Molecular Dynamics Studies on the Lipid Transfer Protein - STARD11

By

Hedda Vik Askeland

Thesis

for the degree of

Master of Science

Department of Chemistry

Faculty of Mathematics and Natural Sciences

University of Bergen (UiB)

Repository: https://git.app.uib.no/reuter-group/MSM_Hedda/-/tree/hedda_upload/notebooks

May 2023

Acknowledgement

This master´s project was conducted as part of the KJEM399 – Master´s thesis, a compulsatory subject in the final year of the Master´s in Chemistry Education program at the University of Bergen (UiB). The project was carried out during the autumn of 2022 and the spring of 2023.

As the writer in this project, my background in environmental technology and molecular modelling helped me to successfully complete this thesis. Furthermore, my knowledge from the following subjects: “Molecular Cell Biology”, “Lipids and Proteins in Biological Membranes”, and “Molecular Modelling” proved invaluable in completing this task. The work was conducted under the Computational Biology Unit (CBU) at the University of Bergen (UiB).

I would like to express my sincere appreciation to my internal supervisor and professor, Nathalie Reuter, and postdoctoral fellow, Mahmoud Moqadam from the Department of Chemistry, for their guidance and support throughout this work. Additionally, I would like to extend my gratitude to Susanna Röblitz from the Department of Informatics for her assistance and guidance with the PyEMMA software, which was used for the analysis of molecular dynamics simulations using Markov State Models.

I am grateful for the support and guidance provided by all these people, whose enthusiasm and knowledge helped me gain a deeper understanding of molecular dynamics in the field of biochemistry.

Hedda Vik Askeland
Bergen, May 2023

Abstract

The STARD11 protein plays a crucial role in the transport of ceramide from the endoplasmic reticulum (ER) to the Golgi apparatus. Ceramide is a lipid molecule that is essential for several cellular processes such as cell growth, differentiation, and programmed cell death (apoptosis). The STARD11 protein has a hydrophobic cavity that can bind to one lipid at a time. However, to accommodate for the ceramide molecule, structural rearrangements of the protein are required. Molecular dynamics studies have shed a light on the conformational changes that occur in the STARD11 protein during ceramide binding. In particular, the movement of the α4-helix and/or the Ω1-loop, and Ω4-loop regions appear to be highly dynamic and play a critical role in the binding process. To further understand the structural basis of ceramide binding by STARD11, we conducted a series of molecular dynamics simulations. Our study involved 15 different systems, including wild-type STARD11 with and without ceramide binding, as well as double and single mutation of specific amino acids. By analysing the simulations, we were able to gain insight into the molecular interactions between STARD11 and ceramide, as well as identify key regions that are essential for the binding process. Our study has revealed interesting finding regarding the structural stability of the wild type holo and apo proteins. Specifically, we observed that the wild type holo protein exhibits a higher degree of structural stability compared to the wild type apo protein. Moreover, our study examined the single mutation on the holo and apo proteins, and found that single mutation had greater impact on the structure of the apo protein compared to the holo form. These findings suggests that the presence of a bound lipid can enhance the stability to the protein, possibly by restricting its flexibility. The MSMs analysis allowed for calculating probabilities and study transition between certain states, and we found that there was a significant likelihood that the W473A/W562A apo protein exhibits a larger opening in the Ω1- Ω4-loop, and α4-helix regions compared to the WT apo protein. These observations are important for understanding the relationship between protein structure and functions, as well as the effect of ligand binding on the protein stability.

Contents

1

Introduction and Objective of the Study

1.1 Lipids

Lipids are fat molecules found in the humans (homo sapiens) and are crucial for controlling what goes in and out of the human cells. Lipids are made up of oxygen, carbon and hydrogen, and their function includes transport and storing of energy, absorbing vitamins, and making hormones. The three main types of lipids in homo sapiens are triglycerides, phospholipids, and sterols where cholesterol is the most prominent. Triglycerides is the most common lipid and can be found in butter, milk, and meat. Phospholipids are water-soluble and are important for building the membrane barrier around the cells in our body. Sterols are the least common type, and our body produces most of it. Cholesterols are crucial for synthesizing reproductive hormones, vitamin D and bile salts [1].

All cells require separation from their surroundings. Lipid bilayers form this barrier between the inside of the cell and its environment, and are composed of different types of lipids, proteins, and carbohydrates. The lipid composition is altered for the cell function, and lipid transfer systems is important for this composition. Lipids trafficking happens both inside and outside the cells. Moving lipids is linked to traffic of membrane vesicles in the eukaryotic cells. Non-vesicular trafficking is also important for areas that are not coupled to vesicular traffic. The non-vesicular trafficking serves multiple roles. First, delivering lipids that cannot be done by vesicular transport. Second, fast changes of the lipid composition. Finally, correcting any unfortunate lipid movements caused by vesicular transport. Experimental evidence has shown that non-vesicular traffic with phospholipids and cholesterol move bidirectionally between the endoplasmic reticulum (ER) and the plasma membrane with a higher speed than vesicular transport would allow. In additional, evidence has revealed that disruption of the vesicular trafficking has little consequence on the transport between ER and plasma membrane [2, p. 85-86].

1.2 Lipids Transfer Proteins

Lipid transfer proteins (LTPs) are proteins which move lipid molecules with non-vesicular transport from the donor to the acceptor membrane. The LTPs move the lipids one at a time using their hydrophobic cavities which allows the lipid to fit inside- covered by a lid. The hydrophobic cavity of the LTPs creates a lower free energy environment for the lipids comparted from the aqueous region. The stoichiometry between the LTPs and lipid molecules is typically one to one, and there are 27 different protein families that transport lipids with their cavities. Some transfer proteins can form open bridges or closed tubes between two membranes that are close together. This allows the lipid molecules to move between membranes without the presence of protein movement. The LTPs cavity bears a resemblance to a box, and the transport of the lipid molecule is divided into eight steps. These steps are listed below [2, p. 85-87]:

- Donor membrane docking
- Lipid extraction
- Donor membrane undocking
- Cytosolic diffusion
- Acceptor membrane docking
- Lipid deposition
- Acceptor membrane undocking
- Further diffusion

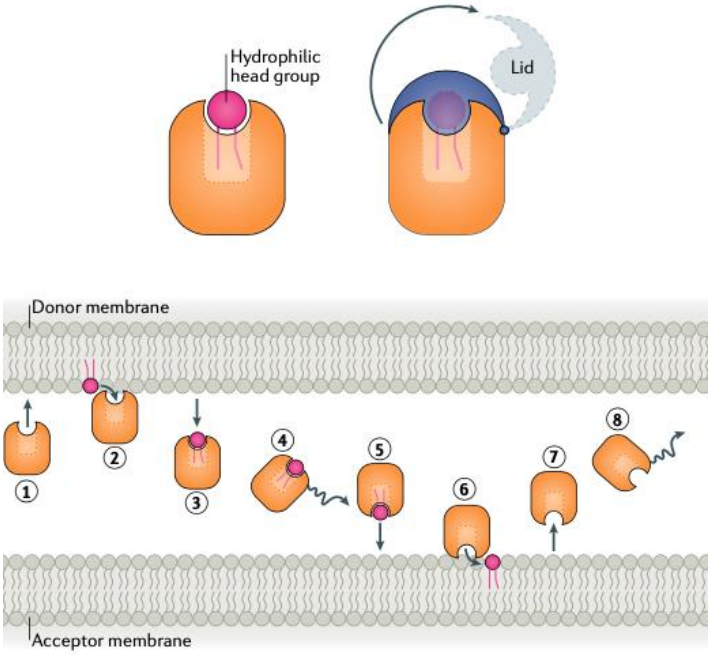

Figure 1.2.1: LTPs (yellow) eight steps for lipid (purple) transport from donor membrane to acceptor membrane. Retrieved from Fig. 1 in [2, p. 86].

1.2.1 Classification of the StAR related LTPs

There are 27 protein families which form lipid transport cavities. The family of StAR related lipid transfer proteins (START) is one of these families [1, p. 87]. The START domain binds and transports glycerolipids, sphingolipids and sterols between membranes. The ligand fits inside the hydrophobic cavity of the domain where it is guarded from the aqueous environment. There are 15 different START domains in humans (Homo sapiens). The START domain protein family is categorized into six sub-families which depends on the sequence homology between the proteins. The protein sharing a common lipid and alike functions are grouped together. The six categories are listed below: [3, p. 85-89].

- The group carrying organelle-bound cholesterol - consists of the two founding members of the START family: STARD1 and STARD3. Both proteins can bind and transport cholesterol.
- The START-only sterol carries group – also consists of proteins that bind to cholesterol (STARD4, STARD5 and STARD6).
- The group carrying phospholipids/ceramide - this subgroup contains STARD2, STARD7, STARD10 and STARD11/CERT/GPBP. The first two proteins transfer

only phophatidylcholine (PC), but STARD10 also binds PC or phosphatidylethanolamine (PE). Unlike the three other members of the group, STARD11 has two additional domains interacting with the membranes, a pleckstrin homology domain (PH) and a FFAT motif.

- The RhoGAP-START group - comprises three proteins: STARD8, STARD12 and STARD13. They all have a sterile alpha motif (SAM) and a Rho-GTP activating protein (Rho-GAP) in addition to their START domain.
- The thioesterase group - includes STARD4 and STARD15. These two proteins are a part of the acyl-CoA thioesterase (ACOT) family which hydrolyzes acyl-CoA molecules.
- STARD9 – is a large protein consisting of a kinesin motor domain, a FHA phosphoprotein binding domain and a START domain.

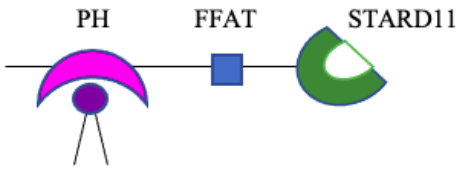

Figure 1.2.2: The STARD11 protein with two additional domains, the PH and the FFAT motif. Modified from Fig. 1 in [3, p. 86].

1.2.2 The Ceramide Transfer Protein (CERT)

The CERT is a multi-domain lipid transfer protein which transports ceramide in a non-vesicular manner from the ER to the Golgi apparatus. The CERT protein is divided into three regions, which are listed below:

- the N-terminal region – contains 120 amino acids, together with the PH domains.
- the middle region – consists of 250 amino acids, including coiled-coil motifs and the FFAT motif.
- the C-terminal region – The START domain, which also consists of 250 amino acids.

The ceramide is converted to a sphingomyelin in the Golgi apparatus. STARD11 recognizes and transports D-erythro-C_{16}-ceramide, dihydroceramide, phyto-ceramide, diacylglycerol, and other ceramides with C14-C20 amide-acyl chains. Ceramide is important for cell growth,

cellular differentiations, and apoptosis. Both a cytosol- dependent and a cytosol- independent ATP pathways have been discovered.

Mutational analysis of ceramide extraction has improved our understanding of ceramide uptake and release by the STARD11 protein [4, p. 488].

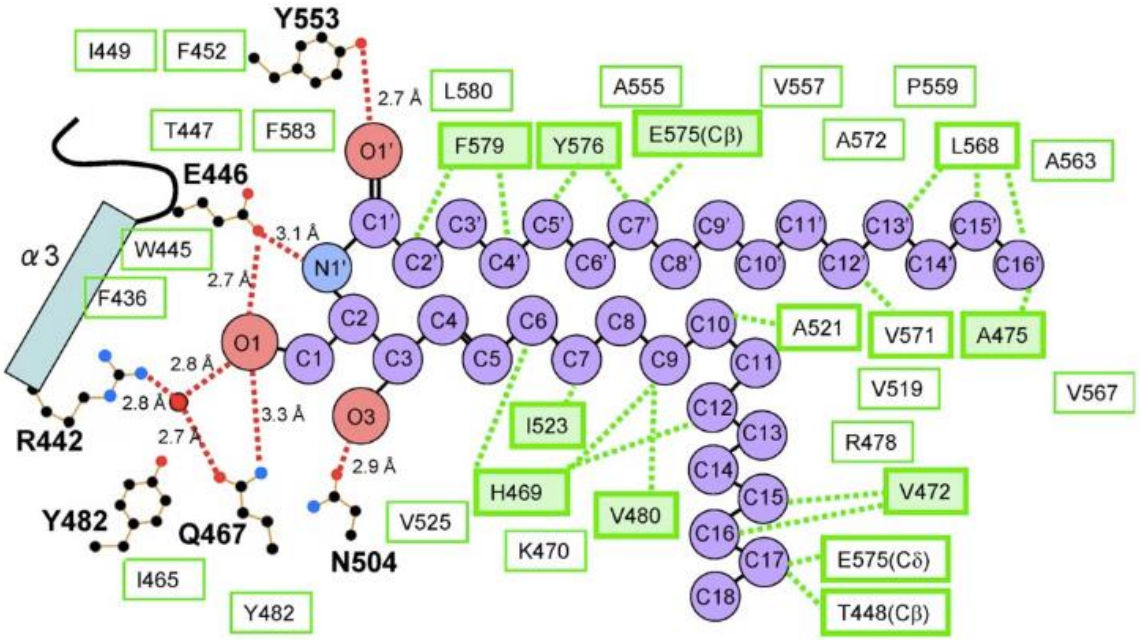

*Figure 1.2.3: Schematic drawing of C_{16}-ceramide, transported by the START/CERT domain. Representation includes hydrogen bonds (red dashed lines), water molecules (red circles), C, N, and O atoms; black, blue, and red dots respectively. Molecules involved in hydrophobicity of the protein cavity are the green boxes. Molecules with direct hydrophobic interactions are labelled with thick green borders including their dashed green lines. Eight of these thick green bordered molecules are also filled with light green inside the box, which represents amino acids residues common to all the C_6-, C_{16}-, C_{18}-*ceramides. *Retrieved from Fig. 2 in [4, p. 489].*

CERT involvement in diseases

Studies on deficient mice have enhanced our understanding of the unique role of ceramide transport between donor to acceptor membranes in the mammalian cells. During embryogenesis process, mice lacking the ER portion of the CERT protein experience heart failure, leading to death. Deficient embryos exhibit various morphological alterations. Furthermore, the cell experience growth of ceramide and sphingomyelin content [3, p. 88]. In

additional to its role in ceramide transport, STARD11 is also involved in the Goodpasture syndrome, cells infected by Chlamydia trachomatis, and has implications in cancer development [3, p. 92].

1.2.3 Structure of STARD11 Domain

The three-dimensional structure of STARD11 in the presence and absence of ceramide, has been resolved by X-ray and is available in the Protein Data Bank (PDB) [5]. The PDB IDs are 2e3m and 2e3q for the forms without ceramide (apo) and with ceramide (holo) respectively. The structure contains a helix- grip fold which consists of an anti-parallel β-sheet with nine β-strands gripped by α-helix 1 and α-helix 2. Between β5/β6 and β7/β8, two Ω-loops (Ω1 and Ω2) are located. The structure also displays two aromatic residues exposed to the protein exterior, tryptophan-473 and tryptophan-562 (Trp-473 and Trp-562). The lipid-binding hydrophobic cavity has room for one lipid molecule. The lipids hydrophilic head is placed in the most inner part of the cavity, and the acyl chains are surrounded by the hydrophilic wall of the cavity. The entrance to this cavity is formed by an α-4 helix and Ω1-loop, serving as a potential gateway for ceramide to enter or exit [4]. This protein region appears to be highly dynamic. The STARD11 protein can also sense the donor and acceptor membrane for lipid delivery. For instance, the Ω1-loop has a hydrophobic area which is thought to be important for ceramide transfer [3, p. 89].

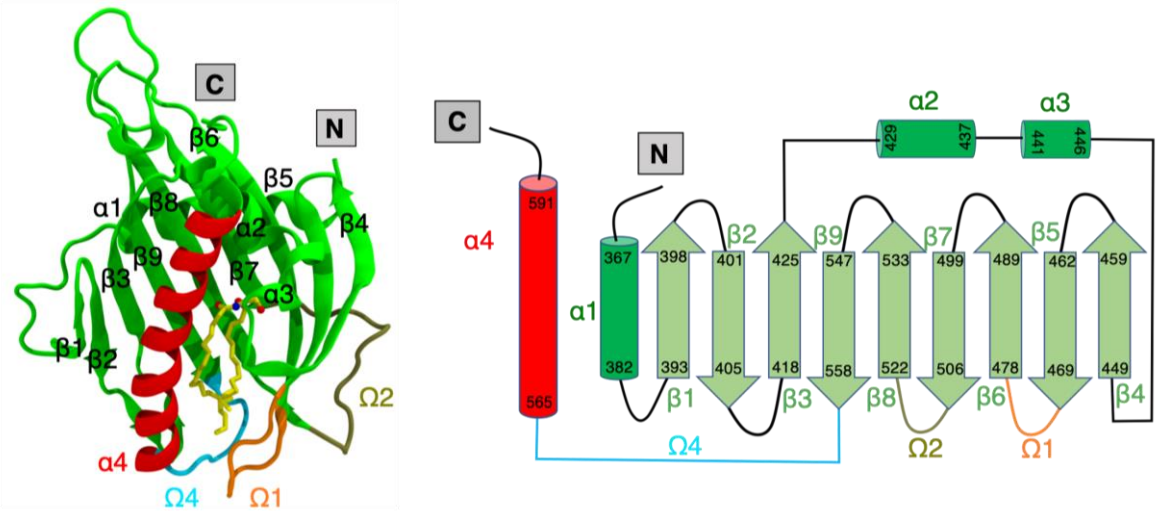

Figure 1.2.4: Structure of the STARD11 domain of the CERT protein. Credit: Mahmoud Moqadam.

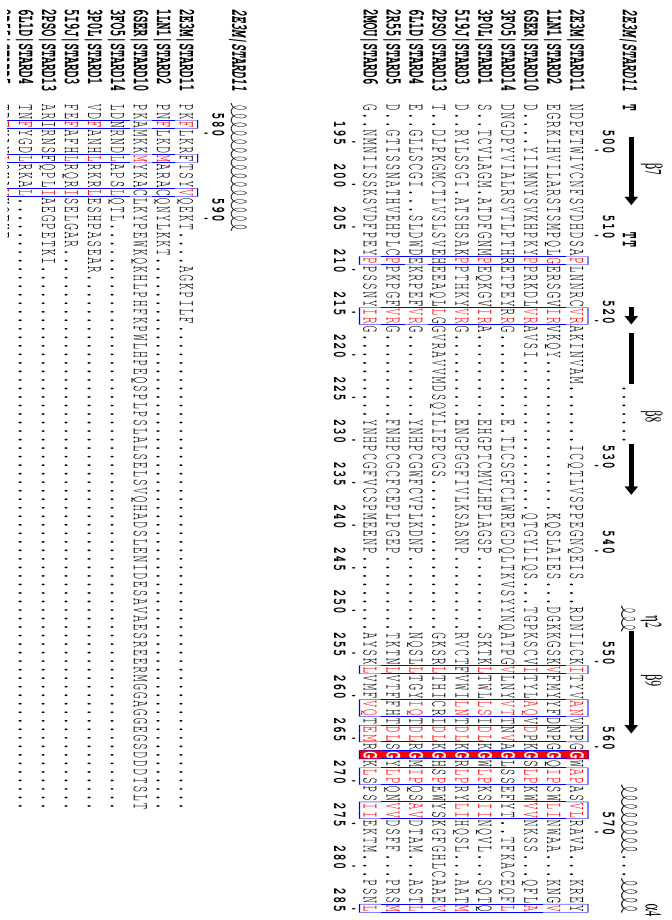

Figure 1.2.5: Alignment of START proteins with the secondary structure of STARD11 (2E3M), and its corresponding residue number. Retrieved from [6].

1.2.4 Transfer Mechanism

The transportation pathway of the CERT protein favours the membrane contacts site (MCS) region within the membrane. The two protein domains interacting with the ER membrane and Golgi membrane is the Pleckstrin homology (PH domain) and FFAT motif, respectively. The FFAT motif binds to the VAP in the ER where it attaches to the ceramide. Whereas the PH domain binds to a phosphoinositide (Pl4P) located in the Golgi apparatus where it delivers the lipid. Most likely the protein act within the MCS and form a “bridge” where it binds to the two membranes simultaneously. If this is the case the protein delivers the ceramide molecule by a “neck-swinging” movement [3, p. 90-91].

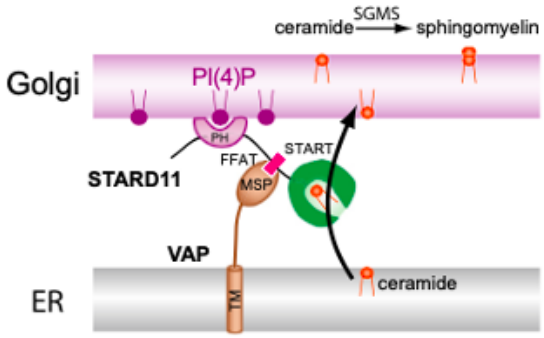

Figure 1.2.6: Schematic representation of molecular mechanism of the CERT protein. CERT transfers ceramide from donor membrane (ER) to acceptor membrane (Golgi) through MCSs regions. Retrieved from Fig. 4 in [3, p. 91].

1.3 Objective of the study

The objective of this project is to enhance our knowledge of the lipid transfer mechanism of the CERT STARD11 domain. More specifically, our focus will be to investigate the region responsible for the uptake and release of the lipid molecule, ceramide. To achieve this, we have conducted mutation analysis, specifically targeting the Ω1- Ω4-loop, and α4-helix regions within the protein domain. To accomplish these goals, molecular dynamics (MD) computer simulations has been done. MD allowed us to simulate the movement of the protein at an atomic level, providing us with information on their structural and dynamic properties. Additionally, Markov state models (MSMs) was used for analysing MD simulations on slower timescales. By using MD and MSMs, we can explore the structural and dynamics properties of the protein-lipid complex.

2

Theoretical Background and Computational Approach

2.1 Molecular Dynamics Simulations

The way STARD11 engages with the donor and acceptor membranes, and how the protein capture and releases ceramide is poorly understood. These mechanisms are challenging to observe experimentally because of the timescale and the atomic level of details needed. To understand these mechanisms, molecular dynamics simulations can be used. MD simulations provide relation between 3-D structure and dynamics by considering the conformational energy of the protein molecule and its environment. In 1977 the first MD simulations of a small protein in vacuum (with 9.2 ps trajectory) was reported [7]. The same protein in water was described eleven years later with a simulation of 210-ps trajectory [8]. The power of computers has increased, and for that reason we can run longer simulations of larger proteins. In addition, the molecular mechanism force field have been improved and extended to more molecules. On the ground of this, we can run more complex systems, including proteins and lipid bilayers. Making improvement on other areas has also increased the simulations stability and accuracy [9].

MD simulations model the thermal fluctuations and relative positions of atoms in a molecule over a certain timescale. Biochemistry simulations is used to increase our insight on ligand binding, enzymatic activities, signalling mechanism and protein folding. Other simulations tools are also important for refinement of electron microscopy, x-ray, and NMR data for more correct structures [10, p. 3].

2.1.1 Basic principles of Molecular Dynamics

MD simulations can be used to compute the properties of classical many-body systems. MD simulations compute the systems equilibrium and transport characteristics following the laws of classical mechanics. The first step in MD simulations is to model a system with N particles

and use Newton´s second law of motion while the system is changing with time. Newton´s second law is as follows:

$$F_i(t) = m_i a_i(t) \tag{2.1.1}$$

$$\frac{d^2 x_i}{dt^2} = \frac{F_{xi}}{m_i} \tag{2.1.2}$$

For a system over a time duration with N particles where each atom is i, m_i represents its mass, a_i its acceleration and the force on each atom is F_i.

When the system has come to an equilibrium the measurements can be done. To observe the quantities in the system after the equilibrium step, we must express this as a function of the positions and momenta of the particles. Average kinetic energy per degree of freedom is as follows:

$$mv_{\propto}^2 = \frac{1}{2} k_b T \tag{2.1.3}$$

Where m stands for mass, v is velocity, k_b represents the Boltzmann factor, and T is the absolute temperature [11, p. 63-64].

MD simulations are dependent on a model for the physical system and choosing the potential energy as a function $V(r_1, \ldots\ldots, r_N)$ of the positions r of the atoms. Forces on the atom i can be derived as the negative gradient of the potential energy:

$$F_i = -\nabla_{ri} V(r_1, \ldots\ldots, r_N) \tag{2.1.4}$$

The sum of the potential energy, V, and the instantaneous kinetic energy, K, is the total energy of the system, E, [12, p. 10]:

$$E = K + V \tag{2.1.5}$$

2.1.2 Integrating Newton´s Second Law of Motion

Once the forces on the atoms have been computed, now Newton´s second law of motion can be integrated. The Verlet algorithm is one of the simplest and best algorithms for integrating the equation of motion and requires two independent initial conditions: positions and velocities. The Verlet algorithm is derived from the Taylor expansion of the particles position, r, around time, t:

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 + \frac{\Delta t^3}{3!}r + O(\Delta t^4) \quad (2.1.6)$$

$$r(t - \Delta t) = r(t) - v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 - \frac{\Delta t^3}{3!}r + O(\Delta t^4) \quad (2.1.7)$$

Adding the two equations gives:

$$r(t + \Delta t) + r(t - \Delta t) = 2r(t) + \frac{f(t)}{m}\Delta t^2 + O(\Delta t^4) \quad (2.1.8)$$

Present time position and velocity are $r(t)$ and $v(t)$ respectively, and $r(t + \Delta t)$ and $r(t - \Delta t)$ are the particles position forward and backward in time. We must consider that the particles new position contains errors of the fourth order, Δt^4, where Δt is the time step of the MD simulation.

The Verlet algorithm is fast, it requires little memory, and its short-term energy conservation is fair. The algorithm also displays long-term energy drift, it is time reversible and an equation which leaves the volume element in phase space unchanged over time. In a MD simulation one should be aware that the aim is not to predict precisely what will happen to a system, but predict statistical predictions, the average behaviour of a known condition in our system (e.g., the total energy) [11, p. 69-74].

A disadvantage with the Verlet algorithm is that velocities are not directly generated to compute the new position, which is sometimes necessary. The velocities are essential for computing the kinetic energy and to test the preservation of total energy with equation (2.1.5). This testing of preservation is important for assessing the evolution of the simulation. Other variants of the latter algorithm have been developed. The velocity Verlet is one of those and has been used in this project.

The velocity Verlet algorithm calculates positions, velocities, and accelerations at time $t + \Delta t$, based on the corresponding quantities at time t using the following procedure [12, p. 16]:

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \left(\frac{1}{2}\right)a(t)\Delta t^2 \quad (2.1.9)$$

$$v\left(t+\tfrac{1}{2}\Delta t\right) = v(t) + \left(\tfrac{1}{2}a(t)\Delta t\right) \quad (2.1.10)$$

$$a(t+\Delta t) = -\left(\frac{1}{m}\right)\nabla V(r(t+\Delta t)) \quad (2.1.11)$$

$$v(t+\Delta t) = v\left(t+\Delta t\frac{1}{2}\right) + \left(\frac{1}{2}a(t+\Delta t\right)\Delta t \quad (2.1.12)$$

2.1.3 Potential Energy Function

The potential energy function is a function for the atomic coordinates in a system where the energy consists of bonded- and non-bonded interactions. The interactions can be represented by the principles of classical mechanics; molecules are treated as a set of spheres connected by springs. The potential energy function, V, can be expressed as a sum of terms representing the *bonded* and *non-bonded* interactions:

$$V(r) = E_{bonded} + E_{unbonded}$$
$$= \Sigma_{bonds}k_i^{bond}\left(r_i - r_{i,0}\right)^2 + \Sigma_{bondangles}k_i^{angle}\left(\alpha_i - \alpha_{i,0}\right)^2 + \Sigma_{torsionangles}k_i^{torsion}$$
$$\left(1-\cos\left(n_i\phi_i\phi_{i,0}\right)\right) + \Sigma_{pairs-ij}\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \Sigma_{pairs-ij}\frac{1}{4\pi\varepsilon_r\varepsilon_0}\frac{Q_iQ_j}{r_{ij}} \quad (2.1.13)$$

The Σ_{bonds} term represents the required energy to stretch or compress a covalent bond pair. The $\Sigma_{bondangles}$ term represents the required energy it takes to bend a bond from its equilibrium angle. The $\Sigma_{torsionangles}$ term is for four atoms in a chain and their rotational sum (dihedral angle) between two planes. The two last terms are for non-bonded interactions and includes van del Waals and electrostatic interactions [10, p. 3-4]. Illustration of the interactions are represented below:

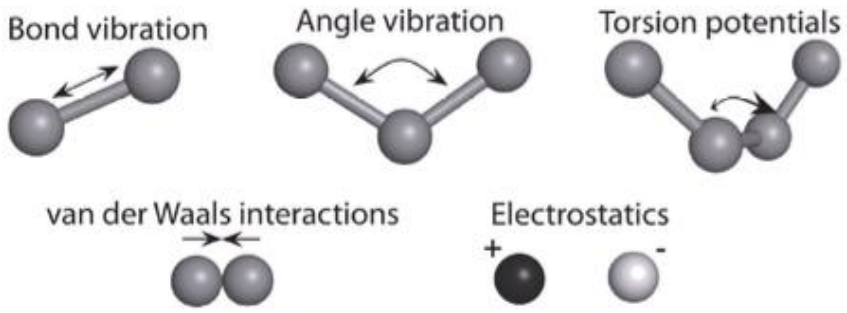

Figure 2.1.1: Illustration of the interactions between atoms which are included in the potential energy function, V. Retrieved from Fig. 1 in [10, p. 4].

Bonded interactions

The potential energy function for bonded interactions includes pair interactions, 3-body, and 4-body interactions. Bond stretching is a term between covalently bound pairs of atoms, bond angle describes the energy associated with changes in the valence angle between 3 atoms, and the dihedral angle term if for the energy associated with the torsion angle defined by 4 atoms.

Nonbonded interactions

The van der Waals interactions are short-range and take place when two atoms or molecules come close together. Short-range means that the distances are within range of the size of the interacting atoms, therefore the neighbouring atoms are often the only ones involved in the calculations. Electrostatic interactions are short range and long-range interactions, and therefore relevant for more than just the neighbouring atoms [14, p. 1].

The Lennard-Jones pair potential is a common interaction model for non-bonded terms and can be described with the following expression:

$$\phi_{LJ}(r) = \Sigma_{pairs-ij} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} = 4\varepsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6}] \quad (2.1.14)$$

The Lennard-Jones potential is computationally intensive because it includes all pair interactions. The first terms describe the repulsion between atoms when they are brought close together, whereas the second term is for attractive forces and is dominating at larger

distances. ε and σ parameters are chosen to match the physical properties of the material and in simulations it is usual to calculate with ε =1 and σ =1 [12, p. 10-11].

The Coloumb potential is another model that describes the electrostatic interactions between two atoms. The following equation represents the Coloumb potential:

$$V_{i,j}^{Coloumb} = \frac{1}{4\pi\varepsilon_r\varepsilon_0}\frac{Q_iQ_j}{r_{ij}} \tag{2.1.15}$$

The Q parameters are two-point charges separated by distance r, and $4\pi\varepsilon_r\varepsilon_0$ is the electric conversions factor [13, p. 70].

2.1.4 Force Fields

Equation (2.1.11) and its parameters k_i, $r_{i,0}$, $\alpha_{i,0}$, φ_i, $\varphi_{i,0}$, A_{ij}, B_{ij} and $\varepsilon_{r,0}$, Q_{ij} are called the force field of the simulation. The accuracy of the simulation depends on how these parameters are defined. The most common all-atom force fields are AMBER [15], CHARMM [16], OPLS [17] and GROMOS [18], meaning they include all the atoms in the simulation and provide a high level of detail. A disadvantage is that simulations with these force fields are computationally demanding [10, p. 4]. In this project CHARMM36m has been used [19][20].

2.1.5 Initial Conditions

The starting point for MD simulations is the equilibration step. Equilibration is used to obtain the thermodynamic properties which should be somewhat alike the properties in experimental conditions. In MD simulations of biomolecules, a common thermodynamic ensemble is NPT. The isothermal isobaric NPT ensemble holds the number of particles N, the pressure P and temperature T constant. To control the P and T additional algorithms called thermostats and barostats must be utilized [21, p. 10].

2.1.6 Periodic Boundary Conditions

Periodic boundary conditions (PBC) are a model used to approximate infinitely large systems and make them traceable for simulations. This is because it is computationally impossible to mimic a real macroscopic system with number of atoms of the order 10^{23}. In PBC, particles are enclosed in a cubic box (called the *central cell*) with water. Without PBC, the water box

edges would experience vacuum. The box is replicated to infinity in three directions in space and for preventing the molecule to see its mirror image the central cell needs to be big enough. If a particle is in position r, one can assume that the particles is repeated infinitely number of times with location:

$$r + la + mb + nc, \quad (l, m, n = -\infty, \infty) \tag{2.1.16}$$

Where a, b, c represents the vector for the edges of the box. These duplicated particles move together, and only one is represented in the computed program. Now the particles interact with their duplicates as well as other particles in their own water box. The surface effects are non-existing, therefore particles can "move through" the box edges. In simulations with PBC, the number of the pairs significantly increases. However, the application of the minimum image criterion prevents this issue by reducing the pair number to a minimum by ensuring that each atom interacts with the closest atom [12, p. 12-13].

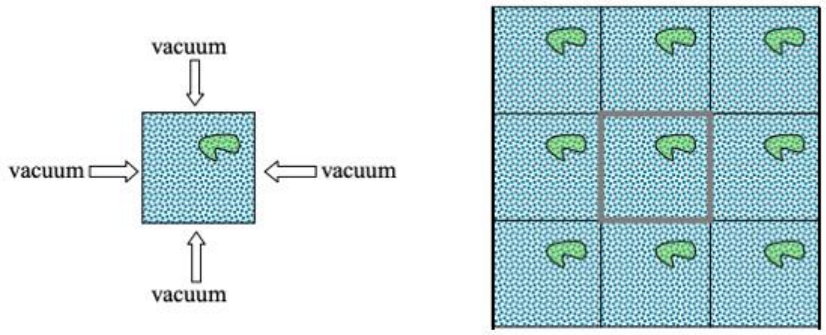

Figure 2.1.1: Illustration of a molecule in a water box, surrounded by vacuum and periodic boundary conditions to the right. Retrieved from Fig. 6 in [10, p. 5].

2.1.7 Solvent Treatment

Biochemistry simulations primarily use water as a solvent. In a simulation cell, the solvent represents many atoms, and therefore the choice of the solvent method is important. MD simulations use methods that handle solvents either explicitly or implicitly. Explicitly solvent treatment includes interactions between all pairs of solute and solvent atoms explicitly. This method is efficient and accurate for long-range interactions in the system. Implicit computed

methods treat the solvents as a continuum, therefore reducing the number of particles in the system and increasing the computational speed compared to the explicit solvent treatment [22, p. 1153].

2.1.8 Limitations of Molecular Dynamics

Molecular dynamics simulations are an essential tool for measuring the statistical properties of classical many-body systems. With a fast-expanding field, MD simulations are becoming more and more like real experiments. However, today's simulations have their limitations which we must assess with careful consideration.

First, the classical approximation should not be used with very light substances such as dihydrogen, helium, and neon. Moreover, if the temperature is sufficiently low the quantum laws should be obeyed. Second, the forces between the atomic coordinates in a system is dependent on the gradient of the potential energy function. Therefore, the accuracy of the MD simulations depends on the precision of the force field and its ability to reproduce the behavior of the system. Finally, it may happen that the time and size of the simulations become a key factor. Different properties have different relaxation times. An example is when a system is in phase transition the system tends to become slow. Therefore, this can become hard to capture in a simulated system. Moreover, the limited size can cause a disadvantage. When the system is in phase transition the correlation length may increase or diverge [12, p. 7-8].

2.1.9 Analysis of Molecular Dynamics Trajectories

RMSD

The root-mean-square-deviation (RMSD) is used to compare two conformations within the protein. RMSD can be used to follow the changes of the protein structure during the simulation by comparing structures along the simulation to a reference structure, for example the starting structure. When the RMSD calculation curve is no longer changing significantly the system can be considered equilibrated. RMSD for atoms in a molecule with respect to a reference structure is calculated by least-square fitting the structure to the reference structure (t_2=0) as follows:

$$RMSD(t) = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i(r_i(t_1) - r_i(t_2))^2} \quad (2.1.17)$$

$M = \sum_{i=1}^{N} m_i$ and $r_i(t)$ is the position of atom i at time t. To avoid noise in the data from the larger fluctuations of the side chains the proteins is typically fitted on the backbone C_α atoms [13, p. 217-218].

RMSF

The root-mean-square-fluctuation (RMSF) is used to describe the average distance between the fluctuation of a group of atoms in the protein within a certain period of the simulation. Therefore, RMSF is a measure of the flexibility of the residues in the protein. The simulation RMSF is calculated as:

$$RMSF_i = \sqrt{< (r_i - < r_i >)^2 >} \quad (2.1.18)$$

The r_i represents the coordinates of atom i and $< r_i >$ is the atoms average position. The RMSF is also typically fitted on the backbone C_α atoms [20].

To examine the accuracy of the atomic fluctuation it is important to compare the simulation RMSF to experimental RMSF results. The experimental B-factor represents the temperature factor for the system and is coupled to the mean-square atomic fluctuations in an isotropic harmonic model. The experimental RMSF is calculated as:

$$RMSF_{E,i} = \sqrt{3B_i/8\pi^2} \quad (2.1.19)$$

B_i is the estimated values from the X-ray B-factor for residue number i [23, p. 191-193].

2.2 Computational Approach

2.2.1 System Setup

The CERT START11 domain is a protein domain involved in the uptake, transportation, and release of ceramides. To better understand the behavior of this protein domain, two molecular structures were taken from the Protein Data Bank (PDB). Both were obtained by experimental X-ray diffraction. The first crystal structure is of the apo form and has 237 residues, and the PDB entry is 2E3M [24]. The second crystal structure of the CERT START11 domain was obtained from PDB entry 2E3Q [25]. The structure, consisting of 235 residues, is bound to C18-ceramide (holo form). However, the first two residues are absent in this conformation, unlike in the apo form. It was chosen to investigate how the protein structure changes when in complex with ceramide. To strengthen the data, two replicas of the structure 2E3M were simulated.

CHARMM-GUI, a web-based tool, was utilized for building the MD simulations systems [24]. The simulations employed the all-atom CHARM36m force field, ensuring accuracy for folded and disordered proteins [26]. The WYF parameter for cation-π interactions was selected, and the systems were maintained at 310 Kelvin. Three 3 K^+- ions were added to neutralize both the 2E3M and 2E3Q structures. The box dimensions along the x, y, and z-axes were 89 Å for the 2E3M complex, and 88 Å for the 2E3Q complex.

To gain a deeper understanding of ceramide uptake, transportation, and release, mutation simulations were performed on the CERT START11 domain. Double mutations were made from tryptophan (W) to alanine (A) on two residues (Trp-473) and Trp-562) for each of the two crystal structures. These mutations are denoted W473A/W562A. Additionally, two separate single mutation analysis were performed on each crystal structure, one for Trp-473 (W473A) and one for Trp-562 (W562A). Table 2.2.1 provides an overview of the 15 MD simulations performed in this project. The table includes information on each simulation's denotation, duration, and the number of atoms involved.

Table 2.2.1: List of the 15 different systems simulated in this project with duration 623-801 ns, including number of atoms for each simulation. WT apo 1, 2, and 3 represents wild type of 2E3M and WT holo 1 and 2 represents wild type of 2E3Q (CERT-STARD bound to ceramide). Double mutation on 2E3M and 2E3Q are denoted W473A/W562A apo 1, W473A/W562A apo 2, W473A/W562A holo 1 and W473A/W562A holo 2. Mutation analysis on structures 2E3M and 2E3Q was done on residues 473 and 562, denoted as W473A apo 1, W473A apo 2, W473A holo 1, W562A apo 1, W562A apo 2 and W562A holo 1.

System number	**Protein data bank ID**	**WT/mutant**	**Duration (ns)**	**Number of atoms**
1	2E3M	WT apo 1	683	66858
2	2E3M	WT apo 2	699	66858
3	2E3M	WT apo 3	699	66867
4	2E3Q	WT holo 1	699	64483
5	2E3Q	WT holo 2	699	64483
6	2E3M	W473A/W562A apo 1	801	66839
7	2E3M	W473A/W562A apo 2	699	66839
8	2E3Q	W473A/W562A holo 1	632	64470
9	2E3Q	W473A/W562A holo 2	699	64461
10	2E3M	W473A apo 1	699	66856
11	2E3M	W473A apo 2	699	66859
12	2E3Q	W473A holo 1	699	64451
13	2E3M	W562A apo 1	699	66880
14	2E3M	W562A apo 2	699	66877
15	2E3Q	W562A holo 1	699	64493

2.2.2 Simulation Parameters

The TIP3P water model was used in NAMD for the simulations. Both the NVT- and NPT ensembles were used for equilibration and dynamic steps. To achieve energy minimization, a conjugate gradient algorithm was used [25]. The Langevin dynamics method was used to maintain a constant temperature of 310 Kelvin. The Nose-Hoover Langevin piston was applied to control fluctuations in the barostat and keep the pressure at 1 atmosphere. The Particle Mesh Ewald (PME) algorithm was used to treat electrostatic interactions and reduce interaction complexity [28]. A cutoff distance of 12 Å was used for Lennard-Jonas interactions, and the pair list was updated every ten steps using a timestep of $\Delta t = 2$ fs.

The velocity Verlet algorithm was used to integrate the equations of motion. All MD simulations were carried out using NAMD 2.13 [26] on the supercomputer BETZY using 4 and 8 nodes with 128 tasks per node.

Various software tools available within the NAMD package were utilized for trajectory analysis, Pymol [27] and VMD [10]. Pymol and VMD were also employed to create figures from the simulation trajectories.

2.2.3 Simulation Trajectory Analysis

Information on conformation changes was obtained through the analysis of root-mean-square deviation (RMSD), which allowed for tracking the overall structural changes of the protein over time compared to a particular structure. A lower RMSD value indicates that the molecules are relatively stable and are not undergoing many changes, while a higher RMSD value indicates that the molecule is undergoing significant conformational changes.

To gain a deeper understanding of the local fluctuations in the protein, we calculated root-mean-square fluctuations (RMSF) along the simulations. For a given protein and simulations, the fluctuations are calculated with respect to the average structure computed from that simulation. This analysis provides more detailed information of the flexibility of specific regions in the protein, allowing for identifying areas that may play a crucial role in the proteins function. Looking closer at residue number 473 and 562 was the focus in the RMSF plots because these amino acids are located within the Ω4-loop and Ω1-loop. These regions

are highlighted in all the RMSF plots. The RMSF values for residue numbers between 367 and 400, which correspond to the N terminus, are not considered informative as this region tends to exhibit high fluctuations in all simulations.

In addition, the distances were calculated between specific pairs of molecules within the protein. This helped determine the precise distances between amino acids and gain insight into the protein´s structure. To investigate the interactions between the Ω4-loop and Ω1-loop regions, which are known to be a part of the lipid-binding cavity region, distance values for residues 473 and 564, as well as 476 and 562, were specifically analysed. These regions were of particular interest in the project, as they could shed a light on the opening of the cavity and ceramide transfer.

2.3 Markov State Models

Markov state models (MSMs) are a tool for presenting and analyzing MD simulations on a slow timescale, making the resulting trajectory less challenging to understand and interpret. MSMs use kinetically relevant states and the timescales between them to build dynamical systems. Protein folding, and ligand binding are examples of typical simulations that can be presented by MSMs. Today, the complexity of the system is limited to around 10-100 residues [28].

2.3.1 Dimensionality Reduction with TICA

TICA (time-lagged independent component analysis) is a linear transformation method that distinguishes itself from PCA by seeking out coordinates that exhibit maximal degree of autocorrelation at a particular time delay, rather than maximal variance. This distinction makes TICA particularly useful for identifying slow components in a set of data. When input data is derived from a Markov process, TICA approximates the eigenfunctions and eigenvalues of the underlying Markov operator in an optimal way [29].

TICA operates on a sequence of multivariate data, X_t, by constructing both the mean-free covariance matrix and the time-lagged covariance matrix as follows:

$$C_0 = (X_t - \mu)^2 (X_t - \mu)$$

$$C_\tau = (X_t - \mu)^2 (X_{t+\tau} - \mu) \tag{2.3.1}$$

The eigenvalue problem is then solved:

$$C_\tau r_i = C_0 \lambda_i r_i \tag{2.3.2}$$

r_i represents the independent components and λ_i are the normalized time-auto correlations. The eigenvalues and relaxation timescale are coupled by:

$$t_i = -\frac{\tau}{\ln |\lambda_i|} \tag{2.3.3}$$

TICA performs a dimension reduction by projecting the input data onto the slowest TICA components [30] [29].

2.3.2 Construction of Microstates

For grouping relevant structures together from the MD trajectories, the clustering method k-means is used. K-means defines geometric boundaries using a structural metric. Implementing a clustering method is important for creating a number of "microstates". Determining the optimal number of cluster centers, denoted by k, is often not immediately obvious. This decision is influenced by factors such as the distribution of our data, as well as the number of dimensions involved in our analysis. PyEMMA employs a machine learning approach that generates an output score (VAMP-2 score) to determine the most appropriate number of clusters. The optimal cluster number is selected based on the point where the VAMP-2 score reaches a saturation point [31].

The aim of the clustering step is to group together molecular configurations that exhibit a high degree of structural similarity. This allows subsequent grouping of microstates into larger kinetic clusters, which is achieved by constructing a transition matrix [28].

2.3.3 Building a Microstate Transition Matrix

A microstate transition matrix is needed for recognizing timescales between kinetical relevant microstates. After constructing a set of microstates using a clustering method, a transition matrix must be built. To construct a transition matrix, the number of transitions between each pair of microstates (i and j) at a specific lag time must be counted. The next step involves normalizing the transition counts [28], which gives the probabilities of transitioning from one microstate (i) to another (j) at a specific lag time. This process is often repeated for various lag times, and a Chapman-Kolmogorov test is utilized to assess whether the system behaves in a Markovian manner, meaning if it is memoryless [30].

2.3.4 Coarse-grained Representation

Once the microstates transition matrix is established a coarse-grained MSM can be constructed. In this step, the transition matrix of microstates is simplified by projecting the microstates to a set of metastable states, a process known as coarse-graining [28]. To build this coarse-grained visualization model, clustering methods which use eigenvalues and eigenvectors to find kinetically similar states are important. The PCCA+, which is a spectral clustering algorithm, looks at the eigenvectors from the transition matrix and uses them to cluster the microstates into a few macrostates [32]. In this way a coarse-grained MSM is computed, which can be interpreted by the human eye.

2.3.5 Analysis of the Coarse-grained Markov State Model

Upon reaching this stage, it is desirable to identify the molecular structures associated with the metastable states that have been identified. To achieve this, representative sample structures are generated for each microstate and stored in a trajectory file, which can be examined visually using external software.

With knowledge of the metastable states obtained through PCCA+, we can also derive the mean first passage times (MFPTs) between them. This coarse-grained representation of the system´s dynamics is more manageable by humans. Despite this simplification, several intriguing properties can still be computed, such as the stationary distribution, which conveys

information about the free energy of each state. This is accomplished by summing all contributions to a coarse-grained state, S_i, as follows [32]:

$$G_i = -k_B T ln \sum_{j \in S_i} \pi_j \quad (2.3.4)$$

3

Results and Discussion

The main objective of this thesis was to investigate the structural changes of the gate to the hydrophobic cavity where ceramide binds. To achieve this, molecular simulations were employed, allowing for the measurement of various quantities over the simulation trajectory in nanosecond. By combining the analysis of RMSD, RMSF, and distance calculations, a comprehensive understanding of the protein´s dynamics and behaviour was developed. Additionally, the impact of introduced mutation on the stability of the systems was assessed. Markov state models (MSMs) served as a valuable tool for presenting and analysing specific MD simulations on a slower timescale. By utilizing MSMs, the resulting trajectories of computed simulation data became clearer to understand and interpret.

On the course of the analysis, plots of the simulations follow a colour scheme of purple for WT apo proteins, and orange for WT holo proteins.

3.1 Validation of Simulations

The first step was to determine the point at which the MD simulations reached equilibration. By identifying the equilibration point, it can be determined whether the simulation has reached a steady state, and if the data obtained is representative of the system.

In addition to determining the equilibration of the MD simulations, it was crucial to evaluate the precision and reliability of both the simulation and experimental RMSF results. By doing so, any potential variation can be identified between the simulated and experimental data. This also helps to ensure that any observations made from the simulation are reliable.

3.1.1 System Equilibration

The RMSD analysis was calculated for the MD simulations fitted to the backbone of the C_α atoms in the proteins with a duration of 632 to 801 ns depending on the simulations (Cf. Table 2.2.1). To compare the influence of the ceramide in the wild-type proteins in apo form and holo form RMSD plots was computed for systems 1-5, 6-9, 10-12 and 13-15.

The RMSD analysis for the complete duration gives information about when the systems are equilibrated. Examining the data after this point ensures that the obtained data is representative of the systems under study. The four different output plots for the systems with their total duration (ns) are shown in Figure 3.1.1, Figure 3.1.2, Figure 3.1.3, and Figure 3.1.4.

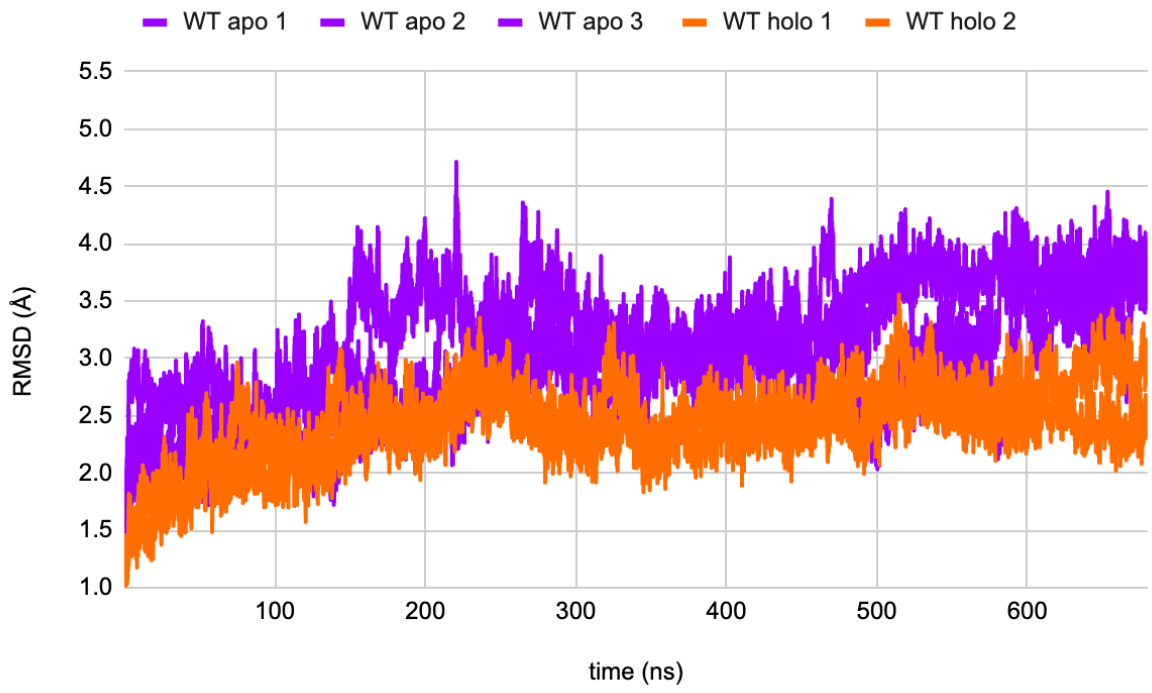

Figure 3.1.1: RMSD plot of WT apo 1, WT apo 2, WT apo 3, WT holo 1 and WT holo 2 with a duration of 699 ns.

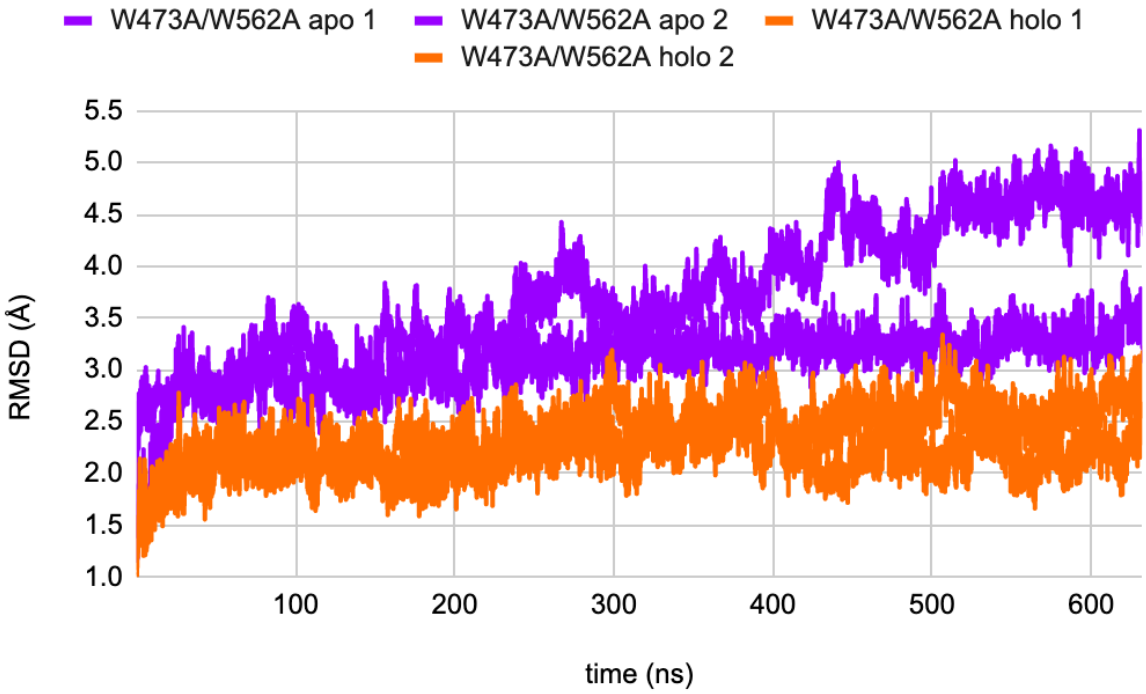

Figure 3.1.2: The RMSD plot for the duration of 623 ns shows the W473/W562A apo 1, W473A/W562A apo 2, W473A/W562A holo 1 and W473A/W562A holo 2 simulations.

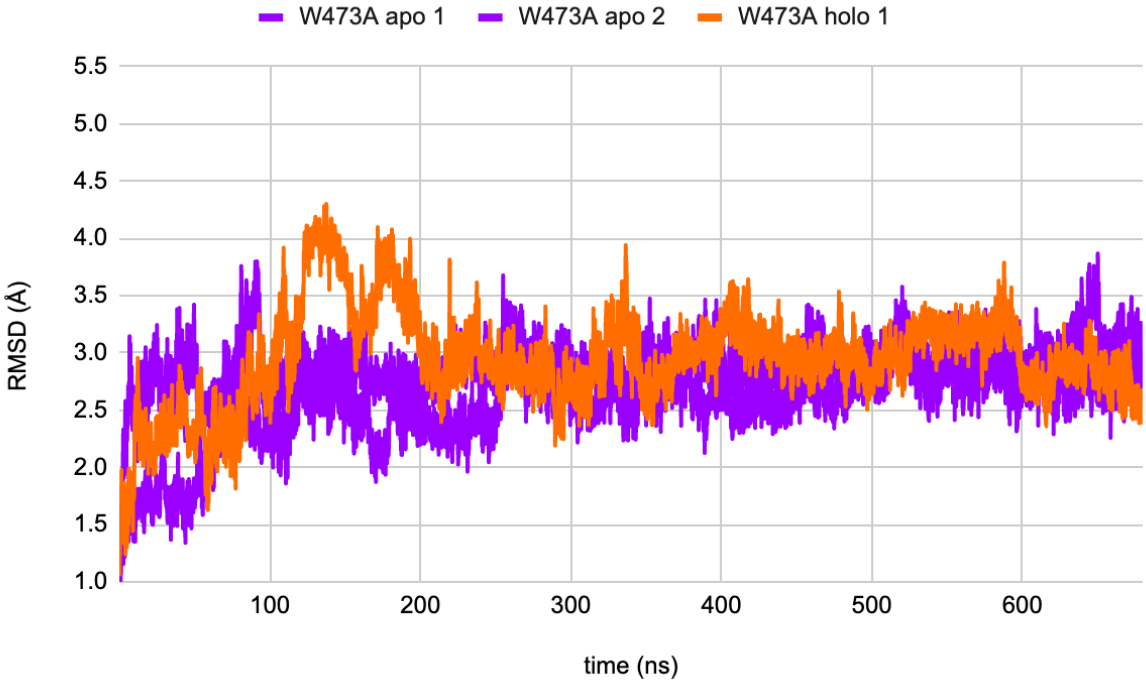

Figure 3.1.3: RMSD plot of W473A apo 1, W473A apo 2, and W473A holo 1 with a duration of 699 ns.

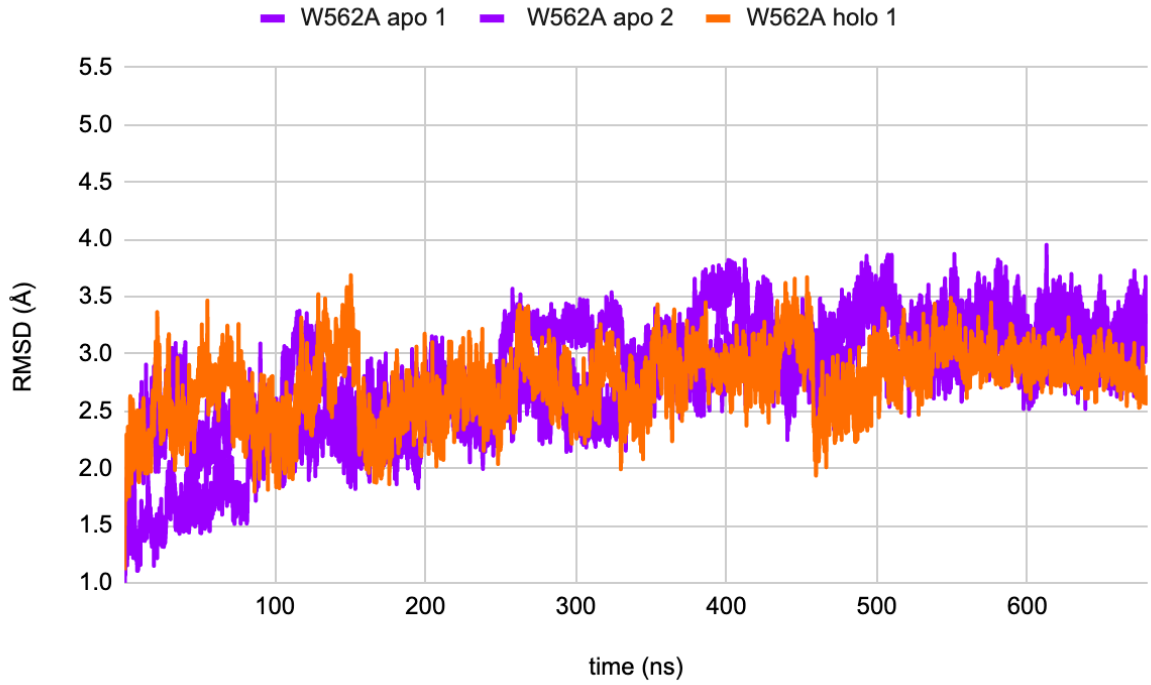

Figure 3.1.4: The RMSD plot for a duration of 699 ns displays the W562A apo 1, W562 apo 2 and W562A holo 1 structures.

The RMSD fluctuations reached a stable state after 250 ns for all the systems except the double mutation plot (Fig. 3.1.2), where we can see that the W473A/W562A apo 1 does not stabilize. Based on this observation, we can conclude that the three systems (Fig. 3.1.1, 3.1.3 and 3.1.4), along with W473A/W562A apo 2, W473A/W562A holo 1 and W473A/W562A holo 2 in Figure 3.1.2, have successfully achieved equilibration. Further analysis could be carried out using simulation data from this point onwards.

3.1.2 Matching Stationery Points for Experimental and Simulation RMSF Data

Experimental RMSF was calculated for WT apo 1, and WT holo 1 using the B-factor from the X-ray structures. Similarly, simulation RMSF was calculated for WT apo 1, and WT holo 1. To ensure the accuracy of atomic fluctuation measurements, it is crucial to compare the simulation RMSF with experimental RMSF results [23]. The experimental values and simulations values were plotted over protein positions to gain a visual understanding of the fluctuation throughout the protein structure. Figure 3.1.5 shows the experimental RMSF values and simulation RMSF values for all regions.

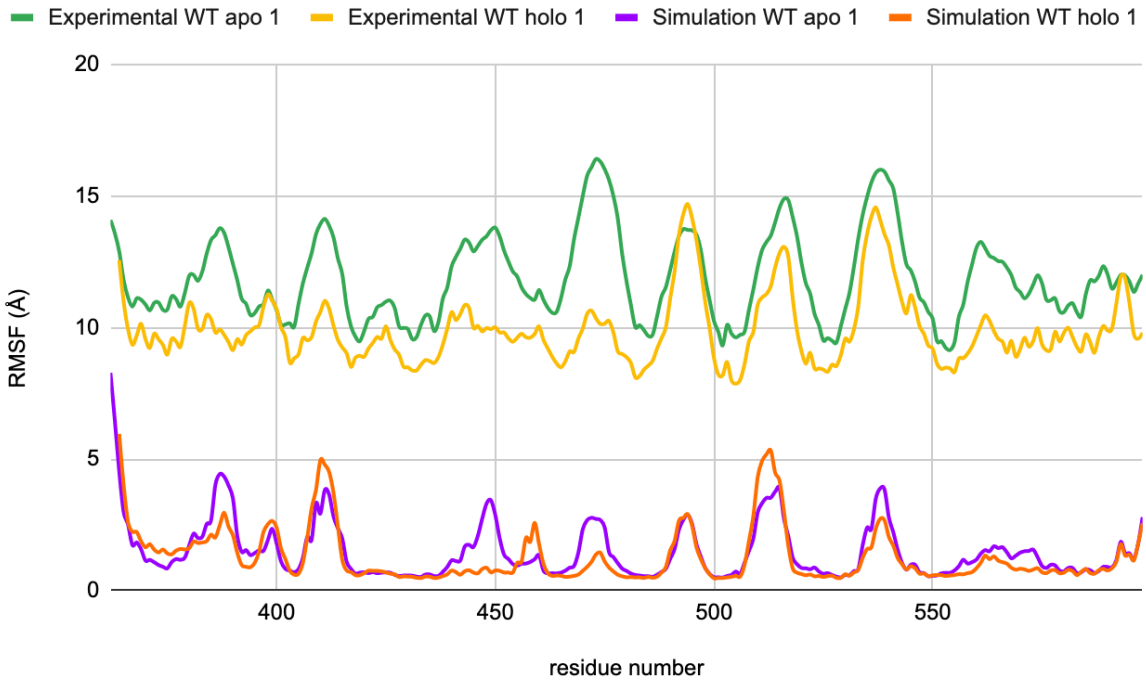

Figure 3.1.5: Experimental RMSF values on Cα for WT apo 1 and WT holo 1, and simulation RMSF values on Cα for WT apo 1 and WT holo 1. Computed for 250-699 ns of simulation. The lines are experimental WT apo 1 (green) experimental WT holo 1 (yellow), simulation WT apo 1 (purple) and simulation WT holo 1 (orange).

The comparison of the experimental and simulation RMSF values has yielded an important observation. Specifically, it was found that the positions of the maxima and minima points in the protein sequence are the same for both datasets. This indicates a level of agreement between the estimated RMSF values obtained from our simulations and those derived experimentally. This agreement between RMSF values is significant because it validates the simulation methodology. The comparison of simulation RMSF and experimental B-factors can enhance the level of confidence in the conclusions derived from the simulations.

3.2 WT Holo Shows Higher Structural Stability Compared to WT Apo

Based on the RMSD analysis presented in section 3.1.1, we selected the data after 250 ns to further investigate the simulation after equilibration was completed. For easy reading of this section, we show here the plots presented in sections 3.2.1 but only for simulation time after 250 ns (Fig. 3.2.1), for wild type simulations for both the apo and holo forms of the protein (systems 1-5).

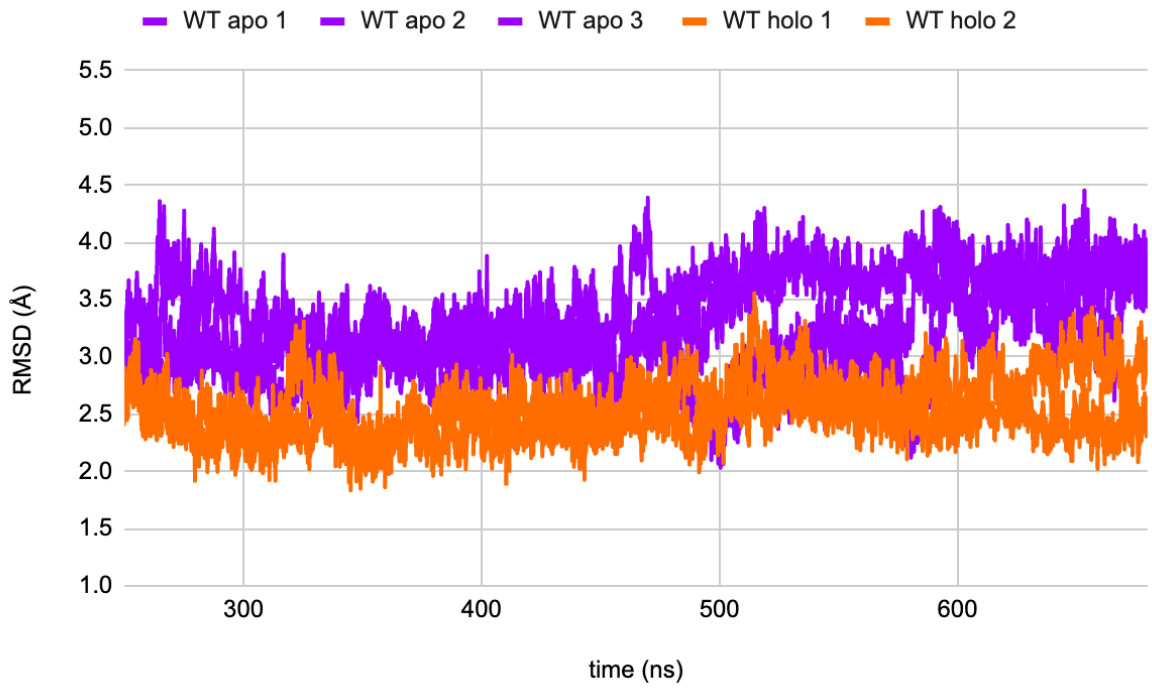

Figure 3.2.1: RMSD plot on Cα-atoms values for WT apo 1, WT apo 2, WT apo 3, WT holo 1 and WT holo 2. Computed for 250-699 ns of simulation.

The analysis revealed interesting differences between the RMSD value for WT apo forms and WT holo forms. Specifically, it was observed that the RMSD values were consistently higher for WT apo 1, 2 and 3 compared to WT holo 1 and 2. These results suggest greater structural stability due to the binding of ligand, which may play a crucial role in stabilizing the protein structure.

The distance between tryptophan 473 and proline 564 amino acids and between serine 476 and tryptophan 562 was plotted over time for WT apo 1, 2 and 3, and WT holo 1 and 2 (Fig. 3.2.2 and 3.2.3). This was performed to gain insight into the distances between loops Ω1 and Ω4 in the protein, which form the entrance to the hydrophobic cavity. High values for those distances are representative of an open gate.

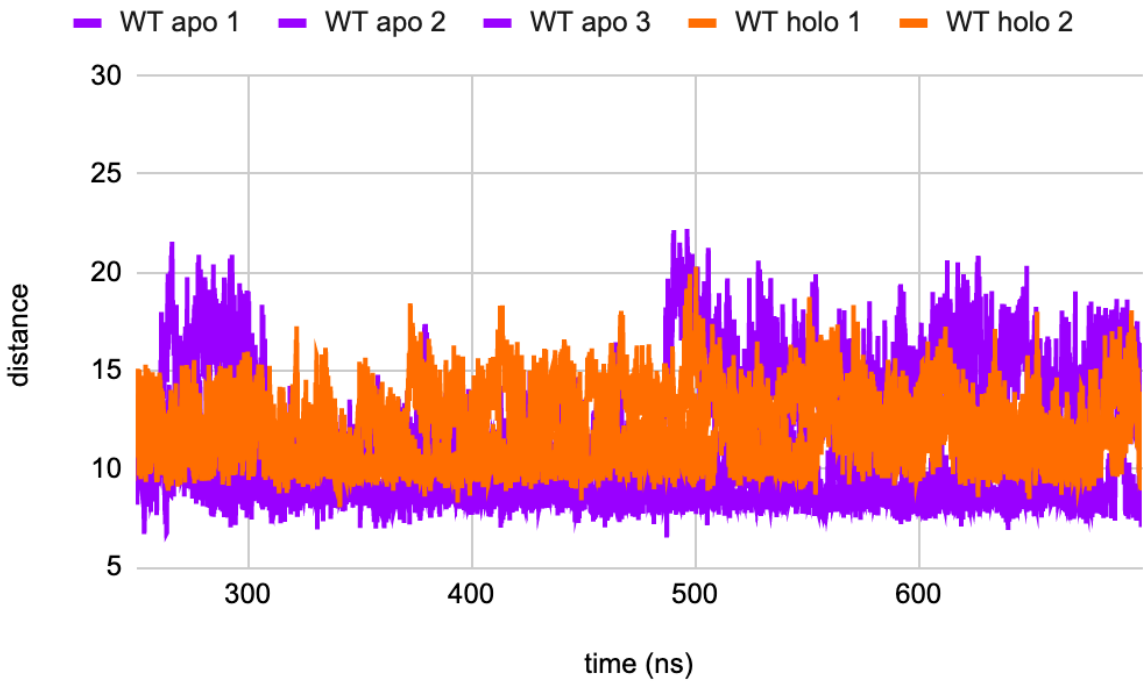

Figure 3.2.2: Plot of distances between tryptophan 473 and proline 564 residues for WT apo 1, WT apo 2, WT apo 3, WT holo 1 and WT holo 2. Computed from 250 to 699 ns of simulation.

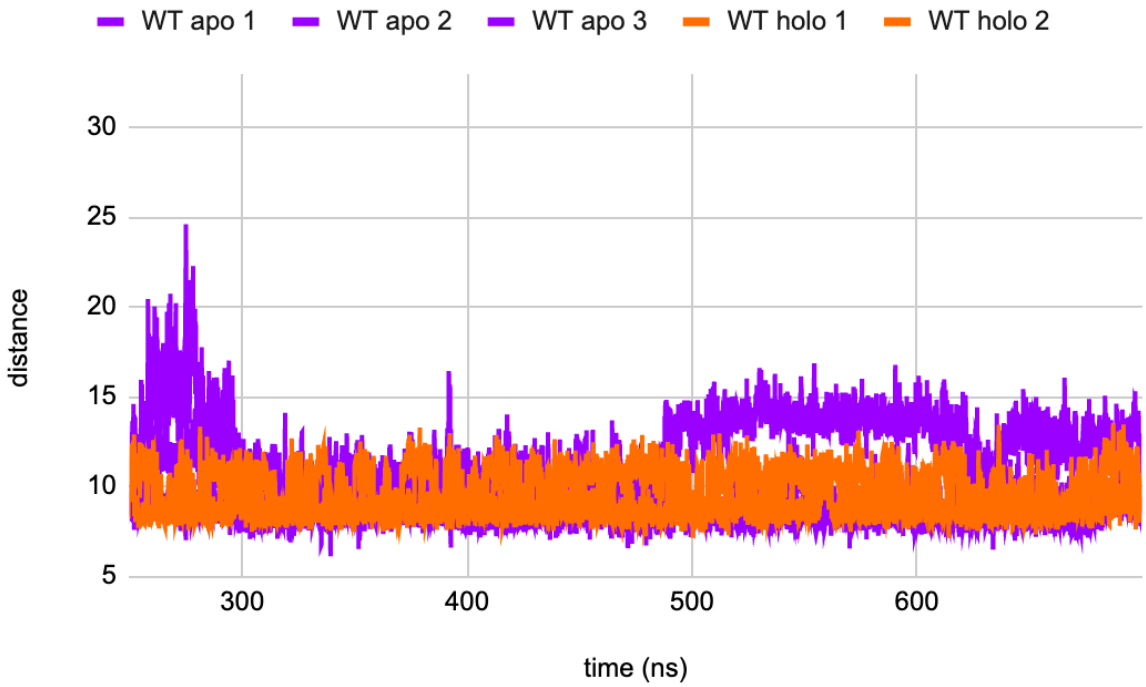

Figure 3.2.3: Plot of distances between serine 476 and tryptophan 562 residues for WT apo 1 WT apo 2, WT apo 3 and WT holo 1 and WT holo 2. Computed for 250-699 ns of simulation.

The distance fluctuations were found to be highest for WT apo 3, Notably, the distance between tryptophan 473 and proline 564 exhibits the highest fluctuation. In contrast, WT apo 1 and WT apo 2 show a similar pattern of distance fluctuation across all residues, suggesting a more stable conformation compared to WT apo 3. However, the results for WT apo 3 indicate that the apo protein may experience an increase in distances between mentioned residues. Unfortunately, we cannot firmly conclude this since it is observed in any of replicas for this apo protein. The plot reveals that in the case of WT holo 1 and 2, the distance between residues 473 and 564 exhibits greater fluctuation compared to WT apo 1 and WT apo 2.

To gain insight on specific parts of the protein, RMSF analysis was conducted. These values were plotted for wild type simulations (systems 1-5) over residue number. Figure 3.2.4 displays RMSF values for WT apo 1, 2 and 3, and WT holo 1 and 2.

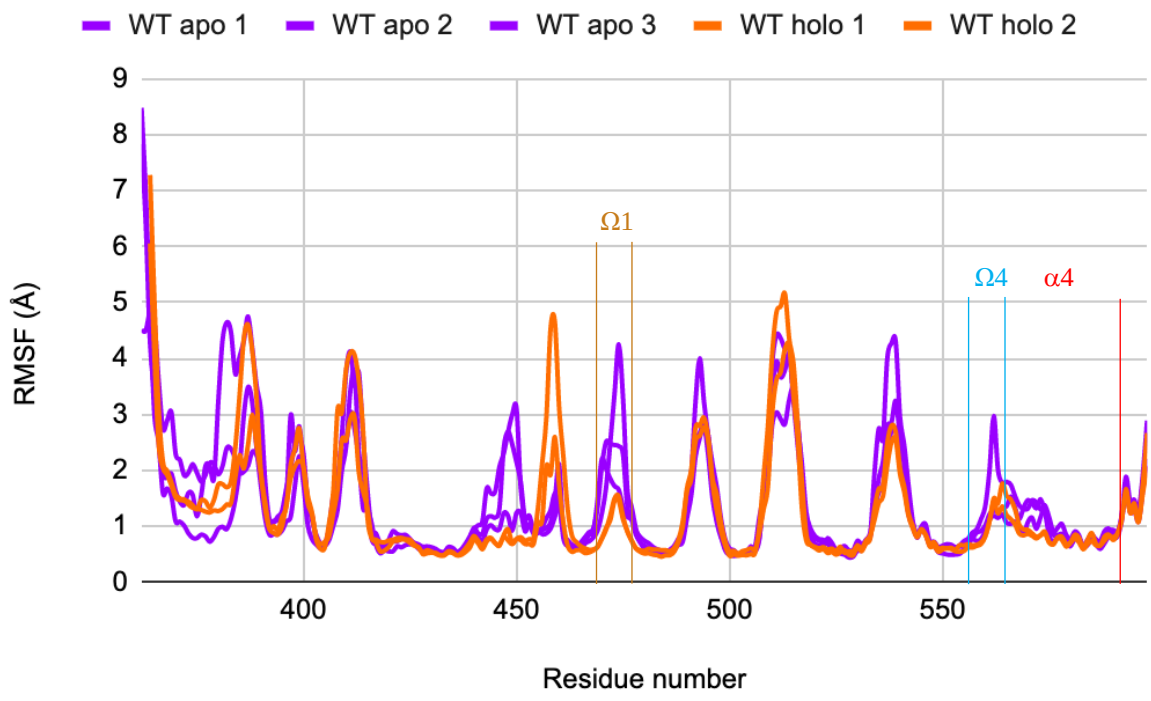

Figure 3.2.4: RMSF plot on Cα values for WT apo 1, WT apo 2, WT holo 1 and WT holo 2. Computed from 250 to 699 ns of simulation.

After analysing output plot (fig. 3.2.4), it can be observed that the RMSF values for the WT apo 3 simulation is higher than WT apo- 1 and 2, and WT holo- 1 and 2 simulations in Ω1- and Ω4- loop regions. Upon closer inspection of WT apo- 1 and 2, and WT holo 1- and 2 the RMSF values are quite similar in the Ω4-loop. In contrast, the WT apo 1 and 2 values are higher than those for WT holo simulations in the Ω1-loop. Other values that stand out for WT

apo- 1 and 2 are at residue number 450, which show an increase compared to WT apo 3. Additionally, WT apo 1 has a higher RMSF value at residue number 539 compared to the other simulations. At residue number 459 we observe that the WT holo 2 simulations have an increase in RMSF values compared to the three other complexes.

In general, the WT apo simulations consistently exhibited higher observed values in the RMSD plots. The distances fluctuations were particularly pronounced in the WT apo 3 system, indicating increased variability between the Ω1-loop and Ω4-loop regions. Moreover, the wild type apo simulations displayed higher and more widespread fluctuations in the RMSF compared to the wild type holo simulations. The higher structural stability observed in the WT holo proteins compared to the WT apo proteins could be attributed to the interaction between the WT holo protein and the bound ceramide molecule. This interaction may provide a higher degree of stability and compactness to the protein. In contrast, the absence of ceramide in the WT apo form causes the protein to adopt a less stable conformation, ultimately leading to higher values. These results highlight the effect of the ceramide presence in the WT holo forms in maintaining stability on the protein.

3.3 Impact of Double Mutation on Apo and Holo STARD11

RMSD analysis from simulations time 250 ns was plotted and output plots of double mutation analysis for apo and holo proteins (systems 6-9) are shown in Figure 3.3.1.

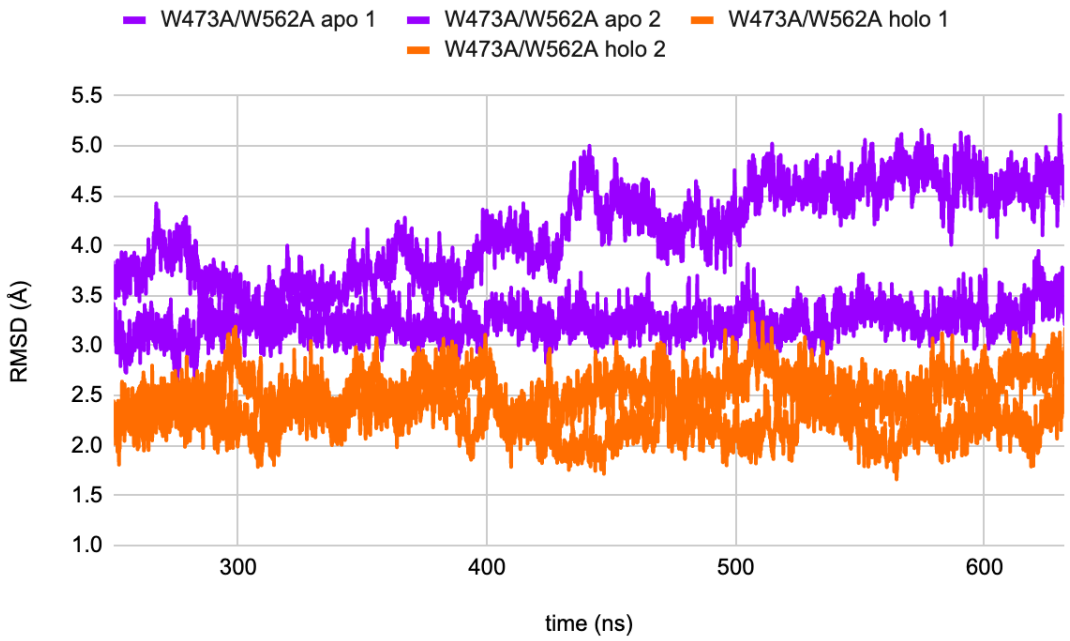

Figure 3.3.1: RMSD plot on Cα-atoms values for W473A/W562A apo 1, W473A/W562A apo 2, W473A/W562A holo 1 and W473A/W562A holo 2. Computed from 250 to 632 ns of simulation.

The results indicates that the RMSD values for the W473A/W562A apo proteins are consistently higher than those for the W473A/W562A holo proteins. Furthermore, the trajectory for W473A/W562A apo 1, fails to reach a stable state after 250 ns. This lack of stabilization suggests that additional simulation time may be necessary to accurately capture the behavior and dynamics of the W473A/W562A apo 1 system.

Distances between alanine 473 and proline 564 was plotted over duration for W473A/W562A apo forms and W473A/W562A holo forms (Fig. 3.3.2). Additionally, distances between serine 476 and alanine 562 are illustrated in Figure 3.3.3.

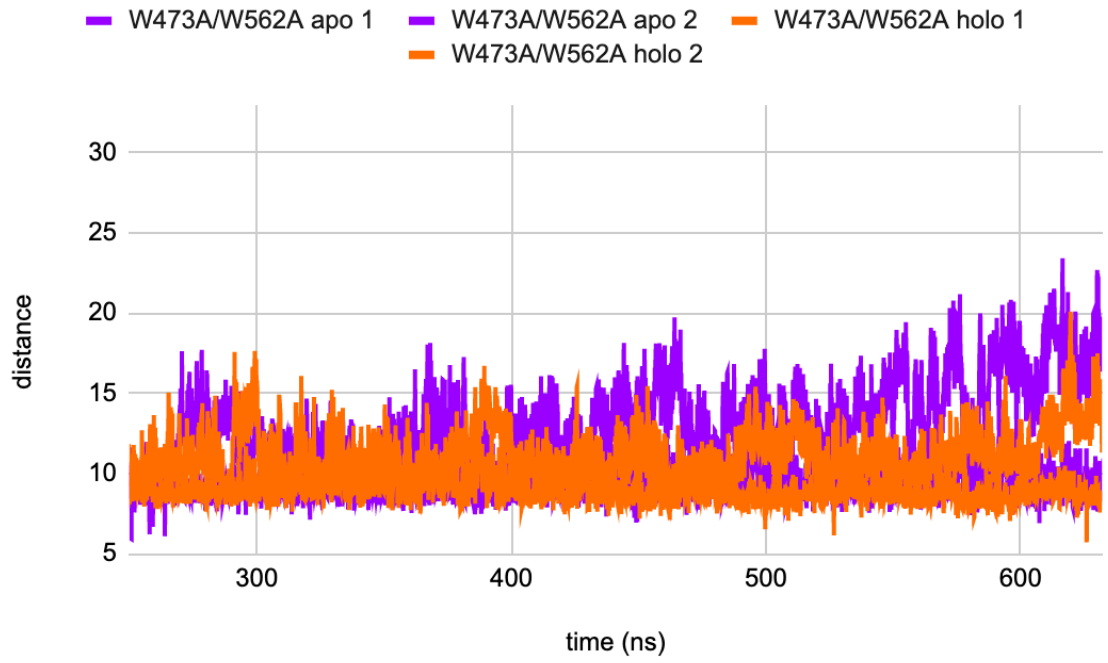

Figure 3.3.2: Plot of distances between alanine 473 and proline 564 residues for W473A/W562A apo 1, W473A/W562A apo 2 and W473A/W562A holo 1 and W473A/W562A holo 2. Computed for 250-632 ns of simulation run.

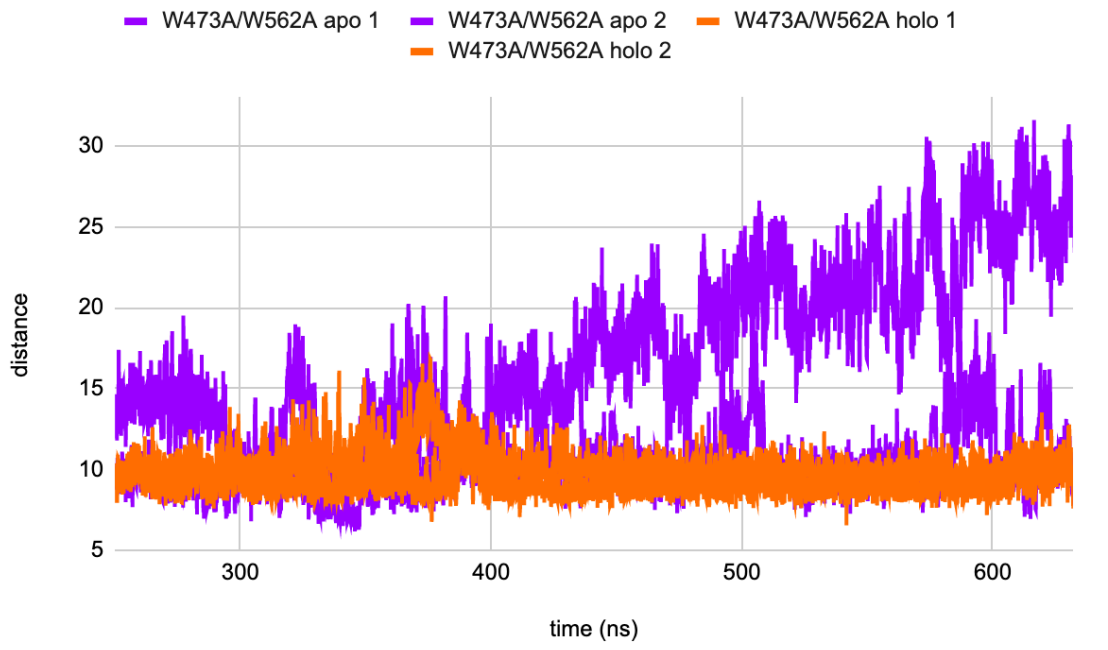

Figure 3.3.3: The plot shows the distances between serine 476 and alanine 562 for four simulations: W473A/W562A apo 1, W473A/W562A apo 2 and W473A/W562A holo 1 and W473A/W562A holo 2. Computed for 250-632 ns of simulation run.

Analysis of the distance charts (Fig. 3.3.2 and Fig. 3.3.3) shows that the distances between alanine 473 and proline 564, as well as the distance between serine 476 and alanine 562 are highest for W473A/W562A apo 1 compared to other simulations. Considering the lack of stability observed in the W473A/W562A apo 1 system, it is important to interpret this with caution, since if may not accurately represent the true behavior of this system. In the other simulations similar fluctuations were observed, indicating consistent behavior across both distances.

Values from RMSF analysis were plotted for double mutation simulations (systems 6-9) with residue numbers on the x-axis. Figure 3.3.4 shows simulation RMSF values for all protein regions of W473A/W564A apo 1, W473A/W564A apo 2, W473A/W564A holo 1 and W473A/W564A holo 2.

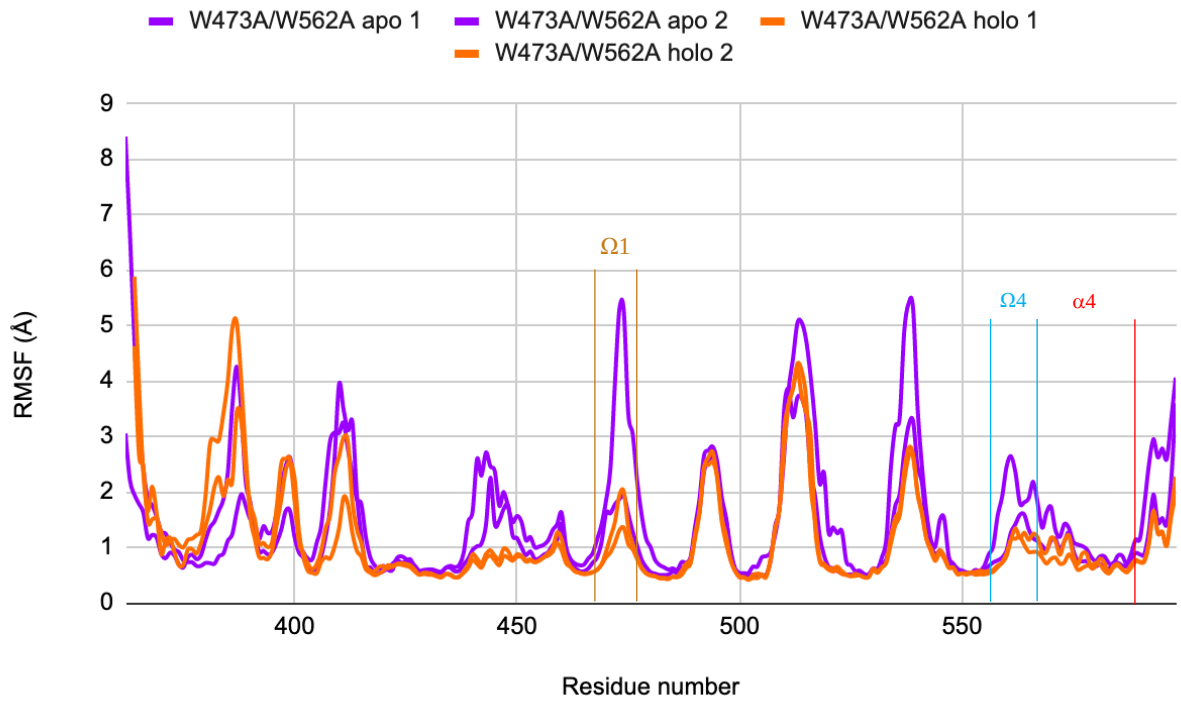

Figure 3.3.4: RMSF plot on Cα values for four simulations: W473A/W562A apo 1, W473A/W562A apo 2, W473A/W562A holo 1 and W473A/W562A holo 2. Computed for 250-632 ns of simulation.

Upon closer examination of the RMSF plot for the double mutation experiments, we can see that the Ω1-loop region exhibits the highest value in the W473A/W562A apo 1 simulation, which is consistent with the distance chart presented in Figure 3.3.2. In contrast, the RMSF value for W473A/W562A apo 2 is only slightly higher than for W473A/W562A holo- 1 and 2 at Ω1-loop region. Furthermore, W473A/W562A apo 1 once again exhibits the highest value

in the Ω4-loop, which is again consistent with the findings from the distance chart presented in Figure 3.3.3. At residue number 539 the W473A/W562A apo 2 displays highest RMSF values. Due to the observed lack of stability in W473A/W562A apo 1 system we must again consider the results of this simulations with caution.

Figure 3.3.5 display the RMSF values for the wild type structures for both the apo and holo forms, as well as the structures with double mutation in the apo and holo forms. This is plotted to compare the wild type apo and wild type apo with double mutation to wild type holo and holo with double mutation.

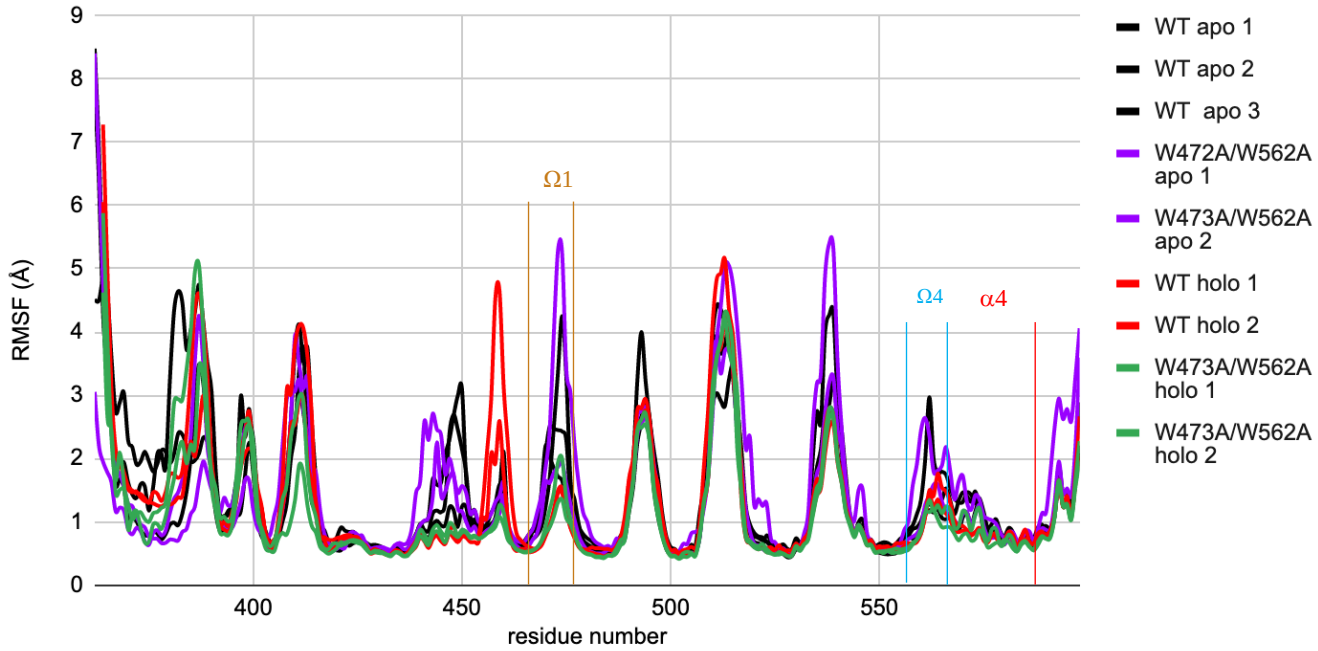

Figure 3.3.5: RMSF plot on Cα values for both the wild types and double mutation simulations in the apo and holo forms: WT apo 1, WT apo 2, WT apo 3, W473A/W562A apo 1, W473A/W562A apo 2, WT holo 1, WT holo 2, W473A/W562A holo 1 and W473A/W562A holo 2. Computed for 250-632 ns of simulation.

By comparing the WT apo and WT holo forms with the double mutation simulations, interesting similarities are observed, despite the observed inaccuracy for the W473A/W562A apo 1 simulation. Specifically, the comparison reveals that the WT apo structures (black lines) and the double mutation apo structures (purple lines) shows a similar pattern when comparing it to the WT holo structures (red lines) and the holo structures with the double mutation (green lines) in Ω1-loop, Ω-4 loop and α4-helix regions. One simulation that stands out is at residue

number 459 for simulations WT holo 2, which show a higher RMSF value compared to all the other simulations. These observation points to a similar conformational change and dynamics of the apo form and the holo form when introducing double mutation on residues 473 and 562. In other words, regardless of whether the proteins are in the apo or holo form, a consistent trend in the conformational dynamics in the Ω1-loop, Ω-4 loop and α4-helix regions are observed between the wild type and double mutation simulations.

3.4 Impact of Single Mutations on Apo and Holo STARD11

RMSD values was plotted with duration on the x-axis. Figure 3.4.1 illustrates the output plot for the mutation on amino acid 473 from tryptophan to alanine in apo- 1 and 2, and holo 1 (systems 10-12). Additionally, Figure 3.4.2 represents the output plot for the mutation on residue number 562 from tryptophan to alanine in apo 1- and 2, and holo 1 (systems 13-15).

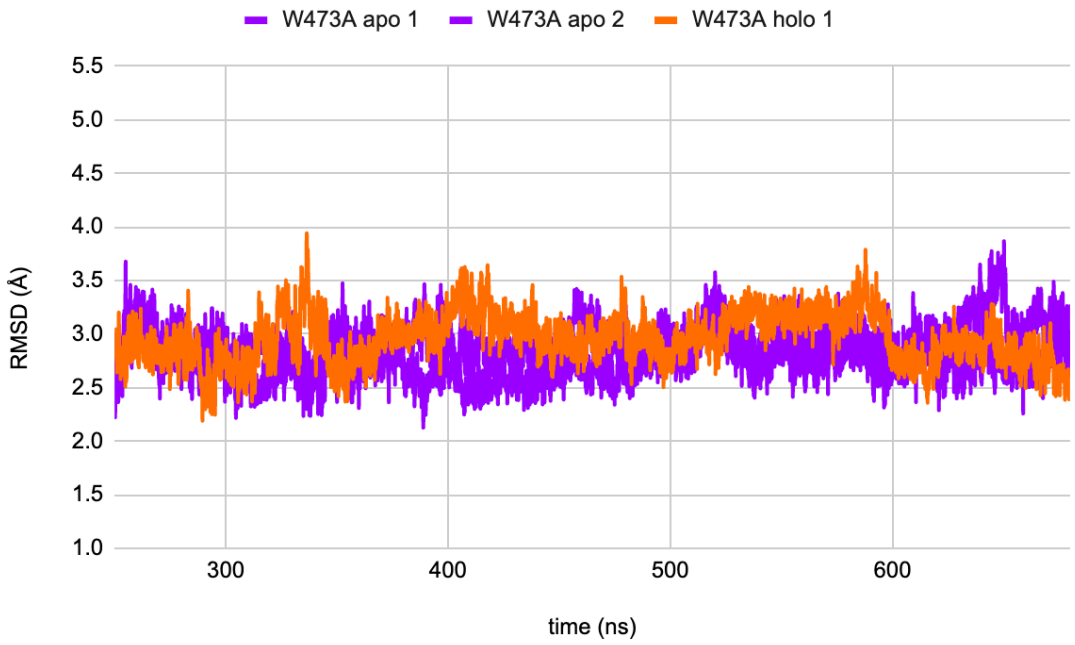

Figure 3.4.1: RMSD plot on Cα-atoms values for W473A apo 1, W473A apo 2 and W473A holo 1. Computed for 250-699 ns of simulation.

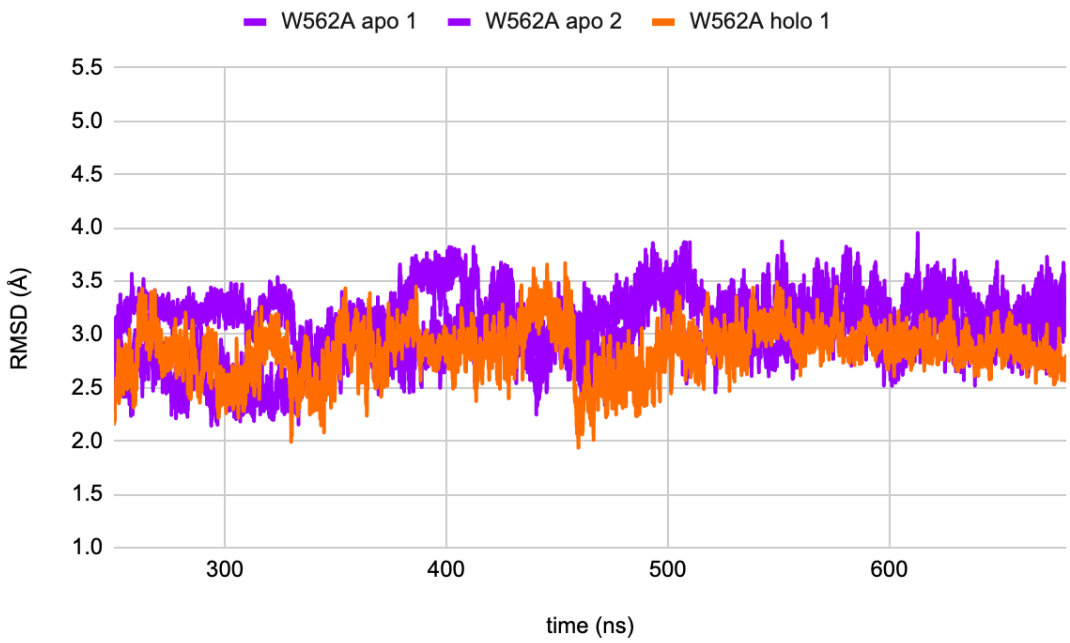

Figure 3.4.2: RMSD plot on Cα-atoms values for W562A apo 1, W562A apo 2 and W562A holo 1. Computed from 250 to 699 ns of simulation.

The RMSD plots generated for the W473A and W562A mutation did not reveal any significant differences in the RMSD fluctuation between the apo or holo complexes. These findings suggest that the impact of a single-residue mutation on the proteins conformational stability is relatively low. The lack of observed differences may be because altering a single tryptophan residue to alanine may not cause significant changes in the overall structural behaviour of the proteins.

Distances between alanine 473 and proline 564 were plotted over the simulation duration for W473A apo- 1 and 2, and W473A holo 1 (Fig. 3.4.3). In addition, distances between serine 476 and tryptophan 562 are illustrated (Fig. 3.4.4) for W473A apo- 1 and 2, and W473A holo 1.

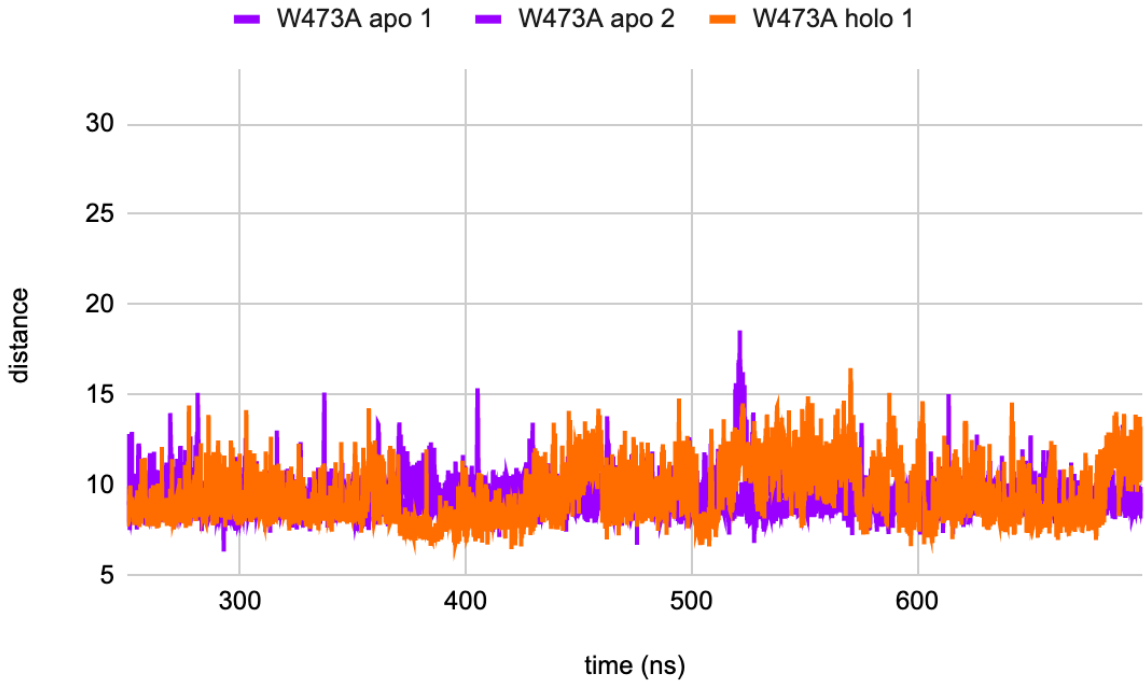

Figure 3.4.3: Plot of distances between alanine 473 and proline 564 residues for W473A apo 1, W473A apo 2 and W473A holo 1. Computed from 250 to 699 ns of simulation.

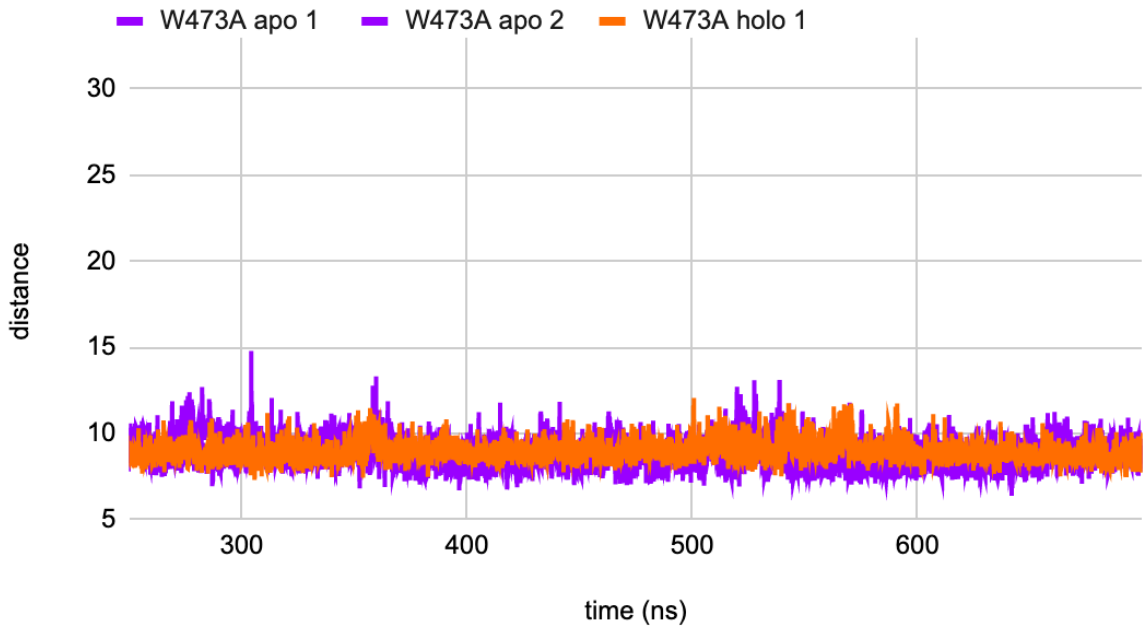

Figure 3.4.4: Distances plot between serine 476 and tryptophan 562 residues for W473A apo 1, W473A apo 2 and W473A holo 1. Computed for 250-699 ns of simulation run.

Examining the distance charts (Fig. 3.4.3) it appears that the distances between the amino acids alanine 473 and proline 564 exhibits a significant variance between 10-15 distance values. This implies that the distance between these amino acids have a significant fluctuates during the simulations. In contrast, the distance plot for serine 476 and tryptophan 562 (Fig. 3.4.4), remain more stable for both the W473A apo- 1 and 2 forms, and the W473A holo 1 form. This suggests that these distances exhibit a consistent level of stability through the simulations.

Distances between tryptophan 473 and proline 564, and distances between serine 476 and alanine 562 were plotted for W562A apo- 1 and 2, and W562A holo 1 (Fig. 3.4.5 and Fig. 3.4.6 respectively).

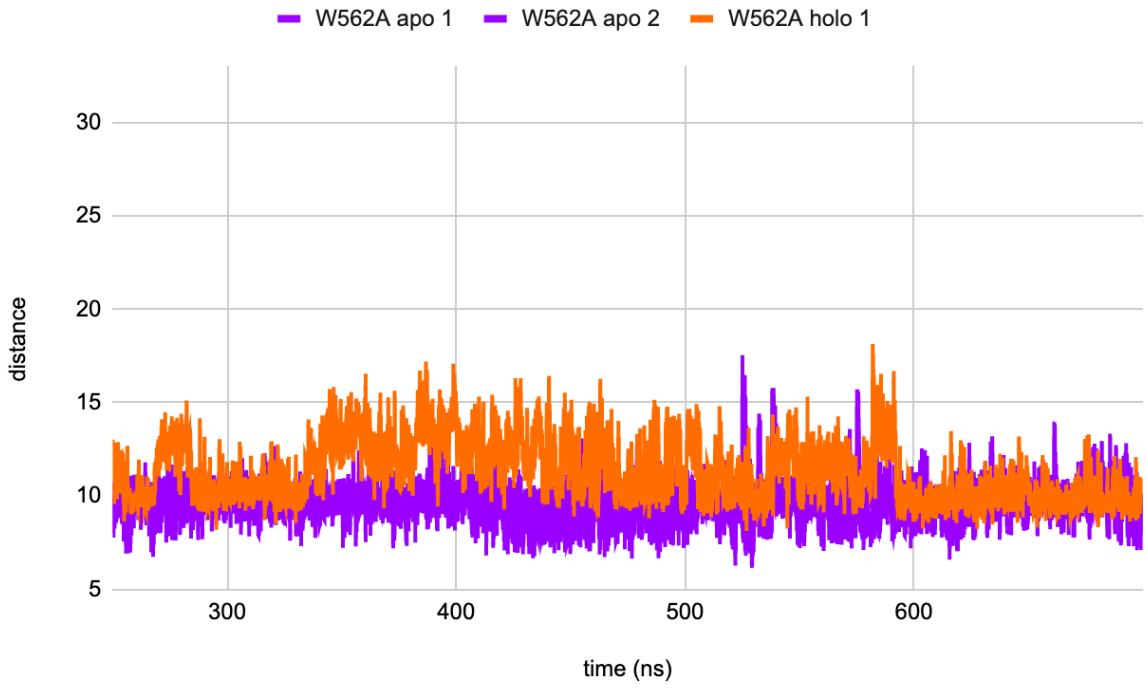

Figure 3.4.5: The plot displays the distances between tryptophan 473 and proline 564 residues for three simulations: W562A apo 1, W562A apo 2 and W562A holo 1. The distances were computed from 250 to 699 ns simulation.

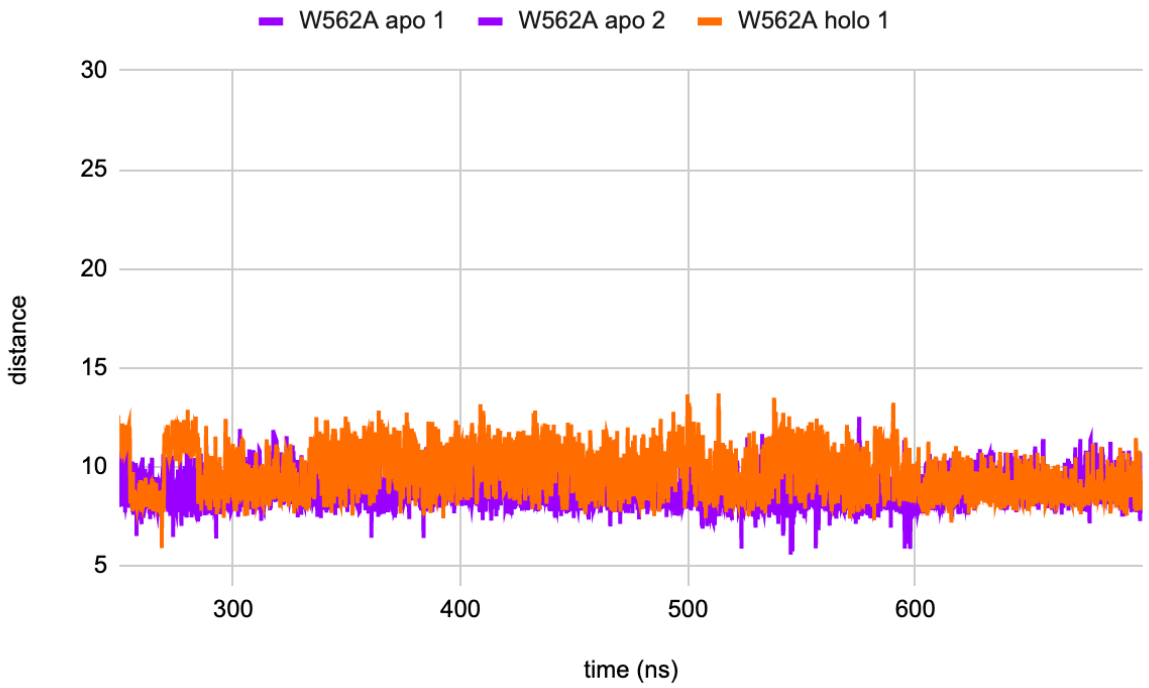

Figure 3.4.6: Plot of distances between serine 476 and alanine 562 residues for W562A apo 1, W562A apo 2 and W562A holo 1. Computed for 250-699 ns of simulation.

The distance chart for W562A mutations (Fig. 3.4.5) shows an increase in distance between tryptophan 473 and proline 564 residues for all the simulations. This increase in distance suggests that the mutation in amino acid 562 may affect the distance between these two residues. In contrast, the distance plot for serine 476 and alanine 562 (Fig. 3.4.6) did not show any significant difference for W562A apo 1, W562 apo 2 or W562A holo 1. This suggests that the W562A mutation may not have a substantial impact on the distance between these residues in the protein structure.

To visualize the results of the RMSF analysis obtained from single mutation simulations on the apo and holo proteins (systems 10-12), data was plotted along the protein positions. Figure 3.4.7 displays the RMSF values of all protein regions for W473A apo 1, W473A apo 2, and W473A holo 1.

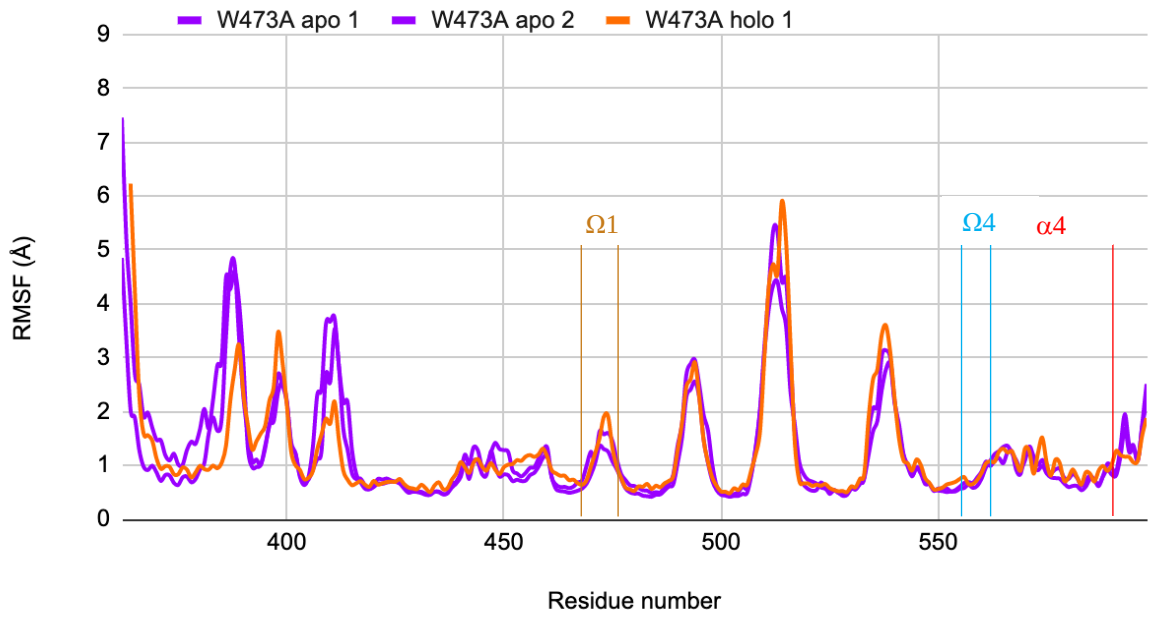

Figure 3.4.7: The plot shows RMSF values for Cα-atoms values for simulations: W473A apo 1, W473A apo 2 and W473A holo 1. Computed for 250-699 ns of simulation.

The RMSF plot for the W473A simulations (Fig. 3.4.7) reveals a consistent fluctuation in atomic positions across the simulations, with no significant deviations observed. This suggest that the RMSF fluctuations between the apo and holo proteins, specifically for the mutation at residue number 473 are relatively low. This indicates that the W473A mutation does not introduce significant changes in the RMSF fluctuation between the apo and holo proteins.

The RMSF analysis data was plotted for single mutation simulations on the apo and holo complexes (systems 13-15). Figure 3.4.8 illustrates the RMSF values for W562A apo 1 and 2, and W562A holo 1.

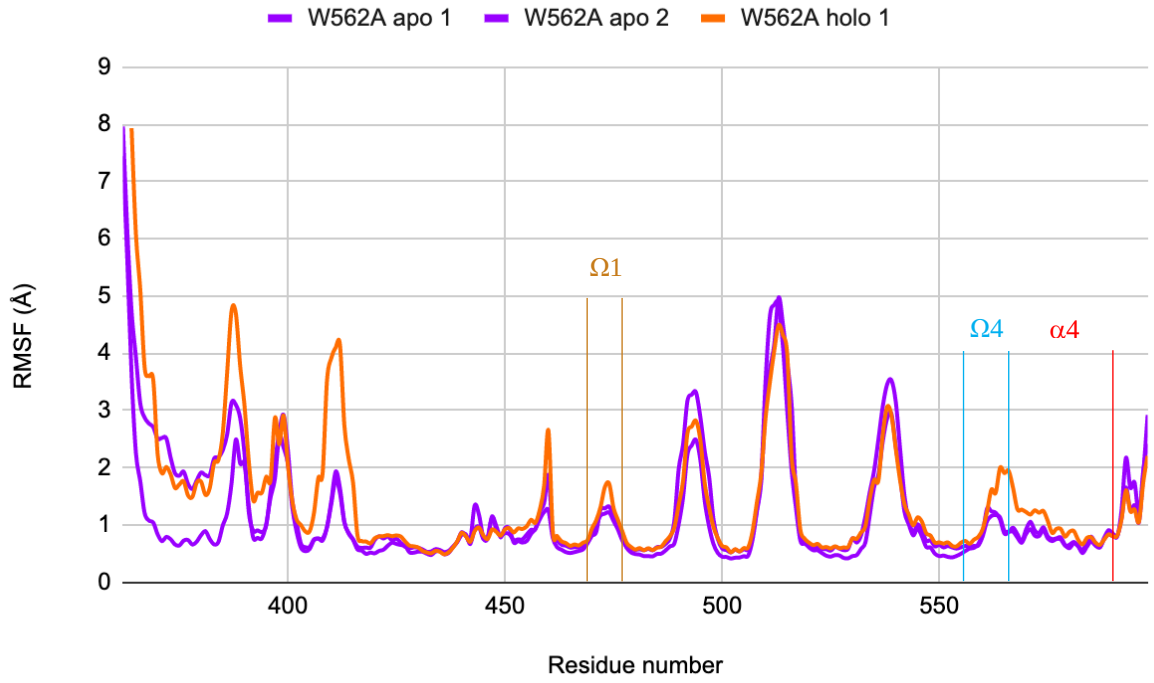

Figure 3.4.8: RMSF plot on Cα-atoms values for W562A apo 1, W562A apo 2 and W562A holo 1. Computed from 250 to 632 ns of simulation.

Generally, the RMSF values for the W562A simulations (Fig. 3.4.8) fluctuate around the same residue numbers. However, around the Ω1- and Ω4-loop the W562A holo 1 simulation shows a slight increase in RMSF value compared to the two other simulations. This corresponds to results from distances between tryptophan 473 and proline 564 (Figure 3.4.5). Overall, the plot shows no other significant deviations between the simulations.

In Figure 3.4.9, the RMSF values for both wild type simulations and single mutation simulations of the apo protein were plotted to obtain a comparison between them (systems 1-3, 10, 11, 13 and 14).

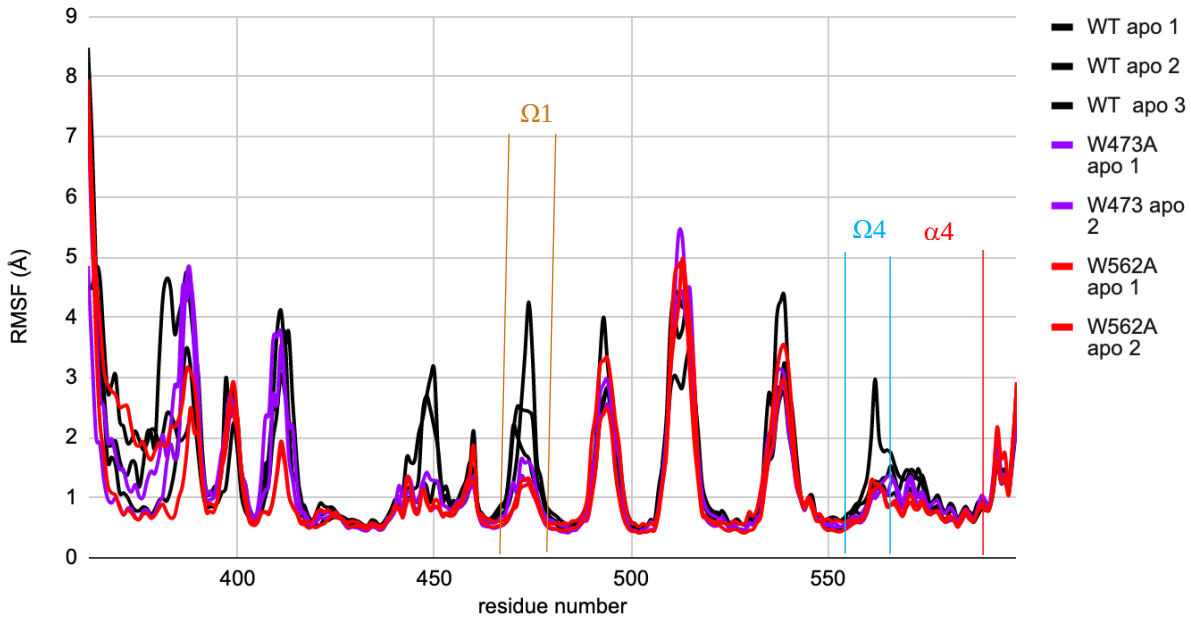

Figure 3.4.9: RMSF output plot on Cα-atoms values for wild type apo proteins. WT apo 1, 2 and 3 are illustrated with black lines. Purple lines indicates W473A apo- 1 and 2. Red lines illustrate W562A apo- 1 and 2. Trajectory is computed for 250-699 ns of simulation run.

This plot highlights some interesting trends. Notably, the fluctuations appear to occur around the same residue numbers for both the wild types and the single mutation simulations. However, an important finding is it exists significant differences between single mutations simulations, represented by the purple and red lines, and the wild type structures, represented with black line. In the Ω1 and Ω4-loop regions, as well as residue number 450, 493, 513 and 539 we can see divergence between these simulations. This indicates that the introduced single mutations has a notable impact on the flexibility of the apo protein. The increased RMSF values suggests that the mutation may introduce structural changes within these regions for the apo protein.

For comparing the RMSF fluctuations on single mutation and the wild type on holo protein Figure 3.4.10 was plotted (systems 4, 5, 12 and 15).

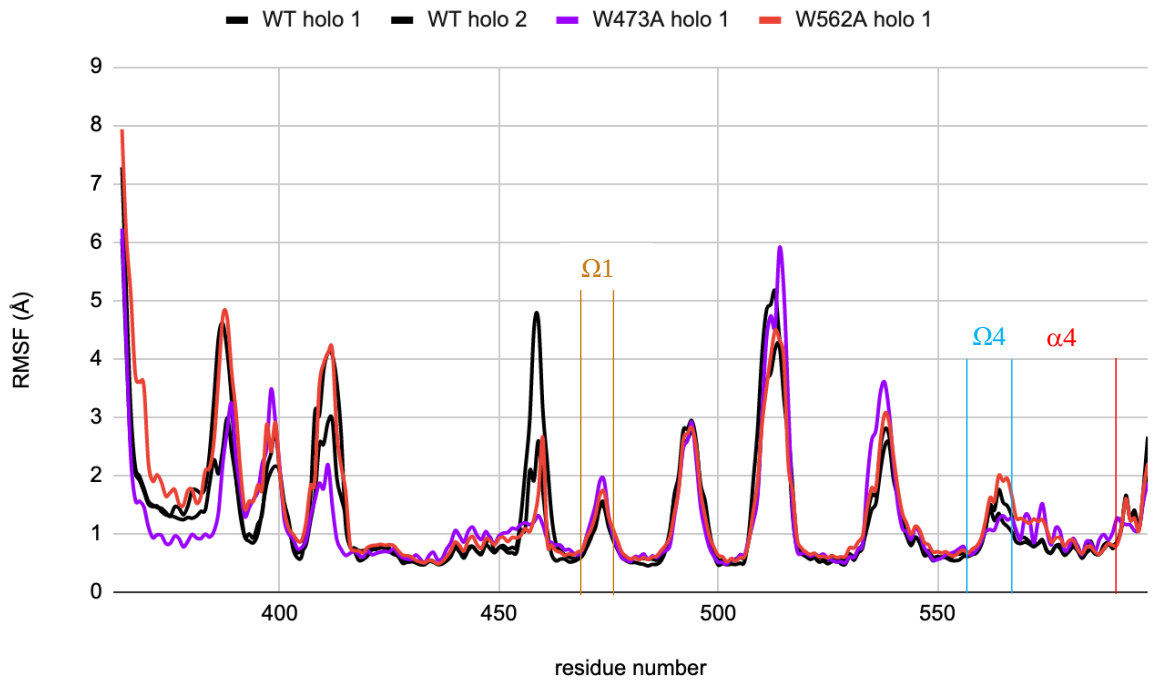

Figure 3.4.10: RMSF output plot on Cα-atoms values for wild type holo and single mutations on holo protein. WT holo 1- and 2 are illustrated with black lines. Purple lines indicates W473A holo 1. Red lines illustrate W562A holo 1. Trajectory is computed for 250-699 ns of simulation run.

Generally, the fluctuations appear to occur around the same residue numbers for both the wild types and the single mutation simulations with no significant difference in RMSF values on Ω1-loop, Ω4-loop. One noteworthy difference is the WT holo 2 at residue number 459 which has a higher RMSF values than the other simulations. This suggests that the introduction of single mutations on the holo complex does not lead to any notable variations in the Ω1-loop and Ω4-loop.

Figure 3.4.9 highlight substantial differences between the apo simulations of single mutation and the wild type structure in the Ω1-loop and Ω4-loop regions, as well as at residue numbers 450, 493, 513, and 539. These findings indicate that the introduced single mutation have a significant effect on the flexibility of the apo protein. The higher RMSF values strongly suggests that the single mutation induce structural alterations within these specific regions in the apo protein without ceramide presence. When comparing these findings to the distance charts in Figure 3.4.3 and Figure 3.4.5, it becomes evident that the distances between alanine 473 and proline 564 exhibits the most fluctuations. Figure 3.4.10 illustrates that there are no significant differences in the RMSF values observed for the Ω1-loop and Ω4-loop between

the wild type holo protein and the single mutations simulations. This indicates that the single mutations do not lead to substantial changes in the flexibility within these two loops. In other word, single mutations on the holo protein in the presence of ceramide do not significantly impact the protein structure.

In summary, the introduced single mutation affects the flexibility of the apo protein, as evidenced by higher RMSF values and structural alterations within specific regions. However, the mutations do not significantly alter the structural flexibility on the holo protein within the two loops. This suggest that ceramide presence in the holo protein may help stabilize the entrance to the hydrophobic cavity.

3.5 Evaluation of Simulation Results

The data obtained from all the 15 simulations has been compiled and presented in Table 3.5.1. The purpose of the table is to obtain a comparative analysis of the RMSD values and distance between the Ω1-loop and Ω4-loop. By examining these parameters, we can assess the variations and similarities in the structural dynamics of the systems.

Table 3.5.1: List of the 15 different systems simulated in this project with its corresponding average and standard deviation for RMSD, distance 1 (distance between resid 473 and resid 564), and distance 2 (distance between resid 476 and resid 562). WT apo 1, 2, and 3 represents wild type of 2E3M and WT holo 1 and 2 represents wild type of 2E3Q (CERT-STARD bound to ceramide). Double mutation on 2E3M and 2E3Q are denoted W473A/W562A apo 1, W473A/W562A apo 2, W473A/W562A holo 1 and W473A/W562A holo 2. Mutation analysis on structures 2E3M and 2E3Q was done on residues 473 and 562, denoted as W473A apo 1, W473A apo 2, W473A holo 1, W562A apo 1, W562A apo 2 and W562A holo 1.

System number	**Protein data bank ID**	**WT/mutant**	**RMSD (Å)**	**Distance 1 (Å)**	**Distance 2 (Å)**
1	2E3M	WT apo 1	3.21 ± 0.43	9.08 ± 1.13	10.11 ± 2.94
2	2E3M	WT apo 2	3.83 ± 0.44	9.49 ± 1.19	8.86 ± 0.66
3	2E3M	WT apo 3	3.06 ± 0.59	12.13 ± 2.98	10.64 ± 2.09
4	2E3Q	WT holo 1	2.36 ± 0.31	11.18 ± 1.59	9.00 ± 0.96
5	2E3Q	WT holo 2	2.49 ± 0.35	11.93 ± 1.91	9.56 ± 1.27
6	2E3M	W473A/W562A apo 1	3.86 ± 0.66	12.34 ± 2.98	16.29 ± 5.90
7	2E3M	W473A/W562A apo 2	3.11 ± 0.31	9.47 ± 1.01	10.34 ± 1.83
8	2E3Q	W473A/W562A holo 1	2.17 ± 0.24	9.06 ± 0.86	10.22 ± 1.04
9	2E3Q	W473A/W562A holo 2	2.43 ± 0.30	10.89 ± 1.55	10.89 ± 1.05
10	2E3M	W473A apo 1	2.94 ± 0.42	8.85 ± 0.48	9.00 ± 0.93
11	2E3M	W473A apo 2	2.83 ± 0.26	9.77 ± 1.01	8.84 ± 0.60
12	2E3Q	W473A holo 1	2.64 ± 0.40	9.62 ± 1.67	9.02 ± 0.75
13	2E3M	W562A apo 1	2.75 ± 0.30	9.81 ± 1.20	8.72 ± 0.90
14	2E3M	W562A apo 2	2.90 ± 0.45	9.59 ± 0.97	8.79 ± 0.65
15	2E3Q	W562A holo 1	2.71 ± 0.53	11.29 ± 1.65	9.66 ± 1.18

By analysing the plot, it is evident that the highest RMSD values are observed in the WT apo simulations (system 1, 2 and 3), with values exceeding 3 Å. However, the RMSD values for the holo proteins of the wild type (system 4 and 5) are significantly lower, with values under 2.5 Å. Examining the distances of these systems, it becomes apparent that distance 1 (between residue 473 and 564) exhibits higher values for WT holo 1 and 2 systems compared to the WT apo 1 and 2. This observation is further supported by Figure 3.2.2, which illustrates the same trend.

In the case of W473A/W562A apo 1, high RMSD values are observed, but this simulation did not reach stabilization. Similarly, W473/W562A apo 2 exhibits RMSD values higher than 3 Å. Comparing the RMSD values between the wild type apo structures and the double mutations apo structures, as well as between the wild type holo and the double mutation holo, it is apparent that the differences are not significant. These observations indicate a similar conformational change between the apo form and the holo form when introducing the double mutation at residue 473 and 562.

Upon comparing the single mutation apo proteins with the wild type apo proteins, lower RMSD values are observed for the W473A apo proteins and W562A proteins compared to the WT apo complexes. Notably, the distances that stand out are for WT apo 3, where the distances between residue 473 and 564, and between residues 476 and 562 are greater than those observed in the W473A apo and W562A apo complexes. When examining the single mutation holo proteins and the wild type holo proteins, no significant deviations in the RMSD values are observed. However, the distance 1 (between residue 473 and residue 564) is significantly lower for W473A holo 1 simulation.

3.6.1 Comparative Analysis of WT apo 1 and W473A/W562 apo 1 through Markov State Models

The Markov State Model analysis conducted in this study aimed to gain a more comprehensive understanding of the movement of the STARD11 protein in key regions, specifically the Ω1-loop, Ω4-loop, and α4-helix. By using MSMs, we were able to generate a transition matrix of these regions within the protein, providing insight into the probability of

the system being in a particular state and switching between states within a certain lag time. To obtain these MSMs, torsions angles were computed for specific regions within the protein. Specifically, the torsion angles Y466 to Y482 were computed for the Ω1-loop regions, and V555 to L580 for the Ω4-loop and α4-helix regions (see Fig. 1.1.4 and Fig. 1.1.5 for further details). Throughout the analysis, we used simulations for the WT apo 1 and W473A/W562A apo 1. The results of the analysis will provide valuable insights into the behavior of STARD11 for wild type, and double mutation without ceramide.

3.6.1 Dimensionality Reduction with TICA

The Time-lagged Independent Components Analysis (TICA) method was used to identify the slow components in a set of data. A total number of 70 TICA components were identified. TICA is a technique that reduces the dimensionality of high-dimensional data while preserving the essential features of the system dynamics. In this project, TICA was used to analyze the protein dynamics data obtained from our simulations. The output of TICA is a kinetic map that provides a visual representation of the slow components of the protein motion. Figure 3.6.1 shows the kinetic map scaling of the WT apo 1 protein and W473A/W562A apo 1, which illustrates the relative motion of different parts of the protein in a two-dimensional map spanned by the first two TICA components. The color scheme in the map represents the density, with yellow indicating regions of high density and purple representing regions of low density.

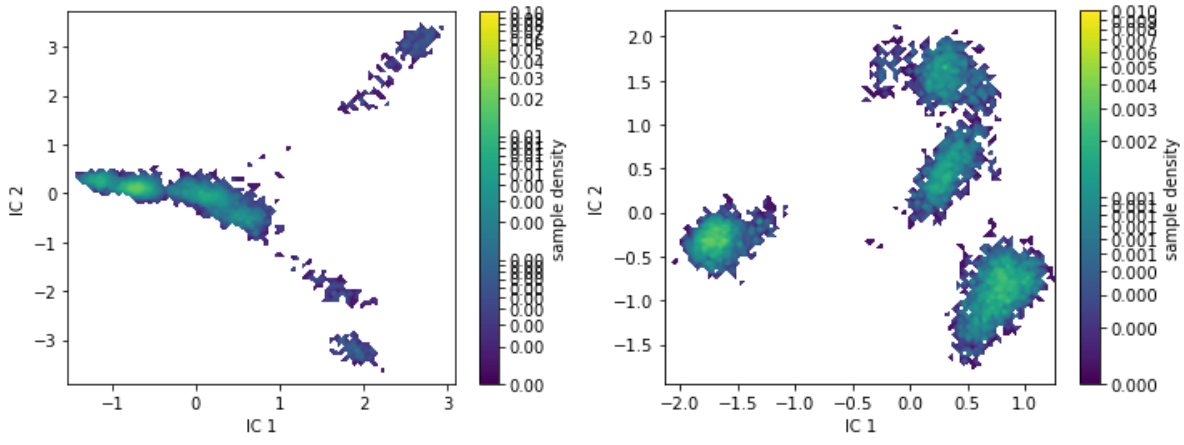

Figure 3.6.1: Visual representation of the slow components in the protein motions of WT apo 1 (left figure) and W473A/W562A apo 1 (right figure) through a kinetic map.

We look at the first two components, IC 2 and IC 1, for both simulations and found that TICA groups together areas with high density, which are probably areas where the systems stay in a relatively stable state for a long time.

3.6.2 Construction of Microstates

In the next step of the analysis, we aim to cluster the TICA data into distinct states using the k-means algorithm. This involves grouping together molecular configurations that exhibit high degree of structural similarity. This clustering into microstates is important for constructing a transition matrix that describes the transitions between different states. In Figure 3.6.2, the VAMP-2 score is plotted against number of cluster centers for both systems. The optimal number of the VAMP-2 score is chosen based on the point where it reaches saturation.

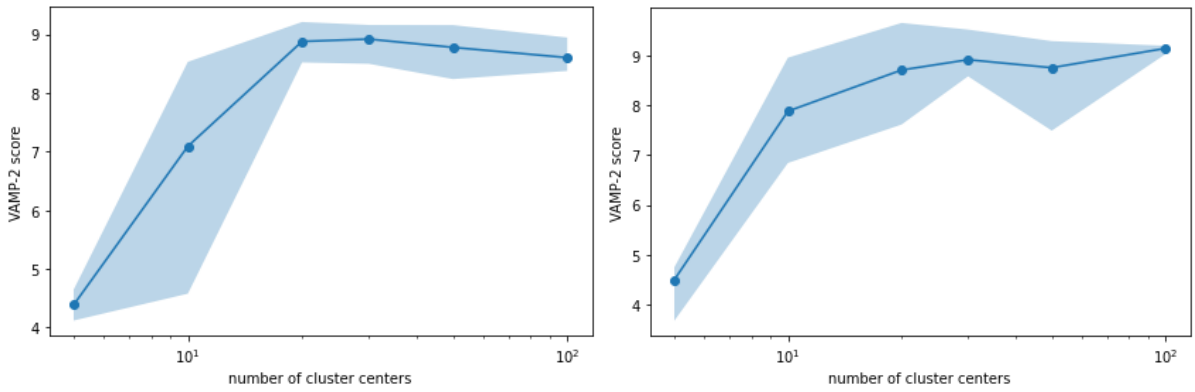

Figure 3.6.2: VAMP-2 score against number of cluster centers for WT apo 1 (left) and W473A/W562A apo 1 (right).

Figure 3.6.2 displays that the VAMP-2 score is saturated at 30 cluster centers for both systems. Therefore, we have determined that 30 will be the optimal number of microstates for further analysis.

In addition to obtaining an optimal VAMP-2 score, it is important to ensure that the resulting microstates reflect physically interesting states. Therefore, after obtaining the microstates centers from k-means, it is important to inspect whether the microstate centers cover the data which were projected onto the independent components.

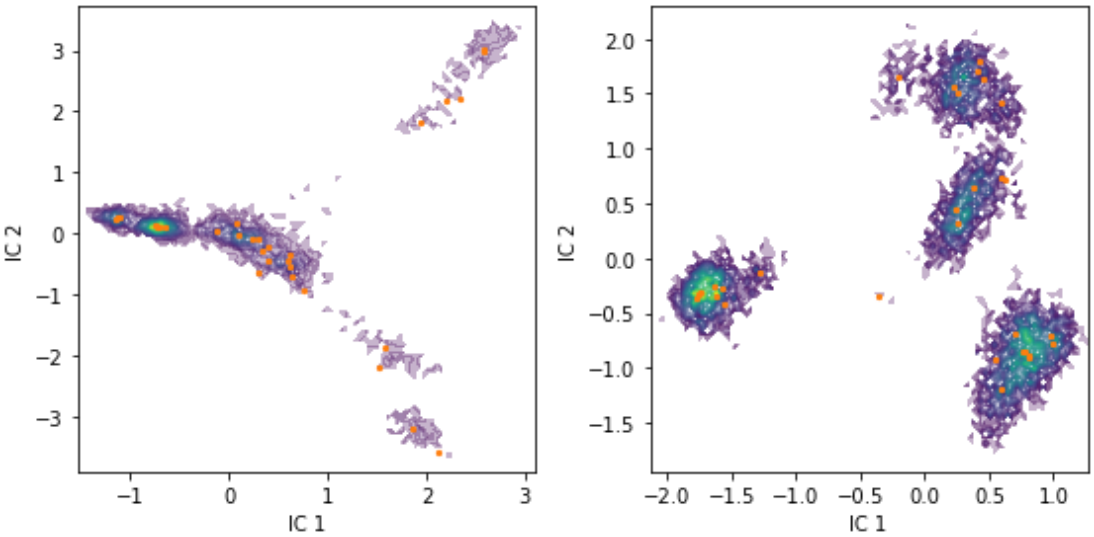

Figure 3.6.3: Plot of microstate centers (orange dots) obtained from k-means clustering for WT apo 1 (left) and W473A/W562A apo 1 (right).

Figure 3.6.3 shows that the clustering of the TICA data was successful in identifying important low-dimensional features for both simulations.

3.6.3 Building a Microstate Transition Matrix

To build a transition matrix, we must choose an appropriate lag time, which can be done by using the implied timescales (ITS). The ITS provide an estimate of the decorrelation times of the slowest processes in the system and should be independent of the chosen lag time. The plot, Figure 3.6.4, displays the convergence of the ITS as a function of lag time for WT apo 1 and W473A/W562A apo 1. The ITS of maximum likelihood MSMs are used to calculate the ITS and are represented by solid lines in the plot. The shaded area represents the confidence intervals, and the dashed lines represent the sample means. By examining the convergence of the ITS, we can choose an appropriate lag time for building the transition matrix which can be used to study the kinetics of the system.

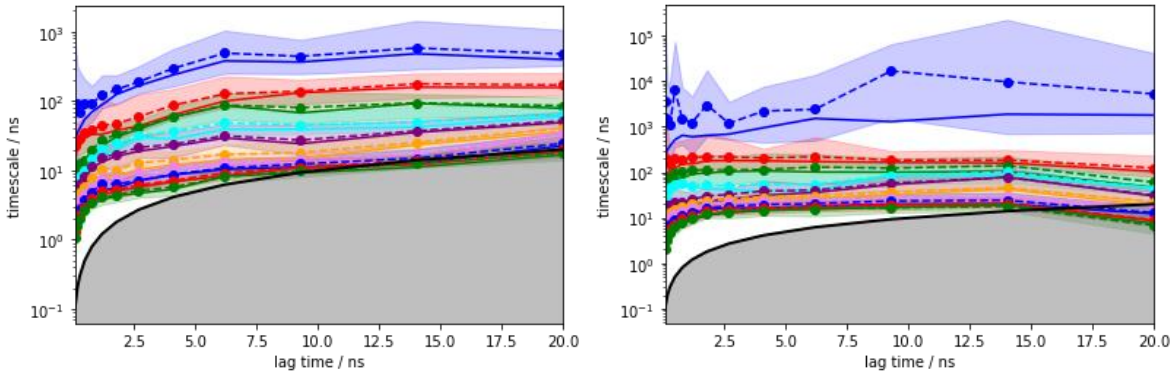

Figure 3.6.4: Convergence of the ITS as a function of lag time for WT apo 1 (left) and W473A/W562A apo 1 (right). Each line in the plots correspond to the eigenvalues for the respective systems.

From this plot we find that the ITS converge quickly on the left side of the figure, indicating that a lag time of 7.5 ns or higher for WT apo 1 is sufficient to build an accurate model. For the right side of the figure the convergence is slower, suggesting that the simulation takes longer time to transition between metastable states. To improve convergence, it is suggested to use a lag time of 15 ns. By increasing the lag time, the system will have a longer time to stabilize, leading to more precise and accurate results. However, in this project we used a lag time of 1 for both systems. While we did explore higher lag times, we found that neither of the resulting clusters contained particularly interesting molecular conformations. However, further investigations into the effect of different lag times could yield valuable insights. The trajectory files with a lag time of 7.5 ns for WT apo 1 can be found in supplemental figures (B.1). From Figure 3.6.4, we can also resolve 4 slow processes (blue, red, green, and cyan curve).

To validate the model, we check the fraction of states and counts to ensure that the data is well-sampled and representative of the systems dynamics.

Fraction of states used = 1.00
Fraction of counts used = 1.00

Based on the fraction of states and counts used, which are both 1.00 for both simulations, we can conclude that the data is well-sampled, meaning that the number of observed transitions between states is sufficiently high to accurately represent the system´s dynamics.

A Chapman-Kolmogorov test checks whether the system under study behaves in a memoryless way. In other words, it checks whether the probability of transitioning from one state to another depends only on the current states, and not on any previous states. Figures 3.6.5 and 3.6.6 represent the results of the Chapman-Kolmogorov test for a MSM with 4 macrostates for WT apo 1 and W473A/W562A apo 1 receptively. The dashed lines correspond to the predicted transition probability from the microstate transition matrix, while the solid lines correspond to the estimated transition probability from the coarse-grained microstate matrix (the MSM). If the two lines overlap, it means that the system passes the test and behaves in a memoryless way. On the other hand, if the two lines do not overlap, it suggests that the system exhibits some form of memory and the MSMs may need to be refined.

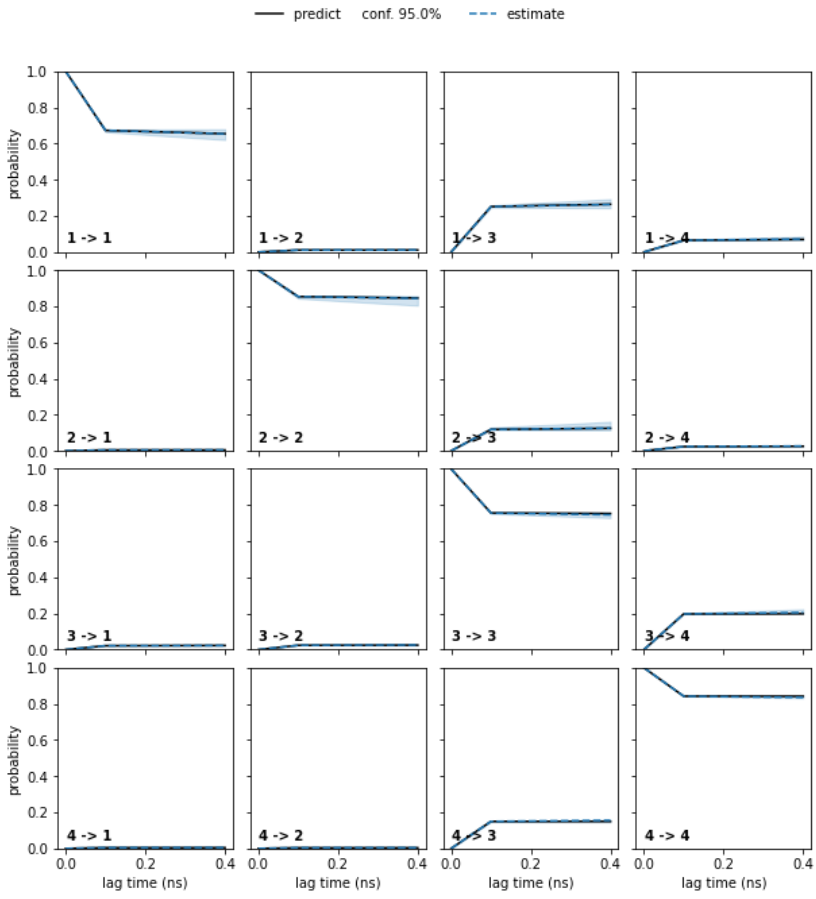

Figure 3.6.5: Chapman-Kolmogorov test for the MSM with 4 macrostates for WT apo 1.

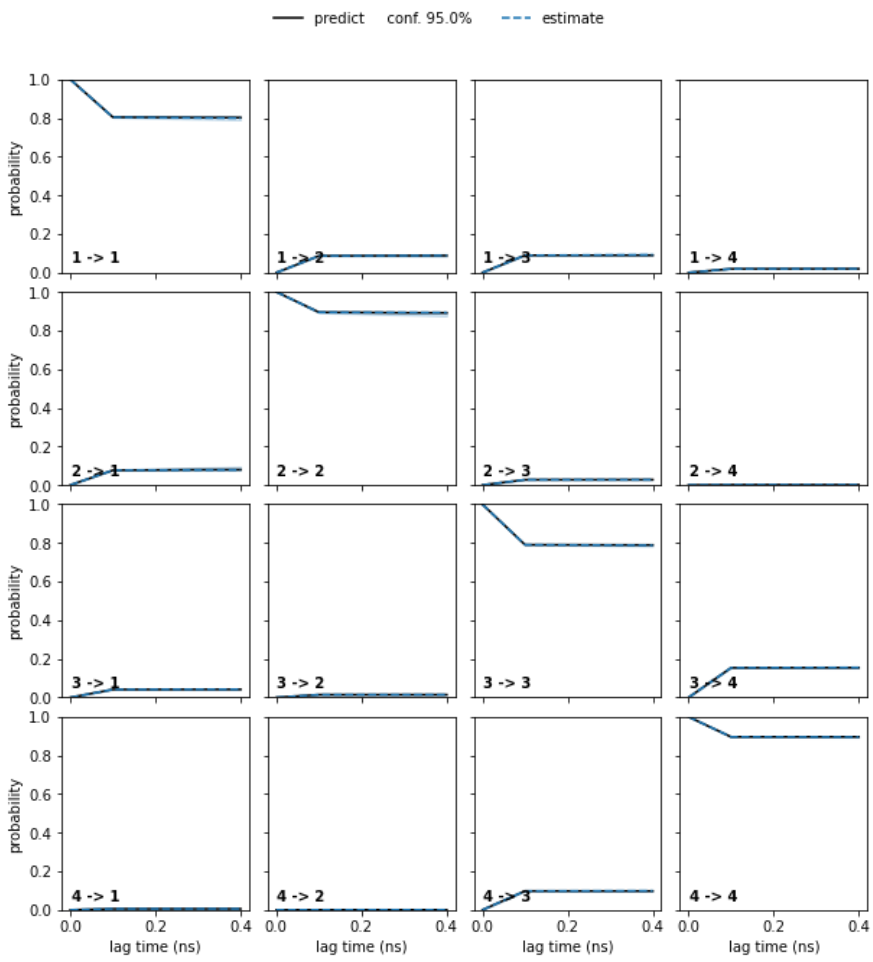

Figure 3.6.6: Chapman-Kolmogorov test for the MSM with 4 macrostates for W473A/W562A apo 1.

From Figure 3.6.5 and Figure 3.6.6 we see that the dashed line and the solid lines overlap, and therefore we have two systems that pass the Chapman-Kolmogorov test.

3.6.4 Coarse-grained Representation

The PCCA+ algorithm is a spectral clustering method for coarse-graining the microstate representation to macrostates and obtaining a clearer interpretation of the clustering results. In Figure 3.6.7, the PCCA+ algorithm has been applied to the microstate transition matrix, and 4 clusters are shown in a two-dimensional plot, where the x and z-axes correspond to the first two TICA components for both of our systems. Different clusters are represented by different colors.

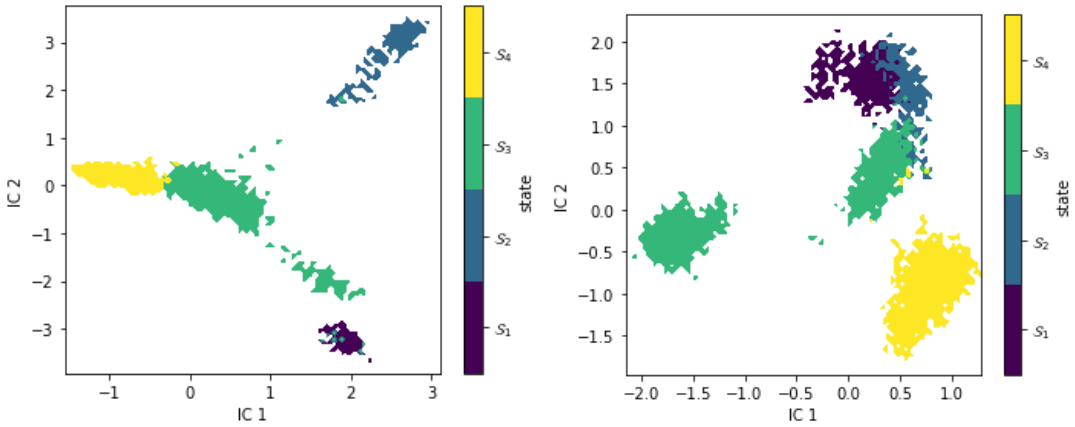

Figure 3.6.7: Visualization of the macrostates over first two TICA projections for WT apo 1 (left) and for W473A/W562A apo 1 (right).

Figure 3.6.7 shows that the clusters clearly separate the states space within the two TICA components for both simulations, demonstrating the effectiveness of the PCCA+ algorithm.

The transition matrix provides information about the probabilities of transitioning between the four microstates where each element represents the probability of transitioning from one microstate to another. Below, we present the microstate transition matrices for two systems: the WT apo 1 (P_1) and the W473A/W562 apo 1 (P_2). The WT apo 1 transition matrix is displayed first, followed by the transition matrix for W473A/W562A apo 1.

$$P_1 = \begin{bmatrix} 9.90 * 10^{-1} & 6.86 * 10^{-6} & 7.61 * 10^{-3} & 2.28 * 10^{-3} \\ 3.29 * 10^{-5} & 9.96 * 10^{-1} & 3.56 * 10^{-3} & 5.37 * 10^{-4} \\ 7.05 * 10^{-4} & 7.88 * 10^{-4} & 9.94 * 10^{-1} & 4.37 * 10^{-3} \\ 1.13 * 10^{-4} & 4.43 * 10^{-5} & 3.39 * 10^{-3} & 9.97 * 10^{-1} \end{bmatrix}$$

$$P_2 = \begin{bmatrix} 9.98 * 10^{-1} & 8.90 * 10^{-4} & 1.03 * 10^{-3} & 1.11 * 10^{-4} \\ 7.97 * 10^{-4} & 9.99 * 10^{-1} & 3.90 * 10^{-5} & -7.81 * 10^{-6} \\ 5.14 * 10^{-4} & 2.02 * 10^{-5} & 9.99 * 10^{-1} & 7.73 * 10^{-4} \\ 4.88 * 10^{-6} & 7.12 * 10^{-7} & 5.17 * 10^{-4} & 9.99 * 10^{-1} \end{bmatrix}$$

Looking at the transition matrix for WT apo 1 and W473A/W562A apo 1, it becomes clear that both systems have a high probability of remaining in one of the macrostates. In other words, State 1, State 2, State 3, and State 4 (the macrostates or clusters) have high probabilities of remaining in their states, with values close to 1. These values indicate that transitions away from these states are unlikely. Moreover, the macrostates of the

W473A/W562A apo 1 system have higher holding probabilities (all values are larger than or equal to 0.998) compared to the macrostates in the WT apo 1 system.

From the transition probability matrix, ee can compute the stationary distribution, which reflects the free energy of states. Lower free energy states are more likely, whereas higher free energy states are less likely. To obtain this, we sum up all the contributions to a coarse-grained state with the help of formula (2.3.4). Table 3.6.1 and Table 3.6.2 provides an overview of our four states and their corresponding free energy for WT apo 1 and W473A/W562A apo 1.

Table 3.6.1: List of the free energy of states for WT apo 1.

State	**π**	**G/kT**
1	0.026	3.641
2	0.072	2.634
3	0.460	0.777
4	0.442	0.816

Table 3.6.2: List of the free energy of states for W473A/W562A apo 1.

State	**π**	**G/kT**
1	0.135	2.003
2	0.145	1.930
3	0.314	1.158
4	0.406	0.902

By examining the two tables, we observe that the states with the lowest free energy are State 3 and State 4 for both systems. These two states are energetically favorable and more stable compared to the other states in these systems.

After identifying the metastable states using the PCCA+ algorithm, we can extract additional information by calculating the mean first passage times (MFPTs) between them. The MFPTs estimate the average time required for the system to transition from one state to another. Table 3.6.3 gives an overview of the MFPTs between each of the four states for WT apo 1. Table 3.6.4 gives MFPTs values for W473A/W562A apo 1.

Table 3.6.3: MFPTs between each of the four states for WT apo 1 (ns):

	1	**2**	**3**	**4**
1	0.00	303.90	9.38	36.72
2	385.63	0.00	22.99	54.11
3	359.74	289.46	0.00	29.71
4	378.83	314.36	19.33	0.00

Table 3.6.4: MFPTs between each of the four states for W473A/W562A apo 1 (ns):

	1	**2**	**3**	**4**
1	0.00	637.76	165.78	413.37
2	115.86	0.00	280.57	528.32
3	460.32	1096.51	0.00	230.42
4	625.30	1261.65	124.04	0.00

Upon examining the MFPTs derived from the transition matrix, several noteworthy values come to light. Specifically, for WT apo 1 it becomes clear that the longest durations are for transitions from State 2, State 3, and State 4 to State 1, with values of 385.63 ns, 359.74 ns, and 378.83 ns, respectively. These values highlight the time it takes for the system to transition from these states back to State 1, suggesting a significant barrier between these states. Furthermore, when considering findings from the transition matrix, we see a significant duration for transitions from State 3 to State 2, with a value of 289.46 ns. The transition from State 1 to State 3 is slower, with a value of 9.38 ns. For W473A/W562A apo 1 we see that transitioning from State 3 to State 2, and from State 4 to State 2 takes the longest with values of 1096.51 ns and 1261 ns, respectively. In view of the transition matrix (P_2) values we see that a transition from State 1 to State 3 has a low duration of 165.78 ns, and from State 1 to State 2 has a significant duration of 637.76 ns.

3.6.5 Analysis of the Coarse-grained Markov State Model

Our metastable structures have been identified, and its therefore important to determine their molecular structures. To do this, we generated a set of representative sample structures for each microstate and saved them in a trajectory file. Visualization of these trajectory files have been loaded to Pymol and are illustrated in Figure 3.6.8.

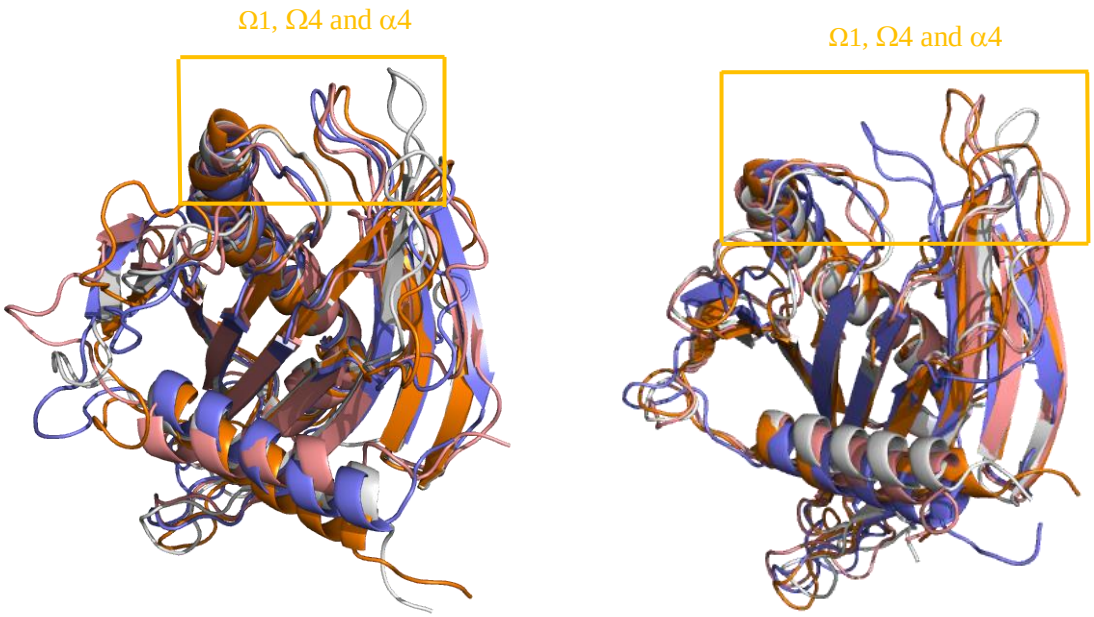

Figure 3.6.8: Visualisation of the four different states for the WT apo 1 protein (left figure) and W473A/W562A apo 1 (right figure) with focus on Ω1-loop, Ω4-loop and α4-helix. Pink represents State 1, white represents State 2, purple represents State 3, and orange represents State 4.

Upon examining the trajectory of both WT apo 1 and W473A/W562A apo 1 systems, it becomes evident that State 2 (white) exhibits the most significant opening among all observed states for both proteins. To gain a more comprehensive understandings of the distances involved, Table 3.6.5 and 3.6.7 have been included. These two tables highlight the distances between residue 473 and 564, as well as the distance between residue 476 and 562. By focusing on these specific distances, we can better analyze the fluctuation between the Ω1-loop and Ω4-loop.

Table 3.6.5: Distance 1 and 2 corresponds to the distances between residue 473 and residue 564, and residue 476 and residue 562, respectively, for the WT apo 1 system across all states:

	State 1 (pink)	**State 2 (white)**	**State 3 (purple)**	**State 4 (orange)**
Distance 1 (Å)	9.6	12.5	8.8	6.9
Distance 2 (Å)	8.7	9.8	7.5	8.2

Table 3.6.6: Distance 1 and 2 corresponds to the distances between residue 473 and residue 564, and residue 476 and residue 562, respectively, for the W473A/W562A apo 1 system across all states:

	State 1 (pink)	**State 2 (white)**	**State 3 (purple)**	**State 4 (orange)**
Distance 1 (Å)	14.5	24.4	8.0	16.8
Distance 2 (Å)	12.6	14.3	9.5	10.7

Based on our data from the table for the WT apo 1 (Table 3.6.5), it is evident that the State 2 (white) exhibits the largest opening among the four states, with a distance of 12.5 Å between residue 473 and 564, and a distance of 9.8 Å between 476 and 562. In other words, State 2 have a significant opening in the regions of Ω1-loop, Ω4-loop and α4-helix. These findings were analysed by the transition matrix (P_1), which indicates that a transition into State 2 most likely happens from State 3 with a MFPTs of 289.46 ns.

After analysing the W473A/W562A apo 1 (Table 3.6.6) it becomes clear that the State 2 (white) exhibits the largest opening here as well, with a distance of 24.4 Å between residue 473 and 564, and a distance of 14.3 Å between residue 476 and 562. This opening is greater compared to WT apo 1. From the transition matrix (P_2) we know that transitioning from State 1 to State 2 has high probability of occurring with a value of $8.90 * 10^{-4}$, and a MFPTs of 637.76 ns. State 1 (pink) and State 4 (orange) also has the larger opening compared to WT apo 1. We know that transitioning into State 1 most likely happens from State 2 with a probability of ($7.97 * 10^{-4}$) and a MFPTs of 115.86 ns, while transitioning into State 4 most likely happens from State 3 with a probability of ($7.73 * 10^{-4}$) and a MFPTs of 230.42 ns. Looking at the stationary density (Table 3.6.1 and 3.6.2) we know that State 4 is much more likely as State 1 and State 2 for both systems.

In summary, our analysis of both WT apo 1 and W473A/W562A apo 1 systems has yielded significant findings, particularly regarding State 2. This state demonstrates the largest opening in our region of interest for the W473A/W562A apo 1 system, indicating significant conformational changes compared to the other states for both systems. Moreover, in the W473A/W562A apo 1 system, we find a significant transition probability from State 1 to State 2, indicating a preferred pathway between these two states. For the WT apo 1 system State 2 also exhibits the largest opening within the same system, with a preferred pathway into State 2 from State 3. These observations highlight the importance of State 2 in both systems and suggest that it plays a significant role in their overall dynamics. Furthermore, comparing the two systems reveals a relatively smaller opening in State 1 and State 4 for WT apo 1. This difference suggests a structural variation between the two systems in that specific region. Indicating that the double mutation, replacing tryptophan to alanine, leads to different structural behavior.

To increase the reliability of our results, it would have been necessary to perform the MSM analysis on replicas of our simulations. However, due to time constraints, we were unable to conduct that analysis on such replicas within the scope of this thesis. Nevertheless, our study provides valuable insight into the dynamics of the protein and lays groundwork for future research.

The code used for this thesis is available on GitHub. You can find it on the following link: https://git.app.uib.no/reuter-group/MSM_Hedda/-/tree/hedda_upload/notebooks

4

Conclusion & Future Research

Molecular dynamics (MD) simulations were conducted to study the transport of ceramide from the endoplasmic reticulum (ER) to the Golgi apparatus, a process dependent on the essential STARD11 protein. The aim was to gain insight into the dynamic behaviour and structural changes of STARD11 during ceramide transport. In total, 15 MD simulations were performed on STARD11, both in the absence (2E3M structure), and presence (2E3Q structure) of ceramide binding. Furthermore, single- and double- mutation experiments were performed on the 2E3M and 2E3Q complexes. The mutation simulations involved substituting tryptophan to alanine amino acid on residue number 473 and 562.

To gain a deeper understanding of the dynamics of STARD11 in the ceramide binding regions, a Markov state models (MSMs) analysis was conducted on the 2E3M crystal structure for the wild type and double mutation complexes. This analysis involved generating transition matrices to calculate the probability of the system being in a specific state and transitioning between them at a certain lag time.

The WT apo simulations exhibited higher structural change as observed in the RMSD analysis. Moreover, the distance fluctuations and the RMSF plots displayed higher variability specifically in the Ω1-loop and Ω4-loop regions compared to the WT holo simulations. These observations will support the idea that these regions are the entrance to the cavity. Overall, WT holo shows higher structural stability in comparison to WT apo form. This increased structural stability can be attributed to the interaction between the WT holo protein and the binding of the ceramide molecule, which likely contributes to a more compact and stable protein structure.

The comparison between the double mutation on apo proteins and holo proteins indicates that there exists a similarity between the simulations in the Ω1-loop, Ω-4 loop and α4-helix regions. Specifically, it suggests that fluctuation observed in the wild type simulations for

both apo and holo proteins, as well as double mutation simulations for apo and holo proteins, exhibit a similar pattern. In other words, there is a consistent trend in the conformational dynamics of these regions between the wild type and double mutation simulations, regardless of whether the proteins are in the apo or holo form.

The introduction of a single mutation affects the flexibility of the apo protein, as indicated by higher RMSF values and structural changes observed in specific regions. However, these single mutations do not significantly alter the flexibility on the Ω1-loop, Ω-4 loop and α4-helix regions in the holo proteins. This suggests that the presence of ceramide in the holo protein contributes to stabilizing of the entrance to hydrophobic cavity.

The MSMs analysis highlights the significance of State 2 in both systems, indicating a preferred pathway for the WT apo 1 protein from State 3 to State 2, while the W473A/W562A apo 1 simulation show a preferred pathway from State 1 to State 2. Notably, State 1 and State 4 exhibit larger openings in the W473A/W562A apo 1 protein, unlike in WT apo 1. These findings suggest structural differences between the wild type apo protein and the apo protein with double mutation in the Ω1-loop, Ω-4 loop and α4-helix regions.

In summary, the findings presented in this thesis indicates that the apo protein exhibits greater structural changes in its wild type forms, as well as when introducing single mutations. On the other hand, the holo protein demonstrates higher structural stability in its wild type form, and with single mutations. These results highlight the role of ceramide binding, specifically in stabilizing the entrance to the hydrophobic cavity.

These observations presented in this project provides valuable insight that will contribute to future experiments focused on understanding the function of STARD11 and its transport of the ceramide molecule. The results obtained from this thesis shed light on specific regions of interest, such as the Ω1-loop, Ω4- loop, and α4-helix. Moreover, this thesis provides information about the impact of single and double mutations on the structure and behavior of STARD11.

5

References

[1] Stephanie Green and Kelli Shallel 2023 Lipids *Nutrition Essentials* (Maricopa)

[2] Wong L H, Gatta A T and Levine T P 2019 Lipid transfer proteins: the lipid commute via shuttles, bridges and tubes *Nat. Rev. Mol. Cell Biol.* **20** 85–101

[3] Alpy F and Tomasetto C 2014 START ships lipids across interorganelle space *Biochimie* **96** 85–95

[4] Kudo N, Kumagai K, Tomishige N, Yamaji T, Wakatsuki S, Nishijima M, Hanada K and Kato R 2008 Structural basis for specific lipid recognition by CERT responsible for nonvesicular trafficking of ceramide *Proc. Natl. Acad. Sci.* **105** 488–93

[5] Armstrong D R, Berrisford J M, Conroy M J, Gutmanas A, Anyango S, Choudhary P, Clark A R, Dana J M, Deshpande M, Dunlop R, Gane P, Gáborová R, Gupta D, Haslam P, Koča J, Mak L, Mir S, Mukhopadhyay A, Nadzirin N, Nair S, Paysan-Lafosse T, Pravda L, Sehnal D, Salih O, Smart O, Tolchard J, Varadi M, Svobodova-Vařeková R, Zaki H, Kleywegt G J and Velankar S 2019 PDBe: improved findability of macromolecular structure data in the PDB *Nucleic Acids Res.* gkz990

[6] Robert X and Gouet P 2014 Deciphering key features in protein structures with the new ENDscript server *Nucleic Acids Res.* **42** W320–4

[7] McCammon J A, Gelin B R and Karplus M 1977 Dynamics of folded proteins *Nature* **267** 585–90

[8] Levitt M and Sharon R 1988 Accurate simulation of protein dynamics in solution. *Proc. Natl. Acad. Sci.* **85** 7557–61

[9] Karplus M and Kuriyan J 2005 Molecular dynamics and protein function *Proc. Natl. Acad. Sci.* **102** 6679–85

[10] Kukol A 2016 NAMD-VMD tutorial

[11] Frenkel D and Smit B 2002 *Understanding molecular simulation: from algorithms to applications* (San Diego, Calif.: Acad. Press)

[12] Furio Ercolessi 1997 *A molecular dynamics primer* (Spring College in Computational Physics)

[13] Anon 2021 Non-Covalent Interactions in Proteins (WORLD SCIENTIFIC) pp 1–26

[14] Mark Abraham 2018 *GROMOCS User Manual*

[15] Cornell W D, Cieplak P, Bayly C I, Gould I R, Merz K M, Ferguson D M, Spellmeyer D C, Fox T, Caldwell J W and Kollman P A 1995 A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules *J. Am. Chem. Soc.* **117** 5179–97

[16] Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I and Mackerell A D 2009 CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields *J. Comput. Chem.* NA-NA

[17] Jorgensen W L, Maxwell D S and Tirado-Rives J 1996 Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids *J. Am. Chem. Soc.* **118** 11225–36

[18] Horta B A C, Merz P T, Fuchs P F J, Dolenc J, Riniker S and Hünenberger P H 2016 A GROMOS-Compatible Force Field for Small Organic Molecules in the Condensed Phase: The 2016H66 Parameter Set *J. Chem. Theory Comput.* **12** 3825–50

[19] Best R B, Zhu X, Shim J, Lopes P E M, Mittal J, Feig M and MacKerell A D 2012 Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles *J. Chem. Theory Comput.* **8** 3257–73

[20] MacKerell A D, Bashford D, Bellott M, Dunbrack R L, Evanseck J D, Field M J, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau F T K, Mattos C, Michnick S, Ngo T, Nguyen D T, Prodhom B, Reiher W E, Roux B, Schlenkrich M, Smith J C, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D and Karplus M 1998 All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins *J. Phys. Chem. B* **102** 3586–616

[21] Lemkul J 2019 From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0] *Living J. Comput. Mol. Sci.* **1**

[22] Anandakrishnan R, Drozdetski A, Walker R C and Onufriev A V 2015 Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations *Biophys. J.* **108** 1153–64

[23] Brooks C L, Karplus M and Pettitt B M 1988 *Proteins: a theoretical perspective of dynamics, structure, and thermodynamics* (New York: Wiley)

[24] Lee J, Patel D S, Ståhle J, Park S-J, Kern N R, Kim S, Lee J, Cheng X, Valvano M A,

Holst O, Knirel Y A, Qi Y, Jo S, Klauda J B, Widmalm G and Im W 2019 CHARMM-GUI *Membrane Builder* for Complex Biological Membrane Simulations with Glycolipids and Lipoglycans *J. Chem. Theory Comput.* **15** 775–86

[25] R. Bernardi et al 2020 NAMD User´s Guide Version 2.14

[26] Phillips J C, Hardy D J, Maia J D C, Stone J E, Ribeiro J V, Bernardi R C, Buch R, Fiorin G, Hénin J, Jiang W, McGreevy R, Melo M C R, Radak B K, Skeel R D, Singharoy A, Wang Y, Roux B, Aksimentiev A, Luthey-Schulten Z, Kalé L V, Schulten K, Chipot C and Tajkhorshid E 2020 Scalable molecular dynamics on CPU and GPU architectures with NAMD *J. Chem. Phys.* **153** 044130

[27] Baugh E H, Lyskov S, Weitzner B D and Gray J J 2011 Real-Time PyMOL Visualization for Rosetta and PyRosetta ed V N Uversky *PLoS ONE* **6** e21931

[28] Pande V S, Beauchamp K and Bowman G R 2010 Everything you wanted to know about Markov State Models but were afraid to ask *Methods* **52** 99–105

[29] Pérez-Hernández G, Paul F, Giorgino T, De Fabritiis G and Noé F 2013 Identification of slow molecular order parameters for Markov model construction *J. Chem. Phys.* **139** 015102

[30] Hoffmann M, Scherer M, Hempel T, Mardt A, de Silva B, Husic B E, Klus S, Wu H, Kutz N, Brunton S L and Noé F 2022 Deeptime: a Python library for machine learning dynamical models from time series data *Mach. Learn. Sci. Technol.* **3** 015009

[31] Wu H and Noé F 2020 Variational Approach for Learning Markov Processes from Time Series Data *J. Nonlinear Sci.* **30** 23–66

[32] Röblitz S and Weber M 2013 Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification *Adv. Data Anal. Classif.* **7** 147–79

Simulation Input Files

A.1 Examples of simulation input files

NAMD uses a set of input files to define and configure a molecular system. Example of both the equilibration step and production step input files are shown below. See NAMD User´s Guide [25] for further details.

Input file for equilibration step:

```
structure          step3_input.psf
coordinates        step3_input.pdb

set temp           310;

set outputname     step4_equilibration;

# read system values written by CHARMM (need to convert uppercases to lowercases)
exec tr "\[:upper:\]" "\[:lower:\]" < ../step3_pbcsetup.str | sed -e "s/ =//g" > step3_input.str

source             step3_input.str

temperature        $temp;

outputName         4.0/$outputname;    # base name for output from this run
                   # NAMD writes two files at the end, final coord and vel
                   # in the format of first-dyn.coor and first-dyn.vel
firsttimestep      0;            #last step of previous run

restartfreq        500;          # 500 steps = every 1ps
dcdfreq            5000;
dcdUnitCell        yes;          # the file will contain unit cell info in the style of
                   # charmm dcd files. if yes, the dcd files will contain
                   # unit cell information in the style of charmm DCD files.
xstFreq            5000;         # XSTFreq: control how often the extended systen configuration
                   # will be appended to the XST file
outputEnergies     125;          # 125 steps = every 0.25ps
                   # The number of timesteps between each energy output of NAMD
outputTiming       1000;         # The number of timesteps between each timing output shows
                   # time per step and time to completion

# Force-Field Parameters
paraTypeCharmm     on;           # We're using charmm type parameter file(s)
                   # multiple definitions may be used but only one file per definition
parameters         toppar/par_all36m_prot.prm
parameters         toppar/par_all36_na.prm
parameters         toppar/par_all36_carb.prm
parameters         toppar/par_all36_lipid.prm
parameters         toppar/par_all36_cgenff.prm
parameters         toppar/par_interface.prm
parameters         toppar/toppar_all36_moreions.str
parameters         toppar/toppar_all36_nano_lig.str
parameters         toppar/toppar_all36_nano_lig_patch.str
parameters         toppar/toppar_all36_synthetic_polymer.str
parameters         toppar/toppar_all36_synthetic_polymer_patch.str
parameters         toppar/toppar_all36_polymer_solvent.str
parameters         toppar/toppar_water_ions.str
parameters         toppar/toppar_dum_noble_gases.str
```

```
parameters              toppar/toppar_ions_won.str
parameters              toppar/toppar_all36_prot_arg0.str
parameters              toppar/toppar_all36_prot_c36m_d_aminoacids.str
parameters              toppar/toppar_all36_prot_fluoro_alkanes.str
parameters              toppar/toppar_all36_prot_heme.str
parameters              toppar/toppar_all36_prot_na_combined.str
parameters              toppar/toppar_all36_prot_retinol.str
parameters              toppar/toppar_all36_prot_model.str
parameters              toppar/toppar_all36_prot_modify_res.str
parameters              toppar/toppar_all36_na_nad_ppi.str
parameters              toppar/toppar_all36_na_rna_modified.str
parameters              toppar/toppar_all36_lipid_sphingo.str
parameters              toppar/toppar_all36_lipid_archaeal.str
parameters              toppar/toppar_all36_lipid_bacterial.str
parameters              toppar/toppar_all36_lipid_cardiolipin.str
parameters              toppar/toppar_all36_lipid_cholesterol.str
parameters              toppar/toppar_all36_lipid_dag.str
parameters              toppar/toppar_all36_lipid_inositol.str
parameters              toppar/toppar_all36_lipid_lnp.str
parameters              toppar/toppar_all36_lipid_lps.str
parameters              toppar/toppar_all36_lipid_mycobacterial.str
parameters              toppar/toppar_all36_lipid_miscellaneous.str
parameters              toppar/toppar_all36_lipid_model.str
parameters              toppar/toppar_all36_lipid_prot.str
parameters              toppar/toppar_all36_lipid_tag.str
parameters              toppar/toppar_all36_lipid_yeast.str
parameters              toppar/toppar_all36_lipid_hmmm.str
parameters              toppar/toppar_all36_lipid_detergent.str
parameters              toppar/toppar_all36_lipid_ether.str
parameters              toppar/toppar_all36_carb_glycolipid.str
parameters              toppar/toppar_all36_carb_glycopeptide.str
parameters              toppar/toppar_all36_carb_imlab.str
parameters              toppar/toppar_all36_label_spin.str
parameters              toppar/toppar_all36_label_fluorophore.str
parameters              toppar/toppar_all36_lipid_cationpi_wyf.str

# Nonbonded Parameters
exclude                 scaled1-4          # non-bonded exclusion policy to use "none,1-2,1-3,1-4, or scaled1-4"
                                           # 1-2: all atoms pairs that are bonded are going to be ignored
                                           # 1-3: 3 consecutively bonded are excluded
                                           # scaled1-4: include all the 1-3, and modified 1-4 interactions
                                           # electrostatic scaled by 1-4scaling factor 1.0
                                           # vdW special 1-4 parameters in charmm parameter file.
1-4scaling              1.0
switching               on
vdwForceSwitching       on;                # New option for force-based switching of vdW
                                           # if both switching and vdwForceSwitching are on CHARMM force
                                           # switching is used for vdW forces.
# You have some freedom choosing the cutoff
cutoff                  12.0;              # may use smaller, maybe 10., with PME
switchdist              10.0;              # cutoff - 2.
                                           # switchdist - where you start to switch
                                           # cutoff - where you stop accounting for nonbond interactions.
                                           # correspondence in charmm:
                                           # (cutnb,ctofnb,ctonnb = pairlistdist,cutoff,switchdist)
pairlistdist            16.0;              # stores the all the pairs with in the distance it should be larger
                                           # than cutoff( + 2.)
stepspercycle           20;                # 20 redo pairlists every ten steps
pairlistsPerCycle       2;                 # 2 is the default
                                           # cycle represents the number of steps between atom reassignments
pairlist will be updated

# Integrator Parameters
timestep                2.0;               # fs/step
rigidBonds              all;               # Bound constraint all bonds involving H are fixed in length
nonbondedFreq           1;                 # nonbonded forces every step
fullElectFrequency      1;                 # PME every step

# Constant Temperature Control ONLY DURING EQUILB
reassignFreq            500;               # reassignFreq:  use this to reassign velocity every 500 steps
reassignTemp            $temp;

# Periodic Boundary conditions. Need this since for a start...
cellBasisVector1     $a   0.0   0.0;       # vector to the next image
cellBasisVector2     0.0    $b   0.0;
cellBasisVector3     0.0   0.0    $c;
cellOrigin           0.0   0.0 $zcen;      # the *center* of the cell

wrapWater               on;                # wrap water to central cell
wrapAll                 on;                # wrap other molecules too
wrapNearest             off;               # use for non-rectangular cells (wrap to the nearest image)
```

```
# PME (for full-system periodic electrostatics)
PME                 yes;
PMEInterpOrder      6;              # interpolation order (spline order 6 in charmm)
PMEGridSpacing      1.0;            # maximum PME grid space / used to calculate grid size
# Constant Temperature Control
langevin            on
langevinDamping     1.0
langevinTemp        $temp
langevinHydrogen    off

constraints         on
consexp             2
consref             restraints/prot_posres.ref
conskfile           restraints/prot_posres.ref
conskcol            B
constraintScaling   1.0

minimize            10000
numsteps            90000000
run                 5000000
```

Input file for production step:

```
structure           step3_input.psf
coordinates         step3_input.pdb

set temp            310;
outputName          5.0/step5_production; # base name for output from this run
                    # NAMD writes two files at the end, final coord and vel
                    # in the format of first-dyn.coor and first-dyn.vel

set inputname       step4_equilibration;
binCoordinates      4.0/$inputname.coor;    # coordinates from last run (binary)
binVelocities       4.0/$inputname.vel;     # velocities from last run (binary)
extendedSystem      4.0/$inputname.xsc;     # cell dimensions from last run (binary)

dcdfreq             50000;
dcdUnitCell         yes;            # the file will contain unit cell info in the style of
files will contain
                    # unit cell information in the style of charmm DCD files.
xstFreq             5000;           # XSTFreq: control how often the extended systen configuration
                    # will be appended to the XST file
outputEnergies      5000;           # 5000 steps = every 10ps
                    # The number of timesteps between each energy output of NAMD
outputTiming        5000;           # The number of timesteps between each timing output shows
on
restartfreq         5000;           # 5000 steps = every 10ps

# Force-Field Parameters
paraTypeCharmm      on;             # We're using charmm type parameter file(s)
                    # multiple definitions may be used but only one file per definition
parameters          toppar/par_all36m_prot.prm
parameters          toppar/par_all36_na.prm
parameters          toppar/par_all36_carb.prm
parameters          toppar/par_all36_lipid.prm
parameters          toppar/par_all36_cgenff.prm
parameters          toppar/par_interface.prm
parameters          toppar/toppar_all36_moreions.str
parameters          toppar/toppar_all36_nano_lig.str
parameters          toppar/toppar_all36_nano_lig_patch.str
parameters          toppar/toppar_all36_synthetic_polymer.str
parameters          toppar/toppar_all36_synthetic_polymer_patch.str
parameters          toppar/toppar_all36_polymer_solvent.str
parameters          toppar/toppar_water_ions.str
parameters          toppar/toppar_dum_noble_gases.str
parameters          toppar/toppar_ions_won.str
parameters          toppar/toppar_all36_prot_arg0.str
parameters          toppar/toppar_all36_prot_c36m_d_aminoacids.str
parameters          toppar/toppar_all36_prot_fluoro_alkanes.str
parameters          toppar/toppar_all36_prot_heme.str
parameters          toppar/toppar_all36_prot_na_combined.str
parameters          toppar/toppar_all36_prot_retinol.str
parameters          toppar/toppar_all36_prot_model.str
parameters          toppar/toppar_all36_prot_modify_res.str
parameters          toppar/toppar_all36_na_nad_ppi.str
parameters          toppar/toppar_all36_na_rna_modified.str
parameters          toppar/toppar_all36_lipid_sphingo.str
parameters          toppar/toppar_all36_lipid_archaeal.str
```

```
parameters              toppar/toppar_all36_lipid_bacterial.str
parameters              toppar/toppar_all36_lipid_cardiolipin.str
parameters              toppar/toppar_all36_lipid_cholesterol.str
parameters              toppar/toppar_all36_lipid_dag.str
parameters              toppar/toppar_all36_lipid_inositol.str
parameters              toppar/toppar_all36_lipid_lnp.str
parameters              toppar/toppar_all36_lipid_lps.str
parameters              toppar/toppar_all36_lipid_mycobacterial.str
parameters              toppar/toppar_all36_lipid_miscellaneous.str
parameters              toppar/toppar_all36_lipid_model.str
parameters              toppar/toppar_all36_lipid_prot.str
parameters              toppar/toppar_all36_lipid_tag.str
parameters              toppar/toppar_all36_lipid_yeast.str
parameters              toppar/toppar_all36_lipid_hmmm.str
parameters              toppar/toppar_all36_lipid_detergent.str
parameters              toppar/toppar_all36_lipid_ether.str
parameters              toppar/toppar_all36_carb_glycolipid.str
parameters              toppar/toppar_all36_carb_glycopeptide.str
parameters              toppar/toppar_all36_carb_imlab.str
parameters              toppar/toppar_all36_label_spin.str
parameters              toppar/toppar_all36_label_fluorophore.str
parameters              toppar/toppar_all36_lipid_cationpi_wyf.str

# Nonbonded Parameters
exclude                 scaled1-4          # non-bonded exclusion policy to use "none,1-2,1-3,1-4,or scaled1-4"
                                           # 1-2: all atoms pairs that are bonded are going to be ignored
                                           # 1-3: 3 consecutively bonded are excluded
                                           # scaled1-4: include all the 1-3, and modified 1-4 interactions
                                           # electrostatic scaled by 1-4scaling factor 1.0
                                           # vdW special 1-4 parameters in charmm parameter file.
1-4scaling              1.0
switching               on
vdwForceSwitching       on;                # New option for force-based switching of vdW
                                           # if both switching and vdwForceSwitching are on CHARMM force
# switching is used for vdW forces.

# You have some freedom choosing the cutoff
cutoff                  12.0;              # may use smaller, maybe 10., with PME
switchdist              10.0;              # cutoff - 2.
                                           # switchdist - where you start to switch
                                           # cutoff - where you stop accounting for nonbond interactions.
                                           # correspondence in charmm:
                                           # (cutnb,ctofnb,ctonnb = pairlistdist,cutoff,switchdist)
pairlistdist            16.0;              # stores the all the pairs with in the distance it should be larger
                                           # than cutoff( + 2.)
stepspercycle           20;                # 20 redo pairlists every ten steps
pairlistsPerCycle       2;                 # 2 is the default
                                           # cycle represents the number of steps between atom reassignments
                                           # this means every 20/2=10 steps the pairlist will be updated

# Integrator Parameters
timestep                2.0;               # fs/step
rigidBonds              all;               # Bound constraint all bonds involving H are fixed in length
nonbondedFreq           1;                 # nonbonded forces every step
fullElectFrequency      1;                 # PME every step

wrapWater               on;                # wrap water to central cell
wrapAll                 on;                # wrap other molecules too
wrapNearest             off;               # use for non-rectangular cells (wrap to the nearest image)

# PME (for full-system periodic electrostatics)
PME                     yes;
PMEInterpOrder          6;                 # interpolation order (spline order 6 in charmm)
PMEGridSpacing          1.0;               # maximum PME grid space / used to calculate grid size

# Constant Pressure Control (variable volume)
useGroupPressure        yes;               # use a hydrogen-group based pseudo-molecular viral to calcualte pressure and
                                           # has less fluctuation, is needed for rigid bonds (rigidBonds/SHAKE)
useFlexibleCell         no;                # yes for anisotropic system like membrane
useConstantRatio        no;                # keeps the ratio of the unit cell in the x-y plane constant A=B

# Constant Temperature Control
langevin                on;                # langevin dynamics
langevinDamping         1.0;               # damping coefficient of 1/ps (keep low)
langevinTemp            $temp;             # random noise at this level
langevinHydrogen        off;               # don't couple bath to hydrogens

# Constant pressure
langevinPiston          on;                # Nose-Hoover Langevin piston pressure control
langevinPistonTarget    1.01325;           # target pressure in bar 1atm = 1.01325bar
langevinPistonPeriod    50.0;              # oscillation period in fs. correspond to pgamma T=50fs=0.05ps
```

```
                                # f=1/T=20.0(pgamma)
langevinPistonDecay     25.0;           # oscillation decay time. smaller value corresponds to larger random
                                # forces and increased coupling to the Langevin temp bath.
                                # Equal or smaller than piston period
langevinPistonTemp      $temp;          # coupled to heat bath
# run
numsteps                500000;         # run stops when this step is reached
run                     10000000;       # 1ns
```

B

Supplemental Figures

B.1 Visualization of trajectory files

Here we present a visualization of the trajectory files for WT apo 1 using a lag time of 7.5 ns for two different clusters.

Figure B.1.1: Visualisation of the two different states for the WT apo 1 protein using a lag time of 75. Green represents State 1 and blue represent State 2.