DEMOCRITUS UNIVERSITY OF THRACE
SCHOOL OF HEALTH SCIENCES
DEPARTMENT OF MOLECULAR BIOLOGY & GENETICS

BSc THESIS

The Repressor of Primer protein: to be native-like, or not to be?
Comparative computational studies of the D30P vs A31P mutants

Svingou Nikoleta, 2658

Academic Supervisor: Nicholas M. Glykos
Associate Professor, Laboratory of Structural and Computational Biology,
Democritus University of Thrace

Alexandroupolis, Greece
March 2026

DEMOCRITUS UNIVERSITY OF THRACE
SCHOOL OF HEALTH SCIENCES
DEPARTMENT OF MOLECULAR BIOLOGY & GENETICS

BSc THESIS

The Repressor of Primer protein: to be native-like, or not to be?
Comparative computational studies of the D30P vs A31P mutants

Svingou Nikoleta, 2658

Academic Supervisor: Nicholas M. Glykos
Associate Professor, Laboratory of Structural and Computational Biology, Democritus University of Thrace

I declare that the present thesis entitled 'The Repressor of Primer protein: to be native-like, or not to be? Comparative computational studies of the D30P vs A31P mutants' is original and was carried out by me personally, as an undergraduate student of the Department of Molecular Biology and Genetics, with Registration Number 2658. I certify that during the preparation and writing of the thesis, all legal requirements were followed, and that the principles of academic ethics and integrity were fully adhered to, which prohibit the falsification of results, the misuse of others' intellectual property, and plagiarism.

Alexandroupolis, Greece
March 2026

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ

ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ & ΓΕΝΕΤΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Η πρωτεΐνη Rop: να είσαι στην φυσική σου μορφή ή να μην είσαι? Συγκριτικές υπολογιστικές μελέτες των μεταλλαγμάτων D30P και A31P

Σβίγγου Νικολέτα, 2658

Επιβλέπων καθηγητής: Νικόλαος Μ. Γλυκός
Αναπληρωτής καθηγητής, Εργαστήριο Δομικής και Υπολογιστικής βιολογίας

Δηλώνω οτι η παρούσα εργασία με τίτλο “Η πρωτεΐνη Rop: Να διατηρείται η φυσική διαμόρφωση ή όχι? Συγκριτικές υπολογιστικές μελέτες των μεταλλαγμάτων D30P και A31P” είναι πρωτότυπη και πραγματοποιήθηκε από́ εμένα προσωπικά́, προπτυχιακό φοιτητή́ του Τμήματος Μοριακής Βιολογίας και Γενετικής, με Αρ. Μητρώου 2658. Βεβαιώνω ότι κατά́ την εκπόνηση της εργασίας και τη συγγραφή́ της τηρήθηκαν τα προβλεπόμενα από́ το νόμο, καθώς και ότι ακολουθήθηκαν πλήρως οι αρχές της ακαδημαϊκής ηθικής και δεοντολογίας, οι οποίες απαγορεύουν την παραποίηση των αποτελεσμάτων, την κατάχρηση της διανοητικής ιδιοκτησίας άλλων και τη λογοκλοπή́.

Αλεξανδρούπολη, Ελλάδα
Μάρτιος 2026

Acknowledgments

First of all, I would like to express my gratitude to my supervisor, Dr Nicholas M. Glykos, for his continuous and unwavering support throughout my thesis. He inspired me to engage with a field that was previously unfamiliar to me, while his guidance was invaluable for the completion of my first research project.

I would also like to thank my friends who have been by my side throughout this journey, offering moments of relief during the most demanding periods of my studies.

Last, but certainly not least, I would like to thank my parents, who have been an endless source of support and comfort. Without them, nothing would be the same, and above all, I would not be the person I am today.

Table of Contents

Acknowledgments..iv

Abstract..vii

Περίληψη..viii

1. Introduction..1

1.1 Proteins..1

1.2 Protein Folding..3

1.2.1 Protein Folding Problem..3

1.2.2 Molecular Dynamics Simulations Aimed at Investigating Protein Folding............4

1.2.3 Force Fields..6

1.3 Structural Motifs: Coiled Coils and 4-α-helical Bundles..7

1.4 The Repressor of Primer Protein (Rop)..8

1.4.1 Molecular Mechanism of Rop Protein in E.coli..8

1.4.2 Structural Characterization of Rop Protein..9

1.4.3 Structural Characterization of Rop Mutants: D30P & A31P..............................10

1.5 Aim and Scope of the Thesis..12

2. Methods..13

2.1 Molecular Dynamics Simulation Systems: Details and Parameters............................13

2.2 Software Used for Molecular Dynamics Trajectory Analysis and Structural Visualization..14

2.2.1 Trajectory Analysis Using Carma and Grcarma..14

2.2.2 XMGR and PyMOL: Plotting and Structural Visualization Tools......................15

2.3 MD Trajectory Analysis: RMSF, RMSD and PCA..15

2.3.1 RMSF Analysis..16

2.3.2 RMSD Analysis..16

2.3.2.1 Statistical Analysis Using the sn Package in R..19

2.3.3 Principal Component Analysis..21

2.3.3.1 Dihedral PCA of the Turn Regions of Both Monomers................................22

2.3.3.2 Cartesian PCA of the Entire Trajectories..24

2.4 Distance and Angle Plots..25

2.5 Ramachandran Plots..26

3. Results.....27
3.1 RMSF Analysis.....27
3.2 RMSD Analysis.....29
3.2.1 Frames Exhibiting the Highest RMSD Values in the Turn Region.....34
3.2.2 RMSD Histograms.....36
3.2.3 Statistical Analysis Based on R Package “sn”.....42
3.3 Dihedral PCA.....53
3.3.1 Application of Dihedral PCA Using Carma and Grcarma.....54
3.3.2 Dihedral PCA Representative Structures.....74
3.4 Cartesian PCA of the Dominant Cluster.....79
3.5 Interpretation of the Different Folding Behavior of the D30P vs A31P Mutants.....82
3.5.1 Hydrogen bonding interactions of Ala31.....82
3.5.2 Ramachandran plots for residues 29-32.....86
4. Discussion – Conclusions.....90
5. References.....91

Abstract

Molecular dynamics simulations constitute a powerful computational tool for the study of the dynamic behavior of biological macromolecules, providing the ability to investigate their folding and stability at the atomic level. In the present thesis, molecular simulations of the native Rop structure, as well as of the D30P and A31P mutants, were studied. Rop is a homodimeric protein found in *Escherichia coli*. Each monomer adopts a coiled-coil supersecondary structure, while the quaternary structure is organized into a 4-α-helical bundle. Experimental data have shown that the substitution with a proline residue at position 30 or 31 in the turn region results in different structural consequences of each mutant. The D30P variant appears to retain a conformation that closely resembles the native one (native-like), whereas the A31P mutant exhibits a significantly altered topology, known as "bisecting U". This thesis presents a comparative study of the three trajectories (native Rop, D30P and A31P) using the GUI Grcarma program, which is based on the molecular dynamics analysis program Carma. The aim of this study is to evaluate the ability of molecular dynamics simulations to reproduce the available experimental data, as well as to attempt to interpret the underlying causes that lead to the preservation of the native protein conformation in one case, whereas the corresponding mutation at the adjacent position results in structural rearrangement and destabilization.

Keywords: Molecular Dynamics simulations, native Rop, D30P, native-like, A31P, bisecting U, folding, point-mutation, α-helix, turn

Περίληψη

Οι προσομοιώσεις μοριακής δυναμικής αποτελούν ένα ισχυρό υπολογιστικό εργαλείο για τη μελέτη της δυναμικής συμπεριφοράς βιολογικών μακρομορίων, παρέχοντας τη δυνατότητα διερεύνησης της αναδίπλωσης και της σταθερότητας σε ατομικό επίπεδο. Στην παρούσα πτυχιακή εργασία μελετήθηκαν οι μοριακές προσομοιώσεις της native δομής της Rop, καθώς και των μεταλλαγμάτων D30P και A31P. Η πρωτεΐνη Rop απαντάται στο βακτήριο *Escherichia coli* και αποτελεί ομοδιμερές, στο οποίο κάθε μονομερές υιοθετεί υπερδευτεροταγή δομή coiled-coil, ενώ η τεταρτοταγής δομή οργανώνεται σε 4-α-ελικοειδές δεμάτιο. Πειραματικά δεδομένα έχουν δείξει οτι η αντικατάσταση με κατάλοιπο προλίνης στη θέση 30 ή 31 της περιοχής της στροφής επιφέρει διαφορετικές συνέπειες στη διαμόρφωση της δομής του κάθε μεταλλάγματος. Το μετάλλαγμα D30P φαίνεται να διατηρεί διαμόρφωση που ομοιάζει σημαντικά με τη native (native-like), ενώ στο μετάλλαγμα A31P παρατηρείται μεταβολή της τοπολογίας και υιοθέτηση νέας διαμόρφωσης, γνωστής ως “bisecting U”. Η παρούσα εργασία περιλαμβάνει συγκριτική ανάλυση των τριών προσομοιώσεων (native Rop, D30P & A31P) με τη χρήση του γραφικού περιβάλλοντος Grcarma, το οποίο βασίζεται στο πρόγραμμα ανάλυσης μοριακών προσομοιώσεων Carma. Στόχος της μελέτης είναι η αξιολόγηση της ικανότητας των μοριακών προσομοιώσεων να αναπαράγουν τα διαθέσιμα πειραματικά αποτελέσματα, καθώς και η απόπειρα ερμηνείας των αιτιών που οδηγούν στη διατήρηση της διαμόρφωσης στη μία περίπτωση, ενώ η αντίστοιχη μετάλλαξη στην αμέσως γειτονική θέση οδηγεί σε δομική αναδιοργάνωση και αποσταθεροποίηση.

Λέξεις κλειδιά: Προσομοιώσεις Μοριακής Δυναμικής, φυσική διαμόρφωση Rop, D30P, A31P, bisecting U, αναδίπλωση, σημειακή μετάλλαξη, α-έλικα, βρόχος

1. Introduction

1.1 Proteins

Proteins constitute the essential macromolecules that form the basis of life, performing the majority of structural, catalytic, and regulatory functions within the cells of all organisms. Amino acids are the building blocks of proteins, as they are connected together through peptide bonds, constructing polypeptide chains that appear in an outstanding variety of sizes and shapes, performing many functions. Proteins are the "workhorses" of living systems, being responsible for a broad range of biological functions due to their structural versatility and conformational changes [1].

Every single amino acid possesses a unique chemical nature defined by its side chain, which determines the physicochemical nature of the polypeptide. These properties navigate the spontaneous folding into a specific three-dimensional conformation. Non-covalent interactions stabilize the final conformation, including hydrogen bonds, hydrophobic interactions, ionic bonds, and van der Waals contacts [2]. This three-dimensional conformation is far from random, as it is the key to the proper function of the molecule and its ability to recognize and interact with other biomolecules.

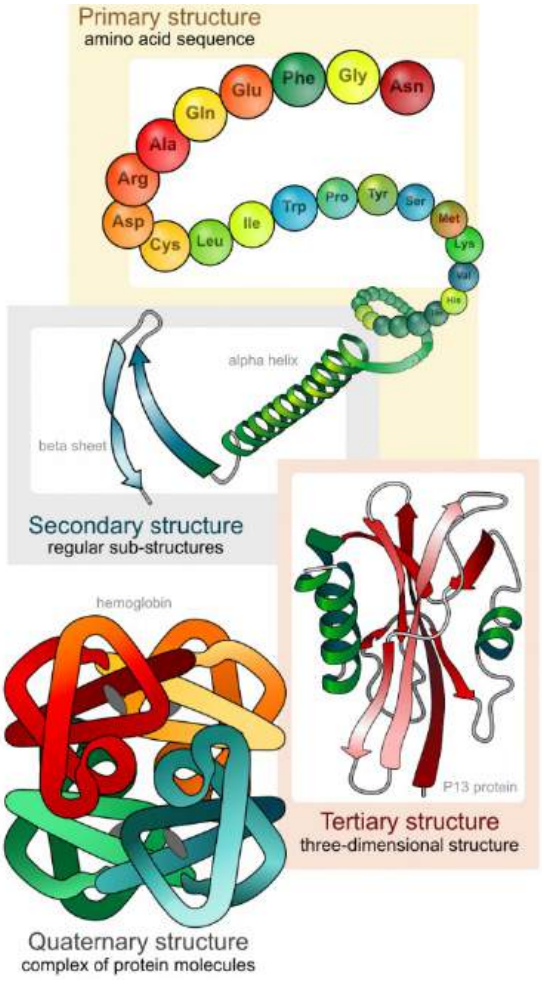

Figure 1: *The four structural stages in the protein folding process (primary, secondary, tertiary and quaternary levels). Reproduced without permission from Lisa Bartee (2019).*

Considering a deeper level, the variety and specificity of proteins represent the complexity of life. Proteins are part of dynamic networks that regulate gene expression, metabolism, and intracellular communication. Even a small alteration in the sequence, which is reflected in the structure may lead to misfolding and, eventually, to a loss of function. Therefore, understanding protein structure has a fundamental role in fields such as molecular biology, medicine and biotechnology [3].

As illustrated in **Figure 1**, the process of protein folding is distinctly organized in a hierarchical manner into four levels. The first of them, known as the *primary structure*, corresponds to the linear amino acids sequence, as determined by the nucleotide sequence of the respective gene. Peptide bonds between amino acids are formed through dehydration reactions, producing one water molecule for each bond.

Hydrogen bonding between backbone atoms of the polypeptide chain generates localized folding motifs that constitute the *secondary structure.* The most common motifs are α-helices and β-strands, which arise from regular coiling or folding of the peptide main chain. The secondary structure depends on the spatial arrangement of main-chain atoms rather than the side-chain positions. This structural organization is followed by the *tertiary structure,* which determines the three-dimensional conformation of the protein. Interactions between distant side chains within the secondary structure elements give rise to a functional conformation. Finally, proteins composed of two or more polypeptide chains display the final structural level, known as the *quaternary structure,* which describes the association of multiple subunits through non-covalent interactions to form a functional complex [4].

1.2 Protein Folding

1.2.1 Protein Folding Problem

The protein folding problem pertains to the manner in which the primary structure of a protein dictates its specific three-dimensional conformation. In other words, it poses the fundamental question of how the linear sequence of amino acids encodes the structural data required for the protein to attain its functional form. After a series of experiments on ribonuclease, Anfinsen (1973) demonstrated that the native configuration of a protein is determined exclusively by its amino acid sequence, implying that all essential instructions for folding are encoded within the linear sequence itself. He also claimed that the native state corresponds to the most thermodynamically favorable conformation, which reflects the lowest total energy. However, due to the immense range of accessible structural states, even for a modest-sized protein, the question of how proteins reach their native structure so rapidly and accurately has challenged scientists for years, remaining one of the most fundamental unsolved issues in molecular biology [5].

Cyrus Levinthal (1968) became known for the *Levinthal paradox*, as he demonstrated that if a protein were to explore all possible conformations randomly to reach its native state, it would take an astronomically long time. In reality, however, proteins fold within milliseconds to seconds. This paradox revealed that protein folding is not a trial-and-error mechanism in which the molecule passes through every possible state. Instead, there exist energetically favorable pathways along which each molecule efficiently finds its stable native conformation. This information led to the concept of the protein folding energy landscape, where folding

proceeds through a funnel-shaped pathway toward the lowest free-energy state [6][7][8].

Later theoretical and computational studies provided an explanation for how this paradox could be understood. Martin Karplus (1997) introduced a perspective that eliminated the concept of random search, proposing that folding is guided by a bias toward the native state over much of the conformational space. Statistical mechanical models and lattice simulations showed that proteins can reach their native conformation through multiple parallel pathways, rather than a single individual route. According to this idea, the *folding funnel* concept was developed, describing how the free energy landscape of a protein narrows progressively toward the native structure. Hence, even though there are numerous possible configurations, rapid folding is enabled, giving science a new perspective [7][9].

The shift from the assumption that protein folding is nearly impossible to the realization that energetic biases and directed pathways make it achievable has been significant in enhancing the field. Although a complete mechanistic explanation of protein folding kinetics and thermodynamics remains an ongoing scientific challenge. The combination of Anfinsen's theory, Levinthal's paradox, and Karplus's energy landscape model provides the theoretical basis for understanding how sequence determines structure [10][11].

1.2.2 Molecular Dynamics Simulations Aimed at Investigating Protein Folding

While the theoretical models described above contribute significantly to the conceptual understanding of the protein folding problem, computational simulations enable an atom-detailed study of the field.

Among these methods, molecular dynamics (MD) is one of the most powerful computational approaches for investigating folding mechanisms. The fundamental concept of MD involves applying Newtonian mechanics to predict the motion of all particles within a molecular system, thereby simulating the folding process over time under the influence of physical forces. These forces are computed as the negative gradient of the system's energy landscape [12][10].

To make this possible, the simulation generates atomic trajectories, which are used for the description of time-dependent coordinates and velocities of each atom in the molecular system. These trajectories provide a detailed representation of molecular motion, making them valuable for future calculations of thermodynamic and structural properties [12].

To accurately represent atomic and molecular movements, it is essential to determine the system's overall energy (given as the sum of kinetic and potential energies). The computation of the kinetic energy is straightforward, as it is based on the masses and velocities of the atoms. However, the potential energy cannot be determined precisely, since it would require solving the Schrödinger's equation, which is not computationally feasible for large biomolecular systems. Therefore, empirical methods known as *force fields* have been developed to approximate potential energy, using simplified equations derived from experiments or quantum mechanical calculations [15][16].

The trajectories discussed earlier contain detailed information about the motion of all atoms over time. By applying statistical mechanics, these microscopic data can be used for calculating macroscopic thermodynamic properties, including temperature, pressure, and internal energy. For obtaining accurately these values, it is essential to describe the energy landscape of the system, which would not be possible without the use of force fields [13].

1.2.3 Force Fields

Force fields describe total potential energy as the sum of two main components: The first includes *bonded interactions,* and the second encompasses non-bonded interactions, as shown below.

Bonded (intramolecular, internal), terms

$$E_{bonded} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{\substack{improper \\ dihedrals}} K_\varphi(\varphi - \varphi_0)^2 + \sum_{dihedrals} \sum_{n=1}^{6} K_{\phi,n}(1 + \cos(n\phi - \delta_n))$$

Nonbonded (intermolecular, external) terms

$$E_{nonbonded} = \sum_{\substack{nonbonded \\ pairs\, ij}} \frac{q_i q_j}{4\pi D r_{ij}} + \sum_{\substack{nonbonded \\ pairs\, ij}} \varepsilon_{ij} \left[\left(\frac{R_{\min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\min,ij}}{r_{ij}} \right)^{6} \right]$$

Figure 2: *Potential energy functions used in molecular mechanics. Adapted without permission from Kenno Vanommeslaeghe, Olgun Guvench and Alexander D MacKerell (2014)*

As illustrated in **Figure 2**, bonded (internal) interactions include bond length variations, angle bending, dihedral or improper rotations. Due to the latter, the maintenance of molecular geometry and the avoidance of non-realistic conformations are achieved [13]. As a result of these internal energy terms, atoms remain connected within the molecule.

On the other hand, non-bonded interactions include van der Waals forces, encompassing permanent dipole-dipole, dipole-induced dipole and London dispersion contributions which are represented by the Lennard-Jones potential, as well as electrostatic interactions.

In conclusion, these components collectively define the potential energy function that determines atomic motion in molecular dynamics simulations.

1.3 Structural Motifs: Coiled Coils and 4-α-helical Bundles

Coiled coils represent one of the characteristic structural arrangements found in proteins, formed by two or more right-handed α-helices that intertwine into a left-handed superhelical structure. Each helix exhibits a specific sequence pattern known as the *heptad repeat,* where residues at the "a" and "d" positions are typically hydrophobic, whereas those at "e" and "g" are commonly charged. Positions "b", "c" and "f" tend to accommodate polar and solvent-exposed residues, forming interactions with the surrounding environment. This periodicity enhances the establishment of a hydrophobic core through the integration of side chains along the α-helix. As suggested by Francis Crick, α-helices are packed together in a specific arrangement known as the *knobs-into-holes* model. According to this model, a side chain from one helix occupies the grooves formed by four side chains of the neighboring helix. The alignment of the two helices is highly specific, with the "d" residues positioned directly opposite one another along the interface [2][17][18].

Besides coiled coils, another common α-helical motif is a 4-α-helical bundle (4HB). It belongs to the quaternary level of protein structure, as it is typically formed by two separate monomers. The 4HB consists of two coiled coil motifs packed together into a compact, cylindrical structure stabilized by a well-organized hydrophobic core [2]. The orientation of the helices is most commonly antiparallel, although parallel arrangements have also been observed, particularly in designed systems [20]. Similar to coiled coil motifs, the main stabilization of the bundle is achieved through

hydrophobic interactions within its core, formed by residues buried inside. Additionally, polar and charged residues contribute to further stabilization by forming both solvent-mediated and electrostatic interactions, such as salt bridges.

The knobs-into-holes model is preserved in both motifs, although the arrangement of the four helices is more symmetrical and compact, which enhances rigidity and stability [20]. Depending on the primary structure of a 4HB and the interhelical angles formed between the helices, the structural topology can vary considerably, adopting left-handed, right-handed or mixed arrangements [21]. This phenomenon is exemplified by the ColE1 Rop protein, in which the topology of the quaternary structure can adopt alternative conformations while remaining comparably stable [22].

1.4 The Repressor of Primer Protein (Rop)

1.4.1 Molecular Mechanism of Rop Protein in *Escherichia coli*

The Repressor of Primer protein, also referred to as RNA-one-modulator (ROM), is a small homodimeric molecule involved in controlling the replication of ColE1 plasmids in *Escherichia coli*. It does not act as a specific RNA-binding protein for the RNAI or RNAII but exhibits high affinity for the region of the RNA complex, known as the “initial kissing complex”. Rop acts as a structural catalytic protein that stabilizes the initially unstable kissing complex, transforming it into a more stable double-helical RNA-protein assembly. The antiparallel helices that form the 4HB topology create a positively charged interface, providing an ideal

environment for double interaction with oppositely charged RNA molecules and stabilizing the intermediate region between them. In conclusion, in the absence of Rop, the interaction between RNA molecules occurs more slowly, leading to an increased production of plasmid copies, whereas the presence of Rop promotes rapid duplex formation, thereby limiting the replication initiation [23][24][25].

1.4.2 Structural Characterization of Rop Protein

At structural level, each Rop monomer is a small α-helical protein composed of only 63 amino acids. The monomer adopts an extremely regular helix-turn-helix motif that contributes to the construction of a superhelical homodimer. The dimer, depicted in **Figure 3**, exhibits a clear and well-organized topology, in which two identical monomers pack together symmetrically through extensive hydrophobic interactions within the core [26]. This arrangement results in a remarkably stable and ordered quaternary structure, establishing Rop as an important model for studying such structural motifs [2]. Along with hydrophobic contacts, electrostatic interactions between residues carrying opposite charges of the heptad repeat further contribute to the protein's overall stability [22] [24].

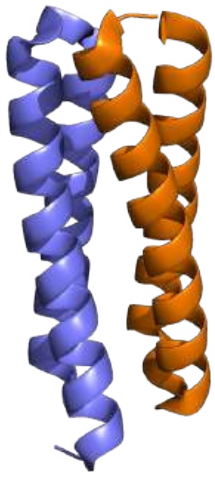

***Figure 3**: 3D representation of the 4-α-helical bundle of Rop native conformation. Visualized using PyMOL. The model was originally reported by Kokkinidis, Banner and Tsernoglou (1992) (PDB ID: 1ROP).*

1.4.3 Structural Characterization of Rop Mutants: D30P & A31P

An experimental study involving the substitution of all amino acids at position 30 was conducted to investigate their contribution in the structural stability of the protein [27]. The findings indicated that replacing aspartic acid with proline at this position has little to no impact on protein stability. The D30P mutant retains a conformation that closely resembles the native structure (native-like), whose overall topology remains unaltered. This implies that introducing a proline residue at position 30, located within the turn region, does not disrupt the protein's native fold [27][28].

On the other hand, the A31P mutant displays a markedly different topology, in which the typical, antiparallel orientated, negative superhelical four-helix bundle of native Rop is transformed into a positive superhelical fold comprising a mixture of parallel and antiparallel helical alignments [28]. This topology is known as *bisecting U* (illustrated in **Figure 4**), characterized by a complete reorganization of the hydrophobic core, exhibiting an alternative pattern of residue packing within the heptad repeat. As demonstrated, the fundamental interactions observed in the knobs-into-holes model are vanished, while unconventional contact patterns such as "dddd", "ggaaa" and "gdd" emerge instead [28]. Crystallographic refinement of the A31P mutant further revealed that this large-scale structural rearrangement is caused by the insertion of a rigid proline ring at position 31 within the turn region [29]. Consequently, the local backbone flexibility increased, causing an overall destabilization in the, otherwise, well-organized Rop structure. The altered turn folding promotes partial unfolding, exposing part of the hydrophobic core into the solvent, contributing to the reduced thermodynamic stability of the A31P

variant relative to the native protein [30]. Despite the structural polymorphism adopted by Rop, its molten-globule characteristics remain essentially unaffected, indicating that, even under different topological arrangements, the protein retains high conformational plasticity while maintaining its overall functional properties [31].

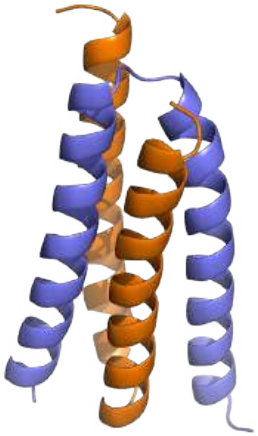

Figure 4: *3D representation of the bisecting U topology of the A31P Rop mutant, visualized using PyMOL. The model was originally reported by Glykos, Cesareni and Kokkinidis (1999). (PDB ID: 1B6Q)*

1.5 Aim and Scope of the Thesis

The present thesis investigates the capability of Molecular Dynamics (MD) simulations to reproduce, in a fully reliable manner, experimental findings concerning the stability of two specific mutants (D30P & A31P) of the Rop protein. In particular, experimental studies have demonstrated that the D30P mutant retains a native-like conformation, although its lower melting temperature value ($Tm = 58.9^{o}C$) compared to the native protein ($Tm = 68.7^{o}C$) indicates reduced structural stability [27]. While previous modeling approaches and structural predictions suggested that the A31P mutant adopts a native-like fold [28], experimental evidence has revealed that its topology rearranges into a right-handed, antiparallel 4-α-helical bundle, known as the "bisecting U" structure [42]. The Tm was found to be significant decreased ($Tm = 43.0^{o}C$), indicating a less compact structure compared to the native protein which, under the same experimental conditions, was determined to be $64.5^{o}C$ [30]. Importantly, the molecular dynamics simulations were not based on the experimentally determined structure, but rather on a hypothetical native-like model carrying the A31P substitution.

Consequently, MD simulations were performed for both mutants, and the derived trajectories were analyzed and compared in an attempt to evaluate the ability of molecular dynamics simulations to reproduce the experimental observations. The title, "The Repressor of Primer protein: to be native-like or not to be?" reflects our aspiration to investigate whether a native-like structural model can explain the experimentally observed behavior of the mutants while focusing on how it affects protein stability in each case.

2. Methods

2.1 Molecular Dynamics Simulation Systems: Details and Parameters

***Table 1**: Structural and simulation details characterizing the three trajectories*

Trajectory	Protein atoms	Box dimensions (nm^3)	Total atoms	Number of frames	Simulation time (μs)
native Rop	1820	6.12 x 6.12 x 6.12	23140	2000000000	20
D30P mutant	1824	6.17 x 6.17 x 6.17	23282	2000000000	20
A31P mutant	1828	6.25 x 6.25 x 6.25	24336	4000000000	40

The *protein.psf* file for each system provides information on the number of protein atoms and indicate that they were generated using VMD in a format compatible with NAMD/X-PLOR. At the beginning of the *gromacs.gro* file, information about solvation and the total number of atoms is provided, while at the end, the box water dimensions are shown, revealing spatial differences among the mutations and their solvation systems. Furthermore, information on the parameters of the MD simulations is provided through *run.mdp* and *md.log* files.

Table 2: Simulation parameters used for the three MD systems (derived from run.mdp & md.log files)

integrator	MD
Start time and time step in ps	0, 0.0050
Number of steps	4000000000
Thermostat type	Nose-Hoover
Thermostat-groups	Protein Water_and_ions
Reference temperature (K)	320
Thermostat time constant	1
Step size for minimization of flexible constraints	0
Cut-off scheme	Verlet: particle based cut-offs
Constraints	All-bonds (lincs algorithm)
Barostat type	C-rescale

Pressure coupling type	isotropic
Reference pressure	1
tau_p	5
software	GROMACS, version 2024.2

2.2 Software Used for Molecular Dynamics Trajectory Analysis and Structural Visualization

2.2.1 Trajectory Analysis Using Carma and Grcarma

Carma is a free, open-source software suitable for analyzing MD simulations [32]. It combines a user-friendly interface with low computational requirements. Developed in the C programming language, it operates through a command line and is designed to read and analyze trajectory data frame by frame, achieving reduced memory consumption. The latest version of Carma is capable of handling PDB, PSF and DCD files formats commonly used to store molecular dynamics trajectories. Notably, it relies on LAPACK libraries for the calculation of eigenvalues and eigenvectors, which are essential for principal component analysis (PCA). One of its most important features is the ability to eliminate overall rotational and translational motions associated with specific atoms or regions of a molecule. Furthermore, it can compute and visualize distance maps and average distance maps (along with their associated rms deviations) for chosen atoms and simulation frames. Carma can also determine the variance-covariance matrix, representing the fluctuations of each picked atom related to its mean position. Additionally, eigenvectors and eigenvalues can be calculated to support further structural analyses. Carma also provides visualization capabilities, making it a practical tool for processing and interpreting data from macromolecular systems [32].

Grcarma is a subsequent and automated version of Carma, developed for the analysis of MD simulations [33]. It functions as a graphical user interface (GUI) based on Carma, written in the perl language, and enables automated analyses through predefined parameters. The software can extract and cluster molecular structures using Dihedral or Cartesian PCA, and it performs and visualizes a wide range of analyses such as RMSD matrices, distance maps, radius of gyration, secondary structures, and entropy calculations. Moreover, it supports the computation of variance-covariance matrices, native-contacts fractions, and various geometric descriptors including bond lengths, angles, and torsions. Designed for all kinds of users, it provides an efficient environment that minimizes user intervention while preserving the computational precision of Carma [33].

2.2.2 XMGR and PyMOL: Plotting and Structural Visualization Tools

Xmgr (latest version: xmgrace) is a two-dimensional plotting program that allows the visualization of data derived from molecular dynamics simulations as well as from other sources. It was used for the representation of RMSF and RMSD data, cumulative plots and lastly for distance and angle distributions of atoms connected through hydrogen bond interactions [40].

PyMOL is an open-source graphics and visualization program developed by Schrödinger that enables the generation of high resolution 3D representations of biomolecules [41]. Due to its ability to produce high quality images, suitable for publications, it was used for the graphical illustration of all protein structures presented in this thesis.

2.3 MD Trajectory Analysis: RMSF, RMSD and PCA

In the context of the MD simulation analysis, each of the three trajectories (native Rop, D30P and A31P), was analyzed using Carma and Grcarma.

2.3.1 RMSF Analysis

The first analysis concerned RMSF (Root-Mean-Square-Fluctuations), which measures the deviation of each atomic position from its average position over time in MD simulations. It provides information about the flexibility of a molecular structure. High RMSF values indicate high mobility, while low values correspond to a more stable state [34].
To create RMSF plots, the initial step was to generate files containing only the information for the Ca atoms.

```
carma -v -w -fit protein.dcd protein.psf
```

Once these files were obtained, the following step was to generate the *carma.superposition.pdb* file, which contains information for multiple models. Its final column has the B-factor values required for RMSF calculation.

```
carma -v -w -col -cov -dot -super -atmid ALLID
CAs_fitted.dcd CAs.psf
```

To avoid superposition from multiple models, only data for model 1 were used, and the final column was extracted. The RMSF plot was then generated using the XMGR program, while minor additions made in GIMP.

2.3.2 RMSD Analysis

The following analysis for comparing the stability among the three systems is provided through the RMSD (Root-Mean-Square-Deviation).

The RMSD represents the average positional deviation of atoms recorded at a specific time point in an MD simulation to a reference structure. Most commonly, the reference structure corresponds to the initial frame of the trajectory or, alternatively, to the experimentally determined crystallographic model. The RMSD provides insight into the overall flexibility over time, as each snapshot recorded corresponds to a specific frame. Taking into consideration the consecutive evolution of the overall motion through frame-by-frame differences, it is possible to investigate how stable the structure is and whether progressive unfolding events are occurring. A continuous increase of RMSD values shows an unstable state that deviates from equilibrium. An RMSD value equal to zero indicates that the two compared conformations are identical, whereas higher RMSD values reflect increasing deviation from structural identity [34].

For a better understanding of the relationships among the three examined systems (native Rop, D30P & A31P), five separate analyses were performed:

1. Considering the total residue number across the two identical chains of the Rop protein
2. Considering all residues except the N & C terminal regions
3. Restricting only to the turn regions of the monomer A and B
4. Focusing exclusively on the turn region of monomer A
5. Focusing exclusively on the turn region of monomer B

To avoid unnecessary information, we used files containing only the Ca atoms. Based on this, the file *carma.fit-rms.dat,* which provides the RMSD values of each frame in the second column, was generated using the following command:

```
carma -v -w -fit -segid A -segid B CAs.dcd Cas.psf
```

To generate the RMSD plot, only the first two columns of the *carma.fit-rms.dat* are required. Using the following command:

```
awk '{print $1/50000, $2}' carma.fit-rms.dat
```

we extracted the corresponding columns, dividing the first one by 50,000 for better numerical convenience on the X axis of the diagram. Notably, the boundaries selection between distinct regions (e.g. turn regions) was based on the RMSF plot. Additionally, the discrepancy in the number of frames for the A31P trajectory is due to the A31P simulation lasting twice the duration of the other simulations.

For the visualization of the frames with the highest RMSD values of the A31P mutant, the *carma.fit-rms.dat* file was used, which contains the RMSD data. They are sorted in ascending order using the command:

```
sort -n -k2
```

Thus, the four frames with the highest RMSD values were selected. However, to avoid representing the same time point, snapshots corresponding to different moments were chosen. For each of the selected frames, the following commands was executed:

```
carma -v -atmid ALLID -pdb -first 1235265 -last 1235265
protein.pdb protein.psf

carma -v -atmid ALLID -pdb -first 1828220 -last 1828220
protein.pdb protein.psf

carma -v -atmid ALLID -pdb -first 3072154 -last 3072154
protein.pdb protein.psf

carma -v -atmid ALLID -pdb -first 3489146 -last 3489146
protein.pdb protein.psf
```

2.3.2.1 Statistical Analysis Using the sn Package in R

In order to enhance our investigation into the wide range of RMSD values, we performed statistical analysis using the **R library "sn",** which provides tools for analyzing and modeling skewed distributions. The following R code was used to generate the RMSD distribution diagrams:

```
#R
> library(sn)
Loading required package: stats4
Package 'sn', 2.1.1 (2023-04-04).
Type 'help(SN)' and 'help("overview-sn")' for basic
information.
The package redefines function 'sd' but its usual working is
unchanged.

Attaching package: ‘sn’

The following object is masked from ‘package:stats’:

    sd
>testdata <- read.table("filename")

>summary( testdata )


V1
 Min.   :0.0000
 1st Qu.:0.8023
 Median :0.9121
 Mean   :0.9774
 3rd Qu.:1.0967
 Max.   :2.8785
>plot(density(testdata$V1))

>mod  <-  selm(V1  ~  1,  data=testdata)              #skewed
distribution

>summary(mod)

Call: selm(formula = V1 ~ 1, data = testdata)
Number of observations: 4000002
Family: SN
Estimation method: MLE
Log-likelihood: 586076.31
Parameter type: CP
```

```
CP residuals:
    Min      1Q  Median      3Q     Max
-0.9829 -0.1806 -0.0708  0.1138  1.8956

Regression coefficients
      estimate   std.err   z-ratio Pr{>|z|}
mean 9.829e-01 1.116e-04 8.806e+03        0

Parameters of the SEC random component
       estimate std.err
s.d.     0.2287       0
gamma1   0.8848       0
>plot(mod)

Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
```

Notably, for fitting the skewness distribution the following command was used, while for the non-skewness fitting, the second one was applied:

```
mod <- selm(V1 ~ 1, data=testdata)
```

```
mod <- selm(V1 ~ 1, data=testdata, fixed.param = list(alpha=0) )
```

Also, to ensure that the Q-Q plots were displayed in the same scale for proper comparison, an alteration within the selm() function was applied, in which the Y-axis limits were manually defined by the user.

2.3.3 Principal Component Analysis

In order to further investigate the structural folding differences between the native conformation and the two specific mutants, we also performed Principal Component Analysis (PCA). PCA is a linear algebraic technique used to detect patterns in data by comparing their similarities and differences. The initial dataset is transformed into a new coordinate system, in which the new variables are linear combinations of the original ones, maintaining the maximum amount of information while reducing dimensionality.

MD simulations generate a large amount of data, as the motion of each atom is described in three dimensions (x,y,z). This technique identifies the principal directions (eigenvectors) and ranking them according to their magnitude (eigenvalues). Focusing on the principal components with the largest eigenvalues preserve low-frequency (high amplitude) motions, while components with the smallest eigenvalues are rejected, as they are associated with noise or local fluctuations [35] [36].

There is also an advanced variant of PCA called Dihedral Principal Component Analysis (dPCA), which takes into account the information from the dihedral angles (φ, ψ) of the protein backbone. Other internal coordinates, such as bond lengths and bond angles, do not undergo significant changes. To avoid issues related to the circularity of dihedral angles, trigonometric functions (sin(φ) & cos(φ)) are used to transform the data into a linear coordinate space. The main advantage of dPCA lies in its ability to provide a clear separation between internal and overall motions, in contrast to Cartesian PCA, in which these two kinds of motion often get mixed up.

dPCA is frequently used for the construction of free energy landscapes of molecules that undergo large structural rearrangements, such as those

occurring in the protein folding process. For a simplified representation of an energy landscape to be considered accurate, it must at least properly generate the number of stable and semi-stable states of the system, including their energies and positions within the landscape. Unfortunately, many of these parameters often get lost by projecting the system into lower dimensional landscapes. However, dPCA allows an accurate representation of the necessary information in lower dimensions by including more principal components in the analysis. Thus, it avoids issues related to empirical parameters such as the number of native contacts that may lead to artifacts and oversimplifications of the free energy landscape. A particularity remarkable feature of this method is its capability of detecting perfectly correlated movements between two atoms that oscillate in the same direction but with a phase difference of 90°. This is possible because dPCA is a non-linear method, based on the dihedral angles and performing a non-linear transformation of the data, revealing associations that would be impossible to detect using a linear approach [37][38][39].

2.3.3.1 Dihedral PCA of the Turn Regions of Both Monomers

To generate two files per analysis, one containing the eigenvalues (*carma.dPCA.eigenvalues.dat*), and a second one containing the eigenvectors (*carma.dPCA.eigenvectors.dat*) as well as the illustration plots of the dominant principal components (PC1, PC2, PC3) the following command was used:

```
carma -v -w -col -3d -segid X -dPCA 5 3 298 protein.dcd
protein.psf
```

For the visualization of the representative structure from each frame of the three trajectories, the *dPCA.representative.cluster.pdb* was used. Additionally, the *dPCA.5-D.superposition.cluster_(...).pdb* file contains all conformations belonging to the same cluster, as defined in the five-dimensional (5D) dPCA analysis. Through this file, it is possible to visualize the internal flexibility that exists within each cluster.

To enable direct comparison among the native, D30P and A31P proteins, it is essential to set a similar color scale, based on the RMSF values of each structure.

Table 3: *Range of RMSF values for turn A and turn B regions derived from the dPCA.superposition.cluster_01.pdb file for the native Rop and the D30P and A31P mutants*

	Turn A	**Turn B**
native	0.25 – 0.50	0.24 – 0.49
D30P	0.24 – 0.57	0.24 – 0.56
A31P	0.31 – 0.84	0.42 – 1.02

In order to ensure that the color coding is common across all systems, the minimum value was set to 0.24, while the maximum to 1.02 (*spectrum b, minimum = 0.24, maximum = 1.02*).

The maximum RMSF value was extracted through the following command:

```
awk '{ print $11 }' dPCA.superposition.cluster_01.pdb |
sort -n | uniq | tail -1
```

while the minimum value was derived using the following one:

```
awk '{ print $11 }' dPCA.superposition.cluster_01.pdb |
sort -n -r | uniq | tail -2
```

Finally, in both dPCA analyses of the D30P and A31P mutants, we additionally performed a 5D clustering analysis to further investigate the distinguishable conformational states within each ensemble. In both cases,

it is essential to compare the 2D variation distribution plots obtained from the 3D analysis with the corresponding plots derived from the 5D analysis, aiming to assess whether the higher-dimensional analysis confirm the previously identified clusters.

For each turn region of D30P and A31P mutant separately, the *plot* program was used with the following commands:

```
plot -k342 < dPCA.5-D.clusters.dat      #5D clustering
plot -k342 < dPCA.clusters.dat          #3D clustering
```

2.3.3.2 Cartesian PCA of the Entire Trajectories

Following the Dihedral PCA, performed separately for Turn A and Turn B regions, the dominant cluster from each analysis was selected for subsequent whole-structure conformational examination. Specifically, the file pairs *dPCA.5-D.fitted.cluster_01.dcd* and *dPCA.5-D.fitted.cluster_01.psf* were used as input for the Fitting analysis.

For each case (Turn A & Turn B based dPCA analysis) the Grcarma tool used to fit only the steady part of each structure (including all amino acids from both monomers, excluding the C-terminals and turn regions).

Subsequently, fitted files, named *carma.fitted.dcd* and *carma.selected_atoms.psf,* were generated and used as input for the Cartesian PCA analysis. The resulting superposed cPCA PDBs depict the entire structure (excluding the tails) of all conformations belonging to the dominant cluster of each trajectory, allowing evaluation of the overall structural flexibility associated with the dominant state of either Turn A or Turn B.

***Table 4**: Range of RMSF values for the entire trajectories (excluding C-terminal residues) derived from the cPCA.superposition.cluster_01.pdb file for the native Rop and the D30P and A31P mutants*

	Total trajectory	
	Turn A	**Turn B**
native	0.35 – 0.99	0.32 – 1.01
D30P	0.38 – 1.27	0.33 – 1.25
A31P	0.39 – 1.88	0.40 – 1.44

For representation on a common color coding scale, the overall lowest and highest RMSF values were used as the limits of the common scale (*spectrum b, minimum = 0.32, maximum = 1.88*).

2.4 Distance and Angle Plots

In order to investigate whether two atoms are connected through hydrogen-bond interaction, it is necessary to calculate two parameters:
i) the distance between the acceptor and donor atoms and
ii) the angle formed by the donor – hydrogen – acceptor atoms (D–H···A).
For these separate calculations, Carma was used using the following commands:
i) `carma -v -dist atom1 atom2 -atmid ALLID protein.dcd protein.psf`
ii) `carma -v -bend atom1 atom2 atom3 -atmid ALLID protein.dcd protein.psf`

2.5 Ramachandran Plots

To generate Ramachandran plots, the mutant residues as well as the two adjacent residues on either side were included. For each residue, a Ramachandran plot was made including all possible pairs of φ and ψ angles from native, D30P and A31P, aiming to compare the differences according the dihedral angles between the mutants.

The Grcarma program was used to produce the *phi_psi_dihedral_segid(A/B).dat* file containing calculated dihedral angles. Then, the relevant values were isolated using the following command:

```
awk '$1 == 29 || $1 == 30 || $1 == 31 || $1 == 32' phi_psi_dihedral_segid(A/B).dat > resi_29-32.dat
```

Subsequently, a plot (using xmgr) was created for each of the selected residues. A two-column format file was necessary to create the plot, in which the φ and ψ values were written. For this, the following command was used for each residue separately:

```
awk '$1 == residue_number {for (i = 2; i < NF; i += 2) print $i, $(i+1)}' resi_29-32.dat > output.dat
```

3. Results

Experimental studies have shown, based on both the melting temperature values and the already obtained structures, that the native protein exhibits the most stable and compact structure, with a Tm value of 68.7°C [27]. The D30P mutant adopts a native-like but less stable structure, as indicated by its Tm value of 58.9°C [27]. Regarding the A31P variant, while the experimental data reveal a topological rearrangement into the bisecting U structure, which is evidently less compact and displays a Tm value of 43°C [30], molecular dynamics simulations were performed using a hypothetical model of A31P, carrying the point mutation at position 31. Therefore, a direct computational comparison of the three systems takes place in order to examine whether molecular dynamics simulations can properly reproduce these experimental observations, while also investigating the stability of the hypothetical native-like A31P model.

3.1 RMSF Analysis

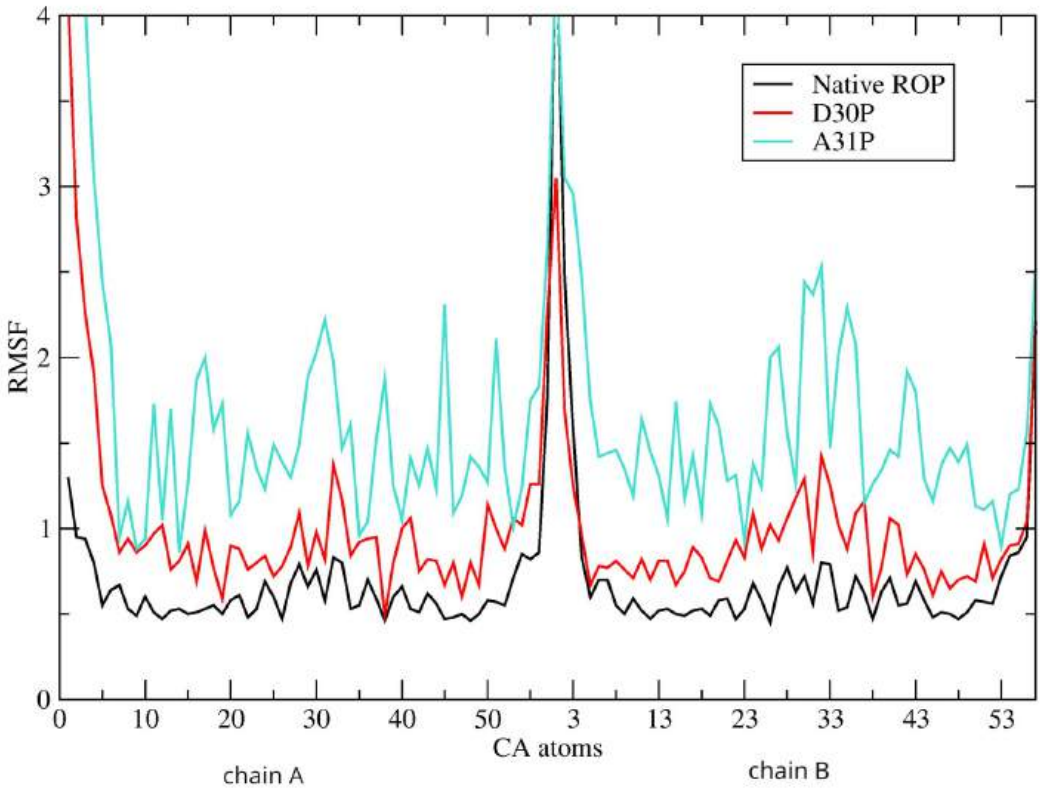

***Figure 5**: Representation of RMSF values of Cα atoms for the three studied systems across the two monomers (chain A & chain B). Produced with xmgr.*

Figure 5 illustrates the variation of RMSF values for each Cα atom in the three studied protein structures: native Rop, D30P and A31P. Each of these proteins is a homo-dimer, and the two polypeptide chains can be clearly distinguished in the plot. Lower RMSF values are observed in the regions corresponding to the α-helices, as expected, since they are stable secondary structures characterized by reduced flexibility. The loop region acting as a bridge between the two constant helices exhibits higher RMSF values, indicating increased mobility in this segment.

A similar fluctuation pattern is observed for all three structures. This observation is consistent with the fact that the D30P mutant adopts a

native-like, but still less compact, conformation than the native protein, since the increased RMSF values are in agreement with the experimental data reflecting slight destabilization.

In contrast, the A31P mutant shows significantly higher RMSF values, indicating greater flexibility and reduced compactness compared to the native conformation. From the RMSF distribution pattern, it can be concluded that the topology adopted by the A31P mutant retains its molten globule characteristics, as it has a correspondence distribution to the native Rop structure. Notably, the exceptionally high RMSF values observed at both the N and C terminals of each monomer are due to their flexibility, which is not related to the examined mutations. Therefore, these regions are not included in the interpretation of the results.

3.2 RMSD Analysis

RMSD plots illustrate the variation in RMSD values for the three structures (native Rop, D30P and A31P) during the MD simulations. Depicting all trajectories in the same plot is extremely useful for comparison, as it graphically represents the deviation of each molecule from its reference structure in a comparable manner. Each MD simulation examines the progressive folding of each molecule, from an initial conformation, called reference structure, to a final one. In the meantime, the evolution of MD is monitored by receiving snapshots. Since a structure remains relatively stable during the simulation, its deviation from the reference structure is small, resulting in low RMSD values.

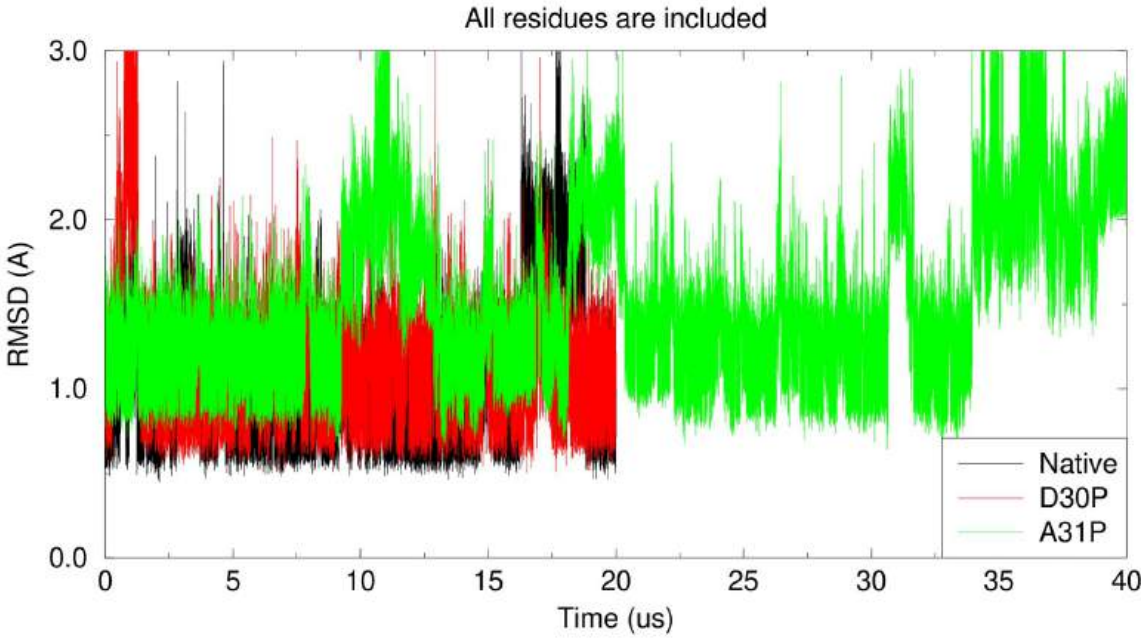

***Figure 6**: Representation of RMSD values of native Rop, D30P and A31P systems over the MD simulation. All residues are included . The plot was produced using the xmgr tool.*

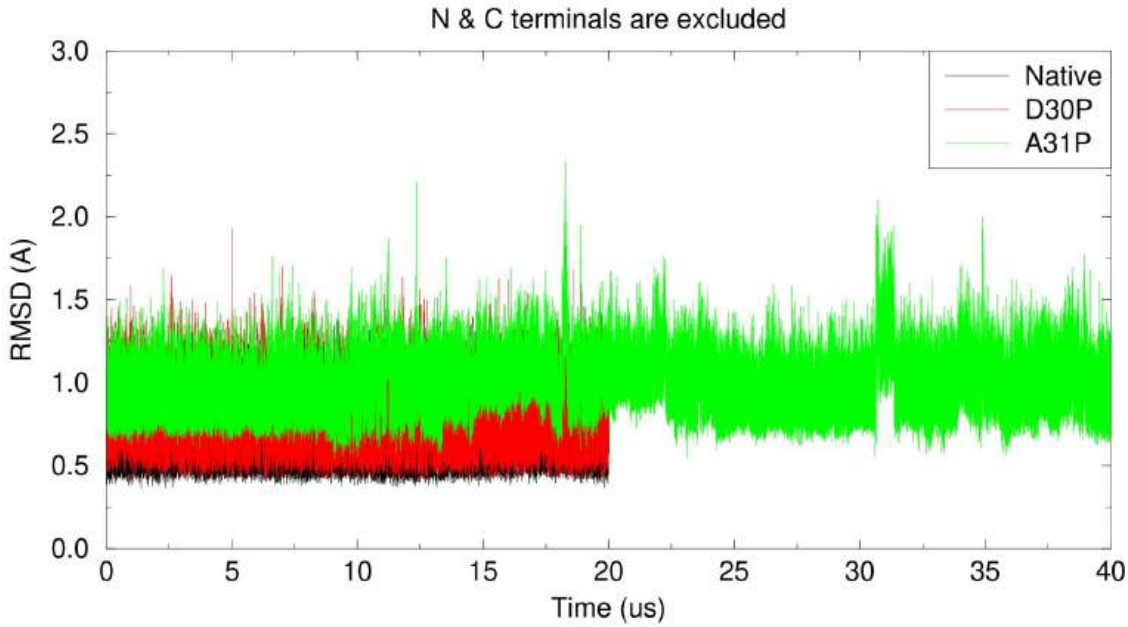

Figure 7: *Representation of RMSD values of native Rop, D30P and A31P systems over the MD simulation. All residues are included, except the N (1-5) and C (55-57) terminals . The plot was produced using the xmgr tool.*

Figure 6 presents the RMSD including the N and C terminal regions of both monomers of each protein. The terminals are flexible as they do not participate in a secondary element, leading to increased mobility in solution. As a result, their positions change in each snapshot, which is reflected in increased RMSD values, giving the impression that the overall structure deviates from the reference one. Including the tails can, therefore, be misleading, since at specific time points RMSD values exceed 3Å. In contrast, **Figure 7** represents the same analysis excluding the tails, in order to focus exclusively on the structural stability of the native Rop and the two mutants, without the influence of terminal flexibility.

The average RMSD value of the native structure, including the terminals, is 1.07Å, while, in their absence, it decreases to 0.69Å. For the D30P mutant, the same value decreases from 1.12Å (with tails) to 0.78Å (without tails), while for the A31P mutant the corresponding reduction is from 1.48Å to 0.96Å. This indicates that the terminal residues

significantly affect the analysis, making it difficult to clearly distinguish the structural variation of each system. Therefore, it can be revealed that the native Rop exhibits the highest stability, since it is characterized by the lowest RMSD values, indicating that each frame deviates only a little from the reference structure over time. The D30P mutant possesses the second place, showing marginally higher RMSD values, confirming the already known native-like conformation. Finally, the A31P mutant displays the highest RMSD values. The partial unfolding caused by this mutation leads to instability, as confirmed by the significant variations in RMSD values. Notably, due to its high mobility, the duration of its MD simulation was twice as long.

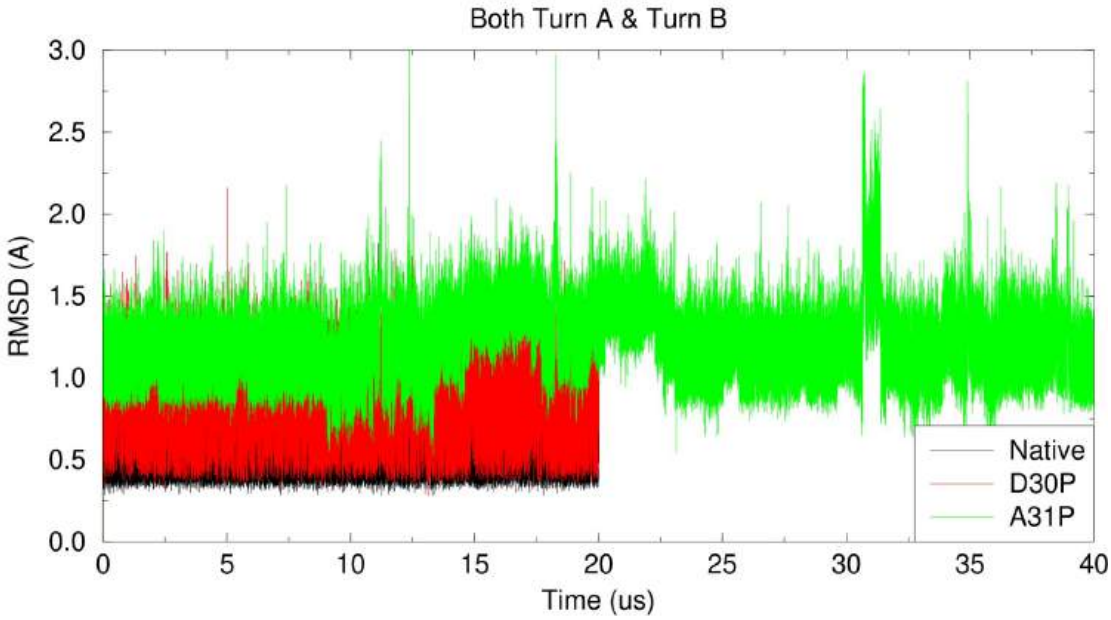

Figure 8: *Representation of RMSD values of native Rop, D30P and A31P systems over the MD simulation. Only the residues belonging to the Turn A (24-39) and Turn B (24-39) regions are included. The plot was produced using the xmgr tool.*

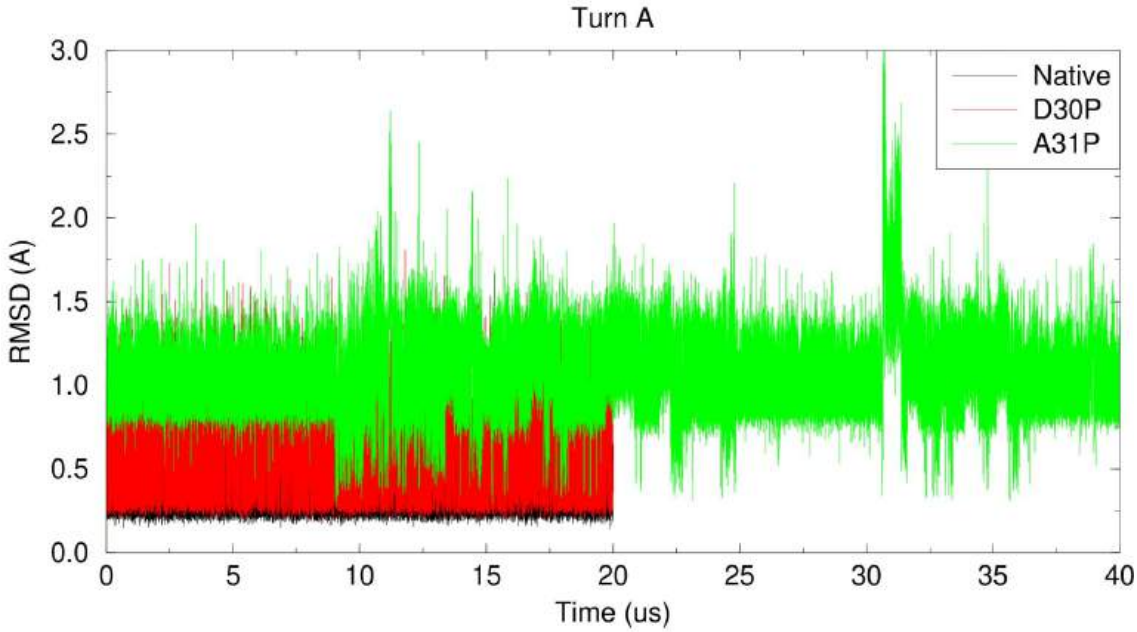

Figure 9: *Representation of RMSD values of native Rop, D30P and A31P systems over the MD simulation. Only the residues belonging to the Turn A (24-39) region are included. The plot was produced using the xmgr tool.*

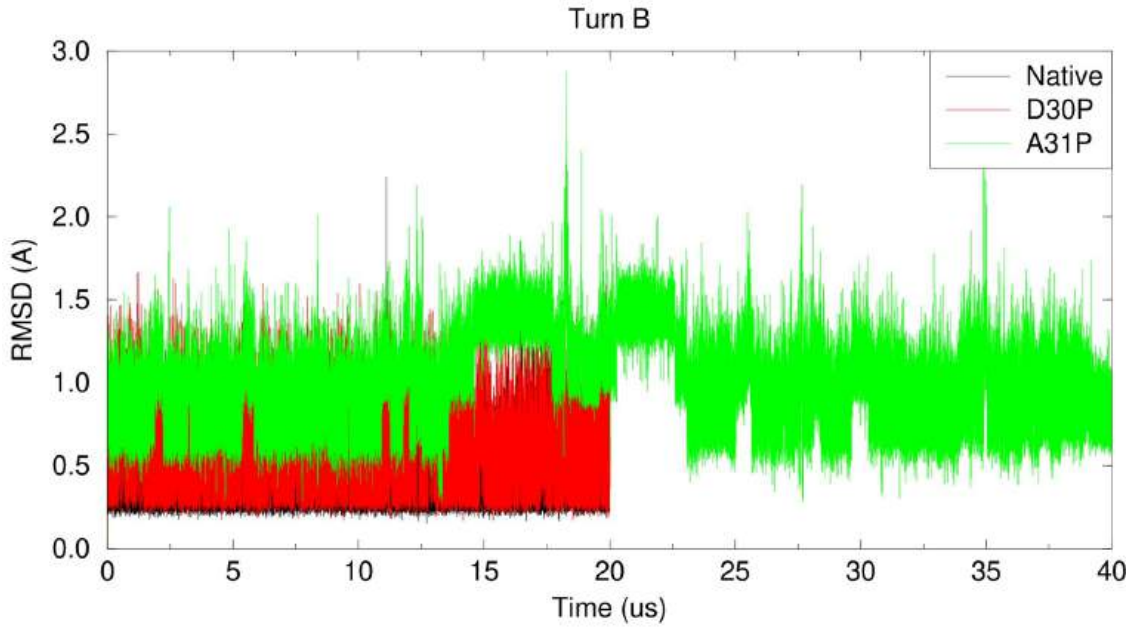

Figure 10: *Representation of RMSD values of native Rop, D30P and A31P systems over the MD simulation. Only the residues belonging to the Turn B (24-39) region are included. The plot was produced using the xmgr tool.*

Figure 8 represents only the residues belonging to the turn regions of both monomers, in order to focus on the area that has the highest flexibility. The point mutations are located within turn region, therefore, this isolation plays an important role in studying how these mutations influence the mobility of the surrounding residues.

It is important to note that the selection of the boundary residues of this region was derived from the RMSF plot (**Figure 5**), which indicates the deviation of each Cα atom from its reference position. Thus, we already know that the regions exhibiting increased RMSF values correspond to the turn regions, as they are more flexible and undergo local rearrangements. For additional clarity, data from study [28] were also used to finalize the determination of these regions.

For the native Rop, the average RMSD value of the turn region is 0.63Å, while for the D30P mutant this value increases to 0.80Å. Comparison of these values with those presented in **Figure 7** indicates that the turn regions are relatively stable, since no substantial increase in RMSD values is observed. However, for the A31P mutant, the average deviation is 1.19Å, compared to 0.96Å in **Figure 7**. The observed increase in RMSD suggests that the mutation in the turn region has altered the structural conformation, resulting in enhanced overall mobility.

Notably, when the two monomers of A31P are analyzed separately, the average RMSD value for Turn A is 1.07Å, whereas for Turn B it is 0.98Å. The small deviations between the two monomers are possibly due to local solvent exposure, changes in hydrogen bonding patterns, or ionic interactions. The difference between the native and D30P systems is negligible (0.02Å). Similarly to the overall structure analysis, the hierarchical stability ranking among the three studied systems remains the same.

3.2.1 Frames Exhibiting the Highest RMSD Values in the Turn Region

Based on **Figure 8**, the four frames of the A31P mutant, displaying the highest RMSD values, were selected for structural visualization using the PyMOL program.

1st conformation
Frame: 1235265 RMSD: 3.0794

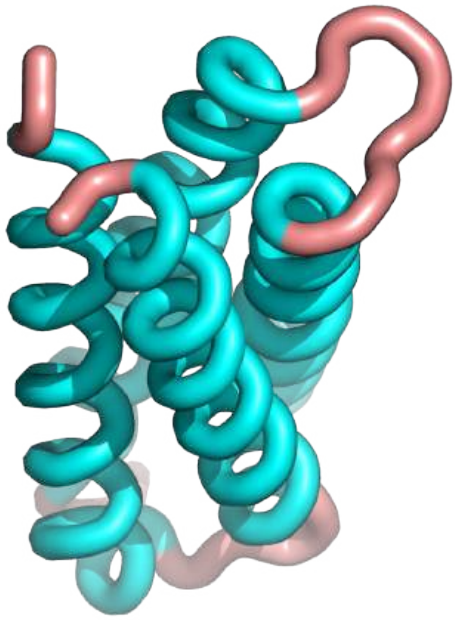

Figure 11: *Structural representation of A31P mutant (frame: 1235265, RMSD: 3.0794Å) using PyMOL. Helices are shown in cyan and loops in pink.*

2nd conformation
Frame: 1828220 RMSD: 2.9763

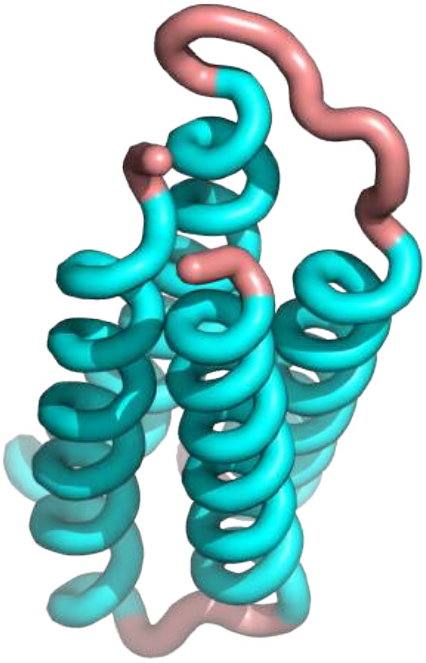

Figure 12: *Structural representation of A31P mutant (frame: 1828220, RMSD: 2.9763Å) using PyMOL. Helices are shown in cyan and loops in pink.*

3rd conformation
Frame: 3072154 RMSD: 2.8653

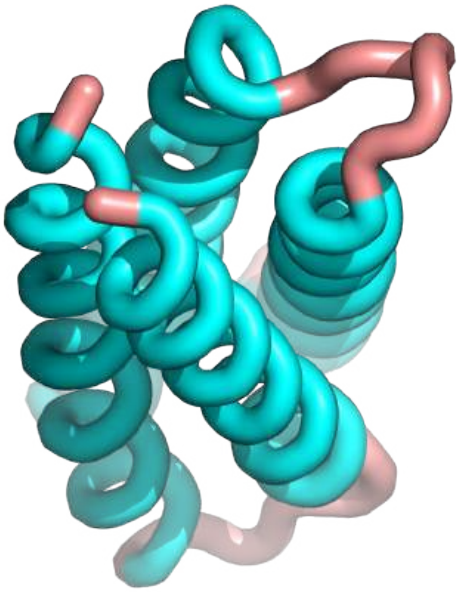

Figure 13: *Structural representation of A31P mutant (frame: 3072154, RMSD: 2.8653Å) using PyMOL. Helices are shown in cyan and loops in pink.*

4th conformation
Frame: 3489146 RMSD: 2.8154

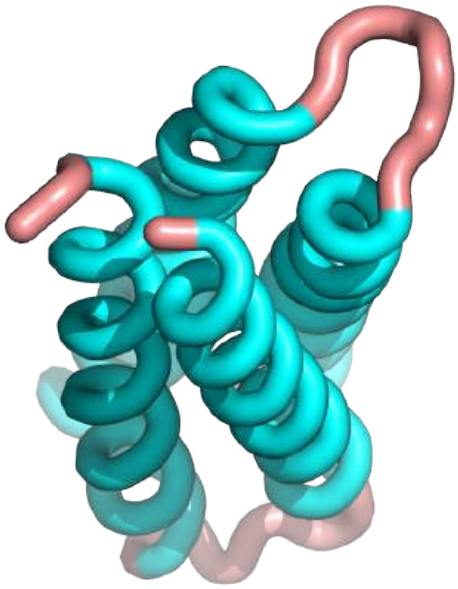

Figure 14: *Structural representation of A31P mutant (frame: 3489146, RMSD: 2.8154Å) using PyMOL. Helices are shown in cyan and loops in pink.*

3.2.2 RMSD Histograms

After obtaining the RMSD plots, the following step is to generate histograms to visualize the distribution of RMSD values. A histogram representation provides insight into the amplitude of each trajectory, since a narrow amplitude indicates a stable structure with a steady behavior, whereas a broad amplitude reflects the presence of multiple conformations which is compatible with a more flexible structure.

Additionally, the histogram peak shows the most frequent RMSD value which reflects how much the dominant conformation deviates from the reference structure. In this way, histograms allow a direct comparison of the distribution of frames according to their RMSD values, as it can be suggested which mutation leads in a more substantial deviation, which variant adopts a structure closer to the native state, and how each mutation influences the overall morphological stability.

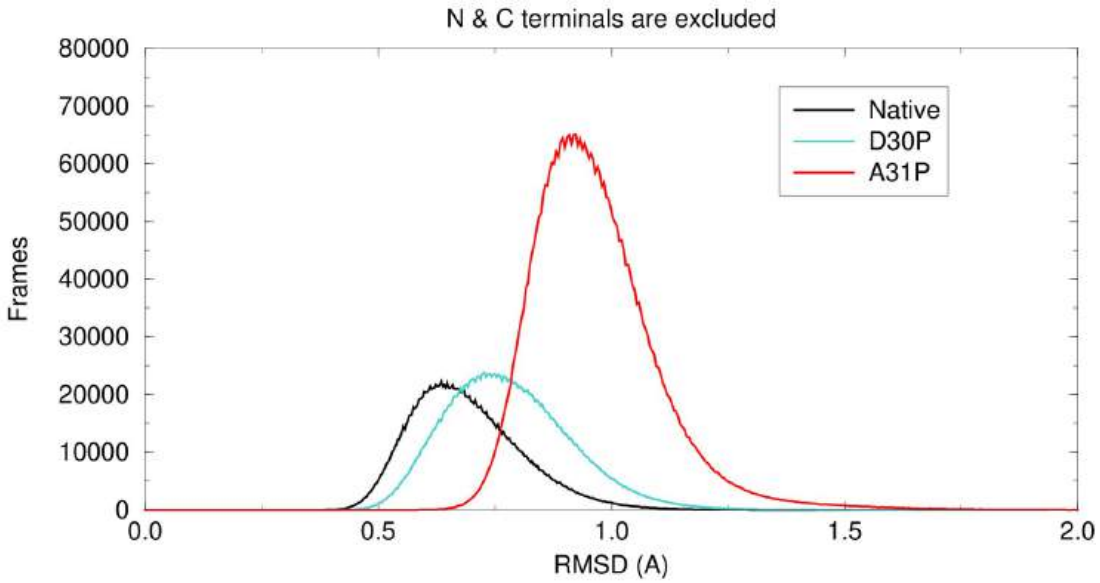

***Figure 15**: Histogram RMSD for the native Rop, D30P and A31P. All residues are included, except the N (1-5) and C (55-57) terminals.*

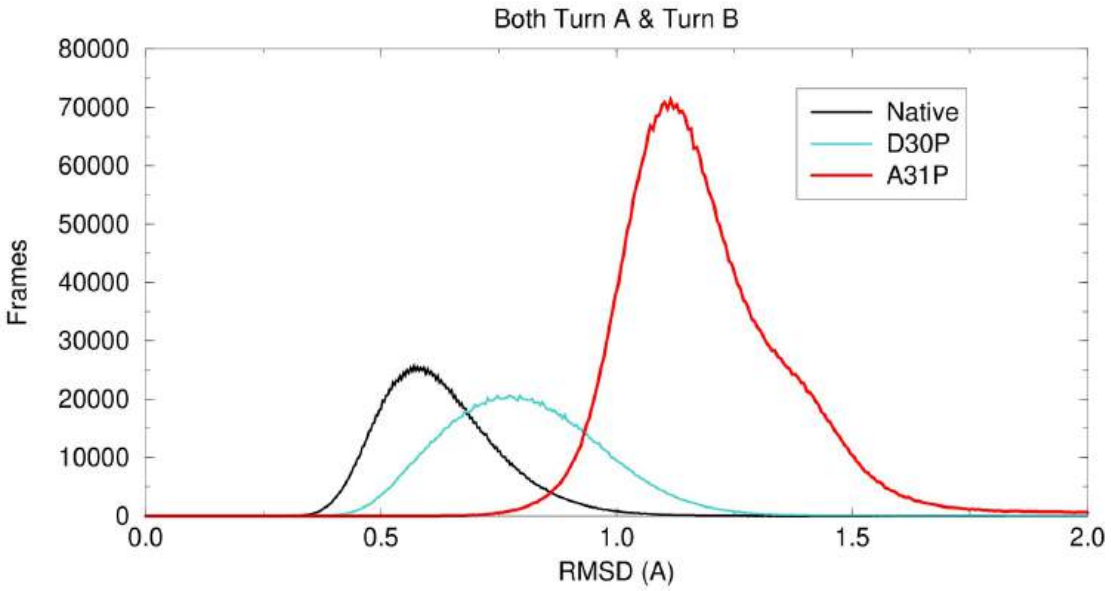

***Figure 16**: Histogram RMSD for the native Rop, D30P and A31P systems. Only the residues of the Turn regions (A & B) (24-39) are represented.*

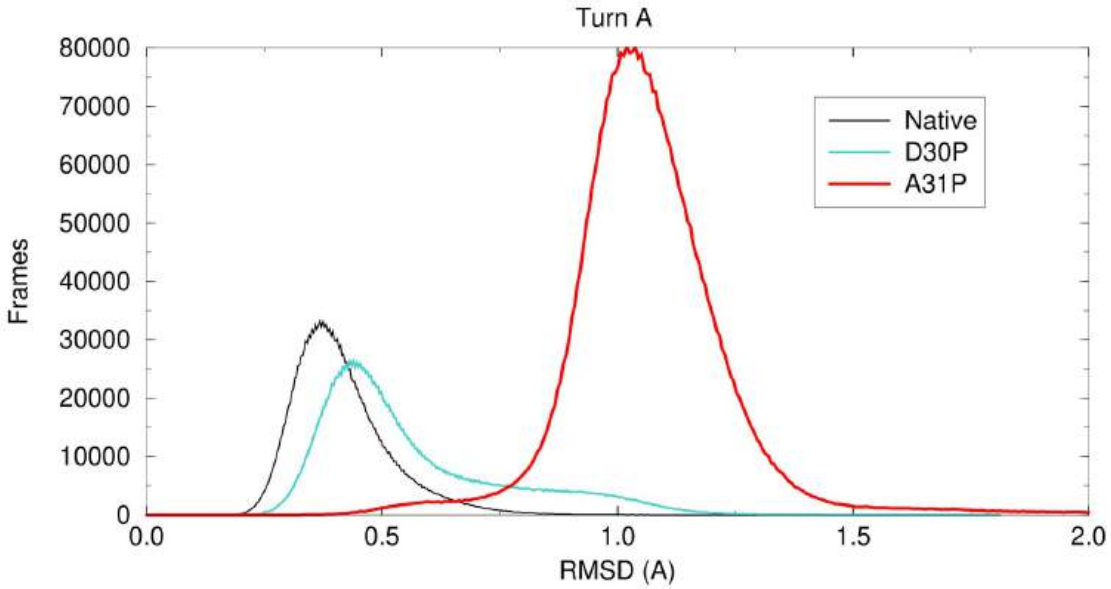

***Figure 17**: Histogram RMSD for the native Rop, D30P and A31P systems. Only the residues of the Turn A are represented.*

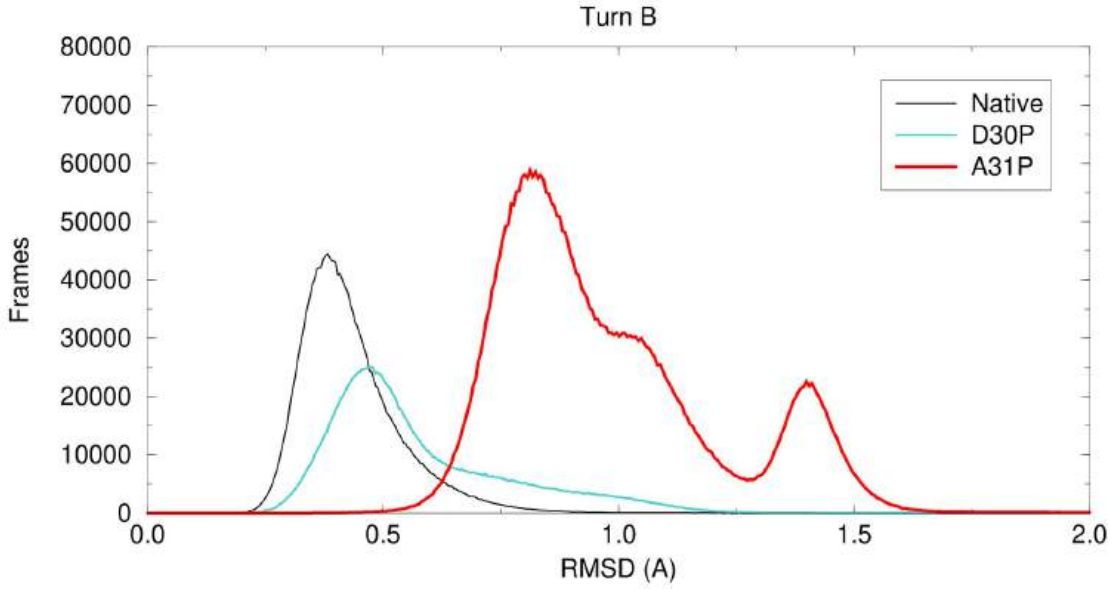

***Figure 18**: Histogram RMSD for the native Rop, D30P and A31P systems. Only the residues of the Turn B are represented.*

Figures 15-18 illustrate the frame distribution for each trajectory (native Rop, D30P and A31P), as it is obtained from the MD simulation. For the native Rop and the D30P mutant, a direct comparison is possible, as both simulations have an equal number of frames (2000001). It is evident that the RMSD range in the native Rop is limited, indicating the relative stability of its overall structure. The D30P curve exhibits a slightly broaden distribution and is marginally shifted to the right, corroborating the native-like character of this mutant, as the deviation from the native structure remains minimal. This is consistent with the experimental results as the thermodynamic stability of D30P is diminished relative to the native form, reflecting by its lower Tm value. In contrast, the comparison with the A31P is indirect because its trajectory contains twice as many frames (4000001). Although the area under the curve corresponds to this increased number of frames, the position and overall features of the distribution remain unaltered, meaning that the mutant retain its molten-globule characteristics. In light

of this, the A31P curve is markedly shifted to the right, confirming that this variant displays higher instability, with RMSD values reaching up to 2.9Å (**Figure 18**). The broaden distribution indicates that, during the MD simulation, the system adopts multiple conformations. The presented illustrations are limited to 2Å to maintain the dynamic range of the plot.

According to accurately compare the three different systems in the same plot, two scripts (1 & 2) were created. **Script 1** normalized all distributions to a maximum of 100, while **script 2** addressed the issue of the different number of frames between the A31P trajectory and the native one. The normalized histograms represented in **Figures 19-22**.

Overall, observations deriving from these plots appear to be in agreement with the experimental data. The narrow RMSD distribution of the native protein represents the most compact and stable system characterized by a high Tm value. At the same time, the D30P mutant exhibits reduced stability consistent with its lower Tm, while still retaining its native-like conformation. The computational model of A31P reproduces the increased instability of the mutant, exhibiting a broad range of accessible conformations. Even if the simulation is based on a native-like conformation of the A31P, the thermodynamic instability still exists reflecting the destabilization caused by the substitution of proline at position 31.

These results are also in agreement with previous computational studies which demonstrated that the hypothetical native-like A31P structure is unstable showing a tendency toward unfolded events in the turn region [28].

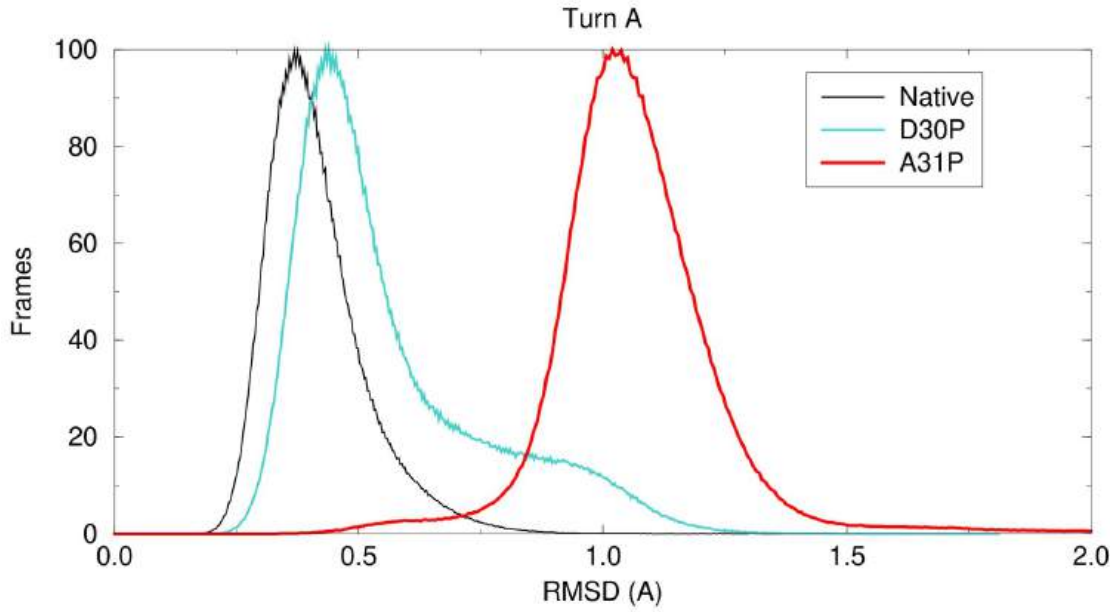

***Figure 19**: Histogram RMSD for the native Rop, D30P and A31P. Only the residues of the Turn A are shown. The distributions have been normalized to a maximum of 100 to allow comparison (script 1).*

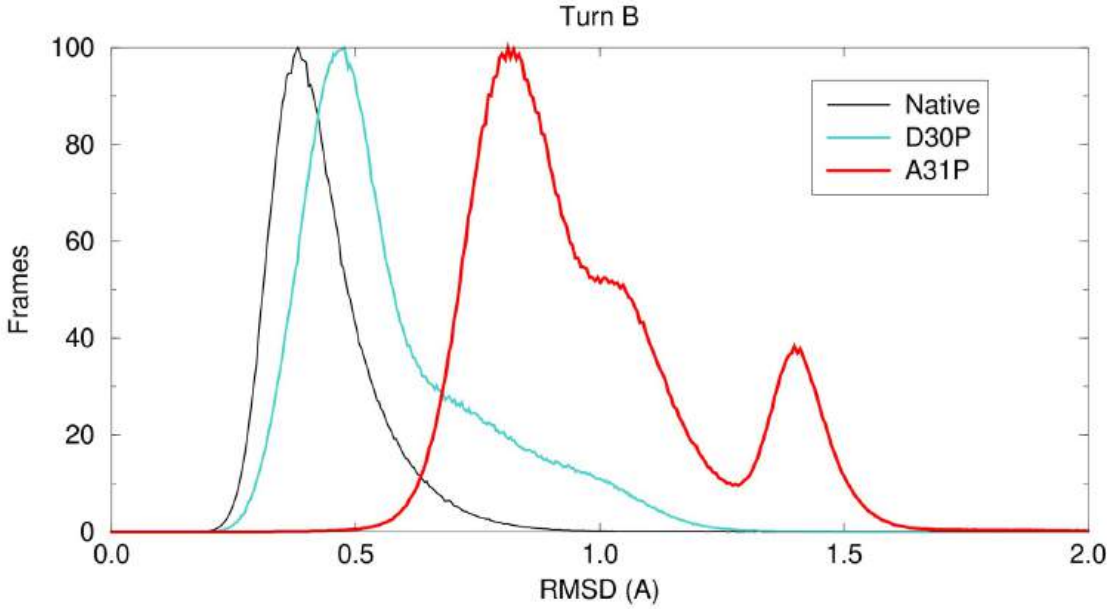

***Figure 20**: Histogram RMSD for the native Rop, D30P and A31P. Only the residues of the Turn B are shown. The distributions have been normalized to a maximum of 100 to allow comparison (script 1).*

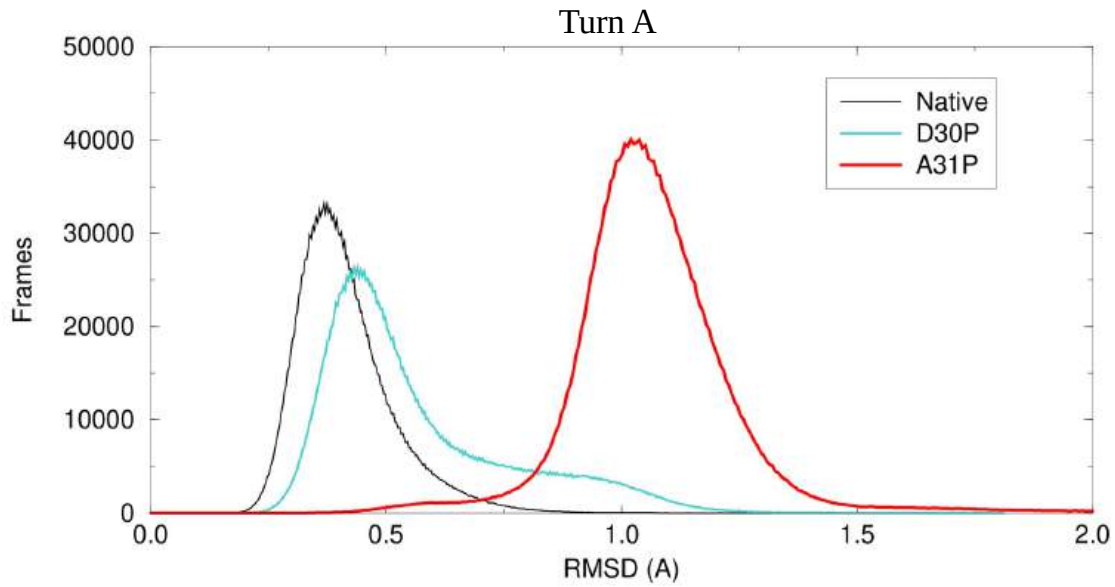

***Figure 21**: Histogram RMSD for the native Rop, D30P and A31P. Only the residues of the Turn A are shown. The distribution of A31P has been normalized to match the number of frames of the compared proteins (script 2).*

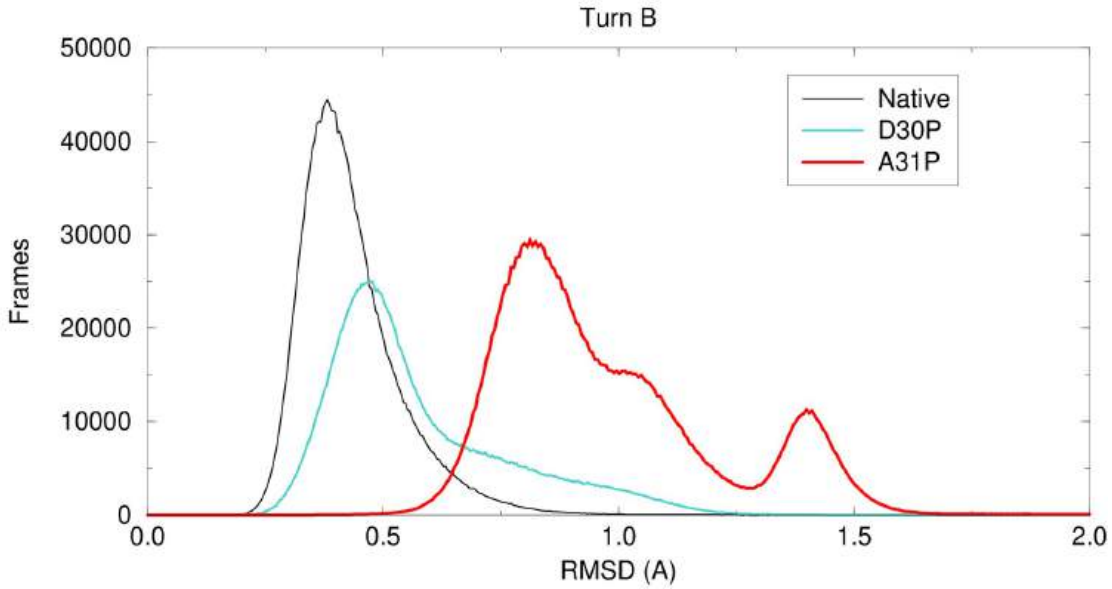

***Figure 22**: Histogram RMSD for the native Rop, D30P and A31P. Only the residues of the Turn B are shown. The distribution of A31P has been normalized to match the number of frames of the compared proteins (script 2).*

3.2.3 Statistical Analysis Based on R Package "sn"

In order to further investigate the obtained results, a series of statistical parameters were calculated to highlight the differences among the systems. For each of the common RMSD histograms, a histogram was generated to emphasis the probability density of the observed variables, illustrating each trajectory separately.

1. *All residues (excluded N & C tails) (Figure 15):*

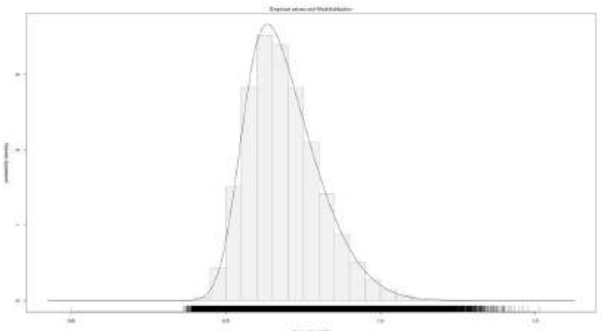

Figure 23*: native Rop*

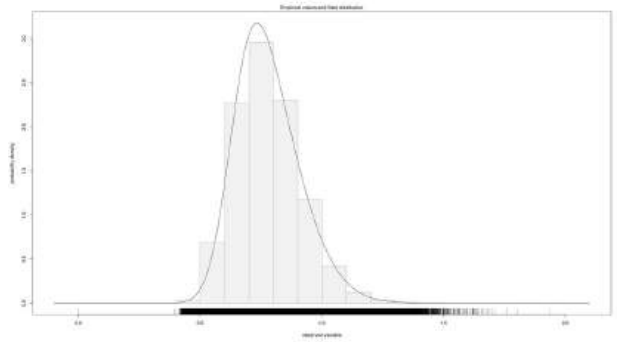

Figure 24*: D30P mutant*

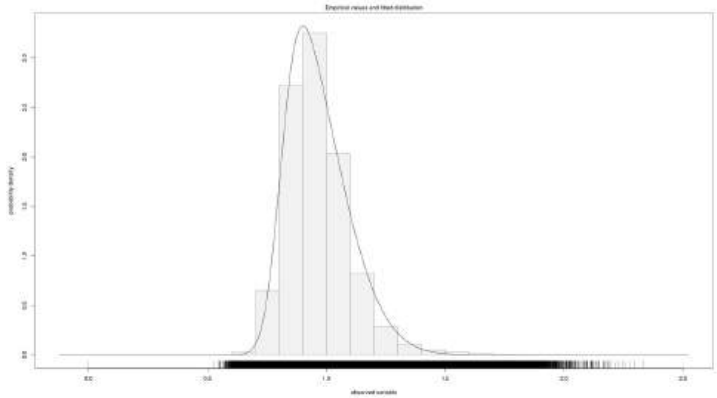

Figure 25*: A31P mutant*

***Table 5**: Statistical parameters describing the flexibility distributions of the overall structures of native Rop, D30P and A31P variants.*

	Native	**D30P**	**A31P**
Mean	0.69	0.78	0.97
s.d.	0.12	0.13	0.13
Gamma1	0.7	0.56	0.73
Log-likelihood (skewed)	1527556.19	1256279.06	2666760.23
LLG/observation	0.76	0.63	0.67
Log-likelihood (non-skewed)	144005.14	1206433.68	2368609.14
LLG/observation	0.07	0.6	0.59

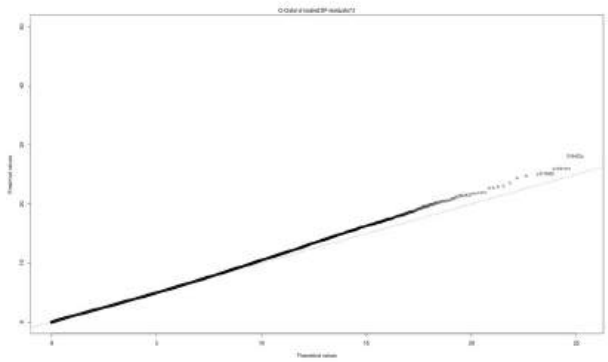

***Figure 26:** Skewed distribution of the native Rop*

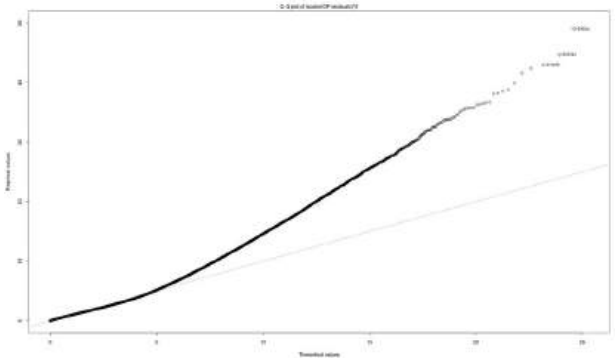

***Figure 27:** Non-skewed distribution of the native Rop*

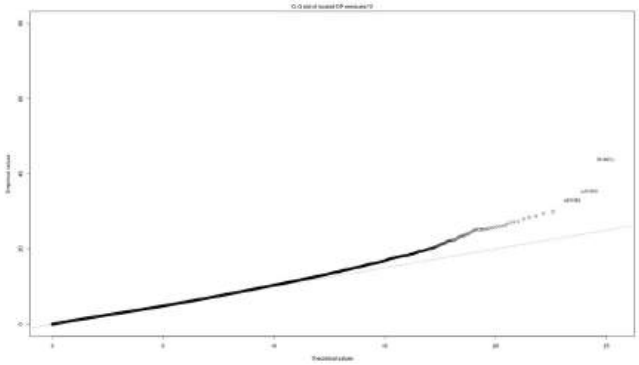

***Figure 28:** Skewed distribution of the D30P*

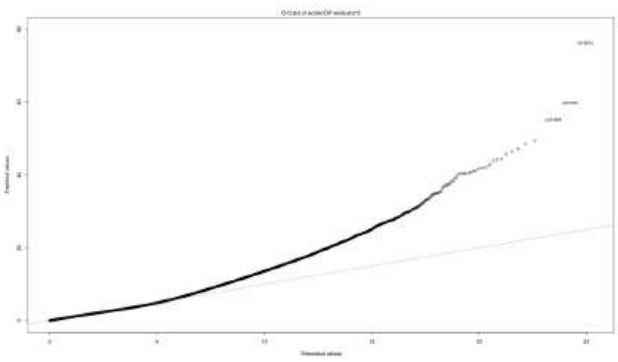

***Figure 29:** Non-skewed distribution of the D30P*

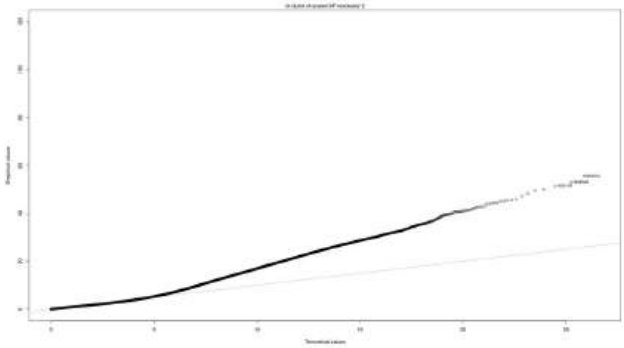

Figure 30: *Skewed distribution of the A31P*

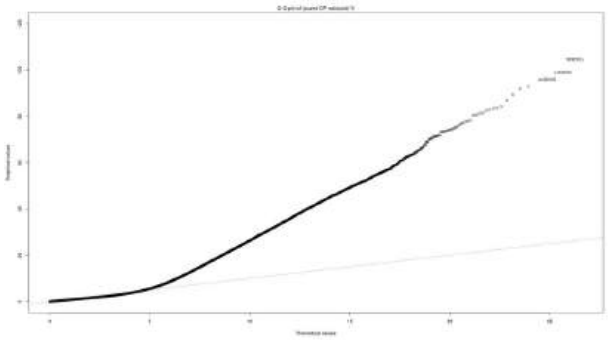

Figure 31: *Non-skewed distribution of the A31P*

As reported in **Table 5**, the mean values, corresponding to the average RMSD of all conformations within the same structure, progressively increase from the Native state to the A31P mutant. The standard deviation value reflects the spread of RMSD values, with the smallest observed for the Native form, while the two variants exhibit similar amplitudes for the overall structure, excluding the tails. The gamma1 values indicate that all distributions are positively skewed, meaning most values are concentrated below the mean, with some larger values extending the tail to the right. Among them, the A31P has the highest skew, revealing a more pronounced right tail. The log-likelihood values show that the skewed model provides a better fit to the data. In particular, for the native Rop, the skewed model fits substantially better, as evidenced by the LLG per observation (0.764 vs 0.072), while for the other two mutants the improvement is modest. Therefore, accounting for skewness is particularly important for the Native Rop, as a symmetrical representation would fail to capture the true distribution.

2. Turn A & Turn B regions (Figure 16):

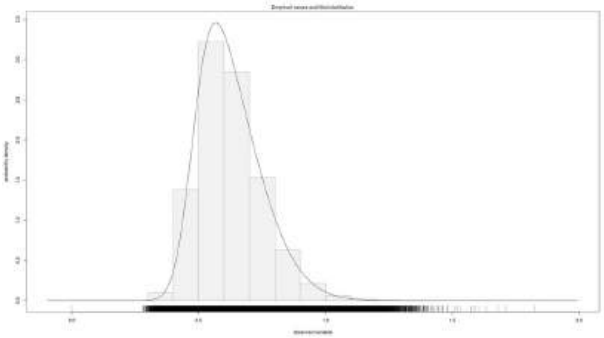

***Figure 32**: native Rop*

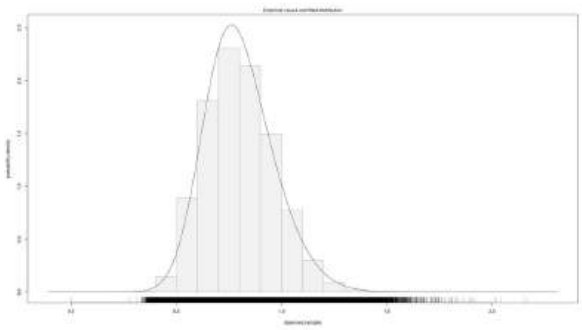

***Figure 33**: D30P mutant*

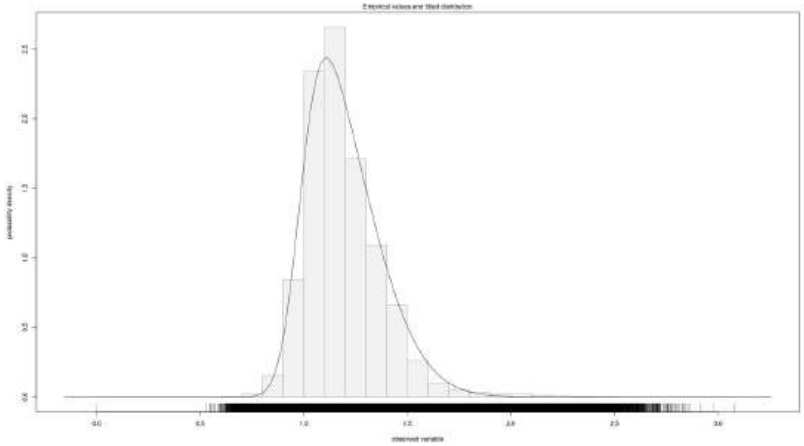

***Figure 34**: A31P mutant*

***Table 6:** Statistical parameters describing the flexibility distributions of the Turn A and B regions of native Rop, D30P and A31P variants.*

	Native	**D30P**	**A31P**
Mean	0.63	0.8	1.2
s.d.	0.12	0.16	0.18
Gamma1	0.7	0.43	0.72
Log-likelihood (skewed)	1413946.12	808145.54	1427874.56
LLG/ observation	0.71	0.4	0.36
Log-likelihood (non-skewed)	1318082.49	782947.85	1091380.44
LLG/ observation	0.66	0.39	0.27

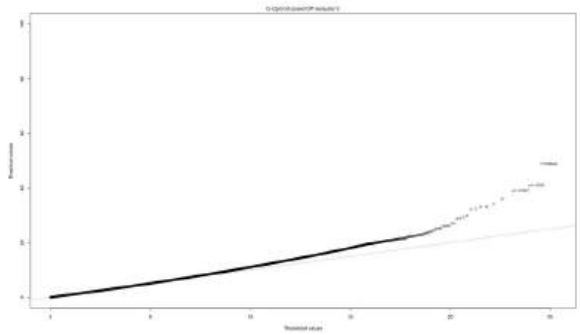

Figure 35: *Skewed distribution of the native Rop*

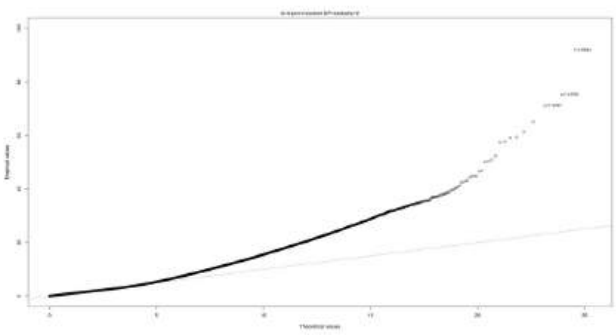

Figure 36: *Non-skewed distribution of the native Rop*

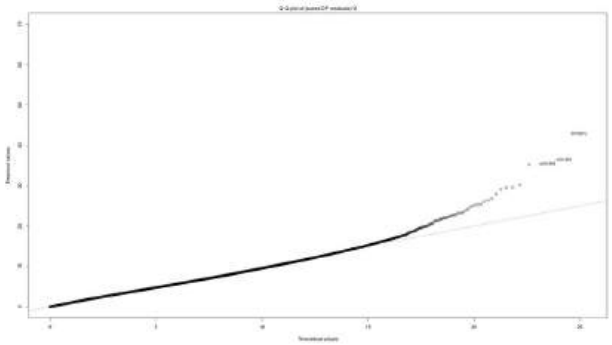

Figure 37: *Skewed distribution of the D30P*

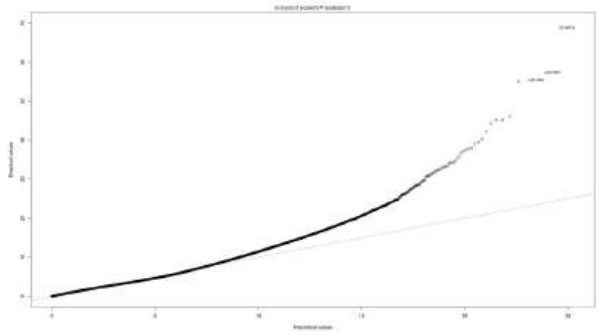

Figure 38: *Non-skewed distribution of the D30P*

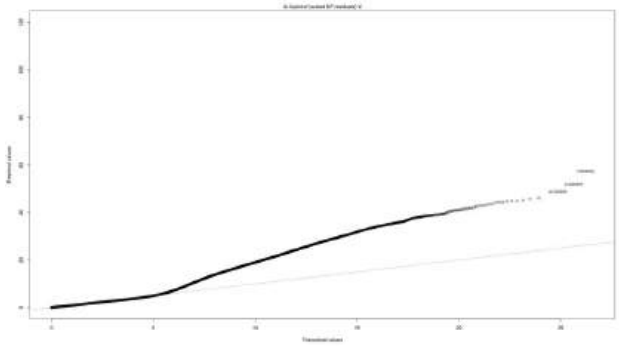

Figure 39: *Skewed distribution of the A31P*

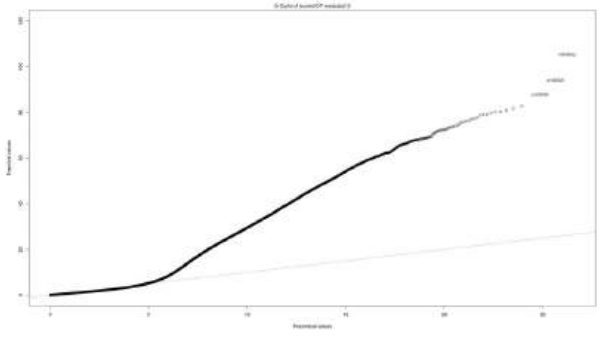

Figure 40: *Non-skewed distribution of the A31P*

Similarly to the previous statistical analysis, in the turns region it is evident that the RMSD mean values follow the same pattern, with the A31P variant showing the highest overall RMSD values and the greatest variability, indicating a less stable structure. At the same time, the LLG/observation values confirm that the skewed distribution model fits the experimental results better across all trajectories. Notably, the most significant deviation is observed for the A31P mutant, suggesting that a symmetrical representation of the data would fail to describe the real distribution. This indicates that the skewed model is more appropriate in this case, which is also evident from the comparison between the skewed and non-skewed distribution diagrams in each trajectory.

3. Turn A region (Figure 17):

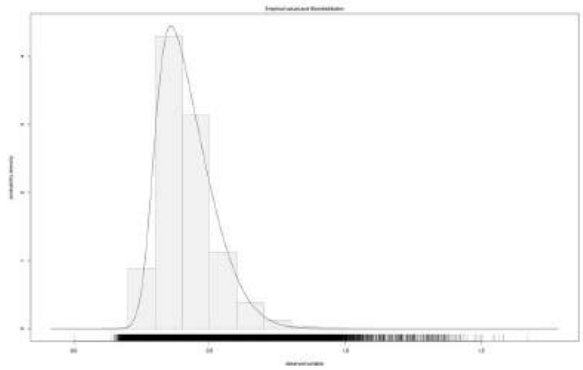

***Figure 41**: native Rop*

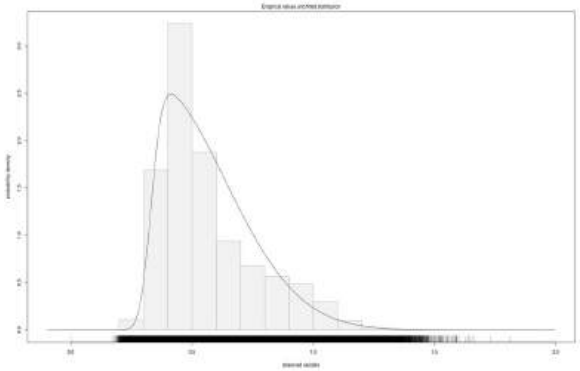

***Figure 42**: D30P mutant*

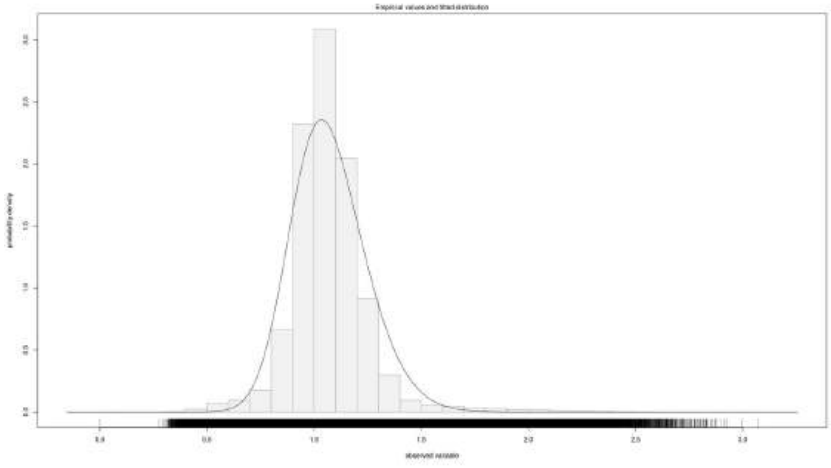

***Figure 43**: A31P mutant*

***Table 7**: Statistical parameters describing the flexibility distributions of Turn A region of native Rop, D30P and A31P variants.*

	Native	**D30P**	**A31P**
Mean	0.42	0.57	1.07
s.d.	0.1	0.19	0.17
Gamma1	0.8	0.94	0.38
Log-likelihood (skewed)	1913868.32	759726.74	1367392.53
LLG/ observation	0.96	0.38	0.34
Log-likelihood (non-skewed)	1731866.97	402305.4	1197384.67
LLG/ observation	0.87	0.2	0.3

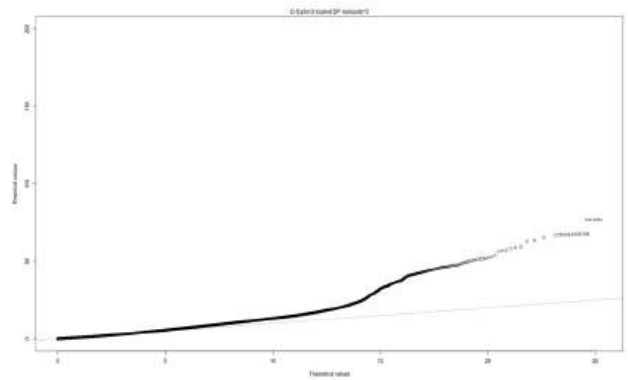

***Figure 44**: Skewed distribution of the native Rop*

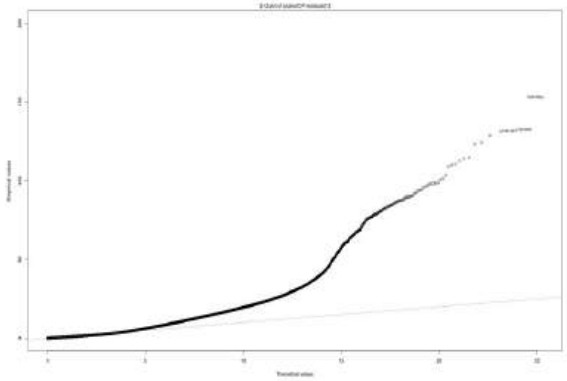

***Figure 45**: Non-skewed distribution of the native Rop*

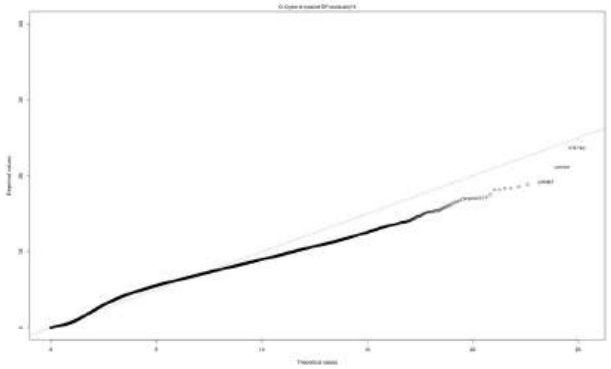

***Figure 46**: Skewed distribution of the D30P*

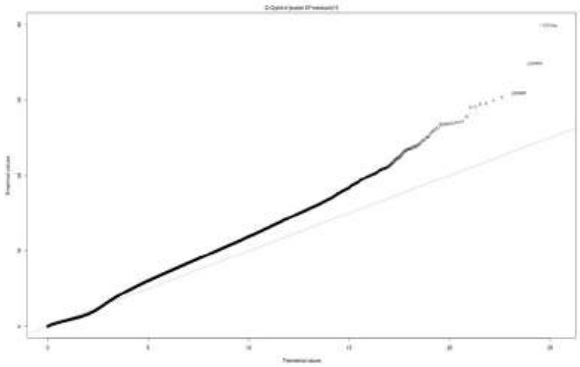

***Figure 47:** Non-skewed distribution of the D30P*

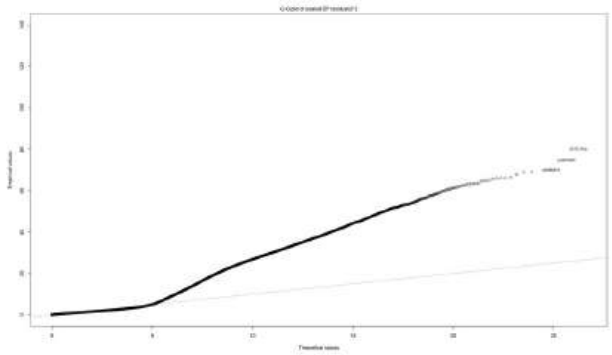

***Figure 48**: Skewed distribution of the A31P*

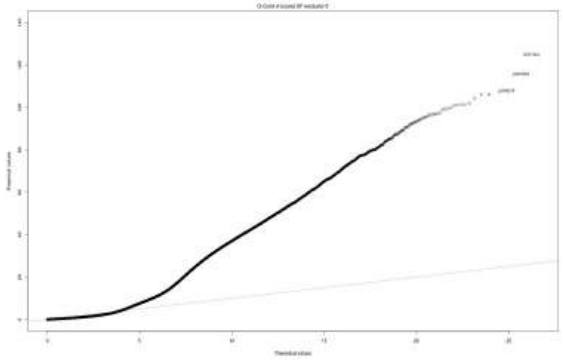

***Figure 49**: Non-skewed distribution of the A31P*

4.Turn B region (Figure 18):

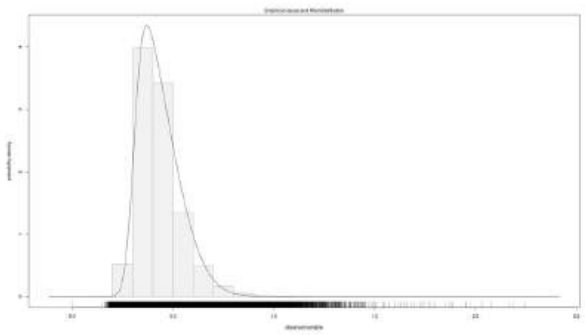

***Figure 50**: native Rop*

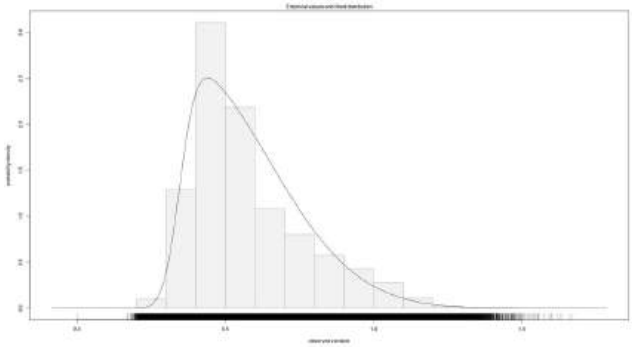

***Figure 51**: : D30P mutant*

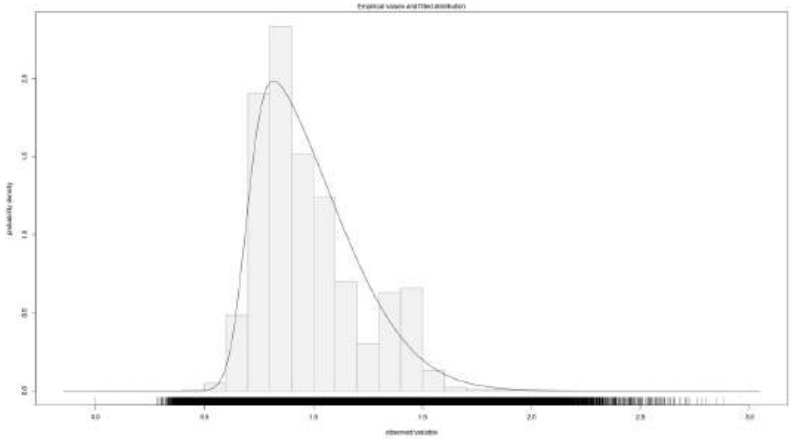

***Figure 52**: A31P mutant*

Table 8: *Statistical parameters describing the flexibility distributions of Turn B region of native Rop, D30P and A31P variants.*

	Native	**D30P**	**A31P**
Mean	0.43	0.58	0.98
s.d.	0.1	0.18	0.23
Gamma1	0.83	0.91	0.88
Log-likelihood (skewed)	1862481.84	764473.8	586076.31
LLG/ observation	0.93	0.38	0.15
Log-likelihood (non-skewed)	1647948.93	467741.84	107393.18
LLG/ observation	0.82	0.23	0.03

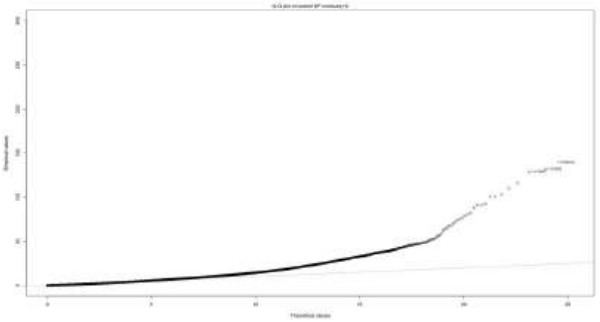

Figure 53*: Skewed distribution of the native Rop*

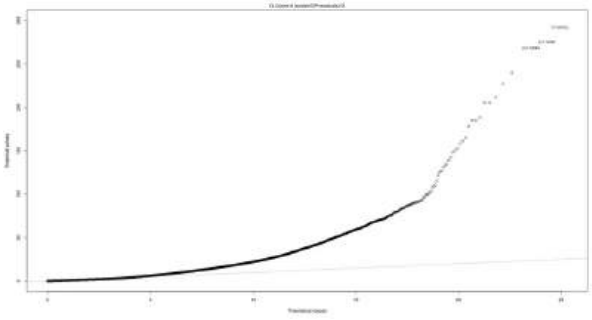

Figure 54*: Non-skewed distribution of the native Rop*

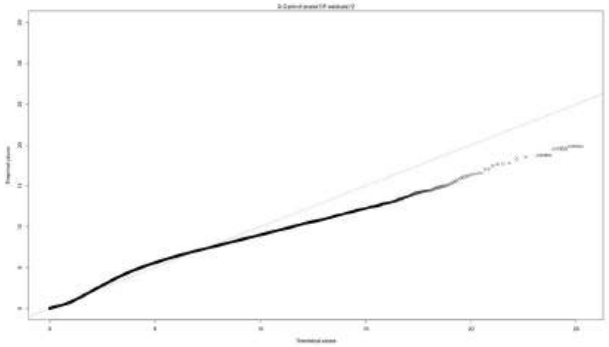

Figure 55*: Skewed distribution of the D30P*

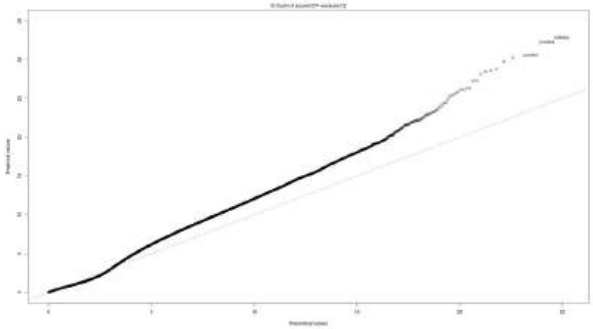

Figure 56*: Non-skewed distribution of the D30P*

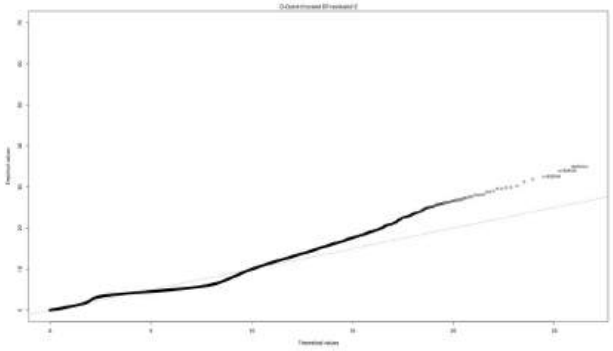

Figure 57*:Skewed distribution of the A31P*

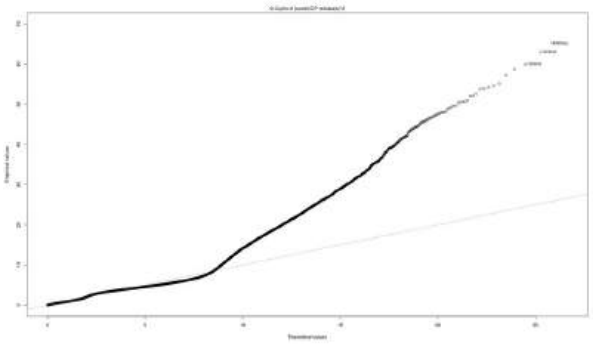

Figure 58: *Non-skewed distribution of the A31P*

So, in the Turn A, as in the Turn B region, separately, the statistical results show a consistent pattern regarding the mean RMSD values, with the A31P mutant exhibiting significantly higher RMSD averages. The difference in standard deviation indicates that the RMSD amplitude is more pronounced in the Turn B region of A31P, while the corresponding values of the other two trajectories are essentially the same. A general assessment of the LLG/observation values shows that, in both regions, the skewed model describes the distribution appropriately. In particular, in all cases the skewed model fits slightly better, except in the Turn B region of the A31P variant, in which the deviation between the skewed and non-skewed distributions is considerable, revealing that the symmetrical model would be unable to describe the current distribution.

Consequently, the overall statistical analysis supports the conclusions derived from the initial histograms (**Figures 15-18**). The average RMSD value increased progressively from the native Rop to the A31P mutant, indicating a hierarchical decrease in structural stability among the three systems. Each distribution exhibits positive skewness along the X-axis. Applying the normal model into the data, the log-likelihood per observation value improved when switching to a skewness model

distribution, indicating that this model provides a better interpretation of the results.

The Q-Q plots clearly confirms the already presented findings, in which in the case of the non-skewed distribution, the data deviate significantly from the theoretical model. These plots also reveal specific frames that deviate remarkably from the expected theory shown by the faint line.

The corresponding frames presented below:

1. <u>All residues (excluded N & C tails)</u> :

- native Rop: 1319400, 1531044 & 1319402
- D30P: 501868, 501863 & 501867
- A31P: 1828193, 1828548 & 1828220

2. <u>Turn A & Turn B regions:</u>

- native Rop: 1112097, 1112092 & 1112084
- D30P : 501868, 501859 & 501867
- A31P: 1235260, 1828220 & 1235265

3. <u>Turn A region:</u>

- native Rop: 1112084, 1112092 & 1112097
- D30P: 263800, 118524 & 125692
- A31P: 1828548, 1828549 & 1828550

4. <u>Turn B region:</u>

- native Rop: 1531231, 1531099 & 1531029
- D30P: 250887, 250890 & 1179718
- A31P: 3066219, 3067285 & 3072154

3.3 Dihedral PCA

Furthermore, in order to investigate how these point-mutations structurally affect the conformation of each variant compared to the native form, dPCA analysis is employed. Projecting the data into a low dimensional space highlights the most significant conformational changes, enabling the visualization of the dynamic landscapes of the native and mutant proteins. This analysis provides information into whether a mutation preserves a native-like conformation or induces overall instability by adopting a different spatial arrangement. Ultimately, dPCA facilitates a deeper understanding of how these specific mutations influence protein flexibility and stability.

3.3.1 Application of Dihedral PCA Using Carma and Grcarma

Native Rop: Turn A

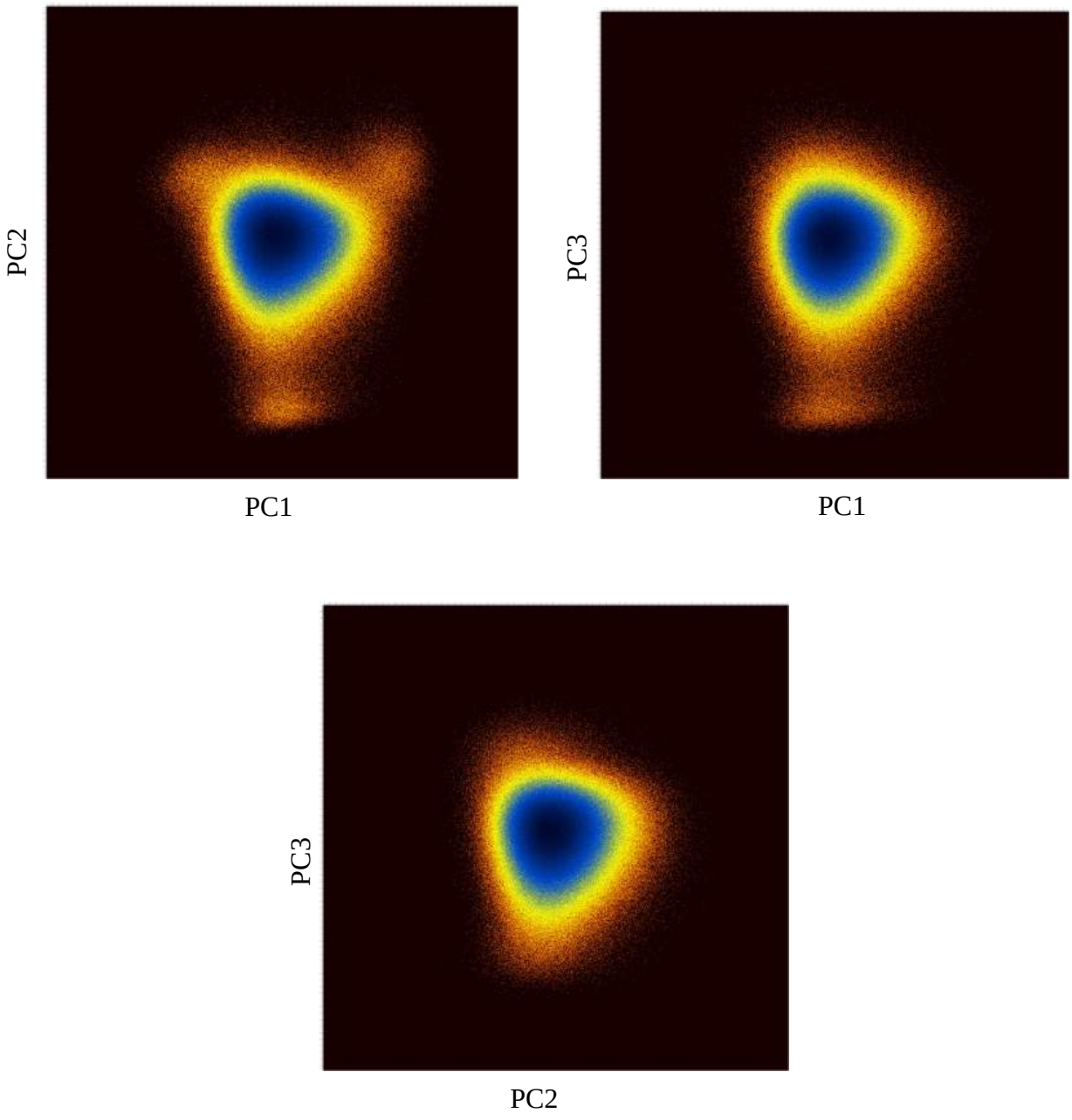

Figure 59: 2D representation plots of the variation distribution for Turn A region residues (24-39) of native Rop. Three pairs of principal components (PC1-PC2, PC1-PC3 & PC2-PC3) are shown.

The cumulative variance curve illustrates how the total dynamical variance of the system is distributed among the principal components. For the Turn A region of the native Rop, the motion is not uniformly distributed across the 60 PCs. Instead, the most dominant PCs have a greater impact on the system, with approximately the first three PCs accounting for about 29% of the total variance, the first four PCs explaining about 35%, and the first five PCs contributing approximately 41%. This smooth and progressively diminishing contribution of each PC indicates that this region undergoes slight rearrangements which are normally caused under natural conditions.

As illustrated in the 2D plots of the three most dominant PCs (**Figure 59**), the density of conformations is concentrated near the center, while fewer structures appear in the periphery. This distribution confirms the previous observations, in which no extensive conformational changes arise. A main cluster is observed, containing the vast majority of conformations (938448), while there is a secondary cluster encompasses only 6 conformations. The two distinct clusters are characterized by a remarkable energetic deviation. For the first cluster, the energy value was calculated as +0.096 kcal/mol**,** while the corresponding value for the second cluster was +7.531 kcal/mol**,** further confirming that most conformations exhibit only minor deviations within a generally compact and stable structure.

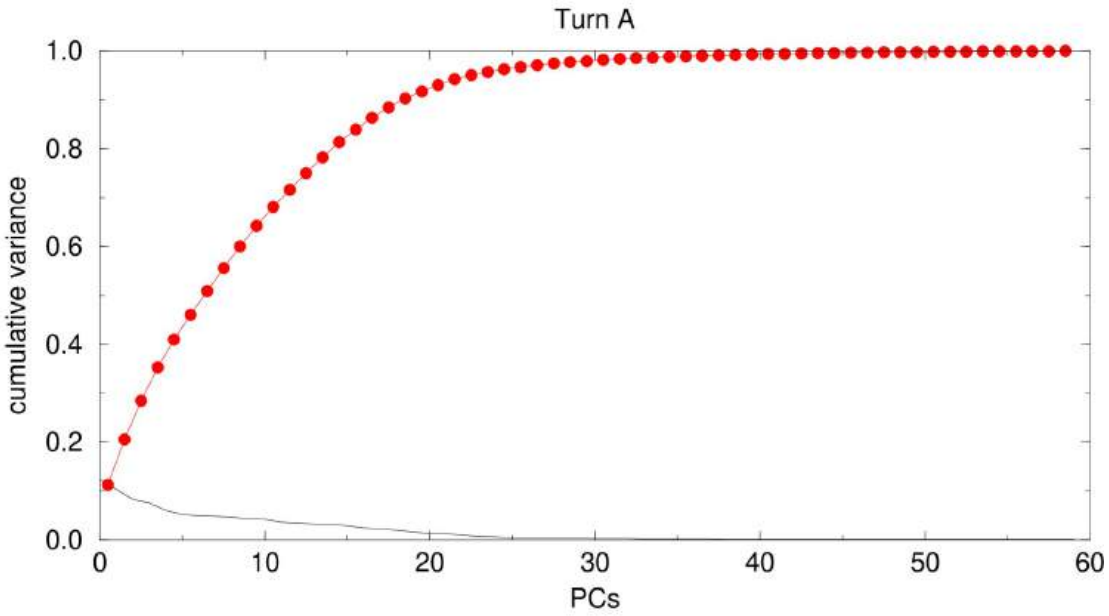

***Figure 60**: The cumulative variance for the Turn A region of native Rop, explained by the principal components, is shown by the red curve. Each filled red circle represents the cumulative contribution up to a given component, while the faint black line depicts the eigenvalue spectrum.*

Native Rop: Turn B

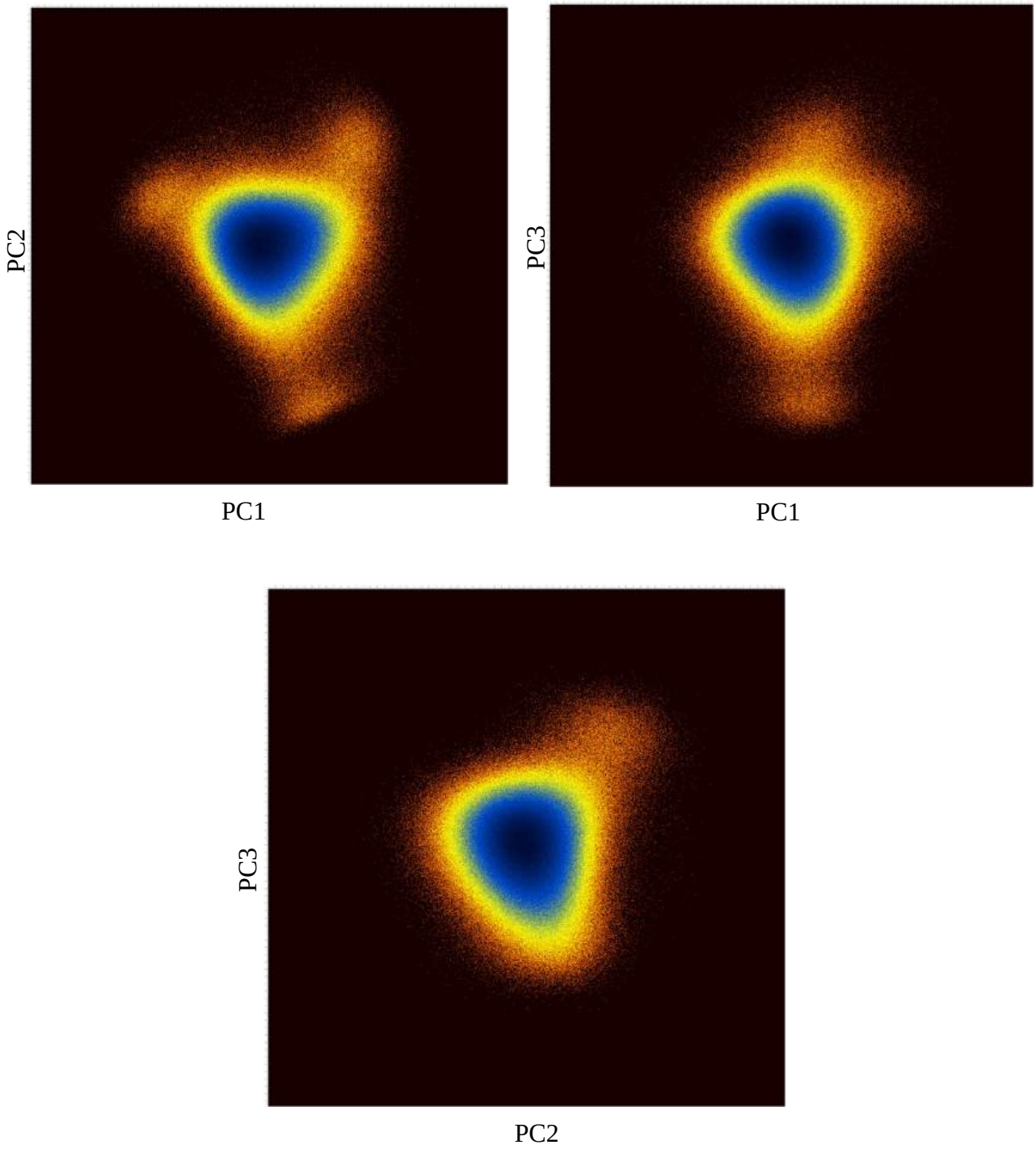

Figure 61: 2D representation plots of the variation distribution for Turn B region residues (24-39) of native Rop. Three pairs of principal components (PC1-PC2, PC1-PC3 & PC2-PC3) are shown.

As expected, similar results were obtained for the Turn B region of native Rop. The first five Principal Components account for approximately 41% of the overall fluctuation, while it is obvious that no significant rearrangements within the loop of the monomer occur. Accordingly to the Turn A, most conformations belong to the main cluster, however, the overall structures distributed within 4 clusters.

***Table 9**: Distribution of frames among the four clusters for the Turn B region in the native Rop trajectory*

Clusters	Frames/cluster	Total frames
Cluster 1	938448	2000001
Cluster 2	17	2000001
Cluster 3	8	2000001
Cluster 4	9	2000001

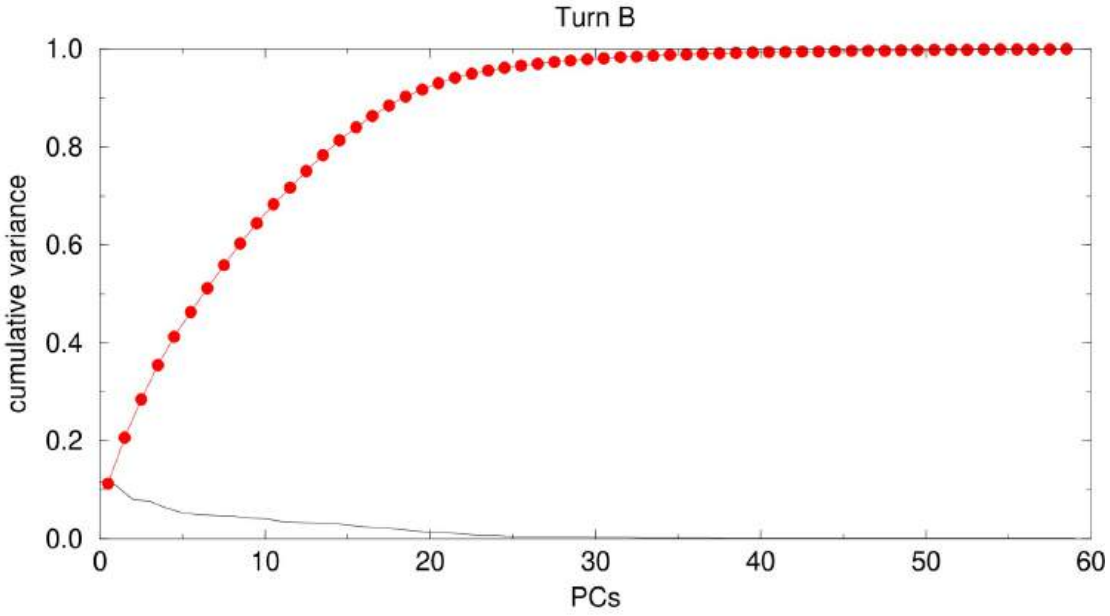

***Figure 62**: The cumulative variance for the Turn B region of native Rop, explained by the principal components, is shown by the red curve. Each filled red circle represents the cumulative contribution up to a given component, while the faint black line depicts the eigenvalue spectrum.*

D30P mutant: Turn A

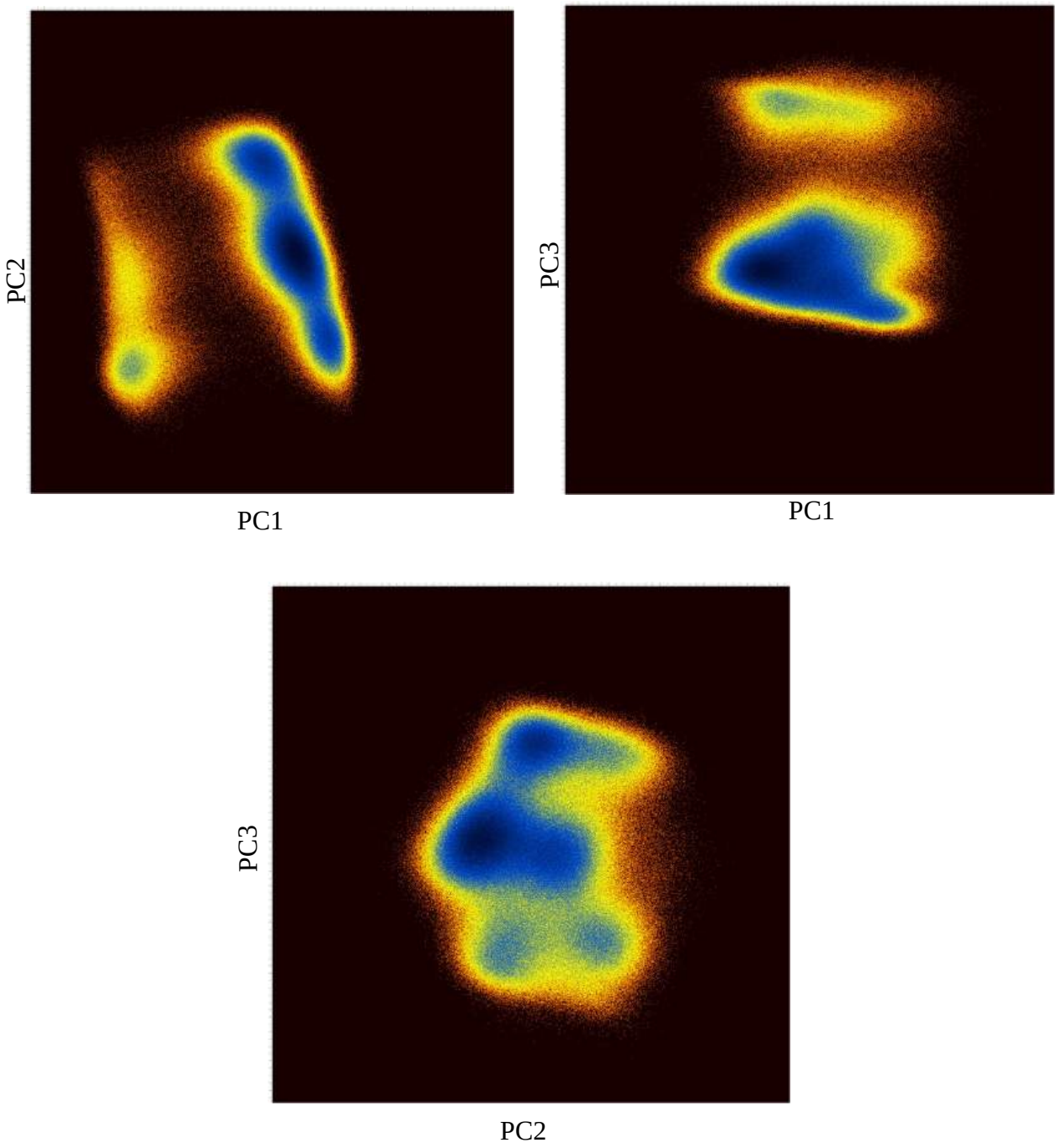

Figure 61: 2D representation plots of the variation distribution for Turn A region residues (24-39) of the D30P mutant. Three pairs of principal components (PC1-PC2, PC1-PC3 & PC2-PC3) are shown.

As shown by the 2D density plots (**Figure 61**), the conformational ensemble is mainly concentrated within the first cluster. Nevertheless, the landscapes appear more extended, capturing a broader surface compared to the native Rop. In order to further investigate the actual number of distinct conformational states represented in the ensemble, a 5D clustering analysis was performed. A direct comparison between the 3D and 5D analysis allows the evaluation of whether the heterogeneity observed in lower-dimensional projections corresponds to distinct conformational states is mainly a consequence of dimensionality reduction. Overall, this comparison provides a more robust view of the true conformational complexity of the system.

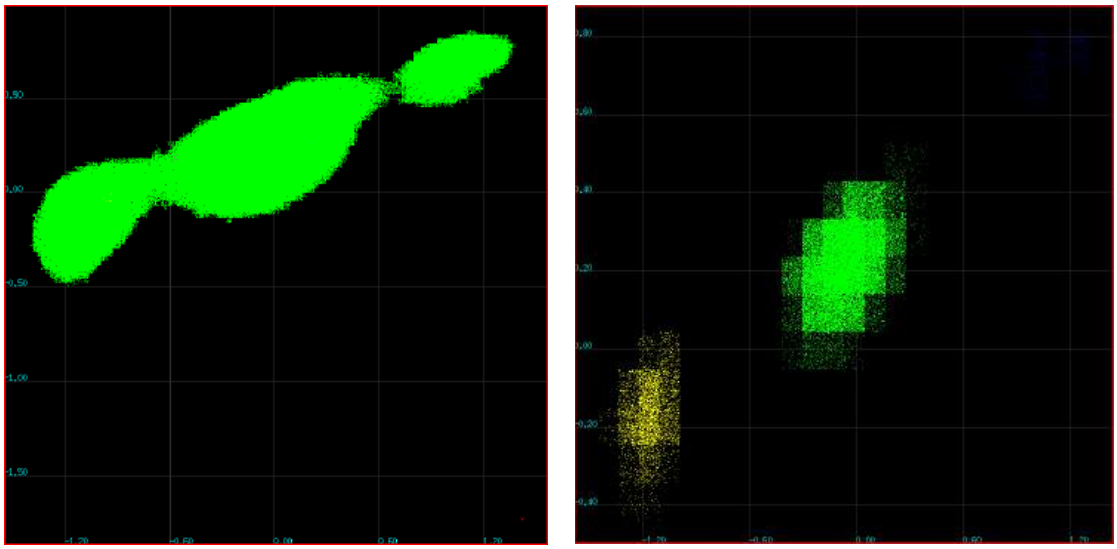

Figure 62: 2D representation plots of the 3D clustering analysis (left panel) and the 5D clustering analysis (right panel), showing the three distinct clusters colored in green, yellow and blue (top-right corner).

The 3D clustering analysis suggests that all conformations are included in a single, expanded cluster, shown in green. In contrast, the 5D clustering analysis reveals three distinct regions dividing the conformations. The majority of them belong to the central, most dominant cluster. Two additional clusters were defined: the second-most populated cluster,

shown in yellow, and a third cluster, colored in blue, containing a notable fraction of conformations.

The 5D distribution of the D30P mutant appears consistent with the cumulative variance (**Figure 63**), as the five most dominant principal components account for approximately 63% of the total variance, indicating that each contributes substantially to the overall motion. On the other hand, the corresponding value for the native Rop is roughly 20% lower, reflecting a more restricted conformational landscape, occupying a single, well-defined cluster. This difference suggests that the D30P variant exhibits greater flexibility, allowing the protein to access multiple conformational states.

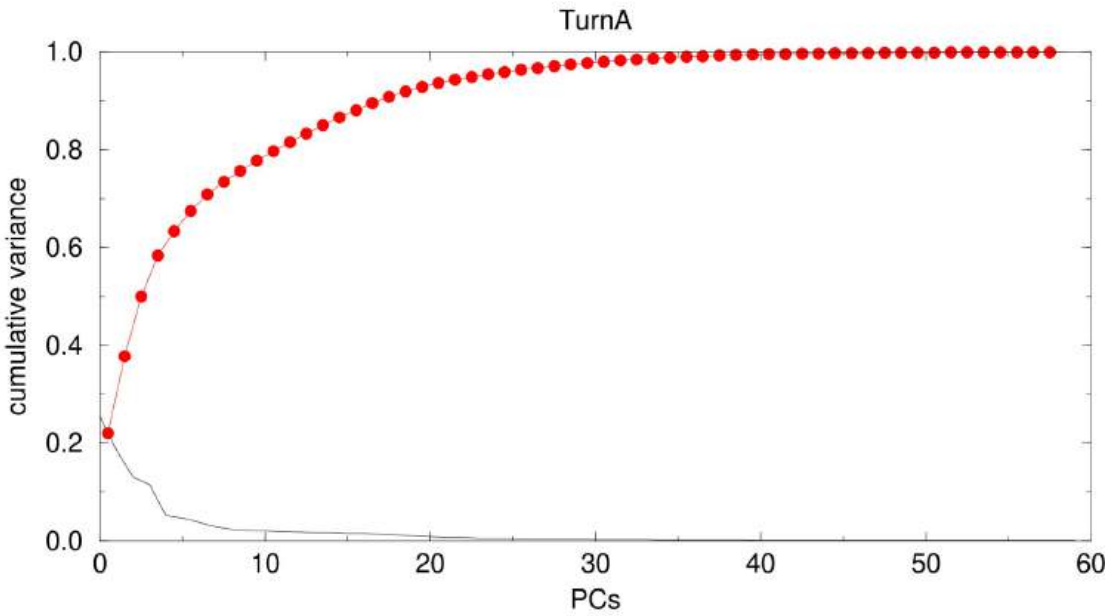

***Figure 63**: The cumulative variance for the Turn A region of D30P variant, explained by the principal components, is shown by the red curve. Each filled red circle represents the cumulative contribution up to a given component, while the faint black line depicts the eigenvalue spectrum.*

In more detail, **Table 10** presents the distribution of frames among the seven distinct clusters The main accumulation of conformations remains within the dominant cluster, while there are complimentary clusters

containing less stable structures characterized by notably higher energy values.

Table 10: Distribution of frames among the seven clusters for the Turn A region in the D30P trajectory

Clusters	**Frames/cluster**	**Total frames**
Cluster 1	827018	2000001
Cluster 2	75	2000001
Cluster 3	3	2000001
Cluster 4	23	2000001
Cluster 5	5	2000001
Cluster 6	7	2000001
Cluster 7	6	2000001

D30P mutant: Turn B

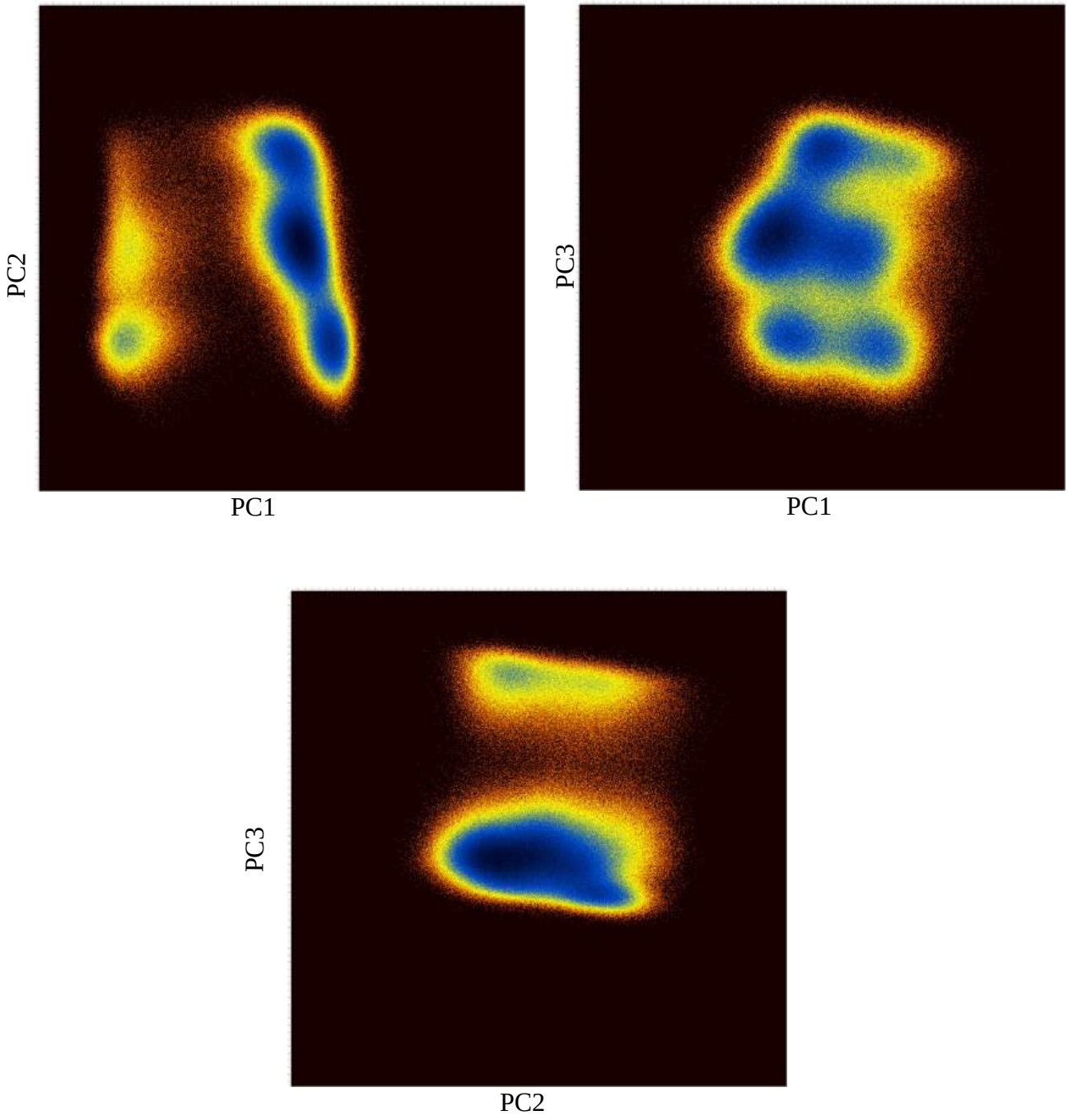

Figure 64: 2D representation plots of the variation distribution for Turn B region residues (24-39) of the D30P mutant. Three pairs of principal components (PC1-PC2, PC1-PC3 & PC2-PC3) are shown.

While a similar overall pattern is observed between the Turn A and Turn B regions in the 2D density plots (**Figures 61 & 64**), there are a clear difference in cluster organization. In the 3D clustering analysis of Turn B, two distinct clusters are identified. Although the general pattern is similar to that of Turn A (**Figure 62**), Turn B shows a dominant cluster containing the majority of conformations, alongside an additional cluster holding a notable percentage. At the same time, in the 5D clustering analysis, the three distinct clusters are maintained, although their distribution exhibits some reshuffling. In particular, the less populated cluster, shown in blue, contains a greater number of structures, as indicated by its brighter color.

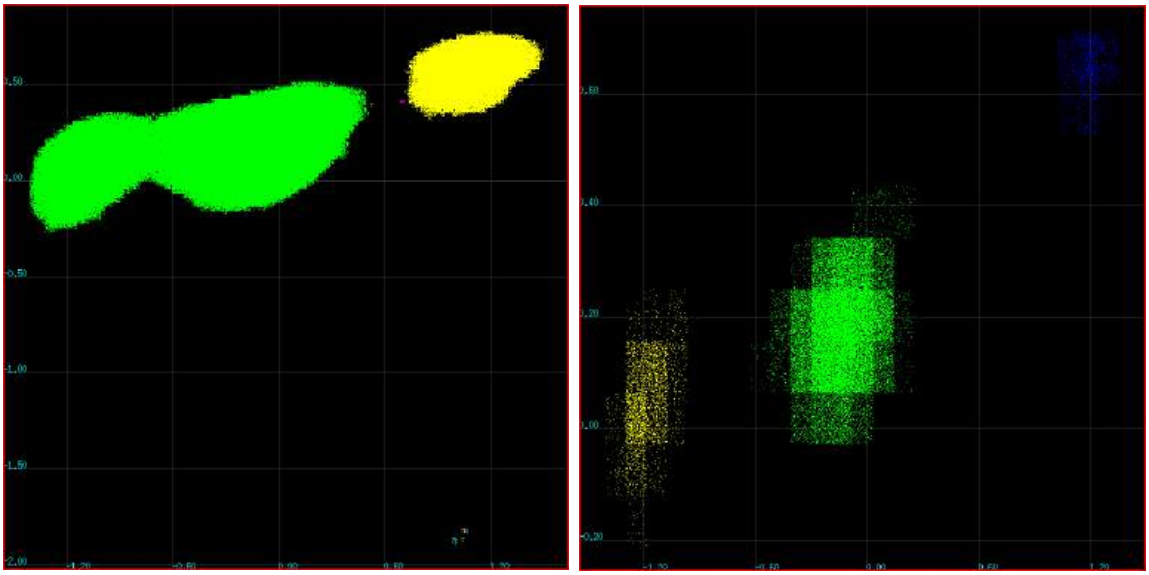

Figure 65: 2D representation plots of the 3D clustering analysis (left panel) and the 5D clustering analysis (right panel), showing the three distinct clusters colored in green, yellow and blue (top-right corner).

The above observations are also supported by the cumulative variance (**Figure 66**), which represents that the first five principal components explain an even higher percentage of the total variance, approximately 68%. This indicates the predominant influence of the first PCs on the

overall motion of the system. The dynamical behavior is therefore not evenly distributed across all modes, instead a small number of PCs describe large-amplitude motions that significantly affect the stability of the trajectory. In combination with the presence of three distinct clusters, it is evident that the Turn region of the D30P variant exhibits reduced compactness compared to the native form.

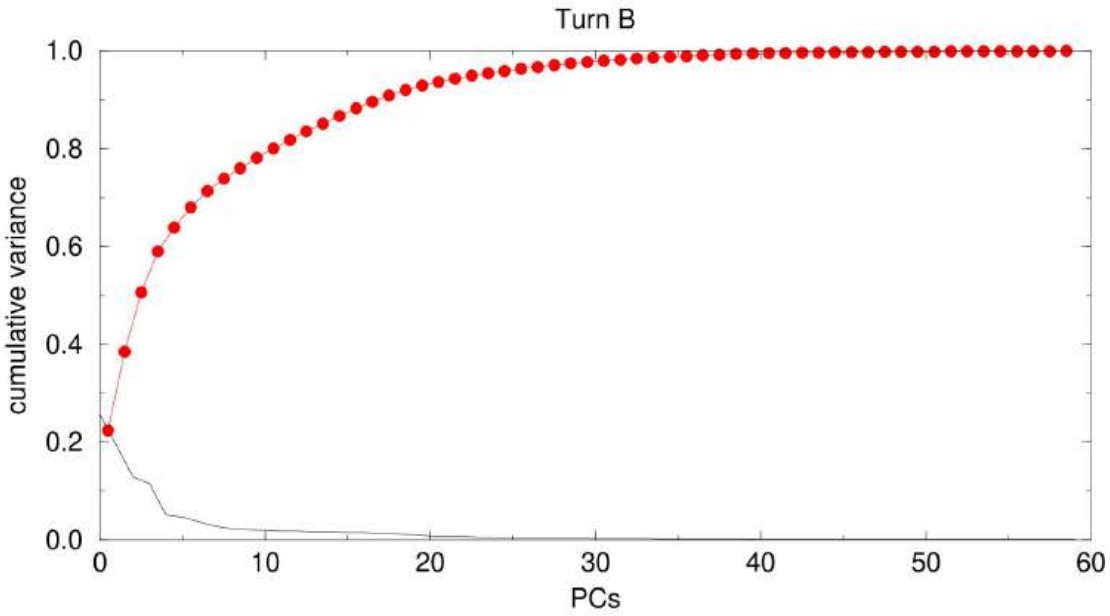

***Figure 66**: The cumulative variance for the Turn B region of D30P variant, explained by the principal components, is shown by the red curve. Each filled red circle represents the cumulative contribution up to a given component, while the faint black line depicts the eigenvalue spectrum.*

In **Table 11**, the detailed distribution of all frames across the total number of clusters is provided. Most conformations are accumulated within the two most dominant clusters, corresponding to the most energetically favorable states. Additionally, eight less-populated clusters are observed, containing conformations exhibiting higher energy values.

***Table 11**: Distribution of frames among the ten clusters for the Turn B region in the D30P trajectory*

Clusters	**Frames/cluster**	**Total frames**
Cluster 1	702372	2000001
Cluster 2	124202	2000001
Cluster 3	19	2000001
Cluster 4	13	2000001
Cluster 5	33	2000001
Cluster 6	6	2000001
Cluster 7	20	2000001
Cluster 8	8	2000001
Cluster 9	6	2000001
Cluster 10	4	2000001

A31P mutant: Turn A

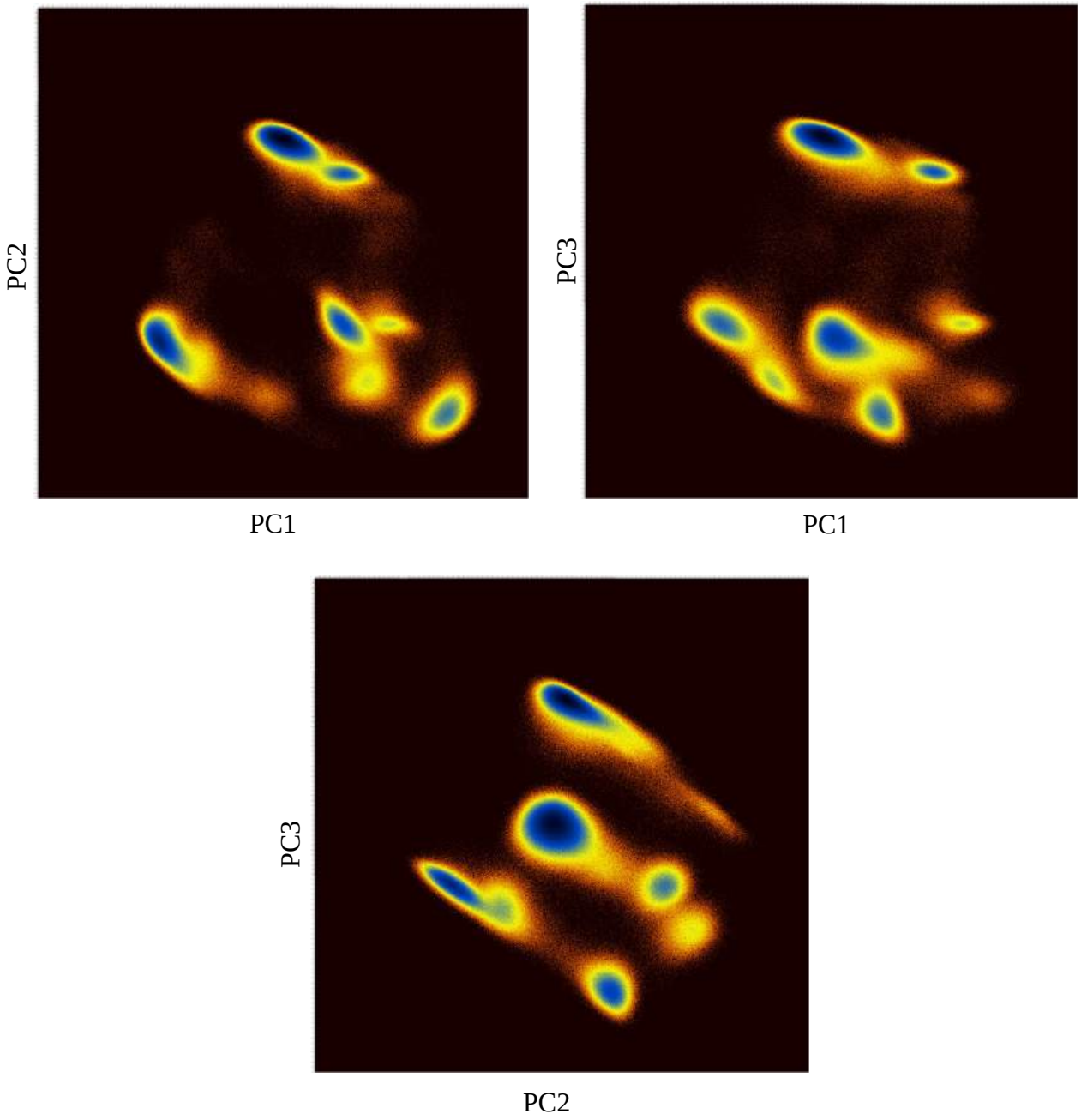

Figure 67: 2D representation plots of the variation distribution for Turn A region residues (24-39) of the A31P mutant. Three pairs of principal components (PC1-PC2, PC1-PC3 & PC2-PC3) are shown.

Attempting to investigate the impact of the A31P mutant on the overall structural topology, density plots were generated, suggesting that the conformations are distributed across multiple clusters. In the 3D clustering analysis, three distinct clusters of comparable size were identified, collectively encompasses all conformations. On the other hand, 5D clustering analysis reveals a more complex organization, with eight clusters of varying sizes. The majority of conformations are accumulated within the first four clusters, whereas the remaining clusters are progressively less-populated, with the last two containing only a marginal number of structures.

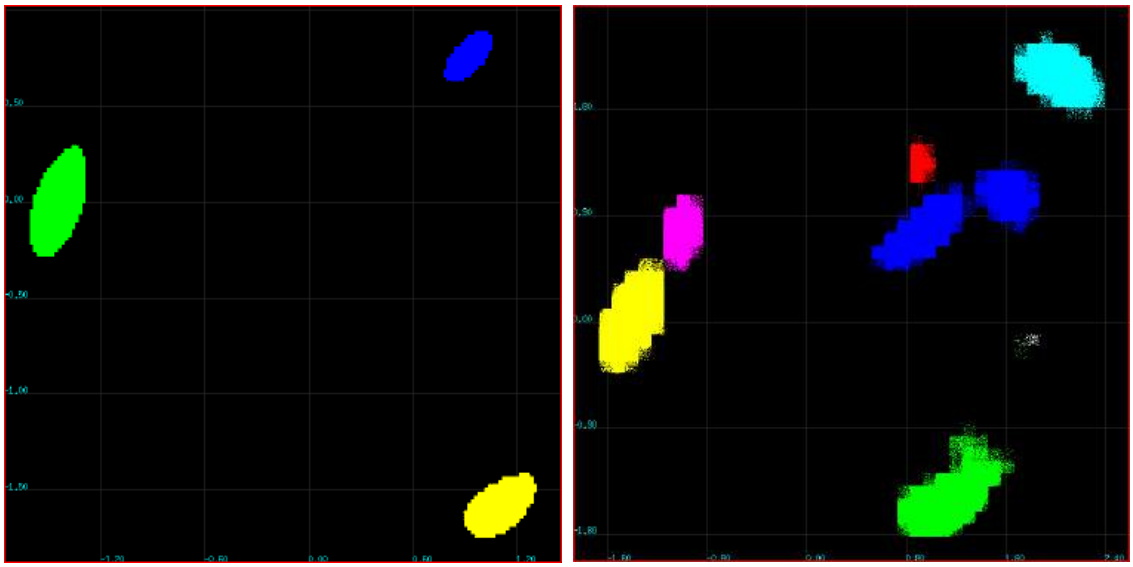

Figure 68: 2D representation plots of the 3D clustering analysis (left panel) and the 5D clustering analysis (right panel), showing the three distinct clusters colored in green, yellow and blue (top-right corner).

The cumulative variance (**Figure 69**) reveals that the five most dominant components account for approximately 70% of the overall motion. This contribution far exceeds that of the native Rop, indicating that this trajectory is characterized by less stable dynamical behavior dominated by high-amplitude motions. These conformations are distributed across multiple energetic states, none of which approach the low energy level

(+0.667 kcal/mol) associated with the cluster occupied by the native conformations. The presence of multiple conformations within higher energy clusters also reflects the increased flexibility of the system. Moreover, the connectivity observed among different clusters suggests that the system can feasible transition between district energetic states, revealing the structural instability of this variant.

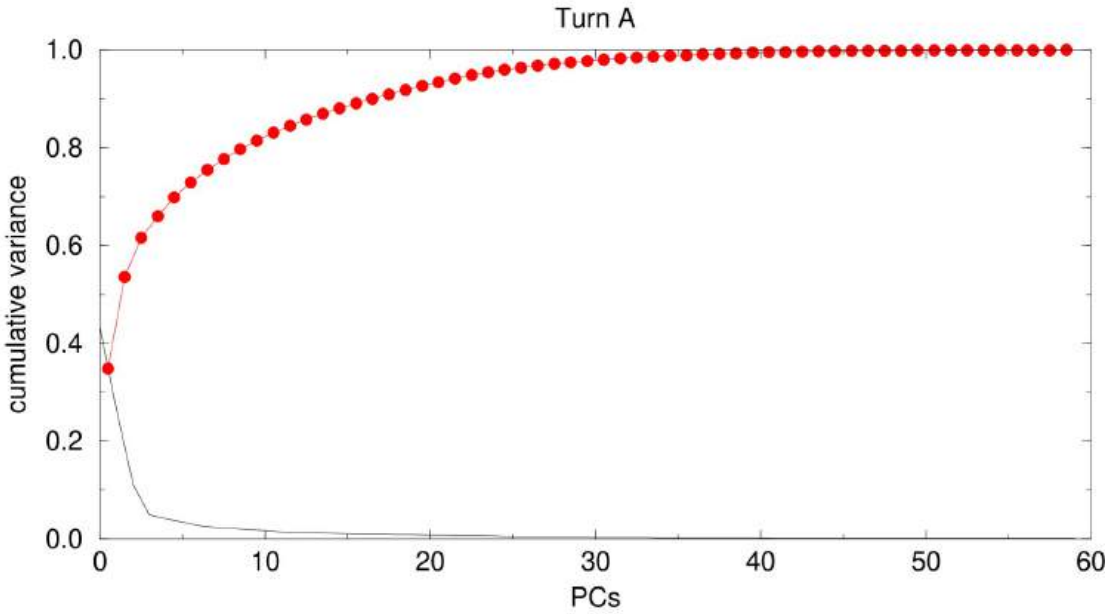

***Figure 69**:The cumulative variance for the Turn A region of A31P variant, explained by the principal components, is shown by the red curve. Each filled red circle represents the cumulative contribution up to a given component, while the faint black line depicts the eigenvalue spectrum.*

As shown in **Table 12**, the total number of frames is divided into three distinct clusters. Each cluster contains a substantial number of conformations, reflecting their distribution according to energy values. The most dominant cluster is characterized by the lowest energy, indicating that the most populated cluster corresponds to the most energetically favorable state.

***Table 12**: Distribution of frames among the three clusters for the Turn A region in the A31P trajectory*

	Frames/cluster		Total frames	Energy (kcal/mol)
Cluster 1	979672	(out of)	4000002	0.67
Cluster 2	357088	(out of)	4000002	1.38
Cluster 3	72144	(out of)	4000002	2.37

A31P mutant: Turn B

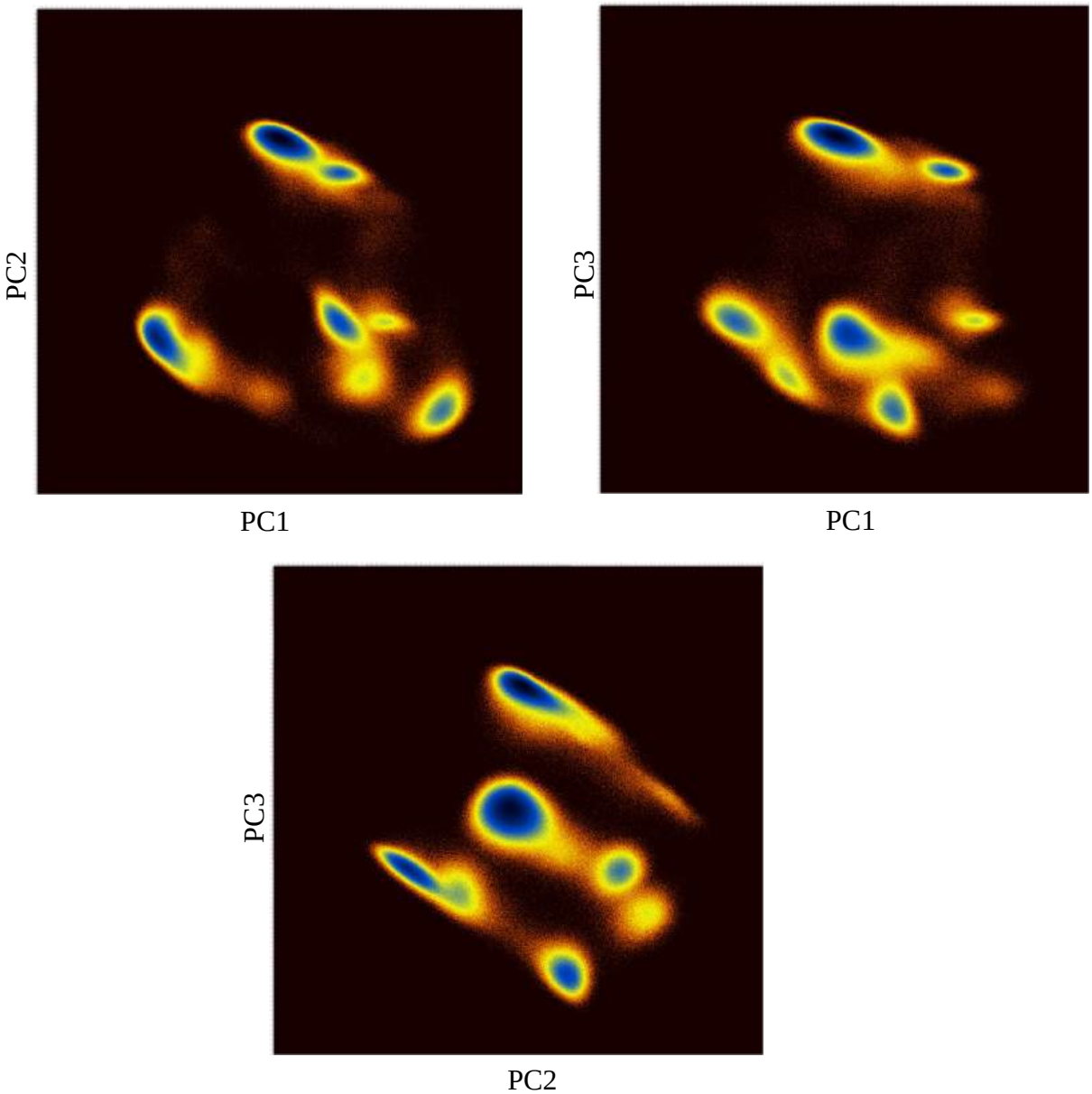

Figure 70: 2D representation plots of the variation distribution for Turn B region residues (24-39) of the A31P mutant. Three pairs of principal components (PC1-PC2, PC1-PC3 & PC2-PC3) are shown.

Similarly to Turn A, the 2D variation distribution plots exhibit multiple clusters, in which the conformations are classified based on their energy states. According to the 3D clustering analysis, the total number of structures is separated into four distinct ensembles, whereas the 5D clustering analysis reveals a broaden range of clusters. As the dimensionality of analysis increases, the classification transitions into a more complex network of states, in which each group differs both morphologically and in population.

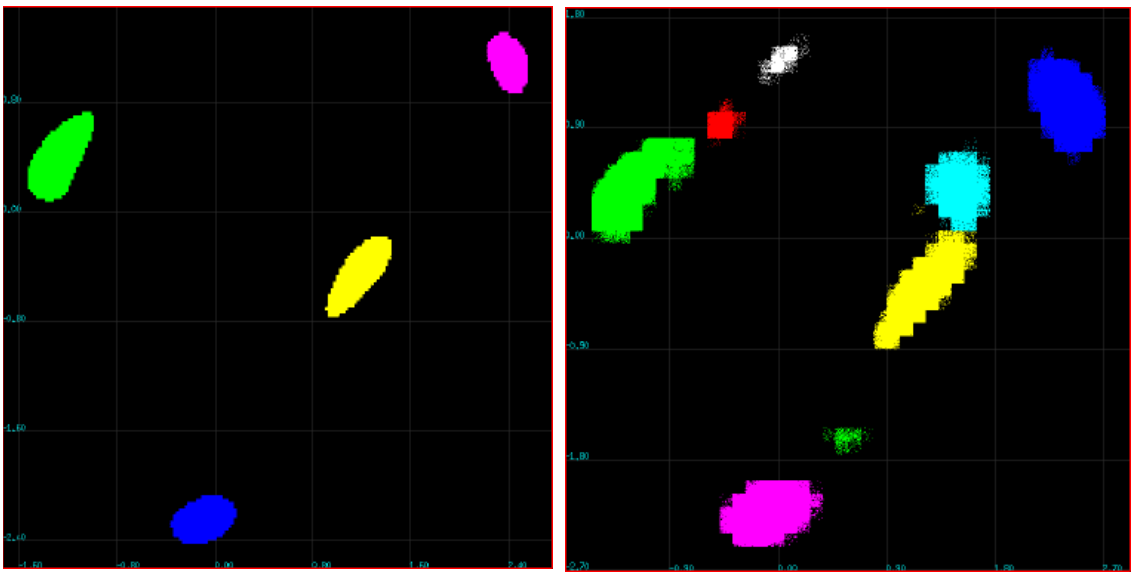

Figure 71: 2D representation plots of the 3D clustering analysis (left panel) and the 5D clustering analysis (right panel), showing the three distinct clusters colored in green, yellow and blue (top-right corner).

The cumulative variance shows, similarly to the Turn A region, that the five most dominant PCs account for approximately 71% of the total fluctuation. This indicates that the Turn B region exhibits extensive internal motion, in contrast to the native state. These findings reinforce each other, supporting the presence of a highly dynamic structure in which each conformational ensemble is characterized by a distinct energy barrier. Collectively, these clusters encompass the full set of conformations adopted by the A31P variant.

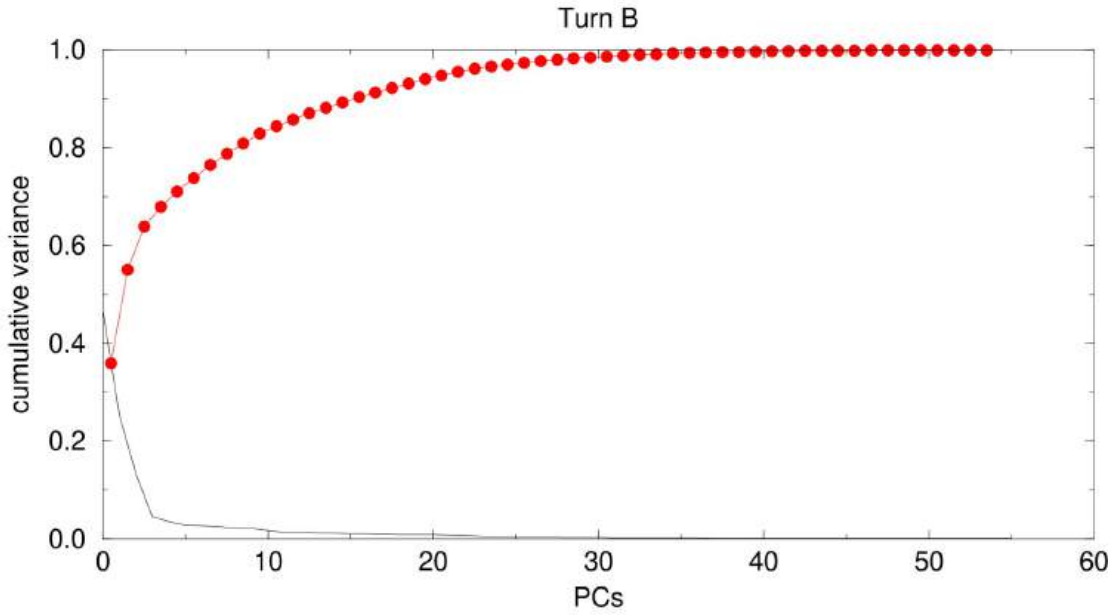

***Figure 72**: The cumulative variance for the Turn B region of A31P variant, explained by the principal components, is shown by the red curve. Each filled red circle represents the cumulative contribution up to a given component, while the faint black line depicts the eigenvalue spectrum.*

Consistent with the previous analyses, the plots combined with the dominant principal components reveal a strong similarity between the Turn B region and the previously described Turn A of the A31P mutant. A slightly discrepancy is observed, as the conformations in the Turn B are distributed among four clusters instead of three.

The energy of the first cluster (+0.285 kcal/mol) appears to be more favorable than that of turn A (+0.667 kcal/mol), suggesting more stable conformations in the most populated cluster of the second monomer. Notably, the upper energy limit of the Turn B region (+1.814 kcal/mol) is significantly lower compared to the corresponding value for Turn A (+2.367 kcal/mol), providing evidence of reduced energy barriers among the Turn B clusters, despite their increased number.

Table 13: Distribution of frames among the four clusters for the Turn B region in the A31P trajectory

	Frames/cluster		**Total frames**	**Energy (kcal/mol)**
Cluster 1	1527836	(out of)	4000002	0.29
Cluster 2	484604	(out of)	4000002	1.17
Cluster 3	263038	(out of)	4000002	1.57
Cluster 4	178661	(out of)	4000002	1.81

Collectively, the results derived from the Dihedral PCA analysis confirm previous observations. Similarly to what had already been observed from the RMSF and RMSD analyses, the native protein appears to be stable with only minor conformational changes mostly taking place within the same energetic basin. The overall motion is distributed smoothly across the PCs confirming the experimentally determined compact structure of the protein.

Consistent with the experimental data, the 5D landscape of D30P demonstrates a broadened distribution compared to the native. Still, there is a central cluster containing the majority of conformations which is important evidence for the preservation of its native-like character, while its lower Tm value is computationally reflected by the increased flexibility of D30P. Additionally, this is also reflected in the distribution across the PCs which is more concentrated within the first components. Taken together, the D30P appears to retain its native-like character while undergoing slight destabilization.

The 5D clustering analysis of the hypothetical native-like A31P structure carrying the point mutation at position 31 reveals increased flexibility, as its conformations occupy multiple accessible energetic states. These results are in agreement with the experimental data, as the A31P mutant is characterized as the most unstable among these systems. At the same time, the variant appears to retain its molten-globule characteristics without complete structural collapse.

3.3.2 Dihedral PCA Representative Structures

The representative structure corresponds to the closest to the average conformation derived from all frames belonging to the same cluster, giving a brief visualization of the alterations caused by the mutations in the turn region.

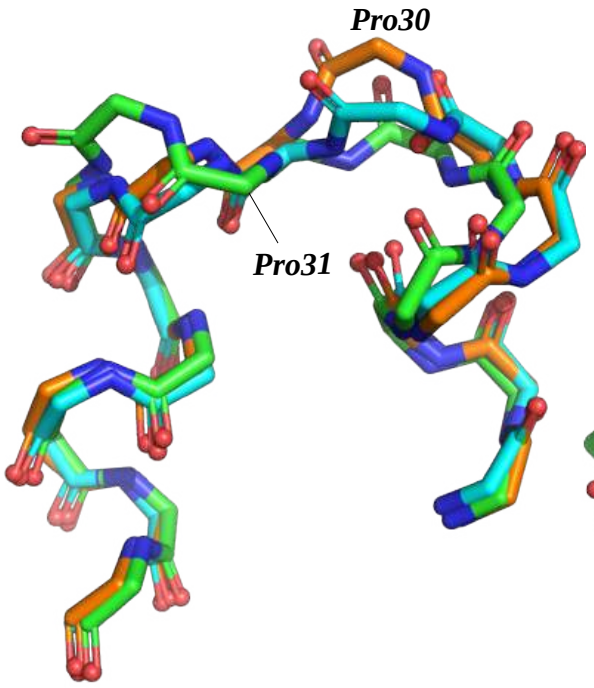

***Figure 73**: Superposition of the native Rop structure (cyan), D30P mutant (orange) and A31P mutant (green), as derived from the MD simulations. Only the backbone atoms are illustrated in the Turn A region. The point-mutation residues are labeled.*

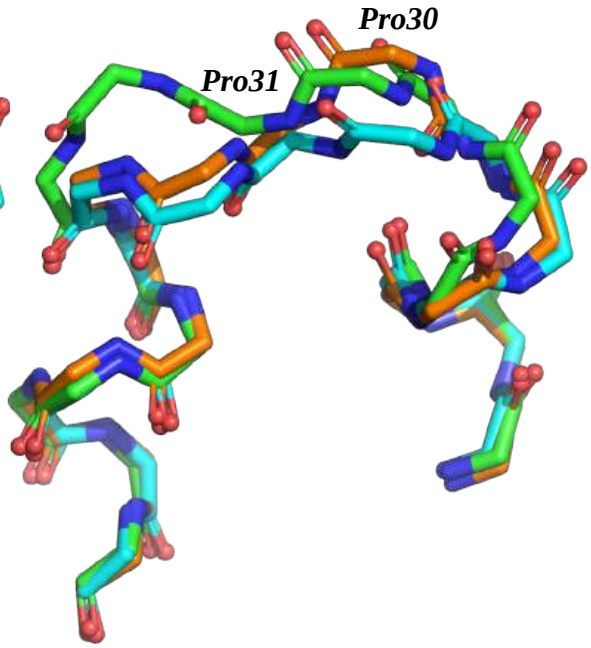

***Figure 74**: Superposition of the native Rop structure (cyan), D30P mutant (orange) and A31P mutant (green), as derived from the MD simulations. Only the backbone atoms are illustrated in the Turn B region. The point-mutation residues are labeled.*

Observation of **Figures 73 & 74** reveals that the insertion of a proline residue at positions 30 and 31 influences the stability of the loop, leading to local structural rearrangements. In particular, the D30P mutant generally follows the pattern of the native conformation, except for the

specific proline residue, which deviates from it by protruding from the rest of the structure.

Following the point-mutation, the overall folding of the D30P structure appears to return to the initial pattern, indicating that the impact of the mutation is limited and does not extend out of its local environment. In the case of A31P mutant, the presence of proline induces structural alterations that extend over a broaden region, including the α-helices. The alignment of the representative conformations reveals that this variant displays noticeable deviations from the reference structure, particularly in the region immediately adjacent to the mutation site, where their overlap has been reduced. These conformational discrepancies reflect a rearrangement of the turn region, which may impair smooth folding.

For direct comparison, the conformations belonging to the main cluster are shown in **Figures 75 and 76**, illustrating the representative structures of native Rop and the two mutants in a common color scale. Through this visualization, the evaluation of internal mobility becomes possible.

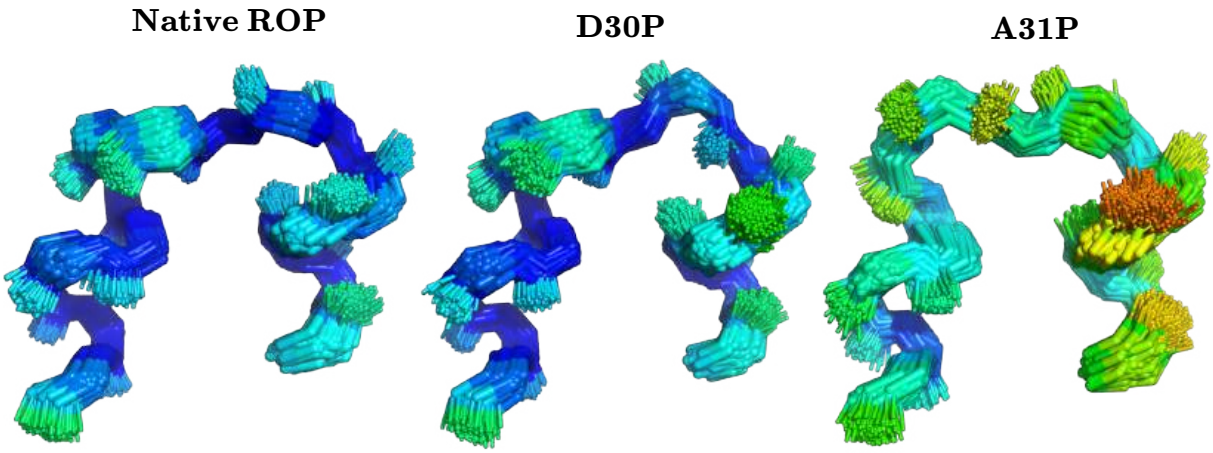

Figure 75: *Superposition of all conformations of the Turn A region belonging to the dominant cluster of native Rop, D30P and A31P. The most flexible segments are shown in red, with the color gradually transitioning to blue according to structural stability. The color scale is common for all illustrations.*

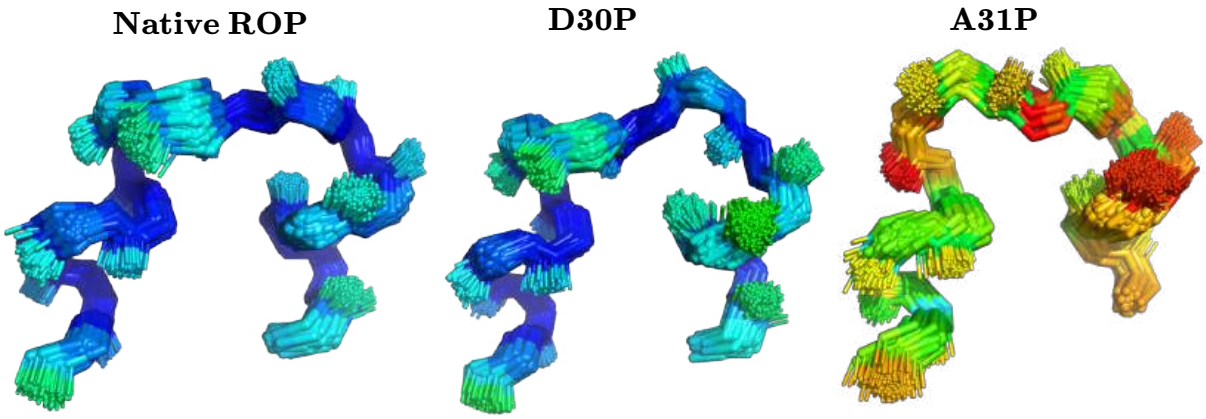

Figure 76: *Superposition of all conformations of the Turn B region belonging to the dominant cluster of native Rop, D30P and A31P. The most flexible segments are shown in red, with the color gradually transitioning to blue according to structural stability. The color scale is common for all illustrations.*

The assessment of flexibility across the three systems indicates that the native Rop exhibits the lowest RMSF values. The abundance of segments colored in cold hues reflects increased structural stability. The D30P variant is characterized by high structural similarity to the native form, deviating only slightly from the original folding. Finally, the A31P mutant adopts the most flexible structure, with warm colors dominating most of the surface, especially in the turn of monomer B. The presence of conformations that deviate significantly within the same cluster suggests that this mutation affects profoundly the overall stability of the structure. Cluster analysis supports that the native Rop fold represents the most stable conformational state, followed by the D30P mutant, whereas the A31P variant exhibits the lowest structural stability.

Finally, the following representation illustrates the flexibility within each trajectory through the superposition of a representative structure from each cluster, reflecting the overall flexibility and stability of each system.

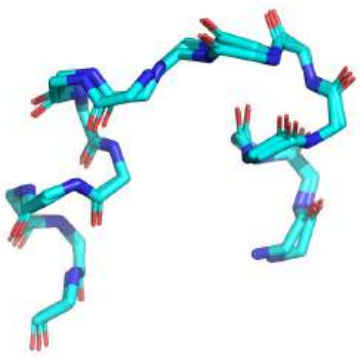

***Figure 77**: Superposition of the representative structures from the two clusters of Native Rop (monomer A). Only the backbone atoms are shown.*

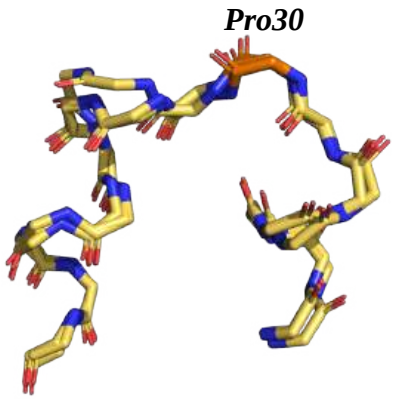

***Figure 78**: Superposition of the representative structures from the three clusters of D30P mutant (monomer A). Only the backbone atoms are shown, with proline residue labeled and highlighted in orange.*

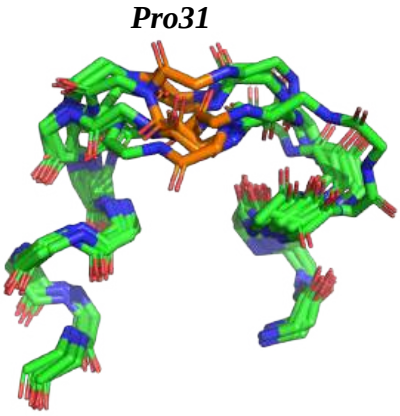

***Figure 79**: Superposition of the representative structures from the eight clusters of A31P mutant (monomer A). Only the backbone atoms are shown, with proline residue labeled and highlighted in orange.*

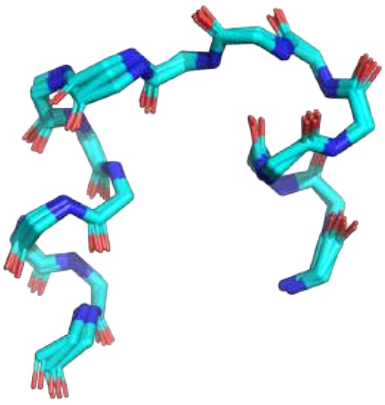

***Figure 80:** Superposition of the representative structures from the four clusters of Native Rop (monomer B). Only the backbone atoms are shown.*

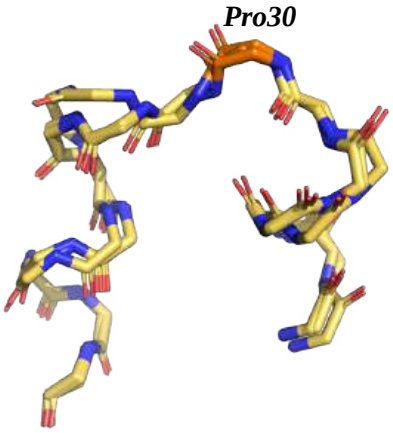

***Figure 81**: Superposition of the representative structures from the three clusters of D30P mutant (monomer B). Only the backbone atoms are shown, with proline residue labeled and highlighted in orange.*

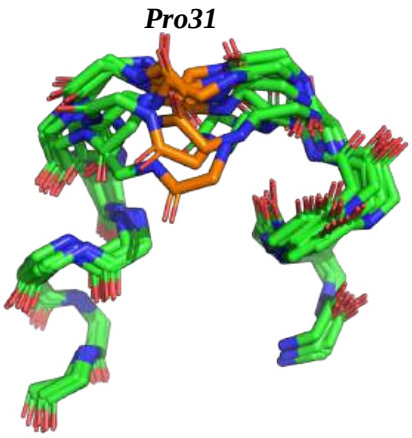

***Figure 82**: Superposition of the representative structures from the nine clusters of A31P mutant (monomer B). Only the backbone atoms are shown, with proline residue labeled and highlighted in orange.*

In both the native Rop and the D30P variant, there are no significant energetic deviations among conformations, leading to their distribution into only two or three clusters. Nevertheless, the native protein exhibits the most highly conserved set of conformations, as depicted in **Figures 77 & 80**. Each representative structure of native Rop is slightly shifted relative to the others, indicating structural stability and compactness. The protein backbone fluctuates within a limited spatial range, resulting in the formation of a small number of clusters, with minor deviations. These results are consistent with the experimental data, reproducing properly the experimentally confirmed high stability of the native form.

The D30P conformations are distributed among three clusters, concluding to three representative structures, as shown in **Figures 78 & 81**. They appear to be closely related to each other rather than a segment in the turn region which distinguishes each conformation. These observations are in agreement with the experiment, as the point mutation affect slightly the stabilization of the overall structure. Therefore, compatible to the previous analyses, the limited deviation further supports the stability of this mutant.

On the other hand, the A31P conformations (**Figures 79 & 82**) occupying multiple distinct energetic states exhibiting greater deviations. The point mutation site distinguishes among each representative structure causing rearrangement of the turn region and concluding to a broadened range of different conformations. The existence of multiple energetic states reflects the flexibility within the system which is compatible with the already taken results as the A31P variant presents the most unstable structure.

Overall, this illustration helps visualize what the dihedral PCA analysis has already shown in the corresponding plots.

3.4 Cartesian PCA of the Dominant Cluster

Isolating the dominant cluster from the dPCA analysis of both Turn A and Turn B, the corresponding frames were subsequently used to perform a cPCA analysis, aiming to examine the impact of the proline insertion on the entire structure rather than focusing solely on the turn region. Therefore, the α-helices which are the steady part of the structures, remained fitted in order to evaluate the flexibility caused by each mutation.

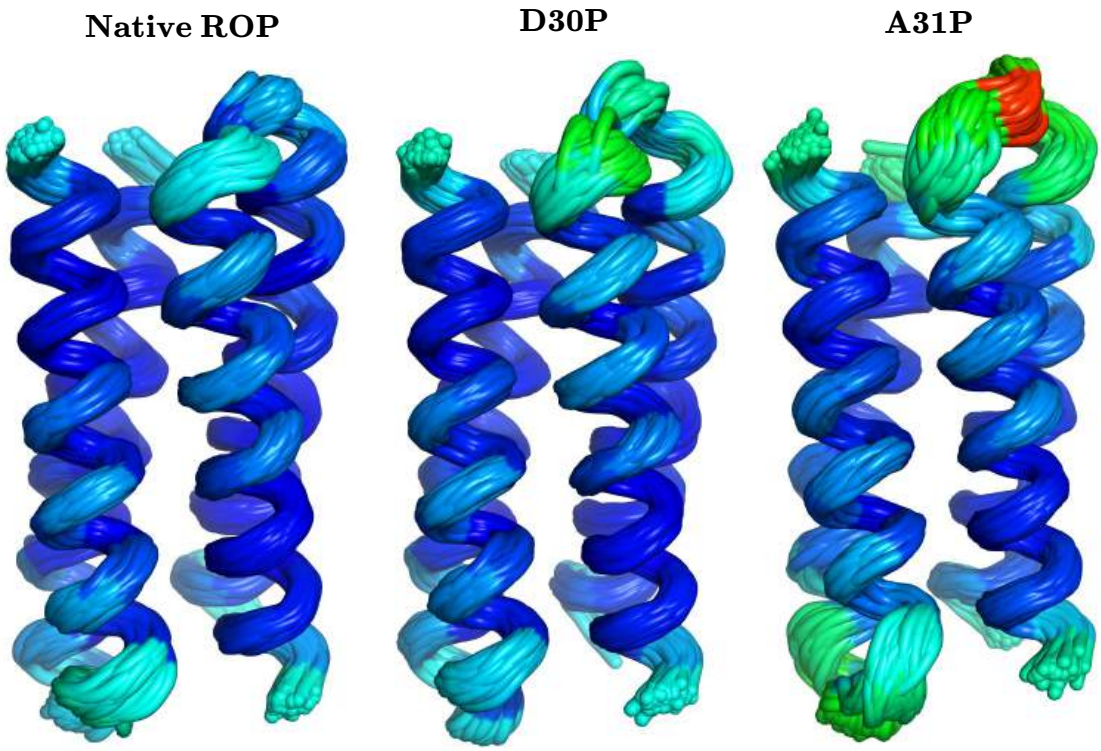

Figure 83: *Superposition of all conformations of the entire structures (excluding C-terminals) belonging to the dominant cluster derived from the dPCA analysis of Turn A for native Rop, D30P and A31P. Structural alignment was performed on the α-helical regions. The most flexible segments are shown in red, with the color gradually transitioning to blue according to structural stability. The color scale is common for all illustrations (Fig 83 & 84).*

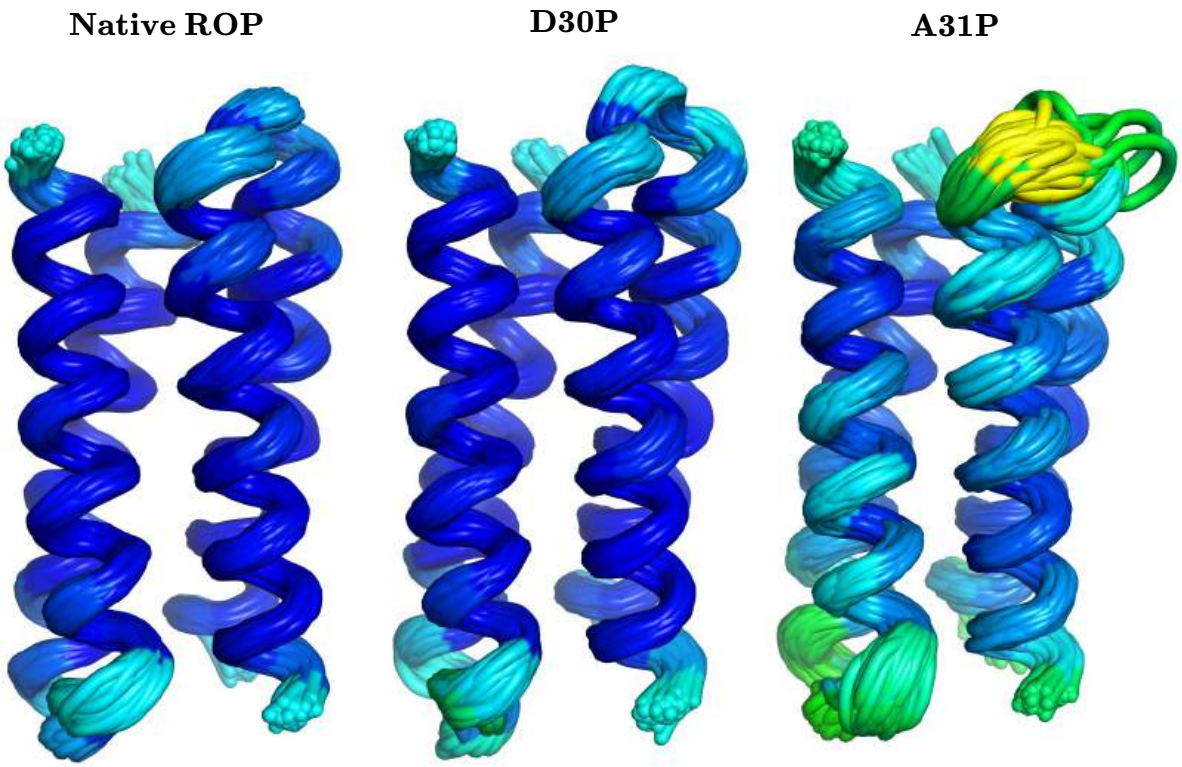

***Figure 84**: Superposition of all conformations of the entire structures (excluding C-terminals) belonging to the dominant cluster derived from the dPCA analysis of Turn B for native Rop, D30P and A31P. Structural alignment was performed on the α-helical regions. The most flexible segments are shown in red, with the color gradually transitioning to blue according to structural stability. The color scale is common for all illustrations (Fig 83 & 84).*

All three trajectories are characterized by a common steady part: α-helices. In both variants, the point-mutation does not significantly affect the folding of the helices. This region demonstrates the lowest RMSF values, while flexibility increases in the turn regions, even in the native state.

The D30P variant exhibits only a minor structural deviation at the mutation site, especially in the representation of **Figure 83**. Immediately after the turn region finishes, the structure is once more associated with low RMSF values, which indicates structural stability. This variant adopts a series of conformations compatible to the native form, as it

depicts in **Figure 84** which are observed only minor discrepancies between the two structures.

On the other hand, the A31P mutant is characterized by more extended partial unfolding in the turn region, particularly at the point-mutation site of Turn A, where the highest RMSF value is observed. In contrast to D30P, the mutation at position 31 seems to influence a broader part of the structure, as the most pronounced discrepancies among conformations of the dominant cluster are detected in both Turn A and Turn B regions.

Moreover, the α-helical segments adjacent to the turns are also affected by this unfolding, as these regions do not immediately adopt their most stable conformations, indicated by their slightly elevated RMSF values. At the same time, as shown in **Figure 84**, all conformations belonging to the dominant cluster from Turn B display increased structural variability in the turn region, capturing a wider range of states without, however, reaching the most extreme RMSF values. Notably, while multiple partially distinct conformations are adopted, one of them is clearly separated from the others beyond the point mutation, forming a loop that is directly connected to the following α-helix with a locally altered orientation.

Overall, the A31P mutant displays consistently elevated RMSF values, even within the helical regions, which indicates increased structural instability.

3.5 Interpretation of the Different Folding Behavior of the D30P vs A31P Mutants

Considering that the substitution of two adjacent residues by proline can affect the topology in such a different way, it arouses interest in a deeper investigation of the structural role of each residue in stabilizing the overall fold of the protein. Focusing on the two replaced residues as well as the previous and the following positions (residues 29,30,31 and 32), a structural comparison between the two mutants was performed in order to understand the cause behind the topological alteration.

3.5.1 Hydrogen bonding interactions of Ala31

Initially and foremost, the most intriguing residue is Ala31, as its replacement constitutes the distinguishing feature of the A31P compared to the native and D30P proteins. Based on this observation, the hydrogen bond connections of Ala31 were studied in order to investigate its contribution to the stability of the α-helical bundle topology.

A hydrogen bond is determined by two parameters:

i) the distance between the donor and acceptor atoms

ii) the angle formed by the donor – hydrogen – acceptor atoms (D–H···A)

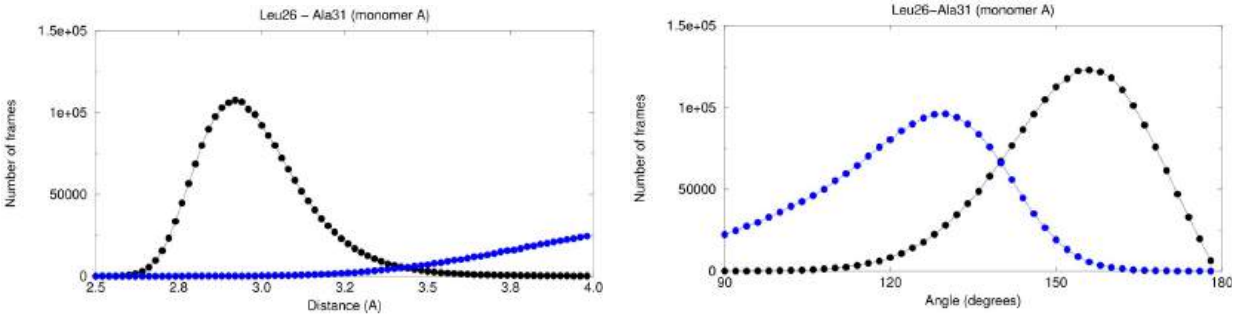

Figure 85*: In the left panel, the distribution of distances between the backbone NH group of Ala31 and the backbone CO group of Leu26 of all conformations of monomer A is shown (black: native Rop, blue: D30P mutant). The right panel illustrates the distribution of the corresponding* N–H···O *angles formed by the same residues.*

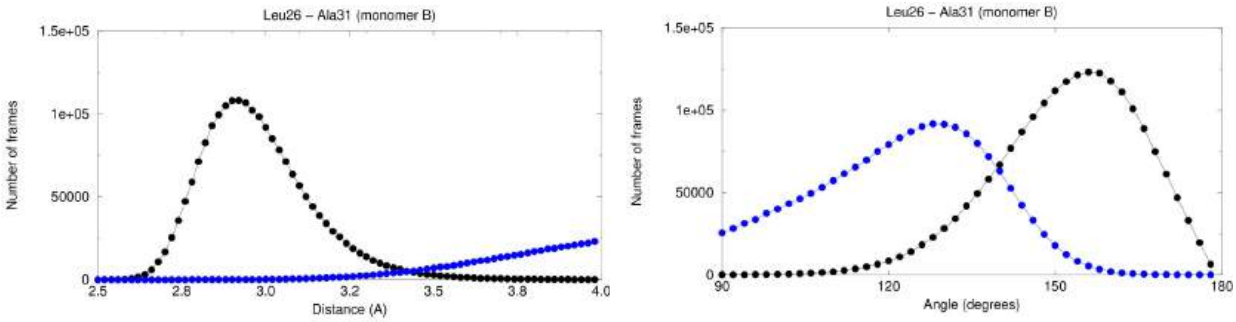

Figure 86: *In the left panel, the distribution of distances between the backbone NH group of Ala31 and the backbone CO group of Leu26 of all conformations of monomer B is shown (black: native Rop, blue: D30P mutant). The right panel illustrates the distribution of the corresponding* N–H···O *angles formed by the same residues.*

Figures 85 and **86** show that the majority of frames of native Rop exhibit a distance of approximately 2.9Å between the NH and CO groups in both monomer A and B, which corresponds to an appropriate distance for a potential strong hydrogen bond. Examination of the angle between these residues reveals values exceeding 150°, indicating that a strong hydrogen bond is formed between the turn region (Ala31) and the first helix of the monomer (Leu26).

The D30P mutant appears to form a weaker hydrogen bond, as the distance between the corresponding atoms is noticeably higher, reaching up to 4.0Å. Additionally, the majority of conformations are concentrated between 120° and 150°, which still corresponds to an acceptable value. Therefore, these observations suggest that the D30P mutant retains a modest hydrogen bond connecting the turn region with the first helix.

In addition, the formation of a simultaneous hydrogen bond between the turn region (Ala31) and the second helix (Ala35) was also examined.

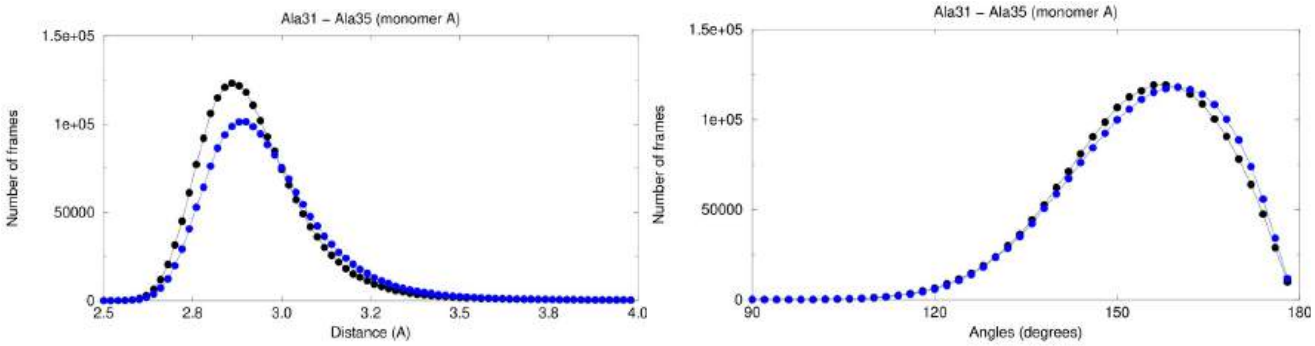

Figure 8***7****: In the left panel, the distribution of distances between the backbone CO group of Ala31 and the backbone NH group of Ala35 of all conformations of monomer A is shown (black: native Rop, blue: D30P mutant). The right panel illustrates the distribution of the corresponding* N–H···O *angles formed by the same residues.*

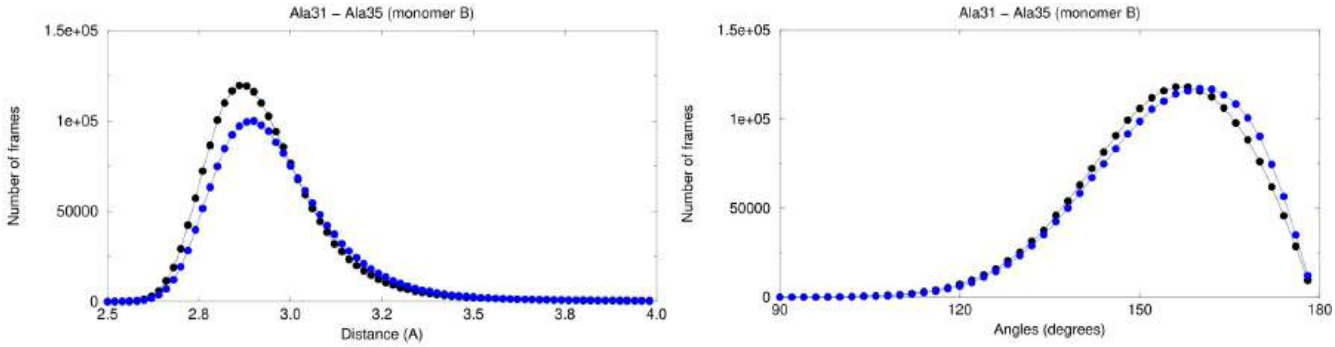

Figure 8***8****: In the left panel, the distribution of distances between the backbone CO group of Ala31 and the backbone NH group of Ala35 of all conformations of monomer B is shown (black: native Rop, blue: D30P mutant). The right panel illustrates the distribution of the corresponding* N–H···O *angles formed by the same residues.*

Figures 87 and **88** illustrate that in the majority of native conformations the appropriate backbone atoms of the alanine residue in the turn region (Ala31), as well as those of the second helix (Ala35), are close enough to form a strong hydrogen bond (approximately 2.9Å). At the same time, the corresponding angle formed by these atoms exceeds 150° in most frames. These observations indicate that these residues are connected through a strong hydrogen bond.

The corresponding curve representing the total number of conformations of the D30P variant is slightly shifted to the right. The peaks of the blue curves indicate that the hydrogen bond not only exists but also remains strong. In the majority of frames the distance remains below 3.0Å, while the angles fall within the range 120° and 150°, exhibiting an appropriate geometry for hydrogen-bond interaction.

In conclusion, the Ala31 residue plays a significant role in stabilizing the native structure. By forming two strong hydrogen bonds simultaneously with both α-helices, it contributes to the proper stabilization of the native fold of the Rop protein. Even in the case where the local environment is slightly altered (D30P variant), Ala31 preserves a strong interaction with the alanine residue of the second helix, while the connection to the first helix is reduced but still remains present. At this point, it has been suggested that its absence induces the deformation of these two stabilizing interactions, which may contribute to the increased flexibility of the A31P variant.

3.5.2 Ramachandran plots for residues 29-32

In order to investigate how each of these mutations influences the spatial arrangement of the polypeptide chain, Ramachandran plots were constructed to visualize the backbone dihedral angles for each residue. These plots allow the examination of how the substitution of proline at positions 30 and 31 may disrupt the topology of the protein. The phi (φ) angle is the dihedral angle around the N – C_α bond and is defined by the atoms $C_{(i-1)} - N_{(i)} - C_{\alpha(i)} - N_{(i+1)}$, whereas the psi (ψ) angle corresponds to the dihedral angle around the C_α – C bond which is defined by the atoms $N_{(i)} - C_{\alpha(i)} - C_{(i)} - N_{(i+1)}$. As the definition of these angles involves atoms belonging to the adjacent residues, it is important to also take into account the residues preceding and following the substituted positions. Therefore, residues 29 and 32 were included in the analysis.

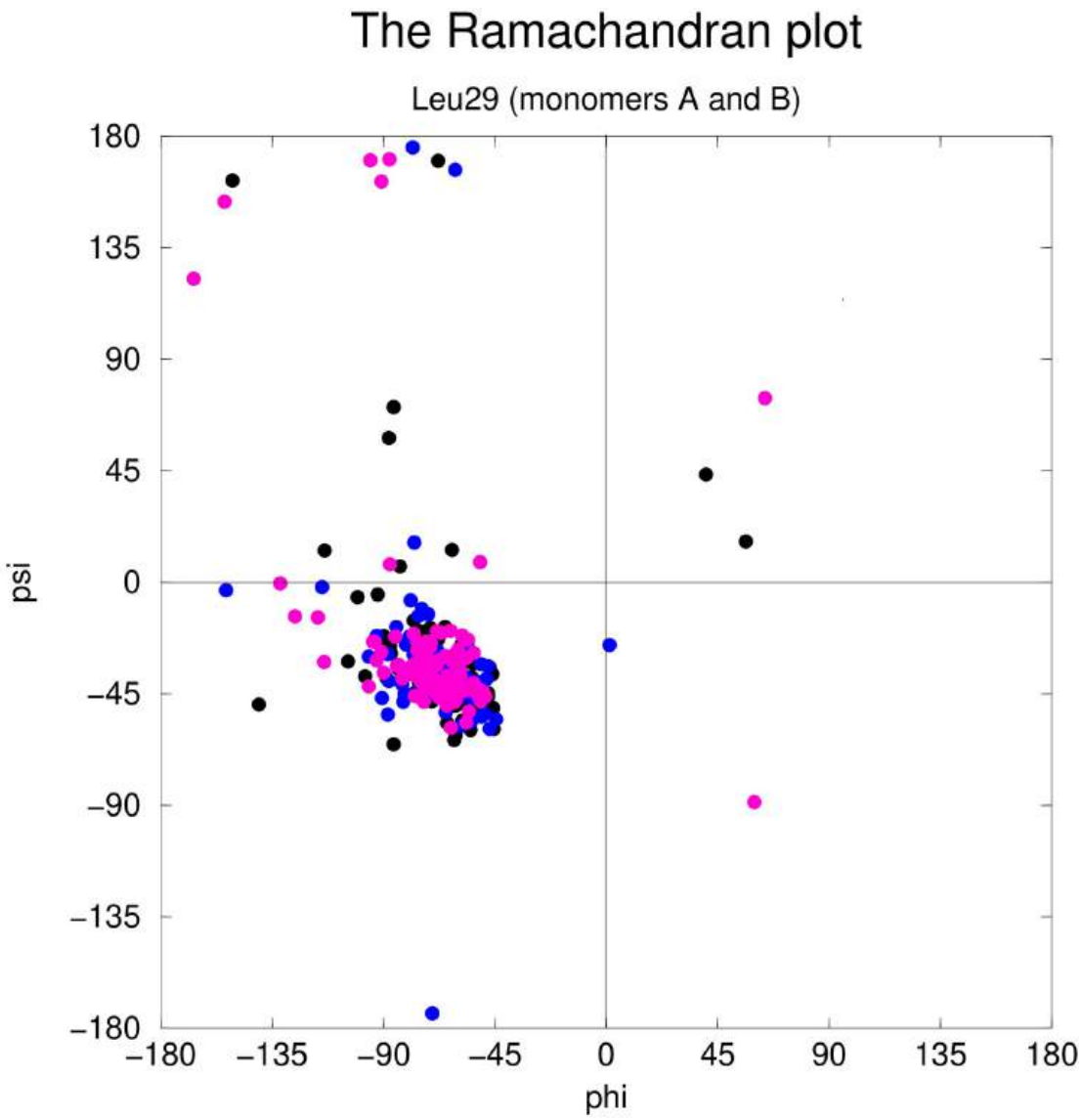

Figure 89: *Ramachandran plot representing all pairs of φ and ψ angles of residue Leu29 in both monomers A and B, belonging to all conformations of native Rop (black), D30P (blue) and A31P (magenta) mutants.*

Figure 89 represents the distribution of the dihedral angles of all possible conformations adopted by Leu29. The majority of conformations are concentrated within the range of φ from -90° to -45°, while ψ fluctuates around -45°. These values correspond to the α-helical region, which is expected since the Rop protein adopts a helical topology. Native and D30P structures exhibit a similar distribution pattern of φ/ψ angles for

Leu29, suggesting that the backbone conformation of the D30P variant remains native-like. On the other hand, the A31P mutant also presents a concentration in the α-helical region but exhibits a broader dispersion. In particular, five distinct outliers reach the highest ψ values, corresponding to the β/extended regions, indicating increased backbone flexibility.

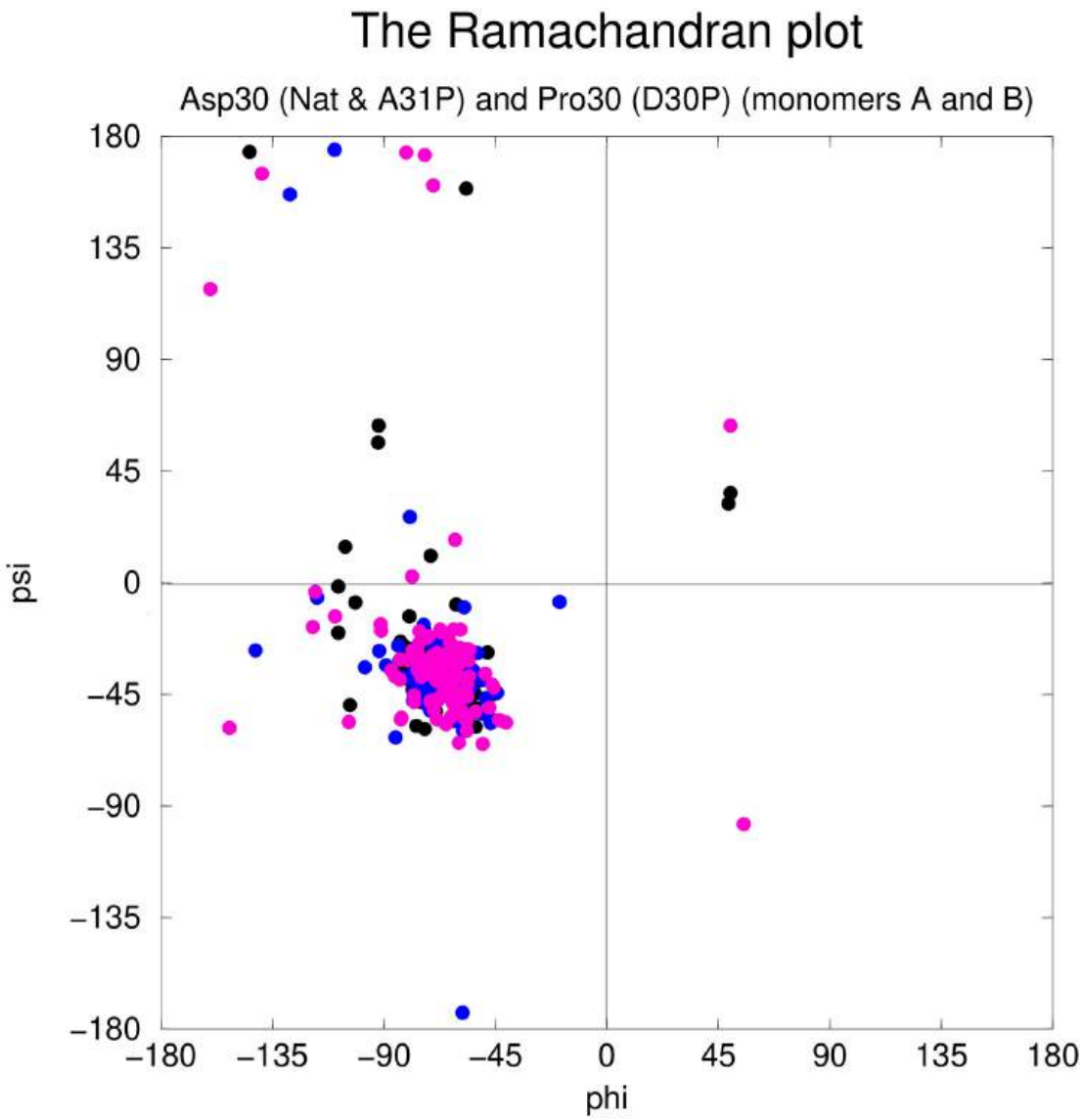

Figure 90: *Ramachandran plot representing all pairs of φ and ψ angles of residues Asp30 and Pro30 in both monomers A and B, belonging to all conformations of native Rop (black), D30P (blue) and A31P (magenta) mutants.*

Similar to position 29, **Figure 90** illustrates that the majority of Asp30 as well as Pro30 conformations present an α-helical geometry. Although proline exhibits structural constrains due to its cyclic side chain, small deviations from the expected region can still be observed. The distribution of φ/ψ angles of Asp30 in the A31P variant remains broadened, reinforcing the previous observation regarding increased backbone plasticity.

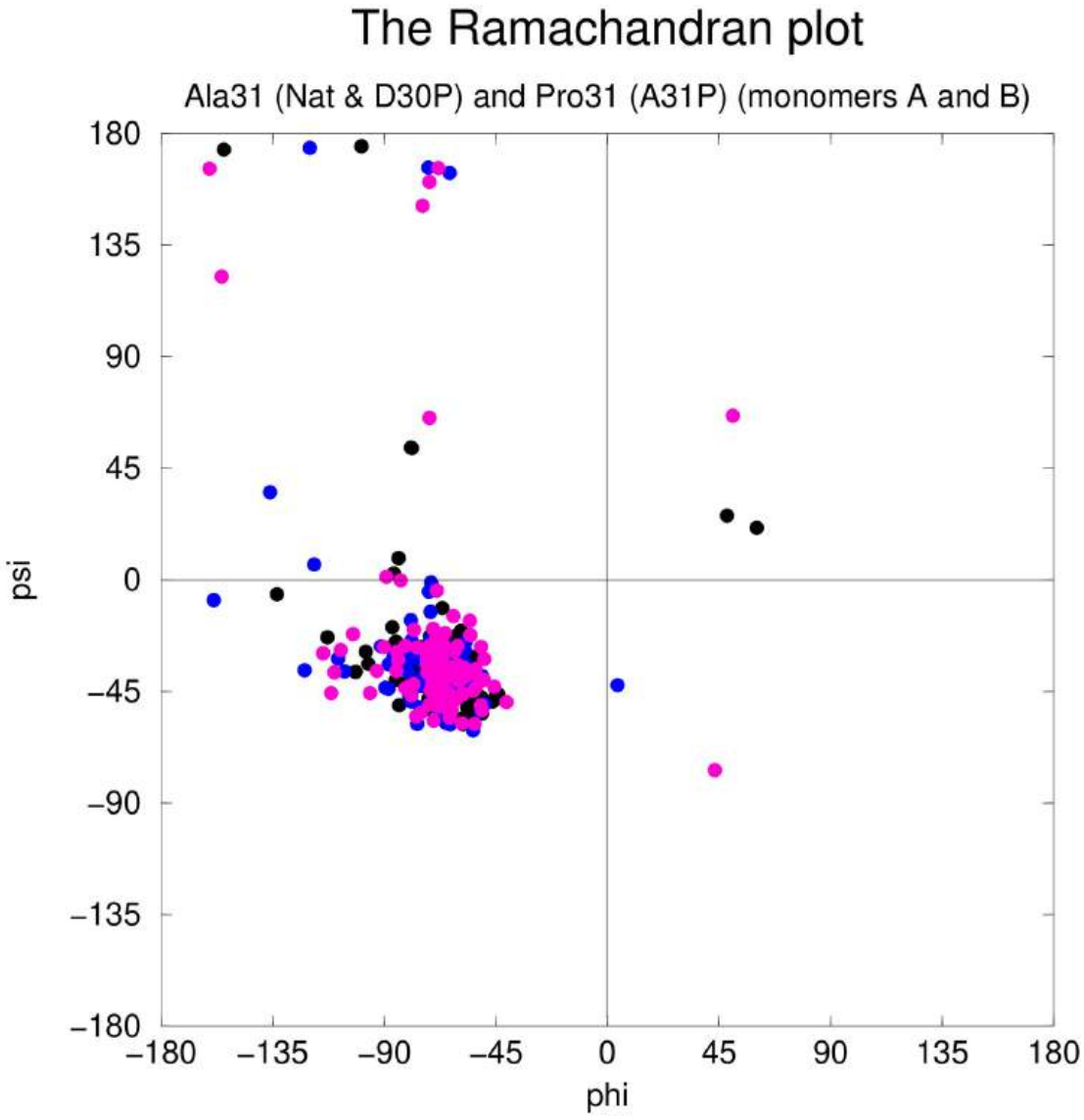

Figure 91: *Ramachandran plot representing all pairs of φ and ψ angles of residues Ala31 and Pro31 in both monomers A and B, belonging to all conformations of native Rop (black), D30P (blue) and A31P (magenta) mutants.*

Residue 31 mostly adopts a helical conformation. However, some distinct pairs of dihedral angles extend to the edges of the Ramachandran plot, revealing the occasional existence of unfolded states of the turn region during the MD simulation. Ala31 in the native Rop exhibits the narrowest distribution, suggesting the most compact backbone geometry. In contrast, the corresponding residue in the D30P mutant displays a few frames deviating from the helical region, reflecting a slight local increase in flexibility caused by the adjacent substitution. Consistently, the A31P variant exhibits the highest number of outliers, further supporting its structural instability.

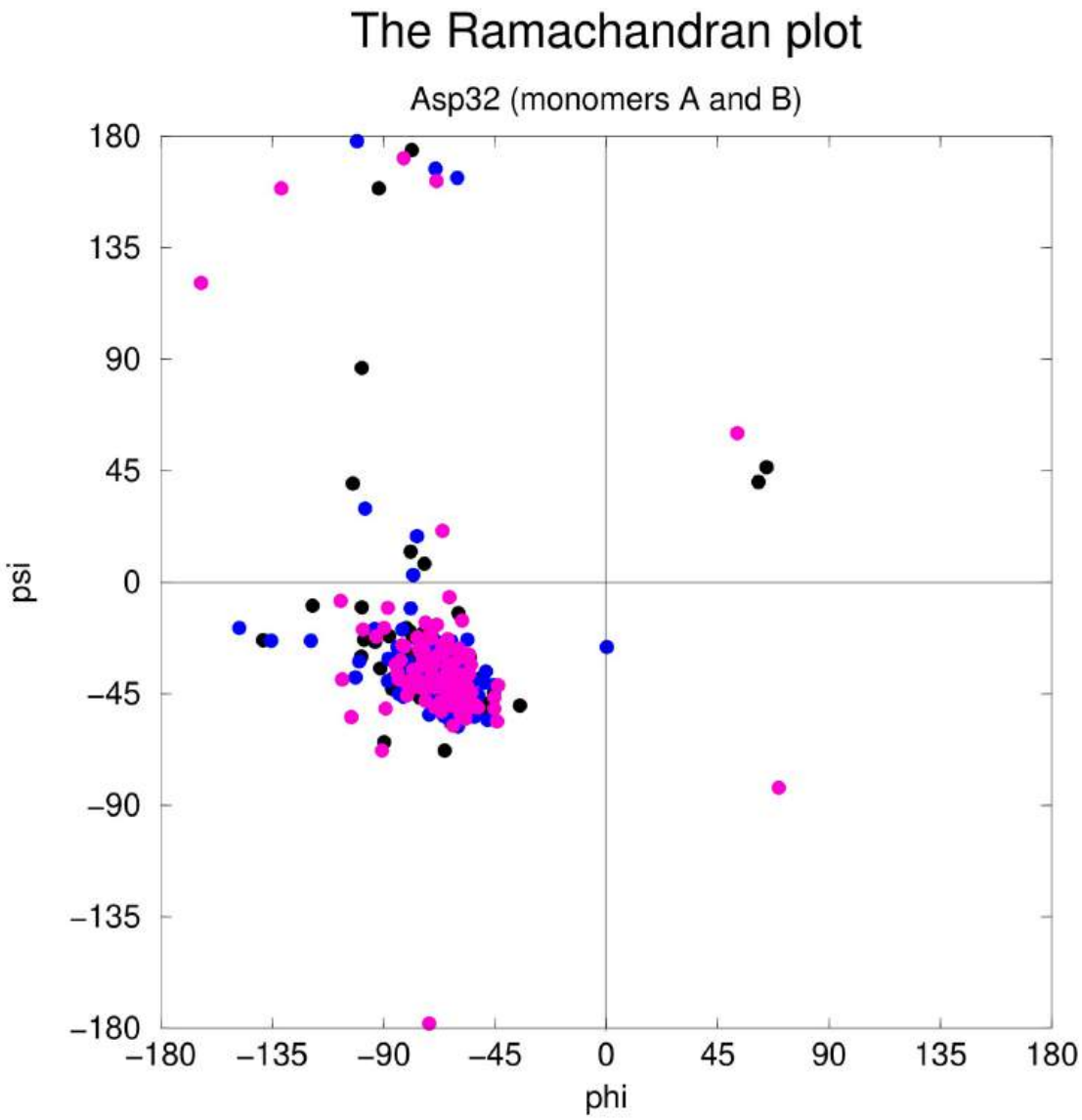

Figure 92: *Ramachandran plot representing all pairs of φ and ψ angles of residue Asp32 in both monomers A and B, belonging to all conformations of native Rop (black), D30P (blue) and A31P (magenta) mutants.*

Regarding the last residue of the examined segment, Asp32 exhibits a similar pattern. The distribution of the A31P variant extends towards the most permissible areas of the plot, indicating increased backbone flexibility. However, the majority of conformations still remain within the α-helical region. Overall, these observations are consistent with those obtained by the previously analyzed residues.

In conclusion, the substitution of proline at two adjacent positions causes a different impact on the folding process of each mutant. Although the two positions are adjacent to each other, this difference may arise from the structural properties of the replaced residues. As demonstrated by the hydrogen bond network of Ala31, this residue plays an important role in stabilizing the native structure. Its absence leads to destabilization due to the inability of the proline ring to replace the structural role of alanine at this key position.

4. Discussion – Conclusions

The present thesis constitutes a comparative computational study of the D30P and A31P mutants. We analyzed three trajectories derived from Molecular Dynamics simulations, including the native structure. The analysis of the native form was an essential step throughout this study, as it was our reference point for evaluating the structural impact of the mutation in each case.

Experimental results already indicate the impact of substituting a proline residue at position 30 (D30P mutant) and 31 (A31P mutant). In case of D30P variant, it has been concluded that it adopts a structure compatible to the native state, known as “native-like”. Despite the insertion of the proline ring in the turn region, the initial form of the 4-α-helical bundle remains unaltered, maintaining the overall architecture [28].

On the other hand, the replacement of the alanine residue by the proline ring in the adjacent position in the amino acid sequence has caused a significant alteration in the overall topology [24][30]. The initially left-handed, all antiparallel 4-α-helical bundle transformed into a right-handed, mixed parallel and antiparallel bundle, known as “bisecting U” topology. This rearrangement leads to a reorganized hydrophobic core characterized by a thermodynamically less stable organization compared to the native form. Crystallographic studies confirm the experimental findings about the different topology adopted by the A31P mutant [28] [42].

Based on these findings, this study aims to examine the ability of Molecular Dynamics simulations in reproducing the experimentally observed conformations. Through the application of multiple structural analyses on the three trajectories, we investigate the folding behavior of

these mutants computationally, while also attempting to explain the structural discrepancy observed between them.

The RMSF distribution indicates the boundaries between the α-helical segments and the turn regions, while also revealing the highly flexible C-terminal residues. Notably, the defined turn region was retained unchanged in all following analyses, in agreement with previously reported studies [28].

An RMSD analysis aimed at evaluating the overall stability of each system. Comparing the entire structures with and without terminal residues, it became clear that their presence introduces additional instability. For this reason, they were excluded to achieve a more reliable evaluation of structural stability. Both the analysis of the entire structures and the separate examination of the turn regions, reveals a hierarchical relationship among the three proteins. Specifically, the native structure exhibits the lowest RMSD values, followed by the D30P mutant, showing slightly higher deviations, suggesting a conformation similar to the native. In contrast, the A31P variant presents substantially higher deviations, indicating increased structural flexibility. These findings are in agreement with the available experimental data [28].

Furthermore, RMSD histograms are meaningful revealing the distribution of sampled conformations. The native Rop exhibits a narrow distribution, indicating limited conformational variability and, therefore, a stable structure. The RMSD values of D30P appear to be slightly shifted to the right, supporting its native-like character. While at first glance the A31P histogram deviates significantly from the others, appearing broader and more extended, this observation is attributed to the larger number of frames in the A31P trajectory. This discrepancy is addressed by normalizing the distributions to a maximum of 100 and adjusting the number of frames to allow direct comparison among the proteins. These

results, along with the statistical analysis performed using the R package sn, confirm the previously described hierarchical stability pattern. In particular, the comparison between skewed and non-skewed models indicates that the RMSD data are better described by skewed distributions, as reflected by the higher log-likelihood per observation values.

The conformational landscapes derived by Dihedral PCA reveal distinct clustering patterns for each system. The total number of native Rop conformations are concentrated within a single cluster, indicating limited structural variability throughout the simulation. The landscape of the D30P mutant appears to be broadened, with additional minor clusters surrounding the dominant one. This pattern suggests modest conformational heterogeneity, which is evidence of native-like behavior. On the other hand, the A31P variant landscape is characterized by multiple clusters, indicating the presence of several distinct conformational states, reflecting increased structural flexibility. The comparison between the three and five dimensional representations of the D30P and A31P mutants reveals an in-depth organization of their conformations. While the three-dimensional clustering analysis grouped together a large number of conformations within the same energetic clusters, the higher-dimensional analysis distinguished additional states that are either connected to each other or remain clearly separated.

The representative structure of D30P suggests that this mutant largely follows the native fold, with only deviations at the mutation site. Additionally, the superposition of all conformations from the dominant cluster exhibits comparable RMSF values to the native form, further supporting its native-like folding. In contrast, the representative A31P conformation, presents a clearly pronounced deviation of the turn region, which also influences the compactness of the upper helical segments.

Overall, the alignment of the representative structures and the superposition of the dominant conformations support the experimental findings: the structure of the D30P mutant remains significantly similar to the native fold, whereas the A31P variant deviates from it.

Subsequently, Cartesian PCA suggests that the α-helices constitute the common stable part among the three systems, remaining largely unaffected by the mutations. Specifically, while the D30P mutant exhibits only a minor structural deviation at the mutation site, the A31P shows local rearrangement in the turn, displaying the highest RMSF values.

Finally, these results are particularly intriguing, as proline substitution at adjacent positions causes markedly different impacts on the folding process of the two proteins. In particular, the substitution of Ala31 seems to significantly influence the structure. It is known that the Ala31 has crucial role in the formation of native topology as it participates forming hydrogen bonds with both helices simultaneously [42]. Specifically, the connection to helix A is formed through a hydrogen bond between the backbone NH group of Ala31 (donor) and the CO of Leu26 (acceptor), whereas with the second helix takes place through a hydrogen bond between the CO group of Ala31 (acceptor) with the backbone NH of Ala35 (donor)[26]. Therefore, the replacement of this specific residue destabilizes the hydrogen bonds network, causing alteration of the overall topology. Our study reproduces the experimental results, confirming the formation of both Leu26 – Ala31 and Ala31 – Ala35 hydrogen bonds in the native structure. The D30P mutant appears to retain the interaction between the turn region (Ala31) and the helix of monomer B (Ala35), as the distance and angle formed by these atoms are consistent with a strong hydrogen-bond interaction. Although some conformations also maintain a permissible N–H···O=C distance and angle between Ala31 and Leu26, the majority fall outside the range required for this interaction, suggesting

either a weak or absent hydrogen bond. On the other hand, both hydrogen bonds are absent from the A31P variant, due to the inability of the N atom of Pro31 to participate in this type of interaction.

A further aspect of this interpretation involves the Ramachandran plots of both the substituted and adjacent positions. The distribution of dihedral angles for Leu29 indicates that most conformations are concentrated within the α-helical region, as expected taking into account the helical topology of the Rop protein. Both the native structure and the D30P variant exhibit outliers in regions surrounding the α-helical area. The A31P mutant contains distinct outliers in the upper region of the plot (β-sheet region), indicating an increased tendency toward disrupted conformations. Previous studies have reported that a significant shift in the ψ angle of Leu29 by 169° in the A31P mutant is associated with a reorientation of the helices, altering the overall topology of the protein [42]. According to this, the presence of outliers in A31P may explain the large rearrangements in topology, while the D30P mutant appears to maintain a more compact folding.

According to the Ramachandran plot for position 30, the majority of conformations are located within the α-helical region, supporting the maintenance of the helical topology. However, the distribution of dihedral angles of the A31P mutant is broadened, revealing increased flexibility of the backbone. This observation is consistent with experimental finding, supporting that changes in the φ and ψ angles by 131° and 119°, respectively, may contribute to the reorganization of the A31P topology [42].

Position 31 presents the most pronounced discrepancy among the three systems. In particular, the distribution of the native Rop is significantly narrower, reflecting a compact and stable helical structure. The D30P mutant displays a broader distribution, suggesting a local increase in

flexibility due to the adjacent mutation, while the distribution of the A31P mutant reveals partially unfolded conformations in the turn region. This pattern is in agreement with the key structural role of Ala31, as it contributes to stabilizing the turn region by simultaneously interacting with both helices, forming the native topology. A similar representation is observed at position 32, suggesting that the instability caused by the mutation affects the adjacent residues.

Taken all these together, the general conclusion may be that molecular dynamics simulations appear to reproduce properly the experimental results. Consistent with the experimental results, the native-like D30P mutant adopts a less stable structure, as indicated by its lower melting temperature (58.9°C) compared to the corresponding native value (68.7°C). This difference is reflected throughout the simulation, as the stability of D30P appears reduced relative to the native protein.

Regarding the A31P variant, experimental studies have revealed a less compact structure, exhibiting the lowest Tm value (43.0°C) and a distinct topology known as the bisecting U fold. However, as already mentioned, the present simulation was performed using a hypothetical native-like A31P model carrying the A31P substitution. Therefore, the main question addressed was whether a native-like conformation would be sufficient to explain the experimentally observed stability or instability of these mutants. The results suggest that, despite preserving an overall native-like fold, the hypothetical A31P model displays increased flexibility, reproducing the experimentally observed destabilization.

These findings highlight the critical role of residue 31 in maintaining the stability of the native Rop topology, while, additionally, it is evident that the adoption of the native-like conformation is not sufficient to ensure native-like stability.

Certainly, several questions remain open. Since the present work focused on a hypothetical native-like A31P model, future studies could involve longer simulations using the experimentally determined bisecting U topology of A31P. Such an approach would allow a deeper investigation of the structure of this mutant, including also the role of the hydrophobic core interactions, providing further insight into the different Rop topologies. In addition, the experimental structure could be used to computationally investigate the folding pathways followed by both the native-like and the non-native forms.

5. References

[1] Elliot J Stollar, David P Smith. Uncovering protein structure. *Essays Biochem* ***2020*** *Oct 8;64(4):649-680.* doi:10.1042/EBC20190042

[2] Carl Ivar Branden, John Tooze. Introduction to Protein Structure. 2nd Edition, **1999**, https://doi.org/10.1201/9781136969898

[3] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. *Molecular Biology of the Cell* ***2003*** *Feb;91(3):401.* doi: 10.1093/aob/mcg023

[4] Yajie Meng, Zhuang Zhang, Chang Zhou, Xianfang Tang, Xinrong Hu, Geng Tian, Jialiang Yang, Yuhua Yao. Protein structure prediction via deep learning: An in-depth review. *Front Pharmacol.* ***2025*** *Apr 3:16:1498662.* doi:10.3389/fphar.2025.1498662

[5] C B Anfinsen. Principles that govern the folding of protein chains. *Science* ***1973*** *Jul 20;181(4096):223-30.* doi:10.1126/science.181.4096.223.

[6] Cyrus Levinthal. Are there pathways for protein folding? *Chim. Phys., Vol. 65,* ***1968****, pp. 44–45.* https://doi.org/10.1051/jcp/1968650044

[7] M Karplus. The Levinthal paradox: Yesterday and today. *Fold Des* ***1997****;2(4):S69-75.* doi:10.1016/s1359-0278(97)00067-9.

[8] R Zwanzig, A Szabo, B Bagchi. Levinthal's paradox. *Proc Natl Acad Sci U S A.* ***1992*** *Jan 1;89(1):20-2.* doi:10.1073/pnas.89.1.20

[9] Martin Karplus. Molecular dynamics simulations of biomolecules. *Acc Chem Res.* **2002** Jun;35(6):321-3. doi:10.1021/ar020082r

[10] Dmitry N Ivankov, Alexei V Finkelstein. Solution of Levinthal's Paradox and a Physical Theory of Protein Folding Times. *Biomolecules* ***2020*** *Feb 6;10(2):250.* doi:10.3390/biom10020250

[11] J M Yon. Protein folding: a perspective for biology, medicine and biotechnology. *Braz J Med Biol Res.* ***2001*** *Apr;34(4):419-35.* doi:10.1590/s0100-879x2001000400001

[12] Michael P. Allen. Introduction to Molecular Dynamics Simulation. *Computational Soft Matter: From Synthetic Polymers to Proteins, Lecture Notes, Norbert Attig, Kurt Binder, Helmut Grubmuller, Kurt Kremer (Eds.), John von Neumann Institute for Computing, Julich, NIC Series, Vol. 23, ISBN 3-00-012641-4, pp. 1-28, 2004.* http://www.fz-juelich.de/nic-series/volume23

[13] Scott A Hollingsworth, Ron O Dror. Molecular dynamics simulation for all. *Neuron.* ***2018*** *Sep 19;99(6):1129-1143.* doi:10.1016/j.neuron.2018.08.011

[14] Kenno Vanommeslaeghe, Olgun Guvench, Alexander D MacKerell Jr. Molecular mechanics. *Curr Pharm Des.* ***2014****;20(20):3281-92.* doi:10.2174/13816128113199990600

[15] Pedro E M Lopes, Olgun Guvench, Alexander D MacKerell Jr. Current status of protein force fields for molecular dynamics simulations. *Methods Mol Biol.* **2015**:1215:47-71. doi:10.1007/978-1-4939-1465-4_3

[16] Zhigeng Jing, Chengwen Liu, Sara Y. Cheng, Rui Qi, Brandon D. Walker, Jean-Philip Piquemal. Polarizable force fields for biomolecular simulations: Recent advances and applications. *Annual Review of Biophysics Volume 48*, **2019**, 371–394. https://doi.org/10.1146/annurev-biophys-070317-033349

[17] Linda Truebestein, Thomas A Leonard. Coiled-coils: The long and short of it. *Bioessays* ***2016*** *Sep;38(9):903-16.* doi:10.1002/bies.201600062 Epub 2016 Aug 5

[18] Andrei N Lupas, Jens Bassler, Stanislaw Dunin-Horkawicz. The Structure and Topology of α-helical coiled coils. *Subcell Biochem.* ***2017****:82:95-129.* doi:10.1007/978-3-319-49674-0_4

[19] Elise A Naudin, et al. From peptides to proteins: coiled-coil tetramers to single-chain 4-helix bundles. *Chem Sci.* 2022 Sep 20;13(38):11330–11340. doi:10.1039/d2sc04479j

[20] Derek N Woolfson. Understanding a protein fold: The physics, chemistry, and biology of α-helical coiled coils. *J Biol Chem.* ***2023*** *Apr;299(4):104579.* doi:10.1016/j.jbc.2023.104579

[21] Aikaterini Kefala, et al. Probing Protein Folding with Sequence-Reversed α-Helical Bundles. *Int J Mol Sci.* ***2021*** *Feb 16;22(4):1955.* doi:10.3390/ijms22041955

[22] Maria Amprazi, et al. Structural plasticity of 4-α-helical bundles exemplified by the puzzle-like molecular assembly of the Rop protein. *Proc Natl Acad Sci U S A.* ***2014*** *Jul 29;111(30):11049-54.* doi:10.1073/pnas.1322065111

[23] M Helmer-Citterich, M M Anceschi, D W Banner, G Cesareni. Control of ColE1 replication: low affinity specific binding of Rop (Rom) to RNAI and RNAII. *EMBO J. 1988 Feb;7(2):557–566. doi:*10.1002/j.1460-2075.1988.tb02845.x

[24] L Castagnoli, M Scarpa, M Kokkinis, D W Banner, D Tsernoglou, G Cesareni. Genetic and structural analysis of the ColE1 ROP (Rom) protein. *EMBO J.* ***1989*** *Feb;8(2):621-9.* doi:10.1002/j.1460-2075.1989.tb03417.x

[25] P F Predki, L M Nayak, M B Gottlieb, L Regan. Dissecting RNA-Protein Interactions: RNA-RNA Recognition by Rop. *Cell* ***1995*** Jan *13*;80*(1)*:*41–50.* https://doi.org/10.1016/0092-8674(95)90449-2.

[26] D W Banner, M Kokkinidis, D Tsernoglou. Structure of the ColE1 Rop Protein at 1.7A Resolution. *J. Mol. Biol.* ***1987*** Aug *5*;*196(3)*:*657–75.* https://doi.org/10.1016/0022-2836(87)90039-8.

[27] Paul F. Predki, Vishal Agrawal, Alex T. Brünger, Lynne Regan. Amino-Acid Substitutions in a Surface Turn Modulate Protein Stability. *Nat. Struct. Biol.* ***1996***, 3, *54–58.* https://doi.org/10.1038/nsb0196-54.

[28] Vouzina, O.-D., Tafanidis, A., & Glykos, N. M. The curious case of A31P, a topology-switching mutant of the Repressor of Primer protein: A molecular dynamics study of its folding and misfolding. *J Chem Inf Model.* ***2024*** *Aug 12;64(15):6081-6091.* doi:10.1021/acs.jcim.4c00575

[29] Glykos, N. M.; Kokkinidis, M. Meaningful Refinement of Polyalanine Models Using Rigid-Body Simulated Annealing: Application to the Structure Determination of the A31p Rop Mutant. *Acta Crystallogr D Biol Crystallogr.* ***1999*** *Jul;55(Pt 7):1301-8.* doi:10.1107/s0907444999004989

[30] Peters, K.; Hinz, H.-J.; Cesareni, G. Introduction of a Proline Residue into Position 31 of the Loop of the Dimeric 4-α-Helical Protein ROP Causes a Drastic Destabilization. *Biol Chem.* ***1997*** *Oct;378(10):1141-52.* doi:10.1515/bchm.1997.378.10.1141

[31] Glykos, N. M.; Kokkindis, M. Structural polymorphism of a marginally stable 4-α-Helical Bundle. Images of a Trapped Molten Globule? *Proteins.* ***2004*** *Aug 15;56(3):420-5.* doi:10.1002/prot.20167

[32] Glykos, N. M. Software news and updates. Carma: a molecular dynamics analysis program. *J Comput Chem.* ***2006*** *Nov 15;27(14):1765-8.* doi:10.1002/jcc.20482

[33] Koukos, I. P.; Glykos, N. M. Grcarma: A fully automated task-oriented interface for the analysis of molecular dynamics trajectories. *J Comput Chem. 2013 Oct 5;34(26):2310-2.* doi:10.1002/jcc.23381

[34] Martinez, L. Automatic Identification of Mobile and Rigid Substructures in Molecular Dynamics Simulations and Fractional Structural Fluctuation Analysis. *PLoS One.* ***2015*** *Mar 27;10(3):e0119264. doi:* 10.1371/journal.pone.0119264

[35] Lindsay I Smith. A tutorial on Principal Components Analysis. ***2002*** *Feb 2*

[36] Jolliffe, Ian T. *Principal Component Analysis*. 2nd ed, Springer, 2002. Springer Series in Statistics. *BnF ISBN*.

[37] David, Charles C., and Donald J. Jacobs. "Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins." *Protein Dynamics*, edited by Dennis R. Livesay, vol. 1084, Humana Press, 2014, pp. 193–226. *DOI.org (Crossref)*, https://doi.org/10.1007/978-1-62703-658-0_11

[38] Altis, Alexandros, et al. "Dihedral Angle Principal Component Analysis of Molecular Dynamics Simulations." *The Journal of Chemical Physics*, vol. 126, no. 24, Jun. **2007**, p. 244111. *DOI.org (Crossref)*, https://doi.org/10.1063/1.2746330.

[39] Altis, Alexandros, et al. "Construction of the Free Energy Landscape of Biomolecules via Dihedral Angle Principal Component Analysis." *The Journal of Chemical Physics*, vol. 128, no. 24, Jun. **2008**, p. 245102. *DOI.org (Crossref)*, https://doi.org/10.1063/1.2945165

[40] Paul J Turner, Grace Development Team, 1991-2007, Grace(xmgr) computer software Portland. Retrieved from https://plasma-gate.weizmann.ac.il/Grace/

[41] Schrödinger. The PyMOL molecular graphics system, Version 3.1.6.1. computer software. Retrieved from https://pymol.org

[42] Glykos, Nicholas M., et al. "Protein Plasticity to the Extreme: Changing the Topology of a 4-α-Helical Bundle with a Single Amino Acid Substitution." *Structure*, vol. 7, no. 6, June **1999**, pp. 597–603. *DOI.org (Crossref)*, https://doi.org/10.1016/S0969-2126(99)80081-1.

Appendix

Script 1

```
   #!/usr/bin/perl -w

    open ( FILE1, $ARGV[0] )     || die "Can not open
$ARGV[0]\n";
    open ( OUT, ">output.dat" )  || die "Can not open
output.dat\n";

   $pos = 0;

      while ($line = <FILE1>)
   {
      chomp $line;
      @values = split(' ' , $line);

       $x[$pos] = $values[0];
       $y[$pos] = $values[1];
       $pos++;
    }
   $y_max = 0;

      for ( $i = 0; $i < $pos; $i ++ )
    {
          if ( $y[$i] > $y_max )

          { $y_max = $y[$i] }
    }
     $factor = 100 / $y_max;

       for ( $i = 0; $i < $pos; $i++ )

         {
              $scaled_y = $y[$i] * $factor;
              print OUT  "$x[$i] $scaled_y\n";

         }

    close ( OUT );
    close ( FILE1 );
```

Script 2

```
   #!/usr/bin/perl -w

 open ( IN, $ARGV[0] )          || die "Can not open
$ARGV[0]\n";
 open ( OUT, ">output.dat" )    || die "Can not open
output.dat\n";

 $pos = 0;

    while ( $line = <IN> )
 {
      chomp $line;
      @values = split ( ' ', $line );

      $x[$pos] = $values[0];
      $y[$pos] = $values[1];
      $pos++;

 }
    for ( $i = 0; $i < $pos; $i ++ )

  {      $y[$i] /= 2;


      print OUT "$x[$i] $y[$i]\n" ;
  }

  close ( OUT );
  close ( IN );
```