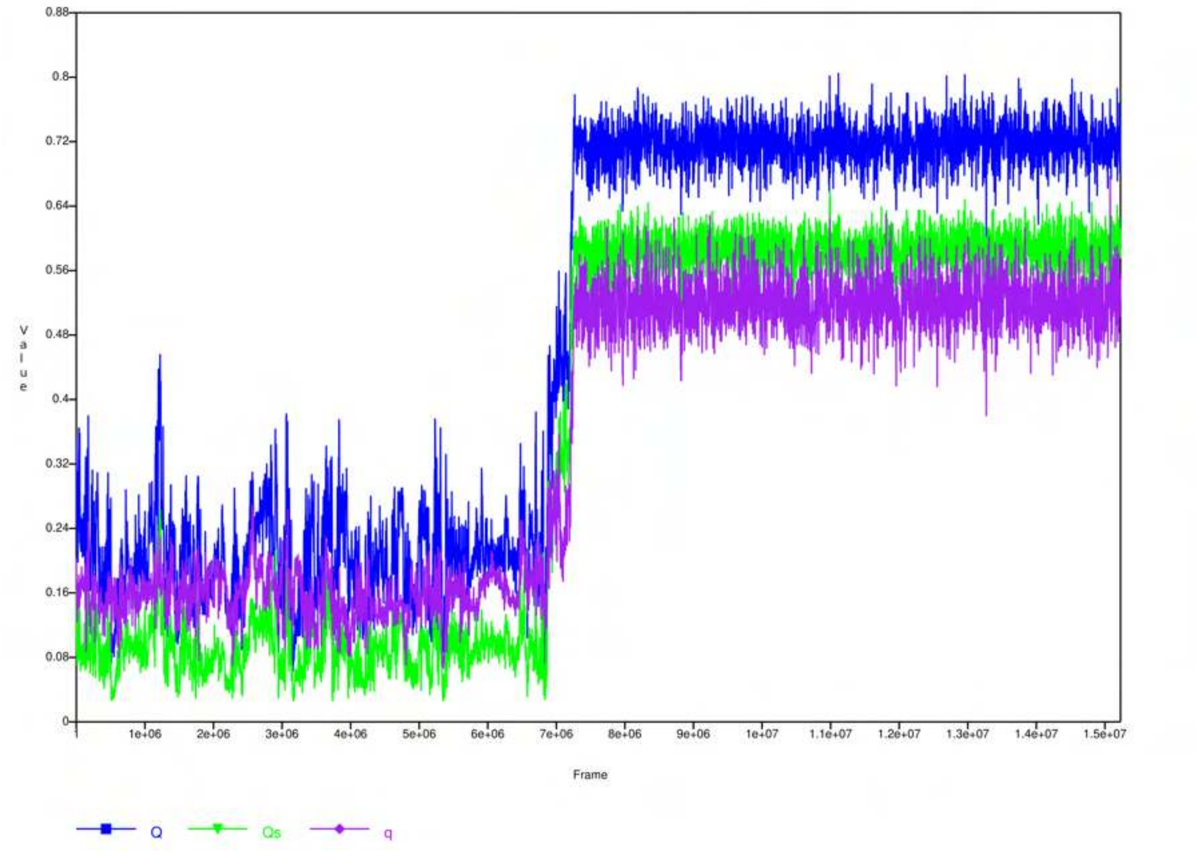


Figure 16 | *Graph of Fraction of Native Contacts: x-axis represents the duration of the whole simulation in frames, y-axis shows the percentages of Q, Qs and q values.*

Continuing our analysis of the trajectory (ff99SB*-ILDN) using the full length, we aim to explore the structure in greater detail. For this purpose, we chose the "Fraction of native contacts" [46] task from the Grcarma's task selection panel with the following specific parameters: backbone atoms, all residues, first frame: 1, last frame: 15229200, step: 1. Additionally, we selected the option "Use PDB file to define the native structure" and as observed in the experimentally-determined structure of the protein, we used "FipWW_backbone34.pdb" as a reference, which is a modified pdb file containing only the backbone atoms of the 34 amino acids.

To further assess the entire trajectory leading up to the successful formation of the final conformation, fraction of native contacts (**Figure 16**) introduce us to three new terms. Specifically, we refer to reaction coordinates which give us insights about the protein's native structure.

The blue line, representing the Q value, plays a crucial role in predicting transitional stages during the simulation. It identifies the bonds within our structure that fall within a defined cutoff distance, relative to the corresponding distances of the “natural” bonds in the reference structure. Essentially, for each frame, it calculates the ratio of these bonds within the cutoff range to the total number of bonds in the reference structure. The range of values varies from 0 (no matching) to 1 (100% matching).

The green line represents Qs, providing more precise information about the distances of the “natural” bonds compared to Q. Specifically, it shows the deviation between our bonds’s distances and their expected distance when the structure is in its native state. The range of values is the same as Q, from 0 (the bond distances are significantly different from the natural ones) to 1 (the bond distances are exactly the same as the natural ones).

To gain more insights from this analysis, we included the pdb file of the protein’s native structure, downloaded from the protein database, to compute the normalized similarity value q. This allows us to compare the native structure from the database to our simulated structure. q uses the same range of values, with 0 indicating very different structures and 1 indicating identical structures.

Observing the graph in **Figure 16**, we notice two distinct phases where the results of all three values align. From the first frame up to the point just before our chain folds into the final conformation, the values for Q range from approximately 0.08 to 0.47, Qs ranges from about 0.04 to 0.26 and q ranges from around 0.08 to 0.26. This indicates that during this phase, a comparison with the reference structure reveals that the bond distances in our structure, which fall within the designated cutoff range, range between 8% to 47% for Q and 4% to 26% for Qs and the values for q suggest that the structure at this stage significantly deviates from the reference structure. In the second phase of the analysis, from the point of achieving the final conformation until the end of the simulation, there is a steep increase in the values of all three reaction coordinates. We have found that a significant portion, ranging from 64% to 80%, of our bonds distances are within the cutoff range (Q value). Furthermore, we have noticed a significant reduction in the deviation between the distances of our bonds and the corresponding distances in the reference structure, leading to a notable increase in the Qs percentage, ranging from 46% to 65%. Finally, the similarity between the two structures increased 40% to 65% (q value).

2.5 Combined outcomes from fraction of native contacts and radius of gyration analyses, provide additional insights

2.5.1 Comparing q vs Rg diagrams derived from full vs excluded terminal residues in ff99SB*-ILDN trajectory

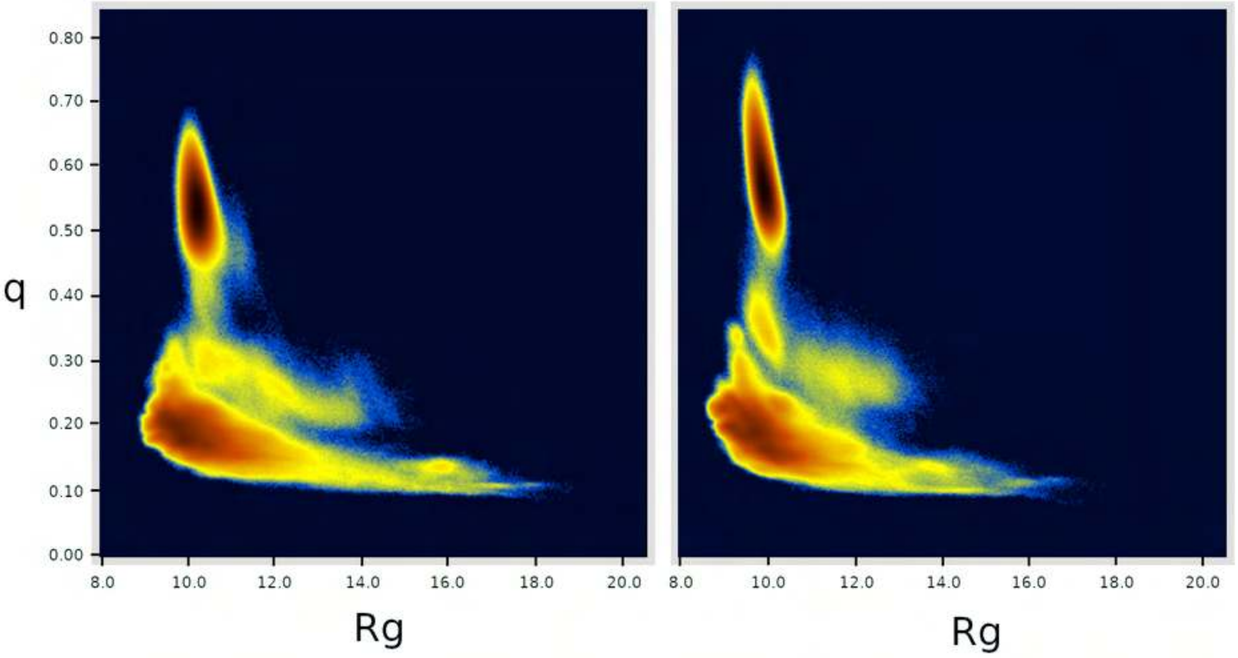


Figure 17 | *q vs Rg graphs, left graph corresponds to the entire sequence, right graph focuses on residues 4 - 31*

In **Figure 17**, our focus was on the ff99SB*-ILDN trajectory. To create these two graphs, we first calculated the fraction of native contacts and the Radius of gyration using the GUI grcarma. Among other outputs, two files were generated: Rgyration.dat and Qfraction.dat. The Rgyration.dat file contains information about the Radius of gyration (Rg), while the Qfraction.dat file contains information about the distances of the contacts (Q and Qs) and the similarity (q) of our structure compared to the reference structure (native structure).

For this specific analysis (**Figure 17**), we disregarded the columns Q and Qs and focused on the column q. We combined Rg and q and created a scatter plot using a specific color code. Each dot in the graph represents one frame, so, each graph contains 15,229,200 dots. These graphs are presented on a logarithmic scale, which effectively highlights both the folded and unfolded conformations. The regions shown in dark red indicate a high concentration of structures in that particular area of the graph, blue represents a lower concentration and yellow indicates an intermediate state.

For the left image, we considered the entire amino acid chain. We used the already generated files Rgyration.dat and Qfraction.dat which are derived from sections 2.3 and 2.4 respectively.

For the right image, we excluded the first three and last three residues. Specifically, during the Radius of gyration task, we selected heavy atoms, residue selection 4 – 31, first frame: 1, last frame: 15229200 and step: 1. During the Fraction of native contacts task, we selected backbone atoms, residue selection 4 – 31, first frame: 1, last frame: 15229200, step: 1. Additionally, we selected the option "Use PDB file to define the native structure" and we used "FipWW_backbone4_31.pdb" as a reference, which is a modified pdb file containing only the backbone atoms of the 28 amino acids. By removing these terminal residues, known for their significant mobility, we try to understand to what extent they influence the results of this analysis.

In both images we observe two regions with dark red. The first one is located at low values of both Rg and q. In this region, we expect to find compact structures due to low Rg values but with low similarity between the natural and the examined structures due to low q values. These structures correspond to transient conformations. The second dark red region is situated at low Rg values and high q values. Here, we anticipate finding compact structures with high similarity between the natural and examined structures. This region contains all the frames where our chain has adopted its final conformation. In fact, the highest concentration of structures is found in this second region, which aligns with the fact that out of the 15,229,200 frames, which is our entire simulation, approximately 7,900,000 frames show our chain in its final conformation.

In both images of **Figure 17**, we observe exactly the same distribution of conformations adopted by the chain, including transient and final conformations. But, there is a difference in q values, with the truncated peptide chain having a maximum q value of 0.8061, whereas the whole peptide chain has a maximum q value of 0.7148. Observing the two graphs, indeed, the second dark red region of the truncated peptide chain is slightly shifted upwards. However, the differences observed between the two images were found to be negligible.

2.5.2 Comparing q vs Rg and Q vs Rg diagrams derived from both ff99SB*-ILDN and ff99SB-ILDN trajectories

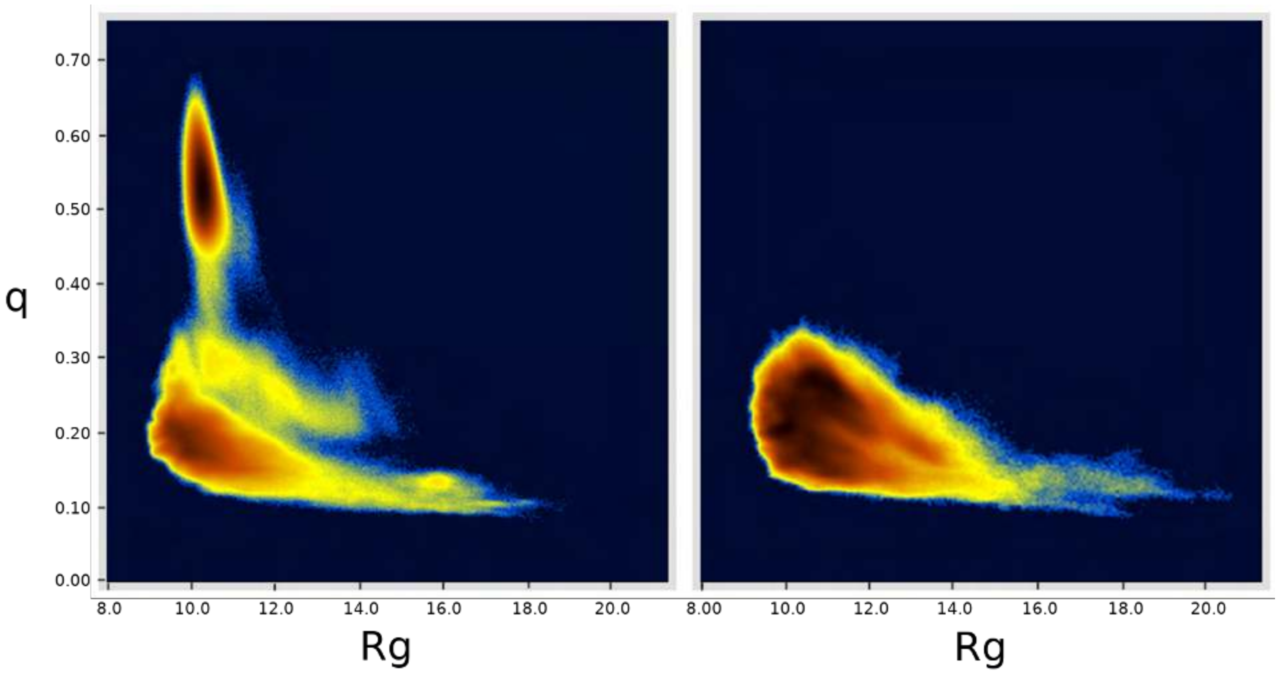


Figure 18 | *q vs Rg graphs, left graph depicts the ff99SB*-ILDN trajectory using the entire sequence, right graph corresponds to the ff99SB-ILDN trajectory using the entire sequence*

For the following analyses, both trajectories, ff99SB*-ILDN and ff99SB-ILDN, were used. In **Figure 18**, we aimed to compare the q vs Rg analysis between the two trajectories. The left graph of **Figure 18** is exactly the same as the one in **Figure 17** (left) and corresponds to the ff99SB*-ILDN trajectory, while the right graph of **Figure 18** corresponds to the ff99SB-ILDN trajectory.

To create the above graph, we executed the Fraction of native contacts and Radius of gyration tasks for our second trajectory (ff99SB-ILDN) and used the same graph from **Figure 17** (left) for our first trajectory (ff99SB*-ILDN). Before running the Fraction of native contacts task for the ff99SB-ILDN trajectory, we had to choose a pdb file to define the native structure. The initial amino acid sequence they attempted to fold using the ff99SB-ILDN force field contained 35 residues. Therefore, the pdb file we chose to use as a reference also had to contain 35 residues. This additional residue in the ff99SB-ILDN pdb file includes an extra GLY at the beginning of the chain that is missing in the other pdb file. The analysis of the reference sequences has been detailed in the introduction, in section 1.2.

Knowing that the first trajectory's simulation (**Figure 18** left) was successful while the second one (**Figure 18** right) was not, we aim to examine the distribution of all conformations in both trajectories and compare them. By comparing them, we observe noticeable differences between the two graphs. The majority of structures in both left and right graphs are found in the low values of the radius of gyration (Rg), indicating that in both graphs, most of the structures are compact, i.e., folded. However, unlike the left graph, which exhibits two distinct dark-red regions, the right graph shows only one region, located at low values of both Rg and q. As we explained in **Figure 17**, these two regions in the left graph, representing low values for both Rg and q and low Rg values with high q values, indicate the existence of transient structures and final conformations, respectively. On the other hand, the concentration of structures in the specific region of the right graph indicates the presence of only transient structures. Furthermore, comparing the intensity of the dark-red color, the number of transient structures identified in the right graph is much higher than those in the left graph.

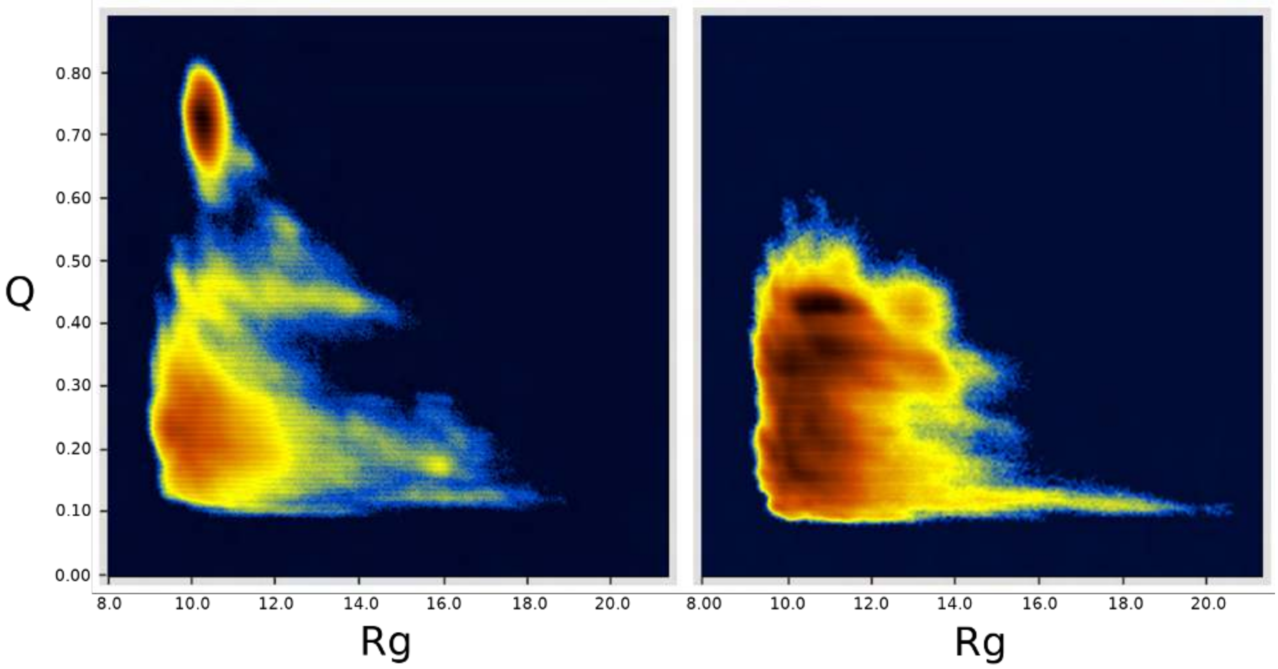


Figure 19 | *Q vs Rg graphs, left graph depicts the ff99SB*-ILDN trajectory using the entire sequence, right graph corresponds to the ff99SB-ILDN trajectory using the entire sequence*

Additionally, in an effort to gain further insights, we created two more graphs using the generated files Rgyration.dat and Qfraction.dat from **Figure 18**. **Figure 19** emerged from combining the radius of gyration (Rg) with the similarity value Q.

At first glance, we observe a similar behavior to that of **Figure 18**. However, upon closer examination, we notice some differences. Specifically, the dark-red region (representing low Rg values with high Q values) on the left side of **Figure 19** appears more concentrated, with Q values ranging from approximately 0.63 to 0.80, while the corresponding region on the left side of **Figure 18** appears more spread out, with q values ranging from approximately 0.45 to 0.67. Therefore, we understand that the high similarity percentage of bond distances (Q) relative to the reference structure does not perfectly align with the similarity percentage of the overall structure (q) compared to the reference.

The graph on the right image confirms the presence of numerous transient structures and indicates a lack of final conformations. However, compared to the corresponding graph in **Figure 18**, the one and only dark-red region in the graph of **Figure 19** is more spread out in space and exhibits a larger range of Q values.

2.6 Calculating histograms and RMS from specific frames, choosing the maximum q, Q and Qs values, gaining insights about RMSD values

From now on, we will focus solely on the ff99SB*-ILDN trajectory. For our next analysis, we want to calculate the RMSD, but this time we will do it based on a specific frame. We will perform this calculation seven times, using a different frame each time to extract as much information as possible. Initially, we need to carry out three calculations. First, we need to find the frame with the highest q value, which means identifying the frame that shows the highest similarity to the structure we used as a reference. Second, to identify the frame with the highest Q value, which represents the frame with the highest percentage of bonds distances that fall within the cutoff range of the corresponding bonds in the reference structure. Third, to find the frame with the highest Qs value, which corresponds to the frame where, overall, our bond distances show the least deviations compared to the bond distances in the reference structure.

We will perform the above process twice, once using the entire peptide chain and the second time choosing residues 4 to 31 of our chain. This specific segment of the chain was chosen based on the analysis of the rmsf diagram of the average protein structure (**Figure 12**). In this diagram, we were able to identify and remove the amino acids located at both ends of our chain that exhibit the highest mobility.

The procedure for extracting these specific results is as follows: Initially, if we haven't done it already, we need to execute the task “Fraction of native contacts” (once using the entire chain and the second time selecting residues 4 - 31). Next, we identify the three frames with the highest values of the three reaction coordinates (Q, Qs and q) from the trajectory of both the entire chain and the truncated chain. Then, we proceed with calculating the RMS from each frame by choosing the task fitting and selecting CA atoms and using the respective frame as a reference. Finally, we calculate the histograms for each frame.

Therefore, for each frame, we will generate two different plots: “RMS from frame” and a histogram. The “RMS from frame” plot is obtained by comparing each frame of the entire simulation with the specific frame we have chosen beforehand. The smaller the difference between the two frames, indicating a higher similarity, the smaller the vertical line appears in our plot. On the other hand, the histogram provides information about the distribution of RMSDs. The horizontal axis represents the RMSD values and is measured in Å, while the vertical axis represents the number of structures from our simulation that have a specific RMSD value. The smaller the RMSD value, the higher the similarity of each structure with the previously selected frame.

Frames selection:

Whole peptide chain:

- The maximum q value was found in frame 7,930,499 and it is 0.7148. In this frame, the Q value is 0.7271 and the Qs value is 0.6224.
- The maximum Q value was located in frame 14,405,337 and it is 0.8424. In this frame, the q value is 0.5787 and the Qs value is 0.6686.
- The maximum Qs value was identified in frame 14,250,453 and it is 0.6946. In this frame, the Q value is 0.8243 and the q value is 0.6386.

Peptide chain consisting of amino acids 4 - 31:

- The maximum q value was observed in frame 12,962,296 and it is 0.8246. In this frame, the Q value is 0.9245 and the Qs value is 0.8392.
- The maximum Q value was found in frame 14,365,495 and it is 1.0000. In this frame, the q value is 0.6690 and the Qs value is 0.8731.
- The maximum Qs value was identified in frame 7,875,920 and it is 0.8939. In this frame, the Q value is 0.9811 and the q value is 0.6095.
- The highest sum of Q, Qs, and q value was obtained in frame 13,523,301 and it is 2.6447. Specifically, the Q value is 0.9811, the Qs value is 0.8744 and the q value is 0.7892.

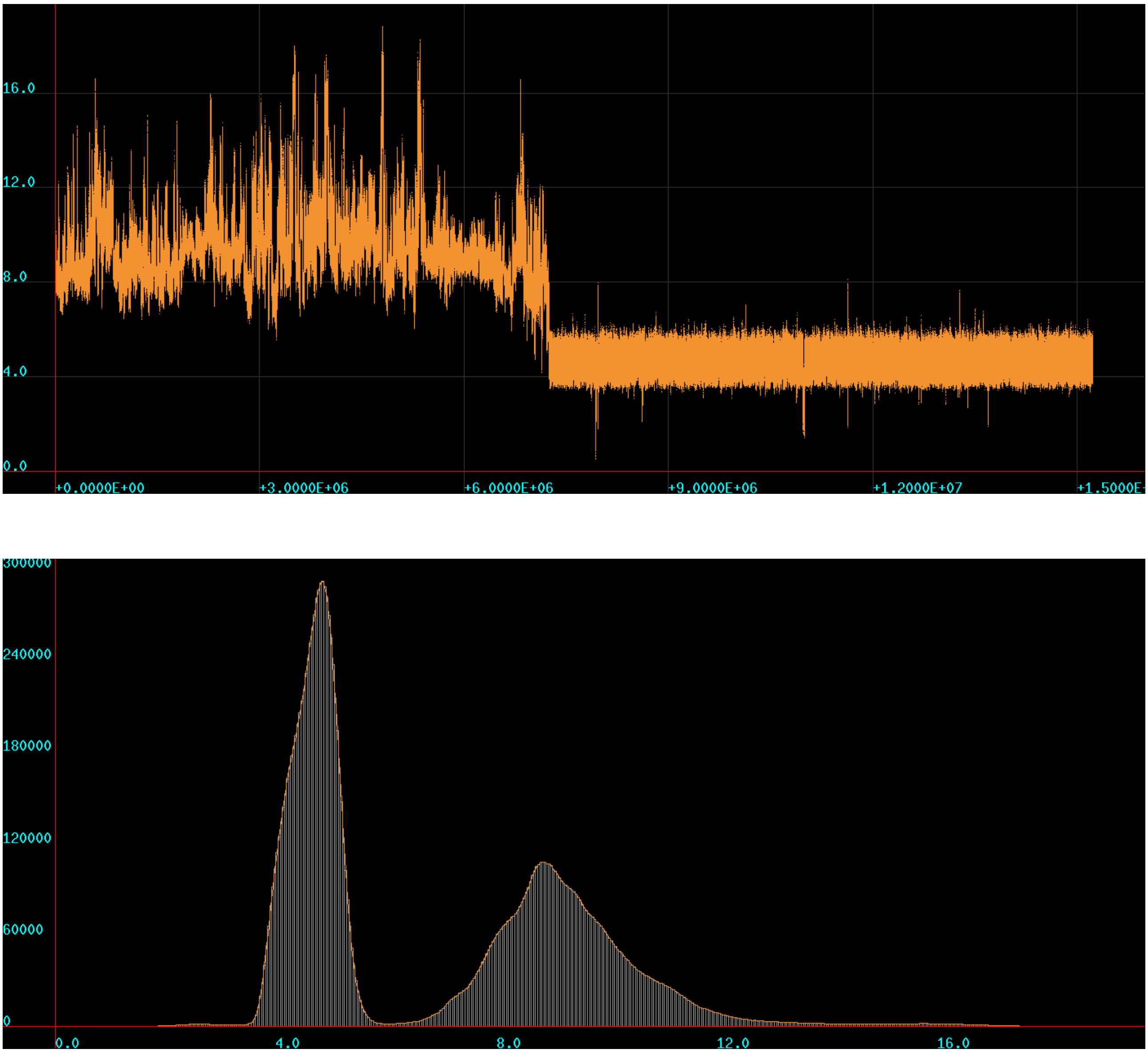


Figure 20 | *Whole peptide chain*: *(top graph) RMS from frame 7930499 (q), (bottom graph) histogram based on frame 7930499 (q)*

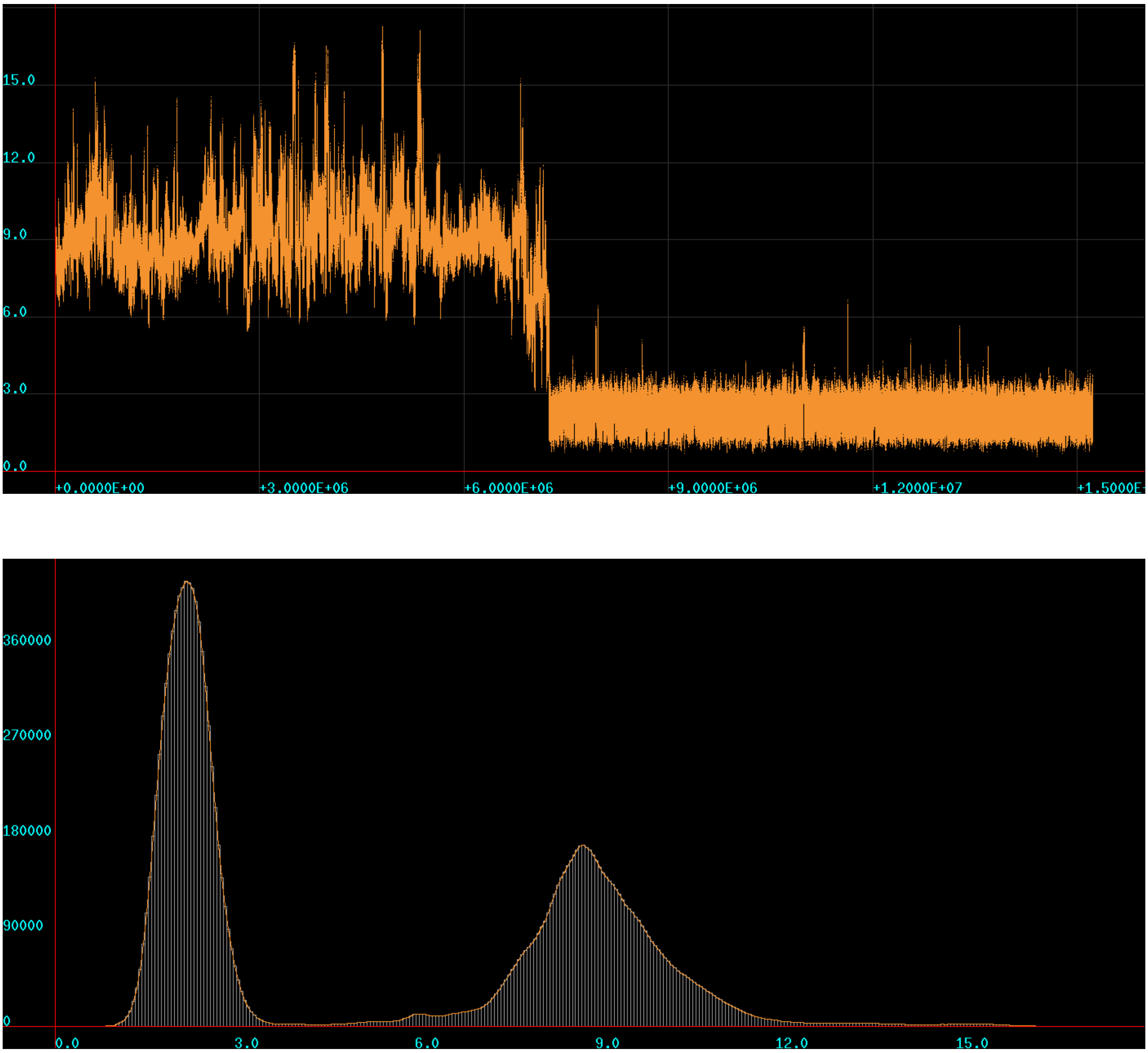


Figure 21 | *Whole peptide chain*: *(top graph) RMS from frame 14405337 (Q), (bottom graph) histogram based on frame 14405337 (Q)*

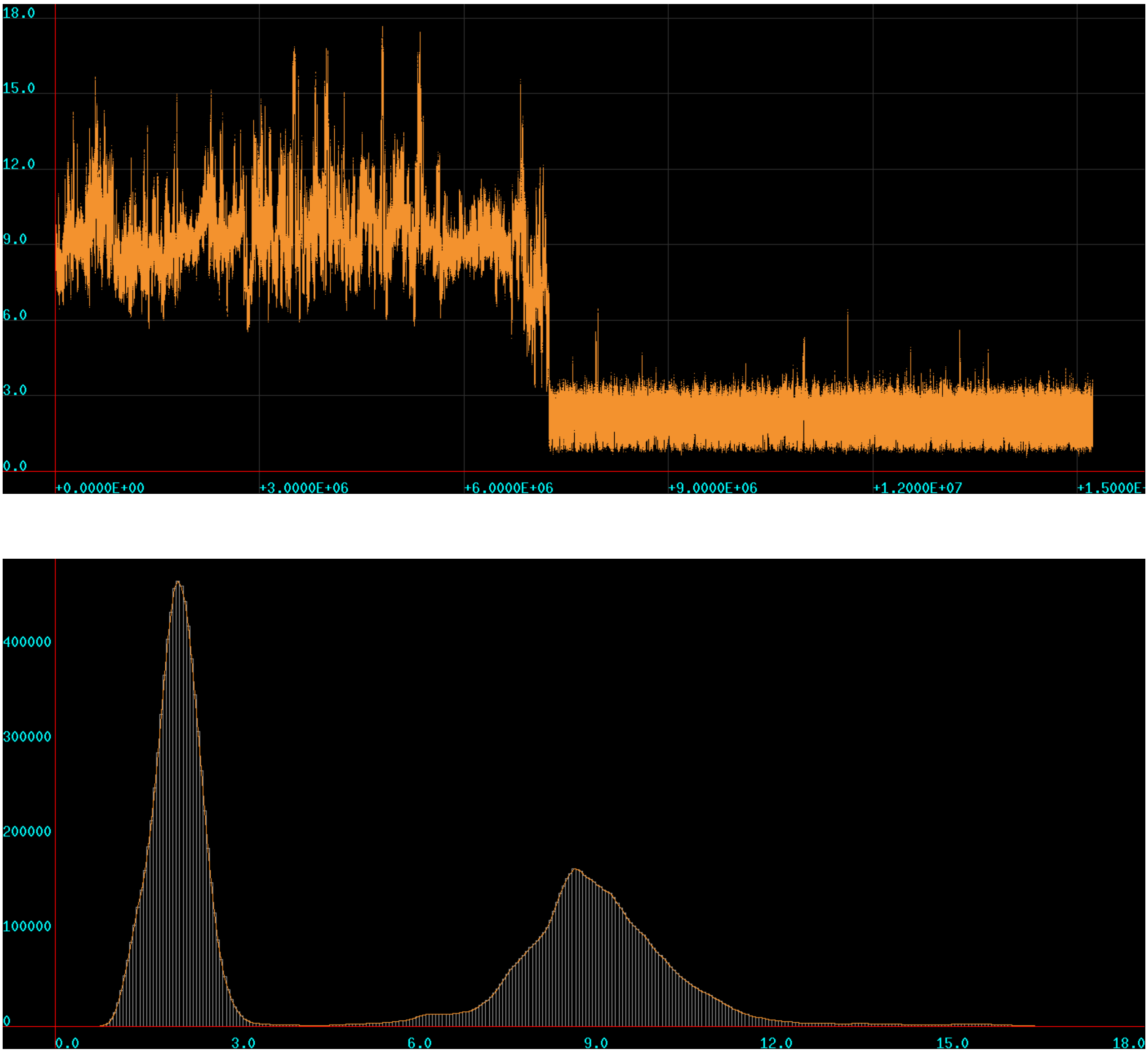


Figure 22 | *Whole peptide chain*: *(top graph) RMS from frame 14250453 (Qs), (bottom graph) histogram based on frame 14250453 (Qs)*

Upon analyzing the diagrams in **Figures 20**, **21** and **22** which include both the RMS from frame plots and the histograms, we observe similarities among all three frames used. Specifically, in the RMS from frame diagrams, we distinguish two main phases. The first phase begins from the beginning of the simulation until just before the final conformation of our protein, approximately up to the 7,200,000th frame. The second phase starts where the first phase ends and coincides with the end of the simulation. The protein's conformation in all three selected frames is very close to the conformation of the reference structure. Therefore, it is logical to observe high RMSD values in the first phase of the diagram and low values in the second phase, during which the chain has reached its final configuration.

In the histograms we created for all three cases, we observe two main peaks. The peak on the right corresponds to the accumulation of transient structures with high RMSD values, i.e., structures that have very low similarity to the frame we used. On the other hand, the peak on the left corresponds to the accumulation of folded structures with low RMSD values, i.e., structures that have acquired the final conformation. Analyzing the left peak of the three histograms, an important insight it offers is its elevated height. The height is noticeably greater compared to the right peak, indicating that in our simulation, there are many more structures with final conformation than transient conformations. The frames where we had the maximum Q and Qs values have values below 3 Å, while the q value has values above 4 Å.

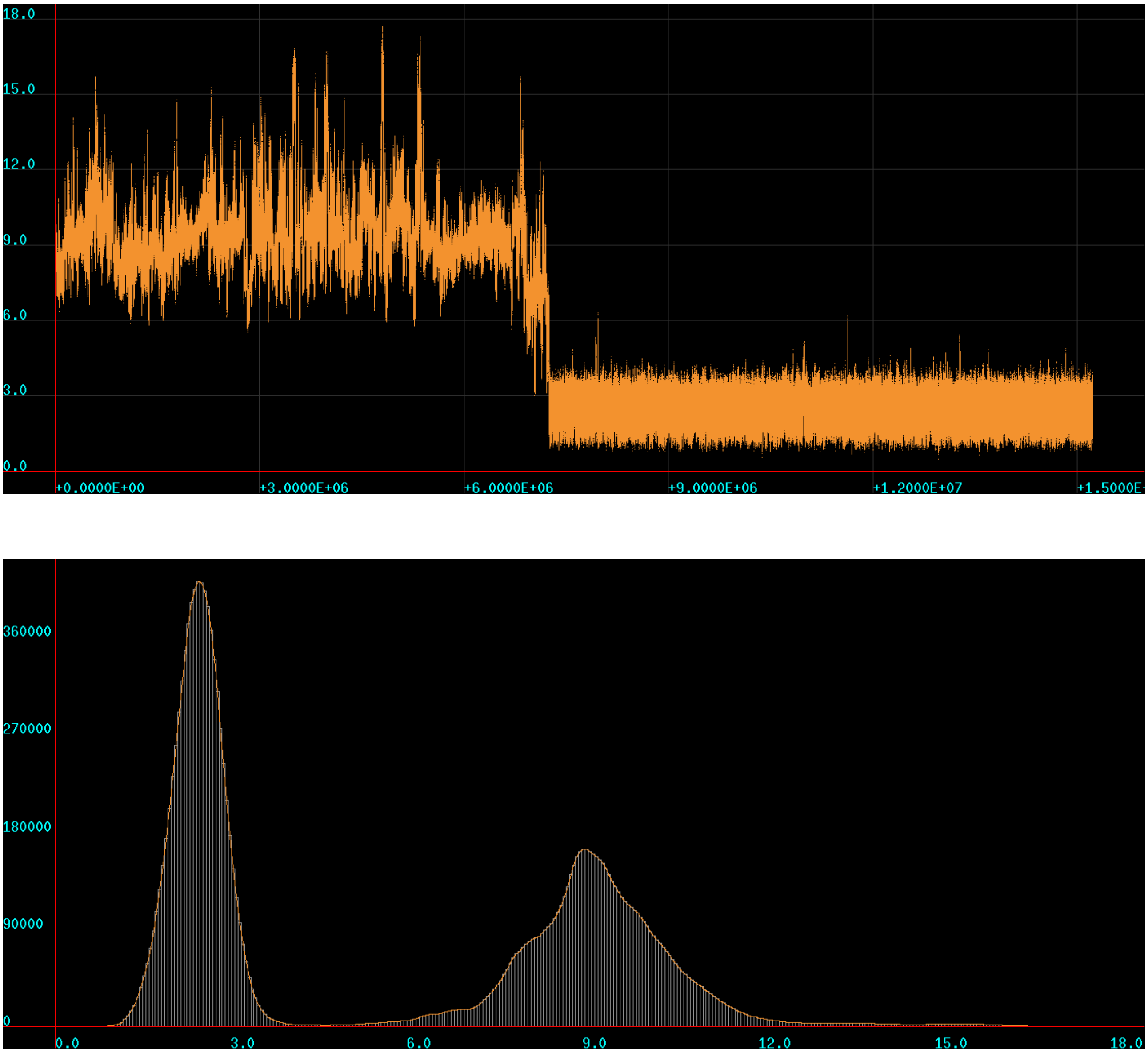


Figure 23 | *Truncated peptide chain using amino acids 4-31: (top graph) RMS from frame 12962296 (q), (bottom graph) histogram based on frame 12962296 (q)*

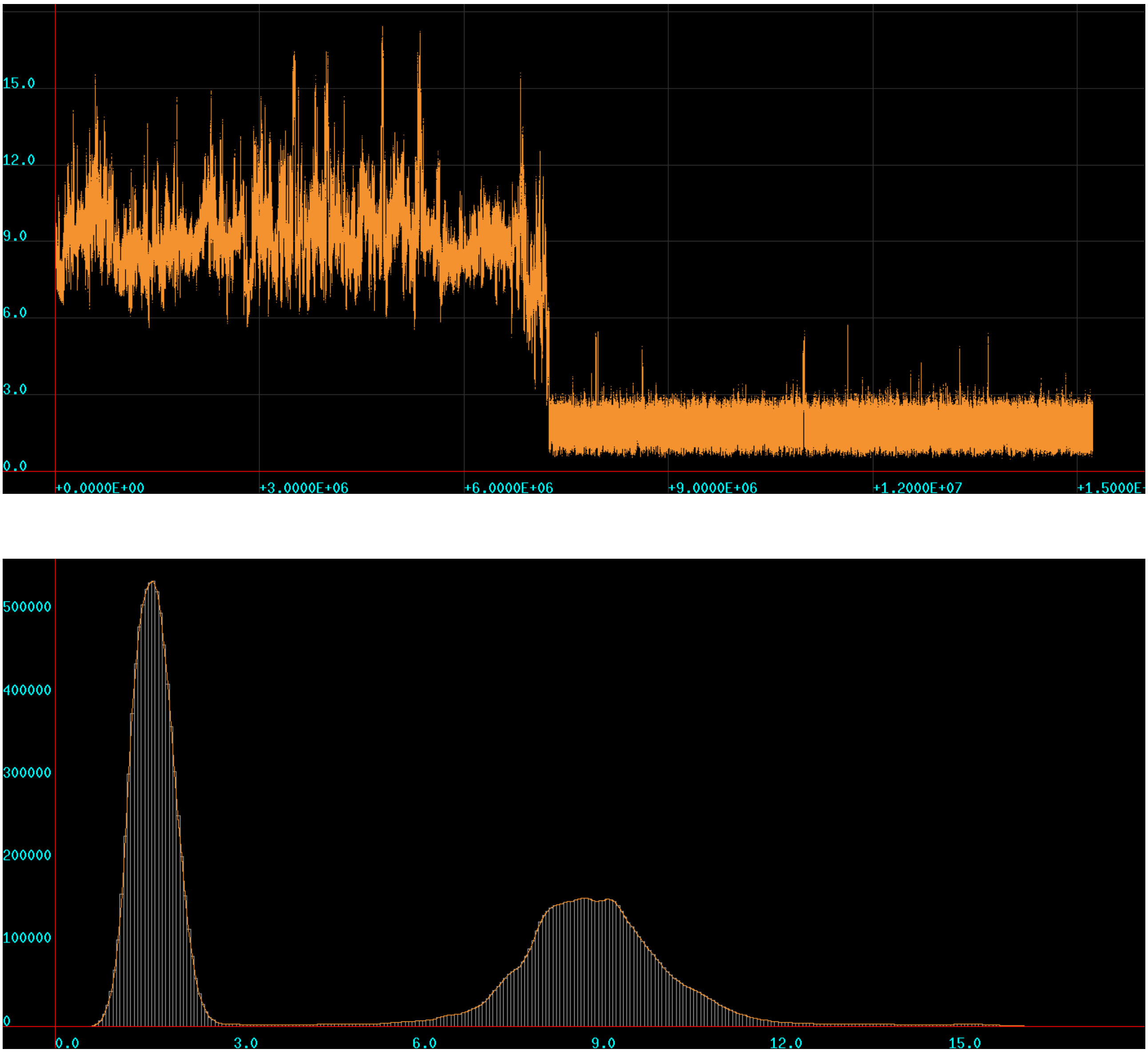
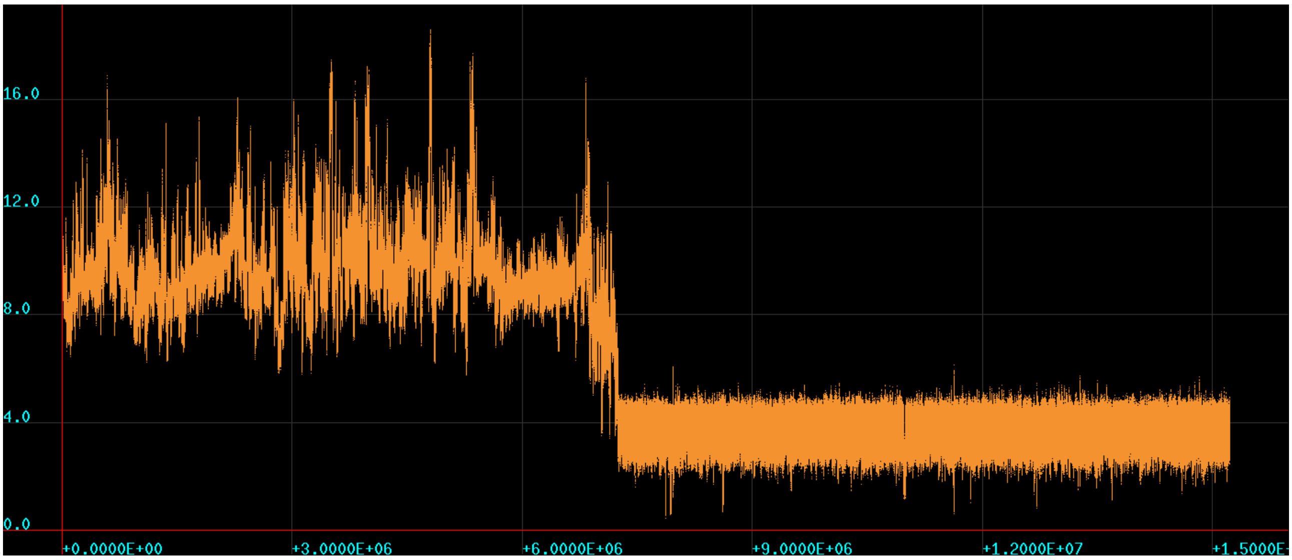


Figure 24 | *Truncated peptide chain using amino acids 4-31: (top graph) RMS from frame 14365495 (Q), (bottom graph) histogram based on frame 14365495 (Q)*



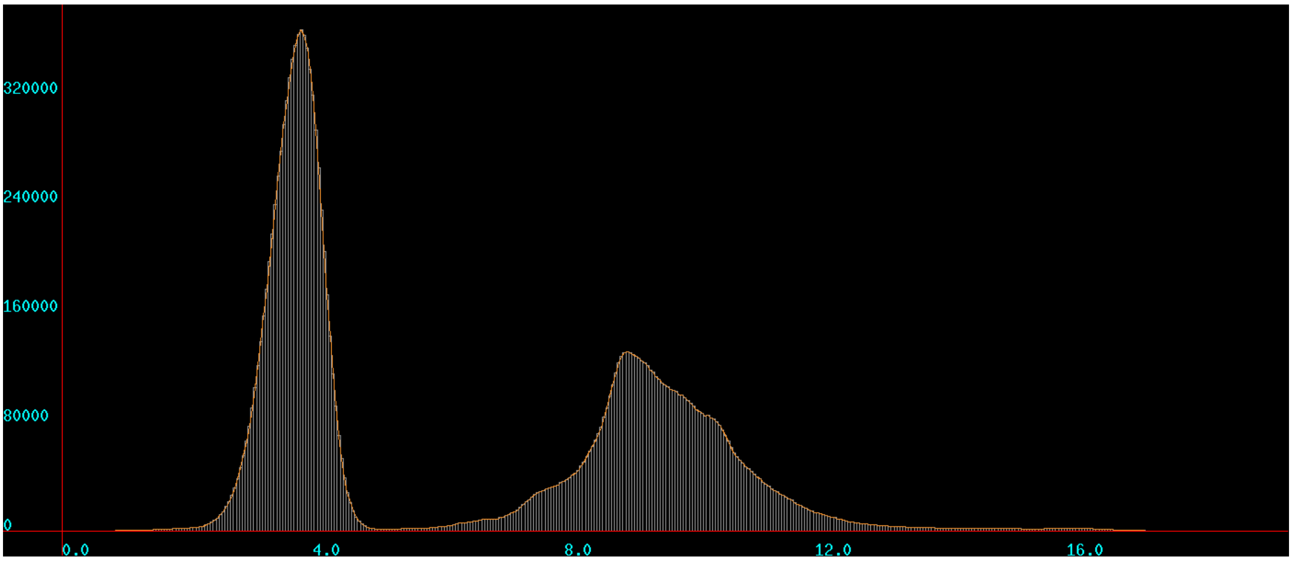


Figure 25 | *Truncated peptide chain using amino acids 4-31: (top graph) RMS from frame 7875920 (Qs), (bottom graph) histogram based on frame 7875920 (Qs)*

The diagrams in **Figures 23, 24** and **25** all exhibit a same pattern. The RMS from frame diagrams consist of two main phases, which are based on the same logic and the histograms also consist of two distinct peaks following the same logic. However, the question we need to answer at this stage is the following: Did we manage, after removing the terminal amino acids, to find a frame that is even closer to the native structure? Upon re-examining the left peaks of the three new histograms, we observe that in all three frames, the accumulation of structures occurs at even lower RMSD values compared to the three previous histograms. Therefore, we are closer to finding the frame we are looking for.

During the search for frames with the maximum values of q, Q and Qs, we recorded the values of all three measures for each frame. Upon reviewing these values, we observe that the maximum value of one measure is not necessarily accompanied by high values of the other two measures. For example, in the case of the peptide chain consisting of amino acids 4-31, in frame 14,365,495, the measure Q has the highest possible value of 1.0000, but the measure q is only 0.6690. Therefore, the next step we can take is to calculate the sum of the values of all three measures for each frame and then identify the frame that exhibits the highest sum.

In **Figure 26**, we generated two supplementary diagrams using the frame with the maximum sum. By comparing the left peak of this histogram with the corresponding left peaks from the previous diagrams, we observe that the RMSD values, where the structures gather, remain consistently low, below 3 Å. However, it's worth noting that these values are not lower than the ones we previously observed in **Figure 24**.

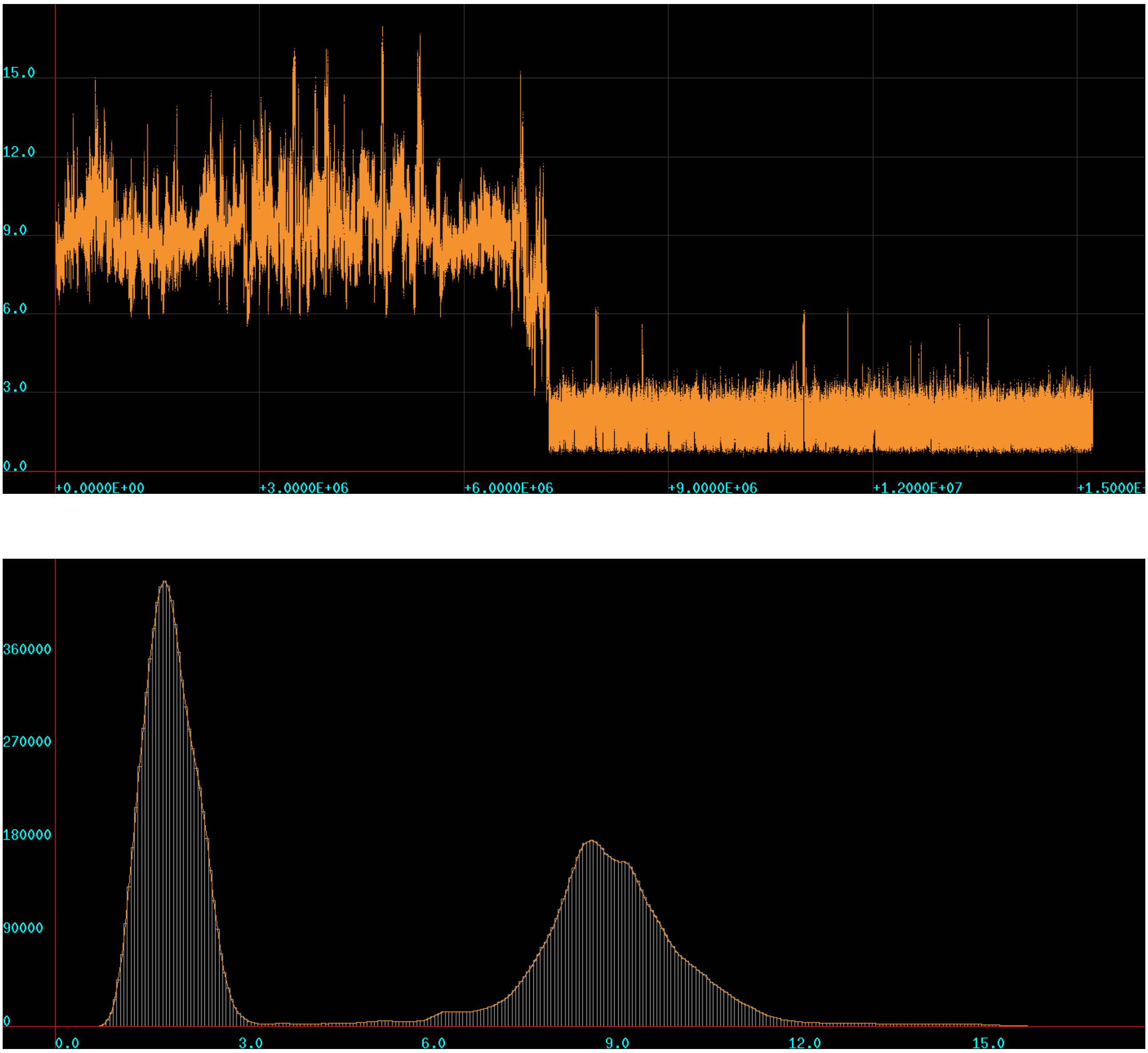


Figure 26 | *Truncated peptide chain using amino acids 4-31: (top graph) RMS from frame 13523301 (maximum sum of q, Q and Qs), (bottom graph) histogram based on frame 13523301 (maximum sum of q, Q and Qs)*

2.7 The significance of dPCA and CPCA for a deeper understanding of internal and overall motions in our system

Understanding the atomic motions/fluctuations in proteins and their correlations with each other is of great importance for gaining insights into protein dynamics. Atoms that exhibit correlated motions and move together in a coordinated manner, can be put to the same collective group. When the fluctuations of an atom, belonging in a specific collective group, are significant, the collective motions become dominant in influencing the behavior of the system. It has been found that the structure of the molecule directly influences the formation of collective groups. Secondary structural components tend to move as collective groups. [47]

Through the statistical tool PCA (Principal Components Analysis), we gain the ability to process large, high-dimensional datasets, by emphasizing their similarities and differences. During data processing using PCA, we encounter two essential terms: "eigenvector" and "eigenvalue", which always appear together as a pair. The eigenvector plays a crucial role in expressing the data in a specific manner and provides information about the direction, slope and magnitude of fluctuation. The eigenvalue associated with a specific eigenvector indicates its significance in interpreting the results and determines its rank. [48], [49]

The eigenvector with the highest eigenvalue is referred to as the principal component 1 (PC1), the eigenvector with the second-highest eigenvalue is called principal component 2 (PC2) and so on. What sets PCA apart and makes it particularly popular in data analysis is its ability to simplify high-dimensional data by ignoring principal components with very low eigenvalues, i.e., low significance, without necessarily distorting the results. Depending on the complexity of the system, a corresponding number of principal components is generated. In the study of molecular dynamics simulations, tens or even hundreds of principal components are generated. By ranking all the principal components from the highest eigenvalue to the lowest and analyzing the impact of each principal component on the overall motion of the system, it has been found that a small subset of principal components is sufficient to describe the motion of the entire system. [48]–[50]

For the analysis of MD simulations, the use of both Dihedral PCA and Cartesian PCA methods is essential. Dihedral PCA, focusing on the dihedral angles of a biomolecule, provides information about local/internal conformational changes, while Cartesian PCA deals with the cartesian coordinates of the atoms and informs us about the collective motions of the entire molecular system. [51], [52]

2.7.1 Dihedral PCA, density distribution of fluctuations using the ff99SB*-ILDN trajectory with step 1

Continuing the analysis of the ff99SB*-ILDN trajectory, we performed the task Dihedral PCA and selected the following parameters: Residue selection from 4 to 31, an upper limit of 150 clusters, a step of 1 and we included all frames from the simulation. During the residue selection, we excluded these 6 amino acids for the same reason as in section 2.6, (to avoid unnecessary noise they might cause). However, in the pdb files generated during this analysis, we instructed the program to include the terminal amino acids to provide more comprehensive representations. In the 3D analysis of dPCA, 4 clusters emerged, while the 5D analysis revealed 147 clusters.

In **Figures 27**, **29** and **31** (on the left), we have the 2D representations of density distribution of fluctuations based on two specific principal components and we compare them with the scatter plot, obtained from the 5D analysis, showing how the two corresponding principal components affect the distribution of the 147 clusters (on the right). These 147 clusters are distinguished by different colors. In **Figure 27**, the plots were generated using the first two principal components, PC1 and PC2. Similarly, in **Figure 29**, the plots were based on PC1 and PC3, while **Figure 31** was created using PC2 and PC3.

Focusing on the left images, what we observe are many small dots, where each dot corresponds to a specific frame. The color representation here informs us about the distribution of fluctuations. Dark blue color indicates a high concentration of specific fluctuations, representing frames with specific folding patterns. Red color denotes very low concentration, while yellow color represents an intermediate state.

In **Figures 28**, **30** and **32**, we revisit the three distributions shown in the left plots in **Figures 27**, **29** and **31**, respectively. These are then compared with three additional scatter plots (on the right) showing the density distribution of the 147 clusters based on their corresponding principal components.

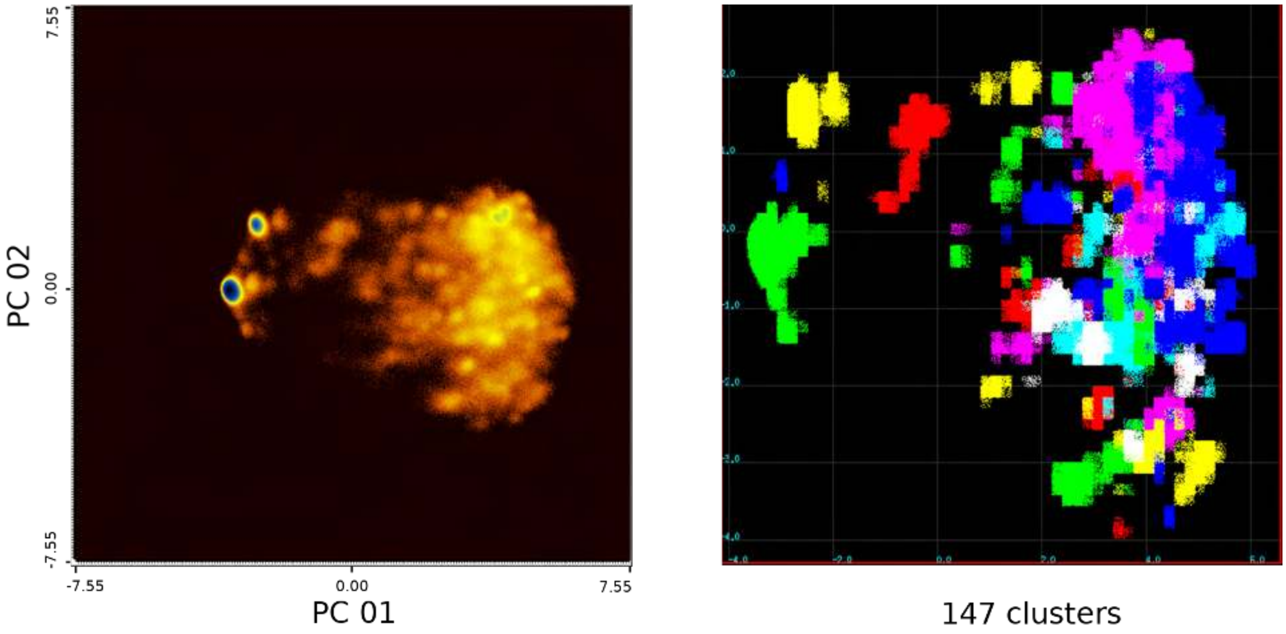


Figure 27 | *2D representations of: density distribution of fluctuations based on PC1 and PC2 (left), distribution of the 147 clusters from the 5D analysis based on PC1 and PC2 (right)*

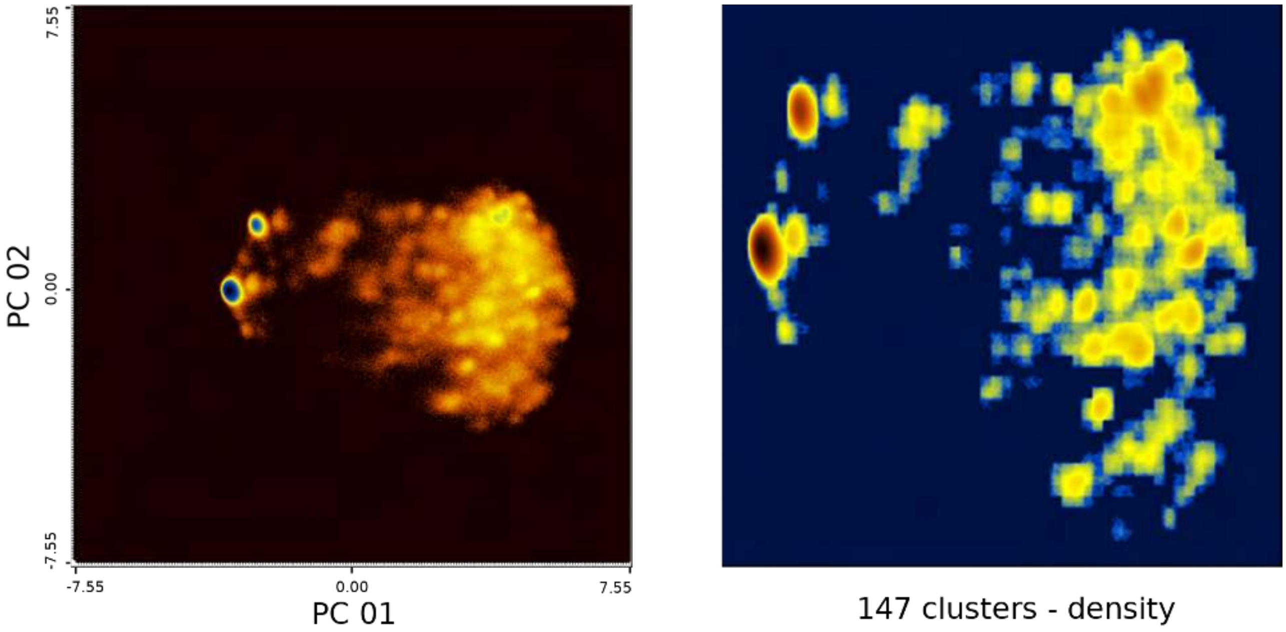


Figure 28 | *2D representations of: density distribution of fluctuations based on PC1 and PC2 (left), density distribution of the 147 clusters from the 5D analysis based on PC1 and PC2 (right)*

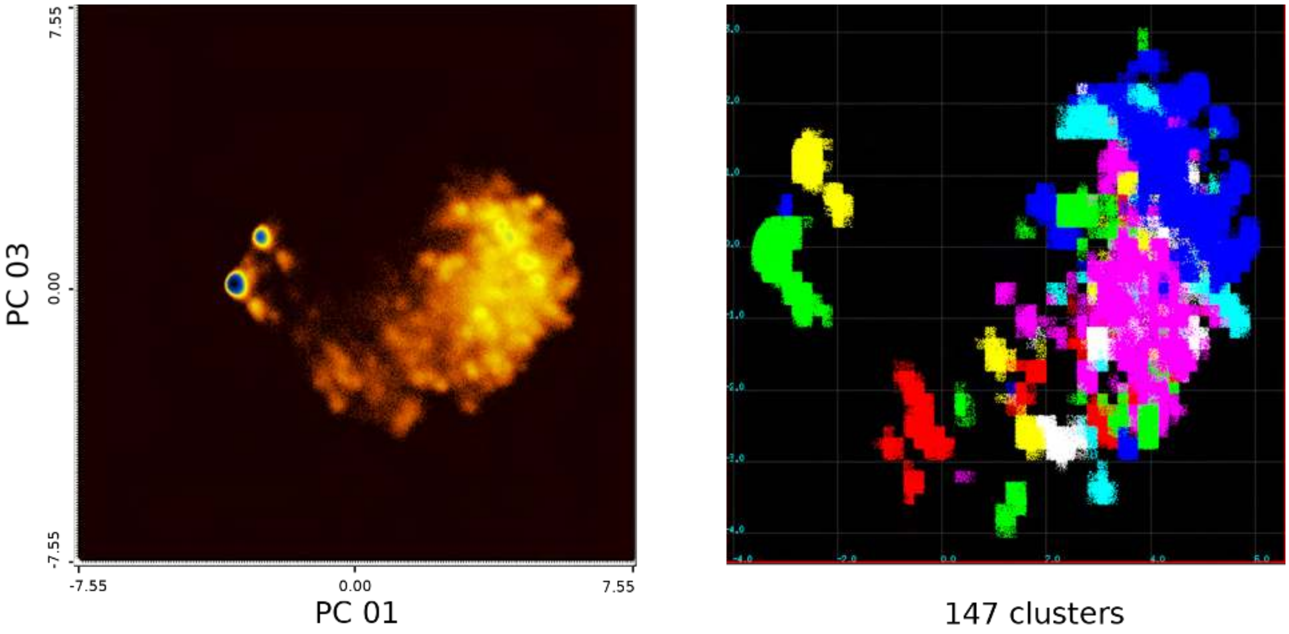


Figure 29 | *2D representations of: density distribution of fluctuations based on PC1 and PC3 (left), distribution of the 147 clusters from the 5D analysis based on PC1 and PC3 (right)*

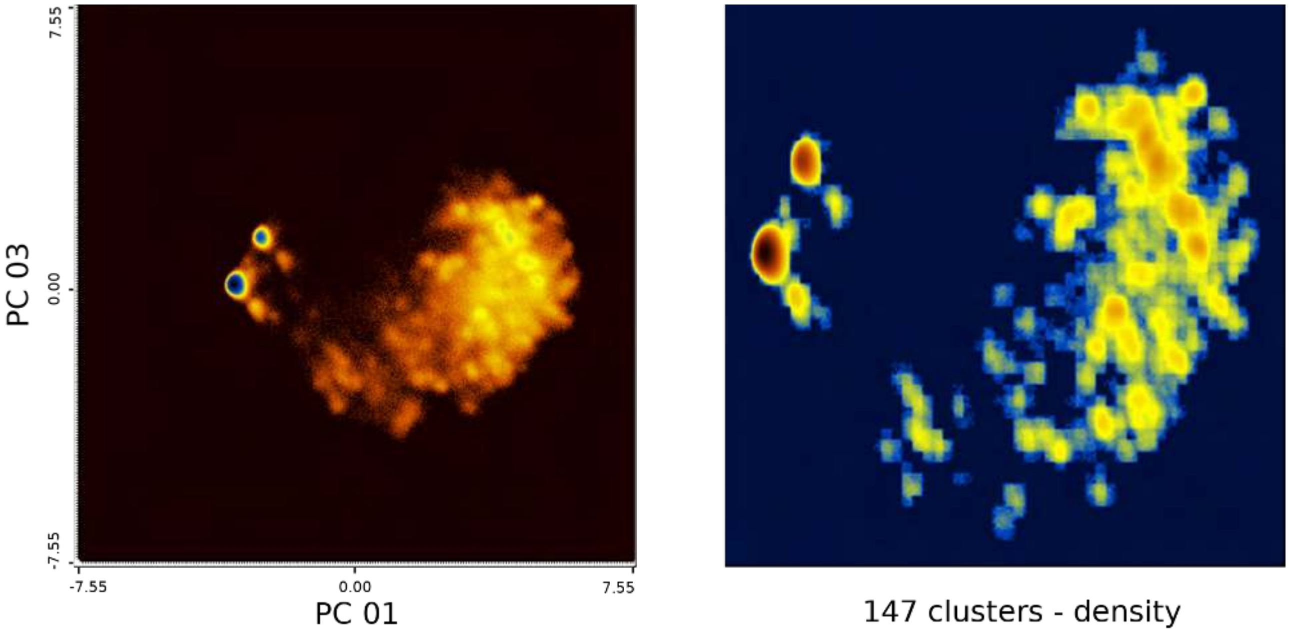


Figure 30 | *2D representations of: density distribution of fluctuations based on PC1 and PC3 (left), density distribution of the 147 clusters from the 5D analysis based on PC1 and PC3 (right)*

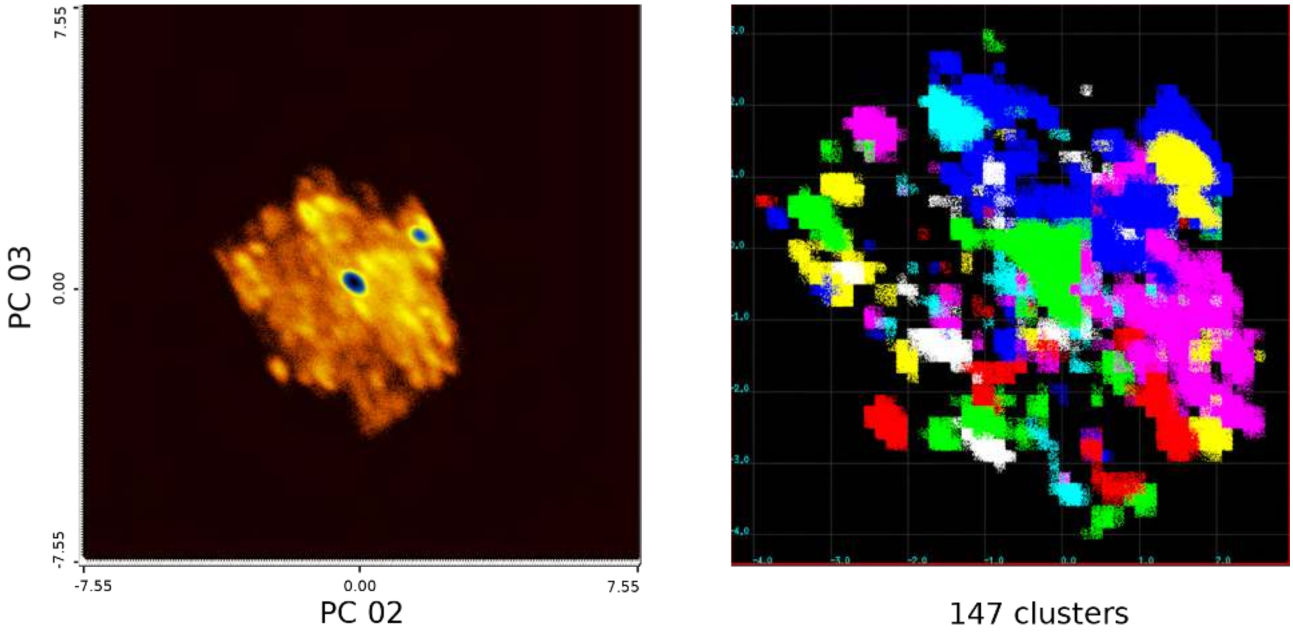


Figure 31 | *2D representations of: density distribution of fluctuations based on PC2 and PC3 (left), distribution of the 147 clusters from the 5D analysis based on PC2 and PC3 (right)*

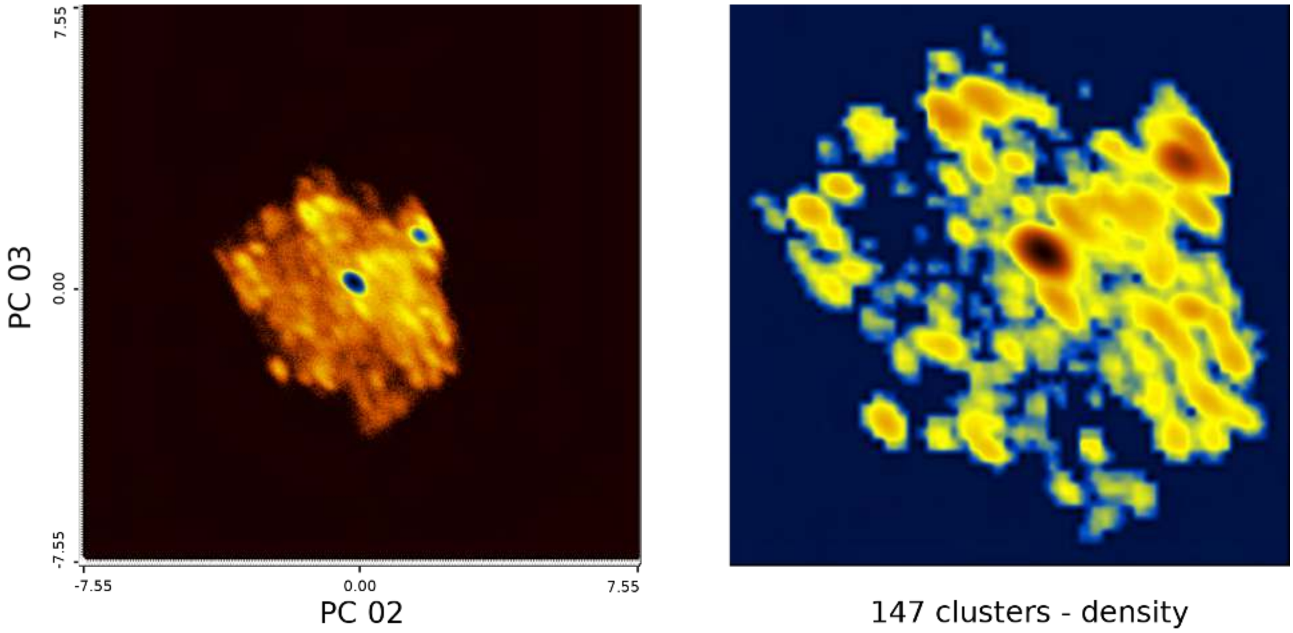


Figure 32 | *2D representations of: density distribution of fluctuations based on PC2 and PC3 (left), density distribution of the 147 clusters from the 5D analysis based on PC2 and PC3 (right)*

The information provided by the three density distributions of fluctuations (PC1 vs PC2, PC1 vs PC3, and PC2 vs PC3) (on the left) indicates the presence of two dark-blue regions with a high accumulation of structures of specific fluctuations. Focusing on **Figure 27**, these two dark-blue regions in the left image correspond to the green and yellow colors in the right plot. According to the information derived from "Scatter plots of categorical data", where in our case the categorical data represent the clusters, a specific set of seven colors is used, following this order: "1" → Green, "2" → Yellow, "3" → Blue, "4" → Magenta, "5" → Cyan, "6" → Red and "7" → White. When there are more than eight categories (clusters), the plot cycles through the same set of colors.

Therefore, by observing both images in **Figure 27**, we can understand that the cluster with the green color can be labeled as cluster 1 and corresponds to the first dark-blue region with the highest accumulation of structures of the same fluctuation. Additionally, the cluster with the yellow color can be labeled as cluster 2 and corresponds to the second dark-blue region.

2.7.2 Dihedral PCA, density distribution of fluctuations using the ff99SB*-ILDN trajectory with step 2

We conducted the dPCA analysis twice, using different step sizes: one with a step size of 1 and the other with a step size of 2. In the previous task, we explored the 2D representations of density distribution of fluctuations in relation to specific pairs of principal components, namely PC1 vs PC2, PC1 vs PC3 and PC2 vs PC3. An important discovery arising from this analysis is the existence of two distinct peaks, representing two conformations with the same fluctuations, that demonstrated extended stability during the simulation. Upon comparing these three distributions (PC1 vs PC2, PC1 vs PC3, and PC2 vs PC3) with their corresponding (PC1' vs PC2', PC1' vs PC3', and PC2' vs PC3') obtained with a step size of 2. Once again, we observed the appearance of two peaks. **Figure 32** illustrates the three distributions derived with a step size of 2.

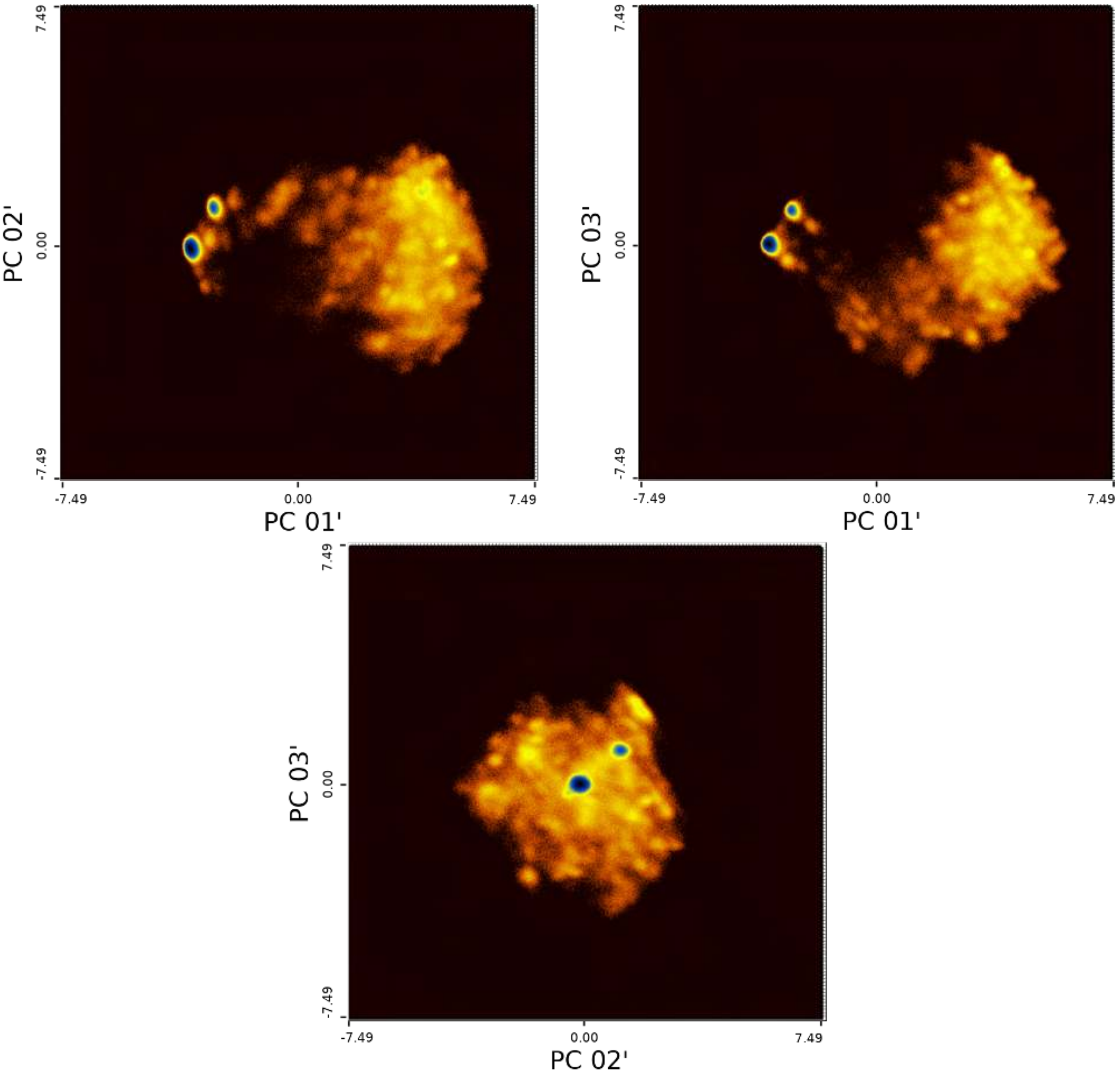


Figure 33 | *2D representation of density distribution of fluctuations PC1' vs PC2', PC1' vs PC3' and PC2' vs PC3'*

Continuing the analysis with a step of 2, we aim to identify the secondary structures associated with these two peaks. To explore this, we examined the dPCA diagrams with a step of 1, which revealed that each peak corresponds to a specific cluster. This observation was further confirmed by the dPCA analysis with a step of 2, where the most prominent peak aligned with cluster 1 and the second peak aligned with cluster 2. These two clusters were found to exhibit a parallel *β*-sheet conformation involving three clones.

The crucial question remains: Do both clusters indeed represent the final conformation adopted by our peptide chain? To address this, we conducted a comparative analysis. We aligned the representative pdb files of clusters 1 and 2 obtained from the 3D analysis using the MM-align protein structural alignment program [53]. Further refinement of the resulting pdb file was performed with PyMOL. The comparison presented in **Figure 34** highlights minor structural differences, which account for the variations observed in their corresponding principal components. Despite these differences, the alignment between the two structures affirms that both clusters represent the chain's final conformation.

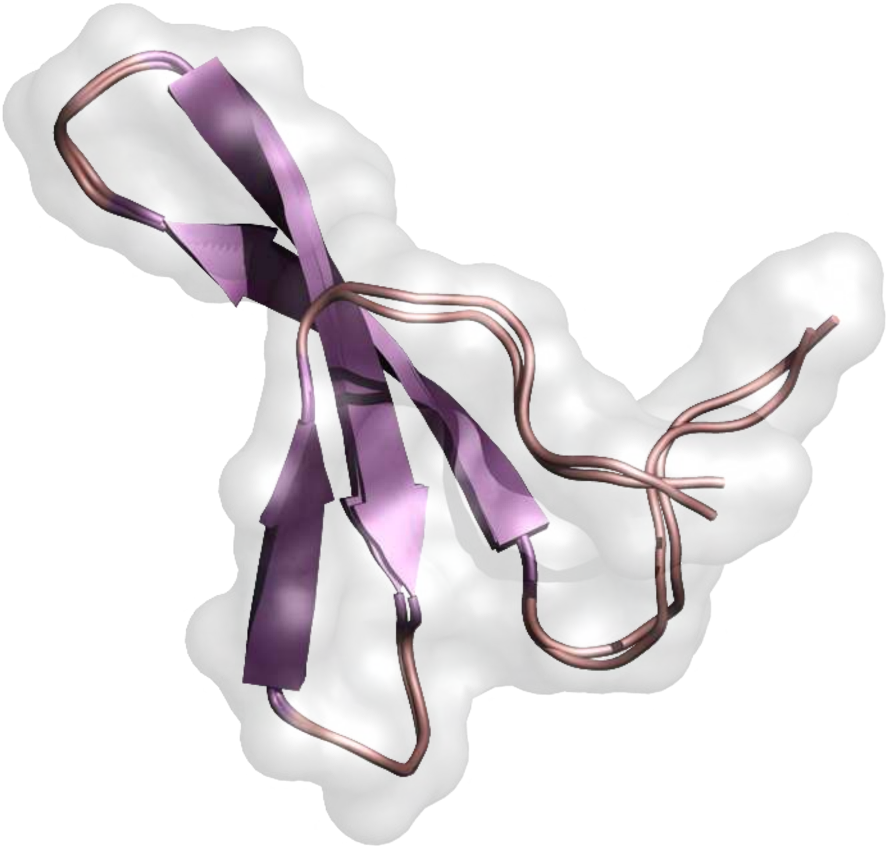
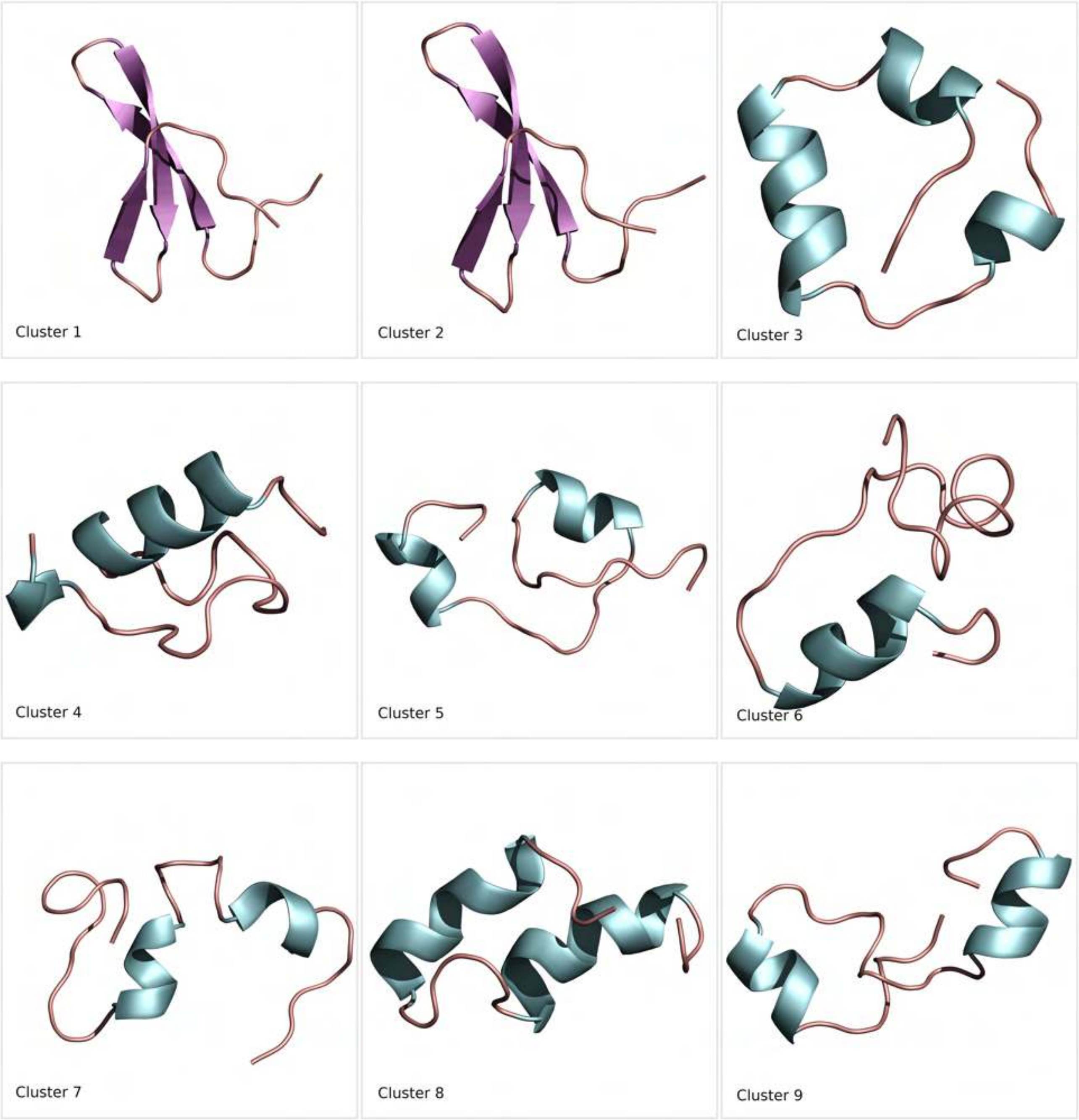
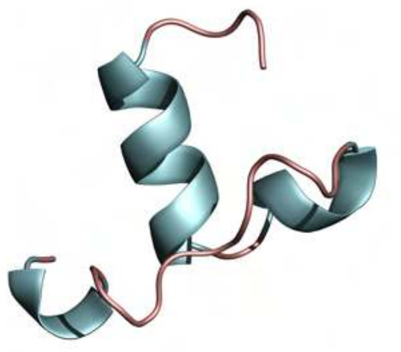


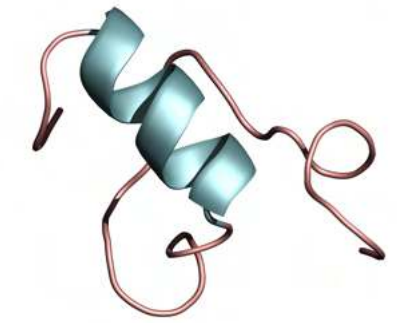
Figure 34 | *Superimposition of the representative structures of clusters 1 and 2 from the 3D analysis. Using pymol, structure is colored by secondary structure: sheet (magenta), helix (cyan) and loop (salmon)*

During the 5D dPCA analysis, we obtained 4 representative structure pdb files corresponding to the 4 clusters generated by the 3D analysis, as well as 98 representative structure pdb files corresponding to the 98 clusters generated by the 5D analysis. The previous examination in **Figure 34** covered the first two clusters from the 3D analysis. Now, in **Figure 35**, we will showcase a selection of 39 representations from the 98 clusters of the 5D analysis.

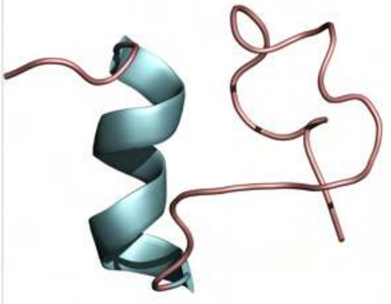




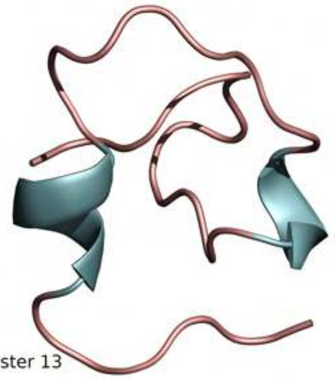
Cluster 10



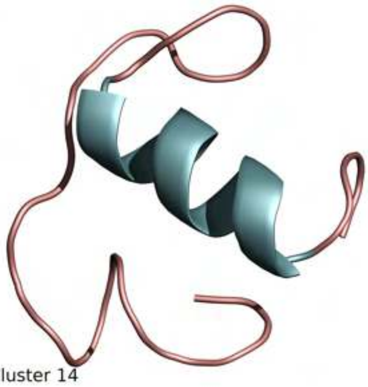
Cluster 11



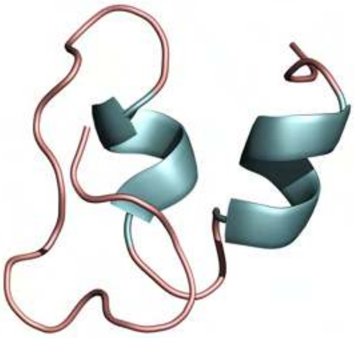
Cluster 12



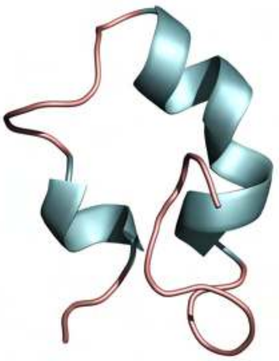
Cluster 13



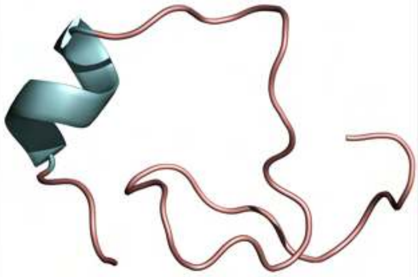
Cluster 14



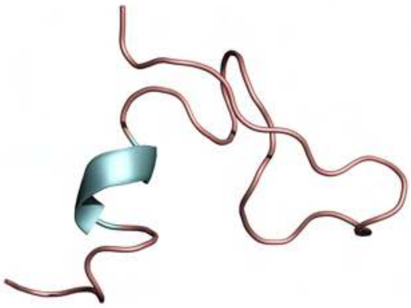
Cluster 15



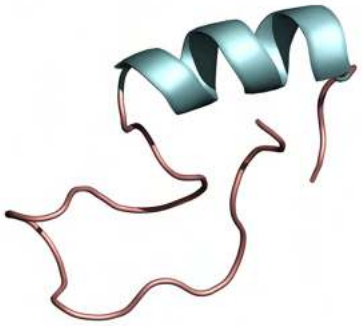
Cluster 16



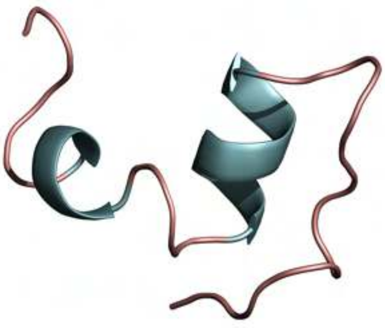
Cluster 17



Cluster 18

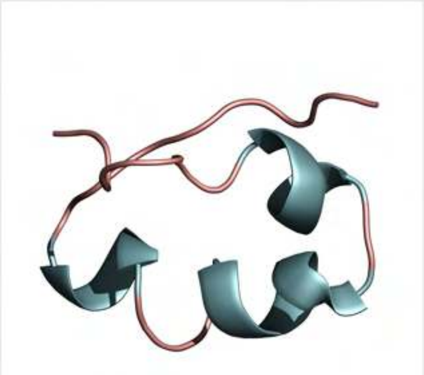


Cluster 19

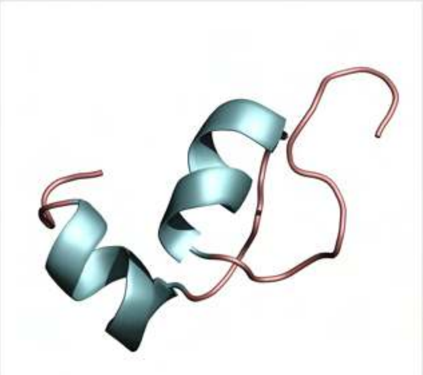


Cluster 20

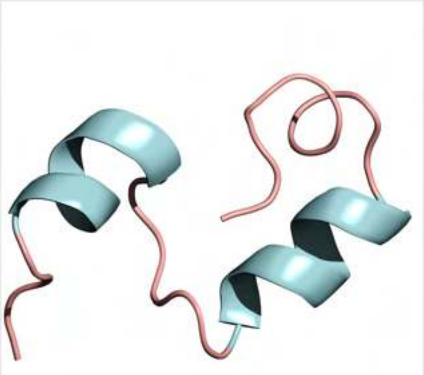
Cluster 24



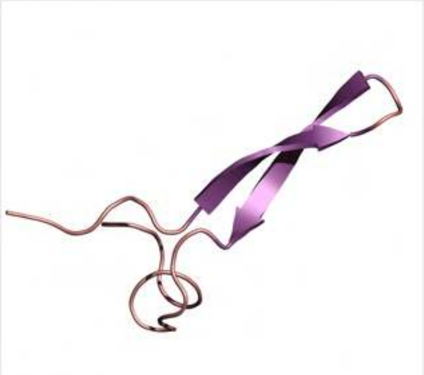
Cluster 26



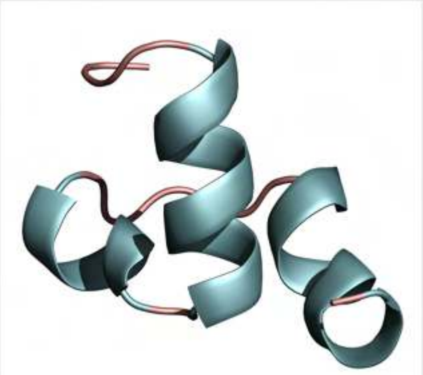
Cluster 27



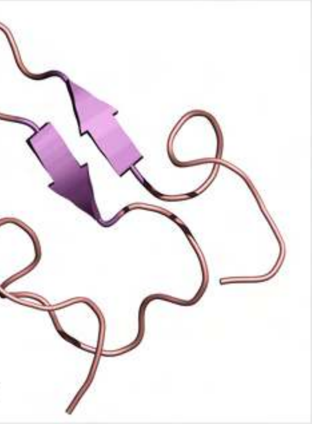
Cluster 28



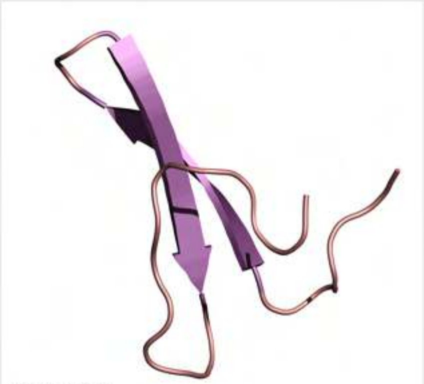
Cluster 29



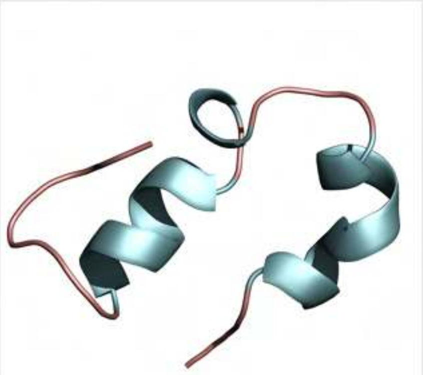
Cluster 31



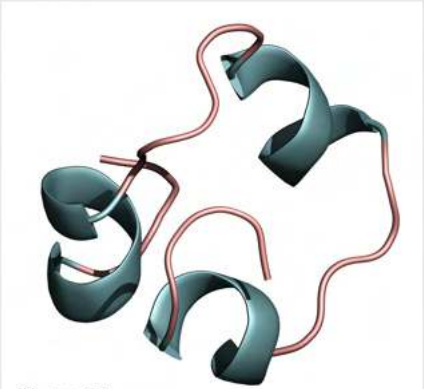
Cluster 35



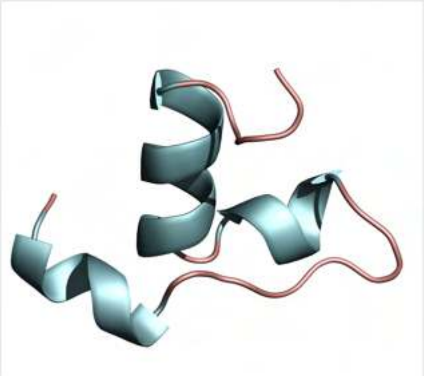
Cluster 36



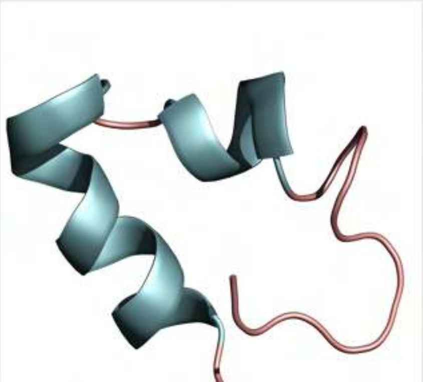
Cluster 44



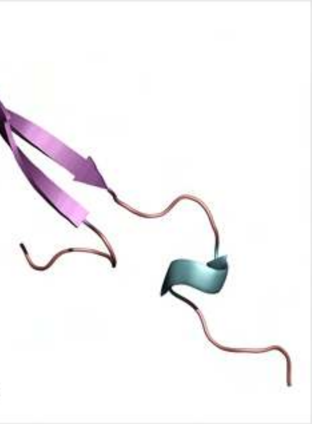
Cluster 51



Cluster 58



Cluster 60



Cluster 70

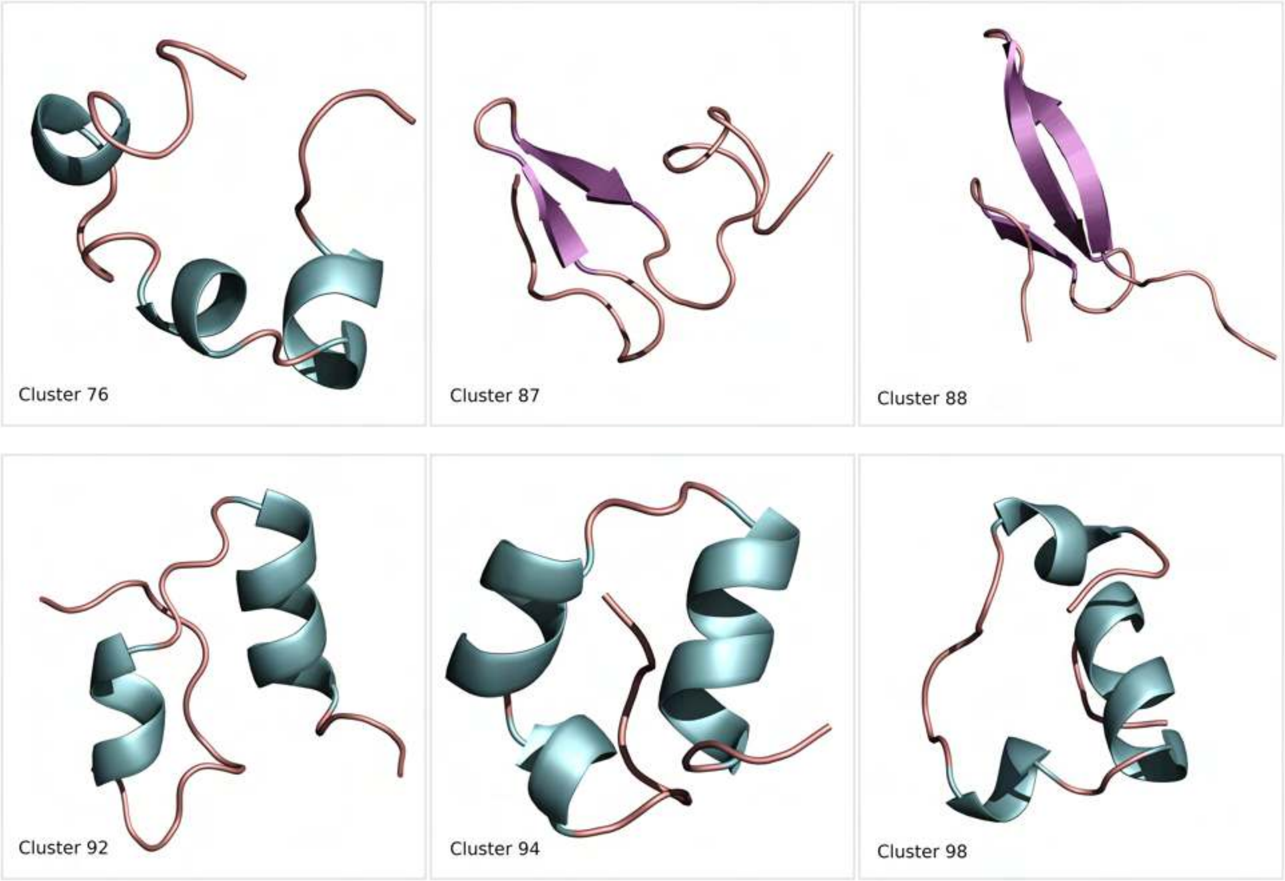


Figure 35 | *39 out of the 98 clusters of the representative structure in the 5D analysis. Using pymol, structure is colored by secondary structure: sheet (magenta), helix (cyan) and loop (salmon)*

In **Figure 35**, the first two clusters, as in the 3D analysis, correspond to the β-sheet conformation, which is the final conformation of our chain. Beyond that, we observe that most clusters involve helical conformations of various lengths, while a few clusters (excluding the first two) specifically, 7 of them, correspond to the β-sheet structures.

2.7.3 Dihedral PCA, density distribution of fluctuations using the ff99SB-ILDN trajectory with step 1

Concluding the dPCA analyses, in order to gain a more comprehensive understanding, we extended the dPCA to the trajectory generated by the ff99SB-ILDN force field. Our parameter selection included: (residue selection: 4 - 31, clusters: 150, first frame: 1, last frame: 10021600 and step: 1). As a result, we obtained 44 clusters from the 3D analysis and 123 clusters from the 5D analysis.

Subsequently, we constructed **Figure 36**, presenting a 2D representation of the density distribution of fluctuations based on the three most significant principal components (PC1 vs PC2, PC1 vs PC3, and PC2 vs PC3). As anticipated, the resulting images confirmed the fact that this specific force field fails to effectively fold our protein. All three representations display numerous peaks, indicating that throughout the simulation, the protein underwent multiple transient conformations, that were maintained for a short period of time.

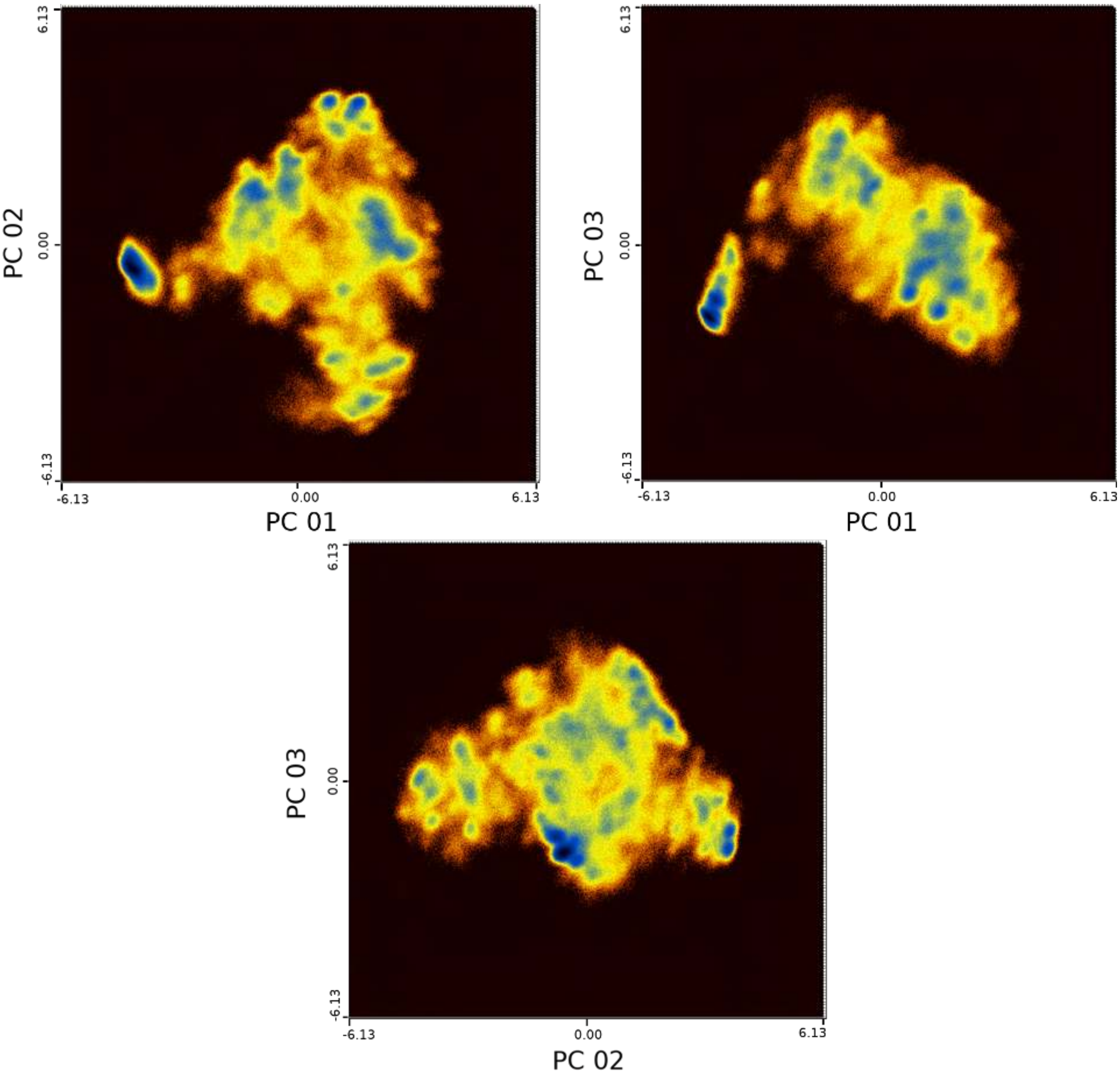


Figure 36 | *2D representation of density distribution of fluctuations: PC1 vs PC2, PC1 vs PC3 and PC2 vs PC3*

2.8 Cartesian PCA, isolating the first principal component and visualizing *β*-sheet fluctuations

Using CPCA, our goal is to be able to observe the fluctuations/motions, known as conformational dynamics, performed by our molecular system as a whole. The entire process is based on the first eigenvector, which has the highest eigenvalue. PC1, with the largest eigenvalue, is the one that distinguishes the folded from the unfolded structures. Therefore, after performing CPCA, we aim to isolate and analyze the first principal component (PC1).

To generate the final image illustrating the overall motion of our system, we employed a combination of the graphical user interface Grcarma and executed some additional commands through the terminal. The step-by-step process is outlined below:

Using grcarma we performed fitting and Cartesian PCA

fitting: (CA atoms, all residues, use frame as reference: 7300000, first frame: 7300000, last frame: 15229200, step 1, use subset of the residues for the fitting: CA, change residue selection: 5 – 30)

If we carry out the fitting without excluding the motion of terminal amino acids (change residue selection: 5 - 30), the resulting final image is difficult to evaluate due to the presence of "noise". Additionally, we have chose to isolate these particular frames that represent only the phase of the simulation where the final conformation is attained. Consequently, the fitting procedure is applied to the "blue box", which we have examined in **Figure 7** (top) and **Figure 9** (top). After the fitting, we obtained a total of 7,929,201 frames.

CPCA: (CA atoms, all residues, 150 clusters, perform five dimensional clustering, 150 clusters, first frame: 1, last frame: 7929201, step: 1)

After completing the fitting and CPCA, we proceeded with the following steps:

Due to the fact that during the previous procedure we selected CA atoms for convenience, we make the following renaming:

- mv carma.selected_atoms.psf CAs.psf

The file carma.fitted_0.dcd was produced using CPCA and includes only the CA atoms of the first cluster. For convenience, we proceed with the following renaming:

```
mv carma.fitted_0.dcd cluster_01.dcd
```

Determine the minimum and maximum observed values of the first principal component:

```
awk '{print $2}' carma.PCA.fluctuations.dat | sort -n | tail -1
```

The maximum value is: 26.0772724

```
awk '{print $2}' carma.PCA.fluctuations.dat | sort -n | head -1
```

The minimum value is: -13.2602043

By considering only the CA atoms and the minimum and maximum values of PC1, we will generate a new DCD file named carma.play.dcd, which will contain solely the PC1 information:

```
carma -verb -write -col -cov -eigen -play 1 -13.2602043 26.0772724 cluster_01.dcd CAs.psf
```

We will generate a PDB file to visualize the motion along the first eigenvector. For the -last option, the total frames in cluster 1 are 2,687,183, so we will use half of this set, which is 1,343,592 frames. Regarding the -step option, we will choose a value that produces around 300 PDB files at the end. Using a step value of 4,570 will result in 294 PDB files:

```
carma -v -w -pdb -last 1343592 -step 4570 carma.play.dcd CAs.psf
cat carma.play.dcd.*.pdb > eigen_01.pdb
rm carma.play.dcd.*.pdb
```

Within the eigen_01.pdb file, there are 294 pdb files, all of which have been superimposed. This superimposition of numerous pdb files, representing frames where the final conformation is reached, allows us to visualize the collective motion exhibited by the molecular structure over time. We proceed with the analysis of the eigen_01.pdb file using PyMOL.

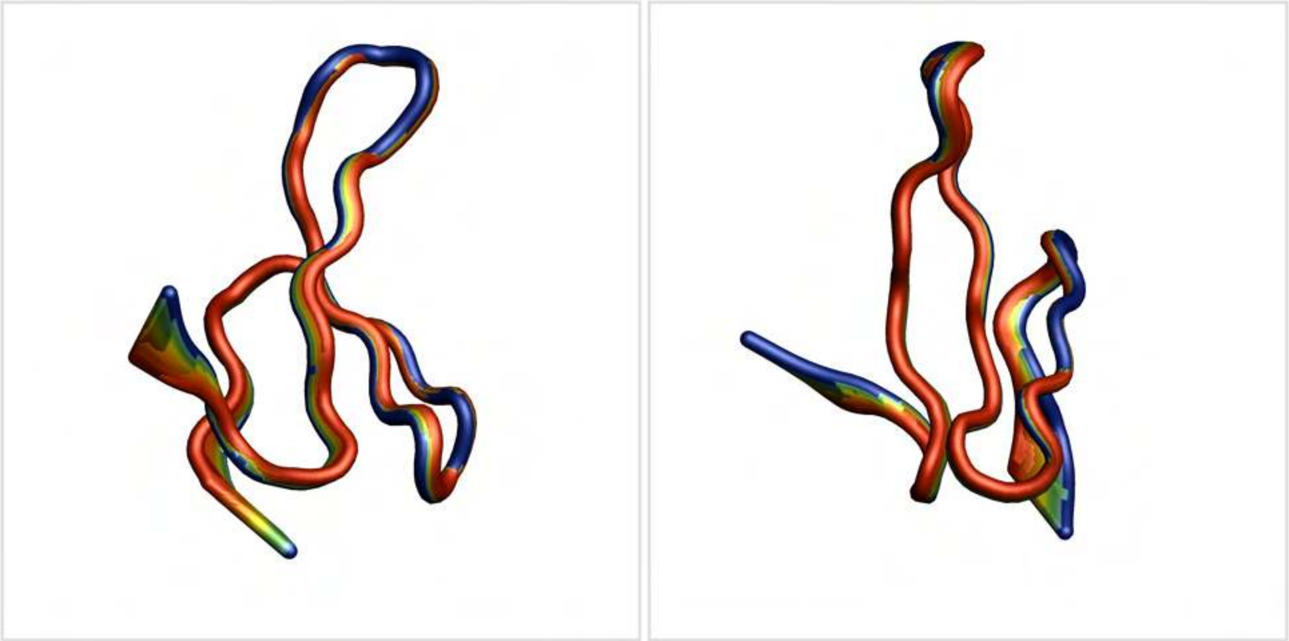


Figure 37 | *Superimposition structures of PC1's 294 pdb files using the entire peptide chain. Using pymol, structure is colored by spectrum: rainbow*

Upon observing the final image in **Figure 37**, it becomes evident that the entire molecular structure, including both flexible and rigid segments, undergoes significant motion during the simulation. The flexible regions, namely the termini, exhibit significant oscillation, indicated by the thickness of the structure. On the other hand, the stable segments, like the *β*-sheet, experience relatively smaller oscillations but are not completely motionless. A closer examination of the three *β*-strands reveals an intriguing behavior, they oscillate in such a way that the distances between them and consequently the hydrogen bonds, remain constant.

We repeated the entire above process one more time, but this time without including the C- and N-terminal residues. As a result, **Figure 38** was generated, in which it is now very clear that the oscillation of the terminal regions is significantly reduced. Regarding the *β*-sheet, the distances between the strands remain constant during its oscillation.

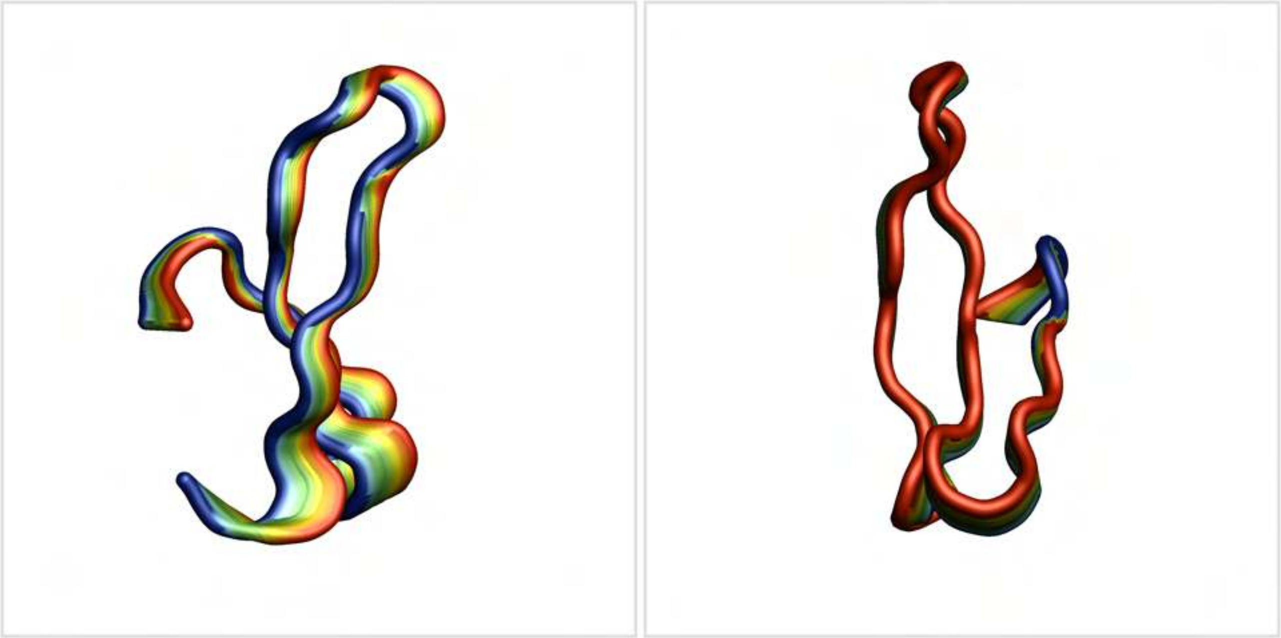
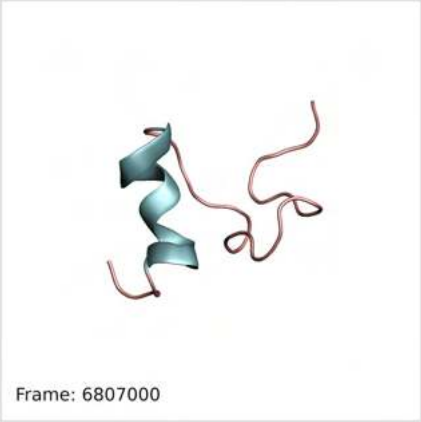


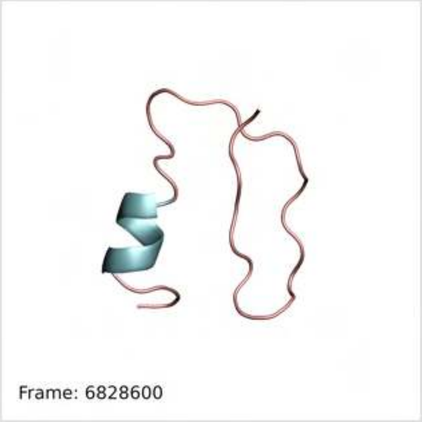
Figure 38 | *Superimposition structures of PC1's 294 pdb files excluding C- and N- terminal residues. Using pymol, structure is colored by spectrum: rainbow*

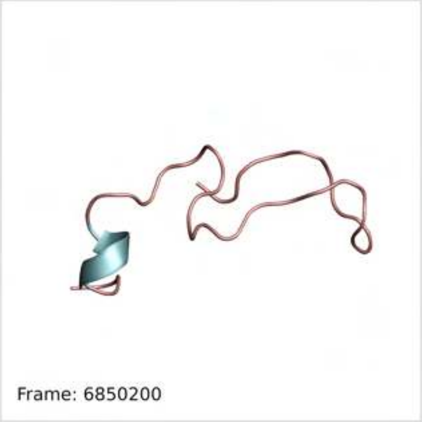
2.9 Visualizing specific folding stages of our protein’s simulation

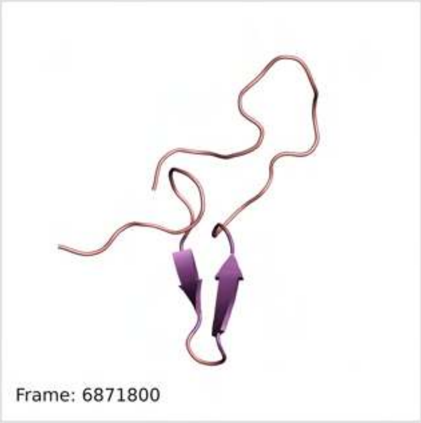
For the last part of our analysis, we chose a visual representation, focusing on specific segments of the simulation. The aim was to observe the different stages of our peptide chain with higher detail, leading to its final conformation.

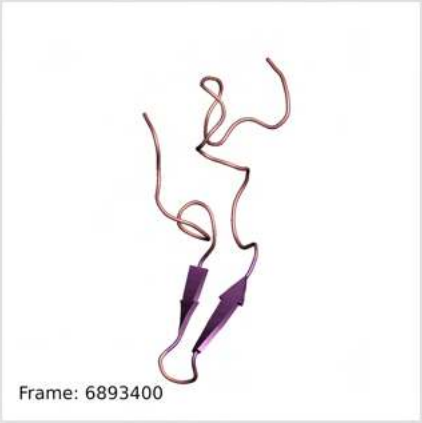
2.9.1 Frames: 6,807,000 – 7,239,000, visualizing the entire process, starting from the formation of the initial *β*-strands, up to the native structure

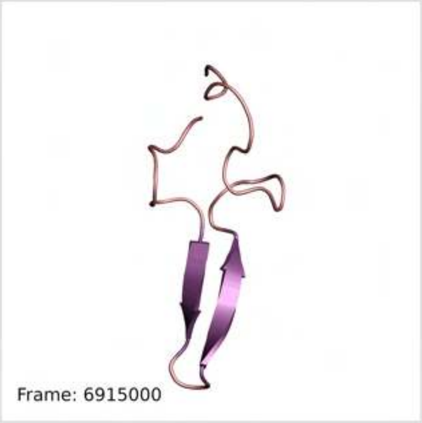


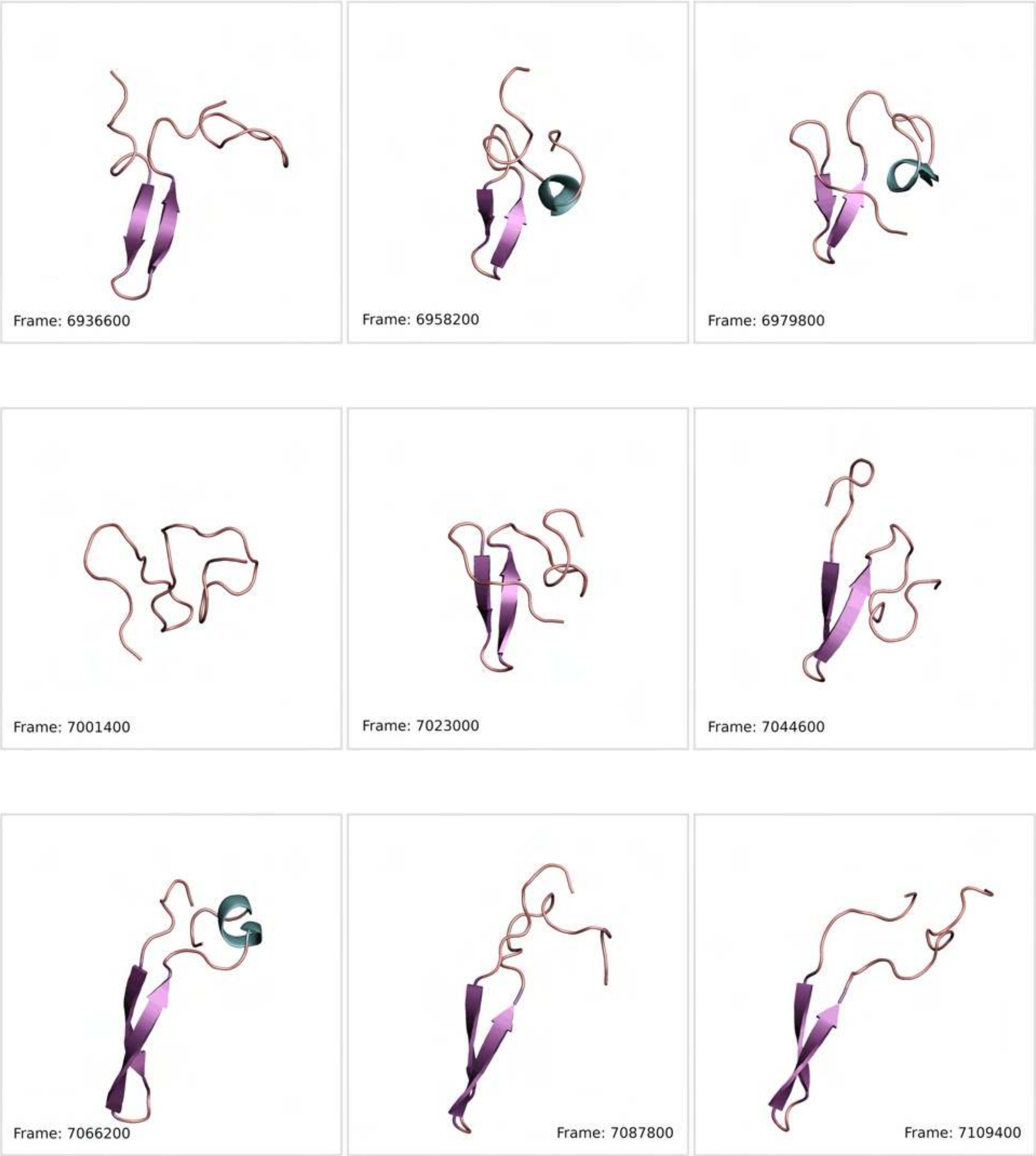











Frame: 6936600
Frame: 6958200
Frame: 6979800
Frame: 7001400
Frame: 7023000
Frame: 7044600
Frame: 7066200
Frame: 7087800
Frame: 7109400

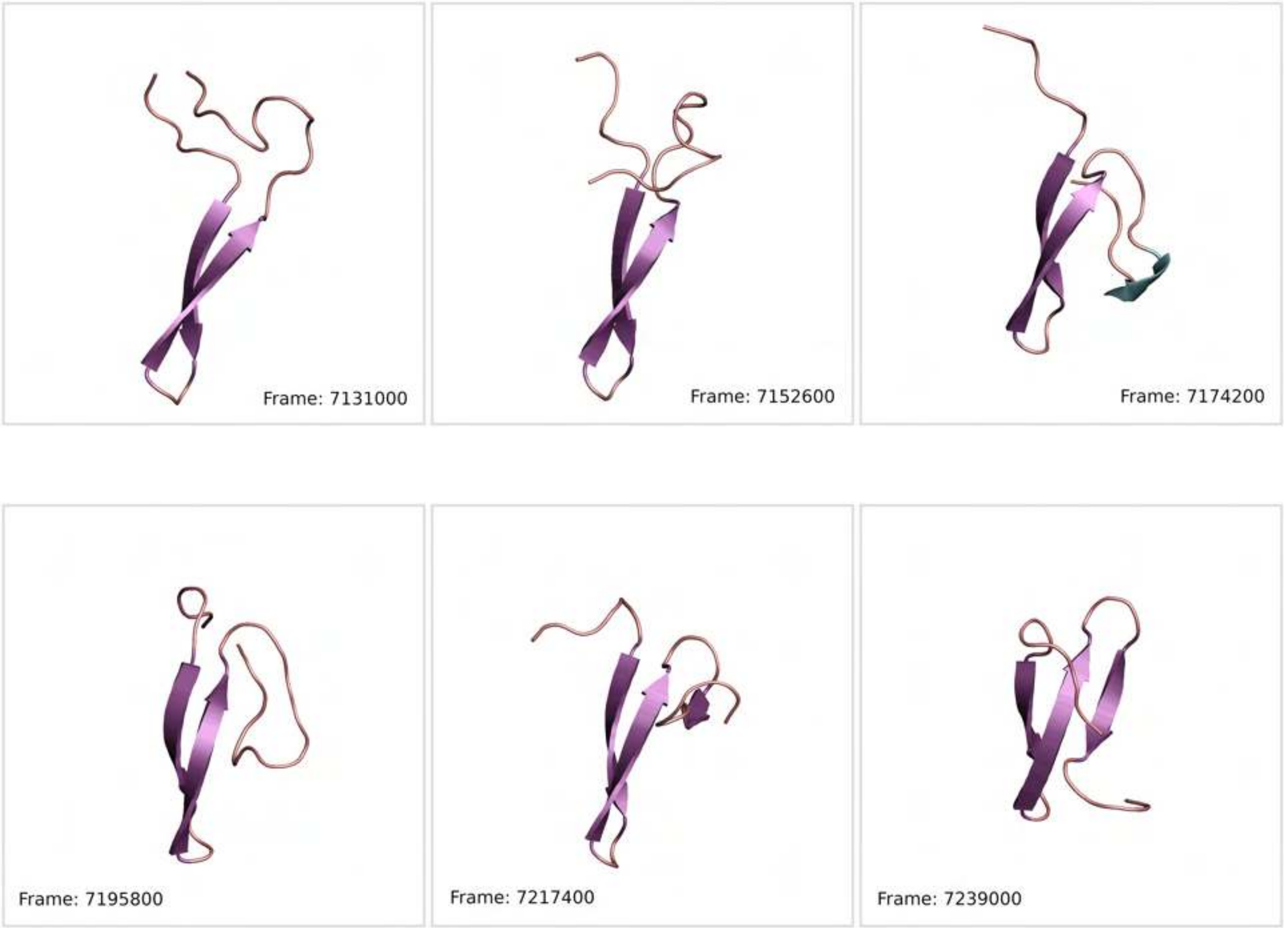


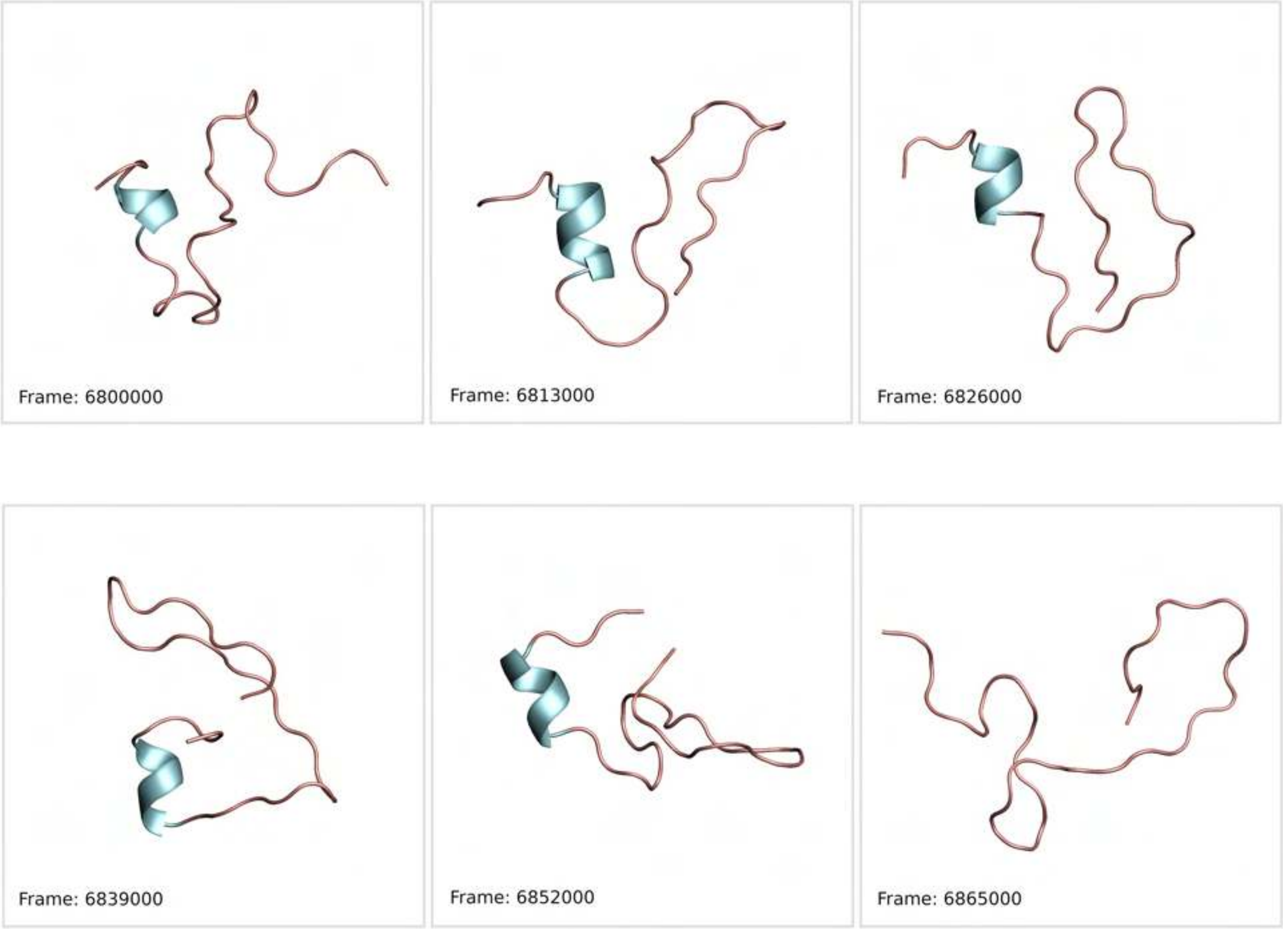
Figure 39 | *21 representative structures showing native's structure formation process starting from the initial β-strands. Using pymol, structure is colored by secondary structure: sheet (magenta), helix (cyan) and loop (salmon)*

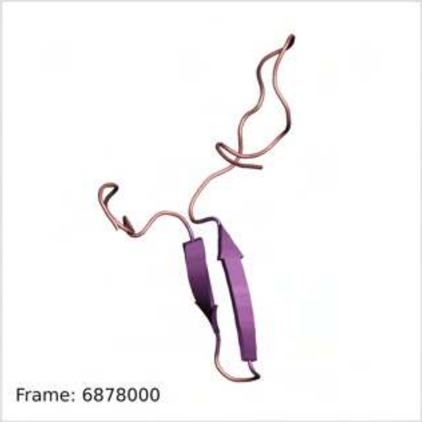
We conducted three visual representations in total. The first representation had lower resolution compared to the other two, allowing us to observe the entire folding process, from the initial formation of the *β*-turn-*β* motif to the creation of the β-sheet. Once again, we utilized the user-friendly Grcarma for this analysis, selecting the task: Extract PDB(s) (parameters: Heavy atoms, all residues, first frame: 6807000, last frame: 7239000 and step 21600). These parameters generated 21 pdb files, effectively describing this specific segment of the simulation. **Figure 39** was then created through further processing using PyMOL.

In the initial stages, before the *β*-turn-*β* motif formation, we observed the presence of various *α*-helices of different lengths. Additionally, up to the establishment of the *β*-turn-*β* motif, the chain once more underwent unfolding and no clear conformation was evident (frame: 7,001,400). This confirms that our peptide chain goes through various folding and unfolding stages before reaching its final conformation, even though it may have acquired the β-turn-β motif in previous frames.

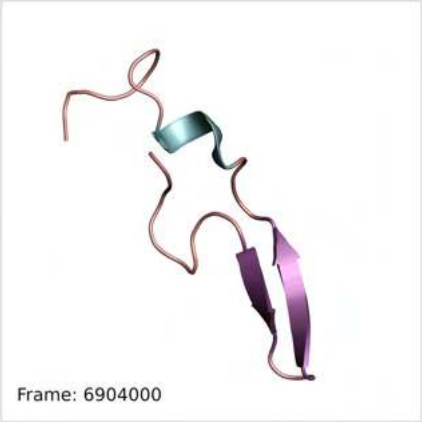
At the early phase of the existence of the two clones, specifically at the C-terminal, where the third β-strand would be formed in the future, we noticed the presence of small transient α-helices. Based on **Figure 39**, we can see that in frame 7,217,400, the third β-strand starts to appear.

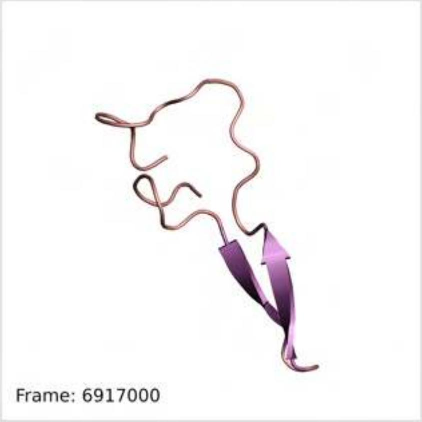
2.9.2 Frames: 6,800,000 – 7,060,000, with higher resolution, visualizing the process before the formation of the first and second *β*-strands until the stabilization of the *β*-turn-*β* motif




Frame: 6878000

Frame: 6891000

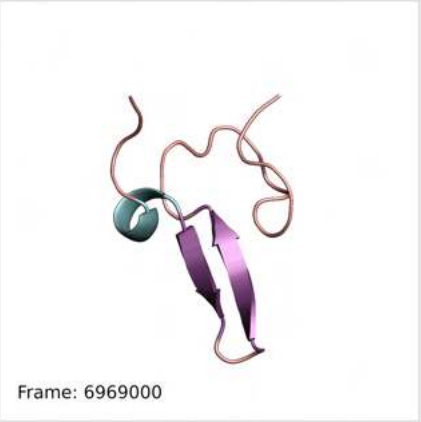

Frame: 6904000

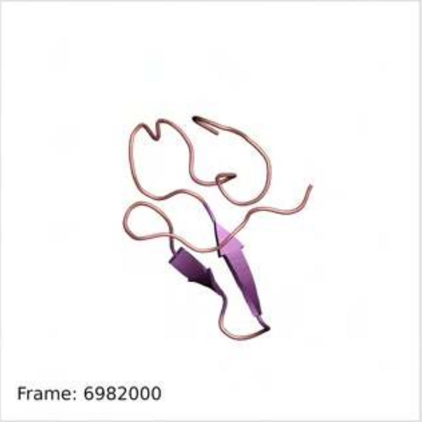

Frame: 6917000

Frame: 6930000

Frame: 6943000

Frame: 6956000


Frame: 6969000


Frame: 6982000

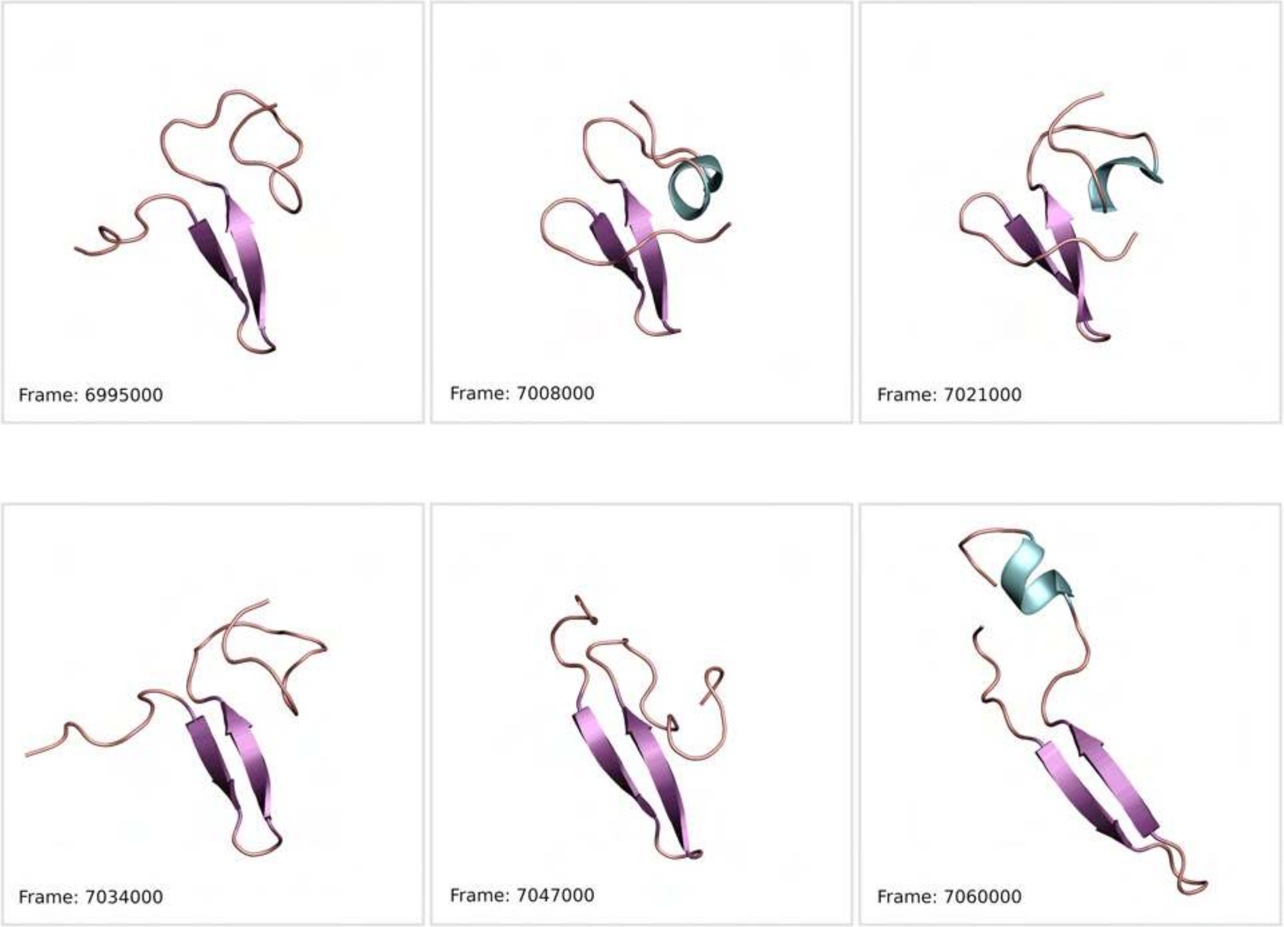
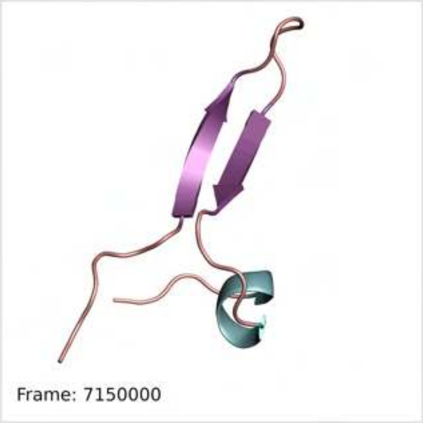


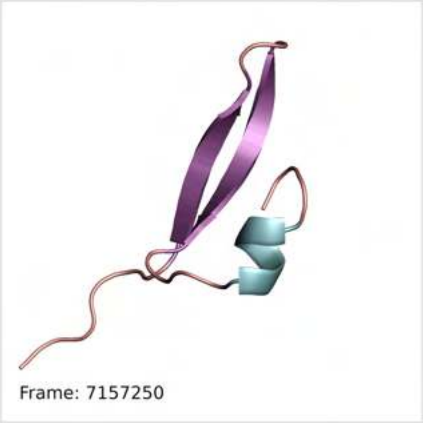
Figure 40 | *21 representative structures showing β-turn-β motif formation process starting from the initial β-strands. Using pymol, structure is colored by secondary structure: sheet (magenta), helix (cyan) and loop (salmon)*

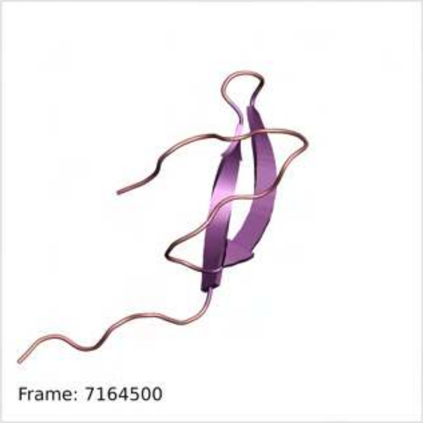
For the second visual representation, we applied higher resolution by selecting a smaller frame range and focusing on the stabilization phase of the *β*-turn-*β* motif. The parameters chosen were (heavy atoms, all residues, first frame: 6800000, last frame: 7060000 and step: 13000). These parameters generated 21 pdb files (**Figure 40**), which were further processed using PyMOL.

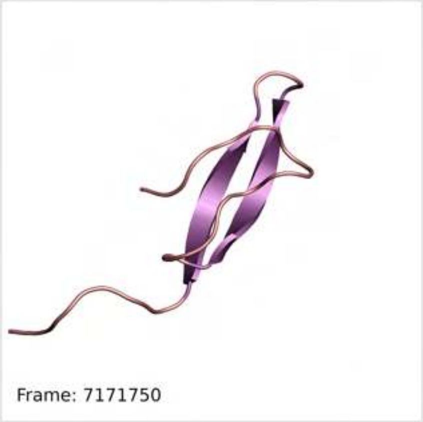
In **Figure 40**, we get a much closer look at the events that occurred in the initial frames of **Figure 39**. The first five images reveal the presence of *α*-helices, followed by the emergence of the first two *β*-strands starting from frame: 6,878,000. Similar to **Figure 39**, we still observe small *α*-helices alongside the *β*-turn-*β* motif, but in this analysis, theses *α*-helices can be detected at both the C- and N-termini of the chain.

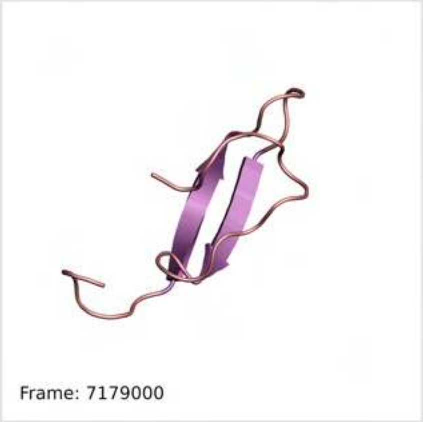
2.9.3 Frames: 7,150,000 – 7,295,000, with higher resolution, visualizing the process from shortly before the third *β*-strand attaches, up to the native structure

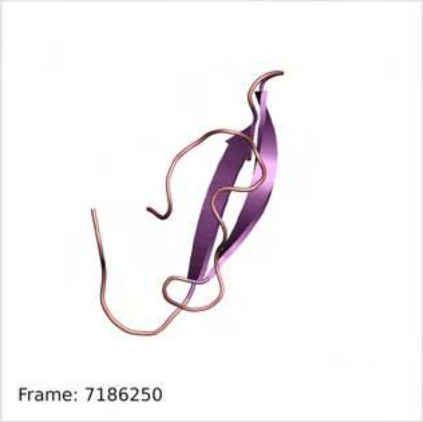


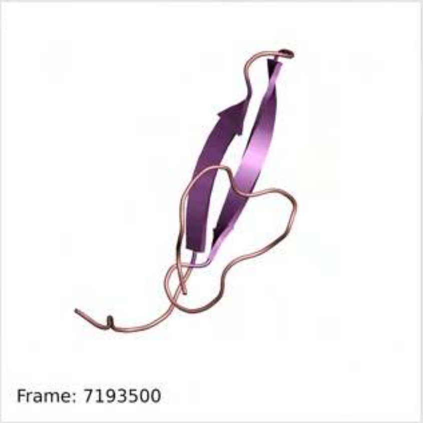


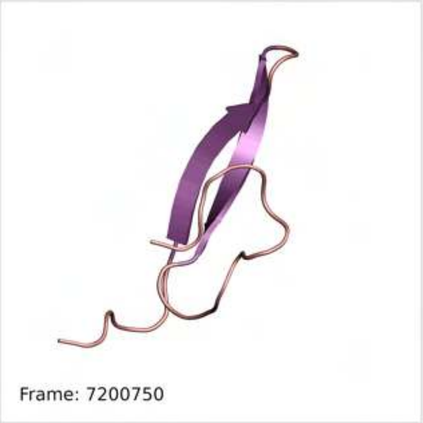


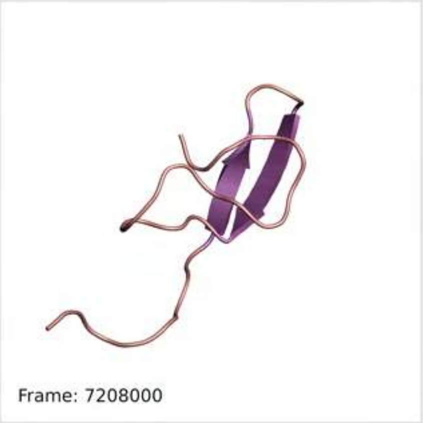












Frame: 7215250

Frame: 7222500

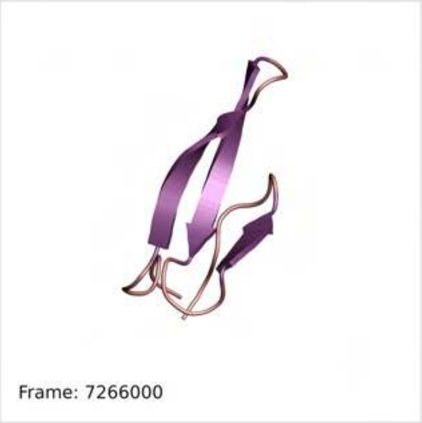
Frame: 7229750

Frame: 7237000

Frame: 7244250

Frame: 7251500

Frame: 7258750


Frame: 7266000

Frame: 7273250

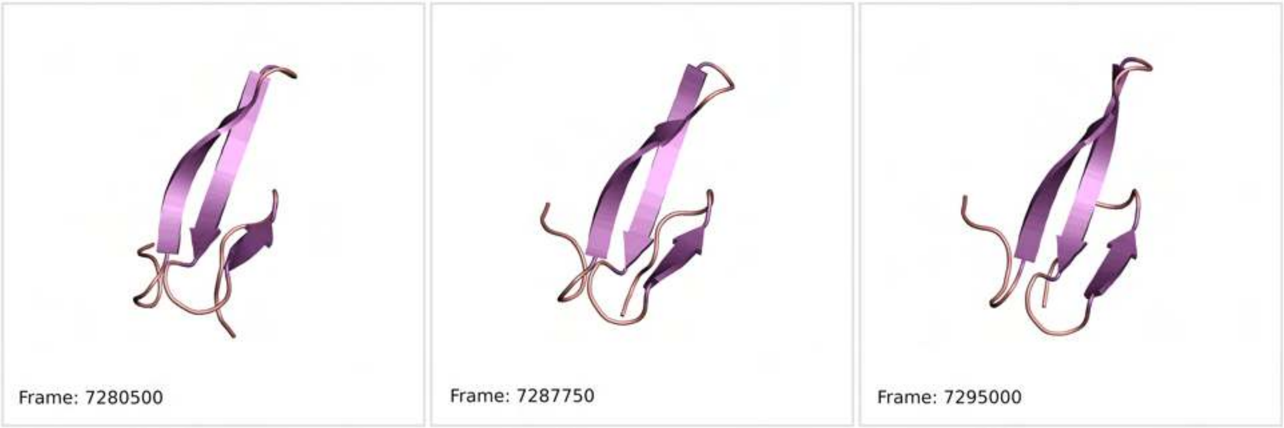


Figure 41 | *21 representative structures showing native's structure formation process starting from shortly before the third β-stand attaches. Using pymol, structure is colored by secondary structure: sheet (magenta), helix (cyan) and loop (salmon)*

For the third and final visual representation, we increased the resolution even further and focused on the stabilization phase of the native state, starting from the attachment of the third *β*-strand. Using Grcarma's task: Extract PDB(s), we selected these specific parameters (heavy atoms, all residues, first frame: 7150000, last frame: 7295000 and step: 7250) and further processed the data using PyMOL.

In **Figure 41**, these selected frames provide a closer look at the spatial stability of the *β*-turn-*β* motif and the fully formed *β*-sheet, while also revealing the dynamic movement of the remaining chain. By carefully aligning the images to display the respective conformations (*β*-sheet and *β*-turn-*β* motif) with consistent orientations, we gain visual insight into the stable and mobile segments and the extent of their motion in space.

Compared to **Figure 39**, in this particular figure, we captured the initial formation of the native state slightly later, at frame: 7,222,500.

3. Conclusions and discussion

In the final part of our thesis, we will present the conclusions that emerged throughout our research and discuss the implications of our findings. Initially, we had two trajectories to analyze, both generated in the effort to simulate the folding process of the same protein, Fip, with the main difference being the applied force field (Amber ff99SB-ILDN and Amber ff99SB*-ILDN). In our laboratory, using the revised Amber ff99SB*-ILDN force field, efforts have already been made to in order to fold various peptides with secondary structures such as α-helices or β-hairpins. [28]–[30] In this thesis, we selected the Fip mutant. The Fip protein serves as a significant representative example of a β-sheet composed of three antiparallel β-strands. Achieving the simulation of its folding from the unfolded state is of utmost importance.

Initially, in both trajectories, we performed two analyses, RMSD matrix and Secondary Structure, with the aim of determining whether and which force fields can successfully fold the chain to resemble the expected native structure. Only the ff99SB*-ILDN trajectory, at approximately 7.2 μs, managed to form a β-sheet composed of three antiparallel strands and it appeared quite prominent. Therefore, in our analyses, we quickly confirmed that the Amber ff99SB-ILDN force field is not capable of successfully folding the Fip protein.

We performed various analyses on this specific trajectory (ff99SB*-ILDN), either covering the entire simulation range or specific parts of it. In the analyses that utilized the full range, we gained a comprehensive view of our peptide chain, acquiring insights not only into the final conformation, the β-sheet, but also the transient structures that emerged throughout. Some of the tasks we conducted using the full length were as follows: **1.** Radius of Gyration: By analyzing all frames, we determined the compactness of our structure on each frame. As expected, the structure was more condensed when it acquired the final conformation, the β-sheet. However, observing the entire diagram, we identified two additional regions with significant compactness. Comparing these frames (**Figure 15**) to the corresponding ones from the secondary structure (**Figure 7**), we noticed that our structure adopted two transient α-helical conformations during those phases. **2.** Fraction of Native Contacts: Using specific reaction coordinates (Q, Qs and q), this analysis provided a detailed and in-depth examination. It identified all the bonds in the structure for each frame of the simulation and informed us (by providing similarity percentages) about their distances compared to the corresponding bonds in the native structure. In this specific analysis, the similarity percentages of the final conformation with the native structure were quite high, confirming the great promise of this particular trajectory.

With the purpose of gaining knowledge about the dynamics and motions of this specific architecture (β-sheet), we isolated and further studied the frames in which we observe the native-like structure. We calculated the covariance matrix, RMS from average, RMSF and performed secondary structure visualization. These analyses provided valuable insights into how the amino acids are correlated with each other, not only at an atomic level (residue vs. residue) but also at a secondary structure level. The neighboring β-strands (β-strand I vs β-strand II and β-strand II vs β-strand III) showed strong correlation, indicating that the motion of one strand follows the motion of the other, while the two terminal strands (β-strand I vs β-strand III) exhibited lesser correlation. Additionally, by observing the fluctuations of each amino acid, we confirmed that the amino acids forming the three β-strands are significantly more stable than those in the two loops. Furthermore, we visualized the fluctuations of the representative structure at a secondary structure level. With these insights at hand, we have gained a preliminary understanding of the dynamic regions and stable components within our structure, as well as the manner in which a β-sheet undergoes motion in space.

During our analysis of the RMSD matrix, we observed that the peptide chain reached its final conformation around the 7,200,000th frame and maintained this conformation for approximately 7,900,000 frames. However, it's important to note that these 7,900,000 frames were not identical, showing variations in the structural fluctuations. In order to identify the frame that best resembles the native structure, the most representative one, we conducted a comprehensive examination by utilizing data from the fraction of native contacts task and we creating histograms and calculating RMSDs from specific frame. The RMSDs from specific frame gave us valuable insights into the deviations exhibited by our frames throughout the simulation. The histograms, which compare each frame to the reference frame, provided us with a clear picture of how these RMSD values are distributed. Through this analysis, we identified the frame closest to the reference structure, which turned out to be the 14,365,495th frame. Intriguingly, during the calculation of the fraction of native contacts using amino acids 4-31 of the peptide chain, this frame demonstrated the highest Q value, indicating its strong resemblance to the native conformation.

To gain a deeper understanding of the atomic and collective motions within our molecular structure, we performed Dihedral PCA and Cartesian PCA analyses. Studying the density distribution of fluctuations from dPCA, we identified two main regions with a high accumulation of structures exhibiting specific fluctuations. Visualizing the pdb files of the representative structures, we deduced that both of these areas, referred to as clusters 1 and 2, represent the final conformation.

During CPCA, we isolated PC1 as it distinguishes the folded structures from the unfolded ones. By generating an image illustrating the overall motions of our system without interference from the unfolded structures, we gained valuable insights into the motion of the entire structure. This particular analysis initially confirmed that the stable regions are within the β-sheet, specifically we are referring to the three β-strands, while the mobile regions are the two termini of the chain. Furthermore, it provided information on the oscillations occurring throughout the entire structure. Oscillations were observed everywhere, with more noticeable oscillations at the two termini and significantly less in the three β-strands. However, the observed oscillations in the three β-strands are not random, they are structured in a way that the distances between them remain constant as they oscillate. This conclusion was not surprising, given that our analysis of the covariance matrix revealed a strong correlation between the amino acids in the neighboring strands. As a result, we expected that the motion of one strand would have a significant impact on the motion of the other, in a similar manner.

In conclusion, we presented three visual representations of specific folding stages during the simulation of our protein, aiming to gain insights into the step-by-step transformation of our chain, starting from the initial establishment of the β-turn-β motif and ultimately leading to the formation of the β-sheet.

The successful folding simulation of the Fip mutant using the Amber ff99SB*-ILDN force field marks a significant milestone in our ongoing quest to find a force field capable of effectively folding peptide chains. However, the journey towards this goal is far from over. Our ultimate goal is to achieve force field transferability, which means the ability to fold all types of secondary structures and their combinations, not just limited to one β-sheet and two-three α-helices. We also aim to successfully fold more complex proteins, larger polypeptides with diverse secondary structures and even proteins with quaternary structures.

This progress represents a crucial step forward in understanding the dynamics and motions of our molecular structure and opens up exciting possibilities for further research and improvement in the field of protein folding simulations.

4. Bibliography

[1] R. Ranganathan, K. P. Lu, T. Hunter, and J. P. Noel, “Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent,” *Cell*, vol. 89, no. 6, pp. 875–886, 1997, doi: 10.1016/S0092-8674(00)80273-1.

[2] M. Sudol, “Structure and function of the WW domain,” *Prog. Biophys. Mol. Biol.*, vol. 65, no. 1–2, pp. 113–132, 1996, doi: 10.1016/S0079-6107(96)00008-9.

[3] M. Jäger, H. Nguyen, M. Dendle, M. Gruebele, and J. W. Kelly, “Influence of hPin1 WW N-terminal domain boundaries on function, protein stability, and folding,” *Protein Sci.*, vol. 16, no. 7, pp. 1495–1501, 2007, doi: 10.1110/ps.072775507.

[4] H. Nguyen, M. Jäger, J. W. Kelly, and M. Gruebele, “Engineering a β-Sheet Protein toward the Folding Speed Limit,” *J. Phys. Chem. B*, vol. 109, no. 32, pp. 15182–15186, Aug. 2005, doi: 10.1021/jp052373y.

[5] F. Liu *et al.*, “An experimental survey of the transition between two-state and downhill protein folding scenarios,” *Proc. Natl. Acad. Sci.*, vol. 105, no. 7, pp. 2369–2374, Feb. 2008, doi: 10.1073/pnas.0711908105.

[6] J. A. Kowalski, K. Liu, and J. W. Kelly, “NMR solution structure of the isolated Apo Pin1 WW domain: Comparison to the x-ray crystal structures of Pin1,” *Biopolymers*, vol. 63, no. 2, pp. 111–121, Feb. 2002, doi: 10.1002/bip.10020.

[7] M. Culka and L. Rulíšek, “Factors Stabilizing β-Sheets in Protein Structures from a Quantum-Chemical Perspective,” *J. Phys. Chem. B*, vol. 123, no. 30, pp. 6453–6461, 2019, doi: 10.1021/acs.jpcb.9b04866.

[8] M. Jäger *et al.*, “Structure–function–folding relationship in a WW domain,” *Proc. Natl. Acad. Sci.*, vol. 103, no. 28, pp. 10648–10653, Jul. 2006, doi: 10.1073/pnas.0600511103.

[9] M. Jäger, H. Nguyen, J. C. Crane, J. W. Kelly, and M. Gruebele, “The folding mechanism of a β-sheet: The WW domain,” *J. Mol. Biol.*, vol. 311, no. 2, pp. 373–393, 2001, doi: 10.1006/jmbi.2001.4873.

[10] P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten, “Ten-microsecond molecular dynamics simulation of a fast-folding WW domain,” *Biophys. J.*, vol. 94, no. 10, pp. L75–L77, 2008, doi: 10.1529/biophysj.108.131565.

[11] A. D. Mackerell, M. Feig, and C. L. Brooks, “Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulation,” *J. Comput. Chem.*, vol. 25, no. 11, pp. 1400–1415, 2004, doi: 10.1002/jcc.20065.

[12] D. L. Ensign and V. S. Pande, “The Fip35 WW Domain Folds with Structural and Mechanistic Heterogeneity in Molecular Dynamics Simulations,” *Biophys. J.*, vol. 96, no. 8, pp. L53–L55, Apr. 2009, doi: 10.1016/j.bpj.2009.01.024.

[13] J. Mittal and R. B. Best, “Tackling Force-Field Bias in Protein Folding Simulations: Folding of Villin HP35 and Pin WW Domains in Explicit Water,” *Biophys. J.*, vol. 99, no. 3, pp. L26–L28, Aug. 2010, doi: 10.1016/j.bpj.2010.05.005.

[14] S. Piana, K. Sarkar, K. Lindorff-Larsen, M. Guo, M. Gruebele, and D. E. Shaw, “Computational Design and Experimental Testing of the Fastest-Folding β-Sheet Protein,” *J. Mol. Biol.*, vol. 405, no. 1, pp. 43–48, Jan. 2011, doi: 10.1016/j.jmb.2010.10.023.

[15] K. Lindorff-Larsen *et al.*, “Improved side-chain torsion potentials for the Amber ff99SB protein force field,” *Proteins Struct. Funct. Bioinforma.*, vol. 78, no. 8, pp. 1950–1958, 2010, doi: 10.1002/prot.22711.

[16] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, “Comparison of multiple Amber force fields and development of improved protein backbone parameters,” *Proteins Struct. Funct. Bioinforma.*, vol. 65, no. 3, pp. 712–725, Nov. 2006, doi: 10.1002/prot.21123.

[17] D. E. Shaw *et al.*, “Atomic-Level Characterization of the Structural Dynamics of Proteins,” *Science (80-.).*, vol. 330, no. 6002, pp. 341–346, Oct. 2010, doi: 10.1126/science.1187409.

[18] D. E. Shaw *et al.*, “Millisecond-scale molecular dynamics simulations on Anton,” in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, New York, NY, USA: ACM, Nov. 2009, pp. 1–11. doi: 10.1145/1654059.1654126.

[19] P. Robustelli, S. Piana, and D. E. Shaw, “Developing a molecular dynamics force field for both folded and disordered protein states,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 21, May 2018, doi: 10.1073/pnas.1800690115.

[20] R. B. Best and G. Hummer, “Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides,” *J. Phys. Chem. B*, vol. 113, no. 26, pp. 9004–9015, 2009, doi: 10.1021/jp901540t.

[21] M. K. Cho, S. H. Chong, S. Shin, and S. Ham, “Site-Specific Backbone and Side-Chain Contributions to Thermodynamic Stabilizing Forces of the WW Domain,” *J. Phys. Chem. B*, vol. 125, no. 26, pp. 7108–7116, 2021, doi: 10.1021/acs.jpcb.1c01725.

[22] R. B. Best, N.-V. Buchete, and G. Hummer, “Are Current Molecular Dynamics Force Fields too Helical?,” *Biophys. J.*, vol. 95, no. 1, pp. L07–L09, Jul. 2008, doi: 10.1529/biophysj.108.132696.

[23] P. S. Georgoulia and N. M. Glykos, “Using J-coupling constants for force field validation: Application to hepta-alanine,” *J. Phys. Chem. B*, vol. 115, no. 51, pp. 15221–15227, 2011, doi: 10.1021/jp209597e.

[24] K. K. Patapati and N. M. Glykos, “Three force fields views of the 3 10 helix,” *Biophys. J.*, vol. 101, no. 7, pp. 1766–1771, 2011, doi: 10.1016/j.bpj.2011.08.044.

[25] P. S. Georgoulia and N. M. Glykos, “On the foldability of tryptophan-containing tetra-and pentapeptides: An exhaustive molecular dynamics study,” *J. Phys. Chem. B*, vol. 117, no. 18, pp. 5522–5532, 2013, doi: 10.1021/jp401239v.

[26] I. Patmanidis and N. M. Glykos, “As good as it gets? Folding molecular dynamics simulations of the LytA choline-binding peptide result to an exceptionally accurate model of the peptide structure,” *J. Mol. Graph. Model.*, vol. 41, pp. 68–71, 2013, doi: 10.1016/j.jmgm.2013.02.004.

[27] P. I. Koukos and N. M. Glykos, “Folding Molecular Dynamics Simulations Accurately Predict the Effect of Mutations on the Stability and Structure of a Vammin-Derived Peptide,” *J. Phys. Chem. B*, vol. 118, no. 34, pp. 10076–10084, Aug. 2014, doi: 10.1021/jp5046113.

[28] A. S. Baltzis and N. M. Glykos, “Characterizing a partially ordered miniprotein through folding molecular dynamics simulations: Comparison with the experimental data,” *Protein Sci.*, vol. 25, no. 3, pp. 587–596, Mar. 2016, doi: 10.1002/pro.2850.

[29] A. P. Serafeim, G. Salamanos, K. K. Patapati, and N. M. Glykos, “Sensitivity of Folding Molecular Dynamics Simulations to even Minor Force Field Changes,” *J. Chem. Inf. Model.*, vol. 56, no. 10, pp. 2035–2041, 2016, doi: 10.1021/acs.jcim.6b00493.

[30] P. S. Georgoulia and N. M. Glykos, “Folding Molecular Dynamics Simulation of a gp41-Derived Peptide Reconcile Divergent Structure Determinations,” *ACS Omega*, vol. 3, no. 11, pp. 14746–14754, Nov. 2018, doi: 10.1021/acsomega.8b01579.

[31] T. Adamidou, K.-O. Arvaniti, and N. M. Glykos, “Folding Simulations of a Nuclear Receptor Box-Containing Peptide Demonstrate the Structural Persistence of the LxxLL Motif Even in the Absence of Its Cognate Receptor,” *J. Phys. Chem. B*, vol. 122, no. 1, pp. 106–116, Jan. 2018, doi: 10.1021/acs.jpcb.7b10292.

[32] I. Stylianakis *et al.*, “The balance between side-chain and backbone-driven association in folding of the <scp>α-helical</scp> influenza A transmembrane peptide,” *J. Comput. Chem.*, vol. 41, no. 25, pp. 2177–2188, Sep. 2020, doi: 10.1002/jcc.26381.

[33] A. Kolocouris, I. Arkin, and N. M. Glykos, “A proof-of-concept study of the secondary structure of influenza A, B M2 and MERS- and SARS-CoV E transmembrane peptides using folding molecular dynamics simulations in a membrane mimetic solvent,” *Phys. Chem. Chem. Phys.*, vol. 24, no. 41, pp. 25391–25402, 2022, doi: 10.1039/D2CP02881F.

[34] I. Gkogka and N. M. Glykos, “Folding molecular dynamics simulation of T-peptide, a <scp>HIV</scp> viral entry inhibitor: Structure, dynamics, and comparison with the experimental data,” *J. Comput. Chem.*, vol. 43, no. 14, pp. 942–952, May 2022, doi: 10.1002/jcc.26850.

[35] J. C. Phillips *et al.*, “Scalable molecular dynamics with NAMD,” *J. Comput. Chem.*, vol. 26, no. 16, pp. 1781–1802, 2005, doi: 10.1002/jcc.20289.

[36] P. I. Koukos and N. M. Glykos, “Grcarma: A fully automated task-oriented interface for the analysis of molecular dynamics trajectories,” *J. Comput. Chem.*, vol. 34, no. 26, pp. 2310–2312, Oct. 2013, doi: 10.1002/jcc.23381.

[37] N. M. Glykos, “Software news and updates carma: A molecular dynamics analysis program,” *J. Comput. Chem.*, vol. 27, no. 14, pp. 1765–1768, Nov. 2006, doi: 10.1002/jcc.20482.

[38] N. Ferguson, C. M. Johnson, M. Macias, H. Oschkinat, and A. Fersht, “Ultrafast folding of WW domains without structured aromatic clusters in the denatured state,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 23, pp. 13002–13007, Nov. 2001, doi: 10.1073/pnas.221467198.

[39] S. Deechongkit, H. Nguyen, E. T. Powers, P. E. Dawson, M. Gruebele, and J. W. Kelly, “Context-dependent contributions of backbone hydrogen bonding to β-sheet folding energetics,” *Nature*, vol. 430, no. 6995, pp. 101–105, 2004, doi: 10.1038/nature02611.

[40] S. V. Krivov, “The Free Energy Landscape Analysis of Protein (FIP35) Folding Dynamics,” *J. Phys. Chem. B*, vol. 115, no. 42, pp. 12315–12324, Oct. 2011, doi: 10.1021/jp208585r.

[41] T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, and V. S. Pande, “Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories,” *J. Am. Chem. Soc.*, vol. 133, no. 45, pp. 18413–18419, Nov. 2011, doi: 10.1021/ja207470h.

[42] C. M. Davis and R. B. Dyer, “Dynamics of an Ultrafast Folding Subdomain in the Context of a Larger Protein Fold,” *J. Am. Chem. Soc.*, vol. 135, no. 51, pp. 19260–19267, Dec. 2013, doi: 10.1021/ja409608r.

[43] C. M. Davis and R. B. Dyer, “The Role of Electrostatic Interactions in Folding of β-Proteins,” *J. Am. Chem. Soc.*, vol. 138, no. 4, pp. 1456–1464, Feb. 2016, doi: 10.1021/jacs.5b13201.

[44] T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*, vol. 21. in Interdisciplinary Applied Mathematics, vol. 21. New York, NY: Springer New York, 2010. doi: 10.1007/978-1-4419-6351-2.

[45] G. Ochbaum and R. Bitton, “Using small-angle X-ray scattering (SAXS) to study the structure of self-assembling biomaterials,” in *Self-assembling Biomaterials*, Elsevier, 2018, pp. 291–304. doi: 10.1016/B978-0-08-102015-9.00015-0.

[46] R. B. Best, G. Hummer, and W. A. Eaton, “Native contacts determine protein folding mechanisms in atomistic simulations,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 44, pp. 17874–17879, Oct. 2013, doi: 10.1073/pnas.1311599110.

[47] T. Ichiye and M. Karplus, “Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations,” *Proteins Struct. Funct. Bioinforma.*, vol. 11, no. 3, pp. 205–217, 1991, doi: 10.1002/prot.340110305.

[48] C. L. Sabharwal and B. Anjum, "Data Reduction and Regression Using Principal Component Analysis in Qualitative Spatial Reasoning and Health Informatics," *Polibits*, vol. 53, pp. 31–42, Jan. 2016, doi: 10.17562/PB-53-3.

[49] L. Smith, "A tutorial on Principal Components Analysis," *Commun. Stat. - Theory Methods*, vol. 17, no. 9, pp. 3157–3175, 1988, [Online]. Available: http://www.mendeley.com/research/computational-genome-analysis-an-introduction-statistics-for-biology-and-health/%5Cnhttp://www.tandfonline.com/doi/abs/10.1080/03610928808829796

[50] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," *Proteins Struct. Funct. Bioinforma.*, vol. 17, no. 4, pp. 412–425, 1993, doi: 10.1002/prot.340170408.

[51] Y. Mu, P. H. Nguyen, and G. Stock, "Energy landscape of a small peptide revealed by dihedral angle principal component analysis," *Proteins Struct. Funct. Genet.*, vol. 58, no. 1, pp. 45–52, 2005, doi: 10.1002/prot.20310.

[52] A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, "Dihedral angle principal component analysis of molecular dynamics simulations," *J. Chem. Phys.*, vol. 126, no. 24, pp. 1–10, 2007, doi: 10.1063/1.2746330.

[53] S. Mukherjee and Y. Zhang, "MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming," *Nucleic Acids Res.*, vol. 37, no. 11, pp. e83–e83, Jun. 2009, doi: 10.1093/nar/gkp318.