The simulations of the 56-residue protein Gb88 in water at two temperatures

Guido Putignano[a], Lorenzo Tarricone[a]

[a] *Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland*

Abstract

The folding stability of a protein plays a pivotal role in comprehending how a protein behaves within a specific environmental context. In our study, we employed the GROMOS program to conduct simulations of the 56-residue protein Gb88 in a water medium at two different temperatures. This investigation aims to gain insights into the protein's folding dynamics and stability. To ensure the robustness of our findings, we first scrutinized the influence of a meticulous equilibration protocol on the stability of our simulation system. Next, we analyzed the variation of simulation parameters over time under the two distinct temperature conditions. Finally, we conducted a comprehensive assessment of the key changes occurring within the system, encompassing metrics such as RMSD (Root Mean Square Deviation), RMSF (Root Mean Square Fluctuation), hydrogen bonds, and protein conformation. Our overarching objective is to contribute to the field of molecular modeling by advancing our understanding of protein stability.

Keywords: Protein Stability, Protein Folding, Simulation, Biomolecular Systems.

1. Topology

In our case, NMR has been used. In order to check that the structure is derived from NMR, there were two main possibilities. The first was to look at the title "Solution NMR structures of two designed proteins with 88% sequence identity but different fold and function". In this case NMR was already defined. The second possibility is to click on the doi file, to be directed to the paper "NMR structures of two designed proteins with high sequence identity but different fold and function" where it's possible to see a section of NMR Spectroscopy. As a paper is revised by other people, we can be almost sure that NMR has been employed.

Q1 When considering the solution, the solvent is a buffer solution consisting of 100 mM potassium phosphate at pH 7.2 with the addition of 10% D2O (deuterium oxide or heavy water). This buffer solution is used for preparing 15N- and 13C/15N-labeled protein samples at concentrations ranging from 0.15 to 0.3 mM. D2O is often added to NMR (nuclear magnetic resonance) experiments to lock the magnetic field and improve the signal-to-noise ratio for certain types of NMR experiments. The temperature mentioned in the provided information refers to the temperature at which the NMR (nuclear magnetic resonance) spectra were recorded for different samples. It states that: NMR spectra of GA88 and GB88 were recorded at 22°C. NMR spectra of GA95 and GB95 were acquired at 20°C. The pressure is not explicitly mentioned in the paper, but it was possible to obtain from the PDB file and it was ambient temperature (1 atm)

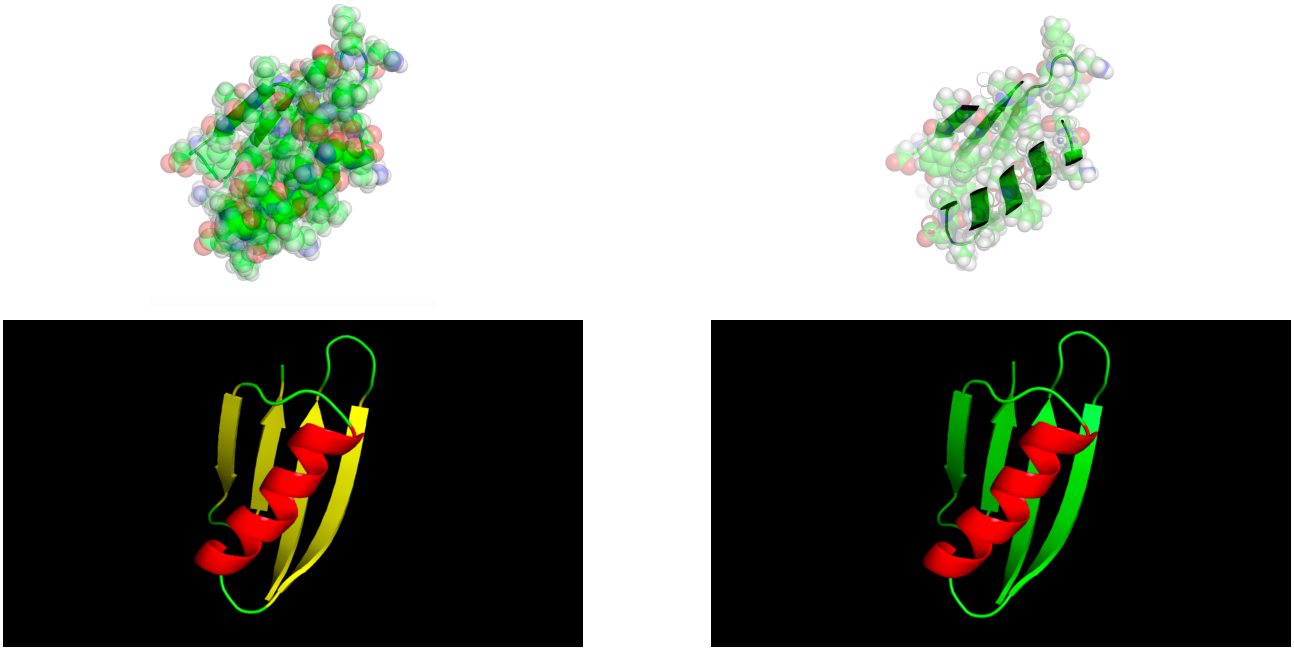


Figure 1: Protein image seen in different ways

NMR analysis, including chemical shift index and NOE data interpretation, revealed the structural organization of GA88. [Q2]It consists of three helices: the first spans residues 9 to 23, the second extends from residues 27 to 34, and the third covers residues 39 to 51. These helices exhibit a compact packing arrangement, where the second helix is antiparallel to both the first and the third.[Q1] In this case there is one main chain as it's possible to see from the image below.

Entity ID: 1

Molecule	Chains	Sequence Length	Organism	Details	Image
Gb88	A	56	synthetic construct	Mutation(s): 0	

Figure 2: Protein structure

The NMR data for the backbone chemical shifts of GB88 reveals the presence of four β-strands and one α-helix, with the secondary structures arranged in the order 1–2—3–4. Detailed analysis of Nuclear Overhauser Effects (NOE) suggests that the β-strands form a four-stranded sheet, with β1 (residues 1–8) and β4 (residues 51–55) serving as the central strands in a parallel arrangement, and β2 (residues 13–20) and β3 (residues 42–46) forming the outer strands in an antiparallel orientation to β1 and β4, respectively. The α-helix (residues 23–36) runs diagonally across the sheet and is connected to β2 via a two-residue loop and to β3 via a five-residue loop.

Modelled Residue Count: 56

The chain of the protein can be obtained from the FASTA file:
TTYKLILNLKQAKEEAIKELVDAATAEKY
FKLYANAKTVEGVWTYKDETKTFTVTE

3-letter Equivalent:

[Q4] NH3+ THR THR TYR LYSH LEU ILE
LEU ASN LEU LYSH GLN ALA LYSH GLU GLU ALA ILE LYSH
GLU LEU VAL ASP ALA ALA THR ALA GLU
LYSH TYR PHE LYSH LEU TYR ALA ASN ALA
LYSH THR VAL GLU GLY VAL TRP
THR TYR LYSH ASP GLU THR LYSH THR PHE
THR VAL THR GLU COO-

Table 1: Protein Structural Information

Parameter	**Value**
Number of Atoms	923
Number of Bonds	930
Number of Model Structures Identified	250 (from the animation of VMD)
Number of Model Structures Submitted	20 (from the animation of VMD)
pH	[Q3] 7 (it's 7.2 in the paper)

In the provided amino acid sequence, we've identified the charges of specific amino acids as follows:

Table 2: Amino Acid Charges

Amino Acid	**Charge**
Lysine (LYSH)	+1
NH3+	+1
COO-	-1
Glutamic Acid (GLU)	-1
Aspartic Acid (ASP)	-1

[Q5] All other amino acids are uncharged at pH 7. Consequently, the net charge (weighting with the abundance of these elements) of the protein is 0.

[Q6] What is possible to see is that the number of atoms goes from 923 to 592. The reasons why it happens are:

1. **Hydrogen Atoms:** PDB files typically exclude hydrogen atoms. However, when creating a molecular topology file, especially for molecular dynamics simulations, hydrogen atoms are often added to the protein structure. This is because hydrogen atoms play a significant role in defining detailed atomic interactions, bond lengths, and angles in force field calculations.

2. **Additional Hydrogen Atoms:** In the case of NMR structures, all hydrogen atoms are typically included, even those that may not be explicitly treated by some force field representations, such as GROMOS (Groningen Machine for Chemical Simulations).

3. **Multiple Structures in PDB Files:** A PDB file may contain multiple structures for the same protein. This is common for NMR structures and can also occur in X-ray structures when different copies of the protein in the crystallographic unit cell are refined independently.

[Q7]The comparative analysis of residue 4 (Lysine) in the Protein Data Bank (PDB) and GROMOS coordinate files reveals notable distinctions. First and foremost, a prominent alteration is observed in the arrangement of atom elements, a phenomenon noted in point 1. In the PDB file, hydrogen atoms are explicitly represented in relation to carbon atoms, while in GROMOS, a united atom representation is employed, resulting in the elimination of hydrogen atoms, to maintain consistency with the united atom model.

This transformation is further substantiated by the observation that nitrogen atoms, specifically Nitrogen N, do not undergo the same united atom representation conversion. In the GROMOS coordinate file, Nitrogen N retains its explicit representation, with the three associated hydrogen atoms (HZ1, HZ2, and HZ3) preserved as individual entities.

[A] Given that there are 918 observables, and considering the simulation involves 923 solute atoms, we can calculate the ratio as follows:

$$\text{Ratio} = \frac{918}{923 \times 3 - 6} \approx 0.332$$

This ratio provides insights into the relationship between the number of observables and the solute atoms in the simulation.

[B] The observable-to-parameter ratio can be determined based on the provided information from the GROMOS molecular topology file. The formula for calculating this ratio is as follows:

$$\text{Observable-to-Parameter Ratio} = \frac{918}{3N' - M' - 6}$$

Where:

N' = Total number of united atoms

M' = Total number of bonds + angles + dihedrals + impropers

Let's calculate the values:

Parameter	Value
N'	592
M'	2186

Now, we can calculate the observable-to-parameter ratio:

$$\text{Observable-to-Parameter Ratio} = \frac{918}{3 \times 592 - 2186 - 6} \approx -2.207$$

This ratio provides an indication of the relationship between the number of observables and the parameters to be determined in the simulation.

[C] Comparison of X-ray Scattering and Neutron Scattering

X-ray Scattering	Neutron Scattering
Detects electron density, primarily for heavier atoms (e.g., C, N, O).	Detects positions of atomic nuclei, including hydrogen.
Insensitive to hydrogen atom positions due to low electron density.	Directly reveals the positions of hydrogen atoms in molecules.

Table 3: Comparison of X-ray and Neutron Scattering Methods

Why Fewer Structures Include Neutron Scattering-derived Hydrogen Atom Coordinates

Challenges in Neutron Scattering	Reasons for Fewer Structures
Experimental Complexity: Requires specialized facilities (nuclear reactors or spallation sources).	Limited access to neutron scattering experiments.
Sample Preparation Challenges: Often involves deuteration (substituting hydrogen with deuterium).	More complex sample preparation compared to X-ray crystallography.
Data Quality and Resolution: Neutron data may have lower resolution and higher background noise.	Challenging to precisely determine hydrogen atom positions.
Cost: Neutron experiments can be more expensive and time-consuming.	Limitation in resources, restricting the number of researchers and institutions.

Table 4: Challenges in Neutron Scattering and Reasons for Fewer Structures

2. Energy Minimization

2.1. Minimisation in the Vacuum

*Q*8 The next part of the work focused on energy minimisation. To make it possible, the algorithm applied in the EM (Energy Minimization) procedure is "steepest descent." The peptide is subjected to energy minimization in a vacuum using the steepest descent algorithm. The minimization process is performed for 20 cycles, each consisting of 2000 steps. *Q*9The potential energy of the system before energy minimization was -2.0788×10^3 kJ/mol and after energy minimization, it was -5.0101×10^3 kJ/mol. In this case, after the energy minimisation there is a decrease of potential Energy and non-bonded energy. *Q*10The exceptionally high potential energy observed in the 2JWU structure within the GROMOS force field can be attributed to two main factors:

1. Differing Force Fields: The structure of the protein, as represented in 2JWU, is generated using a force field that is calibrated differently from the one employed in GROMOS. Force fields are essential for characterizing the potential energy landscape and interactions within a molecular system. When a model relies on a force field calibrated with parameters distinct from those used in GROMOS, it can lead to significant variations in potential energy.

2. Atom Description: During our preprocessing steps, we transitioned from an explicit atom description to an implicit atom representation for some atoms in the protein. This transition involved removing many hydrogen atoms. Consequently, the resulting configuration may not necessarily correspond to the conformation with the lowest potential energy. The absence of explicit hydrogen atoms in the structure can contribute to the observed increase in potential energy.

In summary, the high potential energy in the 2JWU structure within the GROMOS force field can be attributed to both the differences in force field calibration and the transition from explicit to implicit atom descriptions, specifically the removal of hydrogen atoms in the protein structure.

Table 5: Volume Comparison

Property	Value
*Q*11 Volume of the Protein	$30\left(\frac{18}{2}\right)^2 \pi \text{Å}^3$
*Q*12 Volume of the Box	54.9^3Å^3
Space Occupied by the Protein	4.6%

Table 6: Comparison of CPOR Interaction and C-C Bond Force Constant

Parameter	**Value**
Q13Unit of CPOR (CPOR)	kJ/(mol·nm²)
Force Constant for C-C Bond	Value from ifp file

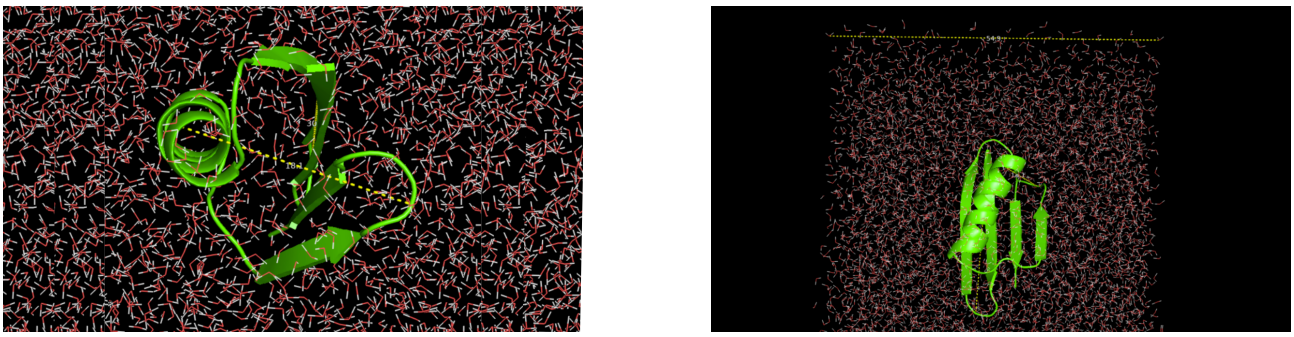


Figure 3: Q11 Representation of the protein in the solution

2.2. Minimisation in the box

When solvating the protein, improper placement and orientation of water molecules can lead to high potential energy. Starting MD simulations under these conditions would quickly raise the temperature at the protein's surface, causing collisions and structural distortion. For this reason, there is the need to have energy minimisation for the second time.

Q14 In GROMOS simulations, the switches NTPORB and NTPORS play a crucial role in controlling how reference positions are managed during the simulation.

NTPORB: When the NTPORB switch is set, it signifies that the reference positions are obtained from a separate file. In this context, the documentation specifies that these reference positions are read from the file "*sim_box_peptide.rpr*" (or "*protein_box.rpr*" in the specific case). Reference positions are typically used in simulations to establish the initial coordinates and constraints for atoms in the system.

NTPORS: Setting NTPORS equal to 0 indicates that the program will not make any adjustments to the reference positions during pressure scaling operations. This means that the reference positions specified in the input file will remain constant throughout the simulation, even if pressure scaling operations are carried out. This can be particularly important when you need to maintain fixed reference positions and prevent them from changing during the course of the simulation.

^{D} The utility of minimization in vacuum can be discussed and we here summarize the arguments in favor of the possible positions

Position	Arguments
The first EM step is actually not needed	• Simplification: The first EM step in vacuum may be redundant for determining the protein's structure in its native environment (water). • Marginal Impact: The influence of the first EM step on final results may be negligible in some cases, depending on the study's focus.
It might be better to skip the first EM	• Resource Efficiency: Skipping the first EM step can save computational resources.
It does not really matter if you skip or not the EM	• Benchmarking: It serves as a reference for comparing the effects of the protein's environment on its structure. • Stability: The first EM step ensures a well-formed initial protein structure, minimizing clashes and providing a stable starting point for subsequent simulations.

3. Thermalisation and Equilibration

The primary goal of thermalisation and equilibration is to release the protein into the simulation environment only after achieving a state of well-equilibrated system conditions at the designated target temperature. This approach ensures that the subsequent production simulation commences from a protein structure that closely aligns with the experimentally derived conformation, mitigating the likelihood of initial structural distortions. [Q15] The decision to independently couple the protein and solvent to distinct heat baths is based on the physical principles governing molecular dynamics simulations. In practice, the solvent experiences more significant fluctuations in temperature, often referred to as cutoff noise, while the exchange in kinetic energy occurs relatively slowly. When both the solute (protein) and solvent are coupled to a shared thermostat, the average system temperature may appear accurate; however, the solute can be colder than the solvent. This phenomenon is recognized as the 'hot-solvent/cold-solute effect.' To address this issue and ensure a more realistic simulation, it is advantageous to employ separate heat reservoirs for the protein and solvent components

Appropriate Combination in a Solution

[E] In the course of thermalizing the system, we undertook the elimination of both translational and rotational motions of the box by imposing a non-spatial constraint multiplier (NSCM) value of -1000. Concurrently, we established a minimum number of degrees of freedom (NDFMIN) at 3. It is worth noting that these selections present an inherent inconsistency, as we are subtracting only 3 from the total degrees of freedom for temperature computation, despite effectively nullifying 6 degrees of freedom by restricting the box from translation and rotation.

The exclusion of rotational degrees of freedom is particularly noteworthy, as the periodic boundary conditions render the rotation of the cell interdependent with other degrees of freedom. To achieve coherence in our parameter choices, two viable alternatives emerge: either adjust NDFMIN to 6 while maintaining NSCM unchanged, or keep NDFMIN unaltered while setting NSCM to a value greater than zero, selectively eliminating only translational degrees of freedom. This ensures a harmonious alignment between the specified parameters and the physical constraints imposed on the system

When considering a change in temperature, it's important to analyse:

1. **System Temperature vs. Time:** The temperature gradually increases during equilibration, signifying successful temperature control to achieve the desired target value.

2. **Total Energy, Kinetic Energy, and Potential Energy vs. Time:** [Q16] The second plot showcases the total energy, total kinetic energy, and total potential energy over time. As the system temperature rises, the kinetic energy

increases proportionally, indicating a higher degree of molecular motion. This behaviour adheres to fundamental principles governing temperature and kinetic energy. Concurrently, the total potential energy also shows an upward trend. This phenomenon is attributed to the increased molecular motion, which results in deviations from equilibrium positions and, consequently, an elevation in potential energy. These observations align with the expected behaviour during equilibration.

Q^{17} If we had to formulate some hypothesis about the further development of the systems at the two different temperatures, we would say that for the simulation at 298K the reached values of potential and kinetic energy would remain the same (accounting for small fluctuations) while for the simulation at 348K I would expect the kinetic energy to almost constant (no increase in temperature) while the potential energy would increase. I expect the increase in potential energy because the protein at denaturation temperature could unfold and therefore change its conformation. This would lead to an increase in potential energy

4. Molecular Dynamics Simulation

Q^{18} The last step is to obtain the molecular dynamics simulation, In this case, there are several differences between the ‘md.imd‘ and ‘equilibration.imd‘ files are as follows:

- In ‘equilibration.imd‘, SHAKE is initialized for both coordinate and velocities, while in ‘md.imd‘, it is not.
- The length of the simulation is 250,000 steps in ‘md.imd‘ compared to 10,000 steps in ‘equilibration.imd‘, with both using the same time steps.
- In the ‘md‘ simulation, the temperature is set to either 298 K or 348 K, while in the ‘equilibration‘ simulation, the temperature is set to 0 K.
- Constraints are not applied in the ‘md‘ simulation, whereas they are in the ‘equilibration‘ simulation.
- The algorithm for pairlists is set to 1 in the ‘md‘ simulation and 0 in the ‘equilibration‘ simulation.
- The ‘NSHAPE‘ value in the NONBONDED block is 3 in the ‘md‘ simulation and -1 in the ‘equilibration‘ simulation.
- Coordinate and energy trajectories are written every 250 steps in the ‘md‘ simulation and every 100 steps in the ‘equilibration‘ simulation.
- The ‘equilibration‘ simulation includes additional blocks such as ‘PROTIONRES‘ and ‘COVALENTFORM‘ that are not present in the ‘md‘ simulation.

Q^{19} The total duration of the simulation is 1 nanosecond (ns), split into two jobs, resulting in each job having a duration of 500 picoseconds (ps). To estimate the time required for this simulation, we can refer to a previous simulation: a 140 ps simulation took 14 hours to complete. Given this reference, we can anticipate the current simulation to last approximately 50 hours.

Q^{20} Concerning the length of each job, both jobs are set to 500 ps for each of the two experimental conditions (temperatures) simulated. With two jobs and a total simulation length of 1 ns, we write the spatial and energy coordinates every 250 steps. Since each step simulates a time interval (δt) of 0.002 ps, the duration of each step is 0.5 ps.

5. Analysis

The analysis of our simulations can help understand more about the evolution of the protein over a long timescale

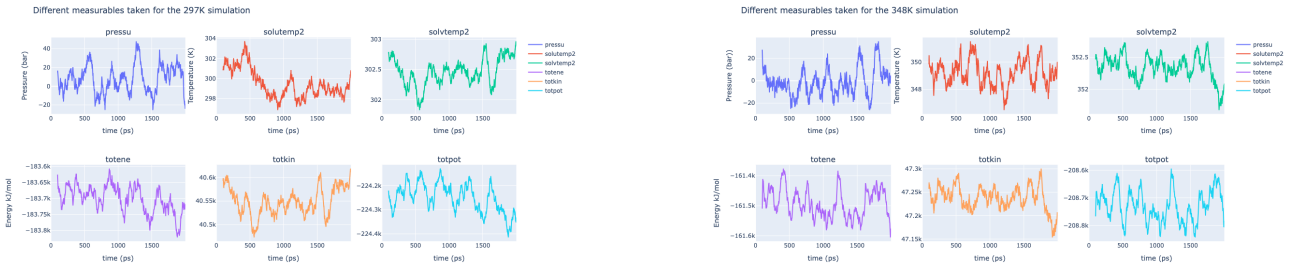


Figure 4: Representation of the protein in the solution

^{Q21} Upon analyzing the time series data for the six quantities, it appears that, in general, these quantities exhibit fluctuations around relatively stable values. However, there are some notable observations:

1. Total Energy: the total energy shows a gradual and consistent decrease in both experimental conditions (even if this change throughout the whole simulation is on the order of $200KJ/mol$ with respect to an absolute value of $\approx 180000KJ/mol$ and therefore might not be meaningful)

2. Solvent and Solute Temperature: The solute reaches the temperature imposed by the heart reservoir quite fast (in less than 1 ns for the 298K experimental condition and in less than 200 ps for the 348K one), while the solvent temperature stays higher than the temperature imposed by the reservoir

These observations suggest that while the majority of quantities fluctuate around stable values, certain changes in the total energy and solvent temperature hint at dynamic processes within the system. Further investigation is required to understand the underlying causes of these trends and their implications on equilibrium. ^{Q22} In both cases, it is evident that the solute temperature (red plot) exhibits larger fluctuations compared to the solvent temperature (green plot).
^{Q23}

- There is a more significant oscillation in pressure in the 298 K simulation, indicating greater variability in this experimental condition.

- The energies, on average, are lower in the 298K simulation, suggesting a lower energy state in this scenario.

- An initial decrease in solute temperature is observed in the 298 K simulation, indicating a cooling trend at the outset of the simulation.

These observations provide insights into the dynamic behavior of the system under different conditions.

5.1. RMSD

^{Q24} For RMSD, the structure of the protein after the (constrained) energy minimization of the protein in water is employed. ^{Q25} The utilization of the @pbc argument serves the purpose of defining the periodic boundary conditions within our model. Its application becomes especially imperative when dealing with scenarios in which the protein resides in close proximity to the boundaries of the simulation box. In such cases, neglecting the consideration of neighboring boxes can result in a significant distortion of metrics such as the Root Mean Square Deviation (RMSD). To ensure the integrity of our analysis, it is crucial to account for these periodic interactions, thus preventing aberrations in the RMSD values due to the partial presence of the protein in neighboring boxes.

^{Q26} It is evident that at elevated temperatures, the Root Mean Square Deviation (RMSD) exhibits notably increased volatility. This phenomenon can be attributed to the heightened kinetic energies of individual atoms within the system. As temperature rises, the thermal motion and kinetic energy of atoms become more pronounced, leading to greater fluctuations in the RMSD values. This observation underscores the intricate relationship between temperature and the dynamic behavior of the system, emphasizing the influence of thermal energy on molecular motion and structural stability.

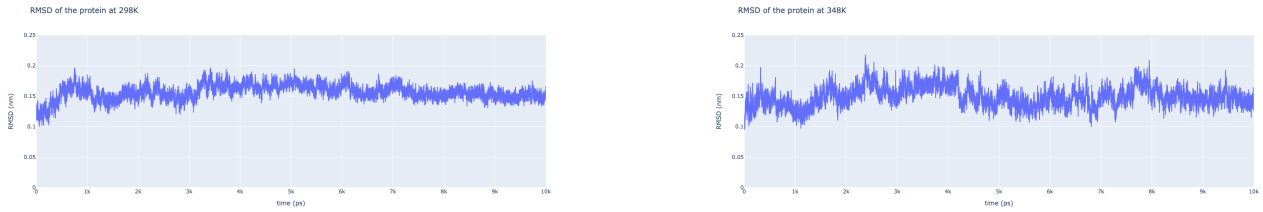


Figure 5: RMSD of the protein at different temperatures

5.2. *RMSF*

The atom-positional root mean square fluctuation (RMSF) gives us information about how locally flexible the protein. *Q27* Generally, we have observed slightly elevated RMSF values, on the order of 0.1 nm, in simulations conducted at higher temperatures. Notably, the positions of the RMSF maxima tend to align. However, a substantial disparity becomes apparent between carbon atoms [16-21] and [46-50] of the protein backbone, where the RMSF values exhibit a pronounced increase in the simulation at 348 K compared to that at 298 K. This distinction underscores the sensitivity of certain regions of the protein to temperature variations, emphasizing the significance of thermal effects on structural dynamics

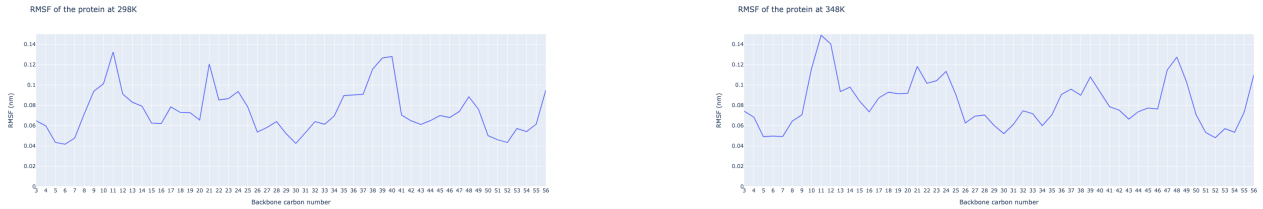


Figure 6: RMSF of the protein at different temperatures

5.3. *Hydrogen Bonds*

Q28 According to the analysis at 348K, the most populated hydrogen bond is as follows:

```
# Two-centered hydrogen bonds:
# HB-ID  Mol  Res DONOR -  Mol  Res ACC Atom D - Atom H ... Atom A DIST ANGLE OCCUR %
303 1 30 PHE - 1 26 ALA 295 N - 296 H - 253 O 0.193 163.487 19608 98.04
```

This hydrogen bond is also the most populated bond at 298K with the following statistics:

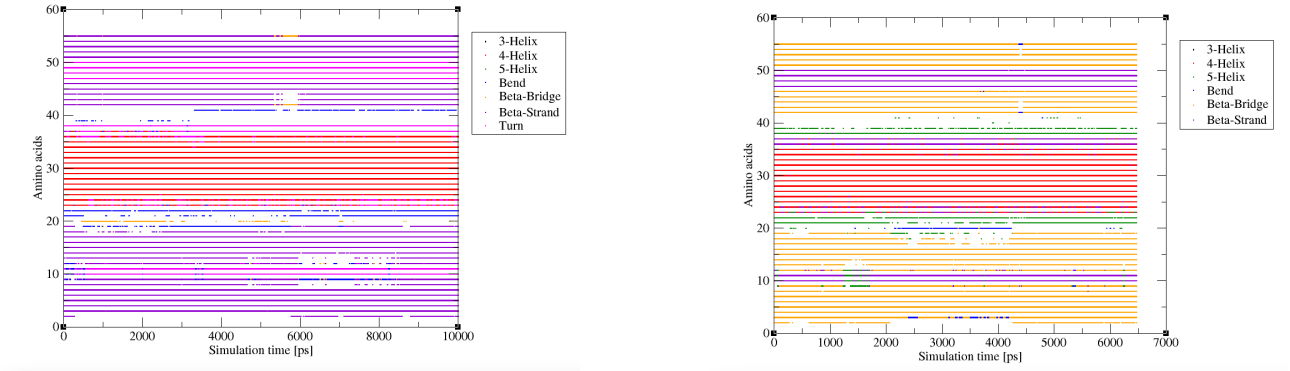
```
1 30 PHE - 1 26 ALA 295 N - 296 H - 253 O 0.192 164.959 19856 99.28
```

It is evident that this hydrogen bond reflects an alpha helix bond, as the distance between the residue number of the donor and the acceptor is 4.

Q29 vFor the simulation at 298K, we can clearly see:

- 14 residues belonging to a single long 4-alpha chain
- 2 groups of beta strands, each comprising around 8 residues
- 2 groups of beta strands, each comprising around 5 residues
- 3 turns of 3 residues
- 2 or 3 bends of 3 residues, approximately

Q30 These observations are consistent with the experimental data. This visualization of the protein provides the first glimpse of its structure, and all the main structural elements mentioned above are present.



(a) Plot describing the classification of the secondary structure of each protein's backbone structure at 298k

(b) Plot describing the classification of the secondary structure of each protein's backbone structure at 348

Figure 7: RMSF of the protein at different temperatures

The graphical representation corroborates the structural attributes identified previously. Notably, it reveals a prominent four-alpha helix alongside four distinct beta strands. However, it is important to acknowledge that certain amino acids pose challenges in terms of classification and precise structural assignment.

Q31 While the primary regions highlighted in both plots remain consistent, it is apparent that there is a denser concentration of mass observed around the central, darker point in the plot. This reinforces our earlier findings, confirming the presence of a substantial alpha helix structure (corresponding to the central region) and the presence of additional beta-sheet structures.Q32 The cloud-like pattern in the top left corner of the plot is indicative of the existence of multiple beta-sheet structures, a count that aligns with our previous determination of four such structures.

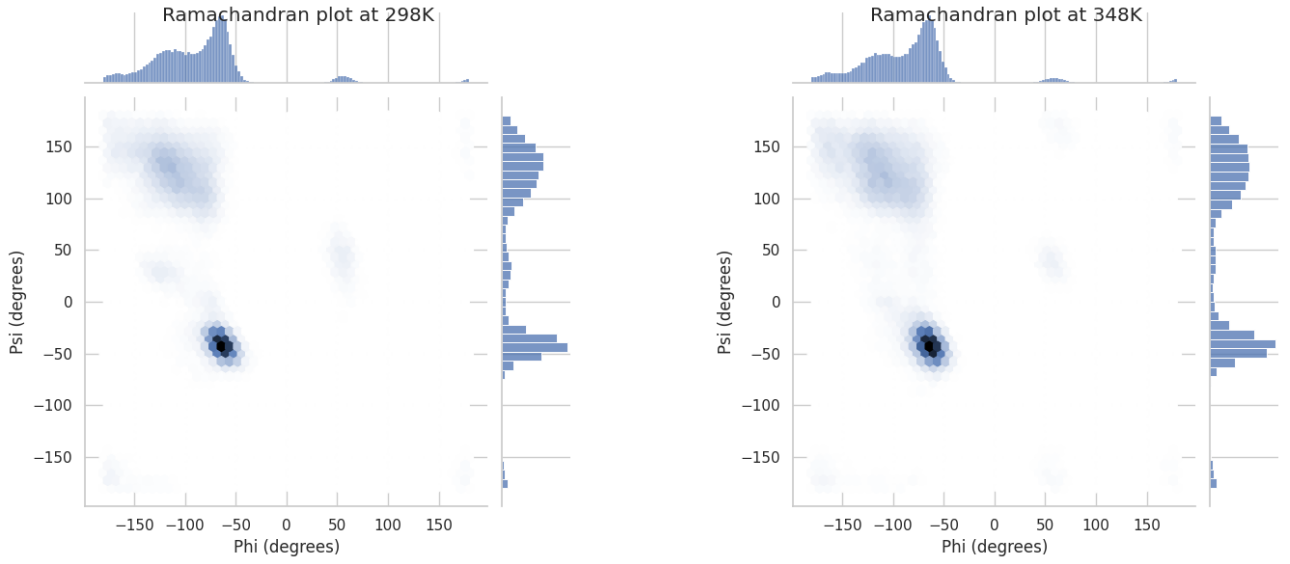


Figure 8: Ramachandran plot of the protein at different temperatures

Q33 At an elevated temperature, the protein exhibits notably increased thermal motion. Additionally, this heightened thermal motion introduces localized, minor deviations from the initial conformation. These deviations have been quantitatively validated through the analysis of Root-Mean-Square Deviation (RMSD) and Root-Mean-Square Fluctuation (RMSF), as discussed earlier. Q34 During the simulation at elevated temperatures, a noticeable departure from the typical

conical structure is observed in the alpha helix. The structural deviation becomes increasingly apparent as the simulation progresses.

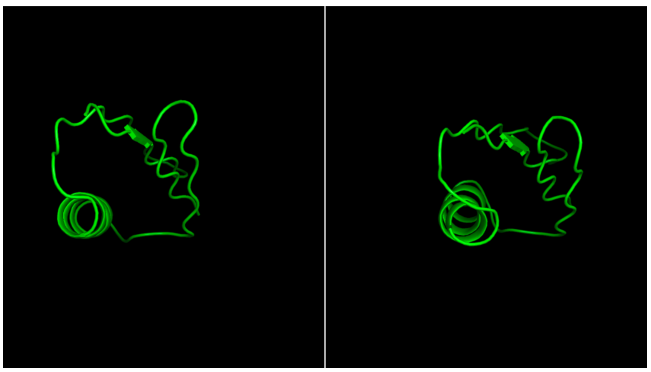


Figure 9: Protein structure. Left: 298K, Right: 348K

Throughout the simulation, the loops connecting the beta sheets exhibit significant mobility, resulting in noticeable tilting and bending of the region where the beta-sheets are situated.

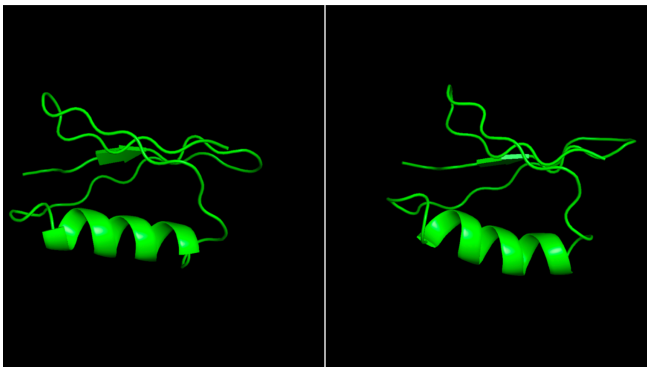


Figure 10: Protein structure. Left: 298K, Right: 348K

Summary 1

You can find the list for the Google Colab Jupyter here.