Molecular Dynamics Simulations of the Water-Soluble Chlorophyll-binding Protein : Identifying Structural Features Responsible for Spectral Dynamics

Martina Mai

A Thesis

in

The Department

of

Physics

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Science (Physics) at

Concordia University

Montréal, Québec, Canada

August 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Martina Mai**

Entitled: **Molecular Dynamics Simulations of the Water-Soluble Chlorophyll-binding Protein : Identifying Structural Features Responsible for Spectral Dynamics**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Physics)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

________________________________ Chair
Dr. Valter Zazubovits

________________________________ Examiner
Dr. Deniz Meneksedag Erol

________________________________ Examiner
Dr. Laszlo Kalman

________________________________ Supervisor
Dr. Valter Zazubovits

________________________________ Co-supervisor
Dr. Rachael (Ré) Mansbach

Approved by ________________________________
Valter Zazubovits, Chair
Department of Physics

____________ 2023 ________________________________
Pascale Sicotte, Dean
Faculty of Arts and Science

Abstract

Molecular Dynamics Simulations of the Water-Soluble Chlorophyll-binding Protein :
Identifying Structural Features Responsible for Spectral Dynamics

Martina Mai

Photosynthesis is the process responsible for nearly all life on Earth. Pigment-protein complexes, fundamental components of photosynthesis, are the subject of many studies to better understand this process and develop novel approaches to harvesting light energy. Observations of line shifts in the single-molecule optical spectra of such complexes through optical spectroscopy indicate structural changes in the chlorophyll environment. Nonetheless, the specific molecular elements responsible for the observed spectral dynamics remain largely unknown. In this project, we have studied the Water-Soluble Chlorophyll-binding Protein to elucidate some of these molecular-level mechanisms. Molecular Dynamics simulations of the complex were conducted at T_1 = 300 K and T_2= 165 K for 1 μs. The vicinity of the chlorophylls was determined by computing the pigments contact map. At 300 K, small conformational changes were discovered, involving side chain rotations of certain residues, primarily the non-polar Leucine and Valine. From those observations, the protein free energy landscape associated with this generalized coordinate was mapped, and the heights of the energy barriers were determined at around 1000-1500 cm^{-1}. This range of values agrees with experimental results. To gain a deeper understanding of these residues dynamics, we performed Dynamical Network Analysis. It revealed high motion correlations between residues located in close proximity, suggesting a similar rate of conformational change among these residues. Through network connectivity analysis, we found a similar side chain rotation in the same residue in each monomer located farther from the pigments. The energy barrier height associated with this residue is also consistent with experimental results.

Acknowledgments

I would like to express my deepest gratitude to my supervisors, Dr. Zazubovits and Dr. Mansbach, for all the knowledge they shared with me, for the patience they have shown while answering every single one of my questions. I am thankful for their kindness and understanding both academically and personally. I sincerely thank them for giving me the opportunity to carry out this research project and benefit from their expertise.

I am grateful to my committee members, Dr. Kalman and Dr. Meneksedag Erol, for their interest in my work, and for the great advice and suggestions they have provided.

I would like to thank my partner, Thomas, for supporting me in this academic journey (and in life) and always believing in me during the stressful times. I am grateful to have found someone who may know me better than myself and who is able to tell when it is time for me to take a break or when I have to face my fears in order to accomplish my goals.

I am thankful to my family for their continuous support in my decision to study abroad and for their unwavering belief in me. Many thanks to my friends here and overseas for their encouragement and for clearing my mind when I needed it the most. Lastly, I express my gratitude to all my labmates for their friendliness and their knowledge. I have learned a lot from each of them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Photosynthesis plays a vital role as the primary process by which energy from sunlight is converted into chemical energy, supporting the majority of life on Earth. Over the years, high-resolution optical spectroscopy methods have been employed for the study of pigment-protein complexes involved in photosynthesis [1]. These complexes can exist in multiple slightly different structural states, represented as minima of the free energy landscape, even when properly folded. Spectroscopy experiments can be used to investigate them. Minor alterations in the protein structure near the pigment molecule lead to shifts in the spectral lines. At cryogenic temperatures, as spectral lines get narrower, these shifts become more observable, rendering them highly sensitive to local protein dynamics. However, the specific structural elements within the pigment-protein complexes that contribute to these observed changes are not well comprehended.

Protein free energy landscapes have multiple hierarchical levels. A simplified representation of this landscape, called the Two-Level System (TLS) [1] [2] [3], depicts two conformational substates with slightly different optical transition frequencies of the pigment embedded in the protein. This TLS model aligns with optical experiments. The relevant conformational changes occur on a nanosecond timescale in the excited state and on an hour timescale in the ground state. Several mechanisms, including TLS resulting from single and cooperative hydrogen bonds, as well as movements of light side-groups, such as hydroxyl and methyl groups, have been proposed to explain these conformational changes. However, the precise location of these entities remains largely unknown. The primary aim of this research project is to utilize Molecular Dynamics (MD) simulations to analyze the dynamics of the Water-Soluble Chlorophyll-binding Protein (WSCP) and identify the generalized coordinate associated with its conformational change to derive its free energy landscape. WSCP is a homotetramer comprising one chlorophyll molecule per monomer, therefore it is a simple enough system for MD simulations.

Although many MD simulations have been conducted on photosynthetic pigment-protein complexes, none have specifically focused on determining multi-well energy landscapes [4] [5] [6] [7] [8] [9]. Previous simulation studies primarily focused on calculating the transition frequencies of pigments in relation to exciton effects, energy transfer, and charge transfer processes. While some distributions of transition frequencies have been reported, evidence of TLS/MLS (multi-well system) structures has not been demonstrated, possibly due to limitations in spectral resolution, simulation time, or too high temperatures used. On the other hand, the presence of TLS/MLS is firmly supported by the observed spectral shifts. These shifts can only be accounted for by the existence of TLS/MLS, thus affirming their presence in the studied complexes.

The objective of this project is to overcome these limitations and enhance our comprehension of photosynthesis at the molecular level. This will be achieved by conducting MD simulations to uncover the reasons behind the conformational changes observed in optical spectroscopy experiments. Gaining insights into the conformational dynamics of pigment-protein complexes will advance our fundamental understanding of photosynthesis and potentially have implications for renewable energy research.

Chapter 2

Pigment-protein complexes

Photosynthesis is a fundamental process operated by plants, algae and cyanobacteria. It converts light energy into chemical energy by means of Pigment-Protein Complexes (PPCs) located in the chloroplast of eukaryotic cells as well as in certain bacteria. In this chapter, we will introduce three major PPCs - Photosystem II (PS II), Cytochrome b_6f, and Photosystem I (PS I) - that are involved in photosynthesis and located in the thylakoid membrane. Additionally, we will discuss WSCP, another PPC that does not play a direct role in photosynthesis reactions.

2.1 Chlorophyll

Pigments are essential molecules for photosynthesis that are responsible for absorbing light energy from the Sun, as well as serving as primary electron donors. The six major pigments are : Chlorophyll a, Chlorophyll b, Pheophytin a, Pheophytin b, Xanthophyll and Carotene. Because each different pigment is responsible for absorbing different wavelengths of light, all are necessary for photosynthetic organisms to collect as much energy as possible. Due to their fundamental function in oxygenic photosynthesis, chlorophyll a and chlorophyll b are the most important pigments.

Chlorophyll (Chl) a, the most abundant chlorophyll, plays a crucial role in photosynthesis. It is involved in both energy transfer and electron transfer processes [10]. During the light harvesting process, Chl a absorbs photons and ultimately this energy get transferred from antenna / Light Harvesting (LH) complexes to reaction centers. In the reaction center, Chl a once again plays a critical role, this time in electron transfer at it is the primary electron donor in the electron transport chain of photosynthesis. On the contrary, Chl b collects energy to pass into Chl a [10].

Structurally, both chlorophylls consist of a central magnesium atom bonded to a porphyrin ring. A side chain known as a phytol chain is attached to the ring. The pigments only differ by a methyl group in Chl a substituted by an aldehyde group in Chl b (Figure 2.1). This structural difference leads to a shift in their absorption spectra as shown in Figure 2.2. In fact, Chl a mostly absorbs in the blue (around 430 nm) and red (around 662 nm) spectral regions. It also absorbs all the other wavelengths of the visible light spectrum except green [11] [12]. That particular light wavelength is reflected by leaves which gives the green color to plants. On the other hand, Chl b absorbs in slightly different spectral regions, mostly around 453 nm and around 642 nm [11] [12].

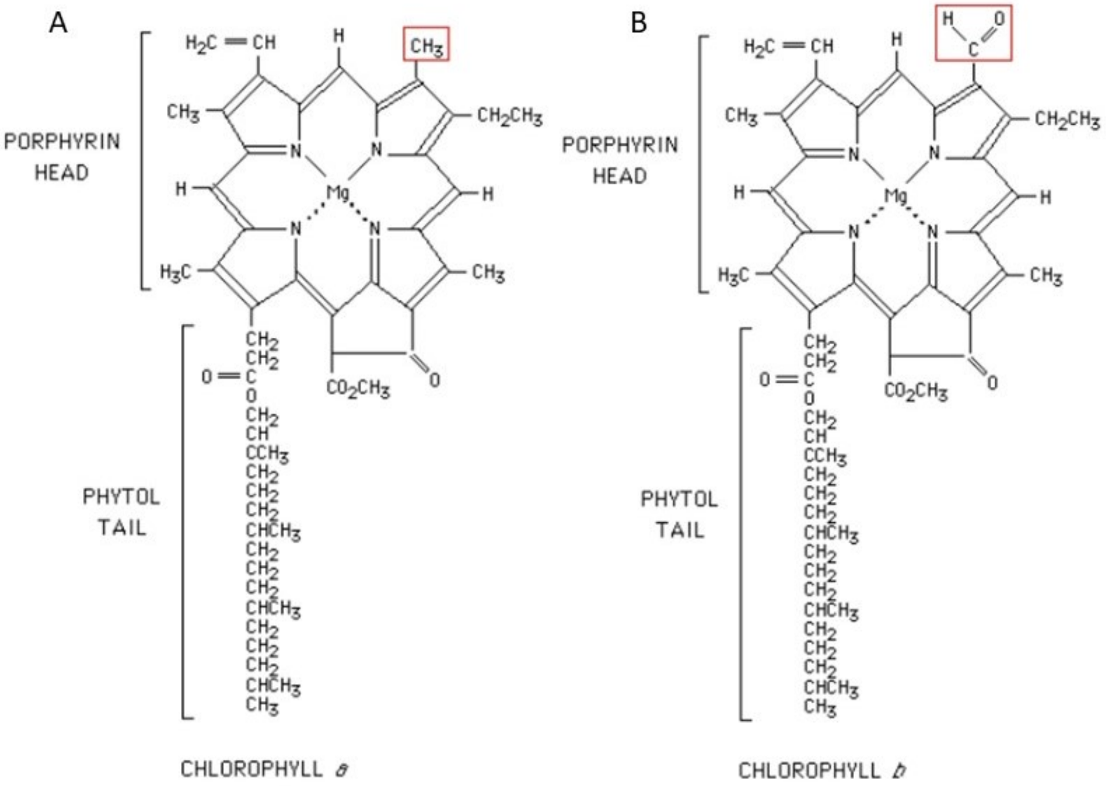


Figure 2.1: Schematic representations of (A) Chlorophyll a and (B) Chlorophyll b. The structural difference is circled in red. [13]

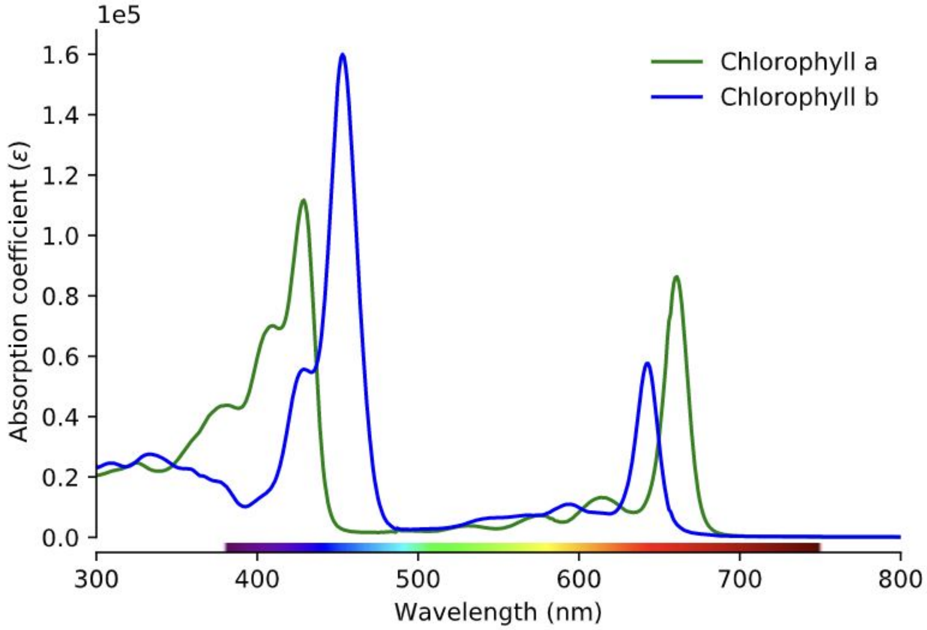


Figure 2.2: Absorption spectra of chlorophyll a (green) and chlorophyll b (blue). [14]

2.2 Photosynthesis

In photosynthesis, plants, algae and cyanobacteria use light energy, water and carbon dioxide to create oxygen and glucose. As mentioned, this process is carried out partly by PPCs : PS II, Cytochrome b_6f and PS I . PPCs are not the only components in photosynthesis but their presence are crucial to ensure conversion of light energy to chemical energy. Without these complexes, photosynthesis would not be possible because the light energy would not be efficiently absorbed and utilized by photosynthetic organisms. The photosynthetic system is depicted in Figure 2.3 and will be discussed in more detail in this section. PS II, struck by photons, passes an electron from Chl to an electron carrier named Plastoquinone. Plastoquinone is responsible for the diffusion of the electron inside the membrane to donate it to the next PPC : Cytochrome b_6f. From there, electron transfer continues to Plastocyanin, which transfers the electron to a Chlorophyll molecule in PS I. The electron is then passed on to Ferredoxin, where it can be used for various metabolic processes.

Each PS is composed of two sub-units : the light-harvesting (LH) antenna and the reaction center. The LH antenna primarily comprises pigments that absorb light energy and become excited.

Subsequently, the pigment electron returns down to the ground state, and the lost energy is acquired by the next chromophore in a chain process known as Excitation Energy Transfer (EET). This process leads to a gradual decrease in electronic excitation energy related to the motion of the energy through the pigments. The excited state energy is transferred from one pigment to another and gradually loses energy due to factors such as energy dissipation through heat and vibrational relaxation. The process ends when it reaches a special pair of Chls in the reaction center.

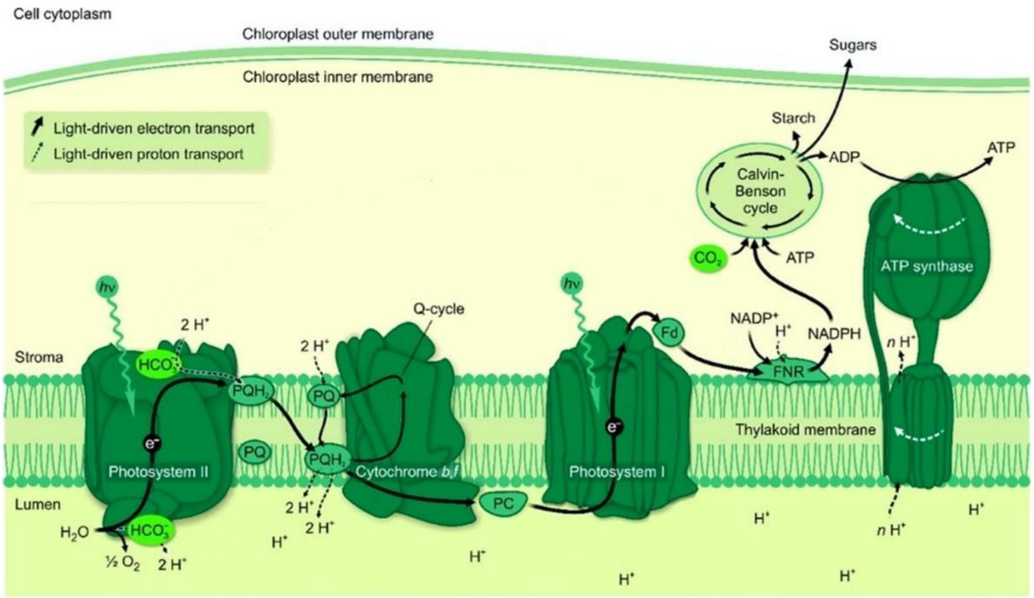


Figure 2.3: Schematic representation of the photosynthetic system, including Photosystem II, Cytochrome b_6f and Photosystem I. [15]

2.2.1 Photosystem II

PS II is responsible for initiating the light-dependent reactions. Energy is transferred through the LH complex until it arrives at its reaction center where energy is transferred to a special pair of Chls named P680 (because of its best absorption at 680 nm). From there, electron transfer starts. As shown in Figure 2.4, P680 transfers an electron to Pheophytin D_1 via Chl D_1, the so called ”accessory chlorophyll”. This electron acceptor passes the electron first to Plastoquinone Q_A then to Plastoquinone Q_B. This process happens twice because Plastoquinone Q_B needs two electrons to be fully reduced and become PQH2. In its reduced form, it can act as a mobile electron carrier to the next component of the system : Cytochrome b_6f. Due to its asymmetry, the reaction center of PS II only supports electron transfer in one branch. As a result, the second branch, does not enable electron transfer from P680 to Plastoquinone Q_B by means of Chl D_2 and Pheophytin D_2. In the meantime, water oxidation takes place in the Oxygen-Evolving Complex (Mn_4CaO_5 cluster). This reaction releases oxygen, protons and electrons according to the equation :

$$2H_2O \rightarrow O_2 + 4H^+ + 4e^- \tag{1}$$

Those electrons are used to reduce $P680^+$.

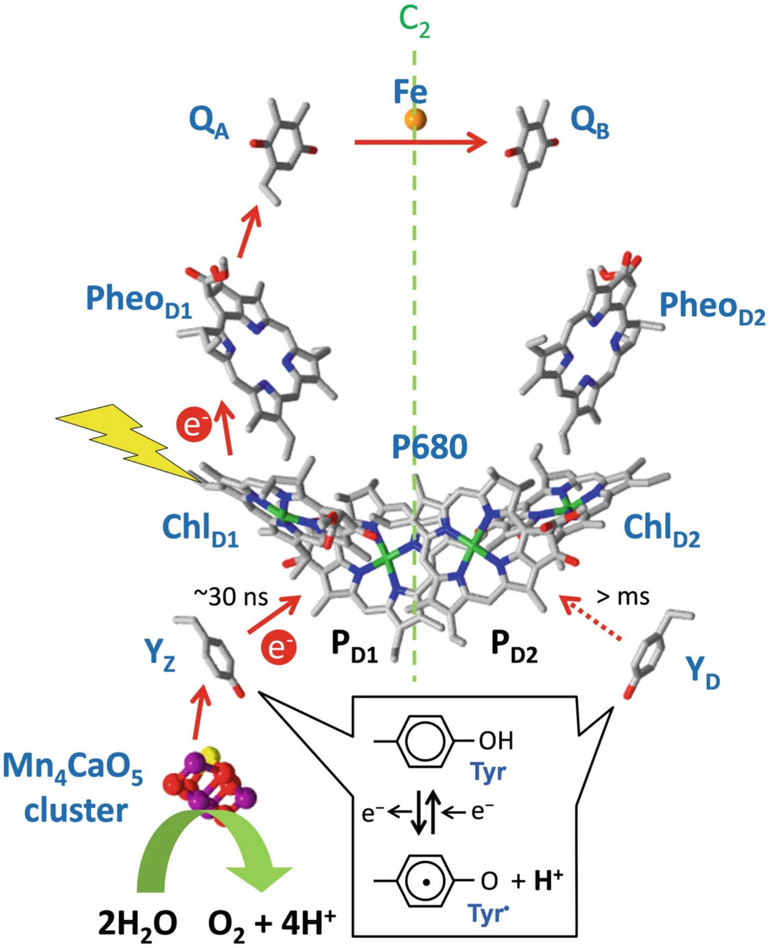


Figure 2.4: Representation of the electron transfer pathway (red arrows) of Photosystem II and the water oxidation reaction by the Mn_4CaO_5 cluster. [16]

To expand the knowledge on PS II, one research group has performed an all-Atom MD simulation at the ns scale. The study focused on the permeability of routes taken by water and oxygen from/to the Oxygen-Evolving Complex [17]. To this day, to achieve simulations of PS II at a µs scale, only coarse-grained [1] MD simulations have been performed [9]. This technique is useful and efficient to reduce the computational cost by removing some atomic details. This work had shown an overall great stability of the PS II core and a highest flexibility on the surface of the complex and near the Plastoquinone exchange cavity [9]. Furthermore, while PS II can function in both monomeric and dimeric arrangements, this study has also shown that in the monomeric form, the complex adopts a surprising conformation and tilts in the thylakoid membrane inducing a buckling of the membrane [9].

To conclude, some MD simulations have been performed on PS II but it remains a challenging task as bonded and non-bonded parameters of many cofactors in PPCs are missing in the force field data.

2.2.2 Cytochrome b$_6$f

The role of Cytochrome b$_6$f is to transfer the electrons from PS II to PS I. Electrons reach Cytochrome b$_6$f with the help of Plastoquinone Q$_B$. Then, they are passed to another mobile electron carrier, Plastocyanin, whose role is to transport them to PS I. This electron transfer is coupled to proton transfer that contributes to proton gradient across the membrane. The function of Chl a and β-carotene pigments in the complex remains unknown. In fact, these pigments cannot be harvesting light because there is no reaction center in the complex. It has also been demonstrated, using site-directed mutagenesis [2] that Chl a and β-carotene pigments in Cytochrome b$_6$f do not participate in electron transfer either [18]. This is due to the pigments not having appropriate electronic properties and not being in the correct position in the protein to facilitate electron transfer. As a result, it is suggested that the pigments may have other functions in the complex.

Structurally, Cytochrome b$_6$f is a dimer with each monomer consisting of 8 sub-units. The four

[1] Coarse-grained Molecular Dynamics is a model used to reduce the computational cost in simulations of large systems. This is achieved by reducing the degrees of freedom and removing fine interaction details.

[2] Site-directed mutagenesis is a technique used to selectively modify amino acid residues in the complex and observed the effect on electron transfer activity.

main sub-units are : Cytochrome b6, Cytochrome f, a Rieske protein [3] and sub-unit IV as shown in Figure 2.5.

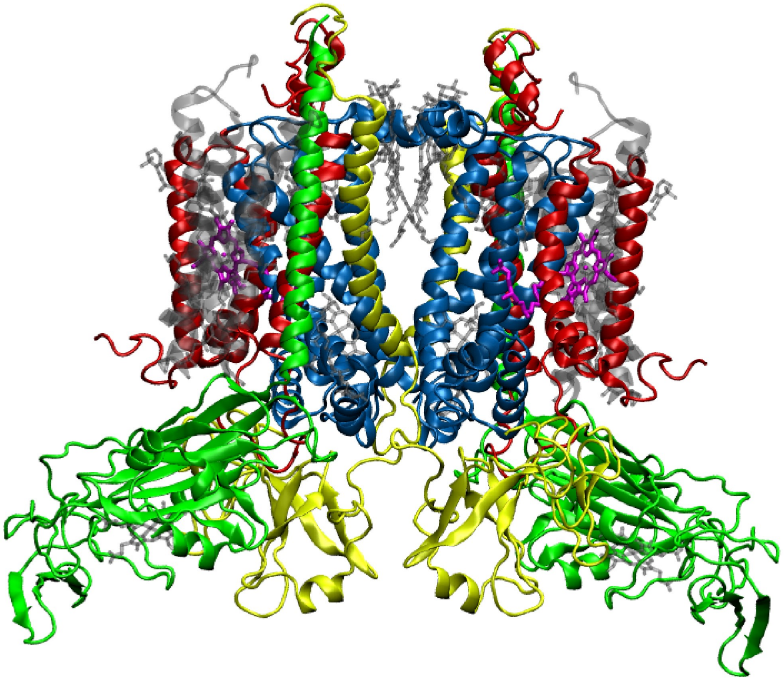


Figure 2.5: Quartenary structure of Cytochrome b_6f from *Mastigocladus laminosus*, PDB : 2E74. Cytochrome b_6 is shown in blue, Cytochrome f in green, the Rieske protein in yellow, sub-unit IV in red and Chlorophylls a in magenta. The other sub-units and cofactors are colored in grey. The figure was rendered using VMD 1.9.4a43.

2.2.3 Photosystem I

PS I is in charge of the transfer of electrons from Plastocyanin to Ferredoxin. The complex is the last PPC involved in the photosynthetic electron transport chain. Analogously to PS II, photons strike PS I inducing energy transfer to the special pair of Chls in the reaction center, named P700. The electrons are then transferred, via several steps (shown in Figure 2.6), to Ferredoxin, which carries them to the NADP reductase.

All-Atom MD Simulations at the nanosecond scale have been performed to study the electron transfer reactions in PS I [20]. It was also shown that PS I exhibits a rigid hydrophobic core and a

[3] Rieske protein is an iron–sulfur protein involved in electron transfer.

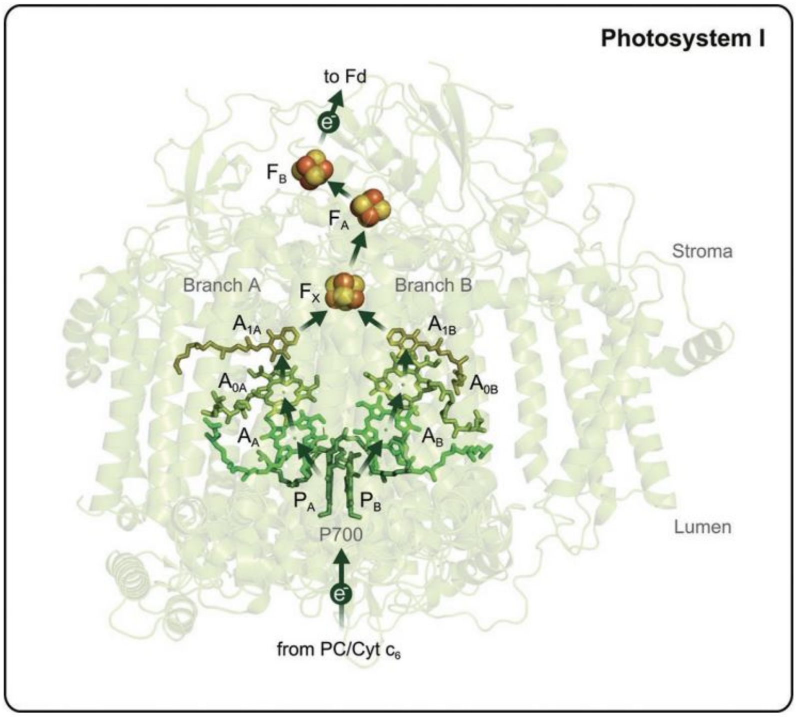


Figure 2.6: Representation of the electron transfer pathways of Photosystem I from Plastocyanin (PC) to Ferrodoxin (Fd). The cofactors are shown in opaque colors and the protein skeleton in a transparent color. The two possible electron transfer pathways (Branch A and Branch B) with equal probability are represented by green arrows. P$_A$ and P$_B$ form the special pair of Chls also referred as P700. A$_A$, A$_{0A}$ are the accessory Chls of Branch A, while A$_B$, A$_{0B}$ are the accessory Chls of Branch B. A$_{1A}$ is the Phylloquinone (electron carrier) of Branch A, while A$_{1A}$ is Phylloquinone of Branch B. F$_X$, F$_A$ and F$_B$ are iron-sulfur centers. [19]

surface with higher mobility. [20]

2.2.4 ATP synthase

ATP synthase is a complex enzyme which plays a crucial role in the synthesis of adenosine triphosphate (ATP), which is the energy product of the light-dependent stage. ATP synthase requires a proton gradient, which is generated from three sources. Firstly, protons are released into the lumen (positive side of the thylakoid membrane) during water oxidation which takes place in PS II (Section 2.2.1). This creates a higher concentration in the lumen compared to the stroma (the negative side of the thylakoid membrane), resulting in the formation of a proton gradient across the thylakoid membrane [21]. Secondly, Cytochrome b$_6$f (Section 2.2.2) pumps protons from the stroma into the lumen , enhancing the proton gradient across the thylakoid membrane. [22]. This proton gradient is formed as electrons flow through the electron transport chain. Lastly, once the electrons reach the NADP reductase (Section 2.2.3), this enzyme reduces NADP$^+$ to NADPH, leading to a decrease in proton concentration in the stroma. The high concentration of protons in the lumen, combined with the low concentration in the stroma, initiates the production of ATP by ATP synthase through the following reaction:

$$ADP + Pi \rightleftharpoons ATP \tag{2}$$

Therefore, the main products of the light-dependent stage are : oxygen (produced by PS II), NADPH (generated by NADP reductase) and ATP (synthesized by ATP synthase). NADPH and ATP, are then used in the light-independent stage of photosynthesis (Calvin cycle) along with carbon dioxide (CO_2) to produce glucose.

2.3 Water Soluble Chlorophyll a-binding Protein

Water-soluble chlorophyll binding proteins (WSCPs) are a group of soluble proteins that are not directly involved in photosynthetic reactions. Rather, they are responsible for binding and stabilizing chlorophyll molecules and act as transporters to shuttle the pigments between different cellular compartments in photosynthetic organisms. WSCPs have been found in a wide range of organisms, including green algae, and higher plants. Unlike most photosynthetic proteins, WSCPs are not bound to thylakoid membranes, but rather float freely in the stroma or cytoplasm of photosynthetic cells. In this section, we will explore the structure and function of WSCP in more detail.

2.3.1 Structure

The complex has a homotetrameric structure in which each sub-unit is composed of one protein chain and one pigment : Chl a or Chl b (Figure 2.7) . The chlorophylls are tightly packed in the complex center and enclosed in a hydrophobic cavity formed by certain amino acids of the monomers [23]. WSCP can be considered as a dimer of dimers where the porphyrin rings of the pigments within a dimer are in close contact (average intra-dimer Mg atoms distance of 10 Å) leading to strong intra-dimer excitonic couplings. The dimers are reported to be weakly coupled together with inter-dimer Mg atoms being on average 20 Å apart [23]. WSCPs are divided into two classes depending on their photoconvertibility[4]. Proteins in WSCP Class I undergo a light-induced absorption change (photoconvertible), whereas proteins in WSCP Class II do not (non-photoconvertible) [24]. Among Class II WSCPs, proteins are separated into two subclasses : Class II-A for proteins with a high ratio (6.3–10.0) of Chl a/b and Class II-B for proteins with a lower ratio (1.0–3.5) of Chl a/b [24] [25].

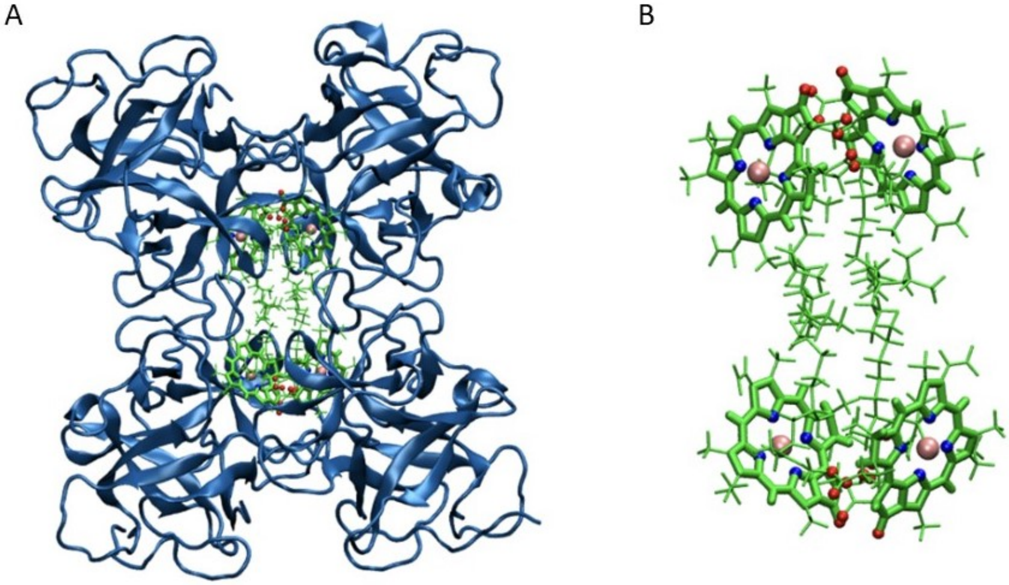


Figure 2.7: Crystal structure of class II-B (Chl a/b ratio around 1.6-1.9) Water-Soluble Chlorophyll-binding Protein (WSCP) from *Lepidium virginicum*. PBD ID : 2DRE. (A) The protein complex is composed of four protein chains (blue) and four chlorophyll a pigments (green) enclosed in a hydrophobic cavity. (B) Closer representation of the four chlorophylls a. Magnesium atoms are colored in pink, nitrogen in blue, oxygen in red and for clarity, hydrogen and carbon in green. The figures were rendered using VMD 1.9.4a43.

The chlorophyll central Mg atom ensures binding between the pigment and its protein sub-unit by displaying a penta-coordinated structure where usually Histidine acts as an axial ligand [26]. However in WSCP, Proline was found to have this function instead of Histidine, as shown in Figure 2.8. The interaction of the Mg atom with the carbonyl side chain of Proline is fundamental for the

[4] Photoconvertibility refers to the ability of a molecule to undergo a change in its properties or state in response to light exposure. This change can involve a conversion of the molecule absorption spectrum, fluorescence, or energy state.

chlorophyll-binding process to the protein; without it the complex falls apart [23]. Chlorophyll, and more precisely its phytol chain, plays a major role in the oligomerization [5] of WSCP : replacing chlorophyll by chlorophyllide (a Chl derivative lacking the phytol tail) prevents oligomerization from taking place [27].

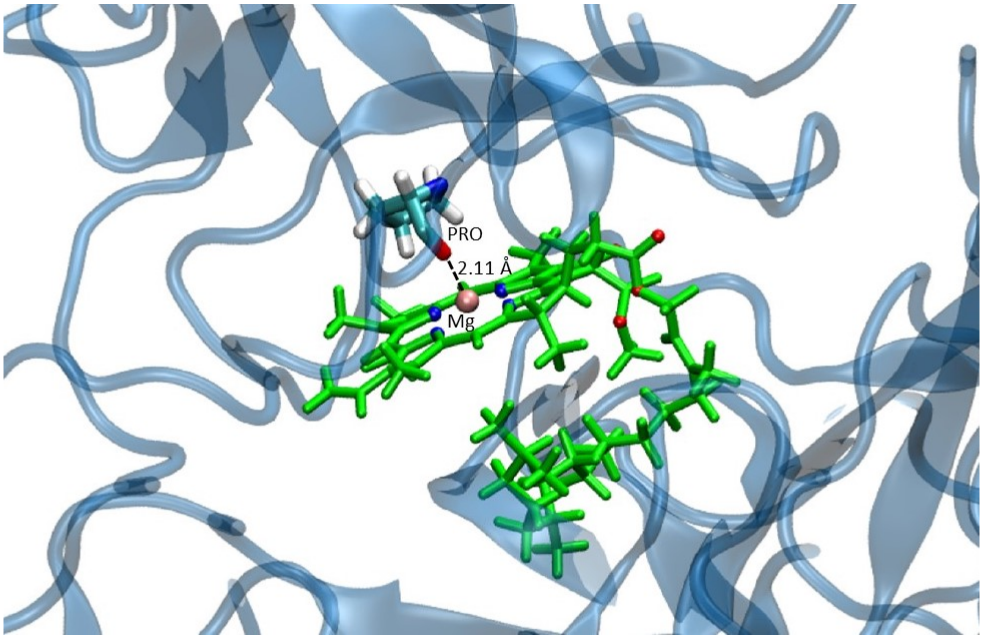


Figure 2.8: Close interaction between the carbonyl side chain of Proline (PRO) and the chlorophyll Mg atom. The distance between the Oxygen atom from the PRO side chain and the Chl Mg atom is labeled. The figure was rendered using VMD 1.9.4a43.

2.3.2 Function

Unlike most of the PPCs, WSCP is not a membrane-bound protein and is not directly involved in photosynthetic reactions due to a combination of factors : the low number of Chls, the absence of other pigments such as carotenoids, and the lack of electron carrier molecules. Consequently, the protein complex can not engage in any electron transfer processes [23]. Thus, its function remains unclear. However, it has been speculated that this protein acts as a scavenger of free Chls, transporting them from the thylakoid membrane to the sites of catabolic reactions [24] [28]. This protein is also known for its high stability against thermal dissociation and denaturation even at

[5] Oligomerization is the process in which monomers are turned into a unique molecule : the oligomer.

100°C [25] [28] [29]. Furthermore, WSCP shows not only a remarkable heat stability but also an unusual photostability, especially considering the absence of the carotenoids that are key components in photoprotection [28]. In addition, many studies have shown that phytol chain conformations of Chls greatly impact the stability of the complex [28] [30].

While the pigments are enclosed in a hydrophobic cavity, a structural analysis of the complex has shown that water molecules could in principle diffuse in the pocket through four small pores [23] . However, an all-atom MD simulation was performed and it was found that the hydrophobic cavity of WSCP was highly stable and rigid, which prevented water molecules from entering the cavity [31]. Additionally, they observed that Chls in the hydrophobic cavity interacted strongly with each other and with the surrounding protein residues, which further hindered water penetration into the cavity [23][31]. Additionally, this research group also performed a series of MD simulations with different mutations in the complex. In one of them, they replaced the Proline residue that interacts with the Mg atom of the Chl with an Alanine residue. They discovered that this mutation significantly destabilized the WSCP protein by disrupting the interaction between the Chl pigments and the surrounding protein residues. This confirms that the interaction between the Chl Mg atom and the Proline is crucial for the stability of the complex [31].

Chapter 3

Protein Dynamics

Protein dynamics refers to the complex process of the movement and changes in the conformation of proteins over time. Proteins are not static structures but rather are dynamic, undergoing fluctuations and changes in their structures and interactions with their environment. Such motion can strongly impact the protein function [32]. Protein dynamics occur on a variety of timescales, from rapid fluctuations in conformation that occur on the timescale of femtoseconds, such as covalent bond vibrations (also referred as molecular vibrations or stretching and bending vibrations) [32], to slower changes that occur on the timescale of hours such as the protein folding and unfolding process for some complexes [33]. Protein dynamics can be influenced by various factors, including changes in the local environment and interactions with other molecules. This chapter will introduce the topic of the protein energy landscape by exploring the folding funnel model and the Two-Level System or double-well potential. Afterwards, the concept and use of the Spectral hole burning experiment will be presented.

3.1 Protein Energy Landscape

The protein energy landscape (PEL), depicted in Figure 3.1, is a model that describes the free energy distribution of all possible conformations of a protein. Due to this flexible structure, proteins can adopt multiple conformations. The PEL represents the protein conformational space as a

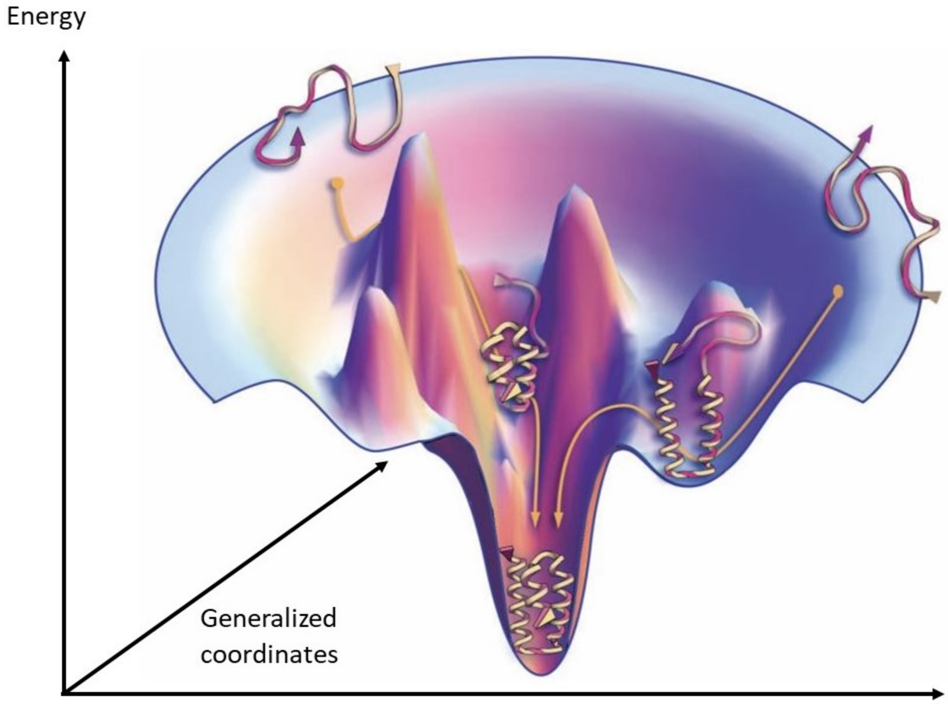


Figure 3.1: Protein energy landscape along two generalized coordinates. The figure is adapted from [34]

multidimensional energy surface, where each point on the surface corresponds to a different conformation of the protein. Each dimension in the energy landscape corresponds to a different protein degree of freedom characterized by a generalized coordinate, also known as a collective variable or an order parameter. The generalized coordinates of a protein can involve various parameters, such as the backbone angle, the side chain orientation, and many others, including combinations of all possible factors. The energy landscape can change as a result of environmental conditions, such as the temperature, the pH and the presence of ligands [35]. The folding funnel model is often employed as a theoretical model to describe the energy landscape of proteins, especially in the context of protein folding. In practice, the PEL can be represented in lower dimensions using dimensionality reduction algorithms. The folding funnel model, on the other hand, focuses on the idea that there is a single lowest energy state (the native state) with intermediate metastable states along the folding pathway. It should be noted that the folding funnel model is more applicable to structured proteins rather than disordered proteins.

3.1.1 Folding Funnel Model

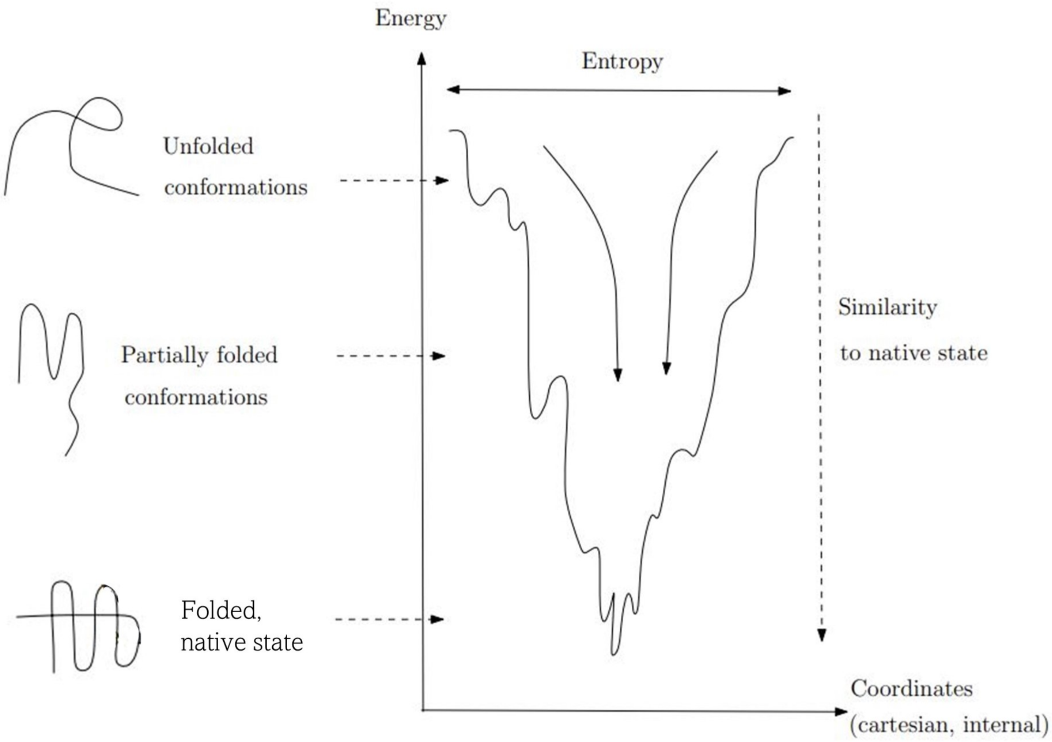


Figure 3.2: Schematic representation of the folding funnel model. The figure is adapted from [36]

The folding funnel model, shown in Figure 3.2, depicts the energy of the protein as a rugged funnel in which the higher energy wells are intermediate energy states that represent conformational substates and the lowest energy well indicates the protein native state. During the folding process, the protein moves along the energy surface. It overcomes energy barriers that separate different conformations, until it reaches the bottom of the funnel which represents the global energy minimum of the landscape and the functional and native state of the protein. The roughness of the funnel corresponds to all possible conformations that a protein can adopt as it folds. As the protein folds, it encounters many local energy minima, which correspond to partially folded intermediate states in which the protein may be trapped [37]. Even the native structure of the protein is not only represented by one state, but rather by several nearly-identical ones [37]. Hence, at cryogenic temperatures, properly folded proteins are studied, in which only those nearly-identical states are present. Consequently, many features of such systems can be captured by the Two Level System

(TLS) or double-well potential.

3.1.2 Two-Level System

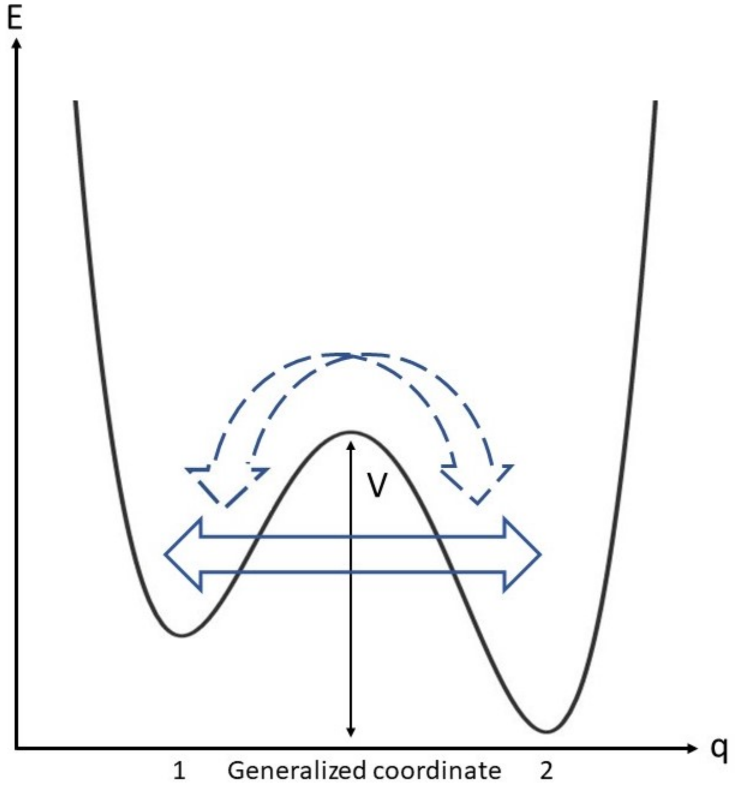


Figure 3.3: Schematic representation of the TLS or double-well potential. Tunneling is represented by the blue solid double arrow and thermally-activated barrier-hopping by the blue dashed double arrow. The potential energy barrier is V.

The Two-Level System (TLS), also known as the double-well potential, is a simplified representation of the protein energy landscape that only considers two protein conformational substates along a generalized coordinate, as shown in Figure 3.3. It is a useful model to study protein dynamics at cryogenic temperatures where the system is nearly frozen and relevant degrees of freedom are reduced. The proteins are then unable to transition over higher barriers and instead tend to remain in lower energy states. However, the system is not completely static due to many protein internal and external factors such as thermal fluctuation and exciton-phonon interactions [38]. Originally, this model was used to describe energy landscapes of glasses, but appears to also provide useful information for proteins [39] [40]. Still, the double-well potential serves as a simplified model, which, while useful, does not provide explicit details about the entities undergoing conformational changes. This lack of explicit detail can be explained by two distinct factors. First, the model assumes a limited number of conformational states available to the system due to low thermal fluctuations or restricted energy levels. Second, the choice of generalized coordinate(s) used in the two-well potential may not directly correspond to the microscopic degrees of freedom responsible for the conformational changes. If the correct microscopic degrees of freedom were appropriately considered, a two-well potential could potentially offer explicit insights into the entities undergoing conformational changes.

The energy wells or states are separated by a potential energy barrier V which represents the activation energy that must be overcome to transition between the two states. The height of this potential energy barrier determines the rate of the transition, and both are greatly influenced by the temperature. However, in the approximation often used for spectroscopy data analysis, the temperature dependence of the barrier heights is usually ignored. Crossing the energy barrier can occur either by thermally-activated barrier-hopping or tunneling. At physiological temperatures, classical barrier-hopping dominates due to the high thermal energy available to the protein. In this process, the protein molecules can move over the energy barrier by random thermal fluctuations. At cryogenic temperatures, however, the thermal energy available to the protein is much lower, making it more difficult for the protein to cross the energy barrier through thermally-activated barrier-hopping. Therefore, at cryogenic temperatures, quantum tunneling becomes a more significant mechanism for crossing the barrier. The tunneling rate Γ can be described using the semiclassical approximation, known as the Wentzel-Kramers-Brillouin (WKB) approximation. The tunneling rate Γ can be expressed as [41] :

$$\Gamma = \omega \; exp(-2\lambda) \tag{3}$$

where ω is the attempt frequency and λ the tunneling parameter. In the case where the generalized coordinate is a variation of angle $\Delta\alpha$, the tunneling parameter λ can be defined as :

$$\lambda = \frac{\Delta\alpha}{\hbar}\sqrt{2IV} \tag{4}$$

$\hbar$ is the Planck constant, I is the moment of inertia, and V is the height of the energy potential barrier. In the case where the generalized coordinate is a distance d, the tunneling parameter λ can be defined as :

$$\lambda = \frac{d}{\hbar}\sqrt{2mV} \tag{5}$$

where m is the effective mass of the involved atoms/proteins. The double-well potential is commonly used to model the spectral dynamics of Pigment-protein complexes, which can be probed experimentally through Spectral Hole Burning techniques [42] [43] [44] [45].

3.2 Spectral Hole Burning

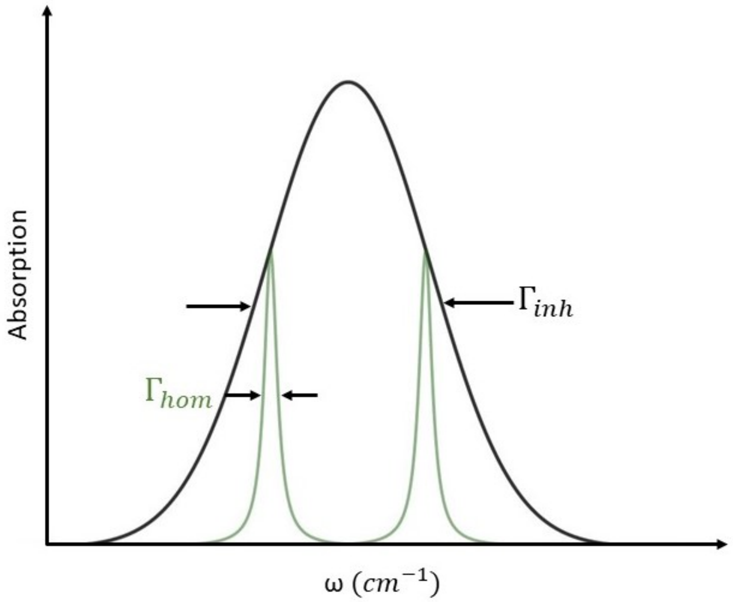


Figure 3.4: Large inhomogeneous absorption spectrum of width Γ_{inh} (black) and several hidden homogeneous bands of width Γ_{hom} (green).

Spectral Hole Burning (SHB) is a technique widely used to study the spectral dynamics of

Pigment-protein complexes. In PPCs, there are small variations, from one complex to another (e.g. one PSI to another PSI), between local environment of the pigment in supposedly structurally identical position, which causes inhomogeneous broadening of the absorption bands, leading to a broad absorption spectrum as depicted in Figure 3.4. As a result, the spectral profile has a wide Gaussian shape [46]. The inhomogeneously broadened band consists of multiple homogeneous bands of individual pigment molecules that contain information about the dynamics of the excited state of the complex.

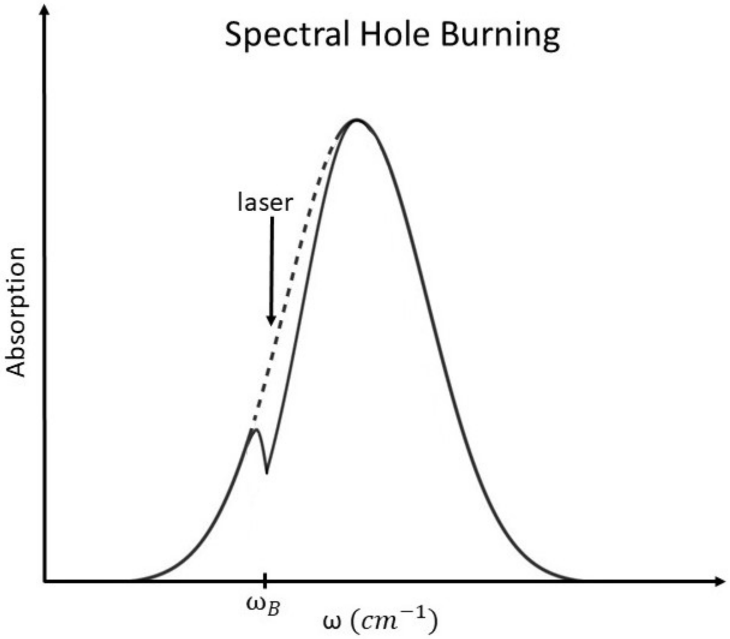


Figure 3.5: Spectral Hole Burning mechanism, with burn frequency ω_B.

Non-Photochemical Hole Burning is a type of SHB in which the environment of the pigments is altered, leaving a hole, as illustrated in Figure 3.5, in the PPC absorption spectrum. This technique is used at low temperature to minimize the width of the homogeneous bands and maintain a stable environment for the chromophores. In this experiment, the PPCs are initially in the first well of the ground state. A narrow-band laser is used to irradiate specific wavelengths of the PPCs absorption spectra. Therefore, the pigments in resonance with the laser frequency ω_B will be found in the first excited state. The excited chromophores then cross the potential energy barrier by thermally-activated barrier-hopping or tunneling to reach the second well of the excited state. At 10 K or

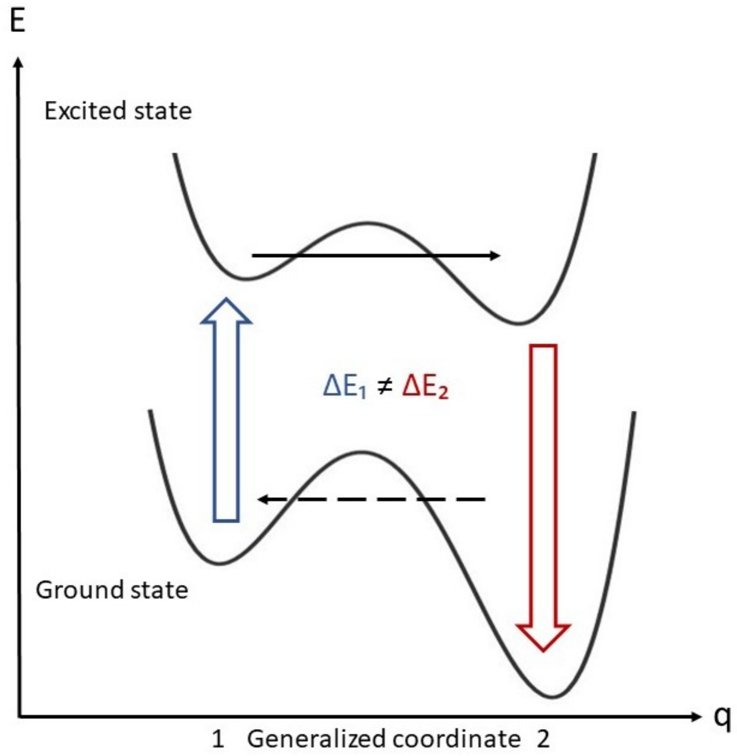


Figure 3.6: Schematic of the Non-photochemical Hole Burning mechanism. Double-wells in the ground state and excited state of the pigment are shown. The pigment excitation process (blue arrow) and relaxation process (red arrow) are depicted. The pigment transition between wells in the excited state is represented by the plain black arrow and the hole recovery by the dashed black arrow.

less, tunneling is more likely to occur. The pigments will eventually lose energy and return to the ground state, but in the second well, where the absorption frequency of the molecule may be slightly different. As a result, a hole appears in the absorption spectrum. After some time, the hole recovers, as the chromophores return to their original well by thermally-activated barrier-hopping or tunneling. By keeping track of the hole recovery, the distribution of the ground state potential barrier height V, which is greater than the barrier height of the excited state, can be obtained. One can also determine the tunneling parameter λ for the ground state and then get some estimate of $\Delta\alpha^2 I$ (see Equation 4) or md^2 (see Equation 5).

To model the conformational dynamics of the complex in Non-photochemical Hole Burning, a double TLS can be used as shown in Figure 3.6 and consists of one double-well for the electronic ground state of the pigment molecule and another one for the electronic excited state of the pigment molecule. Each individual chromophore has a different double TLS based on the slightly varying local environment of the pigments [42]. Using this model, estimations of some parameters such as the potential barrier heights, the effective mass m of the tunneling entity and the distance d between two local minima can be obtained. However, to verify these estimates and determine the specific structural features responsible for spectral dynamics, independent information is needed. Thus, in this project, the Molecular Dynamics simulations technique is employed to investigate this information.

Chapter 4

Research Procedure

The objective of this computational research project is to explore the dynamic characteristics of the Water Soluble Chlorophyll a-binding Protein (WSCP) complex, which was previously introduced in Section 2.3. To accomplish this goal, we used a variety of computational tools. Firstly, the theory behind the major simulation technique (Molecular Dynamics simulation) we employed will be discussed. Then, the theory behind the major analysis technique (Dynamical Network Analysis) we employed will be discussed. Lastly, the specifics of their implementation will be described.

4.1 Molecular Dynamics Simulations

Molecular Dynamics (MD) simulation is a computational technique used to study the behavior and dynamics of molecules at the atomic level, providing insight about the relationships between their structure and function. In the context of biophysical systems, it can predict the evolution of a complex for a fixed amount of time under specific conditions and can capture several biomolecular processes, such as conformational changes and protein folding. MD simulation is a powerful tool that is capable of both confirming experimental results and can also be performed prior to experiments to reduce the range of investigation, as experiments can be expensive, laborious and time consuming. In fact, by simulating molecules in different conditions, understanding about the molecule dynamics are acquired and key variables can be identified. As a result, experimental conditions and setups can be optimized accordingly, focusing on the most promising paths that are most likely to yield meaningful results. MD simulation can also provide information that can be arduous to find with non-computational methods, such as Spectral Hole Burning (Section 3.2), X-ray crystallography [47], NMR spectroscopy [48], and cryo-electron microscopy [49]. For instance, the atomic resolution of MD simulations reveals details of molecule folding pathways, conformational changes, and intermolecular interactions that are challenging to observe experimentally. In this section, both the theory and the protocol of MD simulations will be covered as well as the reproducibility and reliability of the results they produce.

4.1.1 Theory behind Molecular Dynamics Simulations

In order to model the motion and behavior of atoms in a system over time, MD simulations [50] are performed using classical mechanics. Atomic positions are updated over time according to Newton's Second Law :

$$\mathbf{F}_i = m_i\mathbf{a}_i = -\nabla U(\mathbf{r}_i) = -\nabla U(x_i, y_i, z_i) = -(\frac{\partial U}{\partial x_i}\hat{\mathbf{x}} + \frac{\partial U}{\partial y_i}\hat{\mathbf{y}} + \frac{\partial U}{\partial z_i}\hat{\mathbf{z}}) \tag{6}$$

where $\mathbf{F}_i$ is the force on a particle i with a mass m_i and an acceleration $\mathbf{a}_i$. The force is obtained from the negative gradient of the interatomic potential energy U with respect to the Cartesian coordinates (x_i, y_i, z_i) of the particle. The interatomic potential energy U accounts for bonded and non-bonded interactions :

$$U = U_{bonded} + U_{non-bonded} \tag{7}$$

with

$$U_{bonded} = U_{bond} + U_{angle} + U_{dihedral} + U_{improper} + U_{Urey-Bradley} \tag{8}$$

$$U_{non-bonded} = U_{VDW} + U_{electrostatic} \tag{9}$$

These specific terms and the potential energy functions will be discussed in more detail in section 4.1.1.

There are a number of parameters, specifications, and special procedures that are generally undertaken in the context of an MD simulation. The most important points to be considered are described in this subsection. The detailed methodology of the MD simulations performed in this research project will be explained in Section 4.1.

System definition and initialization

Before any simulation can be performed, initial conditions must be determined. In many cases, this includes providing the number of atoms, their initial positions, velocities, masses, and the type of interactions between them. For large biomolecules with experimentally determined structure, this information is usually obtained from the Research Collaboratory for Structural Bioinformatics (RCSB) website[1], or the AlphaFold website[2], which both provide the Protein Data Bank (PDB) files. These structures have been determined using various experimental techniques, such as X-ray crystallography, NMR spectroscopy, and cryoelectron microscopy. For smaller biomolecules such as peptides, the starting structures can be obtained using the PEP-FOLD3 tool [51], or the Rosetta software [52]. The initial velocities of all the particles in the system must be determined. They can be assigned using distributions such as the Maxwell-Boltzmann distribution [53] or a uniform distribution [54].

Force Field Considerations

In order to accurately simulate the behavior of atoms in a system over time, all interactions between the atoms need to be described. A common approach is to describe the behavior of atoms by considering pairwise interactions. To accurately simulate the behavior of atoms in a system over time, these pairwise interactions are described using a potential energy function U. This approach assumes that the interactions between atoms can be effectively captured by two-body force interactions, even though in reality, the interactions involve more complex many-body effects. It is important to note that this is one way of modeling the interactions in MD simulations, and while it is a typical setup, alternative methods using quantum mechanical calculations, such as Density

[1] http://www.rcsb.org/

[2] https://alphafold.ebi.ac.uk/

Functional Theory (DFT) [55] and Quantum Mechanics/Molecular Mechanics (QM/MM) [56] exist, which may account for additional interactions.

In the context of MD simulations, the potential energy function is commonly known as the force field. The force fields differ in several aspects, including the parameterization, and the specific interactions they consider. For this reason, the choice of the force field must be made considering the studied system. Some force fields are more suitable for proteins, nucleic acids, or small molecules. In this project, the CHARMM force field [57] was used as it is one of the force fields recommended for protein studies. Its potential energy function is given by :

$$
\begin{aligned}
U = &\sum_{bonds} k_b(b-b_0)^2 + \sum_{angles} k_\theta(\theta-\theta_0)^2 + \sum_{dihedrals} k_\phi[1+cos(n\phi-\delta)] \\
&+ \sum_{impropers} k_\omega(\omega-\omega_0)^2 + \sum_{Urey-Bradley} k_u(u-u_0)^2 \\
&+ \sum_{Lennard-Jones} \epsilon_{ij}\,[(\frac{R_{min_{ij}}}{r_{ij}})^{12} - 2(\frac{R_{min_{ij}}}{r_{ij}})^{6}] + \sum_{electrostatic} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}
\end{aligned}
\tag{10}
$$

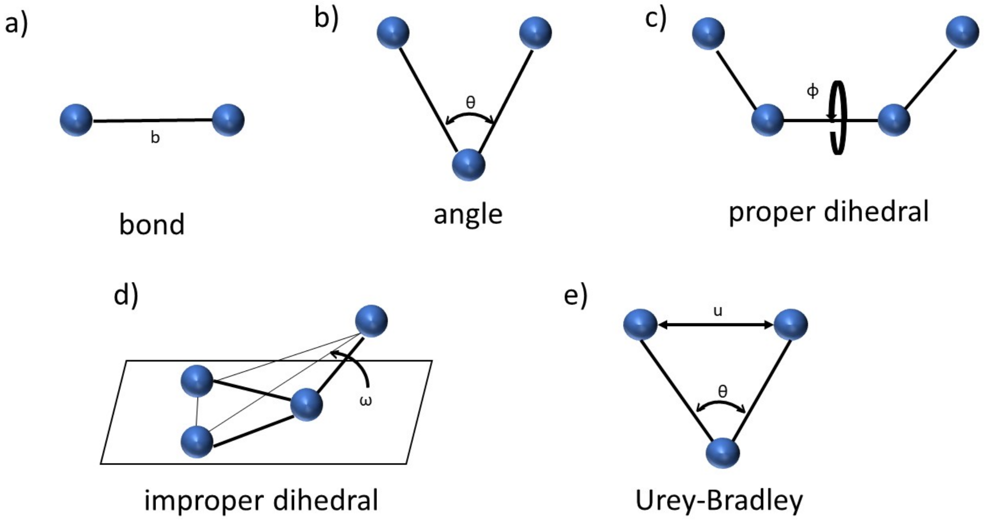


Figure 4.1: Representation of the bonded interaction terms in the CHARMM36 force field. *a)* represents the bond oscillations term, *b)* the angle oscillations term, *c)* the proper dihedral angles term, *d)* the improper dihedral angles term, and *e)* the Urey-Bradley term.

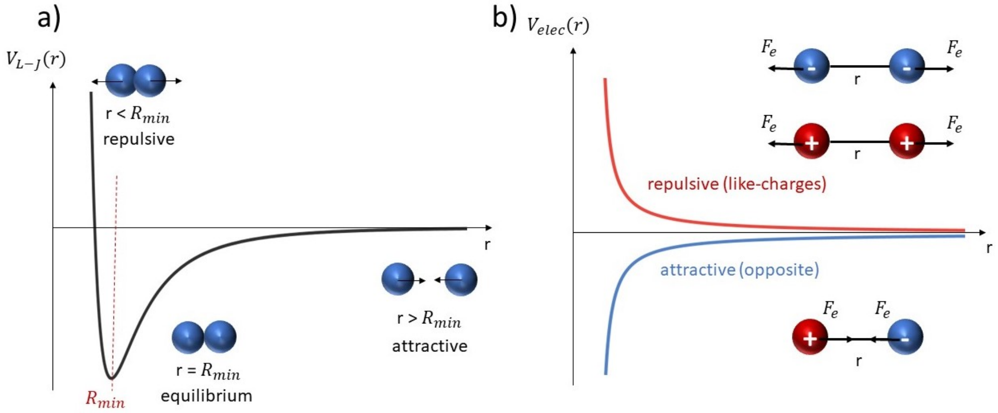


Figure 4.2: Representation of the non-bonded interaction terms in the CHARMM36 force field. *a)* represents the Van der Waals potential term, and *b)* the Coulomb electrostatic term.

The first five terms account for bonded interactions, as illustrated in Figure 4.1. The first term accounts for the bond oscillations around the equilibrium bond length b_0 with the specified bond force constant k_b. The second term describes the angle oscillations around the equilibrium angle θ_0 with the specified angle force constant k_θ. The third term accounts for the dihedral angles where k_ϕ is the amplitude, ϕ is the dihedral angle, n is the periodicity, and δ is the phase. The fourth term accounts for out-of-plane bending (which applies to any set of four atoms that are not successively bonded), where k_ω is the force constant and $\omega - \omega_0$ is the out-of-plane angle. The fifth term is a cross-term that accounts for 1,3 non-bonded interactions not accounted for by the bond and angle terms where k_u is the force constant and u is the distance between the 1,3 atoms. The last two terms account for non-bonded interactions, as illustrated in Figure 4.2. The sixth term is the Lennard-Jones (L-J) potential [58] which approximates the Van der Waals interactions [59]. It consists of two components : a repulsive term and an attractive term. The repulsive term, $(\frac{R_{min_{ij}}}{r_{ij}})^{12}$, accounts for the strong repulsion between atoms when they are very close to each other, preventing their overlap. The attractive term, $(\frac{R_{min_{ij}}}{r_{ij}})^{6}$, represents the weaker, long-range attractive forces between atoms. The last term is the Coulombic potential which describes the electrostatic interactions between pairs of atoms which are represented as point charges, q_i and q_j.

Periodic boundary conditions

Generally, the molecule or molecules are placed into a simulated ”box” modeling environmental quantities such as solvent. The box can have multiple shapes and the box type should be choosen based on the protein complex geometry to accurately represent the system and minimize boundary effects. Some box types can also allow a reduction in computational cost. As an example, the rhombic dodecahedron box (Figure 4.3) is a symmetrical and compact unit cell, considered the smallest among the space-filling unit cells. Compared to a cubic box with the same inter-image distance, the rhombic dodecahedron has a volume that is approximately 71% of the cubic box volume.

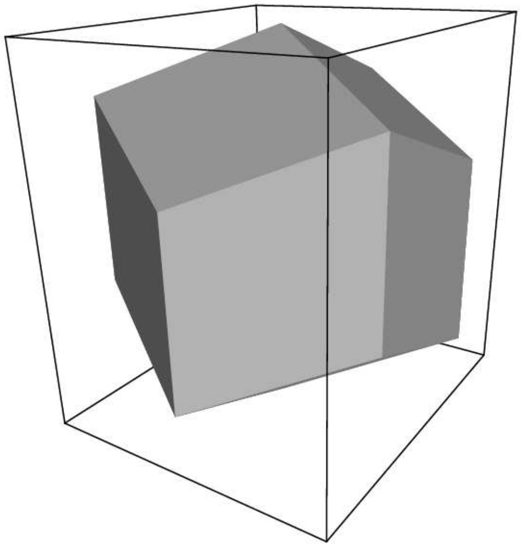


Figure 4.3: Comparison between a cubic box and a rhombic dodecahedron box. [60]

Due to the finite size of the boxes, particles located near the boundaries can have artificial interactions since there are no surrounding particles beyond the box. As a result, the calculations of forces and energies may be inaccurate, as these interactions do not reliably represent the behavior of the actual system anymore. Furthermore, the particles near the boundaries can potentially move out of the box, causing a reduction in the total number of particles within the system. As a consequence, the simulation fails to provide an accurate representation of the system dynamics. To overcome these issues, periodic boundary conditions (PBCs) (Figure 4.4) can be employed.

PBCs allow particles that move out of one side of the box to reappear on the opposite side,

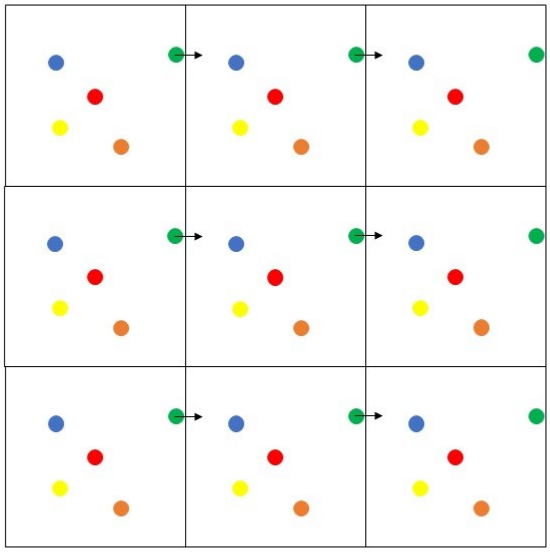


Figure 4.4: Periodic boundary conditions

creating periodic images. This ensures a continuous representation of the system. In addition, PBCs ensure that the total number of particles in the system remains constant throughout the simulation, preserving the accuracy and consistency of the calculations.

The integration time step

Setting an appropriate time step Δt is crucial in MD simulations as it directly affects accuracy and convergence of the system. A smaller time step allows a more precise and accurate simulation as it captures finer details and interactions between atoms, but increases computational costs. On the other hand, a larger time step means that the simulation is taking larger jumps in time, potentially missing important details and interactions between atoms. This can result in an inaccurate representation of the system dynamics and in extreme cases it can cause the simulation to error out because of extreme forces. Thus, selecting a suitable time step is essential in MD simulations. The time step should be small enough relative to the fastest motion. For biomolecular MD simulations, hydrogen atom motion is typically the fastest, therefore a recommended time step is either 1 fs [61] or 2 fs if the hydrogen bonds are constrained [62].

Energy Minimization

Before running an MD simulation, it is important to minimize the energy of the system to remove any steric clashes. This is typically done using an optimization algorithm such as steepest descent algorithm. It uses an iterative approach to gradually minimize the energy of the system. At each iteration, the atomic positions are updated based on the steepest descent direction[3] and step size[4]. The goal of the energy minimization algorithm is to find a configuration of atomic positions $\mathbf{r}_n$ that minimizes the potential energy $U(\mathbf{r})$. In the next paragraphs, the steepest descent algorithm implemented by Gromacs 2021.4 is explained. Molecular dynamics simulation and energy minimization differ in their requirements, with the former necessitating additional information such as temperature, initial particle velocities, and time to accurately simulate the motion of the particles. In contrast, energy minimization does not rely on these factors to achieve its goal.

First, the initial atomic positions $\mathbf{r}_0$ are read from the coordinate file, such as a PDB file. An iteration counter n is set to 0 and a step size h_0 is set. The potential energy $U(\mathbf{r}_n)$ and the forces $\mathbf{F}_n$ on each atom are calculated based on the current atomic positions $\mathbf{r}_n$ and using the choosen force field as described in Equation 10. For normalization, the maximum component of the force vector $\max(|\mathbf{F}_n|)$ is found. The new atomic positions $\mathbf{r}_{n+1}$ can then be calculated using the steepest descent algorithm equation:

$$\mathbf{r}_{n+1} = \mathbf{r}_n - \frac{\mathbf{F}_n}{\max(|\mathbf{F}_n|)} h_n. \tag{11}$$

This equation normalizes the forces $\mathbf{F}_n$ on each atom by the maximum force component $\max(|\mathbf{F}_n|)$, and multiplies it by the step size h_n. The resulting vector is subtracted from the current atomic positions $\mathbf{r}_n$ to get the new atomic positions $\mathbf{r}_{n+1}$. From $\mathbf{r}_{n+1}$, the potential energy $U(\mathbf{r}_{n+1})$ can be obtained. Since the goal is to move the system towards a local energy minimum, the step size h_n is adapted at each iteration based on the change in potential energy. If the new potential energy is lower than the previous potential energy, it means that the algorithm is moving in the right direction, the new positions are accepted and the step size is increased to speed up convergence. If the new potential energy is higher, it means that the algorithm exceed the minimum, and the positions are

[3] The steepest descent direction is the direction in which the potential energy decreases the fastest

[4] The step size is the distance moved along the steepest descent direction at each iteration of the algorithm

rejected and the step size is decreased to avoid moving further away from the minimum. In specific :

If $U(\mathbf{r}_{n+1}) < U(\mathbf{r}_n)$, the new positions are accepted and $h_{n+1} = 1.2h_n$.

If $U(\mathbf{r}_{n+1}) \geq U(\mathbf{r}_n)$, the new positions are rejected and $h_{n+1} = 0.2h_n$.

The iteration counter is then increased by 1, and the algorithm continues until a convergence criterion is satisfied, such as a specified maximum number of iterations or a small enough change in the potential energy function between iterations.

Equilibration of the system

After energy minimization, the equilibration step is performed to ensure that the system has reached a stable state before starting the MD simulation. The purpose of equilibration is to set the system temperature, pressure, and other thermodynamic properties to their desired values, and to allow the system to relax into a more realistic configuration. As an example, equilibration can allow the solvent and ions around the protein to attain more realistic configurations than simple energy minimization. During equilibration, a thermostat and a barostat can be used to control the temperature and the pressure of the system. These tools help maintain the desired conditions and facilitate the relaxation process. Thermostats and barostats can also be employed during the production phase of the simulations to ensure consistency throughout the simulation. During the process, equations of motion are also usually integrated by means of algorithms, and the procedure will be discussed in Section 4.1.1. In our simulations (see the detailed methodology in Section 4.3), the equilibration step was conducted in two phases. The first phase was conducted under the canonical NVT (constant Number of particles, Volume, and Temperature) ensemble to set the desired temperature of the system and to assign the initial velocities of the particles for the MD simulation. To achieve this, multiple thermostats can be used, such as the Bussi-Donadio-Parrinello thermostat [63]. This thermostat works by scaling the velocities of each particle in the system to adjust the temperature. A stochastic term is added to the scaling factor to achieve canonical sampling of the system.

First, the current kinetic energy of the system K is calculated :

$$K = \frac{1}{2}\sum_{i=1}^{N} m_i v_i^2 \tag{12}$$

Then, the velocities of each particle are scaled by a factor λ :

$$v_i(t + \Delta t) = \lambda v_i(t) \tag{13}$$

with λ given by :

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T}\left\{\frac{T_0}{T(t - \frac{1}{2}\Delta t)} - 1\right\}\right]^{1/2} \tag{14}$$

where Δt is the time step, τ_T is the temperature coupling time constant, T_0 is the desired temperature, and $T(t - \frac{1}{2}\Delta t)$ is the temperature of the system at the previous time step.

The new kinetic energy is then modified by a small amount $\mathrm{d}K$:

$$\mathrm{d}K = (K_0 - K)\frac{\mathrm{d}t}{\tau_T} + 2\sqrt{\frac{KK_0}{N_f}}\frac{\mathrm{d}W}{\sqrt{\tau_T}}, \tag{15}$$

where K_0 is the target kinetic energy, $\mathrm{d}W$ is the stochastic term and is a random number drawn from a normal distribution with mean 0 and variance 1, and N_f is the number of degrees of freedom in the system.

The new temperature is obtained using the updated kinetic energy :

$$T = \frac{2(K + \mathrm{d}K)}{3Nk_B} \tag{16}$$

where N is the total number of particles. The steps are repeated at each time step of the simulation.

Once the target temperature is reached, the second phase of the equilibration step was conducted under the NPT (constant Number of particles, Pressure, and Temperature) ensemble to set the desired pressure. To achieve this, several barostats can be used, such as the Berendsen barostat [64]. This barostat is generally used with the Bussi-Donadio-Parrinello thermostat to control both temperature and the pressure simultaneously during the NPT phase.

This barostat works by scaling the volume of the simulation box at each time step to reach the desired pressure. Reducing the size of the box results in increased compression of atoms, leading to an overall higher pressure of the system. On the other hand, increasing the size of the box results in decreased compression of atoms, leading to an overall lower pressure of the system. Analogously to Equation 13, the volume of the box is scale by a factor μ :

$$V_(t + \Delta t) = \mu V_(t) \tag{17}$$

In practice, this is achieved by scaling the positions of the particles instead of the volume of the box. Therefore :

$$r_i(t + \Delta t) = \mu^{\frac{1}{3}} r_i(t) \tag{18}$$

The change of pressure is proportional to the difference in pressure between the target pressure P_0, and the current pressure of the system P :

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} = \frac{\mathbf{P}_0 - \mathbf{P}}{\tau_p} \tag{19}$$

where τ_p is the relaxation time for the barostat. Finally, the explicit expression for the scaling factor λ in the Berendsen barostat is :

$$\mu = 1 - \frac{\beta \Delta t}{3\tau_p}(P_0 - P) \tag{20}$$

where β is the isothermal compressibility. As the pressure difference $(P_0 - P)$ increases, the resulting scaling factor μ also increases.

Production simulation

Finally, the simulation can be run using the chosen simulation method and parameters. The positions and velocities of all particles are updated at each time step by integrating the equations of motions by means of multiple algorithms. The one used in this research project is the leapfrog algorithm (Figure 4.5) [65].

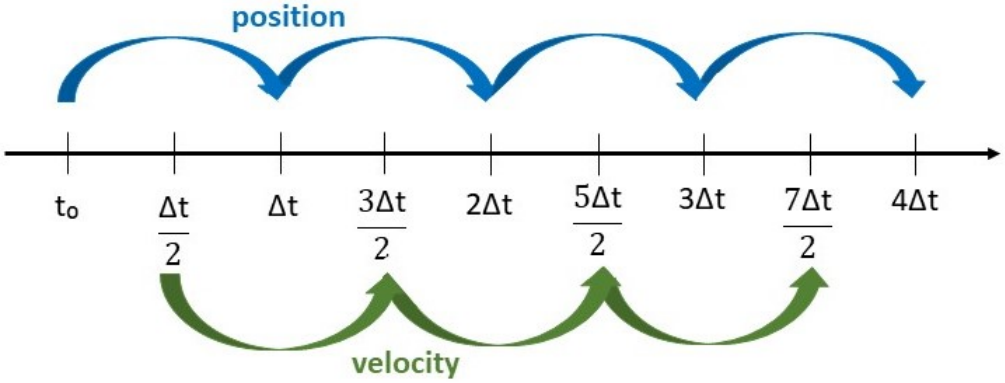


Figure 4.5: Leapfrog algorithm

This algorithm updates the positions and velocities of the particles in discrete time steps. The velocities of the atoms are updated by half a time step $\frac{\Delta t}{2}$ using the forces acting on the particles at their current positions :

$$\mathbf{v}(t + \frac{1}{2}\Delta t) = \mathbf{v}(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}\,\mathbf{F}(t) \tag{21}$$

Then, the positions of the atoms are updated by a full time step Δt using the updated velocities :

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t\ \mathbf{v}(t + \frac{1}{2}\Delta t) \tag{22}$$

Lastly, the velocities are updated again by half a time step using the forces acting on the particles at their new positions :

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t + \frac{1}{2}\Delta t) + \frac{\Delta t}{2m}\,\mathbf{F}(t + \Delta t) \tag{23}$$

The velocities are updated in two different steps to achieve a second-order accuracy instead of a first-order accuracy using other methods such as the Euler algorithm. Trajectory files are generated as the simulation progresses, which is then analyzed to obtain information about the system behavior and properties.

4.1.2 Reproducibility and Reliability

Despite its increasing popularity, MD simulations raise questions about their reliability, reproducibility, and accuracy compared to laboratory experiment results. To ensure reproducibility of the simulations and reliability of the results, it is strongly encouraged to perform multiple replicas of the same simulation. Replicas start with the same initial parameters but are generally generated by assigning different random atomic initial velocities. In theory, simulations of the same system under the same conditions would converge to the same conformational ensemble [66]. As simulations can only run for a finite time, it's not rare to spot diverging outcomes [67].

The various trajectories observed can be explained by the roughness of the free energy landscape formed by multiple local minima separated by high energy barriers discussed in Section 3.1.1 . Complexes can get trapped in local minima for a considerable amount of time or drift away in solution space. Therefore, simulations of the same system can yield inconsistent results, and studying a single replica can lead to false positive or negative results. Positive results refer to findings that support or confirm the initial hypothesis, or expected outcome. On the other hand, negative results refer to findings that do not support or confirm the initial hypothesis, or expected outcome. Therefore, not considering multiple replicas can lead to false positive results that may appear significant in a single or a few very long simulations, but they are not reproducible across a large number of simulations. False negative results can also occur when many short simulations are performed, but they are too short to capture relevant motions or behavior of the system. Significant behavior or interactions that would be observable in longer simulations might not be detected. Such results emphasize the importance to carefully select the simulation timescale to align with the timescale of the process being studied. This ensures that the simulations capture the relevant dynamics and interactions accurately, providing reliable and meaningful results.

Nevertheless, performing multiple replicas is a computationally expensive task that requires a significant amount of computing resources. Consequently, for fixed computational resources it is important to find a balance between simulation length and number of replicas. It is suggested that performing multiple replicas over long single simulations is preferable [66] but there is no universal answer to this question, as it will depend on the specific question and the system being studied.

For example, if the goal is to investigate long-timescale processes, such as protein folding, it may be more advantageous to perform fewer longer simulations rather than multiple short ones. This is because a longer simulation provides more time for the system to explore the full conformational space, which may not be possible in shorter simulations.

Considering all those points and the goal of this research, multiple replicas for each simulation have been performed for this project.

4.2 Dynamical Network Analysis

Dynamical Network Analysis is a field of study that aims to understand the behavior of complex systems that evolve over time. In particular, Dynamical Network Analysis focuses on the dynamic interactions between entities that form networks. Unlike traditional network analysis, which often assumes static relationships between nodes, this approach models the evolution of relationships over time, capturing the dynamics of interactions. When this technique is applied to proteins, it uses trajectory files from MD simulations to determine the correlations in motion between residues at each time step, creating a network. This network can then be analyzed to identify communities, or groups of residues that have similar correlated motions over time. The theory behind Dynamical Network Analysis is discussed in this section.

4.2.1 Theory behind Dynamical Network Analysis

Dynamical Network Analysis uses trajectory files from MD simulations to provide a weighted (or non-weighted) network representation of the protein system. The analysis of the network allows the determination of interactions and correlations between residues in the protein complex. To understand the principle and the theory behind Dynamical Network Analysis, the main notions (node, edge, community, betweenness centrality of an edge) are clarified in this subsection. The detailed methodology of the Dynamical Network Analysis performed in this research project will be explained in Section 4.2.

Node

A node, also referred as a vertex, represents an individual element within a network, that can interact with other nodes in the network. In the context of MD simulations, a node usually represents an atom or an amino acid. In Dynamical Network Analysis, interactions between nodes are analyzed to understand the relationships and dynamics within the network.

Edge

An edge is a connection between two nodes that represents the interaction between them. In weighted networks, the edges have weights associated with them, indicating the strength of the connection between nodes. Some network analyses, such as the ones from [68] focus on analyzing the correlations between nodes that are in close proximity to each other. Therefore, the contacts between nodes in the network are determined based on the pairwise distances between them. Then, correlations between nodes considered in contacts (based on a cut-off distance) are calculated by computing the correlation matrix. To assign weights to the edges, correlation coefficients are derived from the correlation matrix. Those coefficients range from -1 to 1, where a value of 1 indicates a strong positive correlation, a value of -1 indicates a strong negative correlation, and a value close to 0 indicates that there is no correlation between the nodes.

Community

In a network, a community refers to a group of nodes more densely connected to each other than to the nodes outside the community. In the case of a weighted network, are defined using algorithms such as the Louvain heuristics [69]. One essential notion to define communities is the modularity. The modularity Q , is a scalar value between -1 and 1 that quantifies the density of connections inside communities as compared to connections between communities. In weighted networks, it is defined as :

$$Q = \frac{1}{2m} \sum_{i,j} [A_{i,j} - \frac{k_i k_j}{2m}]\ \delta(c_i, c_j) \tag{24}$$

where $A_{i,j}$ is the weight of the edge between node i and node j, $k_i = \sum_j A_{i,j}$ is the sum of the weights of the edges attached to i, c_i represents the community to which i is assigned, and $\delta(c_i, c_j)$ is the Kronecker function.

Initially, each node in the network is assigned to its own community. Then, for each node, the Louvain heuristics algorithm evaluates the change in modularity ΔQ that would occur if the node i were to be moved to the community of one of its neighboring nodes, j. The change in modularity ΔQ is described as :

$$\Delta Q = [\frac{\sum_{in} + 2k_{i,in}}{2m} - (\frac{\sum_{tot} + k_i}{2m})^2] - [\frac{\sum_{in}}{2m} - (\frac{\sum_{tot}}{2m})^2 - (\frac{k_i}{2m})^2] \tag{25}$$

where $\sum_{in}$ is the sum of the weights of the connections inside the community C, $\sum_{tot}$ is the sum of weights of the connections linked to nodes in C, k_i is the sum of weights of the connections links to i, $k_{i,in}$ is the sum of weights of the connections from i to nodes in C and m is the total weight of all connections in the network, regardless of the community assignment. The Louvain heuristics iterates over all the nodes in the network and considers each possibility. Therefore, the change in modularity is calculated for each possibility. The node i is moved to the neighboring community that results in the maximum increase in modularity. If there is no increase possible, the node remains in its current community. Once all nodes have been considered, the algorithm creates a new network where each community is represented as a single node. The process continues until no further improvement can be made. Finally, the algorithm outputs the network partition with the highest modularity.

Betweenness centrality of an edge

The betweenness centrality of an edge $c_B(e)$ measures how often the edge lies on the shortest path between pairs of nodes in a network. It represents the importance of an edge in enabling communication between nodes. Edges with high betweenness are important for maintaining the connectivity and efficiency of the network. They can be obtained using algorithms such as the one from Ulrik Brandes [70], in which the betweenness centrality of an edge is considered as the sum of the fraction of all-pairs shortest paths. The shortest path is the path that minimizes the cumulative weight while traversing from the source node to the destination node in the network. It is defined

using algorithms such as the Dijkstra algorithm [71] which considers the weights assigned to the edges in the network and calculates the path with the minimum total cost.
The betweenness centrality of an edge is defined as :

$$c_B(e) = \sum_{i,j \in V} \frac{\sigma(i,j|e)}{\sigma(i,j)} \tag{26}$$

where V is the set of nodes, $\sigma(i,j)$ is the number of shortest paths, and $\sigma(i,j|e)$ is the number of those paths passing through the edge e.

4.3 Methodology

To answer the question raised in this research project, Molecular Dynamics simulations have been employed to study the dynamics of the WSCP from *Lepidium virginicum*. Once the trajectory files were collected, multiple analyses were performed to investigate the dynamics of the protein complex. Then, Dynamical Network Analysis was performed to study the motion correlations among specific protein residues. The detailed methodology is explained in this section.

4.3.1 Molecular Dynamics Simulations

Molecular dynamics simulations were carried out using Gromacs 2021.4 software and performed on Compute Canada Narval cluster. The applied CHARMM36 force field (july 2021 version) is taken from the Mackerell lab website [72]. The initial crystal structure of class II-B WSCP from *Lepidium virginicum* is obtained from the Protein Data Bank. The protein complex is a homotetramer with each monomer consisting of 180 residues (the protein sequence is displaying in the Appendix A, Figure A.1). The natural Chl a/b ratio of this protein complex is around 1.6-1.9. However, in this specific crystal structure (PDB ID : 2DRE), only chlorophylls a constitute the four pigments located in the center of the protein complex. Although WSCP is not directly involved in photosynthetic chain reactions, it is still an interesting subject of study. In fact, unlike larger PPCs involved in those chain reactions (Section 2.2), and therefore embedded in a thylakoid membrane, WSCP is simpler and easier to study through MD simulations due to its smaller size and water-based environment. This computational cost reduction allows for longer simulation times, making WSCP an ideal model system for initial exploration of the dynamics and interactions between proteins and chlorophylls through MD simulations. In addition, it is worth noting that PS I, Cytochrome b_6f, and PS II each contain several ligands for which no CHARMM36 force field parameters have been found. This complicates the accurate simulation of these complexes. Detailed information on WSCP and its properties were discussed in Section 2.3.

Protein missing residues were built using Modeller program, considering the structure as a homotetramer. The protein chirality and the cis[5] / trans[6] configuration of the peptide bonds were verified using the VMD 1.9.4a43 plugins Chirality 1.4 and Cispeptide 1.4. Chlorophyll a parameters were added to the force field and provided by [73] [74] [75] [76] [77] [78]. Since we are assuming a pH of 7, acidic amino acids (ASP, GLU) are kept deprotonated and basic amino acids (LYS, ARG, and HIS) are kept protonated [31]. An estimation of the pKa values of these residues were checked using the PropKa website [7], and the results are shown in the Appendix A, Figure A.2. Four residues, including GLU 125 in protein chains B and D, and HIS 43 in protein chains B and D, have a pKa value slightly different from the protonation state asssigned to them. The pKa prediction of the other residues matches with the protonation state assigned to them. The deprotonated acidic amino acids and protonated basic amino acids are highlighted in the protein crystal structure shown in the Appendix A, Figure A.3. The resulting negatively charged system was neutralized by replacing 40 random solvent molecules with 40 sodium ions. The system is energy-minimized using the steepest descent algorithm. The protein complex was constrained in a rhombic dodecahedron box with an image distance $d = 11.757\ nm^3$, and filled with TIP3P-CHARMM water molecule model. Periodic boundary conditions were used to ensure a continuous representation of the system, and avoid artificial interactions caused by the finite size of the simulation box.

The objective of this project is to compare the experimentally obtained FEL at cryogenic temperatures with the FEL obtained from simulations. To achieve this, simulations were conducted at two different temperatures: one at room temperature, T_1 = 300 K, and the other at a lower temperature,

[5] Cis configuration refers to the arrangement of two substituents on a molecule with similar groups on the same side of the molecule.

[6] Trans configuration refers to the arrangement of two substituents on a molecule with similar groups on opposite sides of the molecule.

[7] https://www.ddl.unimi.it/vegaol/propka.htm

T_2 = 165 K. It should be noted that T_2 cannot be decreased further, as the melting temperature for the TIP3P-CHARMM water molecule model used in the simulations is $T_{melting}$ = 146 K [79]. The initial velocities of the particles were randomly assigned from a Maxwell-Boltzmann distribution at the given temperatures. For both simulations, equilibration steps (NVT and NPT) ran for 1000 ps each to reach respectively the target temperature and the target pressure. For NVT equilibration, the Bussi-Donadio-Parrinello thermostat [63] was used (referred as the v-scale thermostat in Gromacs). For NPT equilibration, the Berendsen barostat [64] was used and the target pressure was set to 1 bar. To ensure reproducibility and reliability, 5 replicas for each simulation were performed assigning different random initial velocities of the particles for all replicas. The equations of motion were integrated using the leapfrog algorithm. In the case of WSCP, hydrogen bond vibrations are one of the smallest fluctuations in the system, occurring on the timescale of 1 fs. Therefore, the hydrogen bonds were constrained to allow a larger time step of 2 fs to be used without losing significant dynamical information. This choice significantly reduces the computational cost of the simulations while still capturing the essential dynamics of the protein. Each replica was simulated for 1 µs, and coordinates and energies were saved every 10 ps, resulting in 100 000 frames per simulation.

4.3.2 Modeling and Analysis

All visualization tasks were performed using VMD 1.9.4a43. First, the project focused on the vicinity of the chlorophylls a in a protein. This close environment was defined by determining the pigments contact maps with the Contact Map Explorer Python 3.8.10 package. The contact maps allowed for the identification of the protein residues that are in close proximity to the pigment molecules. The all-atom cut-off distance was set to 0.3 nm, which refers to the maximum distance at which two atoms in different molecules (the Chl a and the protein residues) are considered to be in contact with each other. Hydrogen bonds between 2.5-2.9 Å long are known to likely display a double-well potential, which means that the hydrogen atom can switch back and forth between two positions. This phenomenon is related to the hydrogen bond strength, and it has been observed in some experimental and computational studies [80]. Because we were looking for double-well potentials, hydrogen bonds between the chlorophylls and the residues were calculated with the Baker-Hubbard Hydrogen Bond identification function from the Python 3.8.10 library MdTraj 1.9.6. Only hydrogen bonds between 2.5-2.9 Å long with an α angle Donor-Hydrogen-Acceptor greater than 120° are kept for further investigation. Then, analysis calculations to study the protein dynamics, such as the Root-Mean-Square Deviation (RMSD), the Root-Mean-Square Fluctuation (RMSF) and the side chain dihedral angle χ_1, were computed using Gromacs 2021.4 tools (respectively *gmx rms*, *gmx rmsf*, and *gmx angle -type dihedral*).

4.3.3 Dynamical Network Analysis

After having identified some protein residues located near the chlorophylls a that could be responsible for the observed spectral dynamics, the next step involved performing Dynamical Network Analysis of MD simulations to gain further insights. This analysis served two main purposes. Firstly, it was used to determine the correlation in motion between the identified residues, to see if they undergo conformational changes simultaneously or not. Secondly, it aimed to extend the study to the global protein environment by identifying other protein residues that are correlated to the candidates but located farther from the chlorophylls. All network analyses were conducted using Python 3.8.10 libraries Dynetan 1.2.0 and Networkx 2.8.7. The code was obtained from [68]. Each residue in the protein is represented by one node corresponding to its α-carbon atom, while each chlorophyll a is represented by a node corresponding to its Mg atom. The nodes are classified into communities using the Louvain heuristics [69]. The correlation of motion between all the nodes is calculated, and the resulting network can be visualized and analyzed. The trajectory files obtained from the MD simulations contained 100,000 frames each. However, to decrease computational expenses, only half of these frames were used in the network analysis by skipping every other frame, resulting in the use of 50,000 frames per simulation. Due to limiting computational resources, four independent network analyses were performed on the protein complex (one per protein sub-unit) for each replica and for both MD simulations.

Chapter 5

Results

While previous studies have reported simulations that explicitly capture multi-well protein energy landscapes, none have been conducted specifically for proteins involved in photosynthesis [81] [82] [83]. Therefore, the objective of this research project was to conduct all-atom Molecular Dynamics simulations of the Water-Soluble Chlorophyll-a binding protein to identify the structural elements responsible for the observed spectral shifts in protein complexes involved in photosynthesis [84].

Specifically, we performed two MD simulations at temperatures T_1 = 300 K and T_2 = 165 K, following the methodology described in Section 4.3. Firstly, we examined the local environment surrounding the Chlorophyll a pigments, which revealed side chain rotational motions of specific residues within the vicinity of Chls a. This discovery led us to map the multi-well protein free energy landscape associated with this generalized coordinate, exposing temperature-dependent energy barrier heights. Secondly, we employed Dynamical Network Analysis to explore correlations between these residue candidates, revealing that their motions are correlated based on inter-residue distances, with higher correlations observed among nearby residues. Additionally, while examining the dynamics of the entire protein complex using dynamical network analysis, we observed a residue located outside the hydrophobic cavity also undergoing side chain dihedral angle rotations. These findings suggest that the side chain rotational motions of specific residues within the environment of the pigments, and potentially beyond, contribute to the observed spectral shifts in pigment-protein complexes involved in photosynthesis.

5.1 Identification of local environmental contributors to observe Two-Level System signal

We investigated the local environment of the pigments, defined with contact map calculations, to perform a series of analyses on these residues. On the one hand, specific hydrogen bonds between pigments and Chlorophylls a were identified, but we found no evidence that the rearrangements of hydrogen bonds were responsible for the Two-Level System signature. On the other hand, structural analyses, including Root-mean-square deviation (RMSD) and Root-mean-square fluctuation (RMSF), were performed on the residues comprising the local environment. They revealed multimodal signatures in the positional distributions of the side chains of certain residues, manifested as a rotation of the first side chain dihedral angle. As this rotation constitutes a small conformational change, it could be responsible for the experimental observations. Furthermore, the free energy landscape was mapped by considering this dihedral angle as the generalized coordinate, allowing a comparison of energy barrier heights with experimental data.

5.1.1 Definition of local environment via contact between pigments and protein subunits

The studied WSCP complex is a large protein comprising a total of 720 residues. We identified the local protein environment in the neighborhood of the pigment to investigate possible direct interactions responsible for experimental signatures [85]. Consequently, we computed the contacts between the pigments and the residues, allowing us to detect all residues located in the proximity of the pigments and determine the frequency of these contacts. The Chlorophyll a contact occurrence plots were computed following the methodology described in Section 4.3.2. To ensure the inclusion of closely associated residues, an all-atom cut-off distance of 3.0 Å was set between the Chlorophyll a and the residues. As the exact definition of “close environment” was somewhat arbitrary defined, the choice of the cut-off distance was made based on the approximate expected hydrogen bond lengths (Section 5.1.2). By utilizing this cut-off distance, the scale of the contact occurrence plots remained consistent with that of the hydrogen bonds.

Based on the analysis of these contacts, we were able to identify 42 residues per protein chain, totaling 168 residues nearby the Chls a out of the original 720. Furthermore, these residues were categorized based on the type of their side chains into four distinct categories: polar, non-polar, basic, and acidic. This classification provided us with a comprehensive overview of these residues and their probable functions. In the rest of our project, these 168 close residues will be considered for further analysis as the potential local environment might be responsible for the conformational change we investigated.

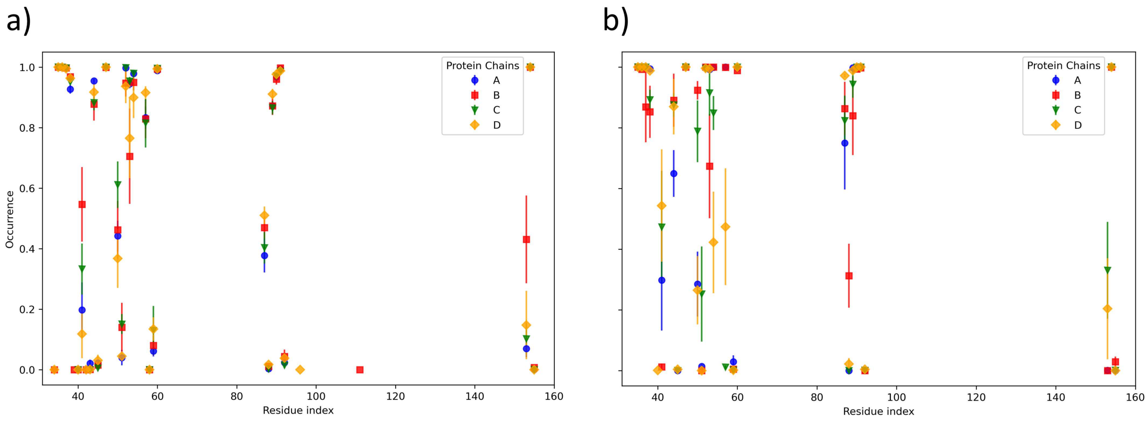


Figure 5.1: Contact occurrence plots between Chlorophyll a and its protein chain, for each monomer of WSCP, at *a)* 300 K, and *b)* 165 K. Error bars represent standard errors across five independent runs.

The contact occurrence plots, depicted in Figure 5.1, show the interactions between Chlorophyll a and its corresponding protein sub-unit, for each monomer of WSCP. The heatmaps of the average occurrences and their associated standard errors at high and low temperatures are displayed in the Appendix B (Figure B.1 and Figure B.2). The results for the contact between each pigment and its three respective distal protein sub-units are not presented because each pigment molecule primarily interacts only with the single most proximal sub-unit. Originally, all residues with an occurrence of a least 0.4 were kept and considered as part of the ligand close environment. Additionally, the reference paper of this crystal structure [23] listed all residues forming the hydrophobic cavity in which Chls a are enclosed in. We realized that eleven of those residues per protein chain (VAL39, ASP40, CYS45, PRO46, PRO55, TYR56, ASP86, SER95, LYS96, SER152, and TPR154), had an occurrence less than 0.4, but we still decided to consider them as potential relevant residues, as their function and dynamics might be related to the presence of Chls a. These residues, despite their lower occurrence, are also part of the hydrophobic cavity and therefore share the same fundamental function as residues with higher occurrence: preventing water molecules from interacting with Chlorophylls a. This function and dynamics could be affecting the presence of conformational change. Additionally, the cut-off distance of 3 Å was set arbitrary and might have been underestimated. Consequently, for the investigation of the nearby Chls a environment at 300 K and 165 K, protein residues 35 to 60, 86 to 96, and 152 to 156 for each protein chain were retained, as seen in Figure 5.2 . This selection criterion significantly reduces the initially large system comprising a total of 720 residues to a smaller domain of 168 residues (42 residues per protein chain).

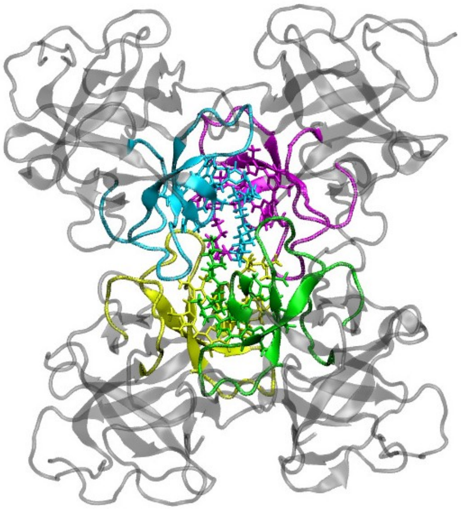


Figure 5.2: Graphical representation of the studied WSCP complex. Far residues are represented in grey, Chls a and their respective 42 close residues are respectively colored in cyan, magenta, yellow and green, for the four monomers of WSCP.

Figure 5.3 displays the classification of the 168 residues based on the types of their side chains. The analysis reveals that 66.7% of the residues are non-polar, 21.4% are polar, while 7.1% are basic, and only 4.8% are acidic. The prevalence of non-polar residues is in line with expectations, as they are hydrophobic residues that contribute to the formation of hydrophobic cavities around Chls a. Conversely, the small percentages of basic and acidic residues are also as anticipated, as these types of residues are hydrophilic and their respective positively and negatively charged side

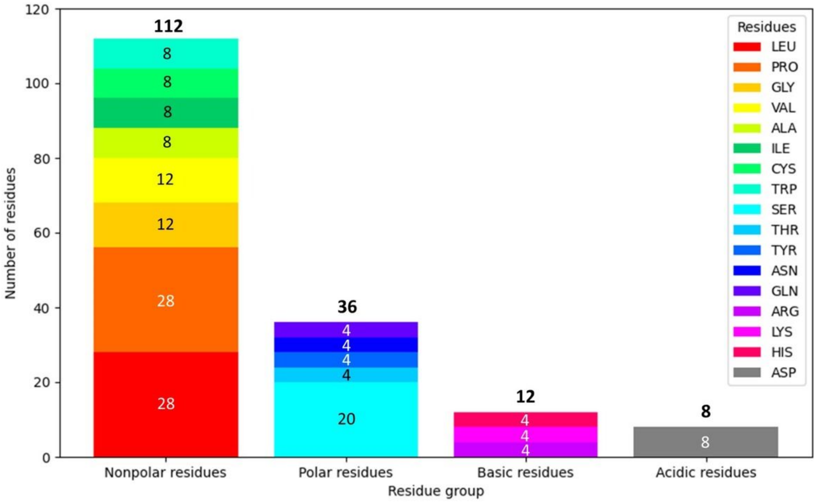


Figure 5.3: Classification of the 168 residues per residue group.

chains are typically engaged in hydrogen bonding with water molecules, therefore, attracting them. We can see that among the 168 residues, the most abundant residue types are the hydrophobic leucine (LEU), with 28 residues and the hydrophobic proline (PRO) with 28 residues as well. It indicates a high contribution of LEU and PRO in the stabilization of the hydrophobic environment and a potential role in facilitating specific protein interactions within that region.

5.1.2 No evidence of multimodal signature in long hydrogen bond configurations

Hydrogen bonds are intermolecular forces that occur when a hydrogen atom, bonded to an electronegative atom like oxygen, nitrogen, or fluorine, attracts another electronegative atom nearby. These bonds are weaker than covalent bonds but stronger than other intermolecular forces. It is well-known that hydrogen bonds within the range of 2.5-2.9 Å exhibit double-well potentials [80]. However, these particular double-well potentials are only accessible via quantum mechanical calculations and not classical MD simulations, thus we cannot find associated energy landscapes or determine the energy barriers. Nevertheless, we still explored hydrogen bonds looking for situations where double-well potentials could be associated with rearrangements of hydrogen bonds. In fact, we considered the possibility that a conformational change, resulting in two distinct positions of the hydrogen atom in relation to the donor and acceptor atoms, might explain the experimentally-observed spectral dynamics. To investigate this, we computed the distance distributions between these atoms for each hydrogen bond. However, our findings did not provide any evidence supporting the hypothesis that these hydrogen bonds were responsible for the observed conformational change.

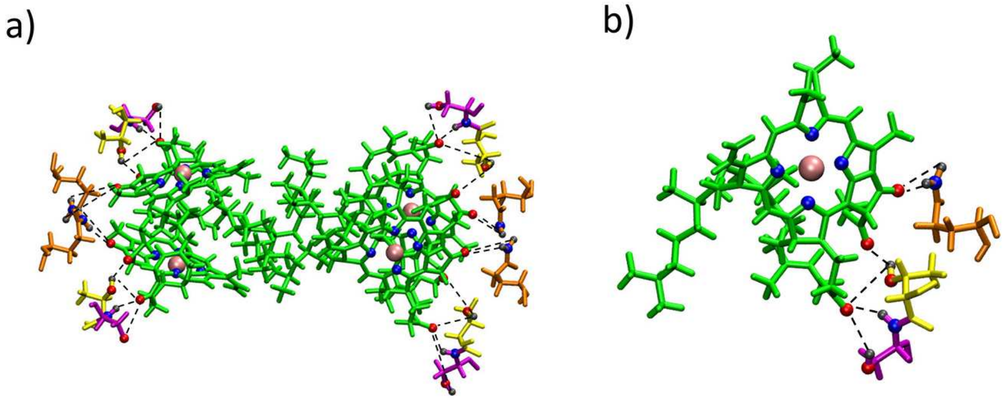


Figure 5.4: Representations of *a)* all Chlorophylls a (green) interacting with their Threonine 52 (yellow), Serine 53 (magenta), and Glutamine 57 (orange). In *b)*, a closer representation is depicted for only one protein sub-unit. Oxygen atoms involved in the H-bonds are colored in red, hydrogens in black, and nitrogens in blue. The central Mg atom of each Chl a is colored in pink and their nitrogen atoms are also shown in blue. All possible H-bonds are marked by black dashed lines.

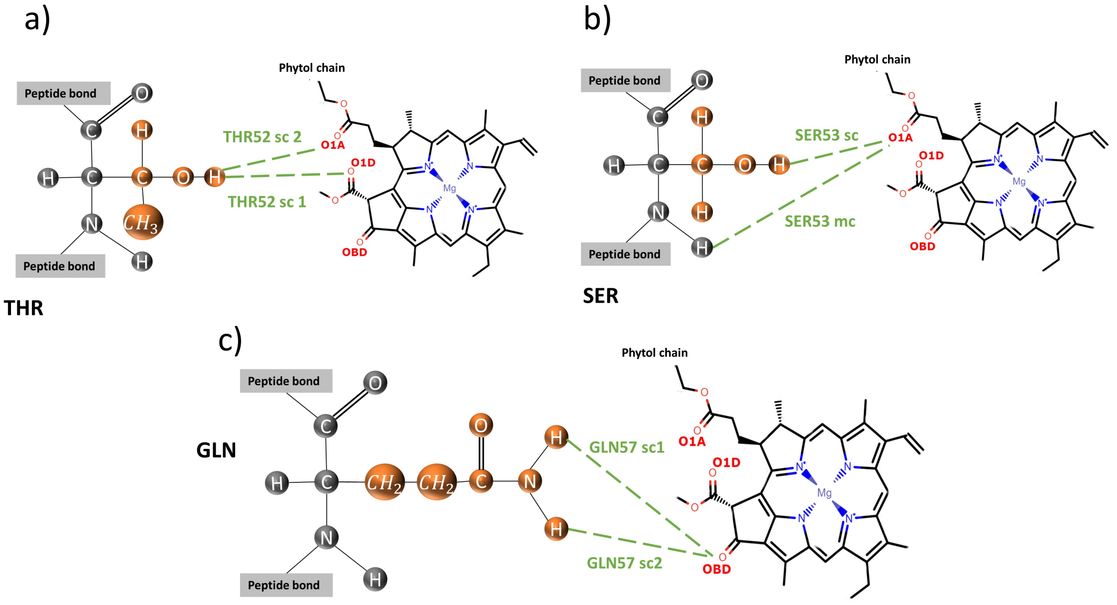


Figure 5.5: Schematic representation of hydrogen bonds (green) between *a)* Threonine 52 and Chlorophyll a, *b)* Serine 53 and Chlorophyll a, and *c)* Glutamine 57 and Chlorophyll a. The backbones of the amino acids are represented in grey, and the side chains in orange.
The figure is adapted from [86].

Hydrogen bonds were computed following the methodology outlined in Section 4.3.2, and only H-bonds between 2.5 and 2.9 Å long were considered. The results revealed that each Chlorophyll a established hydrogen bonds with three specific residues, namely Threonine 52 (THR52), Serine 53 (SER53), and Glutamine 57 (GLN57), as illustrated in Figure 5.4. For all these H-bonds, Chlorophyll a functions as the acceptor, leaving protein residues to serve as the donors. The central magnesium atom in Chlorophyll a does not participate in hydrogen bonding due to its relatively low electronegativity, which results in a weaker attraction for electrons compared to highly electronegative atoms. Each residue establishes two H-bonds with its corresponding pigment, resulting in a total of six distinct hydrogen bonds formed per pigment, as illustrated in Figure 5.5. Specifically, Chl a forms two H-bonds with the hydroxyl side chain group of THR52, respectively referred to as THR52 sc1 and THR52 sc2, where sc stands for side chain. Additionally, Chl a forms one H-bond with the amine backbone group of SER53, denoted as SER53 mc (where mc stands for main chain), and one H-bond with its hydroxyl side chain group, denoted as SER53 sc. Finally, Chl a also forms two H-bonds with the hydrogens in the amide side chain group of GLN57, respectively labeled as

GLN57 sc 1 and GLN57 sc 2. The oxygen atoms of Chl a involved in these H-bonds possess two lone pairs of electrons, enabling them to form two H-bonds simultaneously. As a result, SER53 mc and SER53 sc1 can coexist, as can GLN57 sc1 and GLN57 sc2. However, since the hydrogen atom within the residues only has one electron, it can only form a single H-bond at a time. Consequently, THR52 sc1 and THR52 sc2 cannot coexist.

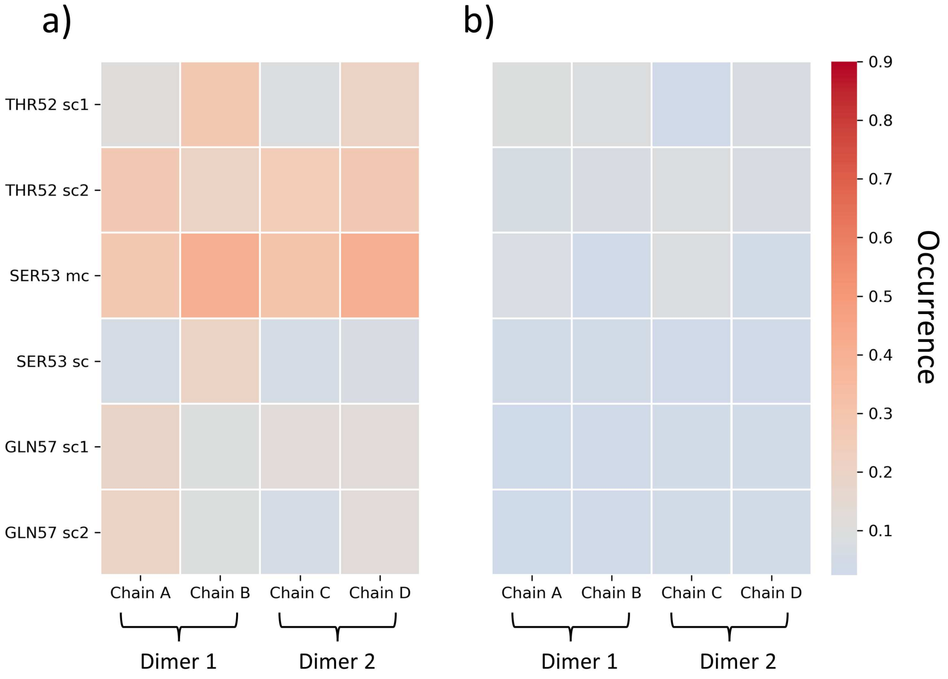


Figure 5.6: Heatmap of the *a)* Mean and *b)* Standard errors of the hydrogen bonds occurrence taken over 5 independent replicas at 300 K.

The occurrence maps of the hydrogen bonds (Figure 5.6 and Figure 5.7) were computed for each protein chain in both simulations conducted at 300 K and 165 K. The data were obtained from five replicas for each simulation and subsequently averaged. The purpose of these maps was to determine the frequency at which hydrogen bonds occur, as transient hydrogen bonds may indicate potential dynamic behavior in the system. At 300 K, we observed that the majority of hydrogen bonds occurred between 10% and 60% of the simulation time, and the occurrence patterns were relatively consistent across the protein chains. However, at 165 K, the results were more varied. Some H-bonds (e.g., THR52 sc2, SER53 sc, and GLN57 sc2) were nearly nonexistent with less than 10%

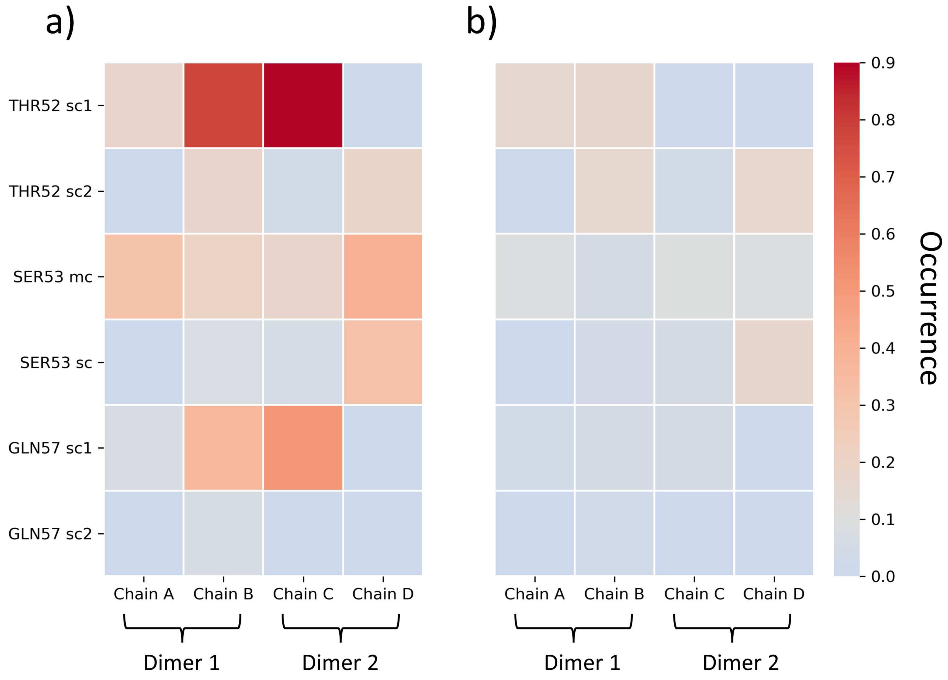


Figure 5.7: Heatmaps of the *a)* Mean and *b)* Standard errors of the hydrogen bonds occurrence taken over 5 independent replicas at 165 K.

occurrence, while others (e.g., THR52 sc1 and GLN57 sc1 in protein chains B and C) were relatively frequent with over 60% occurrence. The temperature-dependent differences in the occurrence of H-bonds can be explained by the reduction in the number of relevant degrees of freedom at 165 K. The protein requires more simulation time to move away from its initial configuration, which affects the occurrence and dynamics of hydrogen bonds. Additionally, there was heterogeneity observed between the protein chains in terms of the occurrence patterns. At 165 K, GLN57 sc 1 in both chain B and chain C is formed around 60% of the time but less than 10% in both chain A and chain D. This variation in frequency aligns with the overall symmetry of the complex. Specifically, the WSCP complex can be viewed as a dimer of dimers, where chain B and chain C represent the same monomer within each dimer, and the same applies to chain A and chain D.

Table 5.1: Average distance between the hydrogen atom and the acceptor atom for six hydrogen bonds, at 300 K. Distances are averaged over five replicas for each protein chain. Standard error of the means are displayed on the right side of the columns.

300 K				
Hydrogen Bonds	Average distance Hydrogen-Acceptor (nm)			
	Chain A	Chain B	Chain C	Chain D
THR52 sc1	0.39 ± 0.04	0.35 ± 0.05	0.43 ± 0.05	0.41 ± 0.05
THR52 sc2	0.40 ± 0.05	0.34 ± 0.06	0.46 ± 0.05	0.42 ± 0.06
SER53 mc	0.29 ± 0.05	0.22 ± 0.03	0.28 ± 0.05	0.22 ± 0.02
SER53 sc	0.42 ± 0.06	0.36 ± 0.05	0.44 ± 0.06	0.38 ± 0.04
GLN57 sc1	0.28 ± 0.05	0.34 ± 0.06	0.30 ± 0.04	0.34 ± 0.05
GLN57 sc 2	0.39 ± 0.03	0.40 ± 0.04	0.41 ± 0.04	0.42 ± 0.06

Table 5.2: Average distance over all frames of the simulations between the hydrogen atom and the acceptor atom (nm) for six hydrogen bonds, at 165 K. Distances are averaged over five replicas for each protein chain. Standard error of the means are displayed on the right side of the columns.

165 K				
Hydrogen Bonds	Average distance Hydrogen-Acceptor (nm)			
	Chain A	Chain B	Chain C	Chain D
THR52 sc1	0.34 ± 0.01	0.33 ± 0.03	0.34 ± 0.01	0.37 ± 0.05
THR52 sc2	0.30 ± 0.02	0.21 ± 0.02	0.18 ± 0.01	0.39 ± 0.04
SER53 mc	0.22 ± 0.01	0.22 ± 0.01	0.25 ± 0.02	0.23 ± 0.03
SER53 sc	0.49 ± 0.02	0.34 ± 0.03	0.40 ± 0.04	0.30 ± 0.05
GLN57 sc1	0.52 ± 0.06	0.22 ± 0.02	0.20 ± 0.01	0.71 ± 0.02
GLN57 sc 2	0.55 ± 0.06	0.35 ± 0.01	0.35 ± 0.01	0.71 ± 0.02

The average distances between the hydrogen atom and the acceptor atom involved in each hydrogen bond were computed over all frames of the simulations and are presented in Table 5.1 and Table 5.2. Observing these tables, we can notice that, in most cases, the average distance exceeds our cut-off distance of 2.9 Å. This discrepancy arises because the averages were calculated over all frames of the simulations, including those where hydrogen bonds are not formed. The larger average distance implies that when the hydrogen bond is not formed, the hydrogen and acceptor atoms are considerably separated, indicating a lack of interaction between them. This observation is further supported by our additional analysis, which revealed unimodal distance distributions for all hydrogen bonds at both high and low temperatures. Therefore, these results suggest that there is no evidence of double-well potentials associated with the hydrogen bonds in our simulations. However, it is important to acknowledge the limitations of our approach. Our MD simulations rely on classical mechanics and employ the classical CHARMM36 force field to approximate the interactions between atoms, including hydrogen bonding. Classical force fields may not fully account for the quantum mechanical effects involved in the formation and dynamics of hydrogen bonds. Quantum effects, such as proton tunneling or delocalization, are not explicitly considered in these classical MD simulations. Considering the limitations of classical force fields, more detailed results could be obtained by employing quantum mechanical calculations, such as Density Functional Theory (DFT) [55]. This method takes into account the quantum effects associated with hydrogen bonding and may provide additional insights into the presence of double-well potentials. In summary, this occurrence H-bonds analysis do not provide sufficient evidence to conclude the presence of a double-well potential associated with those hydrogen bonds.

5.1.3 Identification of side chain dihedral as microscopic collective variable for free energy landscape construction

This section focuses on investigating various structural aspects of the reduced set of 168 residues comprising the close environment of Chls a. This includes the evaluation of Root-mean-square deviation (RMSD), Root-mean-square fluctuation (RMSF), and dihedral angle calculations. These analyses aimed to identify any significant conformational change in the protein. Through them, a generalized coordinate was identified : the rotation of the first side chain dihedral angle of certain amino acids. The free energy landscape associated with the generalized coordinate was characterized, and a comparison was made between the energy barrier heights obtained from the mapping and experimental data. At 300 K, the energy barrier heights were found to fall within a similar range of 1 000 cm^{-1} as the experimental results. However, a discrepancy was observed at 165 K, where the energy barrier heights were smaller than anticipated. This temperature-dependent energy barrier heights evidence is significant because in models used to simulate spectroscopy results [84], it was implicitly assumed that the barriers are temperature-independent, and only the rates change with temperature.

Evidence of bimodal distributions in root-mean-square deviation

The Root-mean-square deviation (RMSD) of the 168 residues were calculated (Figure 5.8) to quantify the positional changes of these residues, from which potential multimodal signatures can be identified. The RMSD measures the deviation of atoms in a molecule relative to a reference frame. For calculations, each frame of the simulation is fitted to a reference structure of the reference frame. In Gromacs 2021.4, the RMSD is defined as [87] :

$$RMSD\ (t_{ref}, t_2) = [\ \frac{1}{M} \sum_{i=1}^{N} m_i \parallel \mathbf{r}_i(t_{ref}) - \mathbf{r}_i(t_2) \parallel^2\ ]^{\frac{1}{2}} \qquad (27)$$

where $M = \sum_{i=1}^{N} m_i$, the mass of all particles in a molecule, $\mathbf{r}_i(t_{ref})$ is the position of a particle i at the reference time t_{ref}, and $\mathbf{r}_i(t_2)$ is the position of a particle i at time t_2.

The RMSD is averaged over the particles, providing time-specific values. In our project, the 168 residues were treated as separate molecules, and all atoms within these residues were considered for

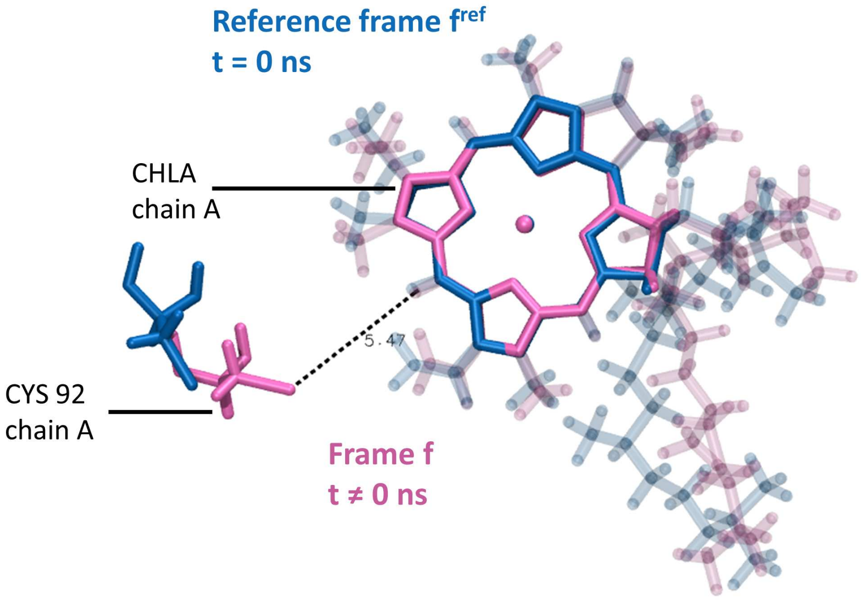


Figure 5.8: Graphical representation of the RMSD calculation of residue CYS 92 in chain A. The reference frame, f^{ref}, is colored in blue, and the frame used for calculation, f, is in pink. The fit is made on the Chl a phrophyrin ring allowing calculation of the deviation in the residue position with respect to Chl a phrophyrin ring. The figure was rendered using VMD 1.9.4a43.

RMSD calculations. The reference frame corresponds to the first frame of the simulation, and the reference structure is the phrophyrin ring of the respective Chl a. This fitting approach was chosen to assess the deviation in position of each residue relative to its corresponding pigment. Among the 168 residues comprising the local environment, 55 exhibited bimodal RMSD distributions corresponding to significant variation in their RMSD values. This variation implies positional changes of these residues throughout the simulations according to a multi-well free energy landscape. Figure 5.9 visually represents the arrangement of these 55 residues within the protein complex. The subsequent paragraphs will present some of these bimodal distributions and will provide a detailed explanation of their characteristics such as the position of the peaks and the distance between them to quantify the range of the positional change. Additionally, the specificities of these 55 residues regarding their type, their residue group and their distance to their corresponding Chl a will also be analyzed to determine if these specificities influenced the shift in the RMSD values.

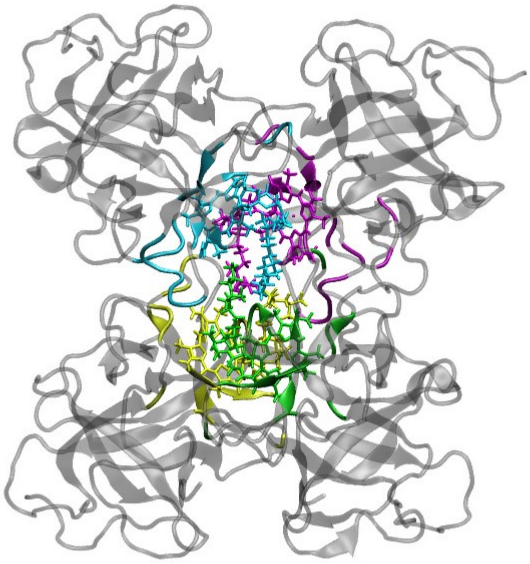


Figure 5.9: Graphical representation of the studied WSCP complex. Far residues are represented in grey, Chls a and their respective residues that show a bimodal RMSD distribution are respectively colored in cyan, magenta, yellow and green, for the four monomers of WSCP.

In Figure 5.10, representative bimodal RMSD distributions are displayed for VAL 50 in chain D and LEU 47 in chain A. The presence of two observable peaks in these distributions suggests the existence of at least two distinct preferred atomic position arrangements. These peaks suggest distinct and energetically more favorable states that the residues can adopt, indicating potential conformational flexibility in these 55 residues. As we can see, not all replicas yield identical results for the same residue. Specifically, at 165 K, replicas #3 and #4 for VAL 50 in chain D exhibit unimodal distributions instead of the expected double-peak distribution. Similarly, at 300 K, replica #3 for LEU 47 in chain A also displays a unimodal distribution. These findings can be attributed to the fact that a simulation time of 1 μs might not be sufficient for convergence between replicas to occur. As explained in Section 4.1.1, replicas begin with different initial particle velocities and require time to explore the entire phase space. In these replicas, certain residues became trapped in a single state, and overcoming these barriers may necessitate a longer simulation time, particularly at lower temperatures. Moreover, Figure 5.10*b)* demonstrates that at 165 K, LEU 47 in chain A (along with many other residues among the 55 identified), does not exhibit a shift in its RMSD distribution,

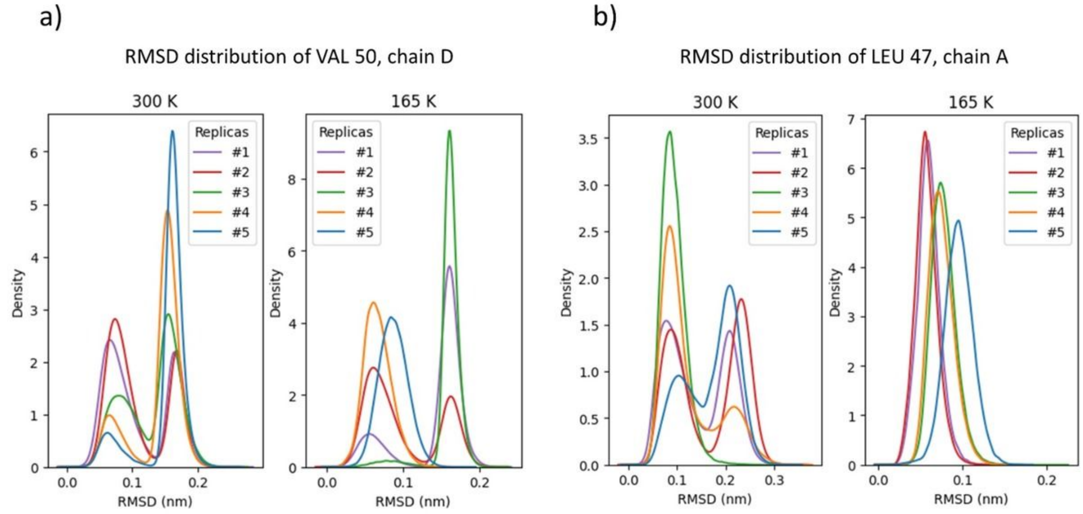


Figure 5.10: RMSD distributions of *a)* VAL 50 in protein chain D, and *b)* LEU 47 in protein chain A, at 300 K and 165 K, for all five replicas.

resulting in a unimodal distribution. However, at 300 K, the same residues depict a double-peak RMSD distribution. This observation can be explained by the reduced number of important degrees of freedom at lower temperatures. In fact, at lower temperatures, the system mobility decreases, requiring longer simulation time to explore the entire phase space.

Table 5.3 presents the average characteristics of the RMSD distributions for the residues, considering only the replicas that exhibit a double-peak distribution. The small standard error values for all characteristics at both temperatures demonstrate an agreement among the replicas, indicating that they have captured the same phenomenon for the residues. Comparing results obtained at 300 K and 165 K, there is a clear shift to lower values at 165K, which makes sense, because there is less available thermal energy. At both temperatures, the average first peak position is below 0.1 nm, indicating a stable initial position [88]. At both temperatures, the second peak position is above 0.2 nm, indicating an apparent shift from the initial position. [88]. Finally, the average distance between the peaks is 0.153 ± 0.006 nm at 300 K, and 0.136 ± 0.004 nm at 165 K, suggesting that the shift is similar for both temperatures. Therefore, it can be concluded that the same structural change is taking place regardless of the temperature and the residue type.

Table 5.3: Characteristics of Bimodal RMSD Distributions. Results are averaged over the 55 residues, at 300 K and 165 K. Standard error of the means taken over 5 independent replicas are displayed on the right side of the columns.

Characteristics of Bimodal RMSD Distributions.		
	300 K	165 K
First peak position (nm)	0.097 ± 0.006	0.073 ± 0.003
Second peak position (nm)	0.250 ± 0.010	0.209 ± 0.008
Distance between peaks (nm)	0.153 ± 0.006	0.136 ± 0.004

The 55 residues were categorized by their side chain category, as shown in Figure 5.11. Comparing this classification to the classification of the initial 168 residues (Figure 5.3 in Section 5.1.1), we observed that the proportion of residues per category is preserved. Among the 55 residues, 65.5% are non-polar, which is similar ratio considering all 168 residues where 66.7% of them were non-polar. Additionally, 25.5% of the 55 residues are polar, compared to 21.4% in the overall 168 residues classification. Furthermore, 9.0% of this reduced set of residues have a basic side chain, while the overall proportion is 7.1%. Notably, none of the original 8 aspartic acid residues displayed a double-peak RMSD distribution, resulting in no acidic residues among the selected 55 residues. Therefore, the main difference observed is the absence of acidic residues. From this classification, it seems that the presence of bimodal RMSD distributions does not appear to be influenced by the side chain group of the residue.

In terms of residue types, the proportions are not necessarily conserved among the reduced set of 55 residues. Among the original 168 close residues, LEU and PRO were the most abundant, each representing 16.7% of the total residues. However, in the subset of residues displaying bimodal RMSD distributions, LEU remains the most abundant residue and now constitutes 34.6% of the total, while PRO accounts for only 7.3%. On the other hand, SER, which previously represented 11.9% of the total residues, now comprises only 5.5%, becoming one of the least present residue. In contrast, VAL, which initially accounted for only 7.1% of the residues, now represents 12.7%. This suggests that LEU and VAL residues play a significant role in this conformational change. It

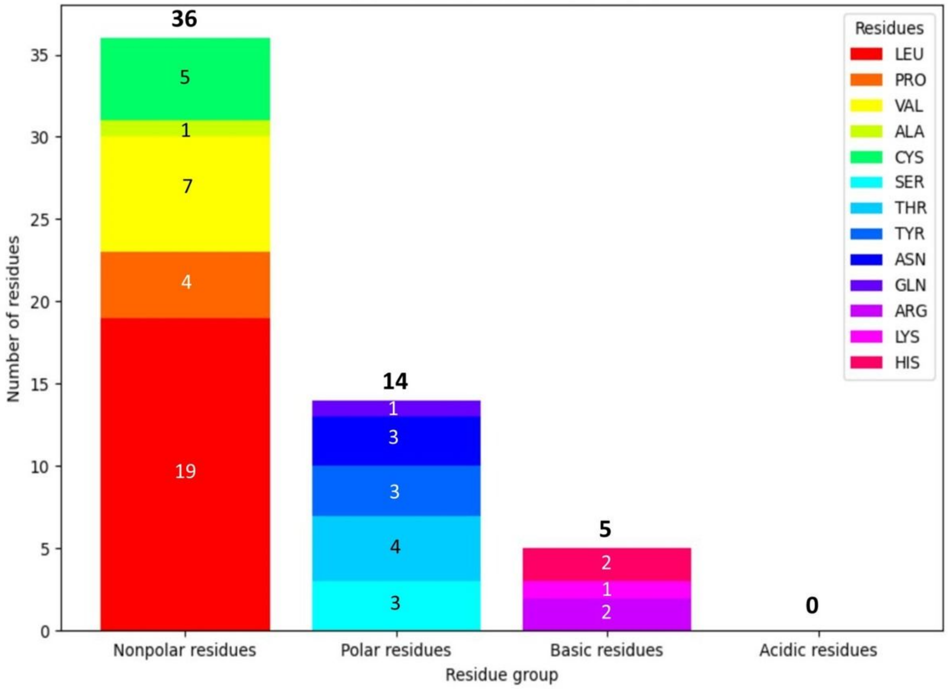


Figure 5.11: Classification of the 55 residues per residue group.

is important to note that although the proportions of individual residue types have changed, they are compensated within their respective residue groups, thereby explaining the consistent proportions of residue groups observed in the previous paragraph.

The 55 residues were also classified by protein chains, as shown in Figure 5.12. Because the four protein chains are chemically identical, it could be expected to observe a even distribution of residues across the protein sub-units since they share the same protein sequence. However, this is not exactly what was observed. While the residues are generally evenly distributed among the four protein chains, there are slight variations. Notably, protein chain C has only 11 residues displaying a bimodal distribution. Additionally, there are differences in the residue composition. For instance, protein chain C is the only chain that lacks PRO residues exhibiting an RMSD shift but is the only chain containing an ALA residue. These subtle differences can be attributed to the fact that although the protein sub-units are chemically identical, they are geometrically not fully identical in the crystal structure [23]. As a result, each protein chain may exhibit slight dynamic variations, leading to the observed distinctions. Additionally, the slight divergence can also be attributed to the insufficient

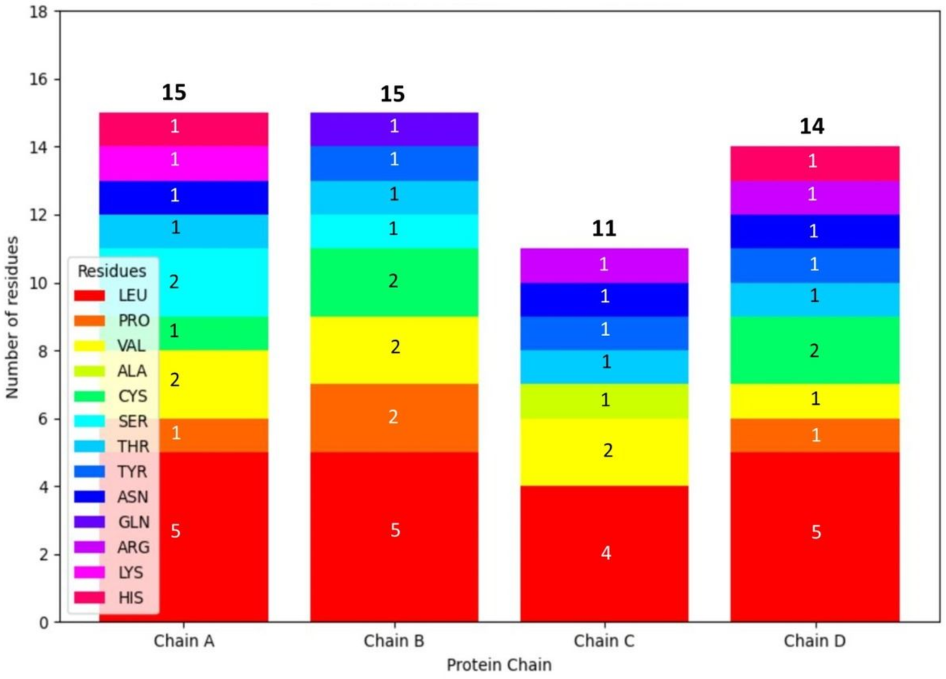


Figure 5.12: Classification of the 55 residues per protein chain.

simulation time of 1 μs to reach convergence between the protein sub-units.

Distances between the center of mass (COM) of each residue and the COM of its corresponding Chl a were calculated. The results are presented in Figure 5.13 and were average over the five replicas, at both temperatures. It can be observed that there is no significant difference between the distances at 300 K and 165 K, indicating that temperature does not affect the overall distance. Additionally, the small standard errors suggest that the position of the residues remains relatively stable throughout the simulations. This noteworthy finding suggests that the earlier detected RMSD shift is likely not caused by a movement in the position of the residues, but rather a smaller conformational change involving specific subsets of atoms within the residues. Furthermore, the distances appear to be disparsed within the range of 0.75 nm to 1.80 nm. Therefore, it seems that the presence of a bimodal RMSD distribution is not influenced by the distance between the residue and its corresponding Chl a.

In summary, the analysis of the RMSD of the 168 nearby residues demonstrated that 55 of them exhibited a bimodal distribution of RMSD characterized by two distinct peaks. This observation

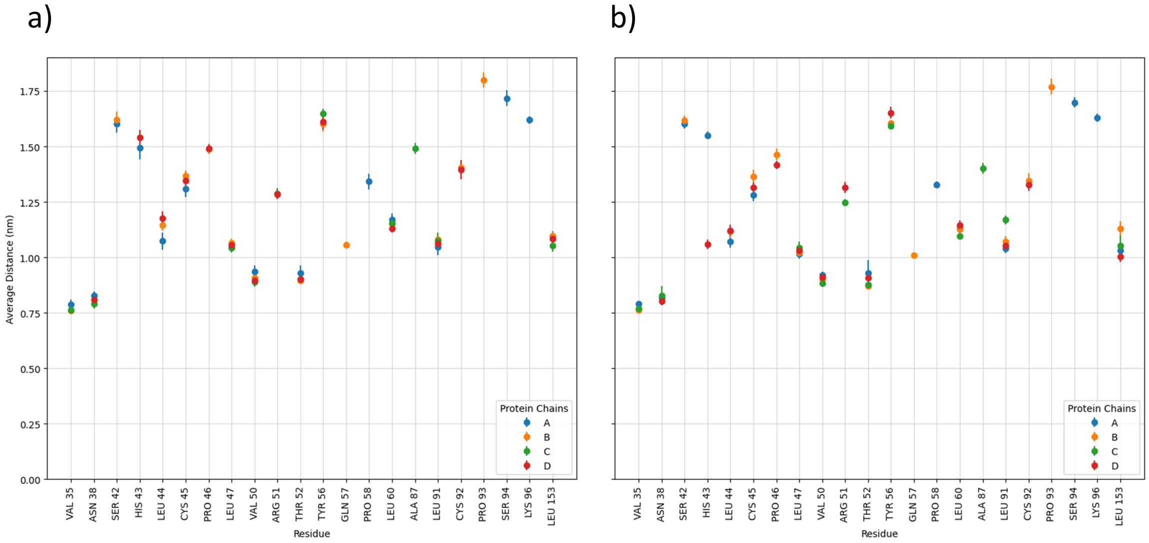


Figure 5.13: Distance between COM of Chlorophylls a and the COM of their respective residues at *a)* 300 K, and *b)* 165 K. Error bars represent standard errors taken over 5 independent replicas.

might suggest the presence of two energetically favored conformations for these residues. The variation in the RMSD values was observed at both temperatures, but with a higher probability at 300 K. Additionally, consistent results were obtained across all four protein chains. Interestingly, the conformational change did not appear to be influenced by the side chain type (non-polar, polar, acidic or basic) of residues or the distance between the residues and their corresponding pigments.

Root-mean-square fluctuations are higher in side chains than in back bone atoms

The Root-mean-square fluctuation (RMSF) of the 168 residues were calculated to quantify the positional changes of individual atoms within each residue. This analysis aimed to gain a detailed understanding of residue dynamics by examining the behavior of individual atoms. The RMSF measures the average deviation of the position of each atom within a molecule throughout simulation, relative to a reference position. In Gromacs 2021.4, the RMSF of a particle i is defined as [89] :

$$RMSF_i = [\ \frac{1}{T}\sum_{t_j=1}^{T} \parallel \mathbf{r}_i(t_j) - \mathbf{r}_i(t_{ref}) \parallel^2\ ]^{\frac{1}{2}} \tag{28}$$

where T is the total simulation time, $\mathbf{r}_i(t_j)$ is the position of a particle i at time t_j, and $\mathbf{r}_i(t_{ref})$ is the position of a particle i at the reference time t_{ref}. The RMSF is averaged over time, giving a single value for each particle i. In our project, the reference position corresponds to the initial position of the residues at the start of the simulations. RMSF calculations were then averaged over the five replicas, at both temperatures. The RMSF analysis revealed that among the 168 residues, the same 55 residues identified in the RMSD analysis exhibited significant fluctuations in the positions of their side chains, while their backbone structure remained relatively stable. This finding confirms the hypothesis proposed in Section 5.1.3, suggesting that the bimodal RMSD distributions of these residues are primarily influenced by the movements of specific subsets of atoms (specifically the side chains), thus being at the origin of the conformational change. The subsequent paragraphs will explain these RMSF results, providing analysis of the observed fluctuations and their implications in the movements of the residues.

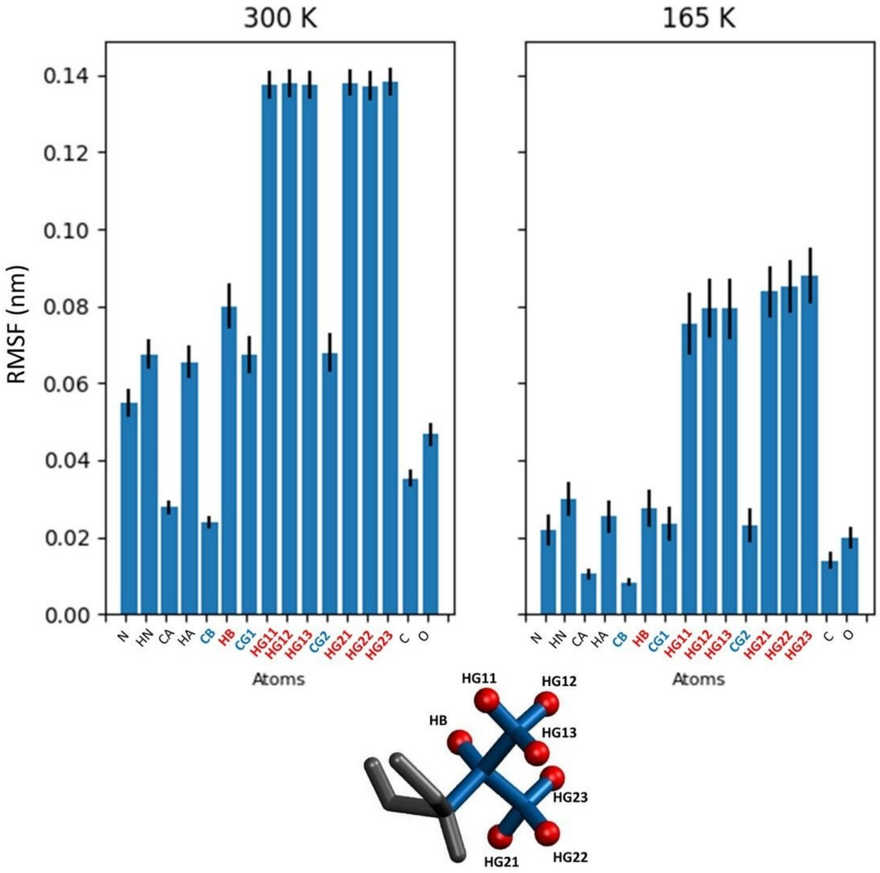


Figure 5.14: Average RMSF values of the 7 Valine residues (VAL 50 in all protein chains and VAL 35 in protein chains A, B, and C), at 300 K and 165 K. Backbone atoms are represented in black, side chain hydrogen atoms in red, and side chain non-hydrogen atoms in blue.

For readability, only results of the 55 residues are shown and the RMSF values were averaged by residue name. Figure 5.14 and Figure 5.15, depict the RMSF values of respectively the 7 VAL (VAL 50 in all protein chains and VAL 35 in protein chains A, B, and C) and the 19 LEU (LEU 47, LEU 60, LEU 91, LEU 153, in all protein chains, and LEU 44 in protein chains A, B and D), that display bimodal RMSD distributions. RMSF plots of the other less abundant residues can be found in the Appendix B (Figure B.3).

We can see that, at 165 K, the RMSF values are lower than at 300 K, this decrease can be attributed to the lower thermal energy available. However, regardless of the temperature, a consistent trend is observed : hydrogen atoms exhibit the highest RMSF values, meaning they exhibit the highest change in position over the course of the simulations. This is totally expected as their smaller

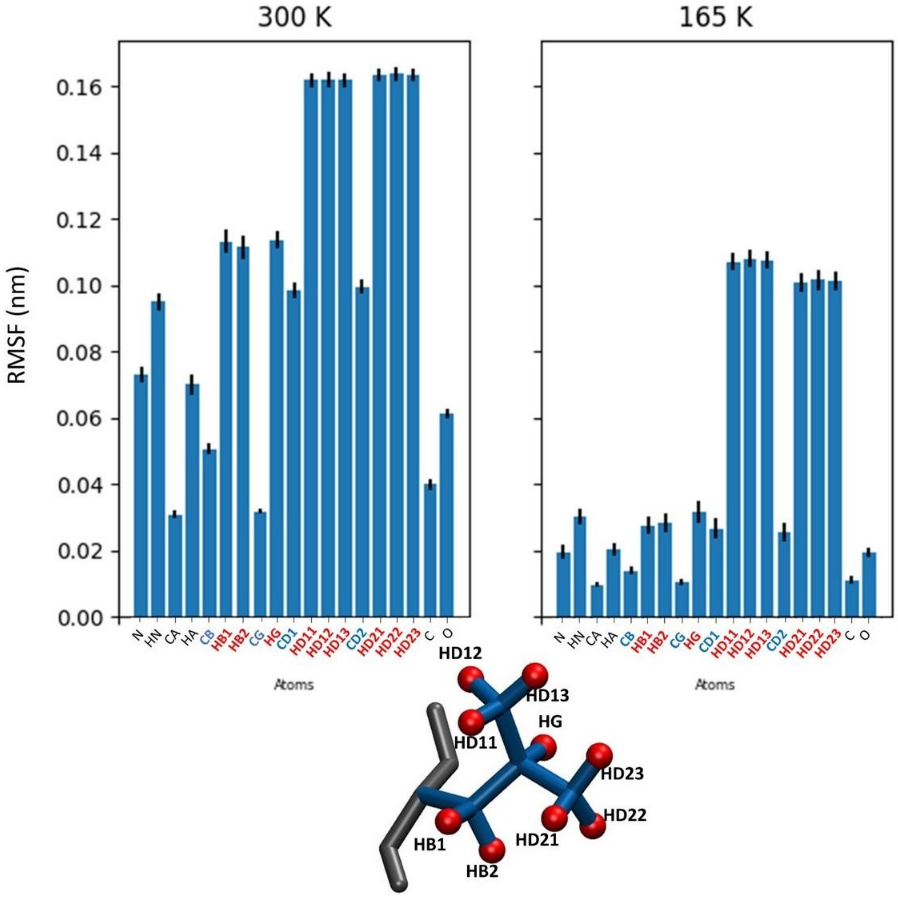


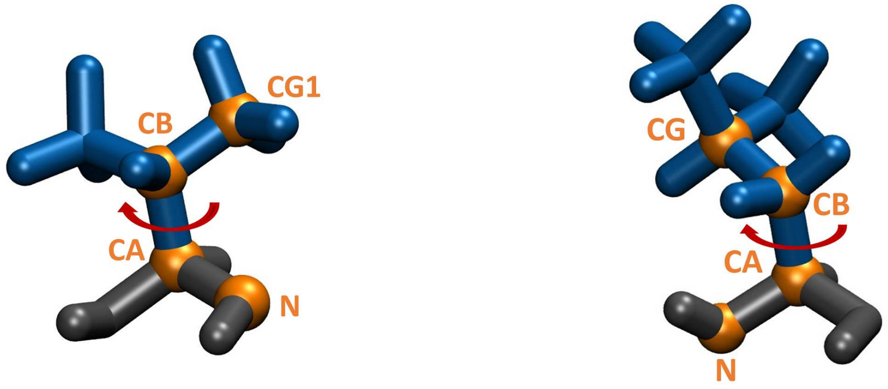
Figure 5.15: Average RMSF values of the 19 Leucine residues (LEU 47, LEU 60, LEU 91, LEU 153, in all protein chains, and LEU 44 in protein chains A, B and D), at 300 K and 165 K. Backbone atoms are represented in black, side chain hydrogen atoms in red, and side chain non-hydrogen atoms in blue.

mass allows them to experience less inertia, making it easier for them to move and respond to external forces. Additionally, their implications in molecular vibrations, such as stretching and bending of chemical bonds, can also contribute to their motion. Nevertheless, the hydrogen atoms (H-atoms) of the side chains still have a significant higher RMSF values than H-atoms of backbone. It implies a greatest motion of the side chain H-atoms over the backbone H-atoms. Therefore, the side chain might be notably more flexible than the backbone. This assumption is supported by the fact, on average, non-heavy atoms in the side chains also exhibit higher RMSF values compared to non-heavy atoms in the backbone, sometimes even being equal or greater than the RMSF values of backbone H-atoms. This is the case for VAL at 300 K : the backbone hydrogens HN and HA have the same

RMSF values of almost 0.07 nm as the side chain carbons CG1 and CG2. This observation is also applicable to LEU at 300 K : the backbone hydrogen HN has the same RMSF value of almost 0.10 nm as the side chain carbons CD1 and CD2. On the other hand, the other LEU backbone hydrogen HA has a significant lower RMSF value close to 0.07 nm. Consequently, the side chains seem to be the subsets of atoms responsible for the bimodal RMSD distributions, which consequently lead to the observed conformational change. Therefore, a more in-depth analysis of the side chains of these 55 residues will be conducted in the next section.

Side chain dihedral angle is a good collective variable for describing multi-modal configurational distributions

Earlier, it was observed, in Section 5.1.3, that certain residues displayed double-peak RMSD distributions, suggesting structural heterogeneity. However, it was also discovered, that the highest RMSF values within these specific 55 residues corresponded to significant fluctuations in the side chain, while the backbone remained relatively stable. In this context, the χ_1 angle became a promising candidate for further investigation. The χ_1 angle is a specific dihedral angle characterizing the rotation of the side chain with respect to the α-carbon (CA) and the nitrogen (N) atoms of the residue backbone, which plays a crucial role in understanding conformational dynamics of the residues. This angle refers to the first side chain dihedral angle in a protein residue. This angle provides insights into how the side chain conformation and orientation change, offering valuable information on the dynamic behavior of these protein residues.



(a) χ_1 angle of the VAL residue.

(b) χ_1 angle of the LEU residue.

Figure 5.16: Graphical representations of the χ_1 angle of the (a) VAL residue, and (b) LEU residue. Backbone atoms are colored in black, side chain atoms in blue, and atoms used for χ_1 angle calculations are colored and labeled in orange. Rotation of the χ_1 angle are represented by a red arrow.

Dihedral angles represent the orientational relationship between four consecutive atoms along a chain. They describe the rotation around a bond connecting two of the four atoms, while the other two atoms define the plane in which the rotation occurs. In the case of χ_1 dihedral angle, the atoms involved are the N and CA atoms of the backbone, and two atoms specific to the side chain of the residue. For VAL, the χ_1 angle is defined by atoms N-CA-CB-CG1, as shown in Figure

5.16a, with CB being the $\beta-$carbon, the first carbon atom in the VAL side chain, and CG1 being the γ_1-carbon, the second carbon atom in the VAL side chain. For LEU, the χ_1 angle is defined by atoms N-CA-CB-CG, as shown in Figure 5.16b, with CG being the $\gamma-$carbon, the second carbon atom in the LEU side chain.

The calculation of χ_1 angles was carried out for all 55 residues using the methodology outlined in Section 4.3.2. However, the proline (PRO) residue has a pyrrolidine unique ring structure, which imposes limitations on the conformational flexibility of its side chain. Consequently, PRO does not possess a χ_1 angle. Instead, calculations for PRO were based on the ω angle, which reflects the rotation around the peptide bond connecting the nitrogen atom of the pyrrolidine ring and the carbon atom of the backbone carbonyl group. This analysis revealed that all 55 residues (expect for the PRO residues which will be discussed later) exhibited a significant variation in the dihedral angle values, which is reflected as trimodal dihedral angle distributions. This variation implies three favored residue conformations corresponding to three specific dihedral angle values. Furthermore, as this angle rotation constitutes a small positional change, it could be responsible for the conformational change observed in experiments. In this section, we will present these distributions and provide potential clues to explain the correlation between the shifts in both the RMSD and the dihedral angle values. Furthermore, we will examine the characteristics of these trimodal distributions which include the positions of the peaks and the angular differences between them.

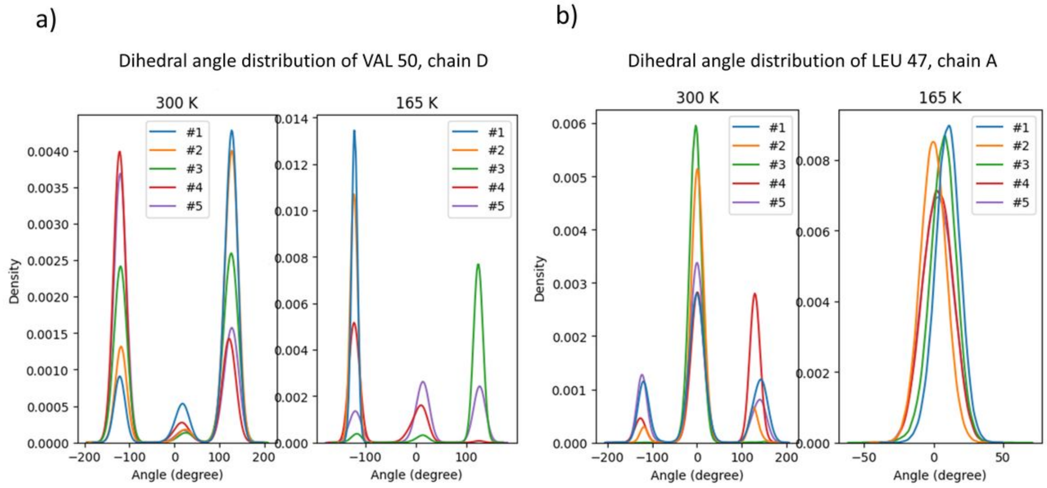


Figure 5.17: Dihedral angle distributions of *a)* VAL 50 in protein chain D, and *b)* LEU 47 in protein chain A, at 300 K and 165 K, for all five replicas.

In Figure 5.17, representative χ_1 angle distributions are displayed for VAL 50 in chain D and LEU 47 in chain A. The presence of three definite peaks in these distributions implies the existence of three distinct conformations associated with different values of the χ_1 angle. Graphical representations of these distinct conformations are illustrated in Figure 5.18 and Figure 5.19. This observation suggests that the side chains of these residues can adopt multiple orientations defined by the χ_1 angle. It indicates the existence of energetically favorable conformations represented by the observed peaks. However, at 165 K, the distributions tend to be more often unimodal or bimodal rather than trimodal. This can explained by the lower probability for the residues to overcome energy barriers at lower temperatures, resulting in fewer observed conformations. Additionally, for the PRO residue, the ω angle distributions exhibit only bimodal patterns at both temperatures, as shown in the Appendix B in Figure B.4 and Figure B.5. This exception can be attributed to its unique pyrrolidine ring structure, which may limit its rotation due to steric hindrance, a smaller number of energetically accessible conformations, or insufficient simulation time.

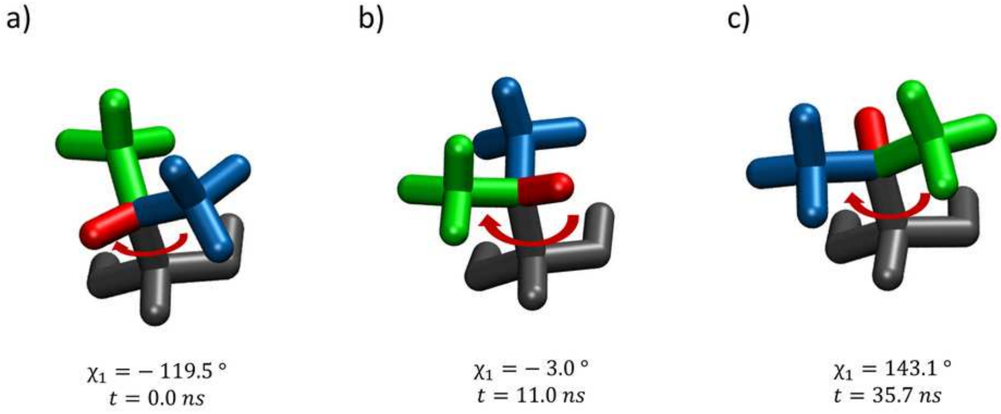


Figure 5.18: Graphical representations of VAL 50 in chain D, in replica #5 of the 300 K simulation at *a)* $t = 0.0\ ns$ with $\chi_1 = -119.5°$, *b)* $t = 11.0\ ns$ with $\chi_1 = -3.0°$, and *c)* $t = 37.7\ ns$ with $\chi_1 = 143.1°$. Backbone atoms are depicted in black, the first side chain group in green, the second in red, and the last in blue. The rotation of the dihedral angle is represented by a red arrow.

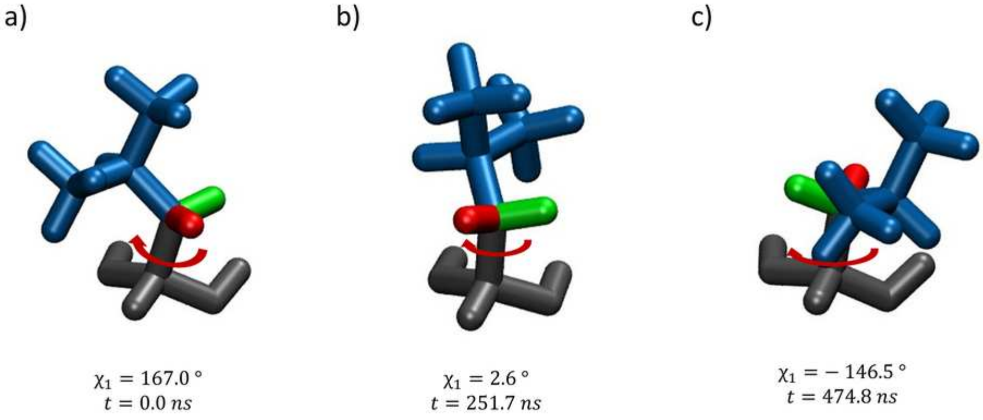


Figure 5.19: Graphical representations of LEU 47 in chain A, in replica #1 of the 300 K simulation at *a)* $t = 0.0\ ns$ with $\chi_1 = 167.0°$, *b)* $t = 251.7\ ns$ with $\chi_1 = 2.6°$, and *c)* $t = 474.8\ ns$ with $\chi_1 = -146.5°$. Backbone atoms are depicted in black, the first side chain group in green, the second in red, and the last in blue. The rotation of the dihedral angle is represented by a red arrow.

Intriguing results were found by comparing the RMSD distributions (Figure 5.10 in Section 5.1.3) to the dihedral angle distributions for the corresponding residues. We observed that replicas displaying a double-peak in the RMSD distributions also exhibited a triple-peak pattern (or sometimes a double-peak pattern at 165 K) in the χ_1 angle distributions. On the other hand, replicas with unimodal RMSD distributions demonstrated unimodal χ_1 angle distributions. An interesting example of this correlation can be seen in the case of LEU 47, in chain A, at 165 K, where all five replicas displayed unimodal RMSD distributions, and the χ_1 angle distributions followed the exact same trend. This suggests that the variation in RMSD values can be attributed to the rotation of the χ_1 angle.

However, it is relevant to note that while the RMSD distributions appear bimodal, the χ_1 angle distributions always exhibit three peaks at 300 K. To understand this discrepancy, it is important to consider the complex relationship between these two measures, which is influenced by various factors. The RMSD reflects the overall structural deviation of the protein, taking into account all residue atoms, while the χ_1 angle focuses solely on the side chain conformation. As a result, the χ_1 angle distribution provides more localized information about the conformational changes occurring in the side chain, which may not be fully captured by the RMSD. This discrepancy can lead to differences in the number of observed peaks. The χ_1 angle value can result in conformations with minimal differences in the overall residue structure, resulting in RMSD values that do not show distinct peaks for these states. Instead, they might appear as a single peak due to their structural similarity. The norm position might also contribute to understanding the variation in peak numbers. If two χ_1 values display minimal deviation from each other in terms of their atomic norm position, it indicates that the corresponding conformations may have negligible differences in the overall protein structure. Consequently, these subtle deviations might not be fully captured in the RMSD calculations, resulting in the observation of only two peaks in the RMSD distribution. However, considering the localized conformational changes reflected in the χ_1 angle, these small differences become distinguishable, leading to the presence of three peaks in the χ_1 angle distribution.

Table 5.4: Characteristics of Trimodal χ_1 Angle Distributions. Results are averaged over all the 55 residues, regardless of their side chain type, at 300 K and 165 K. Standard error of the means taken over 5 independent replicas are displayed on the right side of the columns.

Characteristics of Trimodal Dihedral Angle Distributions.		
	300 K	165 K
First peak position (°)	-132.0 ± 1.3	-128.7 ± 5.4
Second peak position (°)	2.5 ± 1.2	1.8 ± 1.9
Third peak position(°)	126.6 ± 1.7	118.3 ± 2.3
Angular difference first-second peaks (°)	134.5 ± 1.8	130.5 ± 5.2
Angular difference second-third peaks (°)	124.2 ± 1.8	116.5 ± 5.4
Angular difference first-third peaks (°)	101.4 ± 2.3	100.8 ± 9.6

Table 5.4 presents the average characteristics of the χ_1 angle distributions of the residues, considering only the replicas that exhibit a triple-peak distribution. Comparing results obtained at both temperatures, all characteristics fall in the same range of values. The first peak is observed at an angle of -132.0 ± 1.3 ° at 300 K, and at an angle of -128.7 ± 5.4 ° at 165 K. At both temperatures, the second peak is close to 0 °, 2.5 ± 1.2 ° at 300 K, and 1.8 ± 1.9 ° at 165 K. At 300 K, the third peak is observed at 126.6 ± 1.7 °, and at 165 K it is observed at 118.3 ± 2.3 °. These observations suggest that there is consistency of the side chain conformations between the two temperatures. The angular difference first to second peaks and the angular difference second to third peaks values are 134.5 ± 1.8 ° and 124.2 ± 1.8 ° at 300 K, and 130.5 ± 5.2 ° and 116.5 ± 5.4 ° at 165 K. Furthermore, the angular difference first to third peaks is 101.4 ± 2.3 ° at 300 K, and 100.8 ± 9.6 ° at 165 K. Therefore, the peaks are not exactly equally spaced, with the first and third peaks being the closest, considering periodicity of the angles.

In summary, the investigation of the dihedral angle χ_1 in 55 specific protein residues revealed a distribution with three distinct peaks, indicating the existence of three energetically preferred side chain conformations. More precisely, the three distinct conformations adopted by these residues correspond to three specific χ_1 angle values, regardless of the temperature. The correlation between the RMSD and χ_1 angle distributions suggested that the variation in RMSD values could be attributed to the rotation of the χ_1 angle. The analysis of the characteristics of the trimodal distributions at different temperatures showed almost no difference in peak positions and in the angular differences between peaks, indicating temperature-independence in the preferred conformations and side chain orientations. Overall, this analysis revealed the presence of conformational changes in the side chain orientations of the 55 residues of interest.

5.1.4 Protein free energy landscape associated with the χ_1 angle rotation

As discussed in Section 3.1.2, the multidimensional free energy landscape of the protein can be simplified and represented along a single generalized coordinate x. This simplification enables us to gain insights into the system behavior in a reduced-dimensional space. The free energy F along the coordinate x is expressed by the equation [90]:

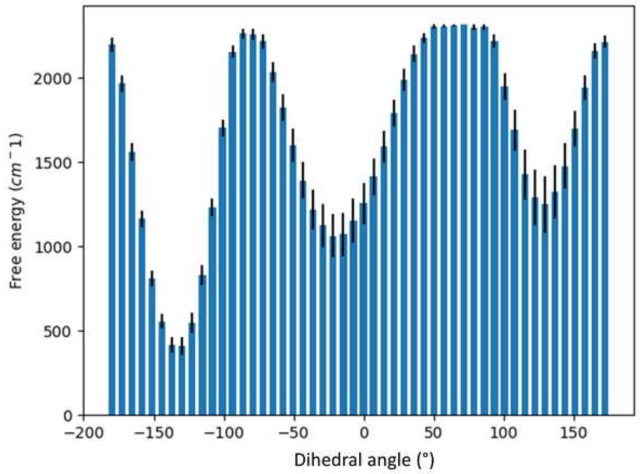
$$F(x) = -k_B T \, ln(P(x)) \tag{29}$$

where k_B represents the Boltzmann constant, T denotes the temperature, and $P(x)$ represents the probability distribution function of x. The formula is derived from the Boltzmann distribution. According to this distribution, the probability $P(x)$ of observing a specific value of x is proportional to the exponential of the negative free energy $F(x)$ associated with that value [90]:

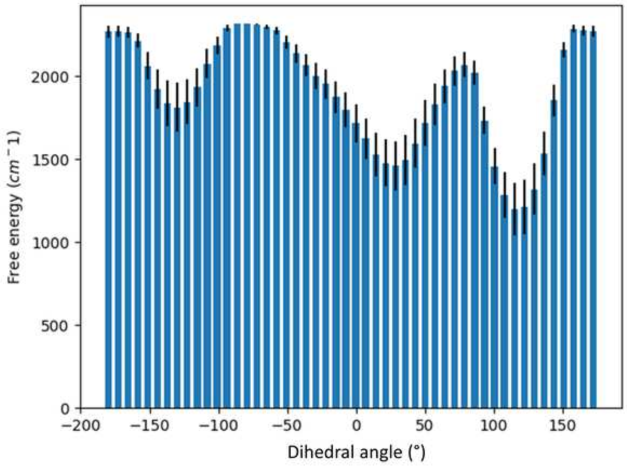
$$P(x) \propto exp(\frac{-F(x)}{k_B T}) \tag{30}$$

By taking the logarithm of both sides of the Boltzmann distribution equation and rearranging the terms, we can derive the simplified Equation 29. To obtain the one-dimensional free energy landscape, the χ_1 values were binned by dividing the range of plausible values (from $-180°$ to $180°$) into bins. For each bin, the probability distribution $P(x)$ was calculated by determining the number of frames that fell within that particular bin and dividing it by the total number of frames (100 000) in the simulation. This normalization step ensures that $P(x)$ represents a valid probability distribution, where the probabilities sum up to 1 over all the bins. It is important to note that any normalization constant causes shift of the landscape up or down in free energy, without changing its shape.

The resulting protein free energy landscapes along the χ_1 angle values at 300 K and 165 K contain three energy wells for the free preferred structural conformations. The heights of the energy barriers were calculated and compared to experimental data. At 300 K, they appeared to fall within a similar range of values. However, as the heights of energy barriers are smaller at 165 K, these values do not fall in agreement with experiments and suggest a temperature-dependent free energy landscape. The energy barriers shown in this thesis are barriers for the pigment-protein system where chlorophyll a is in the ground electronic state. Quantum mechanical calculations would be required to obtain excited-state energy barriers to compare to the dual-Two-Level system model (Section 3.2) used to explain spectral hole burning results.



(a) One-dimensional protein free energy landscape at 300 K.



(b) One-dimensional protein free energy landscape at 165 K.

Figure 5.20: Protein Free energy landscapes associated with the rotation of χ_1 angle at (a) 300 K, and (b) 165 K. Results are averaged over the 7 remaining Valine residues. Error bars represent standard error of the means taken over 5 independent replicas of the free energy values.

Representative protein free energy landscapes associated with the χ_1 angle are illustrated in Figure 5.20 at 300 K, and 165 K. The range of possible χ_1 angle values is similar at both temperatures.

This is in accordance with results found in Section 5.1.3. We can see that all FELs display three energy wells corresponding to three distinct conformations with respect to three distinct χ_1 angle values. This result is also consistent with the trimodal χ_1 angle distributions. In the figures, we can see that regardless of the temperature, the multi-well FEL is not symmetrical. At 300 K, the first energy well is the deepest, indicating its most preferred conformation around $-140°$, while at 165 K, this well is the shallowest. It implies that even though at both temperatures the three same configurations subsist, the most energetically favorable conformation might be temperature-dependent. Nevertheless, at 300 K, the second and third energy wells exhibit similar free energy values, and the same trend persist at 165 K. Quantitative values will be provided in the subsequent paragraph for a more detailed analysis.

Table 5.5: Characteristics of the protein free energy landscapes. Standard error of the means taken over 5 independent replicas are displayed on the right side of the columns.

Characteristics of the protein free energy landscapes		
Energy barrier height (cm^{-1})	300 K	165 K
First-second wells	1530 ± 119	679 ± 146
Second-third wells	1160 ± 149	741 ± 228
First-third well	1384 ± 145	782 ± 176

Average free energy barrier heights are presented in Table 5.5. A 300 K, we can see that energy barrier heights fall within the range of 1100 to 1600 cm^{-1}, which aligns with the values typically observed in spectroscopy experiments [84]. However, an intriguing result emerged from this table : the energy barrier heights at 165 K are approximately half of those observed at 300 K. This result is important because in simple models used to simulate spectroscopy results [84] it was implicitly assumed that the barriers are temperature-independent, and only the transition rates change with temperature. In fact, lower temperatures typically reduce the thermal energy available for the protein to overcome energy barriers, making transitions between conformations less probable. A plausible explanation for this observed discrepancy could be the sampling limitations. The 1 μs

simulations might not have been sufficient to adequately sample the conformational space. Transitions between different states can be rare events that require longer simulation times to capture. It is possible that the 165 K simulations did not adequately explore the conformational space, leading to an underestimation of the energy barriers.

In conclusion, the investigation of the local environment of the pigments in WSCP led to the identification of a specific generalized coordinate that accurately describes a conformational change relevant to the experimental results. Moreover, the energy barrier heights observed at 300 K roughly align with the expected ranges based on experimental data. At 165 K, the energy barrier heights are approximately half of the expected values. To address this disagreement, simulations at 165 K could be extended to potentially reach a convergence. Longer simulation times would allow for a batter exploration of the conformational space and increase the probability of observing transitions between different conformations. By increasing the rate of transitions, it is possible to obtain a more accurate estimation of the energy barrier heights at 165 K. Alternatively, one could also performed Replica Exchange Molecular Dynamics (REMD) simulations to enhance the sampling of the conformational space. In REMD simulations, each replica is simulated at a different temperature, typically spanning a range from low to high temperatures. The replicas exchange information periodically, allowing the system to explore different energy landscapes and overcome energy barriers more efficiently. The goal of REMD simulations is to enhance the sampling of the conformational space and explore the thermodynamic and kinetic properties of the system more effectively than conventional MD simulations at a single temperature. By incorporating temperature exchanges, REMD simulations enable the system to escape from local energy minima, sample different regions of the energy landscape, and enhance the exploration of transition states.

5.2 Dynamical Network analysis determines motion correlations between residues in both the local and global environnment

We employed Dynamical Network analysis to investigate the motion correlations of the 55 residues undergoing conformational changes in the hydrophobic cavity around Chls a. The analysis revealed that high correlations (greater than 0.8) were predominantly observed among residues that are adjacent in the protein sequence, indicating a synchronized motion. Interestingly, residues with the same name or side chain type did not exhibit specific correlation patterns, suggesting that factors other than residue identity play a more significant role in motion correlations. The results indicate that conformational changes can occur non-simultaneously and at varying rates depending primarily on the location of the residues. Furthermore, we expanded our analysis to explore correlations between these 55 nearby residues and those located farther away from the pigments. Our findings revealed that Threonine 180 in each protein sub-unit displayed correlations with approximately 9 of the previously identified residues, regardless of the temperature. The computed χ_1 angle distributions exhibited three peaks at 300 K, indicating the occurrence of conformational change, specifically a rotation around this angle. To further understand the dynamics of Threonine 180, we mapped its protein free energy landscape associated with the rotation of its χ_1 angle, at 300 K. The analysis revealed energy barriers heights in the range of 1 700 to 2 000 cm^{-1}, which are similar to the magnitudes observed in the earlier Section 5.1.4.

5.2.1 Motion correlations between close residues undergoing conformational change

This section focuses on examining the correlations in the motion of the 55 known residues that have been observed to undergo conformational changes, which are reflected in the rotation of their χ_1 angle (Section 5.1.3). The analysis employed a Dynamical Network approach, as described in Section 4.3.3. By analyzing the network, we aimed to identify correlations in the motion of these residues throughout the simulations. These correlations are represented as edges connecting the residues, with the thickness of each edge indicating its normalized weight. An edge with a thickness of 1 indicates that the residues move in a synchronized manner throughout all frames of the simulation. The primary goal was to determine if residues sharing similar properties, such as residue name, residue side chain type, or proximity in protein sequence, exhibited high correlations in their motion. Such correlations could suggest that conformational changes occur simultaneously and at comparable rates between these residues. The findings revealed high correlations (greater than 0.8) among residues that are adjacent in the protein sequence. Conversely, lower correlations (between 0.4 and 0.6) were observed for residues farther apart in the protein sequence. These results indicate that the specific residue name and side chain type do not play a substantial role in determining motion correlations. The most influential factors appear to be the proximity of residues in both the protein sequence and solution space[1]. These results may imply that conformational changes occur simultaneously and at a similar rate for closely located residues but non-simultaneously and at varying rates for farther residues. In other words, sub-groups of residues within the hydrophobic cavity around Chls a may undergo conformational changes independently, with their own distinct kinetics.

Figure 5.21 displays symmetrical heatmaps illustrating the weights of the edges and network representations for each protein chain at 300 K. Similarly, Figure 5.22 presents the corresponding heatmaps and network representations at 165 K. The results presented in this study were averaged over five independent replicas of the simulations. The standard errors of the means, which indicate the variability between replicas, were not displayed as they were too small to be visually discernible. The small magnitude of the standard errors, ranging from 0.001 to 0.010, suggests a high level of agreement between the independent runs of the simulations. This agreement reinforces the consistency of the observed correlations in motion among the residues.

For a given temperature, we can see slight divergence in residue correlations between the protein chains. For instance, in protein chain B, at 300 K, VAL 50 and LEU 153 exhibit an average edge weight of 0.4, while no significant correlation is observed between these residues at the same temperature in the other three protein chains. This slight discrepancy in correlations between the protein chains can be attributed to the relatively short simulation time. Despite these discrepancies, the overall trend in the results remains consistent across the protein chains. Residues located in close proximity to each other, either through direct covalent bonds or within a short distance of up to 5 residues, exhibit correlations greater than 0.8. As an example, this is the case of VAL 50 and

[1] The solution space refers to the set of all possible conformations or structures that a protein can adopt.

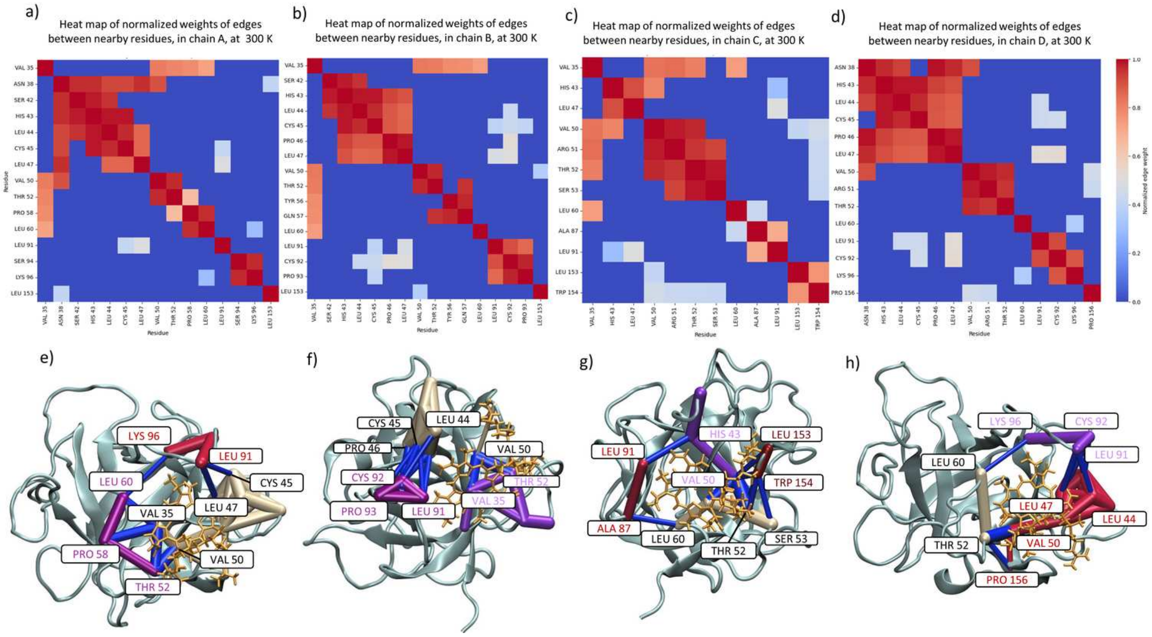


Figure 5.21: a), b), c), and d), represent the heat maps of the normalized weights of edges between residues near Chls a that undergo conformational change, in respectively, protein chains A, B, C, and D, at 300 K. e), f), g), and h) illustrate their respective networks, and residues involved in inter-communities interactions are labeled. Communities are colored in beige, pink, magenta, burgundy and purple. Inter-community links are colored in navy blue. The Chl a is colored in orange and the protein sub-unit in blue. The thicknesses of the edges correspond to the normalized weights of these edges.

THR 52, which display a edge weight greater than 0.8 in all protein chains and at both temperatures. These high correlations observed among nearby residues can be attributed to the covalent bonds, as residues that are directly bonded are typically part of the same secondary structure element, such as a α-helix or β-strand. Covalent bonds restrict the relative motion between these residues, leading to a higher correlation in their motion. Therefore, residues such as LEU 91 and LEU 47, which are farther apart, exhibit a lower correlation of approximately 0.4 to 0.6 in each protein chain. As a consequence these results, residues within the same community tend to be located in close proximity. In network analysis, a community is referred as group of residues more densely connected to each other than to the residues outside the community. In line with this concept, we can observe that residues within the same community, such as LEU 44, cysteine (CYS) 45, and PRO 46 in protein

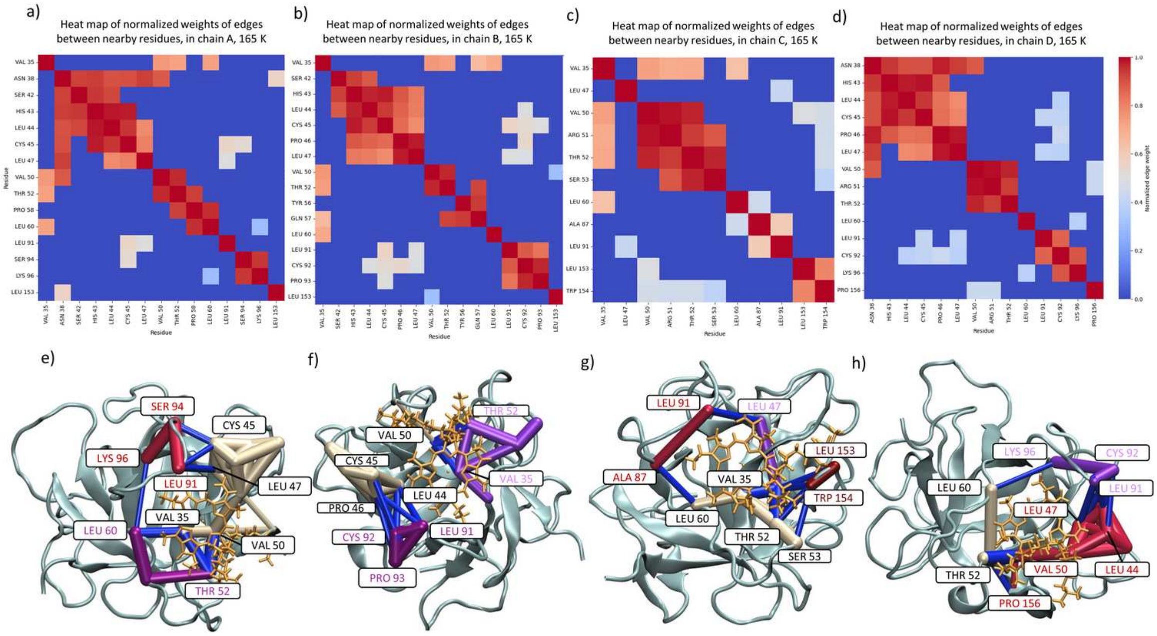


Figure 5.22: a), b), c), and d), represent the heat maps of the normalized weights of edges between residues near Chls a that undergo conformational change, in respectively, protein chains A, B, C, and D, at 165 K. e), f), g), and h) illustrate their respective networks, and residues involved in inter-communities interactions are labeled. Communities are colored in beige, pink, magenta, burgundy and purple. Inter-community links are colored in navy blue. The Chl a is colored in orange and the protein sub-unit in blue. The thicknesses of the edges correspond to the normalized weights of these edges.

chain B, and LEU 91, CYS 92, and lysine (LYS) 96 in chain D, are located in close spatial proximity. This analysis might suggest that residues within the same community might display a rotation of their χ_1 angle at the same rate.

For a given protein chain, a strong similarity is observed between the correlations in motion at 300 K and 165 K. The residues maintain their correlations, with edge weights that are nearly identical between the two temperatures. As an example, in protein chain A, residues such as asparagine (ASN) 38, SER 42, HIS 43, and LEU 44 exhibit edge weights greater than 0.7 between them at both 300 K and 165 K. This finding suggests that the correlations in motion among these residues are not strongly influenced by temperature. It is known that temperature affects the dynamics of protein residues, with lower temperatures typically leading to slower and more restricted motions. However, the consistent correlations observed at different temperatures suggest that the temperature drop affects all the residues in a similar manner.

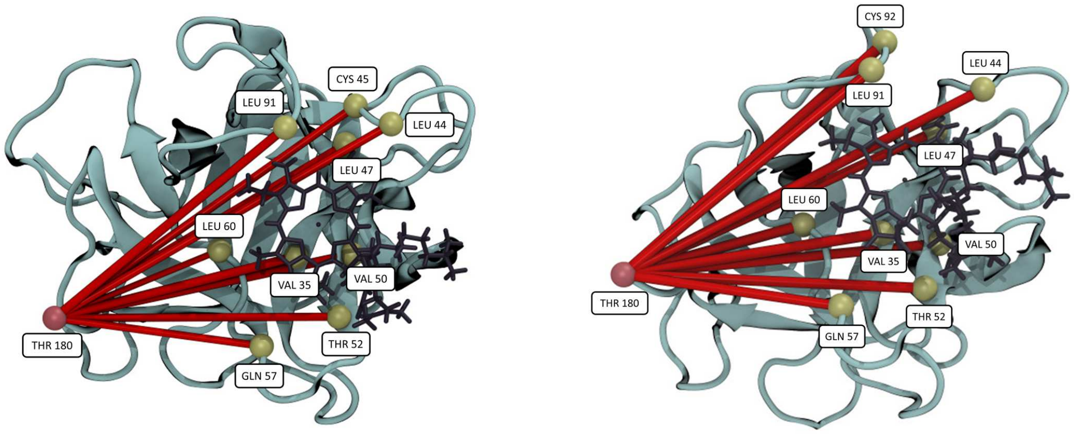
Additionally, the results also indicate that residues with the same name or side chain type (non-polar, polar, acidic or basic) do not exhibit a specific correlation pattern. This can be exemplified by the case of LEU 60 and LEU 91, which are not correlated in any of the protein chains, despite having the same residue name, and therefore the same chemical composition. Similarly, the lack of correlation between VAL 35 and LEU 153, both non-polar residues, as well as between THR 52 and SER 94, both polar residues, supports this finding. These results suggest that the specific residue name or side chain type alone is not a determining factor for correlations in motion.

To summarize, the analysis reveals that despite undergoing the same conformational change (rotation of the χ_1 dihedral angle), the residues do not exhibit specific or strong correlations unless they are in close proximity. This suggests that the conformational change may occur at the same rate for nearby residues but different rates for farther residues. Therefore, there is no evidence for allostery[2] in the nearby 55 residues.

[2] Allostery refers to a regulatory mechanism observed in proteins where the binding of a molecule, called an allosteric effector, at one site on the protein induces a conformational change at a distant site, thereby modulating the protein activity or function.

5.2.2 Identification of another residue that undergoes conformational change through network connectivity

To explore potential correlations and dynamics beyond the 55 identified nearby residues that undergo rotation in their side chain dihedral angle χ_1, we employed Dynamical Network Analysis. The aim was to investigate any correlations between these nearby residues and other residues situated farther away from the pigments. High correlations between these distant residues and the nearby ones could suggest their involvement in similar conformational changes. Our analysis revealed that Threonine (THR) 180 in each protein sub-unit displayed correlation with approximately 9 of the previously identified residues, both at 300 K and 165 K temperatures. At 300 K, the computed χ_1 angle distributions exhibited three peaks. This observation reveals the occurrence of conformational change, specifically rotation of this particular angle. Furthermore, by mapping the protein free energy landscape of this residue at 300 K, we determined that the heights of the energy barriers are within a similar range as those calculated in Section 5.1.4.



(a) Dynamical network of THR 180, in protein chain B, at 300 K.

(b) Dynamical network of THR 180, in protein chain B, at 165 K.

Figure 5.23: Graphical representations of the dynamical network of THR 180 in protein chain B, at (a) 300 K, and (b) 165 K, averaged over the 5 replicas. The protein chain is colored in blue, the Chlorophyll a in black, the $\alpha-$carbon of THR 180 in pink, and the $\alpha-$carbon of the identified nearby residues correlated with THR 180 in green. The edges between THR 180 and the other residues are shown in red, and their thicknesses represent their weights.

Figure 5.23 presents the correlation between THR 180 in protein chain B and the neighbouring residues at both temperatures. Interestingly, we observe that THR 180 exhibits correlations with the similar set of residues, irrespective of the temperature. However, there is one difference between the temperatures. At 300 K, THR 180 shows correlation with CYS 45, while it is not correlated with CYS 92. Conversely, at 165 K, THR 180 is correlated with CYS 92, but not with CYS 45. Furthermore, these residues are not concentrated in the same region of the protein sub-unit but are rather dispersed within the environment of Chlorophyll a. Additionally, these residues do not form a single cohesive community, as they do not exhibit strong correlations with one another. THR 180 is not part of any community of the nearby residues.

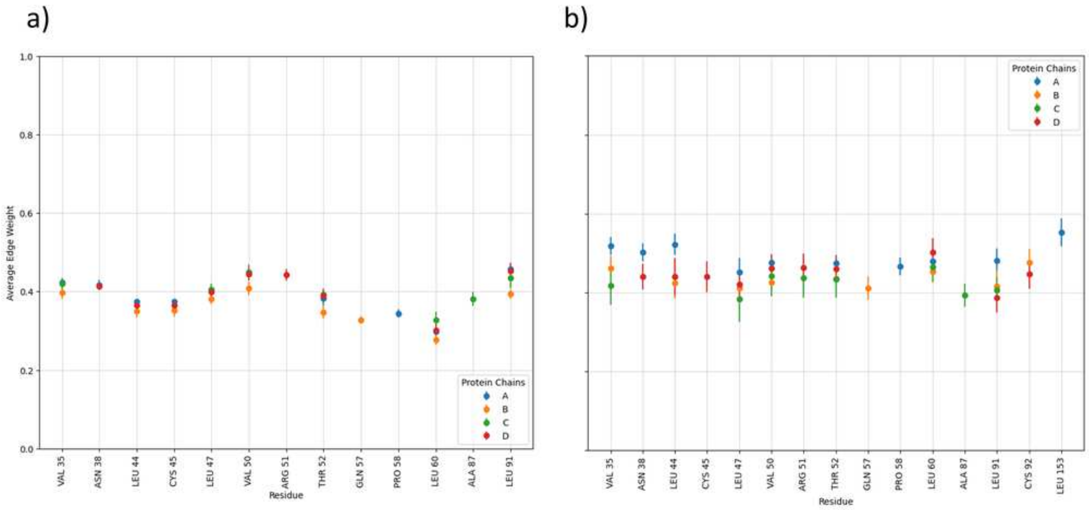


Figure 5.24: Normalized weights of edges between THR 180 and the nearby residues at *a)* 300 K, and *b)* 165 K. Error bars represent standard errors taken over 5 independent replicas.

Figure 5.24 displays the average weights of the edges connecting THR 180 with the neighbouring residues in all protein chains, at 300 K and 165 K. We observe similar average weights for all residues at both temperatures. Specifically, the average edge weight is always around 0.4 at 300 K and 0.5 at 165 K, indicating consistent interaction strengths across the entire group of residues. However, this slight weight difference between temperatures could be due to the effects of decreased thermal energy and slower molecular motion at 165 K, allowing for more stable interactions between the residues. Interestingly, THR 180 does not exhibit any edge weight higher than 0.6 with these residues. This result suggests that only transient correlations exist between them, revealing fluctuations or variability in their correlations throughout the simulations.

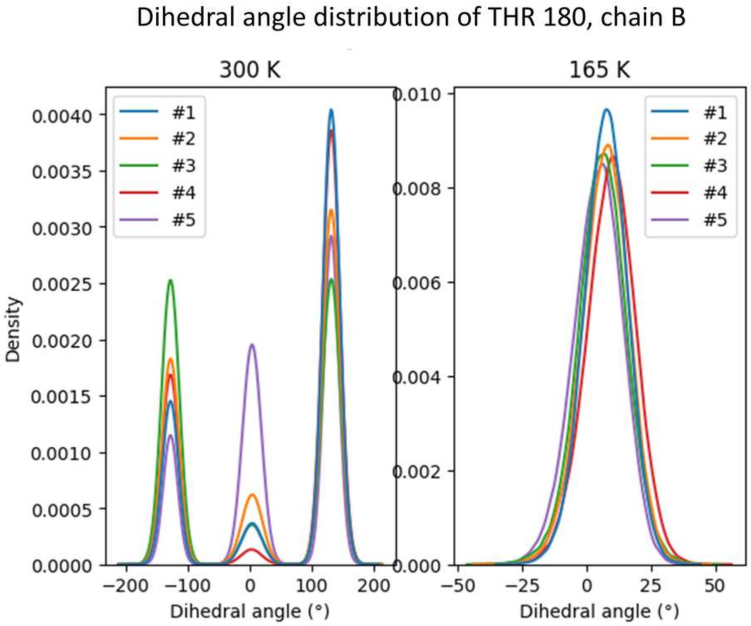


Figure 5.25: Dihedral angle distributions of THR 180 in protein chain B, at 300 K and 165 K, for all five replicas.

The χ_1 angle distribution of THR 180 was computed for all protein chains at both 300 K and 165 K. At 300 K, all the distributions displayed a trimodal pattern, consistent with the findings discussed in Section 5.1.3, where trimodal χ_1 angle distributions were also observed. Furthermore, the average peak positions were found to be approximately -134.0 ± 2.1 °, 1.1 ± 0.9 °, and 127.8 ± 1.9 °, which fall within the same range as the values also reported in Section 5.1.3. These results indicate the presence of the same conformational change, the rotation of the χ_1 angle, occurring in THR 180. In contrast, at 165 K, only single peak distributions were observed. This observation does not contradict the previous statement, as similar findings were also obtained for the nearby residues. This suggests that at lower temperatures, conformational change does not occur for this residues, and longer simulation times are required to overcome the energy barriers and observe the full range of conformational changes.

Figure 5.26 depicts the average one-dimensional protein free energy landscape of THR 180 across all protein chains at 300 K. The three-well energy profile observed in the figure provides further confirmation of the presence of three distinct conformations associated with three distinct

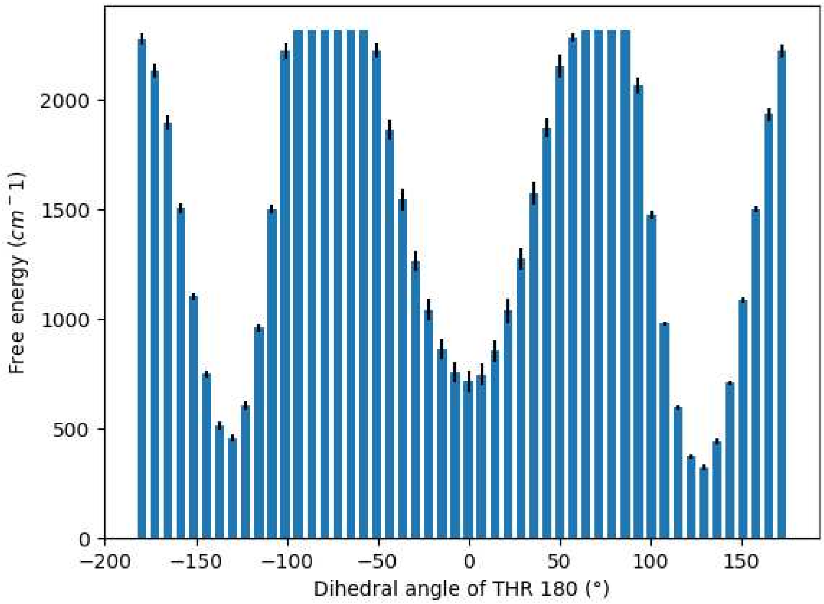


Figure 5.26: Average protein free energy landscape associated with the rotation of χ_1 angle of THR 180, accross all protein chains, and all replicas, at 300 K.

χ_1 angle values. This observation is consistent with the previously observed trimodal χ_1 angle distributions. Furthermore, Table 5.6 presents the average energy barrier heights. These barrier heights, although falling within the same order of magnitude as those found in Section 5.1.4 and in [84] (ranging from 1 000 to 1 500 cm^{-1}), are slightly higher. These higher energy barriers indicate that the transitions between different conformations of THR 180 require more energy to overcome compared to the transitions observed in the reference sources.

Table 5.6: Energy barrier heights of the protein free energy landscape of THR 180 associated with its χ_1 angle at 300 K. Standard error of the means across all protein chains and replicas are displayed on the right side of the column.

Energy barrier heights (cm^{-1})	
First-second wells	1728 ± 33
Second-third wells	2094 ± 30
First-third well	1825 ± 41

In conclusion, through the application of Dynamical Network Analysis, we have identified a new residue, THR 180, located farther away from the pigments, that also undergoes the same conformational change characterized by the rotation of its χ_1 dihedral angle at 300 K. This finding expands our understanding of the conformational dynamics within the WSCP complex and indicate that residues beyond the hydrophobic cavity of the protein can also contribute to the observed conformational changes identified through spectroscopy experiments, despite the lack of evidence for allosteric interactions between them.

Chapter 6

Conclusion and Future Work

Throughout our investigation of the Water-Soluble Chlorophyll-a binding Protein, we have successfully confirmed the existence of small conformational changes observed in optical spectroscopy experiments. These changes primarily involve residues located in the hydrophobic cavity surrounding the pigment molecules, with Leucine and Valine being the predominant non-polar residues involved. The conformational changes manifest as rotations of the side chain dihedral angle χ_1, which exhibit three preferred conformations at approximately $-130°$, $1°$, and $125°$, both at 300 K and 165 K temperatures. Our investigations into the protein free energy landscapes along this generalized coordinate have revealed temperature-dependent potential energy barrier heights. Specifically, at 300 K, the potential energy barriers were found to range between 1 000 and 1 500 cm^{-1}, consistent with experimental observations. However, at 165 K, the potential energy barrier heights decreased to approximately 750 cm^{-1}. This finding contradicts assumptions of the model used to describe spectroscopy experiments. In that model, the free energy landscape is assumed to be temperature-independent, with only the rates of transitions between different conformational states depending on temperature. Additionally, through correlated movements analysis of the identified residues, we observed strong correlations between residues in close proximity, indicating a potentially synchronized conformational change rate among them.

As a future work, the Molecular Dynamics simulations of 1 μs could be extended to achieve better convergence among the five independent replicas, particularly at 165 K, where some replicas encountered challenges in fully exploring the conformational space. This extension would lead to more precise results. Another way to enhance the conformational space sampling is to perform REMD simulations. The analysis of hydrogen bond rearrangements could also be extended to broader range of hydrogen bond lengths, as current range was a result of a confusion between classical and quantum calculations. Furthermore, the energy landscapes obtained from our simulations can be utilized to model the system evolution, incorporating Quantum Mechanical tunneling. Approaches described in [84] [91], which have already been implemented in Dr. Zabuvovits lab software, can be employed for this purpose. Additionally, with the identified generalized coordinate, it would be beneficial to examine and compare the rates of conformational changes between different residues, investigating potential differences or similarities among them. Moreover, conducting simulations on other pigment-protein complexes such as Cytochrome b_6f or the Light harvesting 2 antenna complex of purple bacteria could provide insights into whether similar small conformational changes are observed in these systems. Notably, non-photochemical hole burning data are already available for these two complexes, enabling direct comparisons.

Overall, the research presented in this thesis has contributed to the exploration of phenomena observed in spectroscopy experiments in the field of photosynthesis. Gaining a comprehensive understanding of photosynthesis is crucial, as it not only enables the development of renewable energy solutions but also enhances our understanding of fundamental biological processes.

Bibliography

[1] R. Jankowiak et al. Site selective and single complex laser-based spectroscopies: A window on excited state electronic structure, excitation energy transfer, and electron–phonon coupling of selected photosynthetic complexes. *Chem. Rev.*, 111(8):4546–4598, 2011.

[2] P.W. Anderson et al. Anomalous low-temperature thermal properties of glasses and spin glasses. *Philosophical Magazine*, 25(1):1–9, 1972.

[3] W.A. Phillips et al. Tunneling states in amorphous solids. *J. Low. Temp. Phys.*, 7:351–360, 1972.

[4] X. Wang et al. Open quantum system parameters for light harvesting complexes from molecular dynamics. *Phys. Chem. Chem. Phys.*, 17:25629–25641, 2015.

[5] A. Sisto et al. Atomistic non-adiabatic dynamics of the lh2 complex with a gpu-accelerated ab initio exciton model. *J. Phys. Chem. Chem. Phys.*, 19:14924–14936, 2017.

[6] B.J. Harris et al. All-atom molecular dynamics of a photosystem i/detergent complex. *J. Phys. Chem. B*, 118(40):11633–11645, 2014.

[7] M. Mallus et al. Environmental effects on the dynamics in the light-harvesting complexes lh2 and lh3 based on molecular simulations. *Chem. Phys.*, 515(14):141–151, 2018.

[8] L. Cupellini et al. An ab initio description of the excitonic properties of lh2 and their temperature dependence. *J. Phys. Chem. B*, 120:11348–11359, 2016.

[9] F. J. Van Eerden et al. Molecular dynamics of photosystem ii embedded in the thylakoid membrane. *The Journal of Physical Chemistry B*, 121(15):3237–3249, 2017.

[10] V.O. Voitsekhovskaja et al. Chlorophyll b in angiosperms: Functions in photosynthesis, signaling and ontogenetic regulation. *Journal of Plant Physiology*, 189:51–64, 2015.

[11] V.I. Sugiyama et al. Analyses of absorption and fluorescence spectra of water-soluble chlorophyll proteins, pigment system ii particles and chlorophyll a in diethylether solution by the curve-fitting method. *Biochim Biophys Acta*, 503(1):107–119, 1978.

[12] S.S. Brody. Temperature induced changes in the absorption spectra of porphyridium cruentum and anacystis nidulans. *Zeitschrift fur Naturforschung C*, 36(11–12):1013–1020, 1981.

[13] Chlorophyll a and chlorophyll b representation. https://www.majordifferences.com/2013/05/difference-between-chlorophyll-and.html.

[14] Chlorophyll a and chlorophyll b absorption spectra. https://commons.wikimedia.org/wiki/File:Chlorophyll_Absorption_Spectrum.svg.

[15] L. O. Björn et al. Photosynthetic production of molecular oxygen by water oxidation. *Oxygen*, 2(3):337–347, 2022.

[16] K. Noguchi. Molecular mechanism of asymmetric electron transfer on the electron donor side of photosystem ii. *Advances in Photosynthesis and Respiration*, 47:323–339, 2021.

[17] S. Vasiliev et al. Molecular dynamics simulations reveal highly permeable oxygen exit channels shared with water uptake channels in photosystem ii. *Biochimica et Biophysica Acta*, 1817:1148–1155, 2013.

[18] J. Yan et al. On the structural role of the aromatic residue environment of the chlorophyll a in the cytochrome b6f complex. *Biochemistry*, 47(12):3654–3661, 2008.

[19] D. Shevela et al. Photosynthesis, solar energy for life. 2018.

[20] G. E. Milanovsky et al. Molecular dynamics study of the primary charge separation reactions in photosystem i: Effect of the replacement of the axial ligands to the electron acceptor a(0). *Biochimica et Biophysica Acta*, 1837:1472–1483, 2014.

[21] G. W. Brudvig et al. Water oxidation chemistry of photosystem ii. *Philos Trans R Soc Lond B Biol Sci.*, 363(1494):1211–1219, 2008.

[22] G. Kurisu et al. Structure of the cytochrome b6f complex of oxygenic photosynthesis: tuning the cavity. *Science*, 302(5647):1009–1114, 2003.

[23] D Horigome et al. Structural mechanism and photoprotective function of water-soluble chlorophyll-binding protein. *The Journal of Biochemical Chemistry*, 282(9):6525–6531, 2007.

[24] S. Takahashi et al. Molecular cloning, characterization and analysis of the intracellular localization of a water-soluble chlorophyll-binding protein (wscp) from virginia pepperweed (lepidium virginicum), a unique wscp that preferentially binds chlorophyll b in vitro. *Planta*, 238:1065–1080, 2013.

[25] S. Takahashi et al. The c-terminal extension peptide of non-photoconvertible water-soluble chlorophyll-binding proteins (class ii wscps) affects their solubility and stability: Comparative analyses of the biochemical and chlorophyll-binding properties of recombinant brassica, raphanus and lepidium wscps with or without their c-terminal extension peptides. *Protein J*, 33(1):75–84, 2014.

[26] B.D. Kohorn et al. Replacement of histidines of light harvesting chlorophyll a/b binding protein ii disrupts chlorophyll-protein complex assembly. *Plant Physiology*, 93(1):339–342, 1990.

[27] K. Schmidt et al. Recombinant water-soluble chlorophyll protein from brassica oleracea var. botrys binds various chlorophyll derivatives. *Biochemistry*, 42(24), 2003.

[28] D. M. Palm et al. Water-soluble chlorophyll protein (wscp) stably binds two or four chlorophylls. *Biochemistry*, 56:1726–1736, 2017.

[29] S. Takahashi et al. Isolation, properties and a possible function of a water-soluble chlorophyll a/b-protein from brussels sprouts. *Plant and Cell Physiology*, 38(2):133–138, 1997.

[30] Agostini A. et al. An unusual role for the phytyl chains in the photoprotection of the chlorophylls bound to water-soluble chlorophyll-binding proteins. *Scientific Reports*, 7(7504), 2017.

[31] O. Lemke et al. On the stability of the water-soluble chlorophyll-binding protein (wscp) studied by molecular dynamics simulations. *The Journal of Physical Chemistry B*, 123(50):10594–10604, 2019.

[32] G. Haran et al. How fast are the motions of tertiary-structure elements in proteins? *J. Chem. Phys.*, 153(13), 2020.

[33] A. N. Naganathan et al. Scaling of folding times with protein size. *J. Am. Chem. Soc.*, 127(2):480–481, 2005.

[34] K. A. Dill et al. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[35] E. P. O'Brien et al. Effects of ph on proteins: Predictions for ensemble and single-molecule pulling experiments. *J. Am. Chem. Soc.*, 134(2):979–987, 2012.

[36] F. Cazals et al. Spectral techniques to explore point clouds in euclidean space,with applications to collective coordinates in structural biology. *Nonlinear Computational Geometry*, 2009.

[37] J. N. Onuchic et al. Theory of protein folding: the energy landscape perspective. *Phys. Chem.*, 48:545–600, 1997.

[38] T. P. J. Krüger et al. Fluorescence spectral dynamics of single lhcii trimers. *Biophys J.*, 98(12):3093–3101, 2010.

[39] H. Frauenfelder et al. Conformational substates in proteins. *Biophysics and Biophysical Chemistry*, 17:451–479, 1988.

[40] P. W. Anderson et al. Anomalous low-temperature thermal properties of glasses and spin glasses. *Structure and Properties of Condensed Matter*, 25, 1972.

[41] J. J. Sakurai et al. Modern quantum mechanics. 1985.

[42] M. Najafi et al. Spectral hole burning, recovery, and thermocycling in chlorophyll protein complexes: Distributions of barriers on the protein energy landscape. *J. Phys. Chem. B*, 116:11780–11790, 2012.

[43] M. Najafi et al. Conformational changes in pigmentprotein complexes at low temperaturesspectral memory and a possibility of cooperative effects. *J. Phys. Chem. B*, 119:6930–6940, 2015.

[44] A. Levenberg et al. Probing energy landscapes of cytochrome b6 f with spectral hole burning: Effects of deuterated solvent and detergent. *J. Phys. Chem. B*, 121(42):9848–9858, 2017.

[45] G. Shafiei et al. Evidence of simultaneous spectral hole burning involving two tiers of the protein energy landscape in cytochrome b6 f. *J. Phys. Chem. B*, 123:10930–10938, 2019.

[46] R. Purchase et al. Spectral hole burning: examples from photosynthesis. *Photosynth Res.*, 101(2–3):245–266, 2009.

[47] T.L. Blundell et al. Protein crystallography. *London: Academic Press*, 1976.

[48] J.D. Roberts et al. Nuclear magnetic resonance spectroscopy. *J. Chem. Educ.*, 38(11):581, 1961.

[49] A. Hoenger et al. Cellular tomography. *Adv Protein Chem Struct Biol.*, 82:67–90, 2011.

[50] D. Frenkel et al. Understanding molecular simulation: From algorithms to applications. 1996.

[51] A. Lamiable et al. Pep-fold3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic Acids Research*, 44(W1):W449–W454, 2016.

[52] S. Ó. Conchúir et al. A web resource for standardized benchmark datasets, metrics, and rosetta protocols for macromolecular modeling and design. *PLoS One*, 10(9), 2015.

[53] J. S. Rowlinson et al. The maxwell–boltzmann distribution. *Molecular Physics. An International Journal at the Interface Between Chemistry and Physics*, 103(21-23), 2005.

[54] S. Y. Park et al. Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics*, 150(2):219–230, 2009.

[55] W. Andreoni et al. Dft-based molecular dynamics as a new tool for computational biology: First applications and perspective. *Ibm Journal of Research and Development*, 45(3.4):397–407, 2001.

[56] R. Armunanto et al. Classical and qm/mm molecular dynamics simulations of co2+ in water. *Chemical Physics*, 295(1):63–70, 2003.

[57] X. Zu et al. Recent developments and applications of the charmm force fields. *Wiley Interdiscip Rev Comput Mol Sci.*, 2(1):167–185, 2012.

[58] J.E. Lennard-Jones et al. On the determination of molecular fields. ii. from the equation of state of a gas. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 106(734):463–477, 1924.

[59] V.A. Parsegian et al. Forces: A handbook for biologists, chemists, engineers, and physicists. *Cambridge Univ Press*, 2006.

[60] Comparison between a cubic box and a rhombic dodecahedron box. https://manual.gromacs.org/documentation/2021/reference-manual/algorithms/periodic-boundary-conditions.html.

[61] T. Schlick et al. Biomolecular dynamics at long timesteps: bridging the timescale gap between simulation and experimentation. *Annu. Rev. Biophys. Biomol. Struct.*, 26:181–222, 1997.

[62] J. C. Phillips et al. Scalable molecular dynamics with namd. *J. Comput. Chem.*, 26(16):1781–1802, 2005.

[63] G. Bussi et al. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126, 2007.

[64] H. J. C. Berendsen et al. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[65] R. W. Hockney et al. Quiet high resolution computer models of a plasma. *J. Comp. Phys.*, 14:148–158, 1974.

[66] B. Knapp et al. Avoiding false positive conclusions in molecular simulation: The importance of replicas. *J. Chem. Theory Comput.*, 14:6127–6138, 2018.

[67] A. Grossfield et al. Quantifying uncertainty and sampling quality in biomolecular simulations. *Annual Reports in Computational Chemistry*, 1(5):23–48, 2009.

[68] M. C. R. Melo et al. Generalized correlation-based dynamical network analysis: a new high-performance approach for identifying allosteric communications in molecular dynamics trajectories. *J. Chem. Phys.*, 153, 2020.

[69] Blondel. V. D. et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[70] U. Brandes et al. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[71] T. H. Cormen et al. ”section 24.3: Dijkstra’s algorithm”. introduction to algorithms (second ed.). *MIT Press*, pages 595–601, 2001.

[72] F. B. Best et al. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi1 and chi2 dihedral angles. *Journal of Chemical Theory and Computation*, 8:3257–3273, 2012.

[73] F. Guerra et al. Revised force-field parameters for chlorophyll-a, pheophytin-a andplastoquinone-9. *Journal of Molecular Graphics and Modelling*, 58:30–39, 2015.

[74] N. Foloppe et al. Potential energy function for photosynthetic reaction centre chromophores: energy minimisations of a crystalline bacteriophytin. *Plenum Press, New York*, 1992.

[75] K. Claridge et al. Developing consistent molecular dynamics force fields for biological chromophores via force matching. *The Journal of Physical Chemistry B*, 123(2):428–438, 2019.

[76] K. Kuczera et al. Temperature dependence of the structure and dynamics of myoglobin. *J. Mol. Biol.*, 213:351–373, 1990.

[77] A. Damjanovic et al. Excitons in a photosynthetic light-harvesting system: a combined molecular dynamics, quantum chemistry, and polaron model study. *Phys. Rev. E*, 65:351–373, 2002.

[78] N. Foloppe et al. Structural model of the photosynthetic reaction center of rhodobacter capsulatus. *Proteins*, 2(3):226–244, 1995.

[79] C. Vega et al. The melting temperature of the most common models of water. *J. Chem. Phys.*, 122(11), 2005.

[80] T. Kumagai et al. Direct observation and control of hydrogen-bonddynamics using low-temperature scanningtunneling microscopy. *Progress in Surface Science*, 90(13):239–291, 2015.

[81] I. Tavernelli et al. Protein dynamics, thermal stability, and free-energy landscapes: a molecular dynamics investigation. *E. E. Biophys. J.*, 85(4):2641–2649, 2003.

[82] K. Okazaki et al. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, 103(32):11844–11849, 2006.

[83] M.M Seibert et al. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *J. Mol. Biol.*, 354(1):173–183, 2005.

[84] M. Najafi et al. Monte carlo modeling of spectral diffusion employing multiwell protein energy landscapes: Application to pigmentprotein complexes involved in photosynthesis. *J. Phys. Chem. B*, 119(25):7911–7921, 2015.

[85] A. Trempe. The effect of triplet states on non-photochemical hole burning in cytochrome b6f and modified lh2 complex. (master's thesis). *Concordia University*, 2021.

[86] Chemical structure of chlorophyll a. https://commons.wikimedia.org/wiki/File:Chlorophyll_a.svg?uselang=en#Licensing.

[87] Rmsd definition in gromacs 2021.4. https://manual.gromacs.org/current/reference-manual/analysis/rmsd.html.

[88] I. Kufareva et al. Methods of protein structure comparison. *Methods Mol Biol.*, 857:231–257, 2012.

[89] Rmsf definition in gromacs 2021.4. https://manual.gromacs.org/current/onlinehelp/gmx-rmsf.html.

[90] R. Feynman et al. Statistical mechanics: A set of lectures. 1998.

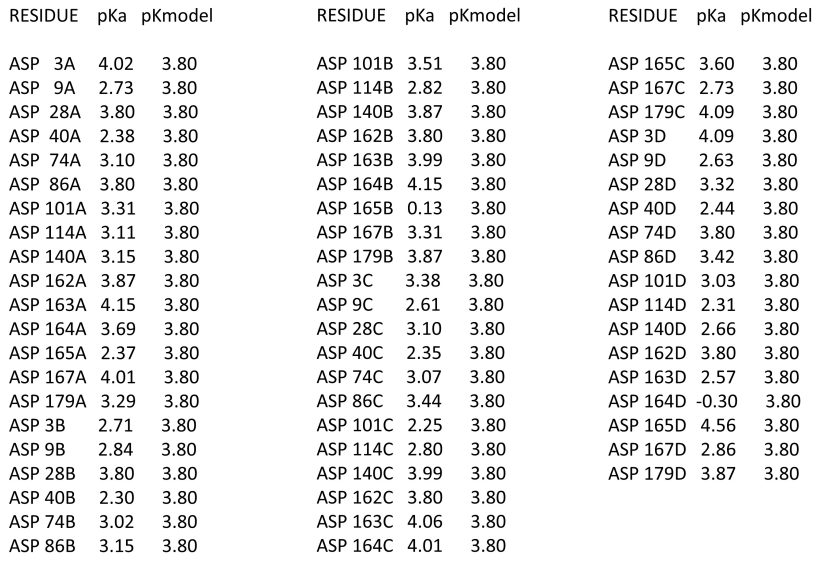
[91] S. Garashchuk et al. Calculation of the quantum-mechanical tunneling in bound potentials. *J. Theor. Chem.*, 10:1–11, 2014.

Appendix A

Supporting information for Chapter 4

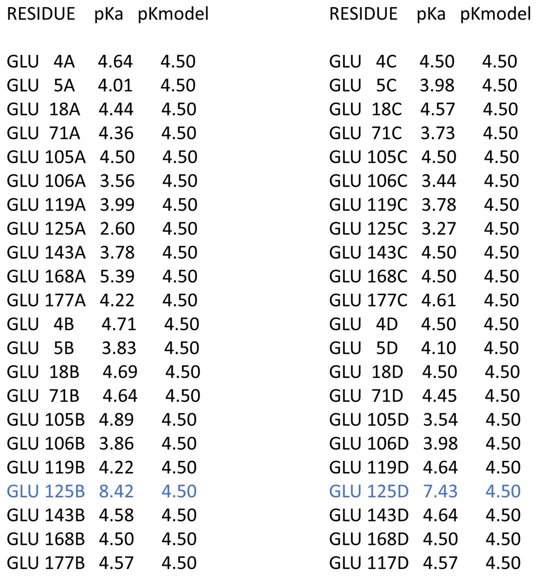
```
1                 10                  20                  30                  40
I N D E E P V K D T N G N P L K I E T R Y F I Q P A S D N N G G G L V P A N V D L S H L C
        50                  60                  70                  80                  90
P L G I V R T S L P Y Q P G L P V T I S T P S S S E G N D V L T N T N I A I T F D A P I W
                  100                 110                 120                 130
L C P S S K T W T V D S S S E E K Y I I T G G D P K S G E S F F R I E K Y G N G K N T Y K
        140                 150                 160                 170                 180
L V R Y D N G E G K S V G S T K S L W G P A L V L N D D D D S D E N A F P I K F R E V D T
```

Figure A.1: Protein sequence of one protein chain of the homotetramer Water-soluble chlorophyll protein, from *Lepidium virginicum*, PDB ID : 2DRE. Acidic amino acids Glumatic acid (E), and Aspartic acid (D) are colored in blue, and are considered deprotonated (negatively charged). Basic amino acids Lysine (K), Arginine (R), and Histidine (H) are colored in red, and are considered protonated (positively charged).



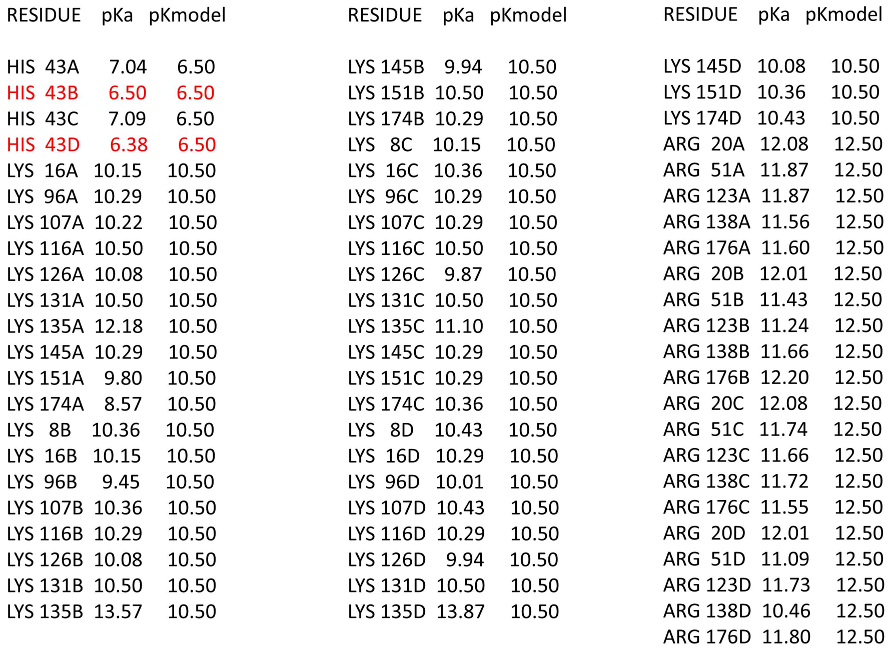
RESIDUE	pKa	pKmodel	RESIDUE	pKa	pKmodel	RESIDUE	pKa	pKmodel
ASP 3A	4.02	3.80	ASP 101B	3.51	3.80	ASP 165C	3.60	3.80
ASP 9A	2.73	3.80	ASP 114B	2.82	3.80	ASP 167C	2.73	3.80
ASP 28A	3.80	3.80	ASP 140B	3.87	3.80	ASP 179C	4.09	3.80
ASP 40A	2.38	3.80	ASP 162B	3.80	3.80	ASP 3D	4.09	3.80
ASP 74A	3.10	3.80	ASP 163B	3.99	3.80	ASP 9D	2.63	3.80
ASP 86A	3.80	3.80	ASP 164B	4.15	3.80	ASP 28D	3.32	3.80
ASP 101A	3.31	3.80	ASP 165B	0.13	3.80	ASP 40D	2.44	3.80
ASP 114A	3.11	3.80	ASP 167B	3.31	3.80	ASP 74D	3.80	3.80
ASP 140A	3.15	3.80	ASP 179B	3.87	3.80	ASP 86D	3.42	3.80
ASP 162A	3.87	3.80	ASP 3C	3.38	3.80	ASP 101D	3.03	3.80
ASP 163A	4.15	3.80	ASP 9C	2.61	3.80	ASP 114D	2.31	3.80
ASP 164A	3.69	3.80	ASP 28C	3.10	3.80	ASP 140D	2.66	3.80
ASP 165A	2.37	3.80	ASP 40C	2.35	3.80	ASP 162D	3.80	3.80
ASP 167A	4.01	3.80	ASP 74C	3.07	3.80	ASP 163D	2.57	3.80
ASP 179A	3.29	3.80	ASP 86C	3.44	3.80	ASP 164D	-0.30	3.80
ASP 3B	2.71	3.80	ASP 101C	2.25	3.80	ASP 165D	4.56	3.80
ASP 9B	2.84	3.80	ASP 114C	2.80	3.80	ASP 167D	2.86	3.80
ASP 28B	3.80	3.80	ASP 140C	3.99	3.80	ASP 179D	3.87	3.80
ASP 40B	2.30	3.80	ASP 162C	3.80	3.80			
ASP 74B	3.02	3.80	ASP 163C	4.06	3.80			
ASP 86B	3.15	3.80	ASP 164C	4.01	3.80			

(a) Estimations of the pKa values for the acidic ASP residues in the homotetramer Water-soluble chlorophyll protein, from *Lepidium virginicum*, PDB ID : 2DRE, obtained from the PropKa website. The first column lists the residues, the second column the estimations of the pKa values, and the third column the theoretical pKa values.



RESIDUE	pKa	pKmodel	RESIDUE	pKa	pKmodel
GLU 4A	4.64	4.50	GLU 4C	4.50	4.50
GLU 5A	4.01	4.50	GLU 5C	3.98	4.50
GLU 18A	4.44	4.50	GLU 18C	4.57	4.50
GLU 71A	4.36	4.50	GLU 71C	3.73	4.50
GLU 105A	4.50	4.50	GLU 105C	4.50	4.50
GLU 106A	3.56	4.50	GLU 106C	3.44	4.50
GLU 119A	3.99	4.50	GLU 119C	3.78	4.50
GLU 125A	2.60	4.50	GLU 125C	3.27	4.50
GLU 143A	3.78	4.50	GLU 143C	4.50	4.50
GLU 168A	5.39	4.50	GLU 168C	4.50	4.50
GLU 177A	4.22	4.50	GLU 177C	4.61	4.50
GLU 4B	4.71	4.50	GLU 4D	4.50	4.50
GLU 5B	3.83	4.50	GLU 5D	4.10	4.50
GLU 18B	4.69	4.50	GLU 18D	4.50	4.50
GLU 71B	4.64	4.50	GLU 71D	4.45	4.50
GLU 105B	4.89	4.50	GLU 105D	3.54	4.50
GLU 106B	3.86	4.50	GLU 106D	3.98	4.50
GLU 119B	4.22	4.50	GLU 119D	4.64	4.50
GLU 125B	8.42	4.50	GLU 125D	7.43	4.50
GLU 143B	4.58	4.50	GLU 143D	4.64	4.50
GLU 168B	4.50	4.50	GLU 168D	4.50	4.50
GLU 177B	4.57	4.50	GLU 117D	4.57	4.50

(b) Estimations of the pKa values for the acidic GLU residues in the homotetramer Water-soluble chlorophyll protein, from *Lepidium virginicum*, PDB ID : 2DRE, obtained from the PropKa website. The first column lists the residues, the second column the estimations of the pKa values, and the third column the theoretical pKa values. GLU 125 in protein chains B and D (in blue) were kept deprotonated in the simulations but could have been kept protonated at pH = 7 based on the estimated pKa values.



RESIDUE	pKa	pKmodel	RESIDUE	pKa	pKmodel	RESIDUE	pKa	pKmodel
HIS 43A	7.04	6.50	LYS 145B	9.94	10.50	LYS 145D	10.08	10.50
HIS 43B	6.50	6.50	LYS 151B	10.50	10.50	LYS 151D	10.36	10.50
HIS 43C	7.09	6.50	LYS 174B	10.29	10.50	LYS 174D	10.43	10.50
HIS 43D	6.38	6.50	LYS 8C	10.15	10.50	ARG 20A	12.08	12.50
LYS 16A	10.15	10.50	LYS 16C	10.36	10.50	ARG 51A	11.87	12.50
LYS 96A	10.29	10.50	LYS 96C	10.29	10.50	ARG 123A	11.87	12.50
LYS 107A	10.22	10.50	LYS 107C	10.29	10.50	ARG 138A	11.56	12.50
LYS 116A	10.50	10.50	LYS 116C	10.50	10.50	ARG 176A	11.60	12.50
LYS 126A	10.08	10.50	LYS 126C	9.87	10.50	ARG 20B	12.01	12.50
LYS 131A	10.50	10.50	LYS 131C	10.50	10.50	ARG 51B	11.43	12.50
LYS 135A	12.18	10.50	LYS 135C	11.10	10.50	ARG 123B	11.24	12.50
LYS 145A	10.29	10.50	LYS 145C	10.29	10.50	ARG 138B	11.66	12.50
LYS 151A	9.80	10.50	LYS 151C	10.29	10.50	ARG 176B	12.20	12.50
LYS 174A	8.57	10.50	LYS 174C	10.36	10.50	ARG 20C	12.08	12.50
LYS 8B	10.36	10.50	LYS 8D	10.43	10.50	ARG 51C	11.74	12.50
LYS 16B	10.15	10.50	LYS 16D	10.29	10.50	ARG 123C	11.66	12.50
LYS 96B	9.45	10.50	LYS 96D	10.01	10.50	ARG 138C	11.72	12.50
LYS 107B	10.36	10.50	LYS 107D	10.43	10.50	ARG 176C	11.55	12.50
LYS 116B	10.29	10.50	LYS 116D	10.29	10.50	ARG 20D	12.01	12.50
LYS 126B	10.08	10.50	LYS 126D	9.94	10.50	ARG 51D	11.09	12.50
LYS 131B	10.50	10.50	LYS 131D	10.50	10.50	ARG 123D	11.73	12.50
LYS 135B	13.57	10.50	LYS 135D	13.87	10.50	ARG 138D	10.46	12.50
						ARG 176D	11.80	12.50

(c) Estimations of the pKa values for the basic HIS, LYS, and ARG residues in the homotetramer Water-soluble chlorophyll protein, from *Lepidium virginicum*, PDB ID : 2DRE, obtained from the PropKa website. The first column lists the residues, the second column the estimations of the pKa values, and the third column the theoretical pKa values. HIS 43 in protein chains B and D (in red) were kept protonated in the simulations but could have been kept deprotonated at pH = 7 based on the estimated pKa values.

Figure A.2: Estimations of the pKa values for the acidic ASP and GLU residues, and the basic HIS, LYS, and ARG residues in the homotetramer Water-soluble chlorophyll protein, from *Lepidium virginicum*, PDB ID : 2DRE, obtained from the PropKa website. The first column lists the residues, the second column the estimations of the pKa values, and the third column the theoretical pKa values.

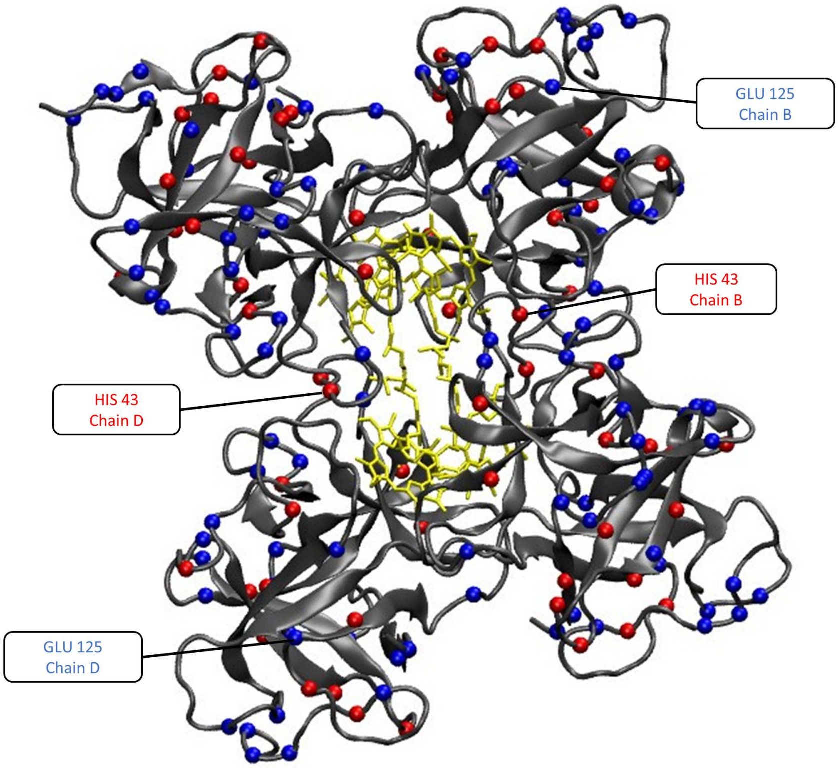


Figure A.3: Crystral structure of the homotetramer Water-soluble chlorophyll protein, from *Lepidium virginicum*, PDB ID : 2DRE. Chlorophylls a are colored in yellow, and the four protein chains in grey. Acidic amino acids are colored in blue and are considered negatively charged, basic amino acids are colored in red and are considered positively charged.

Appendix B

Supporting information for Chapter 5

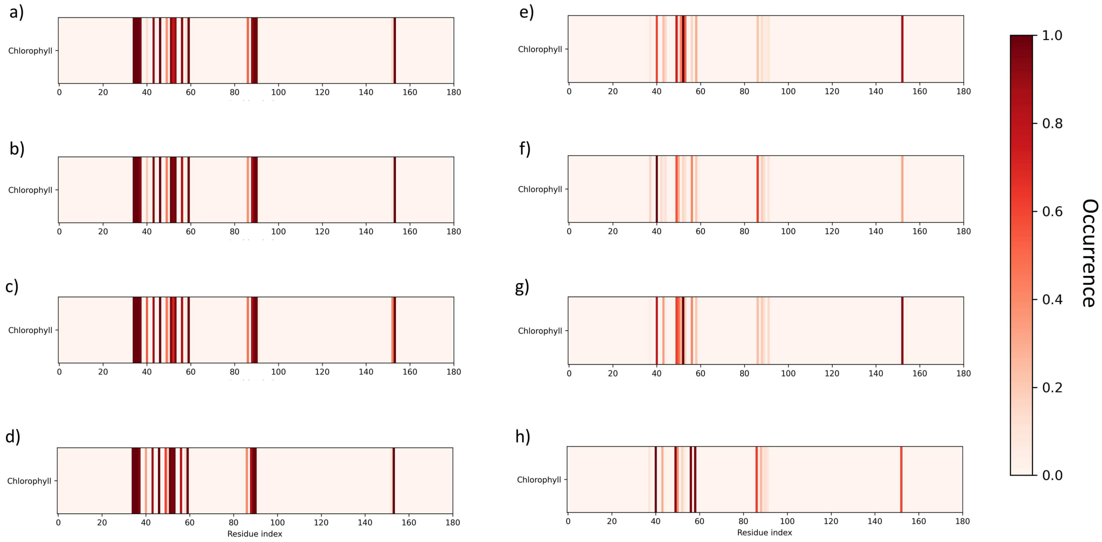


Figure B.1: Heatmaps of the mean contact occurrence between chlorophylls and their respective protein chains *a)* A, *b)* B, *c)* C, and *d)* D at 300 K. The heatmaps of the respective occurrence standard errors are presented in *e)* for protein chain A, *f)* for B, *g)* for C, and *h)* for D.

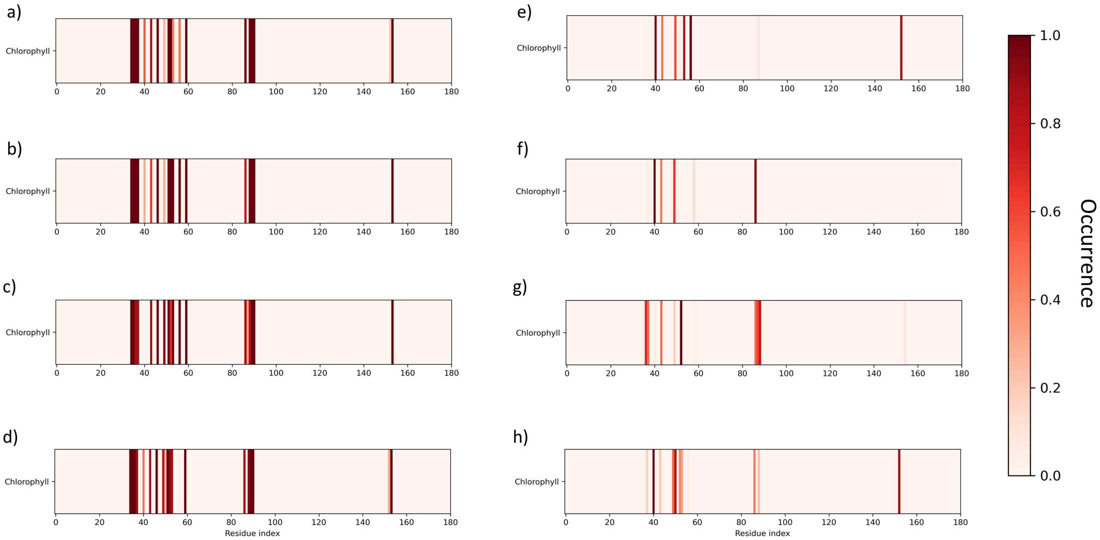
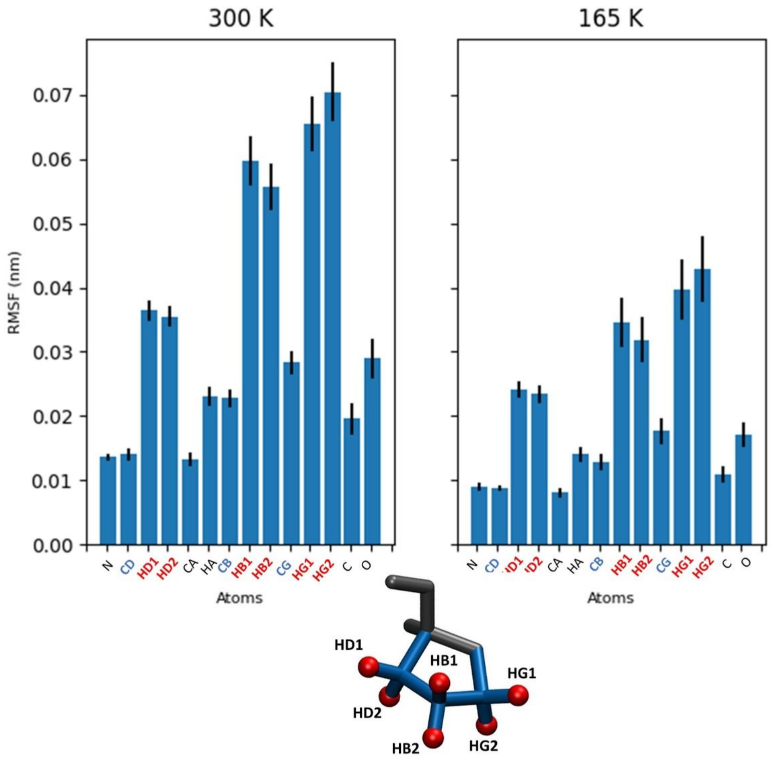
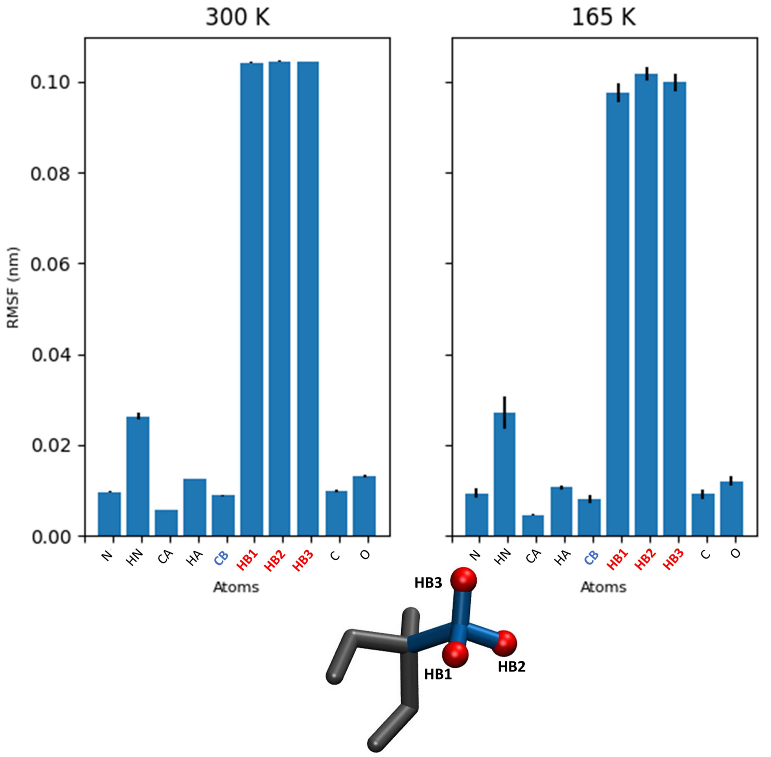


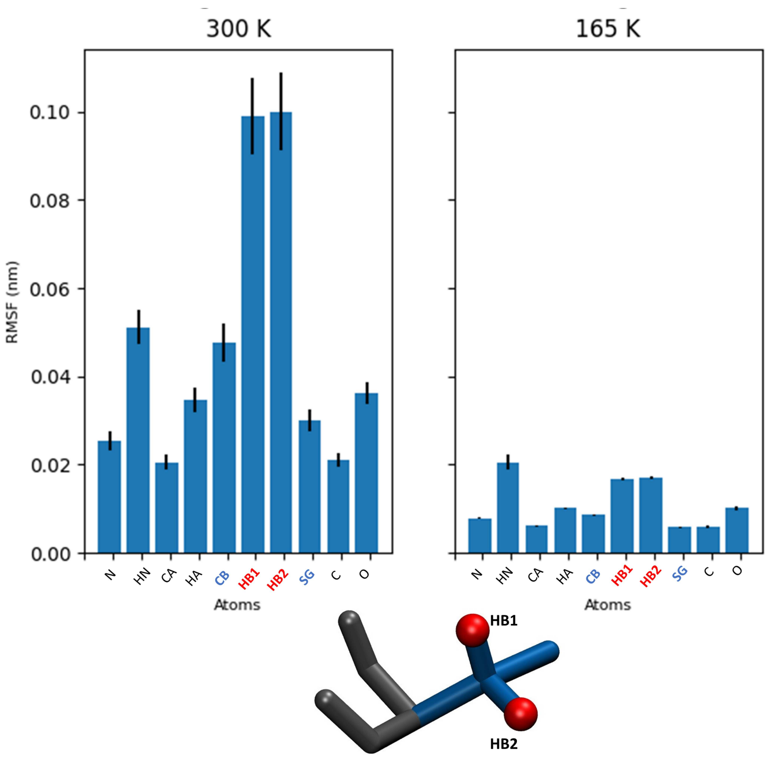
Figure B.2: Heatmaps of the mean contact occurrence between chlorophylls and their respective protein chains *a)* A, *b)* B, *c)* C, and *d)* D at 165 K. The heatmaps of the respective occurrence standard errors are presented in *e)* for protein chain A, *f)* for B, *g)* for C, and *h)* for D.



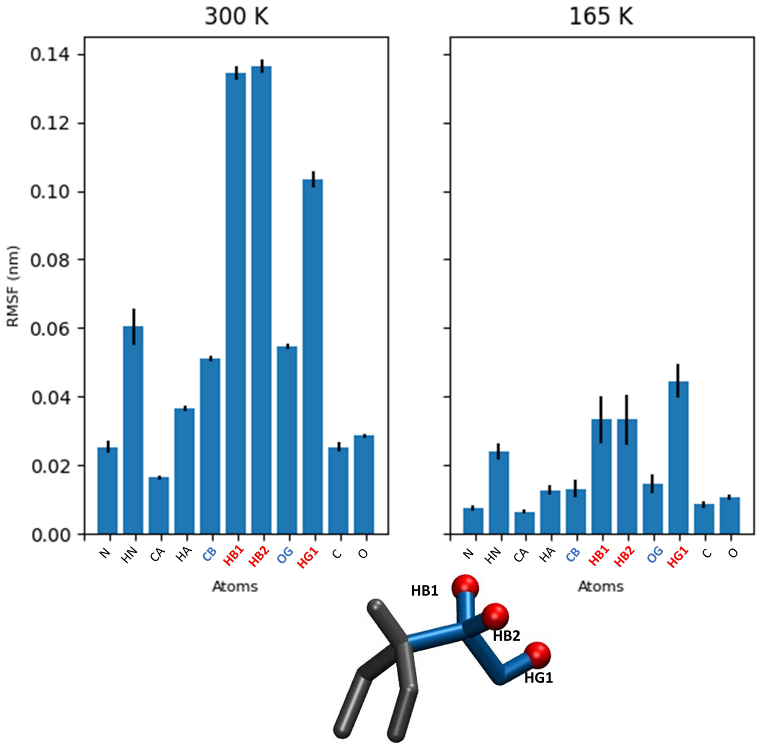
(a) Average RMSF values of the 4 Proline residues, at 300 K and 165 K.



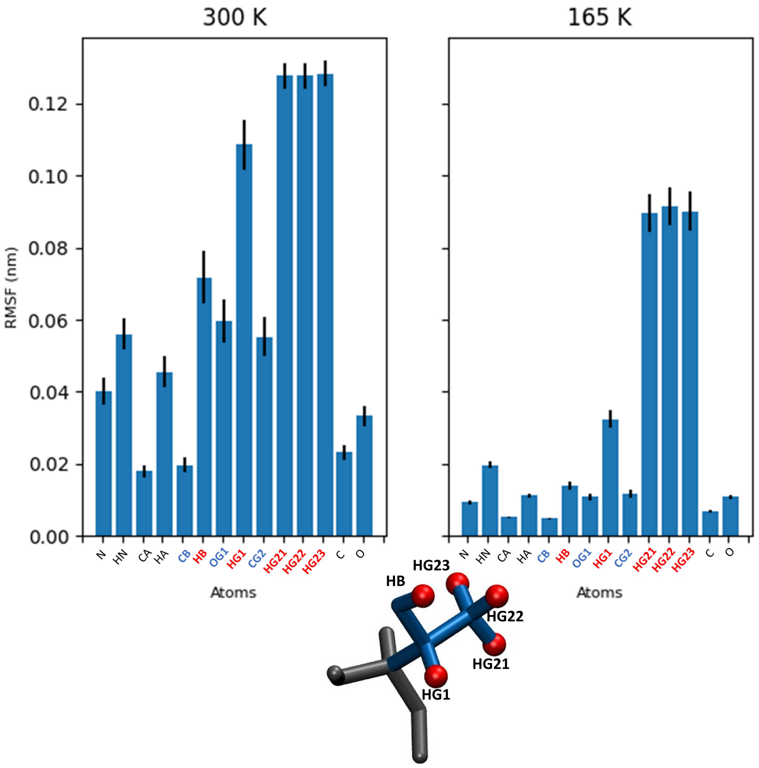
(b) Average RMSF values of the Alanine residue, at 300 K and 165 K.



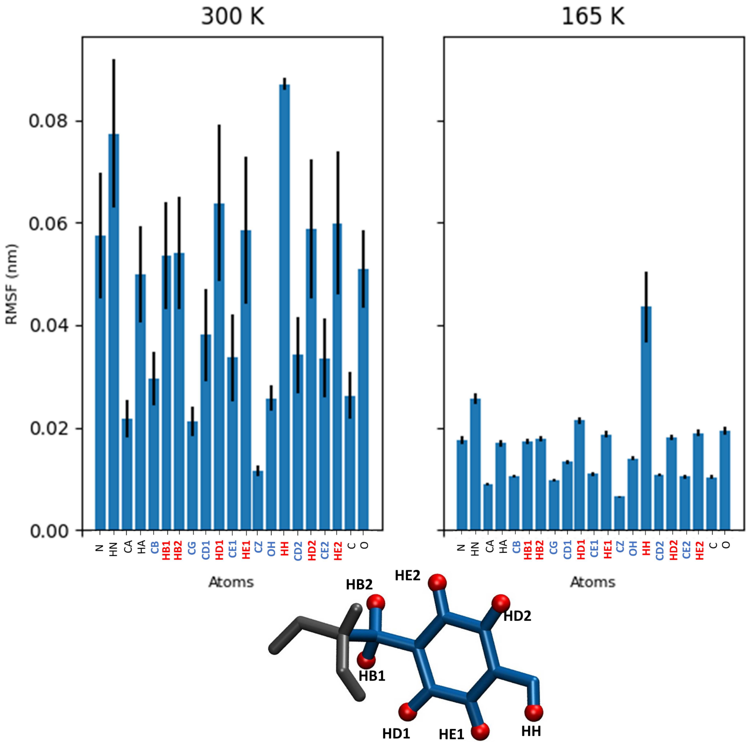
(c) Average RMSF values of the 5 Cysteine residues, at 300 K and 165 K.



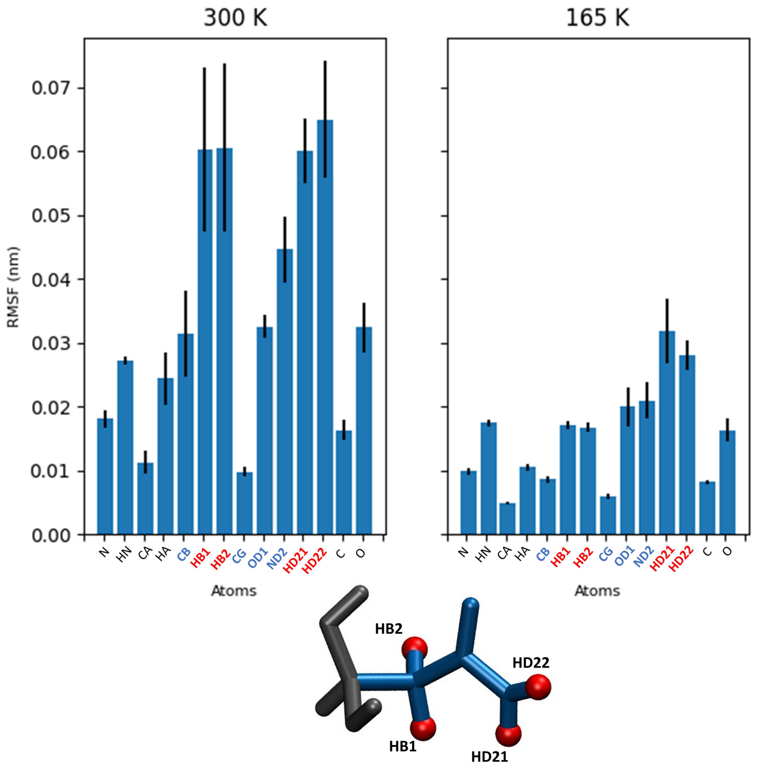
(d) Average RMSF values of the 3 Serine residues, at 300 K and 165 K.



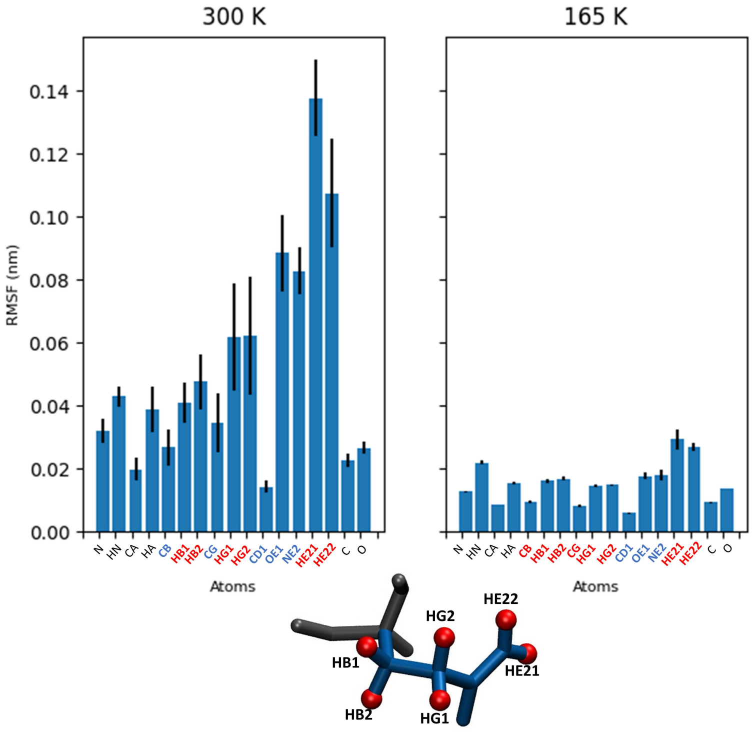
(e) Average RMSF values of the 4 Threonine residues, at 300 K and 165 K.



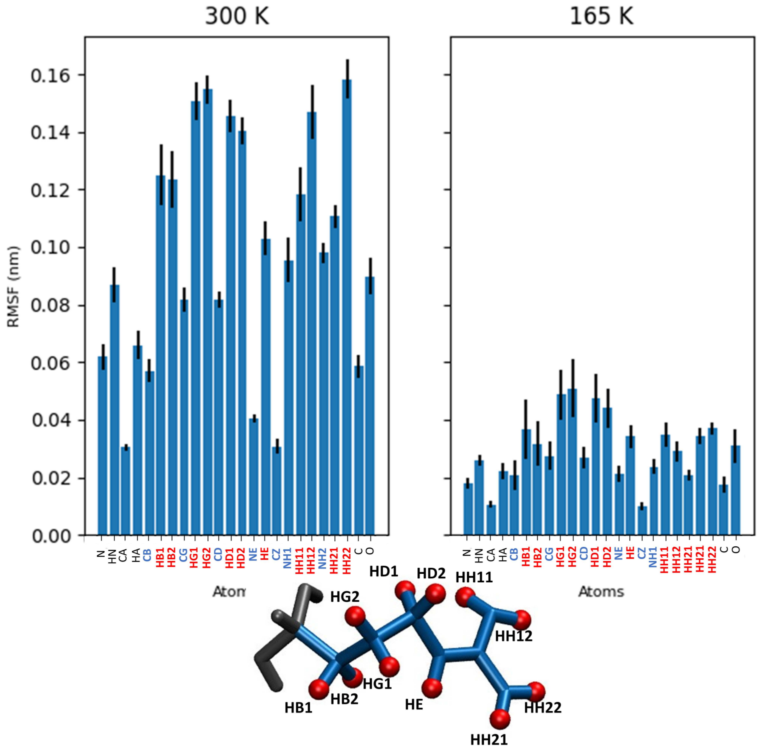
(f) Average RMSF values of the 3 Tyrosine residues, at 300 K and 165 K.



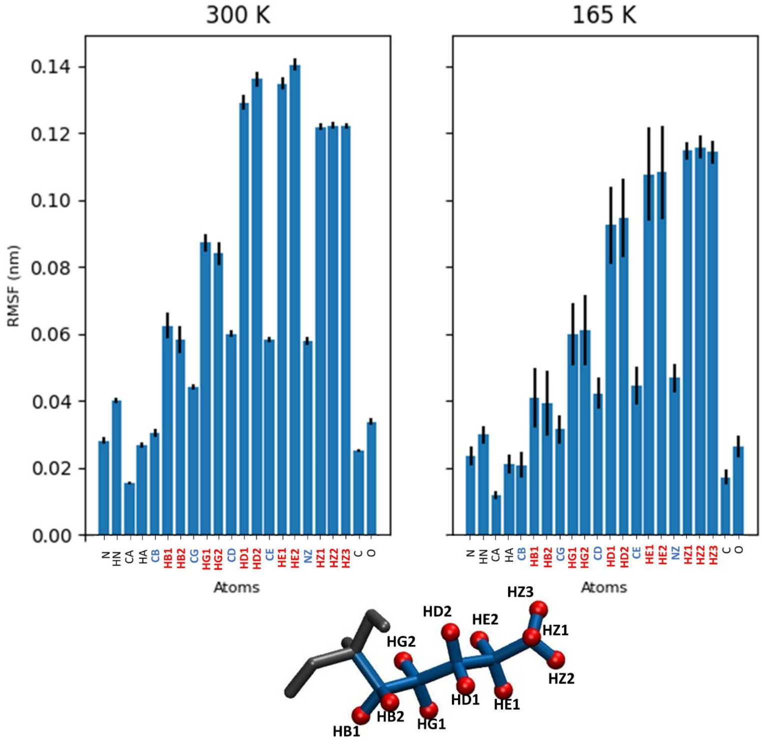
(g) Average RMSF values of the 3 Asparagine residues, at 300 K and 165 K.



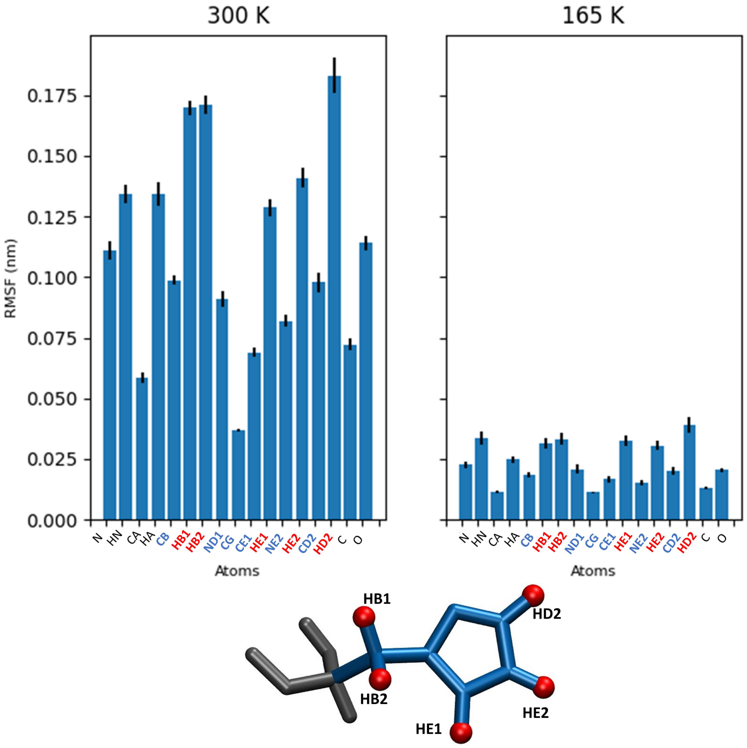
(h) Average RMSF values of the Glutamine residue, at 300 K and 165 K.



(i) Average RMSF values of the Arginine residues, at 300 K and 165 K.



(j) Average RMSF values of the Lysine residue, at 300 K and 165 K.



(k) Average RMSF values of the 2 Histidine residues, at 300 K and 165 K.

Figure B.3: Average RMSF values classified by residue types, at 300 K and 165 K. Backbone atoms are labeled and represented in black, side chain hydrogen atoms in red, and side chain non-hydrogen atoms in blue.

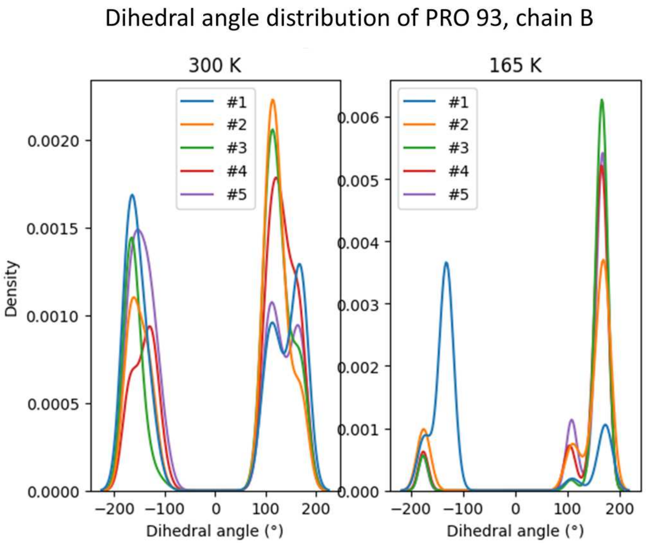


Figure B.4: Dihedral angle distributions of PRO 93, in chain B for all 5 replicas, at 300 K and 165 K.

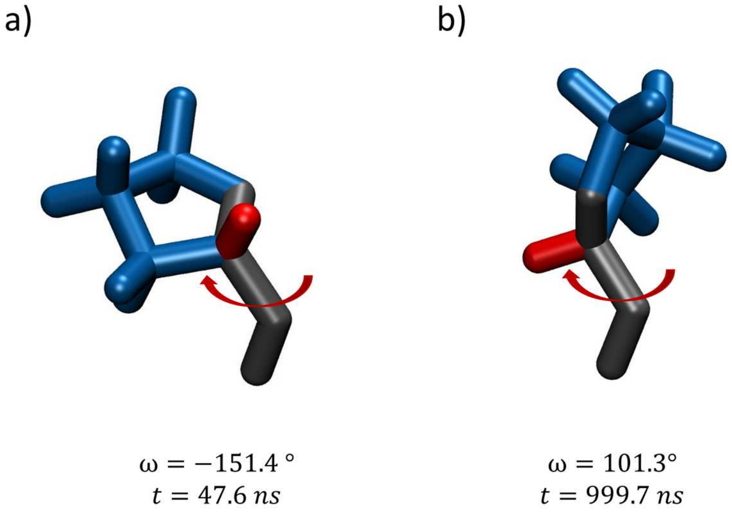


Figure B.5: Graphical representations of PRO 93 in chain B, in replica #2 of the 300 K simulation at *a)* $t = 47.6\ ns$ with $\omega = -151.4°$, *b)* $t = 999.7\ ns$ with $\omega = 101.3°$. Heavy backbone atoms are depicted in black, the hydrogen backbone atom in red, the side chain pyrrolidine ring in blue. The rotation of the dihedral angle is represented by a red arrow.