Graph-Theory Algorithms for Dynamic Hydrogen-Bonded Networks in Proteins and Lipid Membranes

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(*doctor rerum naturalium*)

am Fachbereich Physik
der Freien Universität Berlin

vorgelegt von

Konstantina Karathanou

Berlin, 2023

1. Erstgutachterin: Prof. Dr. Ana-Nicoleta Bondar
2. Zweitgutachter: Prof. Dr. Joachim Heberle

Tag der Disputation: 04.03.2024

“Η αρχή της σοφίας είναι η αναζήτηση”

Σωκράτης

“Wisdom begins in wonder”

Socrates

“Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less”

Marie Curie

To my family.
Thank you for everything.

Στην οικογένειά μου.
Σας ευχαριστώ για όλα.

Abstract

Computer simulations can give essential insights into the dynamics of biomolecular systems but raise significant big data challenges still to be sorted out. To overcome the challenge of large data sets combined with the complexity of biomolecular interactions, I implemented a set of robust algorithms, as part of this doctoral thesis, inspired by graph theory that allows us to use large data sets from atomistic molecular dynamics (MD) simulations and derive simple graphical representations of the hydrogen bond (H-bond) networks of lipid membrane models, proteins in different intermediate states, and of the response of the proteins to mutations. These representations are valuable for the interpretation of data from experiments and computations.

Our algorithms facilitate highly efficient analyses of dynamic H-bond networks at the lipid membrane interface. We introduce the implementation of a Connected Components algorithm to cluster lipid molecules and a Depth First Search (DFS) algorithm that allows us to characterize the topology of dynamic H-bond clusters sampled by lipid headgroups in MD simulations. With the algorithm we developed, we identify the transient sampling of four main types of lipid H-bond clusters: linear, star, circular and extensive networks combining these topologies. Water bridges between lipid headgroups are dynamic with lifetimes lasting for a few picoseconds.

Our algorithms are further extended to study conformational dynamics in proteins. An example is SecA, a protein motor that couples Adenosine triphosphate (ATP) binding and hydrolysis with the pre-protein substrate's translocation through the membrane embedded SecYEG protein translocon. However, the exact mechanism of SecA’s conformational coupling remains unclear. We present a methodology of applying graph-based approaches to characterize the dynamics of the SecA protein motor by computing long-distance H-bond pathways that inter-connect the nucleotide-binding pocket and the pre-protein binding site, shortest-distance routes and centrality measures that reveal amino acids with a central role in the total connectivity of the protein graph. A key finding enabled by the graph-based approach developed as part of this doctoral thesis is that mutations near the nucleotide-binding site associate with modified dynamics at the pre-protein binding domain. Water molecules participate in extended H-bonded water chains contributing to long-distance conformational coupling. Our methodologies are also applied to protein VASA, a DEAD-box enzyme involved in the cell cycle with ATP and Ribonucleic Acid (RNA) binding sites and explore the conformational coupling between the two binding sites and Channelrhodopsin’s C1C2 lipid-protein H-bond molecular dynamics.

Lastly, our algorithms are applied to the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV-2) protein S crystal structures. Protein S undergoes conformational changes and symmetry loss of core H-bonded clusters as it transitions from the closed to the pre-fusion conformation. Our study has identified N501 as a central residue of the H-bond network that interconnects the spike protein S to Angiotensin-Converting Enzyme 2 (ACE2), and that subsequently became mutated into TYR in a new COVID-19 variant.

Zusammenfassung

Computersimulationen können wesentliche Einblicke in die Dynamik biomolekularer Systeme geben, werfen aber auch erhebliche Herausforderungen in Bezug auf große Datenmengen auf, die noch zu bewältigen sind. Um die Herausforderung großer Datenmengen in Verbindung mit der Komplexität biomolekularer Wechselwirkungen zu bewältigen, habe ich im Rahmen dieser Doktorarbeit eine Reihe robuster Algorithmen implementiert, die von der Graphentheorie inspiriert sind und es uns ermöglichen, große Datenmengen aus atomistischen Molekulardynamiksimulationen (MD-Simulationen) zu verwenden und einfache grafische Darstellungen der Wasserstoffbrückenbindungen (H-Bindungen) von Lipidmembranmodellen, Proteinen in verschiedenen Zwischenzuständen und der Reaktion der Proteine auf Mutationen abzuleiten. Diese Darstellungen sind wertvoll für die Interpretation von Daten aus Experimenten und Berechnungen.

Unsere Algorithmen ermöglichen hocheffiziente Analysen von dynamischen H-Bindungsnetzwerken an der Grenzfläche von Lipidmembranen. Wir stellen die Implementierung eines Algorithmus für verbundene Komponenten zum Clustern von H-gebundenen Lipidmolekülen und einen DFS-Algorithmus (Depth First Search) vor, der es uns ermöglicht, die Topologie von dynamischen H-Bindungsclustern zu charakterisieren, die von Lipidkopfgruppen in MD-Simulationen gesampelt werden. Mit dem von uns entwickelten Algorithmus identifizieren wir die vorübergehenden Probenahmen von vier Haupttypen von Lipid-H-Bindungsclustern: lineare, sternförmige, zirkuläre und umfangreiche Netzwerke, die diese Topologien kombinieren. Wasserbrücken zwischen Lipid-Kopfgruppen sind dynamisch und haben eine Lebensdauer in einer Größenordnung von Pikosekunden.

Unsere Algorithmen werden weiter ausgebaut, um die Konformationsdynamik von Proteinen zu untersuchen. Ein Beispiel ist SecA, ein Proteinmotor, der die Bindung und Hydrolyse von Adenosintriphosphat (ATP) mit der Translokation des Präproteinsubstrats durch das in die Membran eingebettete SecYEG-Protein-Translokon verbindet. Der genaue Mechanismus der SecA-Konformationskopplung bleibt jedoch unklar. Wir stellen eine Methode zur Anwendung graphbasierter Ansätze vor, um die Dynamik des SecA-Proteinmotors zu charakterisieren, indem wir die langen H-Bindungen, die die Nukleotid-Bindungstasche und die Prä-Protein-Bindungsstelle miteinander verbinden, sowie die kürzesten Entfernungen und Zentralitätsmaße berechnen, die die Aminosäuren mit einer zentralen Rolle in der Gesamtkonnektivität des Proteingraphen aufzeigen. Eine wichtige Erkenntnis, die durch den im Rahmen dieser Doktorarbeit entwickelten graphbasierten Ansatz ermöglicht wurde, ist, dass Mutationen in der Nähe der Nukleotid-Bindungsstelle mit einer veränderten Dynamik im Bereich der Prä-Proteinbindung einhergehen. Wassermoleküle sind an langen H-gebundenen Wasserketten beteiligt und tragen zur Konformationskopplung über längere Distanzen bei. Unsere Methoden werden auch auf das Protein VASA angewandt, ein am Zellzyklus beteiligtes DEAD-Box-Enzym mit ATP- und RNA-Bindungsstellen, und untersuchen die Konformationskopplung zwischen den beiden Bindungsstellen und die molekulare Dynamik der C1C2-Lipid-Protein-H-Bindung von Kanalrhodopsin.

Zudem werden unsere Algorithmen auf die SARS-COV-2-Protein-S-Kristallstrukturen angewendet. Protein S unterliegt Konformationsänderungen und dem Verlust der Symmetrie der H-gebundenen Kerncluster, wenn es von der geschlossenen in die Präfusionskonformation übergeht. In unserer Studie wurde N501 als zentraler Rest des H-Bindungsnetzwerks identifiziert, das das Spike-Protein S mit dem Angiotensin Converting Enzym 2 (ACE2) verbindet, und das anschließend in einer neuen COVID-19-Variante zu TYR mutiert wurde.

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor, Prof. Ana-Nicoleta Bondar for her guidance, encouragement, and support throughout my entire Ph.D. journey. We worked very closely all these years to answer science questions, write manuscripts, and discuss any open scientific challenge. Prof. Ana-Nicoleta Bondar gave me the freedom I needed to explore the Physics research area and make my own contribution by combining Computer Science and Bioinformatics with the field of Theoretical Molecular Biophysics.

I would also like to thank my second supervisor, Prof. Joachim Heberle, for his support and his group members for the productive exchange of views and knowledge. Our common group meetings helped me to get a wider view of the Biophysics field by attending presentations on various biological systems and methodological approaches.

I express my gratitude for our collaboration to Prof. Anastasios Economou and his lab members at the Laboratory of Molecular Bacteriology (Rega Institute), KU, Leuven in Belgium. It was a great opportunity for me to collaborate with a team of experimentalists, provide results from the computer science perspective, compare, validate our findings, and promote long and valuable scientific discussions.

It is an opportunity for me to thank all my colleagues and friends during these years. Special thanks go to Krzysztof Buzar for his positivity, for working together at the same office during the period of my Ph.D., and for our scientific and encouraging discussions. I thank my fellow labmates, Michalis Lazaratos, Suliman Adam, Federico Guerra, Stefan Milenkovic, Oskar Klaja, Lukas Kemmler, Malte Siemers, Samo Lešnik, Thomas Giannos, Éva Bertalan for the motivating scientific discussions, our collaboration, and the enjoyable moments we shared in the lab.

I would like to thank the Department of Physics of the Freie Universität in Berlin for the financial support. Computations were supported by the ZEDAT cluster of the Freie Universität Berlin, the local cluster of the Department of Physics, the North-German Supercomputing Alliance HLRN, and the computing resources of the AG-Bondar laboratory. I thank Mr. Jens Dreger and other members of the IT Service of the Physics department for their assistance and quick response to issues related to hardware and software programs we used in the lab. I would also like to thank Mrs. Sieglinde Endrias for administrative assistance. She was always kind and supportive of any regulatory process required during my studies. Many thanks to Mr. Andreas Heß for his patience and assistance with doctoral regulations and the submission of this dissertation.

Finally, I would like to thank my family for the love and support they give me throughout my life not only on the bright days but even at the most difficult times. Thank you for everything.

Publications

List of publications, published in peer-reviewed journals, arising from this thesis:

[1] Karathanou, K. and Bondar, A.N., 2018. Dynamic water hydrogen-bond networks at the interface of a lipid membrane containing palmitoyl-oleoyl phosphatidylglycerol. The Journal of Membrane Biology, 251(3), pp.461-473.

[2] Friedman R., Khalid S., Aponte-Santamaría C., Arutyunova E., Becker M., Boyd K.J., Christensen M., Coimbra JTS., Concilio S., Daday C., van Eerden F.J., Fernandes P.A., Gräter F., Hakobyan D., Heuer A., Karathanou K., Keller F., Lemieux M.J., Marrink S.J., May E.R., Mazumdar A., Naftalin R., Pickholz M., Piotto S., Pohl P., Quinn P., Ramos M.J., Schiøtt B., Sengupta D., Sessa L., Vanni S., Zeppelin T., Zoni V., Bondar A.N., Domene C., 2018. Understanding conformational dynamics of complex lipid mixtures relevant to biology. The Journal of membrane biology, 251(5-6), pp.609-631.

[3] Karathanou, K. and Bondar, A.N., 2018. Dynamic hydrogen-bond networks in bacterial protein secretion. FEMS microbiology letters, 365(13), p.fny124.

[4] Karathanou, K. and Bondar, A.N., 2019. Using graphs of dynamic hydrogen-bond networks to dissect conformational coupling in a protein motor. Journal of chemical information and modeling, 59(5), pp.1882-1896.

[5] Siemers, M., Lazaratos, M., Karathanou, K., Guerra, F., Brown, L.S. and Bondar, A.N., 2019. Bridge: A Graph-Based Algorithm to Analyze Dynamic H-Bond Networks in Membrane Proteins. Journal of Chemical Theory and Computation, 15(12), pp.6781-6798.

[6] Lazaratos, M., Karathanou, K. and Bondar, A.N., 2020. Graphs of dynamic H-bond networks: from model proteins to protein complexes in cell signaling. Current Opinion in Structural Biology, 64, pp.79-87.

[7] Karathanou, K.†, Lazaratos, M.†, Bertalan, É., Siemers, M., Buzar, K., Schertler, G.F., Del Val, C. and Bondar, A.N., 2020. A graph-based approach identifies dynamic H-bond communication networks in spike protein S of SARS-CoV-2. *Journal of structural biology*, p.107617. † Equal contribution.

** Our research study of the protein S of SARS-CoV-2 is referred to the News panel on the Freie Universität webpage (https://www.physik.fu-berlin.de/en/news/2020-graph-based-approach-spike-protein-SARS-CoV-2.html).

[8] Krishnamurthy, S., Eleftheriadis, N., Karathanou, K., Smit, J.H., Portaliou, A.G., Chatzi, K.E., Karamanou, S., Bondar, A.N., Gouridis, G. and Economou, A., 2021. A nexus of intrinsic dynamics underlies translocase priming. Structure. 29(8):846-858.e7. doi: 10.1016/j.str.2021.03.015

** Our study is referred in the preview article in Structure: Chen, W. and Komives, E.A., 2021. Open, engage, bind, translocate: The multi-level dynamics of bacterial protein translocation. Structure, 29(8), pp.781-782.

[9] Karathanou, K. and Bondar, A.N., 2021. Conformational coupling via hydrogen-bonding in the DEAD-box protein VASA. Rev. Roum. Chim., 2021, 66(10-11), 845–853 DOI: 10.33224/rrch.2021.66.10-11.08. Dedicated to the memory of Prof. Petre T. Frangopol (1933-2020).

[10] Karathanou, K. and Bondar, A.N., 2022. Algorithm to catalogue topologies of dynamic lipid hydrogen-bond networks. Biochimica et Biophysica Acta (BBA)-Biomembranes, p.183859.

[11] Krishnamurthy, S., Sardis, M.F., Eleftheriadis, N., Chatzi, K.E., Smit, J.H., Karathanou, K., Gouridis, G., Portaliou, A.G., Bondar, A.N., Karamanou, S. and Economou, A., 2022. Preproteins couple the intrinsic dynamics of SecA to its ATPase cycle to translocate via a catch and release mechanism. *Cell Reports*, *38*(6), p.110346.

[12] Jain, H., Karathanou, K. and Bondar, A.N., 2023. Graph-Based Analyses of Dynamic Water-Mediated Hydrogen-Bond Networks in Phosphatidylserine: Cholesterol Membranes. *Biomolecules*, *13*(8), p.1238.

Published Programming Codes

Research in this thesis has led to the development of the following codes and scripts that were made publicly available in Mendeley and Gitlab repositories upon publication of the corresponding original research papers:

Karathanou, Konstantina (2021), "Centrality measures and H-bond clustering in proteins", Mendeley Data, V2, DOI: 10.17632/wbprcvz6h2.2

The dataset includes algorithms to compute, plot, and visualize Betweenness & Degree centrality measures in protein structures. It also includes a workflow to compute and visualize H-bond clusters in protein structures. Code is tested for SARS-CoV-2 spike glycoprotein in open (PDB ID: 6VYB), pre-fusion (PDB ID: 6VSB), and closed conformation (PDB ID: 6VXX).

This set of scripts was developed for publication #7 in the list of own publications arising from the doctoral thesis.

Karathanou, Konstantina (2022), "Graph-based algorithm for common topologies of dynamic lipid clusters", Mendeley Data, V2, DOI: 10.17632/9c7f9vbymh.2

Karathanou, Konstantina (2022), "Graph-based algorithm for common topologies of dynamic lipid clusters", https://gitlab.com/kkarathanou/algorithm-for-lipid-cluster-topologies

The dataset includes (i) scripts to compute H bonds between lipids, H bonds between lipids mediated by waters, and lipid-ions interactions from MD simulations, (ii) scripts to compute H-bond clusters of lipids and identify their cluster topologies based on graph-theory using depth-first search (DFS) algorithm, (iii) scripts to visualize three-dimensional lipid H bond or ion interaction networks with lipids color-coded based on the topology type.

This set of scripts was developed for publication #10 in the list of own publications arising from the doctoral thesis.

Presentations

The Biophysical Society Annual Meeting, February 15-19, 2020 (BPS 2020), San Diego - California. Oral presentation: Karathanou, K., Kemmler, L., Lazaratos, M., Siemers, M. and Bondar, A.N., 2020. Proton Binding at Protein and Membrane Interfaces. Biophysical Journal, 118(3), p.179a.

Bacterial Protein Export 2018 (BPE2018) conference, Leuven, Belgium. Poster presentation: Konstantina Karathanou and Ana-Nicoleta Bondar, Hydrogen Bond Networks and Water Interactions of the SecA protein motor.

Workshop on Computer Simulation and Theory of Macromolecules, Hünfeld, Germany. April 2018. Poster presentation: Konstantina Karathanou and Ana-Nicoleta Bondar, Dynamic water hydrogen-bond networks at an anionic lipid membrane interface.

CECAM meeting in Lugano, Switzerland (January 2018), Frontiers in Computational Biophysics: understanding conformational dynamics of complex lipid mixtures relevant to biology. Oral presentation: Konstantina Karathanou and Ana-Nicoleta Bondar, Dynamic water hydrogen-bond networks at an anionic lipid membrane interface.

CECAM meeting in Bremen, Germany (June 2017), Tackling Complexity of the Nano/Bio Interface - Computational and Experimental Approaches. Poster presentation: Konstantina Karathanou, Ana-Nicoleta Bondar, Efficient analyses of hydrogen-bond networks in large biomolecules.

Attending the Computational Biophysics Workshop (October 2016) at Urbana, Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, US.

Contents

List of Figures

All figures used in this thesis are either original works by the author or are reproduced with permission from the respective copyright holders.

List of Tables

All tables used in this thesis are reproduced with permission from the respective copyright holders.

Acronyms

MD	Molecular dynamics
ACE2	Angiotensin-Converting Enzyme 2
VMD	Visual Molecular Dynamics
AI	Artificial Intelligence
TIP3P	Three-interaction-site water model
H-bond	Hydrogen bond
PS	Phosphatidylserine
PG	Phosphatidylglycerol
PC	Phosphatidylcholine
PE	Phosphatidylethanolamine
DOPC	1,2-dioleoyl-sn-glycero-3-phosphocholine
DMPC	dimyristoylphosphatidylcholine
DPPC	dipalmitoylphosphatidylcholine
CASP	Critical Assessment of protein Structure Prediction
TPs	Transmembrane proteins
TM	Transmembrane
ATP	Adenosine triphosphate
ADP	Adenosine diphosphate
GTP	Guanosine-5'-triphosphate
SRP	Signal recognition particle
SPase	Signal peptidase
XRD	X-ray powder diffraction
NMR	Nuclear Magnetic Resonance
EM	Electron microscopy
RNC	Ribosome nascent chain

ER	Endoplasmic reticulum
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
OST	Oligosaccharyltransferase
NBDs	Nucleotide-Binding Domains
PBD	Protein Binding domain
HWD	Helical Wing Domain
HSD	Helical Scaffold Domain
PBC	Periodic boundary conditions
MTS	Multiple time step
POPC	1-palmitoyl-2-oleoyl-sn-glycero-3-phosphatidylcholine
POPG	1-palmitoyl-2-oleoyl-sn-glycero-3-phosphatidylglycerol
POPE	1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine
POPS	1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine
YOPE	3-palmitoleoyl-2-oleoyl-d-glycero-1-phosphatidylethanolamine
PYPG	1-hexadecanoyl-2-(9Z-hexadecenoyl)-glycero-3-phospho-(1'-sn-glycerol)
PMPE	1-palmitoyl-2-cis-9,10-methylene-hexadecanoic-acid-sn-glycero-3-phosphoethanolamine
PMPG	1-palmitoyl-2-cis-9,10-methylene-hexadecanoicacidglycero-sn-3-phosphoglycerol
QMPE	1-pentadecanoyl-2-cis-9,10-methylene-hexadecanoic-acid-snglycero-3-phosphoethanolamine
OSPE	1-oleoyl-2-palmitoleoyl-snglycero-3-phosphoethanolamine
PSPG	1-palmitoyl-2-palmitoleoyl-snglycero-3-phosphoglycerol
DFS	Depth-First Search
PA	phosphatidic acid
PIP2	phosphatidylinositol 4,5-bisphophate
PDB	Protein Data Bank
DC	Degree centrality
BC	Betweenness centrality

SARS-CoV-2	Severe Acute Respiratory Syndrome
ACE2	Angiotensin Converting Enzyme 2
RBD	Receptor Binding Domain
cryo-EM	Cryo-electron microscopy
C1C2	channelrhodopsin-1–channelrhodopsin-2 chimæra

Chapter 1

“Η παιδεία, καθάπερ ευδαίμων χώρα, πάντα τ’ αγαθά φέρει. ” Σωκράτης

“Education, just like a fertile land, brings all the good” Socrates

1 Introduction

Figures originally published in the Journal of Membrane Biology and the Journal of Chemical Information and Modeling have been reproduced/reprinted in this chapter with permission from J. Membr. Biol. 251 (2018): 461-473. Copyright © 2018 Springer Nature and J. Chem. Inf. Model. 59, no. 5 (2019): 1882-1896. Copyright © 2019, American Chemical Society.

1.1 Biological Background

1.1.1 Cell membranes and model lipid bilayers

Membranes play a crucial role in biology, allowing life as we know it to exist. Essentially, their role is to separate an organism's inside from its outside and regulate the flow of substances with their selective permeability. For example, the plasma membrane, which separates the inside from the outside of the cell, is used to sense external signals [1], screen water and ion transport [2], transfer lipids [3] as well as be a key part in the signal transduction and binding of membrane proteins [4, 5]. On the other hand, intracellular membranes compartmentalize the internal of the cell allowing each organelle (e.g. Golgi complex, mitochondria, etc.) to function in a characteristic manner (Figure 1.1) [6] [7].

Simply put, a biological membrane is a dynamic mixture of lipids and proteins which organizes itself in a way it excludes but it is also excluded, from water. Moreover, membranes allow for the creation of ion gradients across them which give the ability to living organisms to generate energy. This dynamic mixture of lipids and proteins is also able to control the flow of signals between cells by acting as a transmitter, receiver as well as processor of information. The information is typically given in the form of chemical and electrical signals. Furthermore, membranes act as a protective barrier by preventing the transport of unwanted molecules and pathogens into the cell. This is done through different molecular recognition mechanisms (typically protein receptors) which exist at the membrane surface that allow the cell to detect pathogens. Such recognition mechanisms also play a role in signals sent between cells but also other forms of cell-cell interactions.

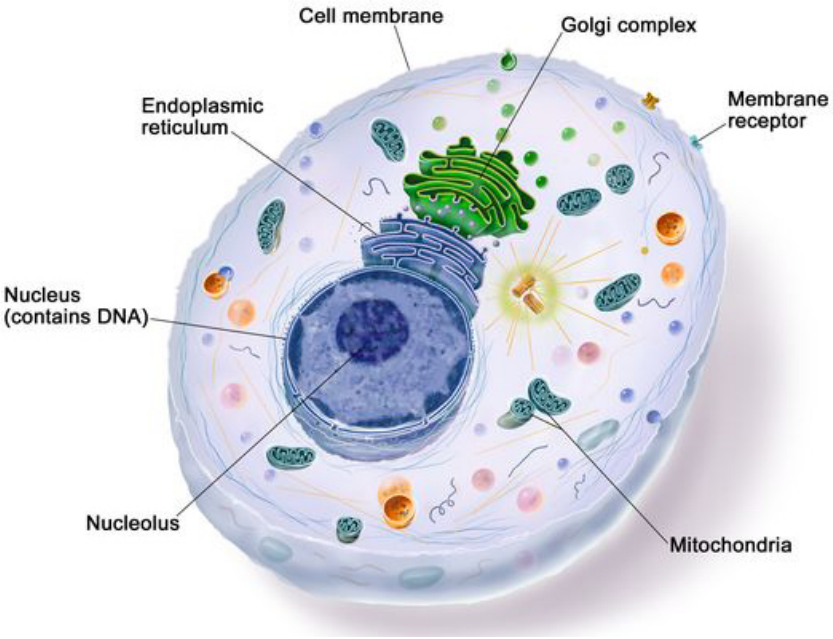


Figure 1.1: A mammalian cell. In this illustration, the outer membrane, as well as the inner organelles, can be seen. The figure is from ref. [8].

In 1972, Singer and Nicolson introduced the fluid mosaic model (Figure 1.2) [9], according to which the membrane consists of proteins and lipids that form a nanometers thin bilayer film with proteins either being on the surface or embedded in the membrane. The membrane is flexible, allowing lateral diffusion of both components (proteins and lipids). Hence, it is characterized as fluid as the bilayer is viscous. The term mosaic comes from the fact that lipids can move in the viscous medium, which is embedded with proteins, resulting in a mosaic of components.

It is common for membranes to contain phospholipids as well as glycolipids and sphingolipids. These molecules are amphipathic, consisting of two components, a phosphate group (hydrophilic head) and fatty acids (hydrophobic tail) (Figure 1.3, Figure 1.4). Due to their amphiphilicity, lipids in polar solvents (e.g., water) will self-assemble in a manner that minimizes unfavorable interactions. Common types of lipids include phosphatidylcholine (PC), phosphatidylglycerol (PG), phosphatidylserine (PS), and phosphatidylethanolamine (PE). In most biological membranes approximately 10-20% of lipids carry a negatively charged head group [10], for example, PS and PG head groups; by contrast, PC lipids are zwitterionic.

Lipids carrying a net negative charge are typically distributed in an asymmetric fashion between the inner and outer leaflets of the membranes, with the inner leaflet being more of these lipids [11] [12]. Under physiological conditions, in mammalian cells, the plasma membrane's inner leaflet is mainly composed of PS and PE, while the outer leaflet is made up of PC and sphingomyelin [13]. Variations in the relative proportion of lipids may influence the activity of proteins [14]. During apoptosis, PS lipids can also be found in the outer leaflet as they translocate from the inner to the outer leaflet of the plasma membrane. Scramblase is an enzyme that facilitates the exposure of PS to the outer leaflet. Since PS carries a negative charge and PC is uncharged, the flipping of PS to the outer leaflet of the plasma membrane results in a change in the membrane's charge. The alteration of the surface charge is characteristic of an apoptotic cell, marks the cell for phagocytosis by macrophages and other types of cells [15], and associates with various human diseases including cancer [16].

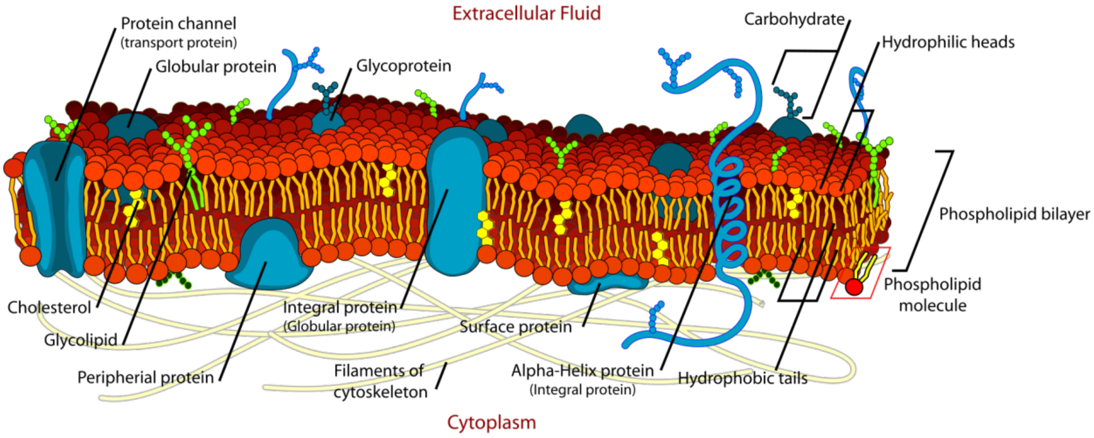


Figure 1.2: Representation of the fluid mosaic model of a cell membrane [8]. The membrane consists of lipid molecules, proteins (surface proteins or transmembrane proteins), and various organic molecules. Within the bilayer, lipids and proteins are free to move through lateral diffusion. The figure is from ref. [17].

Different lengths and saturation of the hydrophobic tails are also important since they affect the packing of phospholipids in the bilayer - which in turn affects the fluidity of the membrane. Typically, the lipid tails are highly fluid, for example in the liquid crystal state the fatty acid tails are disordered and constantly in motion. However, as the temperature decreases the lipid bilayer turns into a crystalline phase in which lipid tails are stretched, oriented, and the van der Waals interactions are maximal. The transition temperature is a characteristic of lipid bilayers with different bilayers having different transition temperatures [18].

The hydrophobic core of the bilayer is approximately 4 nm thick in the fluid state. The precise value of the membrane thickness depends on the type of lipids composing the bilayer. Other than the thickness, characteristics of a lipid bilayer are the area per lipid and the order parameters of the lipid configuration. Such characteristics are typically used to compare results from simulations with experiments. As an example, a lipid bilayer composed of 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC) lipids has an area per lipid of 72.2 Å^2 [19, 20] when measured experimentally, and 71.0±1 Å^2 when computed with atomistic MD simulations [19, 21].The order parameter essentially describes the extent of the ordering of lipids in the bilayer. It can be used to pinpoint possible structural deformations of the bilayer and is a highly important characteristic.

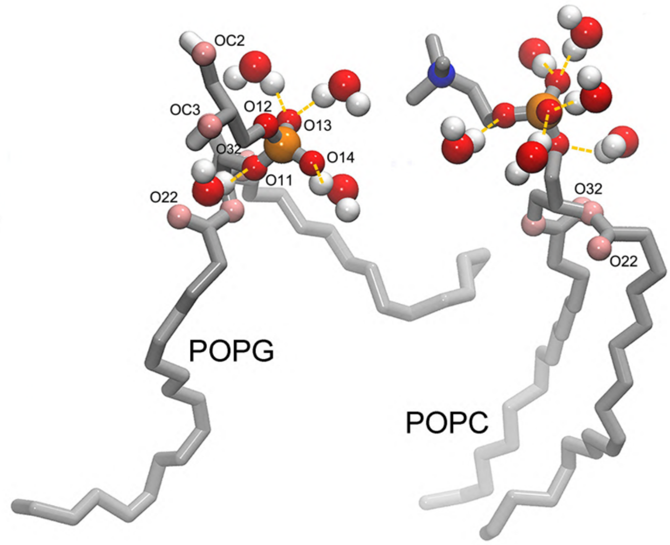


Figure 1.3: H bonding between water molecules and phosphate groups of membrane model composed of 4:1 POPC: POPG lipids. Specific oxygen atoms are marked, and the H bonds between them are depicted using yellow dashed lines. Phospholipid molecules consist of a polar head and two nonpolar tails. The polar head (hydrophilic) is essentially a phosphate group bonded to a glycerol molecule. The nonpolar tails (hydrophobic) contain a fatty acid (saturated or unsaturated) and are composed of hydrocarbon chains (typically 14-22 carbon atoms). The image is from [22], a study presented in this thesis.

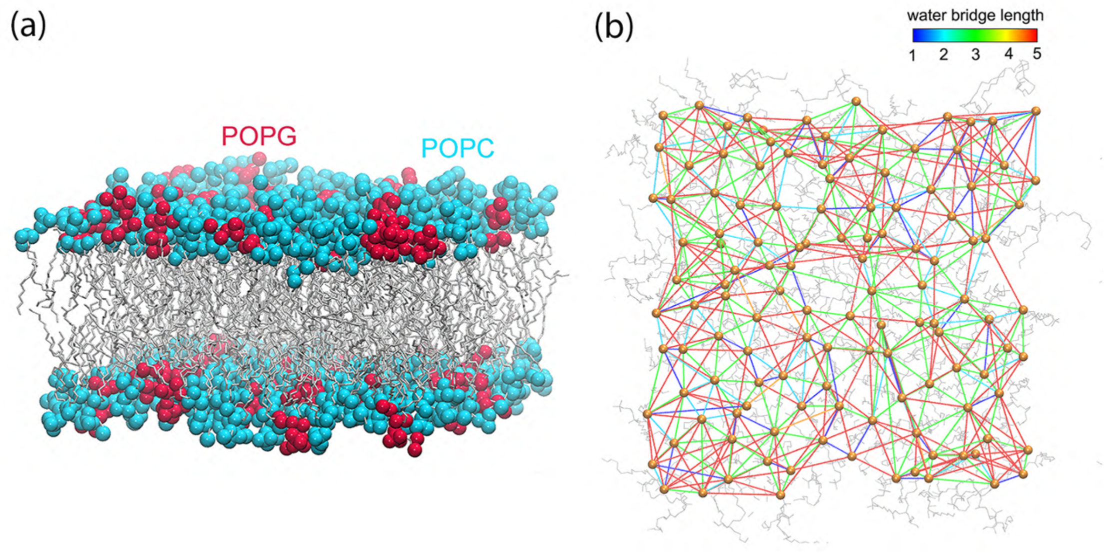


Figure 1.4: Lipid membrane model composed of zwitterionic POPC and anionic POPG lipids. (a) Lipid headgroups are shown as colored van der Waals spheres and lipid alkyl chains are depicted as gray licorice in VMD [23]. (b) Network of water H-bond bridges of various lengths inter-connecting phosphate groups represented as orange van der Waals spheres. Molecular graphics are based on MD simulations from [22], a study presented in this thesis.

The different lipid composition of healthy as compared to diseased cells opens avenues of research to distinguish between healthy neutral cells and cancer cells with altered distribitution of the negatively-charged PS lipids [24] [25] – for example, to develop pharmaceutical targets such as positively charged proteins or drug molecules. Moreover, experiments [26] and computations [27, 28] suggest that anionic lipids form clusters with a high propensity providing a pattern for binding events at the surface of the cell membrane.

Water is a key component affecting the structure and functionality of biological membranes. Water molecules can enter the lipid bilayer, where they interact with lipid oxygen atoms [29-35]. The structure and dynamics of water are affected by its interactions with the lipid headgroups [36, 37] within ~ 10 Å of the membrane surface plane [38]. Water molecules can H bond with the lipid headgroups and form chains of water-mediated bridges [33, 39]. Reduced dynamics of the headgroups [40] and H-bonded waters are slowed down relative to bulk water [35, 36, 41, 42].

Although cell membranes are highly complex, as they are made up of hundreds of diverse lipids and proteins, and other molecules such as cholesterol (or other sterols), model systems used for molecular simulations typically consists of one-two lipid types. In this thesis, bilayers consisting of mixtures of zwitterionic, and anionic lipids in different compositions, and a more complex model of the *Escherichia coli* inner membrane, are modeled and used in MD simulations (Figure 1.4). Graph-theory algorithms are implemented to analyze the intricate dynamics of water H-bond networks in interfaces of lipid membrane models (Figure 1.4), identify the topology of H-bonded lipid/lipid, lipid/water, and ion/lipid clusters, and characterize their dynamics.

1.1.2 Proteins

Proteins are macromolecules that consist of linearly connected amino acids in the form of a polypeptide chain that folds into three-dimensional (3D) structures. Proteins are made up of a group of 20 amino acids that can be found naturally. All amino acids have the same structure consisting of a central α carbon atom creating bonds with an amino group (NH_2), a carboxyl group (COOH), an H atom, and a side-chain group R (Figure 1.5). Essentially, what differentiates amino acids is the side-chain group.

Amino acids are typically identified using three-letter coding and can be separated into subgroups: nonpolar (GLY, ALA, VAL, CYS, PRO, LEU, ILE, MET, TRP, PHE), polar/not charged (SER, THR, TYR, ASN, GLN), polar/charged (LYS, ARG, HIS, ASP, GLU).

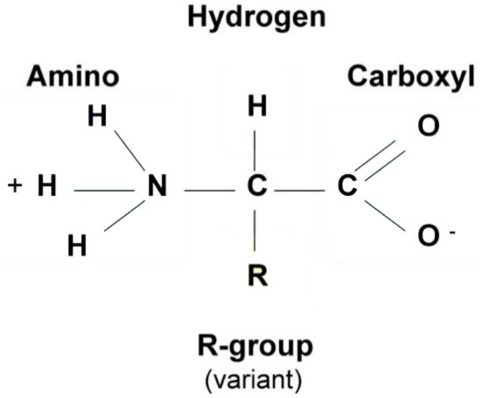


Figure 1.5: Amino acids consist of a central chiral carbon atom (except for glycine), a H atom, an amino group (NH2), a side chain group (R), and a carboxyl group (COOH). The figure is adapted from ref. [43].

In the past, the variety and complexity of proteins couldn’t be explained by scientists as nucleic acids that compose genes were considered very simple molecules. By the second half of the 20th century, numerous protein molecules had been identified and the scientific community had resolved that each one is unique. The composition, sizes, and structure varied wildly while nucleic acids are composed of four nucleotides (Adenine, Thymine or Uracil, Guanine, and Cytosine), and naturally, the question of how such a simple 4-part code could express the diversity of proteins encountered did arise. This single code works because the series of three nucleotides, called codons or triplets, can express one amino acid. Furthermore, the code is degenerate since multiple combinations of three nucleotides can express one single amino acid (e.g., six codons can express serine while only two can express phenylalanine) [44]. In this way, the code for similar proteins can be written in many organisms but also the effect of random mutations is reduced. Moreover, amino acids with larger degeneracy (more codons express the same amino acid) are "cheaper" for the cell to synthesize [45].

The procedure of converting the information hidden in the genetic code into a protein is called expression (Figure 1.6). The first step (transcription) of this process is the conversion of DNA into messenger RNA (mRNA). This is done by another protein, an enzyme called RNA polymerase, which can recognize a sequence of DNA in front of the genes. Next, a short sequence contained in mRNA is recognized by the ribosome (a macromolecular machine that essentially is a protein-RNA complex and synthesizes proteins) and then the mRNA gets translated into a protein [46]. After the translation process, proteins can still be modified in various ways. For instance, proteins can be transferred inside or outside the cell into the extracellular space, depending on where they perform their function. Furthermore, they can simply be left to self-assemble into their folded or functional structure. This conformation is typically the one with the minimum free energy. This structure often changes in shape when the protein interacts with other components within the cell; that change is vital for the protein's function [47]. In essence, amino acids are arranged like beads on a polypeptide chain during translation. That sequence is the primary structure of proteins (Figure 1.8). For most proteins, their folded structure is attained later whereas most proteins like enzymes typically fold into a globular structure. The folded structure (or conformation) is called the tertiary structure (Figure 1.8) [48].

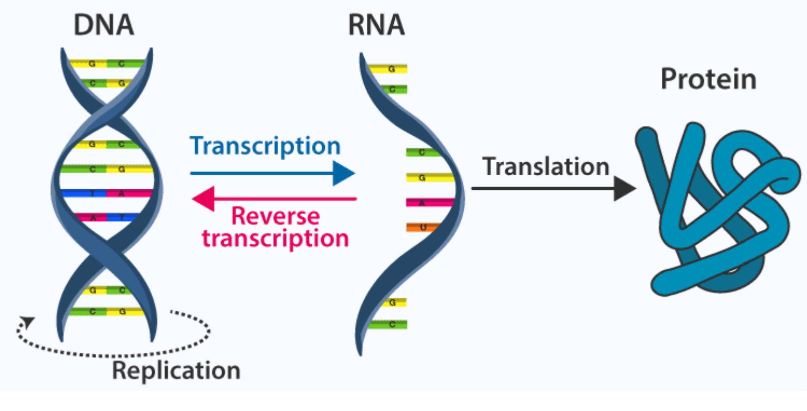


Figure 1.6: Central dogma of biology. DNA serves as the template for RNA synthesis, which in turn is translated into protein through translation.The figure is from ref. [49].

The sequence of amino acids and their size determines the function, size, and shape of each protein. Amino acids are linearly connected into a polymeric chain via peptide bonds formed through a dehydration reaction in which an acid (the carboxyl group of one amino acid) reacts with a base (the amide group of the next amino acid) to form a bond and release a water molecule (Figure 1.7). Each polypeptide will have an unreacted carboxyl group (C terminal) on one end and one unreacted amide group on the other end (N terminal).

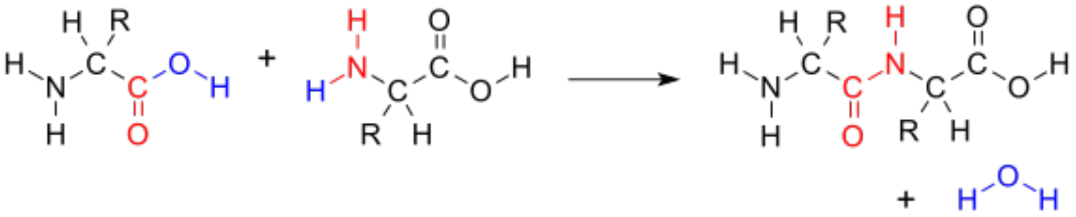


Figure 1.7: Formation of a peptide bond through a dehydration reaction. The figure is from ref. [50].

Polypeptide chains tend to partially form ordered elements (Figure 1.8) in 3D space such as helices or sheet-like arrangements while other parts of the polypeptide chains will remain amorphous to form loops or random-coil bridges between the organized sections. The most commonly observed structures are the α-helix as well as the β-sheet and β-turns [51]. The position of such structural elements can easily be predicted since some amino acids tend to occur in only one type of such secondary structure. That is also valid for specific patterns or sequences of amino acids. However, the estimation of the relative organization of the elements of the secondary structure, which will form the tertiary structure of the protein is challenging (Figure 1.8). The number of ways the secondary structure can be organized into the tertiary is called a fold. To predict the tertiary structure, the potential evolutionary relationship with other proteins of known structure is needed. Typically, knowing the 3D structure of a protein reveals its function which is essential knowledge for the design of efficient and safe drugs. If one needs to study the tertiary structure of the protein, the most common techniques to use are x-ray powder diffraction (XRD) crystallography, electron microscopy (EM), and nuclear magnetic resonance (NMR) spectroscopy [52]. The tertiary structure of proteins can be interesting for drug and medical applications, as mentioned before, but can also be of academic interest since the entire protein structure universe has not been discovered yet [48]. The number of ways the secondary structure can be organized into the tertiary (called a fold) and the size of the 'fold space' is unknown. Hypothetically, when all folds are known, one could explore fold families, discover variations, or try to design new proteins that would not naturally exist.

Based on Cyrus Levinthal, an American molecular biologist, owing to the huge number of degrees of freedom that are in an unfolded polypeptide chain, the protein has an excessive number of potential folds or conformations [53]. Over the years, numerous computational methods were implemented to predict protein structures, but compared to experimental results their accuracy wasn't acceptable except for small molecules. An example is the Critical Assessment of Protein Structure Prediction (CASP) tool launched in 1994 [54]. In 2018, DeepMind company released the AlphaFold tool, a significant breakthrough, which uses artificial intelligence (AI) and deep learning techniques for the prediction of protein structures and an updated version of the program became available in 2020. The program is trained by around 100,000 known human proteins and can predict rapidly the shape of a protein down to atomic accuracy [55, 56].

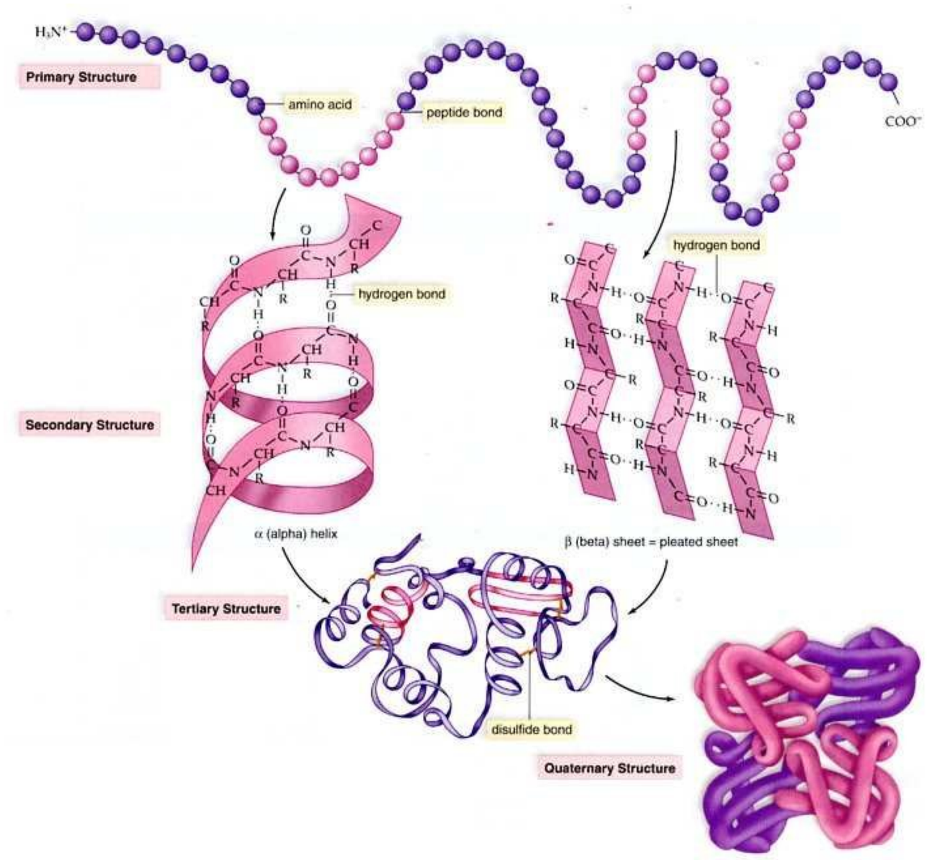


Figure 1.8: There are four levels of structural organization in proteins: primary, which is just the sequence of amino acids, secondary (locally organized regions of the polypeptide chain), tertiary (3D folding of the polypeptide chain), quaternary (a functional form of many polypeptide chains organized into one structure). The figure is from ref. [57].

1.1.3 Membrane Proteins

Proteins that are found in membranes essentially are molecular machines that enable the exchange of signals between the cell and its surroundings as well as the molecules' movement within and outside of the cell. Approximately one in five of all predicted genes express such proteins [58]. The α-helix and the β-barrel are the two primary structural components of membrane proteins. They are created by the polypeptide chain folding due to intramolecular amino acid H bonds. These structures are deeply embedded in the highly complex lipid bilayer. The absence of membrane proteins would render the cellular membrane totally impermeable, isolating cells from their environment and hindering their ability to perform functions such as signaling, nutrient/waste transport, or response to external stimuli.

There are two ways in which a membrane protein can be embedded in the membrane (Figure 1.9) [59]. Firstly, there are the peripheral membrane proteins, which are temporarily linked to the lipid bilayer or to integral proteins (see below). Lipid-anchored proteins belong to the third category. They are present on the membrane's surface and have covalent bonds with the lipids that are incorporated in the bilayer. Such proteins are placed in the membrane close to similar fatty acids (Figure 1.9).

Secondly, there are the integral membrane proteins – these are transmembrane proteins (TPs) that span the cell membrane. Like lipids, TPs are amphiphilic with hydrophilic and hydrophobic regions (Figure 1.10). Hydrophobic regions are placed within the membrane, interacting with the lipid tails where they are not interacting with water molecules. On the other hand, both sides of the membrane expose the hydrophilic regions of integral proteins to the aqueous environment. Some of the TPs can become more hydrophobic by covalently bonding a fatty acid chain that enters the cytosolic monolayer of the bilayer (Figure 1.10). TM proteins are of significant interest for drug discovery as most drugs interact with the membrane protein to achieve their therapeutic outcome. Given their importance, computational tools are essential for predicting and understanding the geometric orientation of TM proteins in relation to the membrane, their number of helices, etc. Significant studies have been performed on the prediction of membrane protein topology [60-62]. A study by [63] underlines that the AlphaFold2 machine learning tool [56] can provide reliable results in membrane proteins and perform well in structural prediction analysis using Artificial Intelligence. Reliable results are concluded after their validation by experiments.

TP proteins are also integral parts of the transport (passive or active) of substances across the cell membrane or membranes around organelles. This process is of paramount importance since cells need to exchange substances with their surroundings to let nutrients in and waste out of them. Briefly, passive transport is the exchange of molecules between the two sides of a biological membrane; this is done through concentration gradients, without an energy input. For example, ion channels may allow the movement of ions following an electrochemical gradient. Active transport is essentially the act of transporting molecules against a concentration gradient which requires energy – which is provided by, e.g., the binding and hydrolysis of adenosine triphosphate (ATP), light (e.g. bacterial proton pump bacteriorhodopsin), or electrochemical gradients [64].

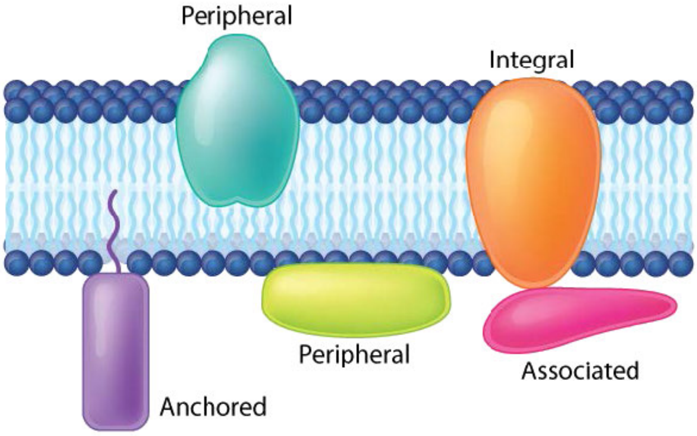


Figure 1.9: Simple illustration of different types of membrane proteins: lipid-anchored, peripheral, and integral proteins. The figure is from ref. [65].

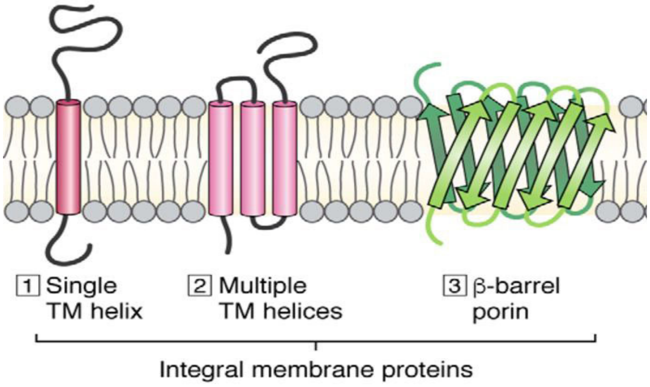


Figure 1.10: Illustrations of different transmembrane proteins: 1) Single α-helix (bitopic), 2) Polytopic α-helix, 3) Polytopic β-sheet. The figure is adapted from ref. [66].

1.1.4 Protein translocation in bacteria

More than one in four of all bacterial proteins are translocated from the inside of the cell (cytoplasm) into the plasma membrane or even across it. A major protein secretion pathway is that of the SecYEG protein conducting channel (also known as the SecYEG translocon).

Protein translocation occurs in two different ways: co-translationally and post-translationally [67]. Secretory proteins—those released by the cell—are often produced in the cytoplasm and transported after their translation. Such proteins contain a N-terminal sequence that needs to be removed at a still undetermined translocation stage [68]. The protein motor SecA ATPase, with the help of SecB chaperone, identifies and attaches to the signal sequence that needs to be removed from the secretory protein, guiding it to Sec translocon. Then the polypeptide gets translocated through a channel able to conduct proteins, with the aid of SecA which uses the energy produced from the ATP hydrolysis reaction [69] [70] (Figure 1.11).

A co-translational mode of translocation is used when it comes to inner membrane proteins. The signal for such a process can take two forms. It can be a signal sequence that is later removed, as described above. Alternatively, it can be a sequence of hydrophobic amino acids at the N-terminus of the nascent protein that becomes the first transmembrane domain of the mature protein. This type of signal is known as a signal anchor.

After exiting the ribosomal tunnel, the signal sequence or anchor is recognized by the generally conserved SRP/SR system. This is where the signal recognition particle (SRP) binds to the nascent protein's signal sequence or anchor. The whole ribosome nascent peptide chain complex targets the membrane, initiating an interaction between the SRP and its membrane-bound counterpart, the SRP receptor (SR). The ribosome-nascent chain (RNC) is transferred to the Sec translocon through a mechanism involving guanine triphosphate GTP hydrolysis in both SRP and its receptor, although this mechanism remains unidentified [71]. Then, the protein is translated and embedded into the membrane by the ribosome, which is attached to the protein conducting channel. Therefore, transmembrane segments need to be pushed into the membrane laterally after being identified by the Sec translocon (Figure 1.12). Such segments can be predicted computationally as they typically consist of hydrophobic sequences of 15-30 residues.

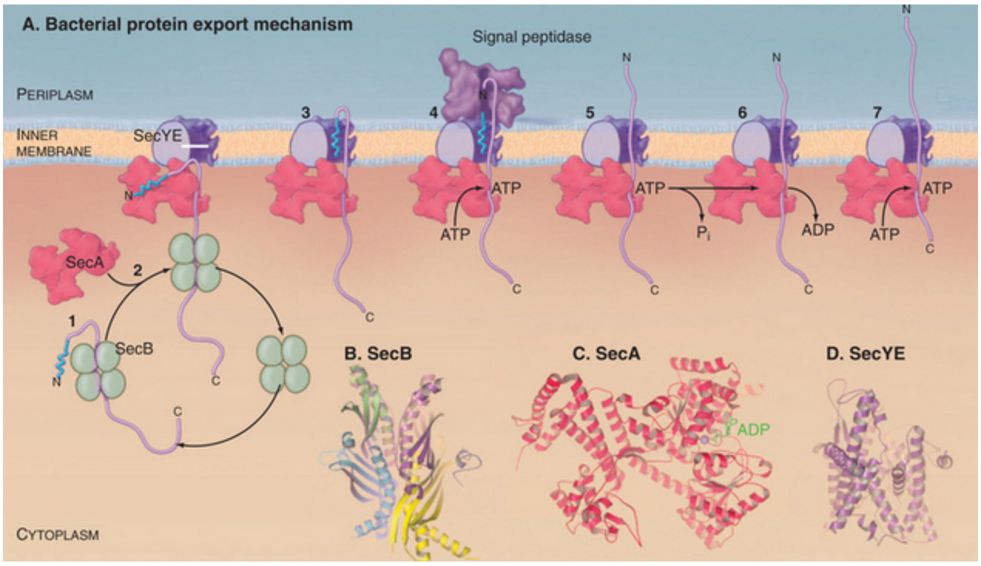


Figure 1.11: Bacterial protein secretion mechanism using the SecYE translocon. A. 1. The polypeptide which is synthesized in the cytoplasm binds to SecB chaperone. 2. SecA links the pre-sequence (illustrated in blue) and binds on the SecYE translocon. 3. The pre-sequence is inserted into the translocon 4. The ATP of SecA enables the insertion of the polypeptide and the signal sequence is cleaved. 5,6,7. Cycles of ATP hydrolysis, as well as the potential of the membrane, drive the peptide through the inner membrane. Ribbon diagrams of B. SecB. C: dimeric SecA. D: SecYE complex translocon. The figure is from ref. [72].

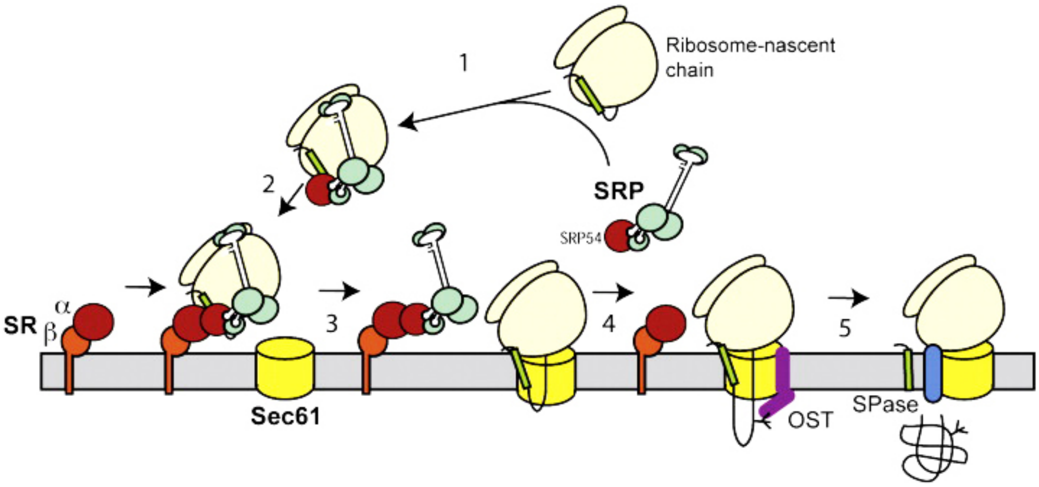


Figure 1.12: Co-translation targeting pathway. The RNC-SRP complex binds to the signaling sequence, attaches to the ER membrane, and enters the translocon pore (Steps 1-3). The nascent chain is translocated through the pore and disengaged from the SR. Signal peptidase (SPase) and oligosaccharyltransferase (OST) enzyme complexes cut away the signal peptide, adding N-linked glycans (Steps 4-5). The process of protein synthesis ends when the polypeptide chain is released from the ribosome, and then transported and folded in the lumen of the ER. The figure is adapted from [73].

1.1.4.1 The SecA protein motor

SecA is a key component part of the Sec protein secretion pathway in bacteria. SecA provides chemo-mechanical coupling, pushing the polypeptide chain through the cytoplasmic membrane. Moreover, it facilitates the transport of secretory proteins from the ribosome to the membrane, either alone or in combination with chaperones. Furthermore, it converts chemical energy to mechanical at the membrane, where preproteins are translocated through the SecYEG channel. In essence, SecA is an enzyme, highly dynamic, exploiting swiveling and dissociation, order-disorder kinetics, transforming from its dimer state to the monomer state, all of which are coupled with its catalytic function. This dynamic nature of SecA is exploited from signal sequences found in the preprotein as well as mature domains in order to control the nanomotor, achieving their translocation using the metabolic energy [70].

SecA is a large protein with several different functional domains (Figure 1.13). There are two Nucleotide-Binding Domains (NBD 1 and 2, Figure 1.13) and nucleotides (ATP or ADP) can bind at their interface. The two NBD domains form the so-called DEAD-motor, the minimal part of SecA. The DEAD-motor can bind and hydrolyze ATP molecules [74]. The name for the DEAD-motor originates from the DExD sequence, (Asp-Glu-x-Asp), present in both DNA and RNA helicases. The binding and hydrolysis of ATP (whereby one molecule of ATP is hydrolyzed by one molecule of water, yielding ADP and inorganic phosphate [75]) releases energies in the range of 11-16 kcal/mol. The domain where the pre-protein binds is called the Protein Binding domain (PBD), which is a continuation of the NBD1 domain as shown in Figure 1.13. PBD is a dynamic region of the SecA protein and during conformational changes, it samples three main position states in relation to the NBD2 and Helical Wing Domain (HWD) [76]. Allosteric coupling takes place between the PBD and NBD. Once the pre-protein binds at the PBD the conformation of the DEAD motor loosens which affects the release of ADP [77]. The Helical Scaffold Domain (HSD) is composed of three helices, two of which are shorter than the other. The long helix will associate with all other domains of SecA, where it participates in clusters formed by dynamic H bonds [78, 79] [80]. On the other hand, the two short helices, commonly reffered as the "two-helix finger" can participate in driving the preprotein through the translocon [81] [82], however, the exact role of the "two-helix finger" in the function of SecA is still under investigation [83] [84] [85]. In some species, SecA might have more insertions in the five functional domains described above, as seen, for example, in one *T. maritima, E. coli* (one insertion), or *T. thermophiles* (two insertions). A well-studied insertion like the ones described above is a variable subdomain known as VAR found in both *T. maritima, E. coli, and T. thermophiles* [86]. In ref. [86] the authors conclude that the VAR is not critical for the function of SecA, although, in *E. coli*, it can modulate the regulation of ATPase activity within the SecA by accelerating ADP release. This could take place because VAR forms H bonds with NBD1 and NBD2 which could affect the rate of ADP release [78].

The study of SecA crystal structures provides crucial information regarding the 3D structure of SecA in several different organisms, such as *Bacillus subtilis* [87, 88] [89], *Escherichia coli* [90], *Mycobacterium tuberculosis* [91], *Thermotoga maritima* [92, 93] and *Thermus thermophiles* [94]. SecA has been mostly captured in its apo or ADP-bound states. The *E.coli* SecA structure in ref. [90] provides some information about the ATP-bound SecA. In this work, coordinates for the ATP molecule are indicated, however, those for the magnesium ion and the majority or the PBD are not provided. On the other hand, the structure of SecA from *E.coli* bacteria is well resolved by Nuclear Magnetic Resonance (NMR) [95]. The structure provided in ref. [95] contains the whole PBD domain and a signal peptide but includes no coordinates for the bound nucleotide.

An energy release from ATP binding to SecA allows the preprotein to translocate 2.5 kDa into the SecYEG. After ATP is hydrolyzed and released into ADP, the pre-protein separates from SecA and inserts itself into SecYEG at a size of 2.5 kDa. This means that each cycle (ATP binding to SecA, ATP hydrolysis) translocates 5kDa (i.e. 20-30 residues) [96] [97]. Thus, in the model described, SecA is inserted into the SecYEG along with the pre-protein, while ATP hydrolysis releases the SecA protein. However, the model is under question because a channel in the internal of SecYEG is too small to accommodate protein domains [98]. An alternative model proposes that SecA uses the "two-helix finger" motion to deliver the preprotein to the SecYEG channel after completing its function inside the cytoplasm [81] [82]. The specific motions at hand are still under question because several experiments have reported translocation of proteins even though the “two-helix finger” domain was artificially immobilized [83].

The study of Zimmer et al. [82] was the first to report the crystal structure of SecA binding on SecY. In the described structure, SecA attaches to SecY in an orientation that is nearly parallel to the membrane. Interaction between residues from SecA protein's PBD domain and SecY protein's loop 8/9 controls binding. Then, the cytoplasmic funnel of SecY receives the loop of SecA's "two-helix finger" domain. It has been suggested that hydrolysis of ATP in SecA protein leads to a palindromic motion of the “two-finger” loop, leading to the translocation of the unfolded polypeptide [84].

Large conformational changes of SecA have been reported when interacting with SecYEG [82] [93] [99, 100]. In the crystal structure of SecA (*Bacillus Subtilis)* obtained using XRD (PDB ID: 1M6N), the PBD is located close to the HWD, a state referred to as the ‘closed’ conformation. When interacting with SecYEG or some substrate protein, the PBD will rotate and translate in order to approach NBD2, and move away from HWD [82] [93] changing into what is called the ‘open’ conformation. The substrate protein and SecA associate with the interface between NBD-1, NBD-2, and PBD [92].

Once the clamp 'opens up' it surrounds the substrate protein, stabilizing its interaction with SecA [82] [101]. This 'opening up' of the clamp leads to activation of the ATPase functionality of SecA through an increase in the rate of nucleotide exchange [101]. Several partly open structures of PBD have been reported in high-resolution structures [93] [95]. In this respect, NMR studies have shown that approximately 10% of the protein is in the closed structure. On the other hand, the remaining 90% assumes the so-called 'partially open' conformation [95] [99].

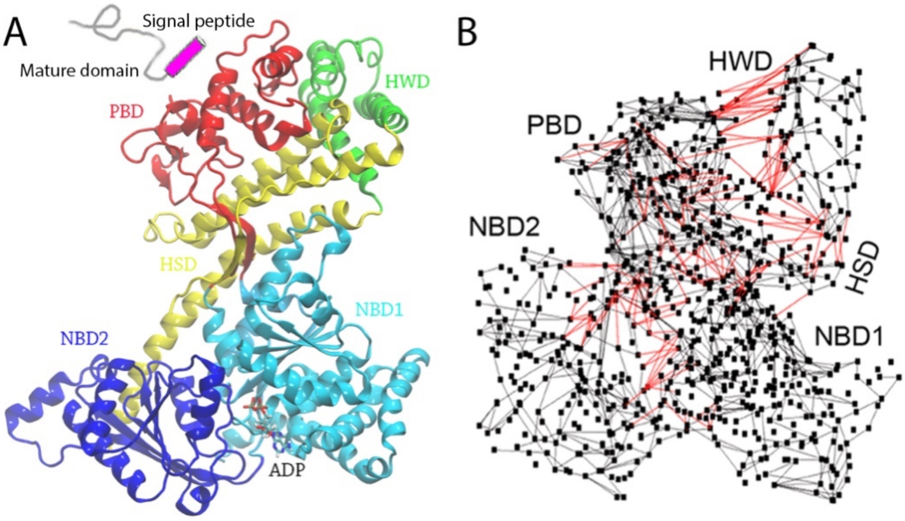


Figure 1.13: Structure and conformation of SecA with an ADP molecule bound. (A) *Bacillus subtilis* SecA crystal structure. Protein Databank ID: 1M74 [87]. Protein is color-coded based on the protein functional domains, NBD1, NBD2, HSD, HWD, and PBD to which pre-protein binds. Molecular graphics were generated using VMD [23]. Coordinates were used as initial conditions for computations in [102]. (B) Network of H bonds depicting interactions between amino acids of the same (gray edges) and different functional domains (red edges). H bonds are sampled during simulations from [102]. At any moment time, there are on average ~200 intradomain and ~60 intradomain H bonds in SecA. Image is from [102], a study presented in this thesis.

In this study, algorithms were implemented to provide simple graphical representations of the complex and dynamic protein H-bond interactions during MD simulations enlightening the complex function of SecA protein [102] [99, 100, 103]. A key finding is that mutations near the region where the nucleotide binds or different nucleotide-bound states of SecA associate with altered dynamics at a long-distance region of the protein, the PBD where the pre-protein binds. Extended pathways of H bonds of different lengths and occupancies were identified. Water molecules contribute significantly to H-bonded water wires that connect different protein functional domains and promote long-distance conformational coupling.

1.1.5 H bonding

H bonds are mainly electrostatic in nature (Figure 1.14), whereby an attractive interaction exists between an H atom, that is covalently bound to a more electronegative atom (typically, in proteins, oxygen (O) or nitrogen (N), but also sulphur (S) –the donor, Dn, hetero-atom), and a second electronegative hetero-atom carrying a lone electron pair (acceptor, Ac). A typical H bond is denoted as Dn−H···Ac, with the solid line indicating a covalent bond and the H bond being shown with a dashed line [104, 105].

H bonds can exist between different molecules (intermolecular) or within the same molecule (intramolecular) [106] [107] [108] [109]. The H-bond binding energy depends on the environment (e.g., polarizability), specific geometry or the nature of the donor and acceptor atoms. This brings them in between van der Waals interactions (weak) and covalent bonds (strong).

H bonding is responsible for structures found in proteins and peptides such as α-helix and β-sheets (Figure 1.15). The biochemical functions of proteins and peptides are strongly connected with their structure; hence, H bonds are highly responsible for the structure since they are typically the driving force behind specific folding. The same holds for the structure of some polymers, synthetic or natural.

Such types of H bonds are typically identified using criteria based on geometry, such as the distance between donor and acceptor atoms, the bond angle, or the distance between the H atom and the acceptor heavy atom. For example, H and N H bonds are typically ranging between 1.56Å and 2.63Å [110] [111] with an angle of 120° [110]. Moreover, the typical H-bond distance between N and O atoms in peptide backbone groups ranges between 2.8 Å to 3.0 Å [112]. On the other hand, for protein structures, where H-atoms are not typically well resolved, the geometric criterion requires that donor and acceptor heavy atoms are separated by 3.5 Å with 90° angle between donor-H- acceptor atoms [110].

In this doctoral thesis, I present the development of efficient algorithms that allow H-bond network analyses from computer MD simulations of experimentally determined protein and lipid bilayer structures.

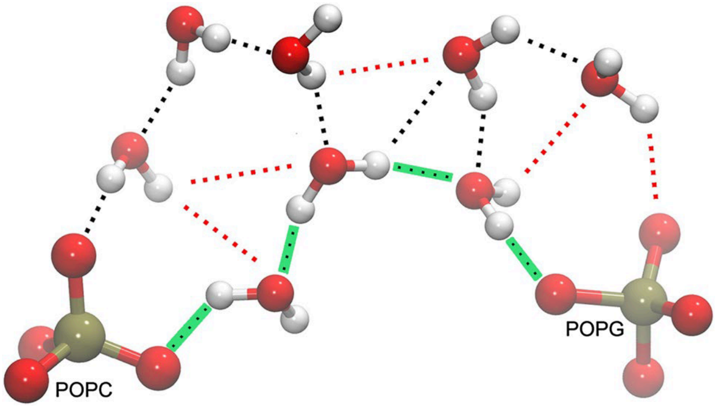


Figure 1.14: Schematic representation of H bonds formed between water molecules and between water and lipids (POPC, POPG). The figure is from analyses by [22], a study presented in this thesis.

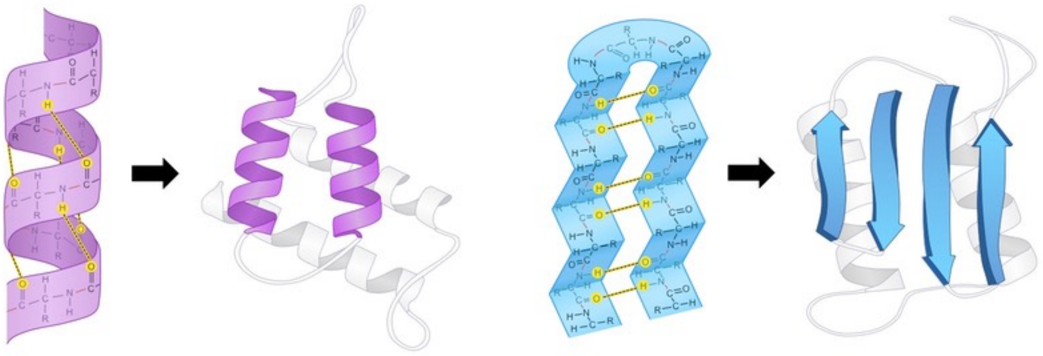


Figure 1.15: H bonds in different secondary protein structures. The α-helix (spiral structure) and β-sheet (pleated structure) motif are commonly found in proteins. The α-helix spiral structure is stabilized by H bonds while the same holds for the β-sheets. One sheet is linked together with another one with H bonds formed between amide and carbonyl groups of the adjacent sheet (or vice versa). Figure is from ref. [113].

Chapter 2

"If I have seen further, it is by standing upon the shoulders of giants" - Isaac Newton

2 Theory

2.1 Molecular Dynamics

High-speed computers have emerged since the late ‘60s and since then have changed science and engineering by introducing computer simulations, an element which can be thought of as being between experiment and theory. In simulations, modeling the system is up to the researchers while calculations are taken care of computers using typical algorithms written in several programming languages. Computer simulations can provide an avenue for exploring experimental systems, leading to a deeper understanding of their function [114].

One of the main computational tools for studying biomolecules is MD simulations. Such an approach allows the study of a system and how it evolves over time, yielding insightful information regarding dynamic processes in biological systems.

MD was initially introduced by Alder and Wainwright towards the end of the 1950s [115] [116] and is founded on the integration of Newton’s second law of motion. By integrating the equations of motion over time, one can calculate the forces on each time step which will lead the algorithm to displace atoms accordingly. Alder and Wainwright studied the interactions of hard spheres in motion simulated for a total of 9.2ps. After that, several systems were investigated using MD simulations; however, the first simulation of a protein was achieved in 1977, when the bovine pancreatic trypsin inhibitor was simulated with a total simulation time of 8.8 ps [117]. Ever since several important improvements have been reported, along with the increase in the impact of computational methods in several areas of science. Improvements in theory, methodology, and hardware translate into MD being increasingly used, especially after the 1990s, to understand the chemistry and biology of proteins. Moreover, during the last 10 years, the booming technological advances in computational speed and data storage volume allowed for simulations to be performed in much larger time scales compared to the picoseconds of the first simulations, allowing access to timescales that are relevant for biological processes [118]. For example, it is not uncommon nowadays to find publications that have studied systems up to micro- or even milli-seconds. These timescales are relevant to observing conformational changes such as side-chain rotations or loop motions [119] [120]. But running longer simulations is not enough to explore a larger conformational space for biomolecules since most of the simulations are just exploring a small region close to some energy minimum which is closest to the initial configurations. One answer to this challenge can be to run in parallel multiple simulations with different initial configurations [121] and then use mathematical tools to analyze the different trajectories and cluster the results [122] [123]. Over the past 50 years, the field of computer simulations has evolved from studying a system of a few independent particles in vacuum for a few picoseconds to studying complex biomolecular systems with millions of particles with or without explicit solvent for millisecond timescales. The field is still evolving, with the advent of Graphical Processing Units (GPU) which in essence contain thousands of computational cores in a cheap, extremely parallelizable architecture that is very efficient for computational processes such as those used in MD simulations [124] [125].

MD simulations of biomolecules can be performed with an atomistic description, or with coarse grain (CG) models. In the latter, atom groups are portrayed as single beads or interaction sites, such that longer timescales or larger systems can be studied. Such methodologies have been applied to, e.g., the study of lipids [126] [127].

MD calculations can explore the energy landscape and find the nearest energy minimum and the geometry that corresponds to that minimum for a given 3D structure that serves as the initial conditions. For that purpose, different optimization algorithms can be used. This is the energy minimization step. The heating phase follows where each atom is given an initial velocity at low temperature and Newton's equations of motion are integrated (Equation 2.1). At a slightly higher temperature, periodically, new velocities are assigned. The simulation is continued until the desired temperature is reached (typically, room temperature). At the equilibration phase that follows, the system is monitored until properties such as structure, temperature, pressure, and energy are stabilized. If a significant temperature fluctuation occurs, velocities are adjusted to restore the intended temperature. During the production phase of the run, the system is simulated for a length of time, which may extend to hundreds of picoseconds, nanoseconds, or longer, and a set of positions (coordinates) for each atom (or CG bead) of the simulation system is obtained as a function of time – the MD trajectory. This is done by solving Newton's second law of motion, which in its differential form reads as follows:

$$m_i \frac{d^2 x_i}{dt^2} = F_i, \qquad i = 1, \dots, N \tag{2.1}$$

The forces can be calculated by differentiating the potential function $E(x_1, x_2, x_3, \ldots, x_N)$ acting on a particle of mass m_i along some coordinate x_i.

$$F_i = -\frac{\mathrm{dE}}{dx_i} \tag{2.2}$$

After each time step, this loop is repeated from the new positions. To ensure accuracy, the time step needs to be less than the fastest dynamics of interest being observed. For example, the fastest motion of an H atom is a vibration with a characteristic time of 13 fs (1 fs = 10^{-15}s of computational time), and generally, a time step in the order of 1 femtosecond is used. Moreover, since the processes of interest (e.g., protein conformational dynamics, ion transport by ion channels and ion pumps) have characteristic time scales in the range of hundreds of nanoseconds, hence, hundreds of millions of time steps are needed for such simulations.

Using concepts and methods from statistical mechanics one can obtain macroscopic observables such as pressure, energy, etc. from the trajectories. By use of MD, a system's macroscopic properties can be explored through simulations on the molecular level, similar to examining the energetics and mechanics of changes in conformation.

2.1.1 MD ensembles

Typically, a system's thermodynamic state is defined by a group of parameters such as the number of particles *N*, temperature *T*, and pressure *P*. The atomic positions and velocities define the microscopic state of the system, which exists in phase space (multi-dimensional). The phase space of a system with *N* particles is a *6N*-dimensional space, where each point represents a system state. Moreover, an ensemble is a group of phase-space points that satisfy specific requirements for a thermodynamic state. Every MD run produces trajectories, which are a sequence of points of the same ensemble representing different conformations of the system.

The ergodic hypothesis suggests that if a system is observed over sufficiently large time scales, the ensemble is representative of the system. An average over all copies in the ensemble permits the calculation of observed macroscopic properties. Hence, the ensemble average of a property A or its expected value can be calculated as a sum over all states *i:*

$$< A > ensemble = \sum_i A_i \, \pi_i \tag{2.3}$$

with π_i being the probability to find the system in state *i* while A_i is the calculated value of the property in the specific state *i*.

The probability function πᵢ is ultimately dictated by the choice of macroscopic properties which conform to all replicas of the system. For example, for a system to follow the canonical ensemble, its temperature (*T*), volume (*V*), and number of particles (*N*) must remain constant. This implies that there is no exchange of mass or particles with the environment (constant *N*), and that the system is coupled to a thermostat (constant *T*) while its size remains fixed (constant *V*). The canonical ensemble is also known as the *NVT* ensemble. In this case, the probability function πᵢ can be defined as:

$$\pi_i = \frac{e^{-\beta E_i}}{Z_{NVT}} \qquad (2.4)$$

with E_i being the total energy in state i, Z_{NVT} is the partition function corresponding to the canonical ensemble and β is the inverse of the thermal energy $\beta = \frac{1}{k_\beta T}$.

Several statistical ensembles can be defined by varying the thermodynamic variables measured as well as the exchanged quantities of the system with the environment:

- *NVT* or Canonical ensemble: The volume (*V*), temperature (*T*), and number of particles (*N*) are all kept at a constant level. This means that to keep the temperature constant, energy is exchanged with the surrounding environment. However, there is no exchange of mass as the number of particles remains constant. Also, the size of the simulation box remains fixed throughout the simulation, implying that its volume does not change.
- *μVT* or Grand canonical ensemble: The system is open with constant temperature, meaning that there is both mass and energy exchange. In such a system, the chemical potential *μ* is constant as well as temperature *T* and volume *V*.
- *NVE* or Microcanonical ensemble: The system is isolated with no mass or energy exchange. The number of particles *N*, volume *V* and energy *E* remain constant.
- *NPT* or Isothermal-isobaric ensemble: The system is in an isolated environment with constant pressure *P*, temperature *T* and number of particles *N*. The system is connected to a thermostat to allow for energy exchange.

2.1.2 Force Fields and the Potential Energy Function

For the algorithm to be able to calculate the forces one first needs to define the potential energy function, *E(r)*. Most MD algorithms describe the potential energy based on five components, with the following basic form:

$$E(r) = E_{bonded} + E_{non-bonded} \tag{2.5}$$

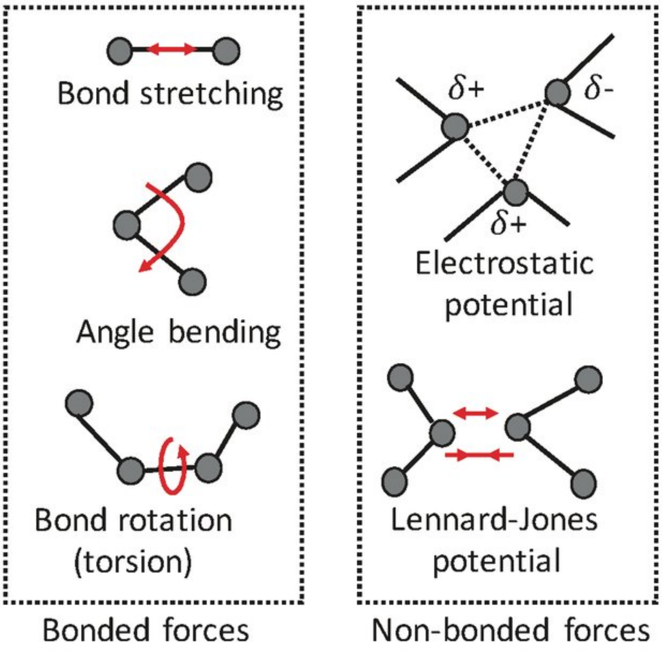


Figure 2.1: Different types of interactions considered in MD algorithms. Typically, the interactions are separated into bonded and non-bonded interactions. The image is from ref. [128].

2.1.2.1 Bonded Interactions

The term corresponding to the bonding interactions describes torsional, bending, and stretching interactions (Figure 2.1). Such interactions are calculated in terms of the deviation from an equilibrium value of a dihedral angle ψ or a bending angle θ or the bond length r, respectively.

$$E_{bonded} = E_{bonds} + E_{angles} + E_{dihedrals} + E_{impropers} \quad (2.6)$$

Bond potentials

Two atoms are termed bonded once they form a chemical bond. Typically, the length of the bond is not constant but slightly oscillates around an equilibrium value. Such interaction is called the bond stretching potential or bond potential. The bond potential can be described in a harmonic form yielding the required increase in energy as the bond length b deviates from the equilibrium value b_0 (Figure 2.1) [129] [130].

$$E_b = k_b \, (b - b_0)^2 \quad (2.7)$$

Here, at each instant, b is the bond length, b_0 is the equilibrium distance between the bonded atoms, and k_b is essentially the amplitude of the harmonic oscillator called the force constant.

Valence angle potentials

The change in conformation of the valence angle is defined between three bonded atoms. Similar to the bond stretching discussed above, the angle oscillates around an equilibrium value θ_0, and this motion is called angle bending (Figure 2.1). The angular potential or valence angle potential describes the deviation from equilibrium and typically has harmonic form [129] [130].

$$E_\theta = k_\theta \, (\theta - \theta_0)^2 \quad (2.8)$$

Here, at each instant, θ is the angle value, θ_0 is the equilibrium angle value, and k_θ is the amplitude of the angular potential or the angular force constant.

Torsional potentials

Torsional potentials are between four atoms that form a dihedral angle (Figure 2.1). Torsional potentials express the deviation of dihedral angles from some equilibrium values due to bond bending or rotation. In general, the torsional potential can be related to the flexibility of molecules. Thus, including torsional potential is of paramount importance in modeling 'organic' amorphous molecules. The torsional terms can be separated into proper and improper dihedrals. Proper potentials are described by a cosine function:

$$E_{\varphi} = k_{\varphi}[1 + \cos(n\varphi\text{-}\delta)], \quad n=1,2,3,4,6 \tag{2.9}$$

with φ being the plane angles formed by the first and last three atoms, δ is the minimum angle energy, n is a term for periodicity and k_{φ} is a force constant.

The improper dihedrals maintain the planarity of specific atoms as well as keep the chirality in reference to a tetrahedral heavy atom. The potential is described with a harmonic function:

$$E_{\omega} = k_{\omega}\ (\omega - \omega_0)^2 \tag{2.10}$$

with ω the instantaneous angle between the plane formed by the central atom and two peripheral ones and the plane defined by peripheral atoms.

2.1.2.2 Non-Bonded Interactions

The non-bonded terms typically represent electrostatic and van der Waals interactions. MD algorithms calculate such terms between all atom pairs (within a specified cutoff distance). These atom pairs can either belong to different molecules or the same one, however, in the case where atoms belong to the same molecule, they should be separated by at least three bonds.

$$E_{non-bonded} = E_{vdw} + E_{elec} \tag{2.11}$$

Such calculations are the most time-consuming step in the MD algorithms since they contain several long-range interactions [131].

van der Waals potentials

The most used potential to describe van der Waals (vdw) interaction is the so-called 12-6 Lennard-Jones (LJ) pair potential which has the following form:

$$E_{vdW} = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2.12}$$

The collision diameter, σ, is the distance at which no potential exists between two particles. Essentially, σ determines the closest distance between two non-bonded particles referred as the vdW radius. The strength of attraction between two particles is indicated by ε, which is the depth of the attractive well in the van der Waals interaction. Lastly, r represents the separation between the point particles (Figure 2.2).

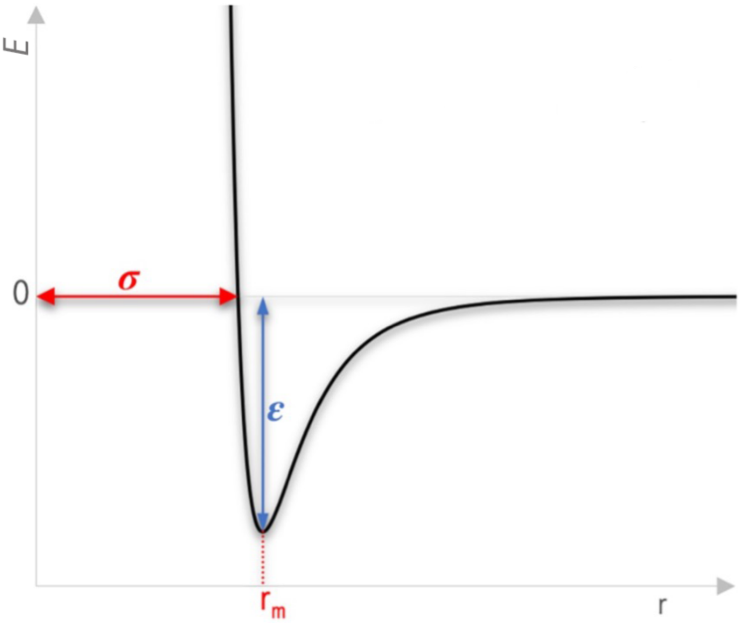


Figure 2.2: Lennard-Jones potential represented schematically, and its parameters are the attractive strength ε and the collision diameter σ. The image is adapted from ref. [132].

In the 12-6 LJ potential, the r^{-12} term provides the repulsive part of the potential, which becomes dominant as the separation distance tends to zero. The r^{-6} term provides the attractive part of the potential which is dominant at larger distances. The schematic in Figure 2.2 shows that the LJ potential approaches zero as the separation distance increases. Interactions are termed short-ranged if they decay faster than r^{d}, with d being the dimensionality of the system, thus vdW interactions are considered short-ranged [133]. In MD simulations short-range interactions are typically calculated over pairs within a specific distance, called the cutoff distance, around each atom. The potential calculations are neglected beyond the cutoff radius, making the simulation algorithm much more efficient.

Electrostatic potentials

The electrostatic potential describes the interaction between two charged particles and follows Coulomb's law:

$$E_{Elec} = \frac{q1q2}{4\pi\varepsilon_0 r_{12}} \tag{2.13}$$

with q_1 and q_2 are the charges carried by the two atoms in the pair and r_{12} is the separation distance. ε_0 is the electric susceptibility of the vacuum (ϵ_0=8.854 × 10^{-12} $C\ ^2J\ ^{-1}mol^{-1}$).

Electrostatic interactions are long-ranged interactions since they decay as r^{-1} and at very large distances tend to zero. In order to calculate the pair interaction of charged particles in a simulation box, the Ewald sum method is used [134]. This method is efficiently running a sum for all interactions between a particle and its periodic image [133]. Similar to vdW potentials, the electrostatic potentials are neglected for distances larger than some cut-off radius.

2.1.3 Periodic Boundary Conditions

The term periodic boundary conditions (PBC) is commonly used in MD simulations. Using such methods, the simulation's box size is effectively increased, enabling the simulation of a small number of atoms to produce results that can be valid in the thermodynamic limit, where statistical mechanics concepts can be used to relate the microscopic results to observed macroscopic thermodynamic quantities [135] [136].

The idea behind PBC is that there is an infinite array of copies or images of the simulation box that repeat periodically in each of the three dimensions. Since this concept is based on periodicity, images are integer multiples of the atomic coordinates from the original simulation box. An image particle enters from the opposite side of the simulation box to replace a particle that leaves through a boundary (Figure 2.3). For such an approximation to work, the simulation box must be chosen wisely in order to avoid self-interaction through the boundary since this will induce the so-called 'finite size' effects on the simulation [135].

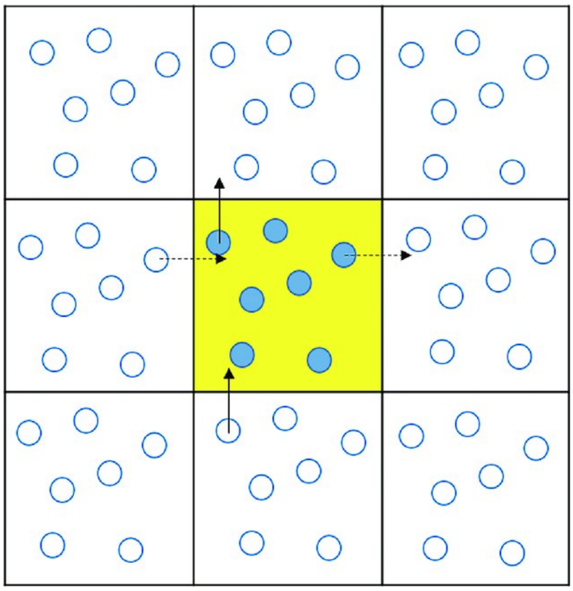


Figure 2.3: Representation of the periodic boundary conditions. For illustration purposes, the idea is shown in two dimensions. The initial simulation box is represented by the central yellow box, and the particles inside it are depicted by filled circles. However, open circles portray the periodic images of these particles in different cells. Bold and dashed arrows indicate the movement of particles in close proximity to the boundary. As a result, while one particle exits the box in one direction, another particle enters the box from the opposite direction. The image is from ref. [137].

2.1.4 Integration Scheme

In MD algorithms the choice of an integration scheme is based on two requirements: conservation of energy and momenta and computational efficiency. Most commonly, Verlet integrators are used, while other approaches are less frequently used and will not be discussed here.

In the Verlet method [138], the first assumption is that positions *r(t+δτ)* and velocities *v(t+δt)* are approximated by Taylor series

$$r(t+\delta t) = r(t) + v\delta t + \frac{a(t)}{2}\delta t^2 + \frac{b(t)}{6}\delta t^3 + \cdots \quad (2.14)$$

$$v(t+\delta t) = v(t) + a\delta t + \frac{b(t)}{2}\delta t^2 + \cdots \quad (2.15)$$

with α(t) being the acceleration and b(t) = $\dot{\alpha}$(t). Equivalently, r(t-δt) is

$$r(t-\delta t) = r(t) - v\delta t + \frac{a(t)}{2}\delta t^2 - \frac{b(t)}{6}\delta t^3 + \cdots \quad (2.16)$$

By considering that $\vec{F} = m\vec{a}$ and substituting the acceleration accordingly but also adding or subtracting equation 2.16 from equation 2.14 yields

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + m^{-1}F(t)\delta t^2 \qquad (2.17)$$

$$v(t) = \frac{r(t+\delta t) - r(t-\delta t)}{2\delta t} \qquad (2.18)$$

This is called the Verlet algorithm, whose advantages are that it is quite simple, time-reversible, and energy-conserving. The distadvantage of this algorithm is that the expressions used for velocities and positions require differences between large and similar numbers, which can lead to numerical inaccuracies. A variation of this method known as the velocity Verlet algorithm [139] takes care of such disadvantages by replacing the substractions in equations 2.17 and 2.18 with sums. Such a method is preferred when using finite precision computers. The expressions for velocities and positions for the velocity Verlet algorithm are:

$$v\left(t+\frac{\delta t}{2}\right) = v(t) + m^{-1}F(t)\frac{\delta t}{2} \qquad (2.19)$$

$$r\left(t+\frac{\delta t}{2}\right) = r(t) + v(t+\frac{\delta t}{2})\delta t \qquad (2.20)$$

$$v(t+\delta t) = v(t+\frac{\delta t}{2}) + m^{-1}F(t+\delta t)\frac{\delta t}{2} \qquad (2.21)$$

The velocity Verlet algorithm is employed in the NAMD code which was used to simulate the systems presented in this thesis.

2.1.5 Short-Range Force Contributions

In MD simulations, most of the computational cost comes from the calculation of non-bonded interactions. In principle, all possible pairwise interactions in a box should be calculated so the time needed for each calculation scales as $O(N^2)$. However, LJ or similar potentials are typically short-ranged (potential decays as $1/r^6$), thus calculating the interaction for distant pairs will not affect the results. It is common, in order to reduce the computational cost, to impose a cutoff distance beyond which the potential is set to zero and the interactions ignored [135] [140]. Conventionally, in PBC-bound systems, the cutoff is such that each atom will only interact with one image of each atom in the box.

However, just applying a cutoff distance to the pair potential calculation does not lead to an important gain in computational efficiency and this is because the distances of all pairs must be calculated and compared with the cutoff distance, which introduces additional calculations, specifically *N(N-1)* additional calculations. To sidestep this problem, algorithms take advantage of the fact that each atom's neighbors are unlikely to change over 10-20 timesteps. Hence, it is easy and computationally efficient to create lists with neighboring atoms with their distance from a central atom being smaller than the cutoff distance. In this manner, distance comparisons are not as frequently calculated. In several MD algorithms, atoms just beyond the cutoff distance are also kept track off, and are being used in the pair potential calculations only if they move within the cutoff distance [135, 140].

Introducing a cutoff distance in the calculations makes the potential energy discontinuous at the cutoff and hence the force. To prevent such problems, most MD algorithms multiply the potential with a potential that reaches zero in a smooth manner close to the cutoff. This switching of the potential is only introduced close to the cutoff.

2.1.6 Long-Range Force Contributions

The contribution to the potential due to electrostatic interactions is such that a cutoff distance cannot be applied without introducing false dynamics. Thus, the full electrostatic calculation scales as $O(N^2)$. Once PBC is introduced the system can be approximated as infinitely periodic, which can be used to calculate the electrostatic potential. The Ewald summation method, which was first introduced in 1921 [134] is based on the decomposition of the potential into short- and long-range contributions. The long-range contributions are efficiently calculated as a sum over the Fourier transforms of the potential and the charge density, which rapidly converges and can be truncated with small numerical error. This truncation significantly decreases the computational workload.

In the Ewald summation method, it is required that atomic charges are screened to diminish rapid changes at small separation distances which leads to the real space summation converging rapidly. This is achieved by introducing a Gaussian charge distribution of width β and equal magnitude centered at each atomic position, with the charge of the Gaussian being opposite to that of the atom. To correct for the charge screening an oppositely charged but otherwise identical Gaussian distribution is introduced into the second term. However, the second term cannot be summed efficiently in real space and is thus Fourier transformed, summed in the reciprocal space where the summation can be efficiently done, and then converted back to real space by means of the inverse Fourier transform. Finally, a third term is introduced to correct for the self-interaction of the Gaussian distributions. The Ewald summation method, even though it is more complicated, yields a gain in computational efficiency since it converges as $O(N^{3/2})$ instead of $O(N^2)$ scaling of the simple Coulombic summation.

The Ewald summation method can be improved by incorporating the particle-mesh Ewald (PME) method [141]. The PME method essentially interpolates charges on a 3D mesh which accelerates the summation of the second term of the Ewald summation method. PME scales as *O(N lnN)* even better compared to the original Ewald summation allowing for routine simulations of charged systems without cutoff for systems under PBC. One can also fine-tune the width of the Gaussian distributions since increasing the width leads to faster convergence of the sum in real space but slows down in reciprocal space.

2.1.7 Multiple Time Step Algorithms

Multiple time step algorithms (MTS) are highly efficient since they take advantage of the fact that long-range forces typically vary slowly with time. Thus, an algorithm that calculates such long-range forces less frequently than short-range forces which are varying rapidly with time will be more computationally efficient [142] [143]. For most MTS algorithms, forces are divided into three categories. The first category is one of the rapidly updated forces, which typically includes bonded forces (typically updated every femtosecond). The second category includes forces generated from non-bonded interactions within a specific distance (usually the cutoff distance). These interactions are typically short-range interactions (LJ-like or electrostatic interaction of close atoms) and are not updated as often. The third category includes long-range forces, such as electrostatic forces between distant atoms that are not updated frequently.

2.1.8 Thermostat Algorithms

A thermostat is an algorithm that keeps the temperature of the simulation box fixed [144-146]. Thermostats are typically used to: (i) match experimental conditions (ii) conduct temperature-dependent studies of different processes or phenomena, such as the determination of the glass transition temperature, and (iii) facilitate equilibration or conformational search with, for example, high-temperature dynamics or simulated annealing. The reference temperature T_0 is defined as the temperature of the heat bath with which the system is in touch. A sufficiently good thermostat algorithm should keep the temperature of the system close to the required one at a given timescale.

The widely used Nosé-Hoover thermostat obeys the following equations:

$$\vec{r_i} = \frac{\vec{p_i}}{m_i} \qquad (2.22)$$

$$\vec{p_i} = \vec{F_i}(\vec{r_1}, \ldots, \vec{r_1}) - \zeta\vec{p_i} \qquad (2.23)$$

$$\dot{\zeta} = \frac{1}{Q}\left[\sum_i \frac{\vec{p_i^2}}{2m_i} - 3Nk_BT\right] \qquad (2.24)$$

with ζ being a thermodynamic friction coefficient that drives the exchange of heat among the system and the thermostat bath while Q dictates the strength of the thermostat. By following these equations, the thermostat allows us to study the system and its evolution within the canonical ensemble. Comparing if a specific thermostat will yield more realistic or more accurate dynamics is empirical, it is within reason to assume that: (i) thermostats that fix temperature to a specific value are less likely to describe the dynamics realistically in comparison to thermostats that allow the temperature to slightly fluctuate around the reference value. (ii) thermostats that allow for temperature fluctuations will realistically describe the dynamics when the fluctuations take place within the simulation timeframe and the dynamics are following a continuous velocity trajectory (smooth dynamics).

2.1.8.1 Langevin Thermostat

In the Langevin thermostat, the motion of large atoms is considered through a continuum of smaller atoms. The larger particles impulse smaller particles which in turn build a damping force to the momenta, $-\gamma p_i$, where γ is the Langevin friction coefficient applied to system atoms, *i.* The smaller atoms have kinetic energy and interact randomly with the larger particles. The Langevin equation for a single particle is given by:

$$m_i \frac{d^2x_it}{dt^2} = F_i\{x_it\} - \gamma_i \frac{dx_it}{dt} m_i + R_it \qquad (2.25)$$

The second term represents the frictional damping force that is applied to the particle with frictional coefficient $\gamma_i m_i$. The third term represents Gaussian distribution random forces that act on the particle (e.g., solvent interaction). Combining the two terms, the kinetic energy is maintained to keep the system at a constant temperature and the correct canonical ensemble is given. In simulations presented in this thesis, the Langevin dynamics scheme is used. High damping constant can significantly slow the dynamics of the system. We use a collision frequency of 5 ps^{-1}.

2.1.9 Barostat Algorithms

Andersen in 1980 described a way to extend MD algorithms to systems that cannot be described by the microcanonical (*NVE*) ensemble [145]. It was proven by Andersen that with a modification to the Langrangian of the system one could impose a constant external pressure *P* on the system. In that case, the system volume *V* is not constant but fluctuates to keep the mechanical equilibrium between the internal (system) pressure and the external (applied) one. In essence, the volume dynamics of a system are regulated by the barostat, which is analogous to having a free "piston" of any "mass" in the system. Although the ensemble averages are not dependent on the piston mass, they affect the response time to volume fluctuations.

The approach by Andersen was modified later to include a thermostat [145]. Nosé [144] and Hoover [147] [148] proposed an *NPT* (isobaric-isothermal) algorithm which was based on Andersen's piston's method for keeping pressure in check while it was using Nosé's thermostatting. Other algorithms for the *NPT* ensemble were developed later on, however, the choice is subjective to the observed results.

In this work, we use the Langevin piston Nose-Hoover method, which combines the Nose-Hoover method for constant pressure [149] with piston fluctuation control using Langevin dynamics [150]. The algorithm will drive the system to pressure equilibration (i.e., external pressure will be equal to internal pressure). The approach is similar to the analogous Nosé-Hoover thermostat which uses a separate set of variables.

2.2 Graph theory

Networks and connections are everywhere in our everyday life: rail tracks and road networks, landlines, internet lines, electronic circuits, and even molecular bonds. There also exist social networks including friends and families. All these examples can be portrayed using graph theory.

Graphs were first mentioned in 1736 by Leonhard Euler in his publication on the Seven Bridges of Königsberg. The city of Königsberg included two islands which were situated in the two rivers on both sides of the city. Both islands were connected to each other but also to both sides of the river. The problem that required a solution was to find a path in which each bridge was crossed only once, with the starting and ending points not necessarily the same. Nowadays, such a problem is known as proving the existence of a Eulerian path. Regarding the case of Königsberg, Euler proved in his paper that such a path could not exist and gave birth to basic graph theory [151].

Euler's solution was to model the two islands and each side of the river as points (nodes or vertices) while the bridges were lines (edges). A simple definition of a graph *G* is *G= (U, R),* suggesting that graph *G* is defined by a set of edges (*R*) and vertices (*U*) as shown in Figure 2.4. Using this simple representation, one can solve problems like the Eulerian path problem or the Hamiltonian path problem, which are closely related since they require a path to go to every single node of the graph.

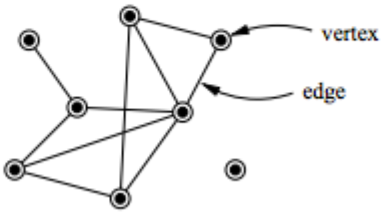


Figure 2.4: Example of a graph consisting of eight vertices and ten edges. The image is from ref. [152].

2.2.1 Graph types and algorithms

As more complicated graph-related problems arose, an extension of the definition of a graph was required [153-155]. These extensions include assigning an attribute to a vertex, a weight to an edge, or directionality to the edges. Imagine the map of a country. On this map, cities could be modeled as vertices, roads as edges and the weight of each edge could model the distance between vertices. Moreover, an attribute that is assigned to a vertex can be the population of each city or in the case of one-way streets, directionality to the edges can be assigned.

Once weights are assigned to the edges, the graph is called a weighted graph, and the weight indicates the cost of reaching from one node to the other. A weighted graph may include directionality or not.

In terms of mathematics, an edge can be expressed through a pair of vertices. For instance, an edge *R* with directionality, called a directed edge, connects two vertices, *v* and *u*, with the starting point being at *v* and the endpoint at *u*. This edge can be expressed as *R= (u, v),* or an edge with opposite directionality will be *R= (v, u).* A directed graph is also known as a digraph. Last but not least, an edge with no directionality or undirected edge can be expressed as *R={u,v}={v,u}.* In essence, an undirected edge consists of two directed edges pointing in the opposite direction of each other (Figure 2.5).

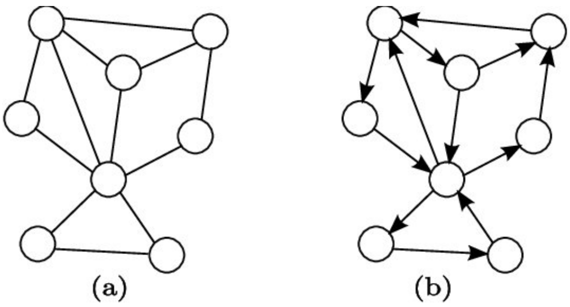


Figure 2.5: Examples of (a) undirected graphs and (b) directed graphs. Reproduced from [156].

In a graph, a vertex v can be connected to itself by an edge creating loops. If a pair of vertices gets connected by multiple edges, those edges are called parallel and, in that case, the graph is called a multigraph. A graph that has no loops and with one edge, at most, between any two vertices, is a simple graph.

The number of neighboring vertices that a vertex v has is its degree; this is indicated by the symbol deg(v). Every vertex in a simple graph with N vertices has a degree of:

$$\deg(v) \leq N - 1 \;\forall v \in G \qquad (2.25)$$

An edge can be formed between any of the vertices, but the vertices cannot form edges with themselves. Hence, deg(v) must be up to the number of vertices minus one, which excludes the self-vertex since it cannot form an edge by itself. If loops are included, then it is not a simple graph.

When all vertices of the graph have the same degree, the graph is called regular. A graph is called complete when only one edge joins each pair of vertices in the graph. In the case a vertex can be visited from any other vertex in the graph then it is a connected graph. In connected graphs, there is at least one path or edge between every couple of vertices while for a disconnected graph, any path doesn't exist between every pair of vertices.

A graph is called bipartite if the vertices can be divided into two separate subsets, *U1* and *U2*, that are not empty and vertices from the same subset are not connected by an edge. Instead, each edge links a vertex in *U1* to a vertex in *U2,* creating a partition between the two subsets. The bipartition of G is simply $U = U1 \cup U2$. A bipartite graph can be considered complete in which every vertex in the U_1 set is joined to each vertex in U_2 by one edge (Figure 2.6).

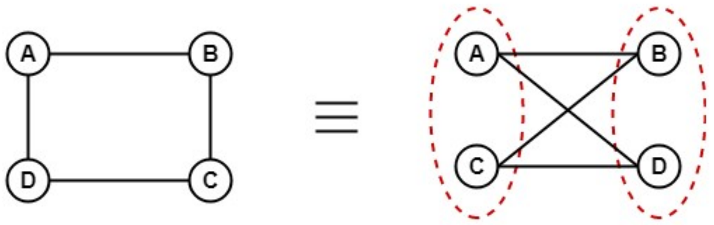


Figure 2.6: Schematic of a bipartite graph. The image is from [157].

A graph is called a planar graph when a plane can be illustrated without any edges crossing except at a vertex to which they are incident. A graph that does not satisfy these conditions is called a non-planar graph.

A graph is termed path or linear graph when its vertices are listed in series, i.e., u_1, u_2, ...u_n, and in a manner where the edges are {u_i, u_{i+1}} with i=1,2…., n-1. In other words, a path with at least two vertices has two vertices with degree 1 (terminal) and is connected, while all others have degree 2 (see Figure 2.7).

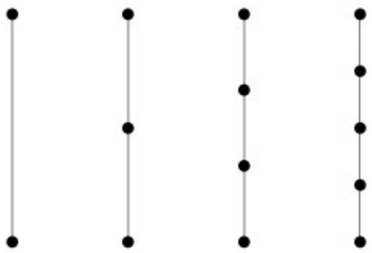


Figure 2.7: Representation of different linear graphs. The image is from [158].

Star graphs are complete bipartite graphs. A star graph with n vertices is expressed as S_n. The central (core) vertex has a degree of n-1, while all other vertices have a degree of 1 (Figure 2.8).

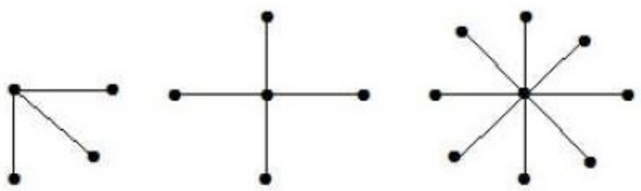


Figure 2.8: Different depictions of star graphs. Image is from [159].

A graph is called a cycle graph if it has n vertices and n edges that form a closed path and is denoted as C_n. In cycle graphs, the degree of each vertex is 2 (Figure 2.9). If a graph does not contain any cycles, it is called acyclic.

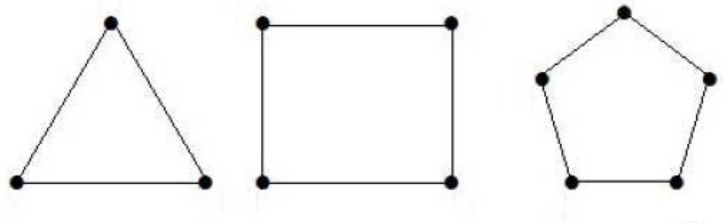
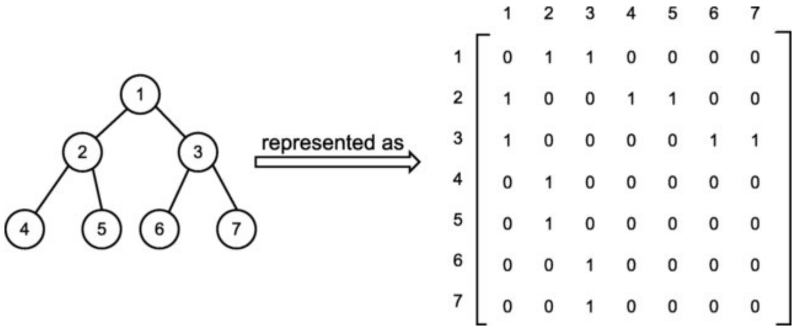


Figure 2.9: Various types of cyclic graphs. Image is from [160].

A graph can be described through its adjacency (or connection) matrix. Each row and column of the adjacent matrix represents a pair of vertices (v_i, v_j) and has a value of 1 if the vertices vi and vj are adjacent and 0 otherwise (Figure 2.10). Thus, for a simple graph without self-loops, all diagonal elements must be 0. Moreover, for an undirected graph, the adjacency matrix will be symmetric.



$$\begin{array}{c|ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 7 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}$$

Figure 2.10: An example of a graph and its representation as adjacency matrix. The image is from [161].

2.2.2 Shortest path problem

2.2.2.1 Dijkstra’s Algorithm

As we’ve seen above, a path is a series of edges joining a sequence of vertices, either finite or infinite. Computing the shortest path over a network is of interest to several different areas. In this problem, the target is to detect the shortest path. This essentially means that a sequence of adjacent nodes must be found, with minimal total edge cost [162]. Typically, computational efficiency is important making the problem more challenging.

Several algorithms have been suggested for detecting the shortest path, thus, it is important to consider which algorithm one chooses for a specific problem since it can vary based on the application at hand. For example, in transportation planning, several thousand of shortest paths might need calculation, and this is over an area covered by millions of nodes. Hence, detecting and storing all possibilities in the computer memory is not practical since Terabytes of memory might be required in a short amount of time. In other cases, the nature of the route costs can make this approach quite impractical [162]. Such algorithms have been applied by several researchers to a variety of problems such as the one discussed above [163] [164].

The Dijkstra algorithm is the original algorithm for identifying shortest paths [165-167] and was found to be one of the most efficient algorithms confirmed by benchmarking studies [168]. The algorithm computes the shortest path in a graph G with weighted edges w from a source node s to a destination node t (Figure 2.11).

Its working principle is based on an iterative process during which it examines the closest but not-yet-examined node to the starting node. Then it adds its successors to the set of nodes that are being examined. This step effectively divides the graph into two sets: S and S'. The nodes belonging to set S have the shortest path to the starting node, while the nodes belonging to set S' do not have the shortest path detected. In the first step, S' includes all the nodes of the system. As the iterative process continues, nodes are removed from S' and moved to S. Nodes are added to set S based on a priority queue that assigns distance labels to remaining nodes in set S'. These labels represent the cost of the current shortest path to the start node. Let's consider node u, which has the highest rank in the priority queue. After reviewing it, we add it to S, and loosen its out-links. If the total of the distance label of u plus the cost of the out-links (u, v) is less than the distance label for v, we update the estimated distance of node v. After iterating, the algorithm examines the node at the top of the queue until the queue is empty or the goal is achieved. By finding the shortest path trees from a single source to all other nodes, this algorithm solves single-source shortest path problems.

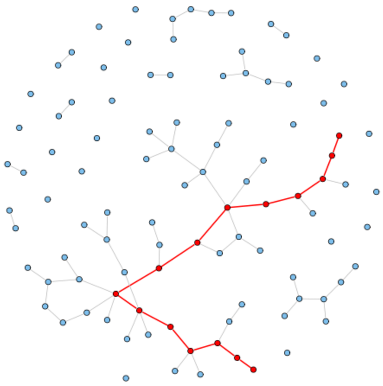


Figure 2.11: The shortest path connecting a source and an end node. The image is from [169].

The pseudo-code of Dijkstra's algorithm [170] is below.

```
“Function Dijkstra (G, start)
    $d[start] = 0$
    $S = \emptyset$
    $S' = U \in G$
    while $S' \neq \emptyset$
        do $u = Min(S')$
        $S = S \cup \{u\}$
        for each link $(u, v)$ outgoing from $u$
            do if $d[v] > d[u] + w(u, v)$ // Relax $(u, v)$
                then $d[v] = d[u] + w(u, v)$
                $Previous[v] = u$”
```

2.2.3 Centrality measures

In most networks, some of the vertices or edges can be more important or influential than others. For this centrality to be quantified, indexes were introduced. Jordan was the first to introduce the concept of the centrality of graphs [171]. Several ways exist to quantify the relative "importance" of a network node, hence different motivations will lead to different centrality measures. Centrality has been used in several fields such as chemistry [172], psychology [173], sociology [174], geography [175], game theory [176], and several other fields.

2.2.3.1 Betweenness Centrality

One of the most popular indicators of a node's centrality, which was introduced by Anthonisse [177] but was popularized by Freeman [178], is the betweenness centrality. The betweenness measure is a metric that indicates how often a vertex lies on the shortest path between two other vertices. In communication networks, a vertex can gain authority or significance by controlling the information flow (Figure 2.12). A vertex can either block or facilitate information flow between all vertices {x,y} where there is a unique path connecting them. Naturally, one can define the number of such {x,y}= pairs as the betweenness index *B(v).* For a given graph G and two vertices called s and t, the value of $\sigma_{s,t}$ represents the number of paths between s and t that have a length of *d(s,t).* When we consider a vertex v that is not part of {s, t}, then $\sigma_{s,t}$ represents the number of shortest paths between s and t that pass through v. Therefore, the expression $\sigma_{s,t}(\mathrm{v})/\sigma_{s,t}$ yields the number of shortest paths between s and t that pass through the vertex v.

$$B(v) = B(v; G) = \sum_{s \neq v \neq t} \frac{\sigma_{s,t}(\mathrm{v})}{\sigma_{s,t}} \qquad (2.25)$$

The normalized *BC* of node v_i can be calculated by dividing its betweenness value by the total number of node pairs in the system that exclude v_i.

For the normalization, in undirected graphs with *N* nodes, the *BC* is divided by $\frac{1}{2}(|N|-1)(|N|-2)$ and by $(|N|-1)(|N|-2)$ in directed graphs. In directed graphs, the normalization term is double that used in undirected graphs. This is because, in a directed network, the path s to t could be different from the path from t to s.

2.2.3.1 Degree centrality

One of the most basic centrality measures is the degree of centrality (*DC*) of a vertex. It is defined as the degree of a vertex (deg(v)) but only applies to undirected graphs. For directed graphs, there are two ways to calculate the degree of centrality: in-degree (deg-(v)) and out-degree (deg+(v)). It's worth noting that the degree of centrality is a local measure since it only takes into account the number of neighbors a vertex has (Figure 2.12). When Freeman centralization is applied to degree centrality, it results in the following definition from [179]. The normalized value of *DC* of a given vertex v_i can be computed by dividing its *DC* by the maximum number of possible edges to it, which is *N*-1, with *N* being the total number of vertices.

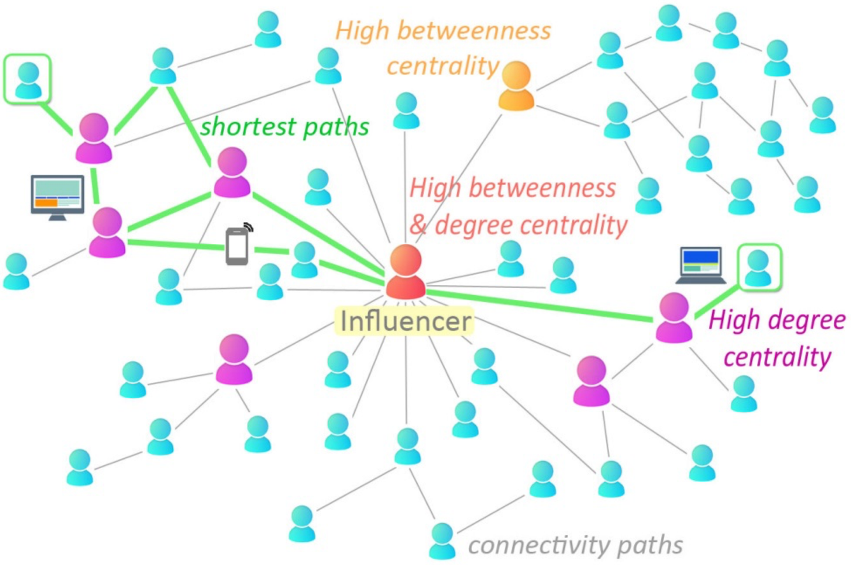


Figure 2.12: Depiction of different centrality measures that were used to analyze H-bond networks. A high degree centrality group will have many connections around it, while a high betweenness centrality group will be part of several short-distance paths. A group with high both degree centrality and betweenness centrality is considered an influencer or a connection hub. The illustration was created for the scope of this thesis.

2.2.4 Connected components

Detecting the Connected Components (CC) is an integral preprocessing step for many graph algorithms. In an undirected graph *G = (U, R),* a connected component is a subset C that satisfies two requirements. Firstly, all vertices in C can be reached from any other vertex in C. Secondly, there are no edges connecting vertices of different components.

The problem at hand is to detect the number of such components in a graph, tag each component with a unique ID and label each vertex in the graph according to its component ID. A graph with multiple connected components is shown in Figure 2.13. When it comes to directed graphs, a Strongly Connected Component refers to a set of vertices that is maximal and ensures that every vertex in the set can be reached from any other vertex [180].

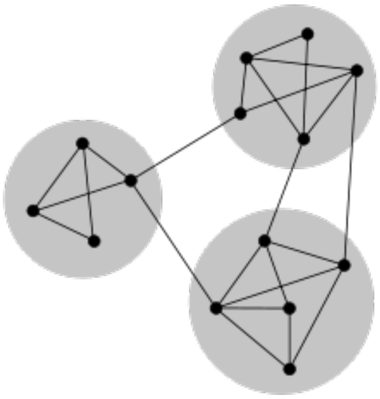


Figure 2.13: Representation of connected components in a graph. The image is from [181].

2.2.4.1 Depth-first search (DFS) algorithm

The DFS algorithm accepts as input a graph G and yields its predecessor subgraph. Moreover, it assigns two timestamps to each vertex, one corresponding to its discovery and one corresponding to the finishing time [180]. The algorithm, at first, considers each vertex as not discovered as well as sets the parent of each vertex to null. As the first step, the algorithm begins by selecting one vertex, marking it as discovered but unfinished, and giving it a discovery timestamp of 0. The algorithm works by recursively assigning an appropriate discovery time, represented by variable d[v], to each vertex in the set Adj[u] that has not been discovered yet. The algorithm increases the time variable at every step. If a vertex v has no discovery descendant, the algorithm goes back to investigate v's ancestor and assigns v the proper finishing time.

After all of u's descendants are completed, u is considered finished. The algorithm ends when there are no more active vertices in the graph; if there are, it loops back and continues.

The initialization of the algorithm takes a time complexity of Θ(n) because each vertex is examined once to be classified as "not yet discovered". While examining each vertex's adjacent vertices, the recursive part of the algorithm needs to cross each edge twice, giving it a time complexity of Θ(m). Thus, the total time complexity of the algorithm is Θ (m+n).

In this thesis, we use graph theory to implement algorithms that allow us to interpret the complex data sets that we derive from MD simulations for various biological systems. We convert the data into simple graphical representations, and we analyze the graphs by applying graph-theory algorithms. For example, H-bond networks, clustering of lipids or amino-acids, centrality measurements, long-distance pathways, shortest paths, the role of water through interactions with lipids or proteins, and the topology of interactions are some examples of the analysis that can be revealed by the algorithms presented in this thesis. Our representations are essential for the interpretation of data from experiments but also give a new perspective for the analysis of complex biological data to both experimentalists and theoreticians.

Chapter 3

“Το μυστικό της αλλαγής είναι να επικεντρωθείς, όχι στο να πολεμήσεις το παλιό, αλλά στο να χτίσεις το νέο” Σωκράτης

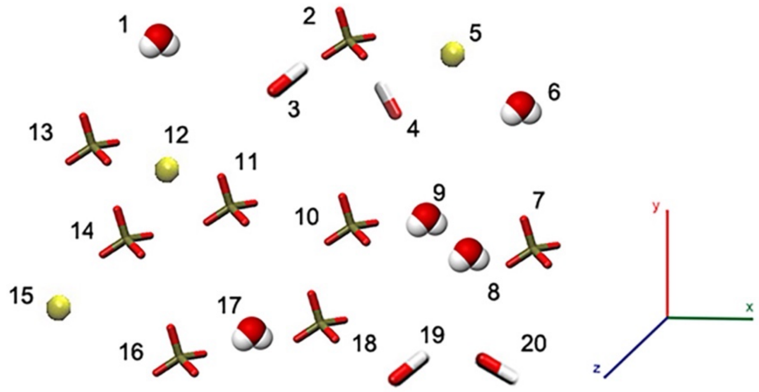
“The secret of change is to focus all of your energy, not on fighting the old, but on building the new.” Socrates

3 Methodology

Figures and tables presented in this chapter have been reproduced/reprinted with permission from J. Membr. Biol. 251 (2018): 461-473. Copyright © 2018 Springer Nature, J. Membr. Biol. 251 (2018): 609-631. Copyright © 2018, The Author(s) – Publisher Springer Nature, FEMS, 365, no. 13 (2018): fny124. Copyright © 2018, © FEMS 2018, J. Chem. Inf. Model. 59, no. 5 (2019): 1882-1896. Copyright © 2019, American Chemical Society, BBA-Biomembranes 1864, no. 4 (2022): 183859. Copyright © 2022 Elsevier B.V., COSB, 64 (2020): 79-87. Copyright © 2020 Elsevier Ltd., and J. Struct. Biol. 212, no. 2 (2020): 107617. Copyright © 2020 The Authors. Published by Elsevier Inc.

3.1 Analyses of lipid H-bond dynamics

We create undirected graphs $G = (U, R)$, where the nodes of the graph, U, are lipid molecules and R stands for the edges of the graph. We treat lipid phosphate groups as nodes in the H-bond network, and we calculate H bonds regarding all O atoms to be either H-bond donors or acceptors. The geometric criterion for the H-bond calculations is the distance between H and the acceptor heavy atom to be less or equal to 2.5 Å [22]. Direct H bonds between lipids, water-mediated bridges, or ion-mediated bridges between two lipids constitute the graph's edges. System coordinates are read for each step of the simulation, the H-bond criterion is applied and text-based tables containing information on H-bond partners and interaction distances are created in the initial step of our algorithm (Figure 3.2, Figure 3.3). At each step of the simulation, we monitor the distances between H-bonded pairs and use this information to create adjacency or connection matrices. These binary matrices are used to represent the connections between groups (Figure 3.1). They are set to 1 if every pair in the system is connected and 0 otherwise. We visualize our results using adjacency matrices, which allow us to create graphs with nodes representing lipid headgroups and edges representing H bonds between lipid headgroups.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Figure 3.1: Algorithm for detecting H bonds. (a) The coordinates of each molecule are used to detect H bonds between molecules and construct an adjacency matrix [22]. (b) Every lipid molecule that could form an H bond is regarded as a node in the network, and edges represent the presence or not (1 or 0) of H bonds connecting the nodes. For example, phosphate with index 2 H bonds with the hydrogens of waters with indexes 3 and 4. The procedure follows the MD trajectory and adjacency matrices are built for each time step of the simulation. The illustration was created for the scope of this thesis.

3.2 Identifying dynamic lipid clusters

A cluster, or network of lipids, is made up of a subset of nodes and edges, all of which are interconnected. Using adjacency matrices, for each node, all potential H-bond connections are recorded, and paths of connected nodes are created for each simulation step with the purpose of locating H-bonded clusters (Figure 3.3). The number of nodes interconnected by H bonds (direct between lipids or water-mediated bridges between lipid phosphates) and/or a sodium ion bridge determines the size of each network component, namely, the size of the lipid clusters.

To cluster H-bonded lipids, we implemented the Network Components algorithm, and we conducted Connected Component searches based on the Depth-First Search (DFS) algorithm [180] (Section 2.2.4.1), extensively employed in algorithms in graph theory. After the calculation of H bonds and of the adjacency matrix, a graph of connections is built. The algorithm starts from a starting/root node of the graph and finds all H-bond paths that start from that node. To locate all the components, it launches a new search from a node that was not included in a previously discovered component. The process is repeated until all nodes are reported as visited (Figure 3.2). In more detail, a binary vector of length n, where n denotes the total nodes in the network, was at first initialized with zeros to indicate that all nodes have not yet been visited. For each node j, if j wasn't visited before, connecting nodes to j were searched to the adjacency matrix. The search was continued repetitively for each connecting node to j, for finding their own interactions. All node connections were being added to a member's list and marked as visited nodes. Finally, when all nodes were visited, the total number and the size of each network component was given as a result by the member's list length (Figure 3.2).

As a result, in our graphs, we create paths of linked nodes and compute the number of clusters discovered and the list of H-bonded nodes. These results allow us to further analyze the dynamics of each cluster. For the network visualization, we create planar and circular connectivity graphs.

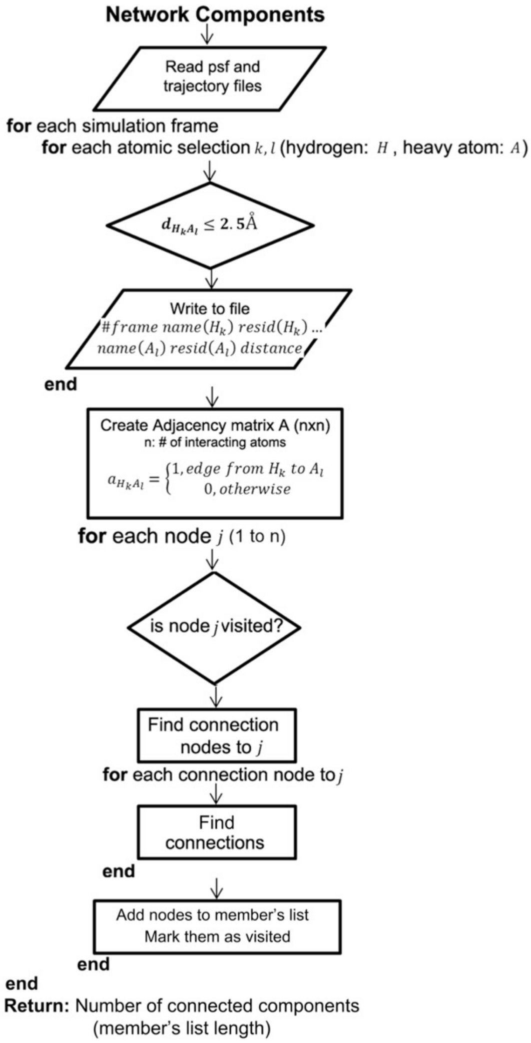

Network Components
Read psf and trajectory files
for each simulation frame
for each atomic selection k, l (hydrogen: H , heavy atom: A)
$d_{H_kA_l} \leq 2.5$Å
Write to file
$\#frame\ name(H_k)\ resid(H_k) \ldots$
$name(A_l)\ resid(A_l)\ distance$
end
Create Adjacency matrix A (nxn)
n: # of interacting atoms
$a_{H_kA_l} = \begin{cases} 1, edge\ from\ H_k\ to\ A_l \\ 0, otherwise \end{cases}$
for each node j (1 to n)
is node j visited?
Find connection nodes to j
for each connection node to j
Find connections
end
Add nodes to member's list
Mark them as visited
end
end
Return: Number of connected components (member's list length)

Figure 3.2: Flow chart of the network's component algorithm used for the clustering of lipid molecules. At the first step of the algorithm, the atomic coordinates are read for each simulation step. Then, the H-bond criterion is applied and tables with details on H-bond partners and interaction distances are written in files. From the resulted data, adjacency matrices are built. To find the network components, the DFS algorithm is applied until all nodes are discovered and marked as visited. The paths of linked nodes constitute the lipid clusters. The illustration was created for the scope of this thesis.

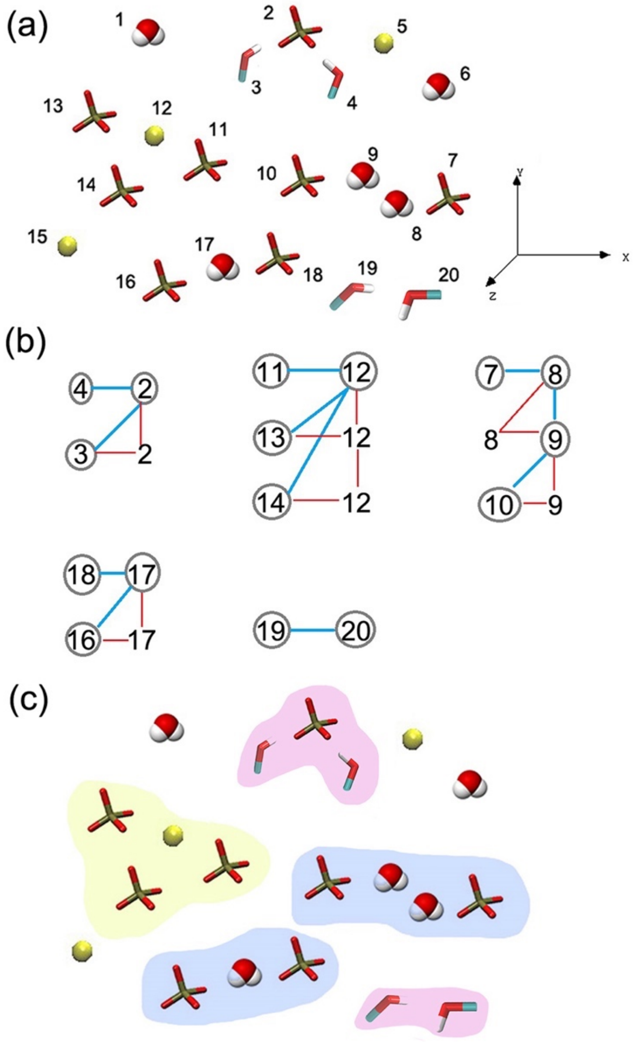


Figure 3.3: Algorithm process visualized to identify and characterize lipid clusters [22]. (a) The system's coordinates are read, H bonds are calculated, and adjacency matrices are constructed. (b) Paths of interacted nodes are calculated based on DFS, and each H-bonded cluster is found and displayed. (c) Visual representation of the ion-mediated interactions (in yellow), the direct H bonds (in pink), and water molecules (1-2) that interconnect lipid phosphates (in blue). For the molecular visualizations, VMD [23] was used. Image is from [22], a study presented in this thesis.

3.3 Water H-bond bridges between lipids

A computational approach was developed to detect H-bonded bridges among lipids and evaluate their dynamics. The steps of the algorithm are described below.

Water H-bond bridges between lipid phosphate atoms were investigated independently for bridges containing 1-5 waters. In the first step, waters were chosen in an H-bond distance from the oxygens of each lipid phosphate group. When searching for one-water bridges, the algorithm's second stage involves the detection of water hydrogens (H) that are H bonded to two different phosphate oxygens (O), at the same time. When looking for a bridge of length two to five, we select the waters picked in the first step as input to the second step of the algorithm and perform again the process of searching for H-bonded waters $k - 1$ times, where k indicates the maximum number of waters permitted in the bridge. The last water of each bridge is checked to H bond to another lipid's phosphate group. Following this process, water chains of various lengths that interconnect different lipid phosphates are formed. The algorithm searches backward the water H-bond connections and gives the paths of pairs of bridged lipids.

As a lipid pair can be H bonded with bridges of different lengths, our algorithm extracts distinct water chains of the minimum length. The H-bond occupancy of each bridge is calculated to assess the dynamics of the water bridges. The occupancy gives the percentage of the simulation time length during a specific water bridge is present.

The DFS algorithm was used to exclude paths that include cyclic connections instead of linear (Figure 3.4b). To detect cycles in a graph, the algorithm identifies back edges, which connect a node to one of its ancestor nodes [180]. The longest linear path length was kept after all resulting path lengths were computed and cycle edges were removed one at a time if a cyclic path was part of a path with a linear length greater than three.

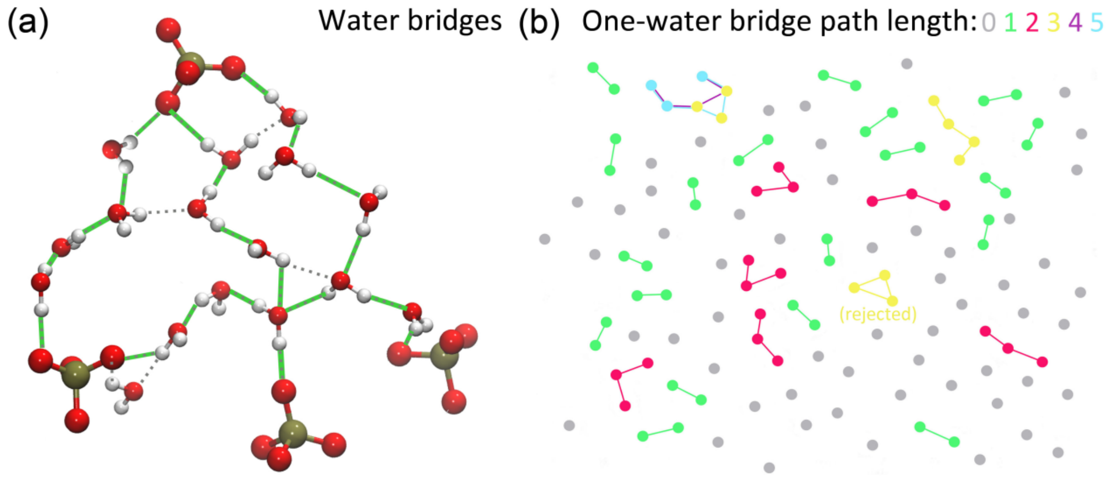


Figure 3.4: Waters interconnecting lipid phosphate groups. (a) The algorithm extracts distinct water bridges where the distance between the lipid pairs is the shortest. The optimal paths are depicted in green. (b) One-water paths link phosphate groups on the extracellular side of the

bilayer. Phosphorus atoms are shown as spheres and color-coded based on the length of the water bridges. Linear pathways are investigated, whereas cyclic paths are excluded from our computations [182]. Image is from [182], a study presented in this thesis.

3.4 Residence times of water H-bond bridges

To calculate the average lifetime of the water bridges that H bond lipids or protein amino acids, *NVE* simulations were performed. We computed the residence time correlation function $C_R(t)$ according to the formula [183]:

$$C_R(t) = \frac{1}{N(t)}\sum_{t_0}\sum_{j=1}^{N_w} p_{R,j}(t_0, t_0 + t) \quad (3.1)$$

where $p_{R,j}(t_0, t_0 + t)$ equals 1 if water j continuously H bonds lipid phosphate groups or protein amino acid residues over the time interval t, starting from an arbitrary time origin t_0, and 0 otherwise [183]. N_w stands for the total number of waters that can form H-bond bridges in the system. The function is calculated with a moving time origin t_0 starting from 0, an internal increment $\Delta t = 1\text{ps}$, and $N(t)$ is the number of fragments corresponding to each time length *t*. Each increment splits the sequence of the whole 1ns time to $N(t)$=$1000ps - \Delta t + 1$ segments (Figure 3.5).

The normalized residence correlation function $C_R(t)/C_R(0)$ is then adapted to a stretched exponential function (Kohlrausch– Williams–Watts, KWW),

$$\frac{C_R(t)}{C_R(O)} = e^{-\left(\frac{t}{\tau}\right)\lambda} \quad (3.2)$$

and the mean residence time is determined from the integral of the KWW [183].

$$\langle \tau_R \rangle = \frac{\tau}{\lambda}\Gamma\left(\frac{1}{\lambda}\right) \quad (3.3)$$

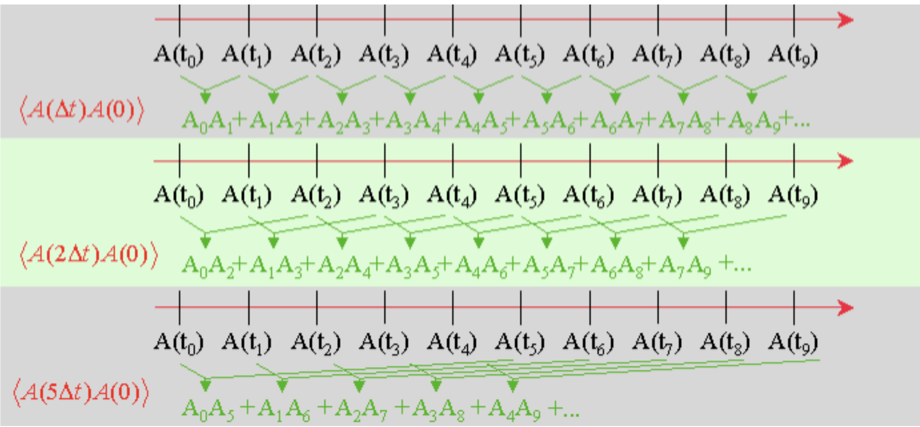

A(t0) A(t1) A(t2) A(t3) A(t4) A(t5) A(t6) A(t7) A(t8) A(t9)
⟨A(Δt)A(0)⟩
A0A1+A1A2+A2A3+A3A4+A4A5+A5A6+A6A7+A7A8+A8A9+...
⟨A(2Δt)A(0)⟩
A0A2+A1A3+A2A4+A3A5+A4A6+A5A7+A6A8+A7A9 +...
⟨A(5Δt)A(0)⟩
A0A5 +A1A6 +A2A7 +A3A8 +A4A9 +...

Figure 3.5: Schematic representation of the calculation of residence times for a sequence of time intervals. Each time in simulation serves as a new time origin for the time correlation function (*t=nΔT*). Here, we show an example of *n=1, 2,* and *5*. The maximum number of n is the total simulation time and for this case, the corresponding number of segments, N(t) = 1. The image is adapted from [184].

Our water residence time calculation was verified by reproducing a simulation of native-state α-lactalbumin (PDB ID: 1A4V) in a water box using the simulation protocol from [183]. We used the first hydration shell to calculate the lifetime of waters within 4 Å of protein. We calculated a value of 25.1ps, compared to 25.9ps from [183].

Additionally, in proteins, such as SecA, we followed the same procedure for the calculation of the residence times of waters close to the protein surface and report the water residence times per amino acid residue (Figure 3.6)[102].

Our algorithms are written in Tcl [185], that rely on the graphical interface of the VMD software [23], and support graphical visualizations of the H-bond networks. Further data processing was executed using Matlab [186].

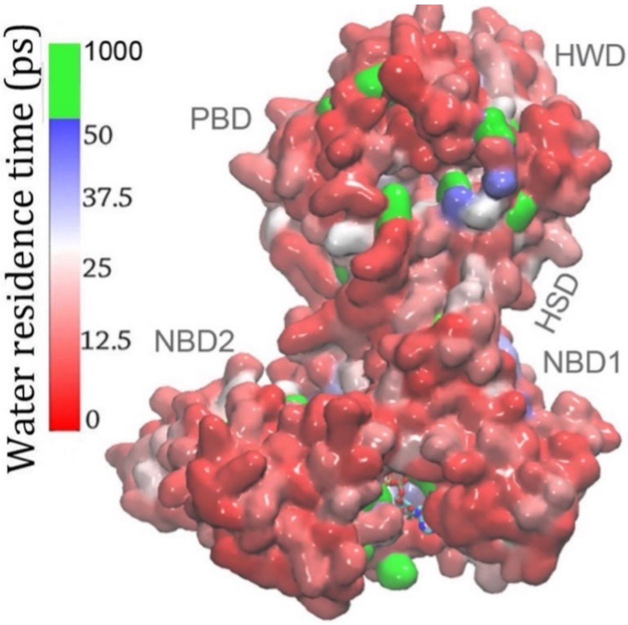


Figure 3.6: Example of our algorithm to compute water dynamics at the SecA's interface. We calculate the average residence times of water molecules per amino acid residue of SecA at the first hydration layer and color code the surface of SecA according to our results. With green we show sites of stable water molecules. Calculations are based on [102]. Image is from [102], a study presented in this thesis.

3.5 H-bond topology paths

To compute the connections of H bonds in our membrane models, we used our algorithm shown in Section 3.2 [22]. The algorithm calculates the network components and obtains the details of lipid molecules engaged in clustering for each simulation step in our analyses.

In our graphs, the nodes are the lipid molecules, and the edges may be i) direct H bonds between lipid headgroups, ii) one-water H bond bridges between lipid phosphates or iii) ions interacting closely -within 4 Å- with lipid phosphate groups. Edges have no orientation, so our graphs are undirected. A cluster is a smaller graph of the entire network, a subgraph, which consists of nodes and edges that are connected to each other. The H-bond criteria are the same used in Section 3.1.

Our lipid cluster analyses indicate membrane interfaces host interactions that create paths of connected lipid molecules [187]. The number of edges linked to a node defines the degree, D, of that node. Topology is the geometric arrangement of the components of the graphs, and it defines how the communication is set up between the nodes.

We provide a DFS implementation for lipid H-bond clusters and their topologies [187]. A scheme is illustrated below and presents graphically how the algorithm works.

The algorithm starts from a node which is the starting or source point of a current path and visits all the nodes along that route. When all nodes have been explored, it returns on the same path to trace nodes not previously seen. The algorithm then chooses the next unexplored path following the same procedure. The calculation is completed, when the entire graph has been examined [180].

As a result, in our graphs, we create pathways of linked nodes and find the number of components/clusters discovered, a list of the ids of the H-bonded lipids, and the occupancy of each cluster. The algorithm involves eight steps, which are depicted in Figure 3.7, and outlined below.

In the initial step of our computation, the search begins at source node A and explores any unexplored node nearby, identified here as node B (Figure 3.7). Node B has been marked as visited, and node C next to it is being explored in the second step. When there are many surrounding nodes, for example, node C relates to nodes G and E, the algorithm selects the next node according to the node names' alphabetical order. In the third step, node E is explored before node G, starting with node C. The procedure is iterated until all the path nodes have been visited (steps 2-4). The algorithm then proceeds backward along the same route to locate unexplored nodes (steps 5-7). The procedure is terminated until all nodes are marked as explored (step 8).

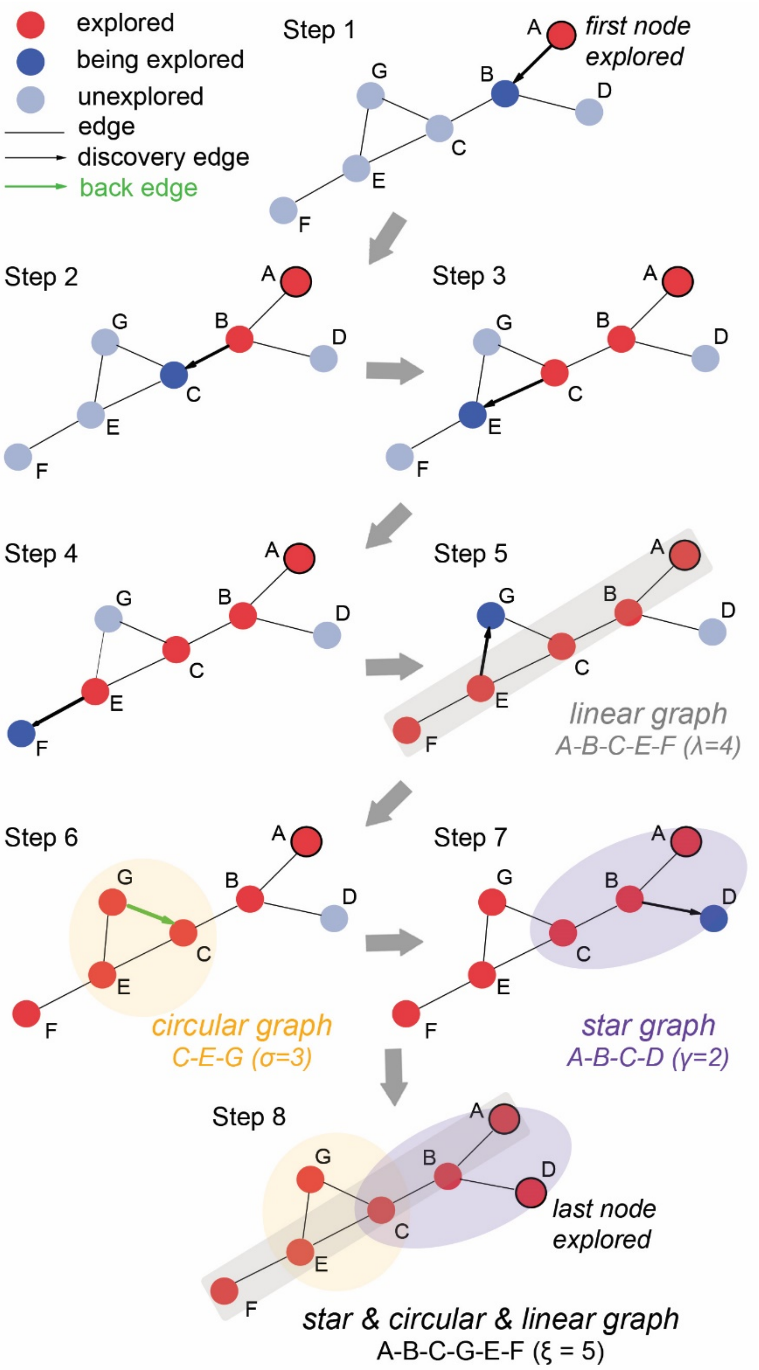


Figure 3.7: The DFS algorithm was used to locate and describe lipid clusters. Colors of cluster nodes are based on the stage of the search, with red indicating nodes that have previously been searched, blue indicating nodes that are being investigated in the current step of the algorithm, and light blue indicating nodes that have not yet been explored. The edges colored black and containing an arrow represent the current search direction; the green edge indicates a back edge in step 6. The rest of the edges are gray [187]. Image is from [187], a study presented in this thesis.

3.6 Topologies of linear, star, and circular graphs and combinations thereof

We calculated and illustrated lipid clusters from multiple biomolecular simulations. We analyzed cluster topologies, which we labeled as linear, star, and circular. Figure 3.8 depicts these paths and their combinations. These three schemes are fundamental because they describe single lines (linear), circles, and branches (star nodes), and they permit us to extract additional complicated schemes described in graph theory. We categorize all paths based on their length (Figure 3.8) [187].

When two nodes have degree (D) equal to 1 and all nodes in between have degree of value 2, the path is considered linear. In other words, a path is linear if all its nodes and edges are on a single line (Figure 3.7, step 5). When one internal node is connected to all other nodes, the path is referred to as a star path. In such a scenario, just one node (the central) has a D greater than one. A star graph has been spotted when 1 central node has D = N – 1 and the remaining N - 1 nodes have D = 1, where N stands for the total number of nodes in the graph (Figure 3.7, step 7).

When all nodes of a path are linked in a closed chain, the path is termed circular. As a result, the smallest number of nodes required for a path to be circular is three. We sought back edges to locate a circular graph. When a node links to its ancestor via an edge in the graph discovered with DFS, there are back edges. The principal distinction between edges linked to nodes in a circular and a star graph is that in the circular graph, we find edges linked to previously explored nodes, namely, everything accessible from these nodes was explored earlier (Figure 3.7, step 6).

As DFS explores each node, we keep track of the number of edges on it. We also keep a record of nodes that have three or more edges. This helps us identify complex combinations, like a star and a linear graph (Figure 3.7, step 8) [187].

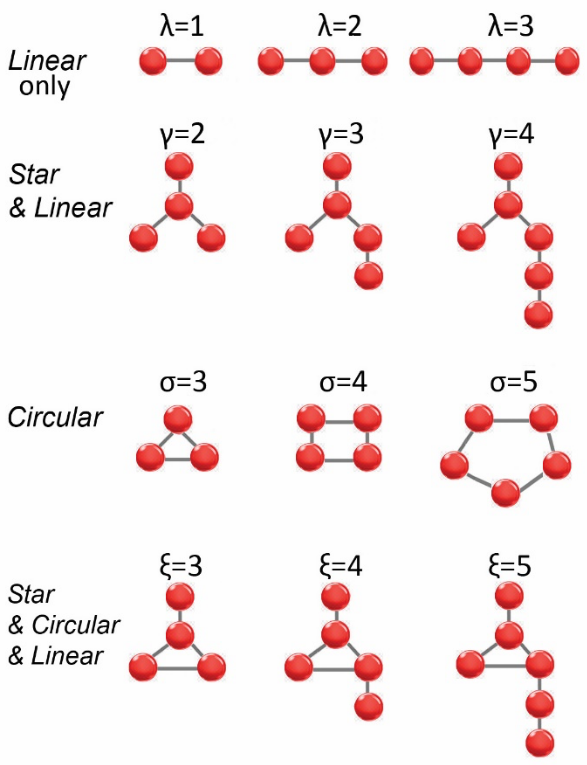


Figure 3.8: Membrane's lipids topologies illustration. The number of nodes determine the cluster size in the detected topologies [187]. Image is from [187], a study presented in this thesis.

In Figure 3.7, it is shown that the algorithm discovered the 4-length A-B-C-E-F of linearly connected nodes. Step 6 involved finding a circular path with the nodes C, E, and G connected in a closed chain. The circular path is detectable via searching for back edges, or in other words, edges that link to previously explored nodes (green arrow at step 6 in Figure 3.7). In step 7, a star graph containing nodes A-B-C-D was identified. This is the scenario when we detect edges to nodes that DFS is still exploring. Step 8 presents a complex mix of linear, star, and circular graphs, with the longest linear path being identified and reported. If complicated combinations are detected, we discard small branches from star paths and retain solely the edge that links the maximum length path to the end node, for circular paths.

The illustration in Figure 3.8 gives examples of the lengths of the topologies found in our simulations. The Greek letters λ, γ, σ, and ξ, respectively, correspond to the size of clusters with a linear lipid arrangement, a combination of star and linear, circular, and, a pattern of star, circular, and linear topology, respectively. L stands for the length of linear paths, including linear segments, in both Star & Linear and Star & Circular & Linear paths (Figure 3.8).

3.7 Algorithm implementation to detect and analyze lipid topologies

The topology analysis script uses the DFS algorithm to cluster H-bonded lipid molecules and classify types of topologies. Our algorithm reads a column-oriented delimited file with the following structure: Column 1 is the simulation time step, columns 2-4 and columns 5-7 the atom names, resnames, and resids of each H-bonded lipid pair respectively, and column 8 the H-bond distance between pairs in Å (optional). Data rows are divided based upon the simulation time so that we have groups of data, each one contains the H-bond information of each simulation time.

For each simulation time, adjacency matrices are created to represent the H bonds between different groups. These matrices are binary and have a value of 1 if the pair in the system is connected, and 0 if there is no connection. Our findings can be projected onto graphs G using adjacency matrices, with lipid molecules as nodes and H bonds as edges. We start the search of graph G at node 1 using a depth-first search (DFS) algorithm. This algorithm thoroughly searches all the nodes along the current path. Once all nodes have been visited, the algorithm backtracks to locate unvisited nodes. After visiting every node on the current path, the algorithm chooses the next unexplored path. That means that the search restarts when remained undiscovered nodes are unreachable from the discovered nodes. The new start node is the node with smallest index that is still undiscovered. Once the examination of the entire graph is complete, the calculation comes to an end. We detect clusters as groups of H-bonded lipids per simulation time and we save the lipid ids per cluster.

We compute the degree centrality (*DC*) of each node in the graph. The degree D of a node is computed by the sum of edges connected to that node. The *DC* is used as a validation for the topology type computation that follows. The geometry of nodes connected to edges in our graphs is refered as the path's topology.

When nodes are connected in a closed loop, a circular path is formed. The degree D of each node is 2. To detect circular graphs, we monitor edges with the dfsearch flag in Matlab 'Edges connected to a finished node'. Identifying a circular graph involves looking for back edges. When a node links to its ancestor through an edge in the graph discovered using DFS, back edges are found.

When each node in a path connects to a single internal node, it forms a star. In a star, only one node can have a degree greater than 1. To discover star graphs, we monitor edges with the dfsearch flag in Matlab 'Edges connected to a previously discovered node'. We keep track of the number of edges to each node being investigated by DFS. Internal nodes of star graphs have at least three edges.

When comparing edges connected to nodes in a circular graph versus a star graph, the key difference is that in the circular graph, we can identify edges that are connected to nodes that have already been visited. This means that we have already discovered everything that is accessible from these nodes.

To detect complex combinations of linear & star & circular, the above two criteria must be met at the same time. The path is categorized as linear if we do not find any internal node in a star graph or a circular path. A linear path is defined as a path where every node has a degree of either 1 or 2, and all edges and nodes lie on a straight line.

For each simulation time step, we calculate the length of each H-bonded lipid cluster defined as the longest number of edges between a starting and ending node. The cluster size is the number of nodes that are linked by H bonds.

We write to .dat and .mat files results of the lipid clusters (cluster size and total number of lipids in clusters per simulation time) and cluster topologies (lipid nodes that belong to clusters, degree centrality for each lipid node, size, and topology type for each cluster per simulation time).

The topology visualization script uses the results from the above analysis to visualize in VMD H-bond network of lipids colored according to graph topology type (circular, star & linear, linear & star & circular, linear). The user gives as input the .pdb file of the system; it can be a coordinate snapshot or an average structure from the simulation. The user also selects the simulation time frame to be used for the visualization of lipid clusters. The script reads the H bonds to create a network in VMD with graphic lines connecting lipid ids based on their coordinates (Figure 3.10). Phosphorus atoms are used as the representative node for each lipid. H bonds are colored as grey graphic lines.

From resulting files, we extract the lipid ids, the cluster they belong (Figure 3.11, Figure 3.12), and the topology type and we create graphic spheres for lipid nodes color-coded based on the following flags: 0 is for circular (yellow), 1 for star & linear (red), 10 for combined linear & star & circular (blue), and 2 for linear (green) (Figure 3.10).

Examples of output vmd images rendered after we run the script are in the topology_analysis folder (.jpgs).

Sample_folder contains output results after running the scripts to detect graphs of H bonds formed between lipids in a pure POPE and 3:1 POPE: POPG lipid bilayer. Results are in the output folder. Images from the visualization of topologies from different simulation times are in .tga and .jpg format. The output folder contains a Readme file with details about input/output and script runs.

Workflow is generated and tested in MATLAB R2017b [186] and VMD 1.9.3[23] and released as GitLab and Mendeley repository. The steps of the algorithm are illustrated in Figure 3.9 below.

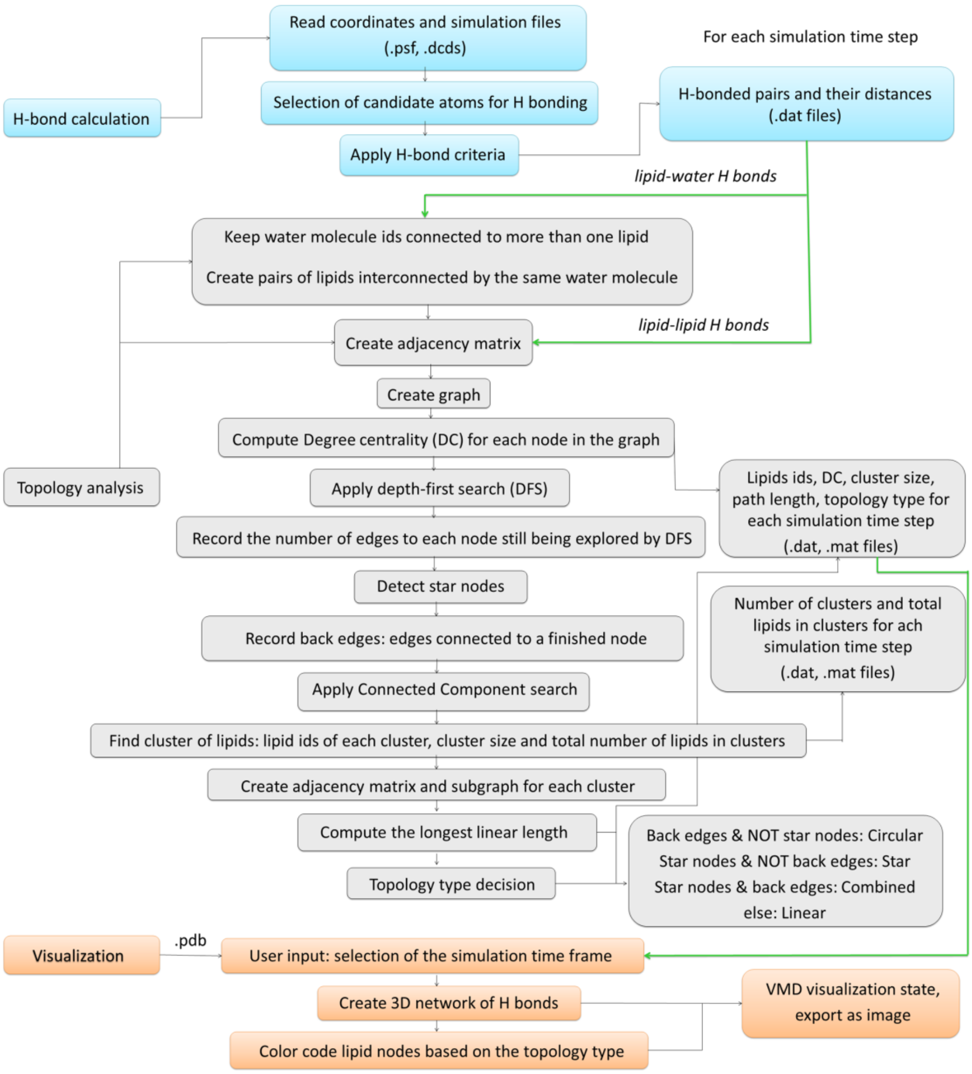

H-bond calculation
Read coordinates and simulation files (.psf, .dcds)
Selection of candidate atoms for H bonding
Apply H-bond criteria
For each simulation time step
H-bonded pairs and their distances (.dat files)
lipid-water H bonds
Keep water molecule ids connected to more than one lipid
Create pairs of lipids interconnected by the same water molecule
lipid-lipid H bonds
Create adjacency matrix
Create graph
Compute Degree centrality (DC) for each node in the graph
Topology analysis
Apply depth-first search (DFS)
Lipids ids, DC, cluster size, path length, topology type for each simulation time step (.dat, .mat files)
Record the number of edges to each node still being explored by DFS
Detect star nodes
Number of clusters and total lipids in clusters for ach simulation time step (.dat, .mat files)
Record back edges: edges connected to a finished node
Apply Connected Component search
Find cluster of lipids: lipid ids of each cluster, cluster size and total number of lipids in clusters
Create adjacency matrix and subgraph for each cluster
Compute the longest linear length
Topology type decision
Back edges & NOT star nodes: Circular
Star nodes & NOT back edges: Star
Star nodes & back edges: Combined
else: Linear
Visualization
.pdb
User input: selection of the simulation time frame
Create 3D network of H bonds
Color code lipid nodes based on the topology type
VMD visualization state, export as image

Figure 3.9: Schematic depiction of the basic steps of the algorithm used to classify lipid H-bond cluster topologies [187]. Image is from [187], a study presented in this thesis.

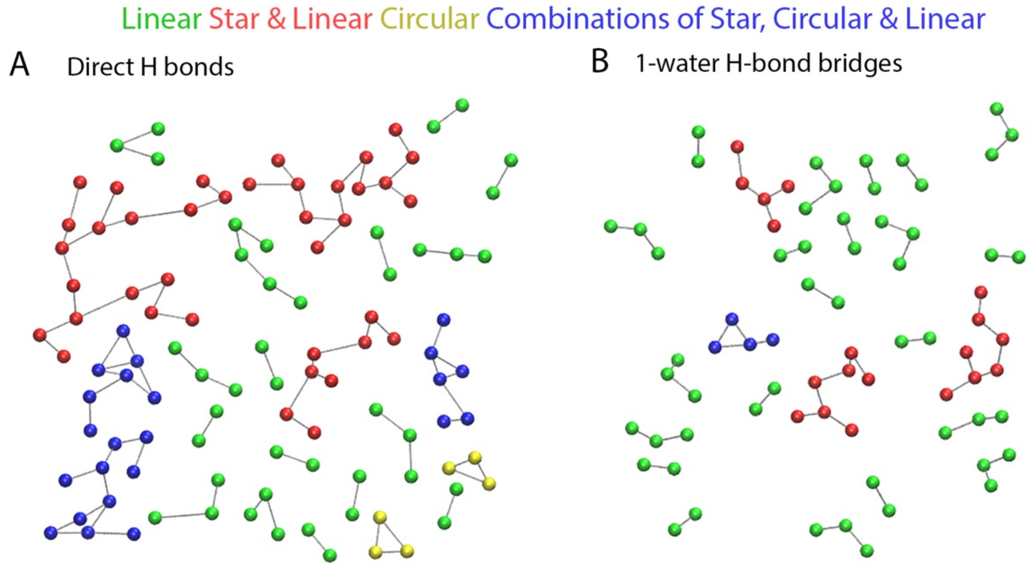

Linear Star & Linear Circular Combinations of Star, Circular & Linear
A
Direct H bonds
B
1-water H-bond bridges

Figure 3.10: Examples of graphical outputs in VMD [23] from our code implementation in TCL [185]. Images show topologies of (A) direct H bonds, and (B) 1-water H-bond bridges between lipids using coordinate snapshots from our MD simulations. Topologies are color-coded based on categories of linear, star & linear, circular, and combinations thereof. Spheres represent lipid headgroups. The image was created for the scope of this thesis.

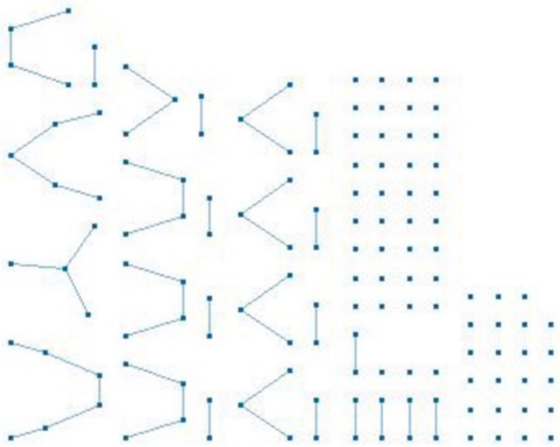


Figure 3.11: Example of the lipid clustering as exported from our codes in Matlab [186]. Dots represent the nodes (lipid molecules) and the lines represent the edges (H bonding) between the nodes. The image was created for the scope of this thesis.

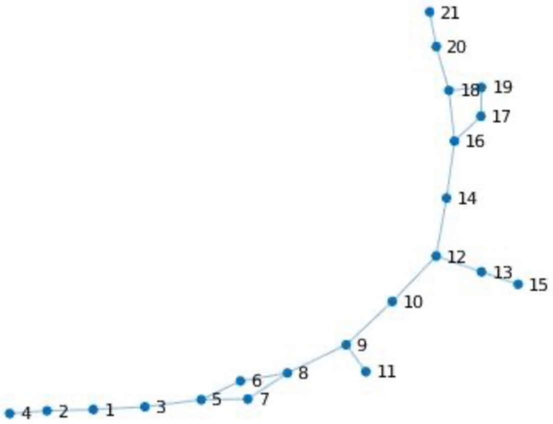


Figure 3.12: Example of the lipid clustering between POPE lipids as exported from our codes in Matlab [186]. Dots represent the nodes (lipid molecules) and the lines represent the edges (H bonding) between the nodes. Numbers represent the lipid ids. The image was created for the scope of this thesis.

3.8 Analysis of H-bond networks in proteins

We developed an algorithm to derive graphical network representations of H bonds between amino acids and characterize the strength of the connections between them and the centrality of each node (Figure 3.13).

In the first step of the algorithm, we create arrays of H-bonded atom pairs and the distances of their interactions. The data obtained in from step 1 is utilized to construct the H-bond network in step 2. For simplicity, each H-bonding group is joined by a single edge linking Cα atoms. Therefore, Cα atoms are employed as the typical node of every protein group. The H bonds are then classified based on their average length, as determined by the simulation time. When the H-bond distance is $d_{\mathrm{HA}} \leq 1.7$ Å, 1.7 Å < $d_{\mathrm{HA}} \leq 1.9$ Å, and 1.9 Å < $d_{\mathrm{HA}} \leq 2.5$ Å, the H bonds are classified as strong, medium, and weak respectively.

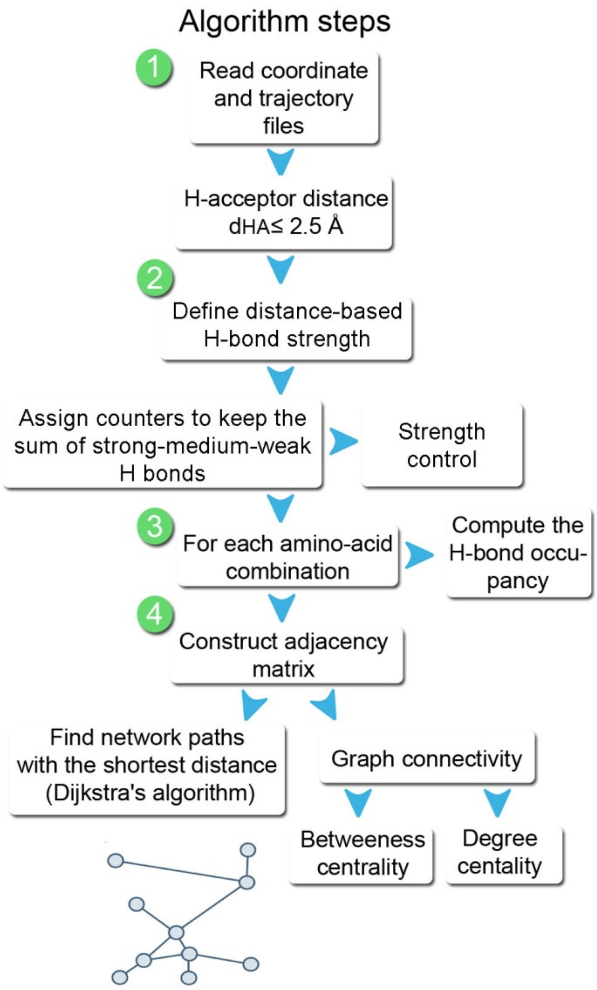

Algorithm steps
1
Read coordinate and trajectory files
H-acceptor distance dHA≤ 2.5 Å
2
Define distance-based H-bond strength
Assign counters to keep the sum of strong-medium-weak H bonds
Strength control
3
For each amino-acid combination
Compute the H-bond occu-pancy
4
Construct adjacency matrix
Find network paths with the shortest distance (Dijkstra's algorithm)
Graph connectivity
Betweeness centrality
Degree centality

Figure 3.13: Algorithm utilized to generate protein H-bond network visualizations based on occupancies, H-bond strength, shortest distance pathways and centrality measurements [188]. Image is from [102], a study presented in this thesis.

To simplify the illustration of H bonds, we first calculate the total number of strong, medium, and weak H bonds sampled durong the simulation. The largest sum of them is then employed to classify that specific H bond as strong, medium, or weak. When two or all three sums previously mentioned resulted in similar values, we continued in the following way. Firstly, we compare the total number of strong H bonds with the 25 % of whichever of the remaining two total numbers of weak and medium H bonds. The H bond is deemed as strong if the sum of the strong H bonds is more than 25% of the sum of the medium and weak H bonds. In case the previous criterion is not satisfied, and the sum of medium H bonds exceeds 25% of that estimated for strong and weak H bonds, the H bond is considered as medium. The H bond is, otherwise, regarded as weak. This analysis, which is included in the second step of the algorithm, is designated as 'Strength control', in Figure 3.13.

The third step of the algorithm calculates the occupancies for each H-bond pair. Occupancy is defined as the proportion of time that the H-bond requirement is met during the analyzed trajectory segment. The strength of H bonding between two amino acid residues is determined by the H bond with the highest occupancy if a pair of amino acids form multiple H bonds through different atoms during a simulation.Data regarding the occupancy and the total strength of the H bonds are then projected onto the protein H-bond network (Figure 3.14).

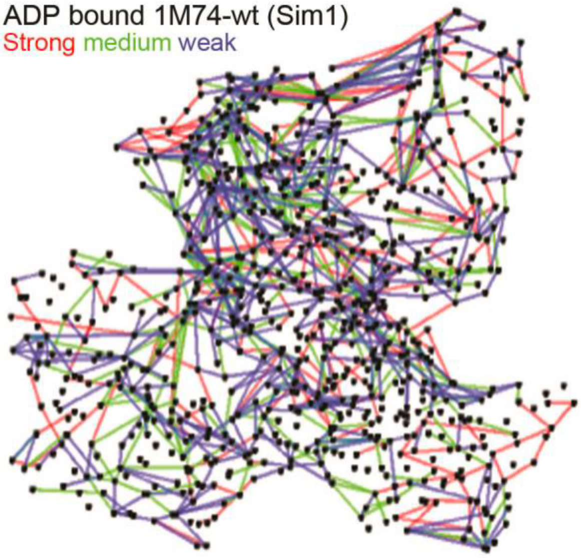


Figure 3.14: Network representation of the SecA protein based on the H-bond strength criterion during the MD simulation. Black dots represent protein amino acids and lines represent H-bond interactions between the nodes colored according to their strength (strong, medium, or weak). Analysis is presented on [102]. The image is from [102], a study presented in this thesis.

3.9 H-bond pathways with the shortest distance

Dijkstra's algorithm (Section 2.2.2, Figure 3.15) was employed to extract shortest-distance H-bond pathways on a graph *G = (U (nodes), R (edges))*. Weights between edges must be positive. For protein analyses, weights are the distances between amino acids, as computed using the strength definition from the H-bond analysis (Section 3.8). The algorithm begins by initializing a source and an end node, and then searches for the shortest route between them.

The following is an example of finding the shortest path from vertex A to every other vertex.

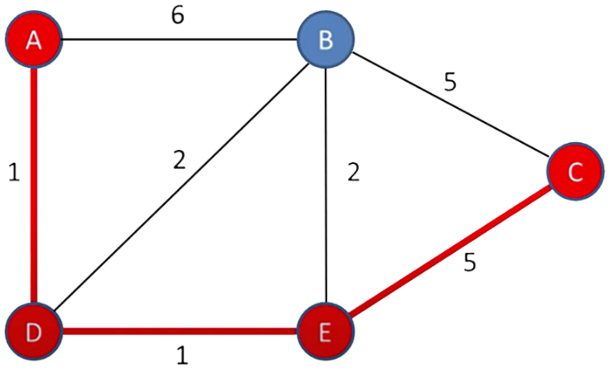


Figure 3.15: The shortest path between any two vertices in a graph can be found using Dijkstra's algorithm. A collection of data is produced by Dijkstra's shortest path algorithm, which contains the shortest paths between each vertex in the graph and the starting/source vertex. Here, the shortest path is A-D-E-C. The example is adapted from [189].

We consider A as the start vertex. The distance from A to A is 0, and the distances to all other vertices from A are unknown, therefore we set a value of ∞ (infinity). The algorithm visits the unvisited vertex with the smallest known distance from the start vertex. First time round, this is the start vertex itself, A. For the current vertex, the algorithm examines its unvisited neighbors, which are B and D, and calculates the distance of each neighbor from A. If the calculated distance is less than the known distance, the shortest distance is updated. The previous vertex for each of the updated distances is updated. The current vertex, A, is added to the list of visited vertices. Next, the vertex with the smallest known distance from the start vertex is visited. In our case, this is vertex D. For D, we examine its unvisited neighbors, B and E. For the current vertex, the distance of B and E from the start vertex is calculated. It is distance 3 for vertex B and 2 for E. Again, if the calculated distance is less than the known distance, the shortest distance is updated. The previous vertex for each of the updated distances is updated. The current vertex, D, is added to the list of visited vertices. Next, the vertex with the smallest known distance from the start vertex is visited. In our case, this is vertex E. The procedure is repeated until all vertices are visited. In this example, the shortest path is A-D-E-C.

We have applied the shortest distance path algorithm to simulations of lipid bilayer systems and proteins. We can identify paths of the shortest distance between two nodes (source and end node) (Figure 4.30), between one node and a set of nodes (Figure 3.16), and between nodes setting an intermediate node (or set of nodes) (Figure 3.17).

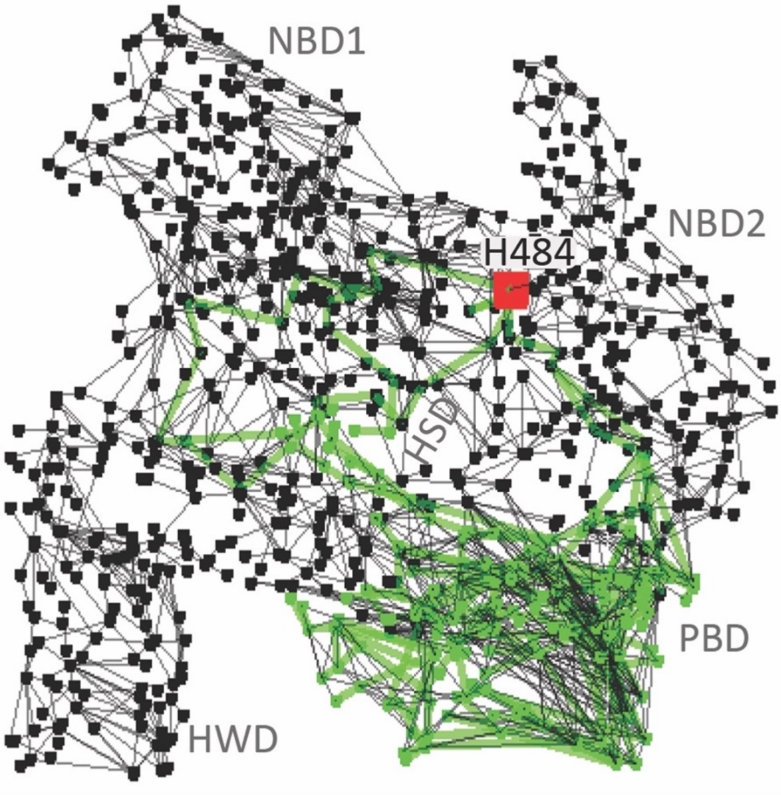


Figure 3.16: Identifying H-bond paths between the H484 (NBD2) and the PBD interface. (A) Graph of all H bonds in ecSecA2VDA. Thin black lines indicate H bonds sampled at least once during the last 200 ns of the simulation. The thick color lines are the shortest paths that connect NBD2-H484 to PBD. The shortest distance paths obtained using Dijkstra's algorithm are shown in green. There are 149 pathways connecting H484 to PBD. Each path consists of an average of 9 H-bonded amino acids. The image was prepared with MATLAB R2017b using the last coordinate snapshot of the simulation. Similarly, from E487 to PBD, there are 149 pathways, each one consisting of an average of 10 H-bonded amino acids. Data are from the last 200ns of the ecSecA2VDA simulation. Data are from [99]. Image is generated for the scope of this thesis.

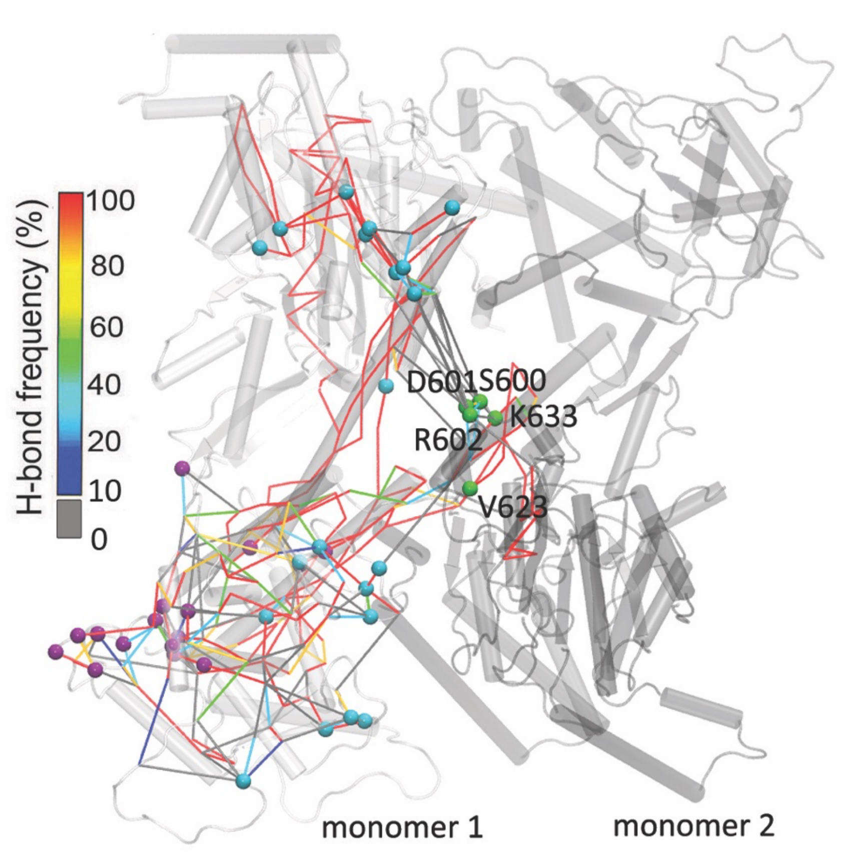


Figure 3.17: Shortest-distance analysis of the two monomers of dimer 1NL3 SecA. We select a specific set of nodes (source-intermediate-end) and we calculate all the paths connecting those nodes. Here, we detect water-mediated H-bond communication between the two monomers. Lines show the H-bond frequency of each sampled H bond. Data are from [100]. Image is generated for the scope of this thesis.

3.10 Graph connectivity: betweenness and degree centrality

The theory and equations of the Betweenness (*BC*) and Degree centrality (*DC*) in graphs [190, 191] [179] [192] is explained in Section 2.2.3.

In a graph, a node is described by *BC* values when it mediates among two other nodes linked by shortest-distance paths. *BC* affects the graph's total connectivity, since discarding nodes with high *BC* values may trigger the network to disconnect.

The number of edges that are connected to a node gives the *DC* of that node. *DC* gives a measure of local connectivities in the graph.

We present a schematic example (Figure 3.18) of the algorithms used in our H-bond graphs, featuring a small selection of amino acids from a protein network (Figure

3.18A). Using as source node the amino acid residue E208 and an end/destination node the R517, we compute the shortest-distance H-bond path using Dijkstra's algorithm. We find an optimal path with a total distance of 5.5 Å and two intermediate nodes, the D215 and S211 (Figure 3.18B).

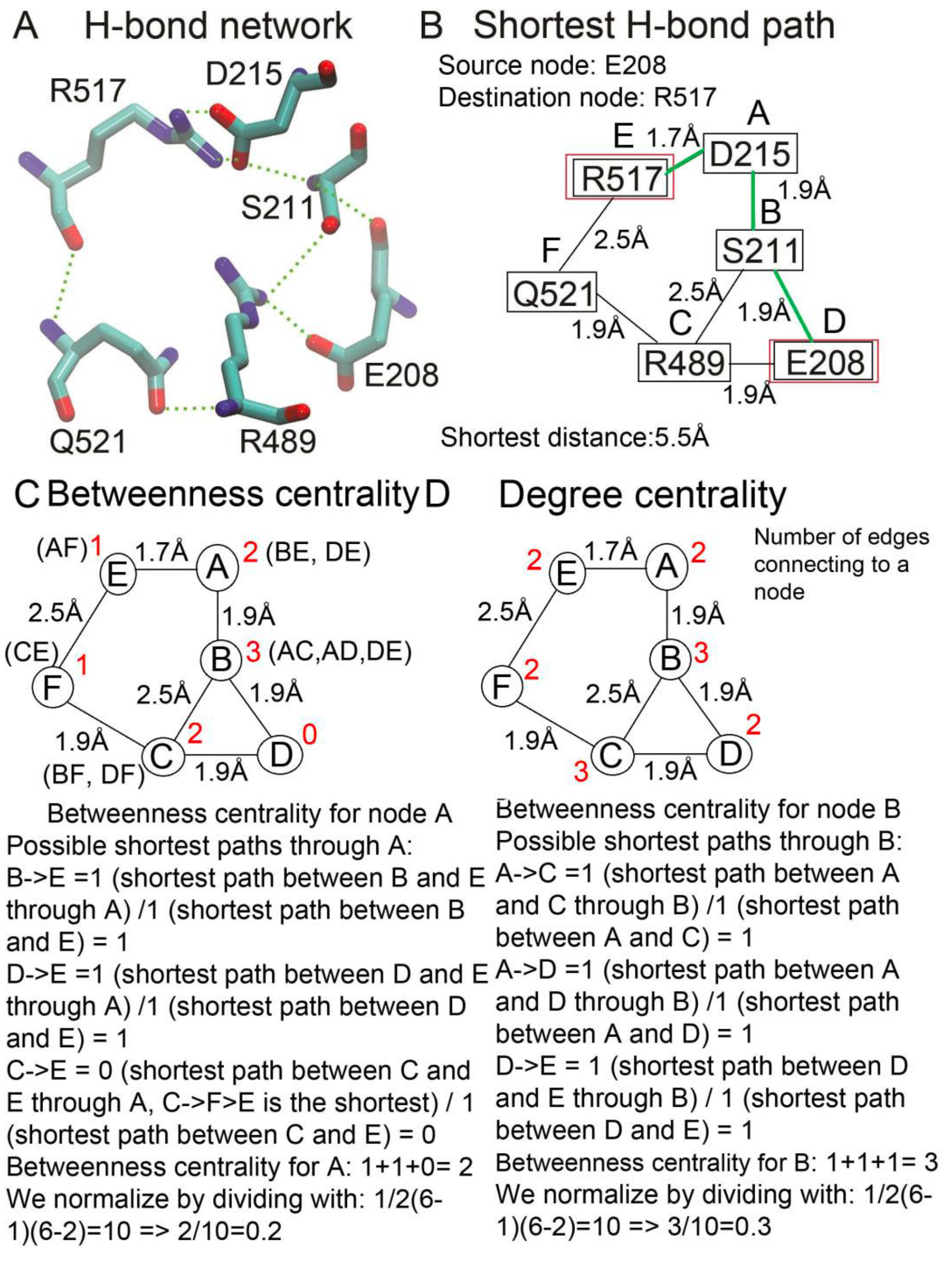
To compute the *BC* centrality for node A, we check the possible shortest paths through node A, and we apply the equation 2.25. The first path is between B and E through A. The *BC* is computed as the number of shortest paths between B and E through A (it is equal to 1) divided by the total shortest paths between nodes B and E (it is equal to 1). The result is 1.

The next nodes to check are D and E through node A. The *BC* is computed as the number of shortest paths between D and E through A (it is equal to 1) divided by the total shortest paths between nodes D and E (it is equal to 1). The result is 1.

We then check the nodes C and E through node A. The *BC* is computed as the number of shortest paths between C and E through A (it is equal to 0, optimal path is the C-F-E) divided by the total shortest paths between nodes C and E (it is equal to 1). The result is 0.

Conclusively, the *BC* of node A is the sum of the previous results: 1+1+0 =2. According to Section 2.2.3, the normalized *BC* in our undirected graph is the *BC* divided by $\frac{1}{2}(|N|-1)(|N|-2)$. Here, N=6. The result is : 2/ (1/2*(6-1)(6-2)) = 2/10 = 0.2.

The *DC* values are shown in Figure 3.18D. An analysis of the *BC* of node B is depicted in the illustration. The procedure is as the computing of the *BC* of node A and it is presented above.



Betweenness centrality for node A
Possible shortest paths through A:
B->E =1 (shortest path between B and E through A) /1 (shortest path between B and E) = 1
D->E =1 (shortest path between D and E through A) /1 (shortest path between D and E) = 1
C->E = 0 (shortest path between C and E through A, C->F>E is the shortest) / 1 (shortest path between C and E) = 0
Betweenness centrality for A: 1+1+0= 2
We normalize by dividing with: 1/2(6-1)(6-2)=10 => 2/10=0.2

Betweenness centrality for node B
Possible shortest paths through B:
A->C =1 (shortest path between A and C through B) /1 (shortest path between A and C) = 1
A->D =1 (shortest path between A and D through B) /1 (shortest path between A and D) = 1
D->E = 1 (shortest path between D and E through B) / 1 (shortest path between D and E) = 1
Betweenness centrality for B: 1+1+1= 3
We normalize by dividing with: 1/2(6-1)(6-2)=10 => 3/10=0.3

Figure 3.18: Schematic representation of the centrality algorithms in an H-bond subgraph. (A) H bonds between E208, S211, D215, R489, R517 and Q521 are shown as green dashed lines and protein groups as bonds. (B) Calculation of the shortest distance from E208 to R517. Distances are based on our strength control (see Section 3.8). (C) *BC* calculation from the graph and explanation of the results for selected nodes. (D) *DC* calculation from the graph of H-bond interactions [188]. Image is from [102], a study presented in this thesis.

3.11 Water-mediated H bonds between protein functional domains

Using our residence times code, we computed the water lifetimes for each protein amino acid residue. We searched for H-bond bridges linking amino acids from different protein regions, identifying them based on the proximity of water molecules to specific residues for a long simulation time. More specifically, our procedure is applied to SecA protein to explore H-bond pathways and characterize its long-distance conformational coupling.

We developed an algorithm that uses H-bonded water chains with a maximum length of L=5 water molecules to find the shortest pathways between two protein domains. When length $L = 1$ is detectable, two protein groups form H bonds with the same water molecule. The method proceeds by considering the waters identified in the first step as nodes for the new search for H-bonded waters, looking for water bridges with L greater than one. For each pair of protein domains interconnected by H-bonded water bridges, only that with the shortest distance is kept.

To examine how dynamic are the waters that belong to H-bonded chains linking protein groups, we recorded their occupancy computed as the percentage of the simulation time a water bridge of a specified length is detected (Figure 3.19).

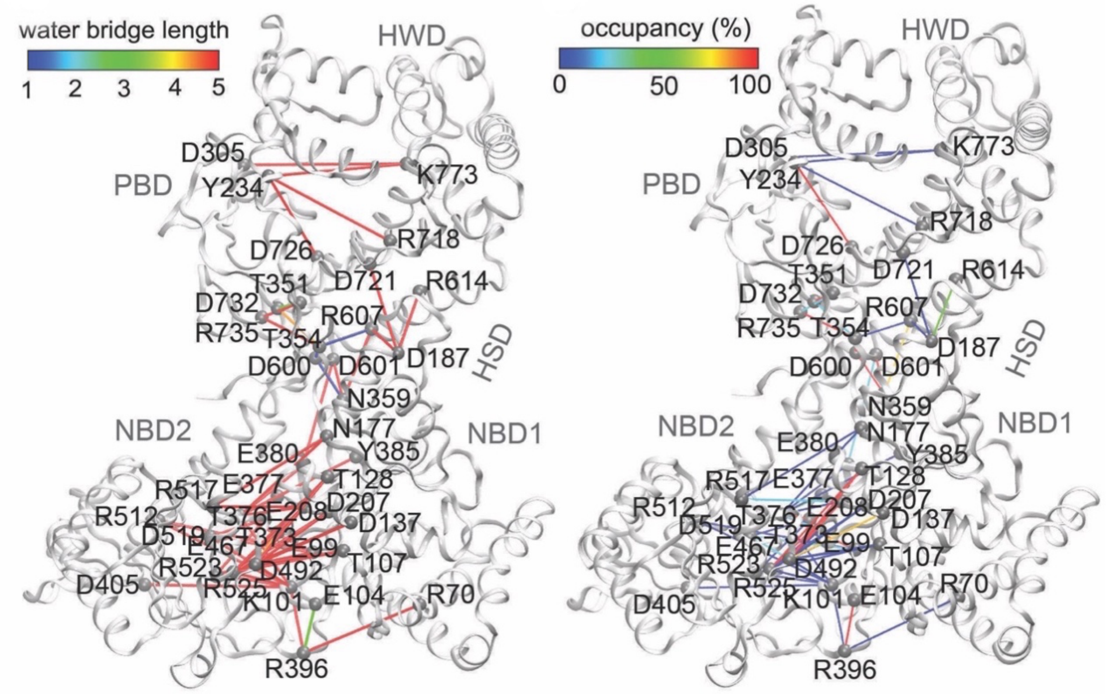


Figure 3.19: Example of algorithm calculation for networks of H-bonded water bridges linking amino-acid residues from distinct SecA protein domains. (Left) Inter-domain water-mediated bridges based on bridge length L. (Right) Inter-domain bridges based on water bridge occupancy. The calculations are based on [102]. The image is adapted from [102], a study presented in this thesis.

3.12 Comparison of H-bond networks

Graph representations of H-bonding interactions can give essential insights into a molecular system. The comparison of graphs of proteins of different organisms or between the wild-type and mutations or between mutations is not always easy by examining separate graphs and interactions. For that reason, we implemented an algorithm to illustrate the comparison of H bonds of different systems (≥2) in one network.

After the calculation of H bonds separately for each simulation, we create a joined table of all interactions (rows) of all systems (columns). We mark it as 1 (or other weight based on the visualization) if each H bond is present in one column and 0 if it is absent. The same procedure is followed for all other columns. In that way, we create binary tables showing the H-bonding sampling or not in all systems.

For the visualization, we choose a coordinate snapshot of one molecular system, and we map all the interactions based on what question we want to answer (Figure 3.20). There are many modifications to the code. For example, we can map amino acid residues as nodes with red edges all H bonds present in wild type but not in mutations and with gray H bonds present only in mutations. Additionally, we can show edges color-coded based on the number of mutations that are present or differences in H-bond occupancy or distances in different simulation systems. We can also color-code specific nodes to show the sampling of H bonds in different simulations.

Code was written in Matlab [186] and Tcl [185]. Visualizations were performed in VMD [23].

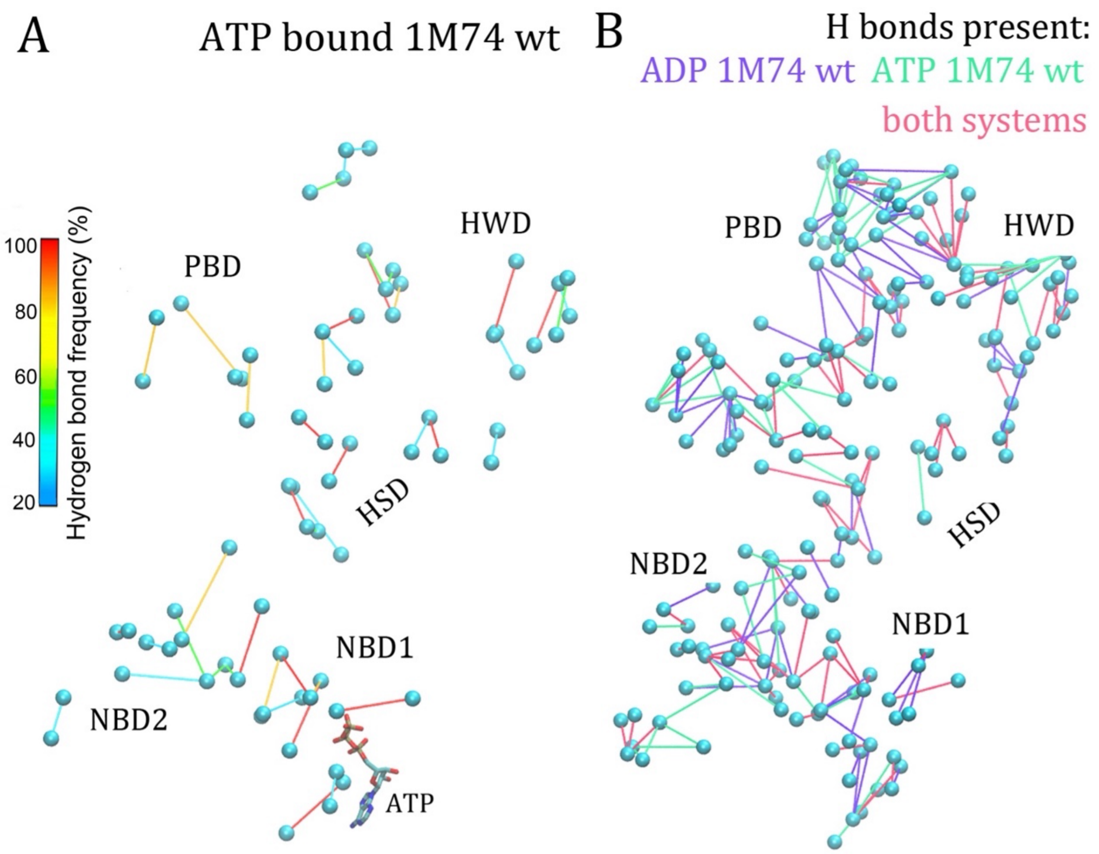


Figure 3.20: ATP binding alters inter-domain H bonding in SecA. (A) Inter-domain H bonding computed for ATP-bound SecA from Sim7. Each line represents H bonds colored according to the occupancy levels, and interactions are shown between the Cα protein atoms represented as small spheres. (B) Inter-domain H bonds in ADP-bound vs. ATP-bound SecA. H bonds present only in ADP-bound SecA, only in ATP- bound SecA, or in both simulations are colored blue, green and red, respectively. For the locations of the Cα we used the last coordinate snapshot of Sim7 of [102]. Image is generated for the scope of this thesis.

3.13 H bonding in crystal structures vs. MD simulations

We wrote a script to calculate and visualize two-dimensional H-bond maps color-coded according to the occupancy of each H bond. We performed two separate analyses, one for the SecA crystal structure (PDB ID: 1M74) only and the other for an MD trajectory of the protein (Figure 3.21). Our analysis is important as it shows that MD simulations can give a more precise picture of the interactions and their dynamics compared to a crystal structure. However, the availability of more crystal structures that were solved under various circumstances could offer a more thorough understanding of the H bonds in complicated H-bond clusters.

Our implementation allows the graphical comparison between interactions in crystal structures and MD simulations. Figure 3.22 depicts a shortest-distance pathway in crystal structure 1M74 which is continuous only if we allow water dynamics in our calculations. On the other hand, the same calculation from an MD simulation of the same protein system reveals a complex and dynamic network of many interactions connecting with many possible ways the same nodes (amino acids). This comparison is essential and combined with centrality measurements presents the importance of MD simulations to give us more detailed information about the interactions and conformational dynamics of our systems (Figure 3.23).

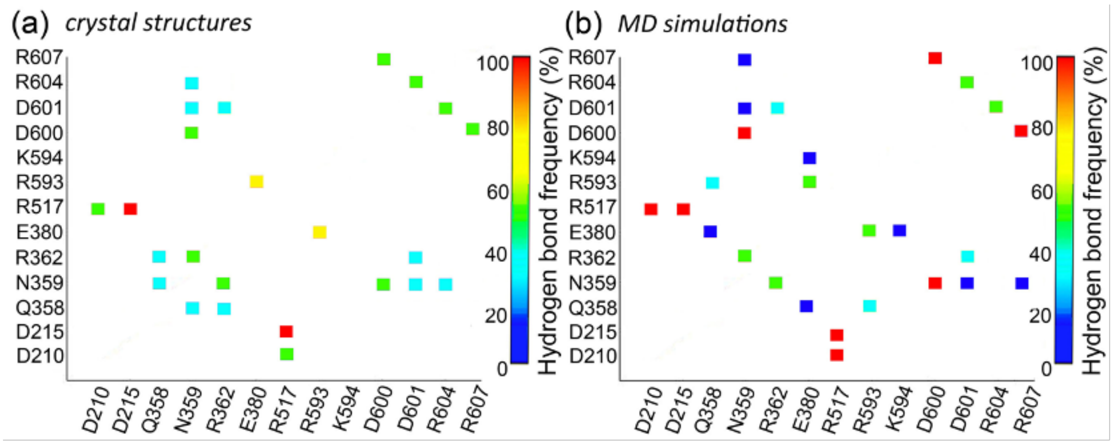


Figure 3.21: H-bond maps of *B. subtilis* ADP-bound SecA computed from (a) crystal structure PDB ID:1M74, and (b) MD simulations. Image is from [103], a publication arising from this thesis.

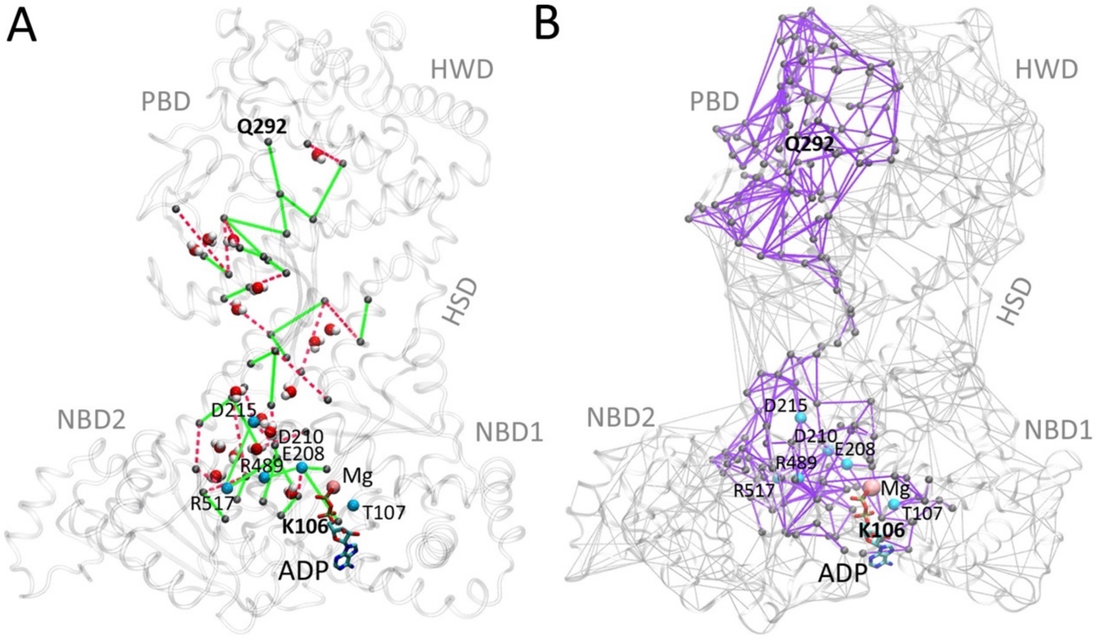


Figure 3.22: H-bond networks computed from crystal structure compared to MD simulations of SecA protein with PDB ID:1M74. (A) H-bond path from K106 of NBD1 to Q292 of the PBD, is discontinued between HSD and PBD/HWD. We included water dynamics in our graph analysis allowing for a continuous path between the two starting nodes. (B) A dense network of shortest-distance H bonds between the same nodes K106 and Q292 but using 100ns MD trajectory. Image is adapted from [193], a publication arising from this thesis.

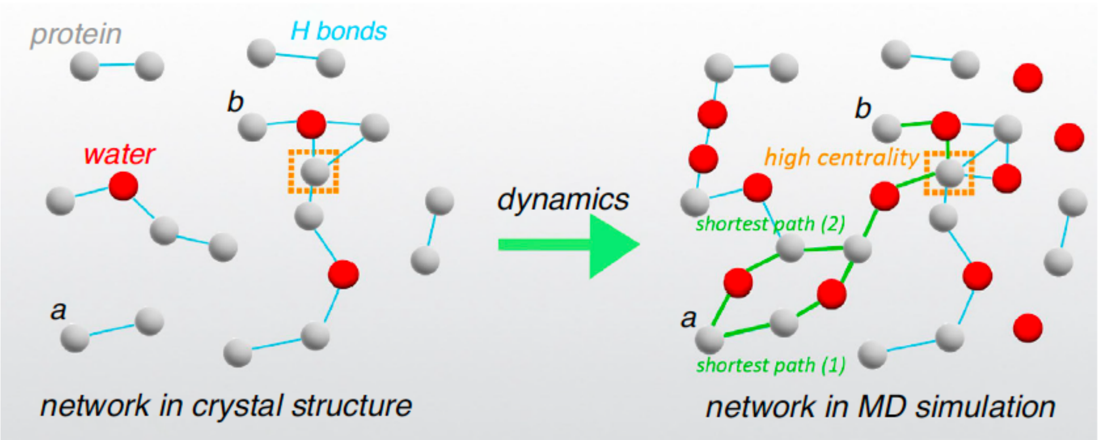
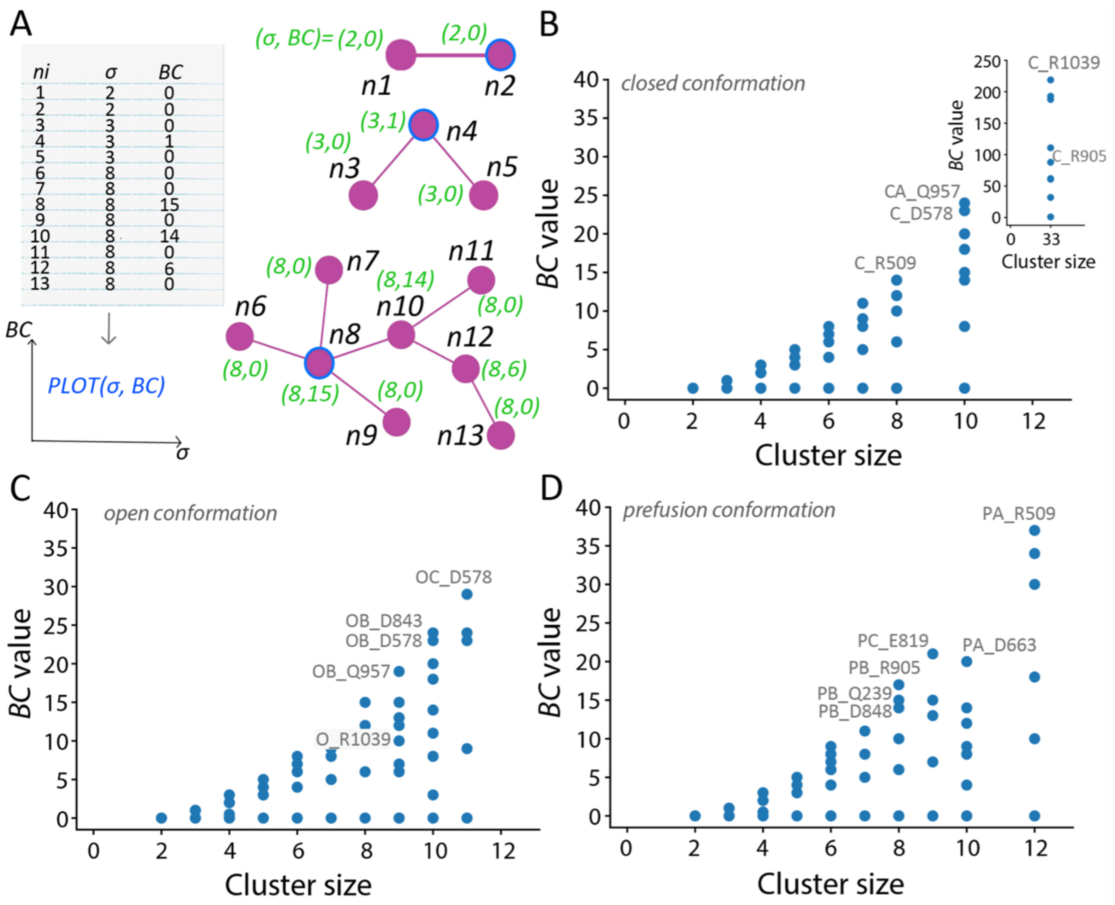


Figure 3.23: Algorithms inspired by graph theory. The H-bond network may be seen statically in crystal structures but fluctuations in atomic coordinated from MD simulations allow for detecting dynamic and transient H-bond complex networks. H-bond graphs enable accurate calculation of the whole network of H bonds and identification of groups that are vital for the network when combined with centrality measurements. Image is from [193], a publication arising from this thesis.

3.14 Correlation between H-bond cluster size and centrality

Centrality values within a cluster can be influenced by the size and shape of the cluster. H-bonding groups that are components of large clusters often have high centrality values. Nodes located centrally in a H-bond cluster can have higher centrality scores than nodes at the periphery. This is explained as follows: a central node can be a component of several shortest-distance H-bond pathways, and therefore have high *BC*. Clusters with somewhat large cluster size can also show low *BC* values, including *BC* = 0. These H-bonding groups are located at the periphery of the graph (Figure 3.24). When a specific H-bonding group has significant difference between two systems, it is possible that it is located differently in the cluster, or the size of the cluster varies.



ni	σ	BC
1	2	0
2	2	0
3	3	0
4	3	1
5	3	0
6	8	0
7	8	0
8	8	15
9	8	0
10	8	14
11	8	0
12	8	6
13	8	0

Figure 3.24: H-bond cluster size and centrality in protein S. (A) Protein clusters with different cluster sizes, σ and *BC* values. We present the full list of (σ, *BC*) pairs and we illustrate each cluster. (B-D) *BC* and σ values for all H-bond clusters calculated for the closed (panel B), open (panel C), and pre-fusion conformations (panel D) of the protein S [194]. We note that in most

cases when σ increases, then *BC* is also elevated depending on each cluster shape. Image is from [194], a study presented in this thesis.

3.15 Computational efficiency of our algorithm for lipid H-bond clusters & topologies

To test the computational efficiency of our algorithm, we calculated the direct lipid H bonds (tcl script) and the topology analysis code (Matlab script) in a segment of 10 coordinate snapshots from a simulation of a hydrated POPE membrane environment (65,922 atoms) and of a 3:1 POPE: POPG mixed lipid bilayer (65,545 atoms). From the 10 coordinate snapshots, we calculated the average computing time for 1 coordinate set. Tests were performed on a single core in the following systems, Intel(R) Xeon(R) CPU E5-2660, Intel(R) Xeon(R) CPU E3-1240 V2, Intel(R) Core(TM) i5-2400 CPU, Intel(R) Xeon(R) CPU W3550, Intel(R) Core(TM) i7-7700 CPU (Table 1). Our analyses indicate that our algorithm is efficient and suitable for longer simulations and wider membrane patches.

Table 1: Computational efficiency of the DFS algorithm for analyses of H-bond topologies. Benchmarks are given for pure POPE and 3:1 POPE:POPG simulations from [187], a study presented in this thesis.

	Pure POPE (65,922 atoms)		3:1 POPE: POPG (65,545 atoms)	
	Lipid H bonds	Topology analysis	Lipid H bonds	Topology analysis
Intel(R) Xeon(R) CPU E5-2660	3.1 sec	0.06 sec	3.1 sec	0.06 sec
Intel(R) Xeon(R) CPU E3-1240 V2	1.8 sec	0.03 sec	1.7 sec	0.03 sec
Intel(R) Core(TM) i5-2400 CPU	2.0 sec	0.03 sec	2.0 sec	0.03 sec
Intel(R) Xeon(R) CPU W3550	2.5 sec	0.04 sec	2.5 sec	0.04 sec
Intel(R) Core(TM) i7-7700 CPU	1.3 sec	0.02 sec	1.3 sec	0.02 sec

Chapter 4

"The most beautiful experience we can have is the mysterious. It is the fundamental emotion that stands at the cradle of true art and true science."

— Albert Einstein, The World As I See It

4 Results & Discussion

4.1 Dynamic H-Bond Networks at Negatively Charged Lipid Membrane Interfaces

This chapter is based on the following publication:

Karathanou, K. and Bondar, A.N., 2018. Dynamic water hydrogen-bond networks at the interface of a lipid membrane containing palmitoyl-oleoyl phosphatidylglycerol. The Journal of Membrane Biology, 251(3), pp.461-473

Figures and tables originally published in the Journal of Membrane Biology have been reproduced/reprinted with permission from J. Membr. Biol. 251 (2018): 461-473. Copyright © 2018 Springer Nature.

4.1.1 Introduction

To shed light on the complex procedures conducted in our cells [1, 3-5, 185], lipid bilayers of various lipid ratios are used for computations in theory and experiments. The lipid membrane's composition influences the surrounding proteins' activity as well as the binding events that occur at the membrane surface [14, 195]. There are diseases where the lipid concentration is found to be altered [16, 196-198] thus opening avenues to the binding of target-specific proteins or drugs to membranes [25, 199, 200]. A key point here is to understand the way water molecules interact with lipids and with the protein.

Experiments [26] and computations [27, 28] suggest that anionic lipids form clusters with a high propensity. Computer simulations of a POPG bilayer indicate strong lipid-lipid and water-lipid H bonds and the formation of lipid-sodium bridges. The stability of those bridges leads to the formation of coordinated sodium-mediated clusters that force lipids to move together influencing the dynamics of the whole bilayer [27].

A simulation study of a binary mixture of POPC lipids containing 23% anionic POPG lipids revealed nearly the same average structural properties as for the pure POPC. Differences were observed including strong H-bonding potential of the anionic POPG lipids and an increased ordering of waters at the hydrophobic-hydrophilic and headgroup-water interfaces of the mixed bilayer [201].

To investigate how soluble SecA protein binds to SecY protein translocons integrated in nanodiscs, experiments were conducted using a combination of PC and PG lipids. It is shown that the the binding of SecA to the translocon is enhanced through the combination of PG lipids with PC lipids or the presence of *E. coli* lipids in nanodiscs [5]. SecA is a protein with multiple groups that can form H bonds with anionic lipids. This observation raises the question of whether dynamic lipid/water H bonds could contribute in the binding of SecA or proteins involved in the bacterial protein transfer mechanism.

Water molecules insert deeply into the lipid bilayer forming interactions with deeper buried lipid oxygens [29-35]. The structure and dynamics of water are affected by its interactions with the lipid headgroups [36, 37] within ~ 10 Å of the membrane surface plane [38]. Water molecules can H bond with the lipid headgroups and form chains of water-mediated bridges [33, 39]. Reduced dynamics of the headgroups [40] and H-bonded waters are shown relative to bulk water [35, 36, 41, 42]. Waters near the membrane interface have slower dynamics, relating them to proton transfer events [35]. Additionally, ions affect the membrane bilayer structure and thereby the interaction of the membrane with other molecules, including water and proteins [202, 203].

In this study, two 100-ns MD simulations were performed for anionic membrane models, consisting of 4:1 POPC: POPG and 5:1 POPC: POPG, ions, and water. Similar lipid concentrations were studied in [204] and [205] as models of the bacterial membrane. We investigated the water H bonding at the interface of bilayers and characterized their dynamics. The properties of anionic lipid membranes composed of lipid mixtures and interactions between charged and polar phospholipids with sodium ions are studied.

We find that the anionic membrane interfaces are characterized by dynamic clusters of lipid molecules whose headgroups are directly H bonded or bridged via water or sodium ions. Water bridging dominates, followed by direct H bonding and, to a much lesser extent, ion bridging.

Phosphate groups of POPG lipids have, on average, slightly fewer H bonds with waters than POPC but additional H bonding is provided by hydroxyl groups of POPG glycerols. We observe that approximately 50% of the lipids in each leaflet of both bilayers participate in dynamic H bonding facilitated by one-water bridges. Our topology path algorithm revealed linear paths of interconnected lipids by one-water H-bonded chains with lengths of up to 5 individual one-water bridges. In the two simulations, the proportion of lipids increases to 85-86% when accounting for two-water chains between lipid phosphates. Roughly 92% to 95% of the lipids on each leaflet can be connected by three, four, or five water-mediated bridges.The visualized H-bond networks of waters inter-connecting lipid phosphates reflect our findings giving a picture of dynamic and dense network connections. Anionic lipids are found to form transient clusters of 2-3 POPG lipids direct H bonded to each other or mediated by positively charged ions and water molecules involving on average half of anionic lipids in total interactions on each leaflet.

In summary, we use our in-house algorithm implementations and graph theory to analyze the complex interactions derived from atomistic MD simulations of the two membrane models. With our algorithms, we can detect the lipid clusters, and the water or sodium residence times, and visualize their geometric arrangement (topology) in each bilayer leaflet giving essential insights into the H-bond interactions and possible patterns of connections. The extent of studies involving interactions of protein and lipids is of interest giving a better understanding of the interplay of the composition complexity and local clustering of lipids and proteins and the H bonding of water molecules.

4.1.2 MD Simulations Protocol

CHARMM-GUI was utilized to build two zwitterionic lipid bilayer models, comprised of 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphatidylcholine (POPC), and negatively charged 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphatidylglycerol (POPG) lipids with a 4:1 and 5:1 POPC: POPG ratio, respectively [206, 207]. The CHARMM 36 potential energy function was utilized to characterize the lipid molecules [208-210], while TIP3P was employed to model the water molecules [211]. To achieve a neutral charge in the system, ions were introduced. The CHARMM-GUI protocol for velocity rescaling was used to achieve geometry optimization and primary equilibration of each system. The *NPT* ensemble (constant number of particles: *N,* constant temperature: *T* =303.15 K, and constant pressure: *P* = 1 bar) was used, a Langevin dynamics scheme with a collision frequency of 5 ps^{-1} and a Nosé–Hoover Langevin piston [150] was utilized throughout equilibration and production runs. The smooth-particle mesh Ewald method was applied for the calculation of Coulomb interactions [141] [212]. A switching function 10-12 Å was employed for real-space interactions. SHAKE was used for fixing the lengths of bonds that involve H atoms [213]. A 1 fs integration step was employed during equilibration and the first 1 ns of production runs, and a reversible multiple time-step algorithm was used for production runs [143]. Every 10 ps, coordinates were stored. The final 50 ns of each *NPT* simulation were used to calculate H bonding and average characteristics. To examine the fast dynamics of waters that H-bond lipid phosphate groups, we performed 5 *NVE* simulations under constant volume (*V*) and constant energy (*E*) conditions by using an integration step of 1 fs and recording the coordinates every 10 fs for each simulation of 1 ns. Each *NVE* run was executed independently from the end of the *NPT* simulations. NAMD was operated to perform all the simulations [214, 215].

4.1.3 Results and Discussion

4.1.3.1 H-bonding analyses

Our implementations were used to calculate the dynamic H bonds in the two simulated systems. In the 4:1 POPC: POPG bilayer, the average number of water H bonds is found ~5.2 ± 0.1 and 4.6 ± 0.2 per POPC and POPG phosphate groups, respectively (Figure 4.1). For the 5:1 POPC: POPG bilayer, we calculated 5.1 ± 0.1 per POPC and 4.6 ± 0.2 per POPG lipid phosphate group. Values are similar between the two systems.

We examined the contribution of each oxygen atom of the phosphate and glycerol group to water H bonding. We found that most of the water H bonds are formed by the non-ester O13 and O14 oxygens (Figure 4.1b), with an average of ~2.0 water H bonds. However, O11 and O12 oxygens (Figure 4.1b) have on average ~0.6 water H bonds. Our results are confirmed by previous studies on dipalmitoylphosphatidylcholine, DPPC [29], and dimyristoylphosphatidylcholine, DMPC [216].

Table **2** summarizes below the average number of water H bonds of each lipid oxygen depicted in Figure 4.1b, for the two bilayer models.

Table 2: On average, the number of H bonds computed for POPC and POPG lipids' selected oxygen atoms. Figure 4.1b shows labeled the oxygen atoms presented in this table. Table is from [22], a study presented in this thesis.

Lipids	Average number of H bonds							
	O11	O12	O13	O14	O22	O32	OC2	OC3
4:1 POPC:POPG								
POPC	0.6±0.04	0.7±0.03	2±0.06	2±0.05	0.8±0.04	0.8±0.04	-	-
POPG	0.6±0.08	0.4±0.07	2±0.1	2±0.1	0.7±0.08	0.8±0.09	0.8±0.1	1±0.1
POPC+POPG	0.6±0.04	0.6±0.03	2±0.05	2±0.05	0.8±0.04	0.8±0.04	0.8±0.1	1±0.1
5:1 POPC:POPG								
POPC	0.6±0.04	0.7±0.03	2±0.06	2±0.06	0.8±0.05	0.8±0.04	-	-
POPG	0.6±0.08	0.4±0.08	2±0.1	2±0.1	0.8±0.09	0.8±0.09	0.8±0.1	1±0.1
POPC+POPG	0.6±0.04	0.6±0.03	2±0.06	2±0.05	0.8±0.04	0.8±0.04	0.8±0.1	1±0.1

For the POPG lipids, additional H bonds are formed between the hydroxyl groups and water (Figure 4.1). Hydroxyl groups can act as H-bond acceptors to water with an average of ~0.7 ± 0.1 H bonds per lipid but additionally as H-bond donors giving a total average of ~2.5 ± 0.1. POPG lipids, through hydroxyl groups, make inter-lipid H bonds with phosphates of both POPC and POPG lipids with an average of ~0.2 ± 0.1 H bonds per glycerol group. This result clarifies why POPG phosphate groups form fewer H bonds on average with water compared to POPC. The average number of H bonds between glycerol groups of POPG lipids was found ~0.04 ± 0.1 per POPG lipid, as there are a few anionic compared to zwitterionic lipids in the bilayer (POPC: POPG ratio 4:1 and 5:1). No significant difference was obtained between the two anionic membrane models.

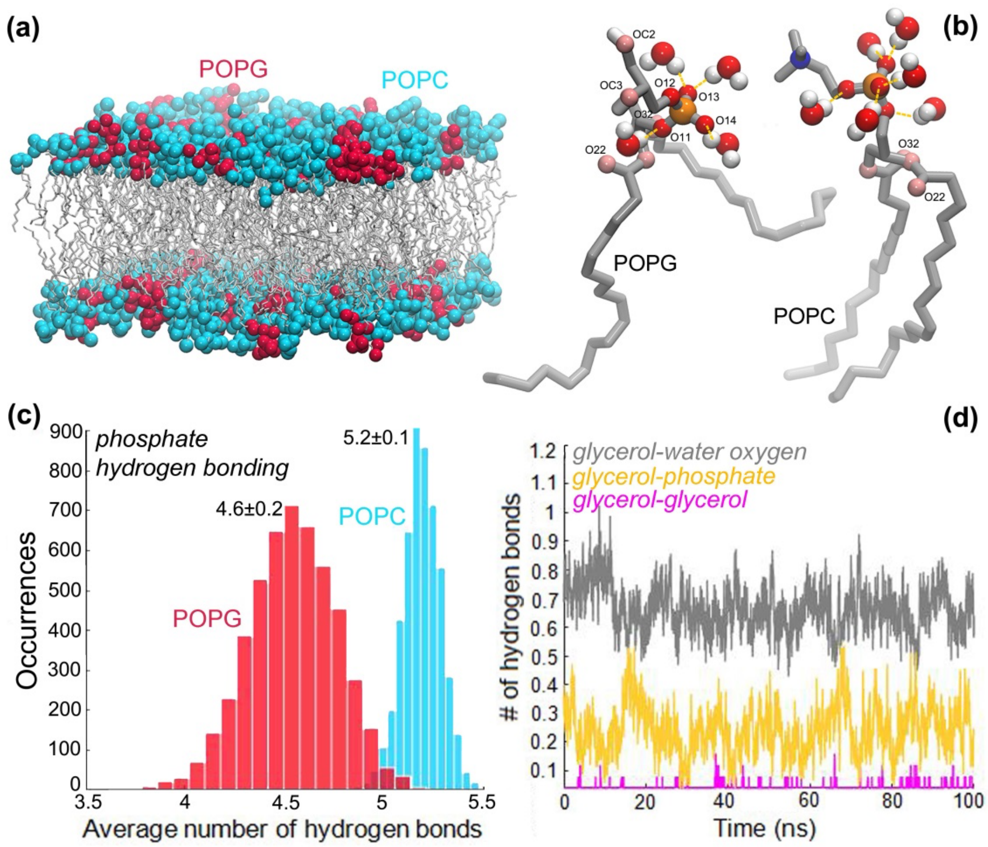


Figure 4.1: H bonding at an anionic lipid membrane made up of a 4:1 ratio of POPC and POPG lipids. (a) Cyan and red van der Waals spheres represent POPC and POPG lipid headgroups, respectively, whereas gray bonds represent lipid alkyl chains. Molecular graphic was generated in VMD using a coordinate simulation snapshot of the system [23]. (b) Water H bonding at a POPC and POPG lipid's phosphate group region. Selected oxygen atoms have been tagged and yellow dashed lines indicate H bonding. (c) Histogram of the average number of water H bonds per phosphate group of POPC and POPG computed from the last 50 nanoseconds of the simulation. (d) Plots of the average number of H bonds per glycerol group formed during simulations connecting glycerol–water, glycerol-phosphate, and glycerol–glycerol colored gray, yellow, and magenta, respectively. [22]. Image is from [22], a study presented in this thesis.

4.1.3.2 Dynamic lipid clusters

H-bonded water bridges can be formed among two phospholipids at the same time [40]. We performed calculations for one to five water H-bond bridges between phosphate groups of both POPC and POPG lipids.

For the 4:1 and 5:1 POPC: POPG membrane models, we found that ~50% of lipids participate in one-water H-bond water bridges based on the *NPT* simulations (Figure 4.2b). In the case of two-water bridges, ~85% of lipid phosphates are found to be paired by waters, on average. In the case of three-, four-, and five-water bridges, the percentage of lipids is increased to 92%, 94%, and 95%, respectively. Percentage results are calculated for each bilayer leaflet and average values are reported (Figure 4.4).

We run our residence times code for the one-water mediated water bridges using the 1ns simulations in *NVE.* We found that water bridges are dynamic with a lifetime of ~4ps in the two systems studied (Figure 4.2c). Short lifetimes of the one-water phosphate bridges are reported in the study of DMPC lipids by [40].

An essential point to address is the importance of the simulation conditions and the selection of the simulation time step to the calculation of water residence times (equations in Section 3.4). When using the *NPT* simulation trajectories having a 10ps time step, we concluded that the average water residence times of the one-water/phosphate H-bond bridges rise incorrectly to 60 ps. How the residence times rely on the simulation time step with which water H bonding is calculated during simulations is highlighted in the study of the bovine pancreatic trypsin inhibitor (BPTI) by [217].

Water-H bonding forms clusters on each leaflet of the bilayer, causing around half of POPC lipids to interact with other POPC lipids. These clusters usually consist of 3-5 POPC lipids.

Experiments and computations have suggested that POPG lipids form intra-POPG and inter-lipid H bonds with a high propensity [26, 27, 201, 218]. Our calculations for the two anionic membrane models are based on the Network Component's algorithm as illustrated in Figure 3.2. POPG lipids through their hydroxyl and phosphate groups may direct H bonded to other POPG lipids, bridge to other POPG lipids by H bonding waters, or by interacting with cations, forming clusters (Figure 4.3).

According to our calculations, around 20% of the anionic POPG lipids are engaged in one-water H-bond bridges with other POPG lipids forming clusters that consist of ~3-4 lipids on each leaflet of the 4:1 and 5:1 POPC: POPG bilayers, on average (Figure 4.3). We calculated the lifetimes of these clusters and we found that they are highly dynamic, around 2.3 ps. Except for water bridging, POPG lipids can also form clusters with other POPG lipids by H bonding mediated by the POPG glycerol (Figure 4.4b). These clusters formed by direct H bonds between POPGs contain around 13-15% of the total number of lipids with a size of 2-3 lipids and a lifetime of 35ps in 5:1 and 44ps in 4:1 POPG: POPG.

POPG lipids form clusters by interactions with sodium ions (Figure 4.3). We define as ion-lipid interaction when the Na^+ is within 3.3Å of the phosphate group of lipids in the bilayer. That distance is according to the radius of the first shell of Na+ ions. On average, there are 1-2 clusters consisting of 2-3 POPG lipids on each bilayer leaflet (Figure 4.3). Around 12% and 14% of POPG lipids interact with other POPG lipids by ion interactions in the 4:1 POPC: POPG and 5:1 POPC: POPG membrane, respectively.

It is shown that sodium ions can penetrate the anionic 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine (POPS) lipid membranes linking lipid headgroups [31]. The size of the anionic clusters formed by ions is consistent with studies on DPPC and DPPS [219]. The ion concentration and its nature are important parameters to the number, size, and dynamics of clusters that can be shaped during MD simulations of the lipid bilayers.

Sodium-mediated clusters are very dynamic with lifetimes as low as ~0.3ps. We found a few clusters with higher lifetimes of about 49ps or 107ps, and 118ps in the 4:1 POPC: POPG. As an example, we illustrate the cluster with a lifetime of 107ps in Figure 4.5. There are two POPG lipids that interact closely with sodium. This geometry allows the POPGs to be also mediated by H-bonded waters. The water bridges are small with a length of 1-2 waters in the chain.

4.1.3.3 Linear pathways of one-water mediated bridges

Phosphate pairs can form pathways of linked lipids through one-water-mediated bridges (Figure 4.2a). We study the length, L, of the linear paths described as the number of distinct one-water phosphate bridges in the path. For instance, L=3 indicates that 4 lipids are connected by three one-water phosphate H-bond bridges. The network component analysis was subsequently employed to determine linear one-water bridge paths with L > 1 (Figure 4.2a) while excluding circular paths from the search (see Section 3.3).

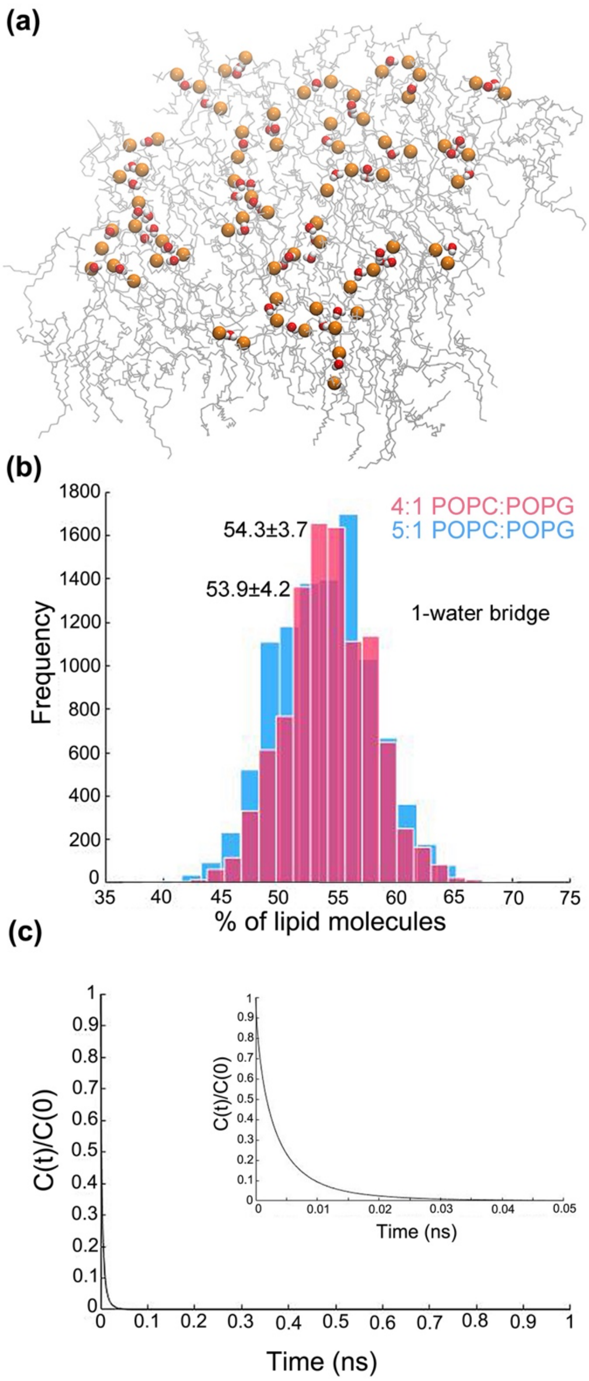


Figure 4.2: Lipid phosphates connected via water H-bond bridges. (a) Molecular graphics of the 5:1 POPC: POPG bilayer showing one-water H bonding between POPC or POPG phosphate groups on the extracellular side of the bilayer. Each phosphate group is represented by a phosphorus atom depicted here as an orange sphere. (b) Histogram representation that shows the percentage of lipids participating in one-water H-bond bridges during the last 50 ns of each simulation. (c) Normalized residence-time correlation function of one-water bridges between phosphate groups in 5:1 POPC/POPG membrane. (Inset) Residence-time correlation function, as shown for the first 50ps of the *NVE* simulation for one-water bridges in 5:1 membrane, rapidly decays on a picosecond timescale [22]. Image is adapted from [22], a study presented in this thesis.

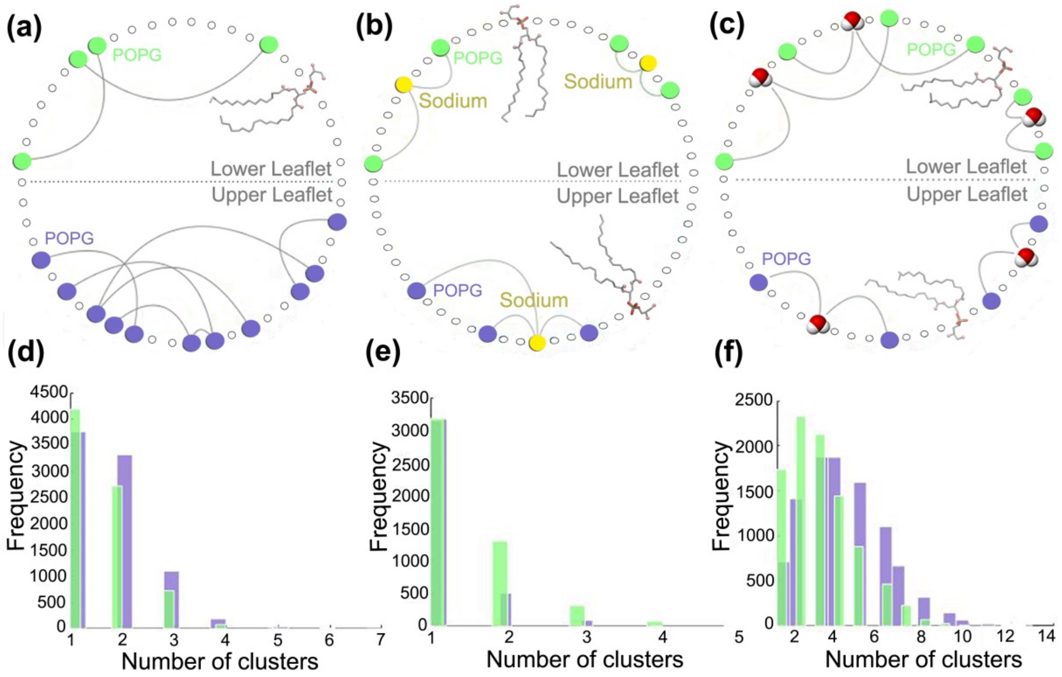


Figure 4.3: POPG lipid clusters in the 4:1 POPC: POPG membrane model. Panels a-c depict circular connective networks of direct (panel a), sodium-mediated (panel b), and one-water mediated clusters (panel c) between POPG lipids from one coordinate snapshot of the *NPT* simulation. Green and purple circles for the extracellular and the cytoplasmic side of the bilayer, respectively, show POPG lipids that interact with each other via H bonds. The number of circles in the 4:1 POPC: POPG system correlates with the total amount of anionic lipids. The histograms in panels d-f show the number of POPG clusters mediated by direct POPG interactions (panel d) vs by sodium ions (panel e) or by one water molecule (panel f). Average values calculated for the extracellular and the cytoplasmic side of the bilayer during the last 50ns of the simulation are represented by green and purple bars, respectively [22]. Figure panels a-c were generated for the scope of this thesis. Panels d-f are from [22], a study presented in this thesis.

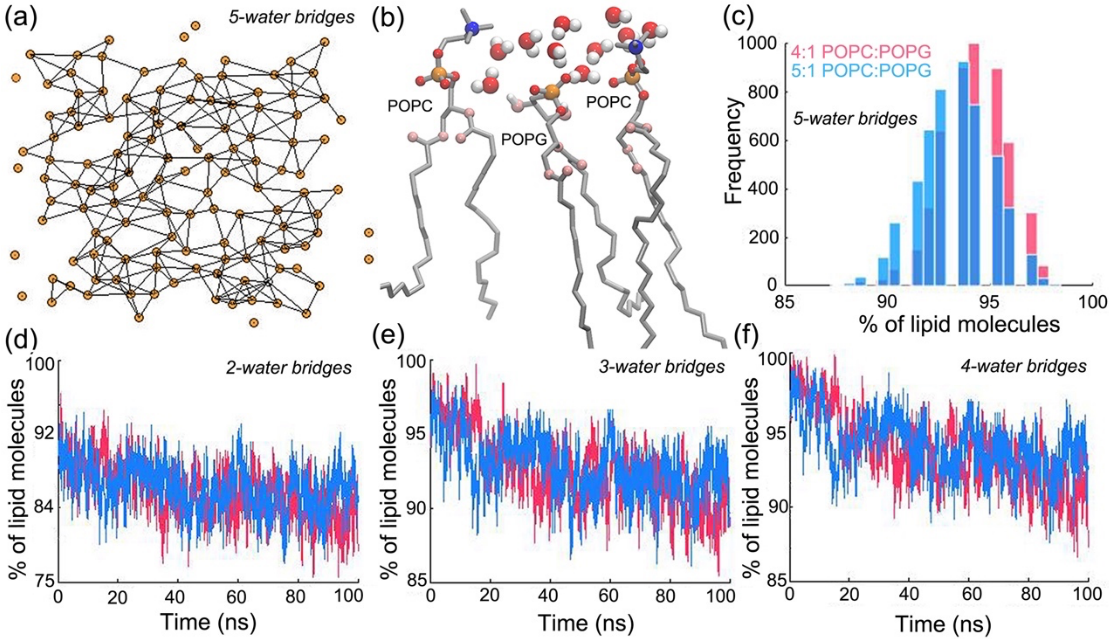


Figure 4.4: Water H-bond bridges inter-connecting phosphate groups. (a) One leaflet from the 4:1 POPC:POPG membrane is visualized as a network with orange spheres representing each phosphate group. The visualization shows five water molecules involved in H bonding between lipid phosphate groups after 100 ns of the simulation. (b) Molecular graphics illustrating 5-water bridges between POPC and POPG phosphate groups. (c) Histogram representation of the percentage of lipids involved in 5-water H-bond bridges for each system during the last 50ns of each simulation. Panels d-f give the percentage of lipid molecules involved in H-bond interactions mediated by respectively 2-, 3-, and 4- water bridges. The data for the 1-water bridges are given in Figure 4.2 [22]. Image is from [22], a study presented in this thesis.

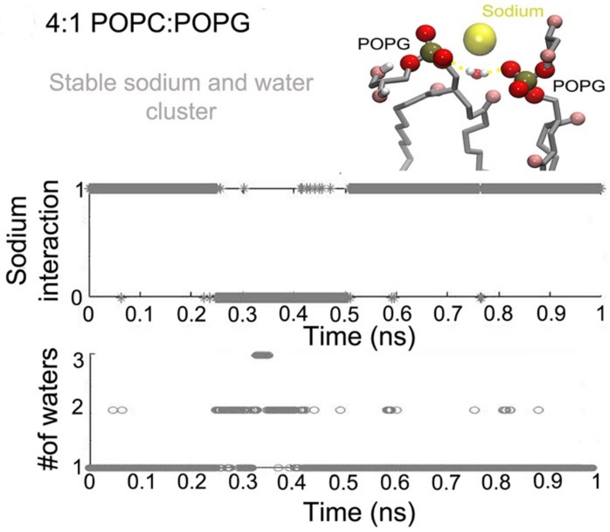


Figure 4.5: Sodium-mediated interaction stabilizing two POPG phosphate groups in the 4:1 POPC: POPG membrane model. Phosphates are additionally H bonded by water bridges of lengths 1 and 2. The number of water bridges connecting the two POPG lipids and the sodium interaction time series are displayed. For clarity, we present information from coordinate

images collected from the *NVE* simulation every 1 ps [22]. Image is from [22], a study presented in this thesis.

Longer linear pathways are quite rare; most paths have one or two lengths (Figure 4.7). There is a substantial number (about 20) of one-water phosphate bridges with L = 1 during the 100 ns simulation (Figure 4.6). Linear paths with L = 2 and L=3 are less common, and their number can be ~ 5 and ~ 2–3, respectively (Figure 4.6). Long linear clusters of one-water H-bond bridges with L = 4 and L = 5 are seldom observed, implying that these paths are uncommon.

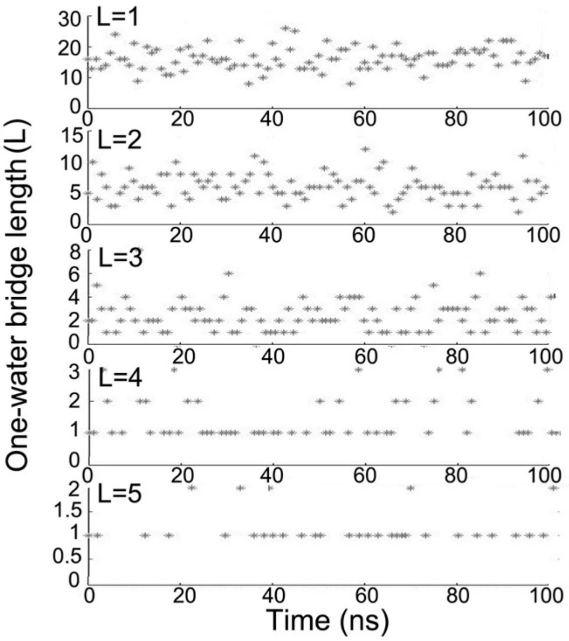


Figure 4.6: Linear path lengths of one-water-mediated bridges in 4:1 POPC: POPG membrane model. The number of pathways with lengths, 0<L≤ 5, is shown as time series. Here, we present for clarity results from coordinate images read every 1 ns [22] from a total 100ns simulation time. Image is from [22], a study presented in this thesis.

4.1.3.4 Dynamic networks of H-bonded lipids

We utilize a protocol to extract graphs of the lipid phosphate groups (nodes) connected by H bonds (edges). For each lipid pair, there are many water bridges with various lengths during the simulation. Our algorithm keeps the water bridge length with the highest occupancy. The water's H-bonding networks were analyzed independently for two *NPT* simulations lasting 100 ns each, as well as for *NVE* simulations.

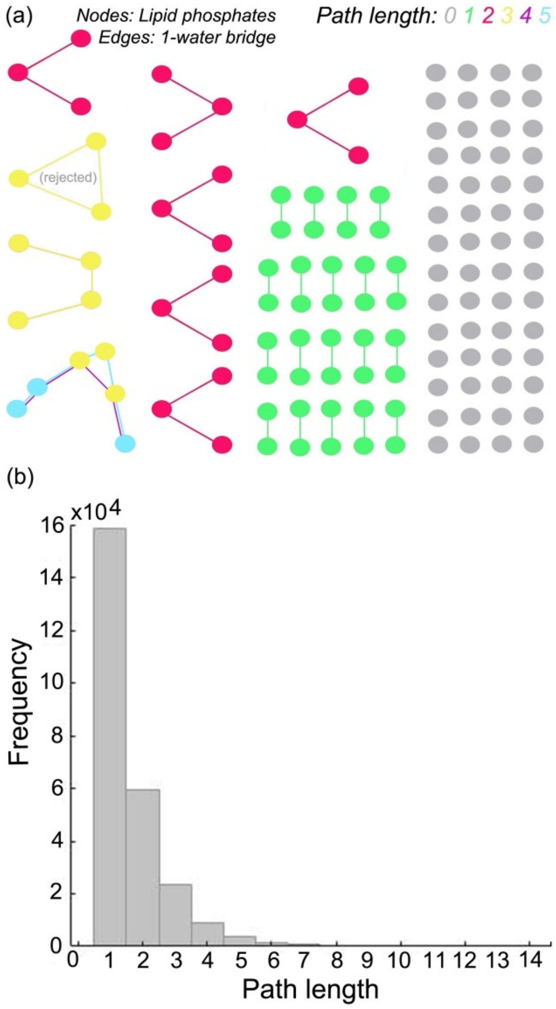


Figure 4.7: Linear pathway lengths of one-water H-bond bridges between phosphate groups in a 4:1 POPC:POPG bilayer. (a) Topology representation from a coordinate image obtained from the simulation that depicts one-water-mediated lipid pathways in the extracellular side of the bilayer. Path lengths L are found to range from 0 (no water-mediated bridge) to 5 (a linear path made up of five water-mediated bridges). The spheres are colored-coded and represent lipid phosphates involved in linear paths with lengths L of 0, 1, 2, 3, 4, and 5, respectively. Paths that are cyclic were ruled out during the path search. (b) The number of pathways with lengths, L, is shown as a histogram. The study was carried out for the 100 ns length of the *NPT* simulation. [22]. Figure panel a is from [22], a study presented in this thesis. Figure panel b is generated for the scope of this thesis.

Based on our results, the lipid headgroup region is described as a complex web of dynamic water-mediated bridges connecting the lipid phosphate groups. In a picosecond timeframe, the water bridges break and are reconstructed quickly (Figure 4.8). Based on the data analyses, it has been observed that the movements of lipids are closely associated with the dynamics of H-bonded water chains. When two lipid molecules come close to each other, H-bonded water chains of varying lengths can connect the lipids (Figure 4.8c, d). During the simulation, water molecules formed bridges between the two groups of lipid phosphates. These bridges consisted of one to five water molecules and existed for varying percentages of time ranging from 33.8% to 8.4%.

We investigated the dynamics of water bridges between POPC, POPC and POPG, and POPG lipids using the last 100ps of the five *NVE* trajectories of 1ns each, with coordinates saved every 10 fs. We used an occupancy cutoff of 30% to include in the analysis. This gave us ~20% of the total number of bridges.

Water bridges with 1-2 hydrogen-bonded water molecules have high occupancies in lipid pairs composed of POPC and POPG lipids, with up to 71% occupancy in a single water bridge between two POPC lipids (Figure 4.9). On average, larger H-bonded water bridges of 3–5 waters have lower occupancies (36–47 %) (Figure 4.9). The occupancy values of the water-mediated bridges for the three types of lipid pairs are substantially identical (Figure 4.9), irrespective of how long is the water-mediated lipid bridge.

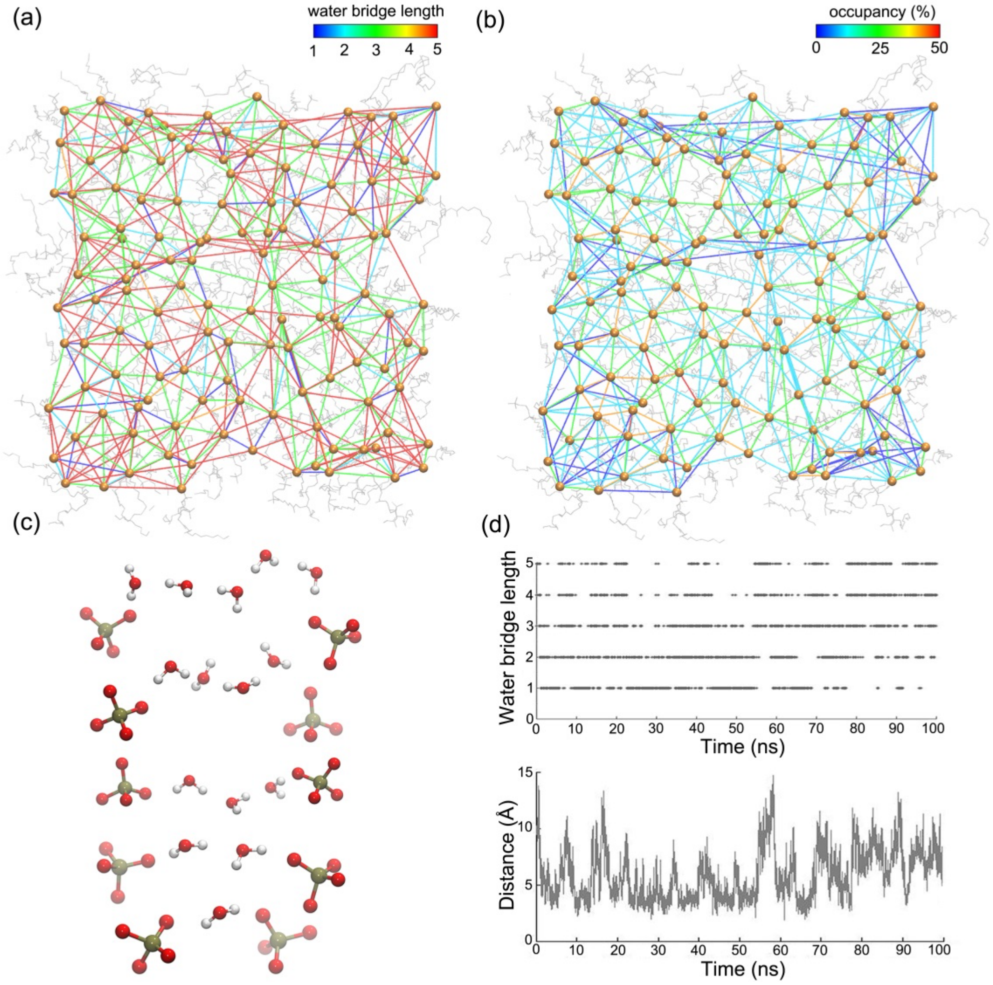


Figure 4.8: Dynamic water-mediated H-bond networks at the surface of the 4:1 POPC: POPG bilayer. Network visualization based on (a) water bridge lengths. Each edge in the graph represents water H bonds and is color-coded based on the bridge length. We consider up to eight connections per lipid, and the largest occupancy bridges are displayed, for simplicity.

Network illustration based on (b) water bridge occupancies from blue (0% occupancy) to red (50% occupancy). The networks are overlaid onto a simulation-derived typical lipid structure. Analyses are from the last 50ns of the *NPT* simulation. (c) Molecular visualization of two lipid phosphates and the water chains that H bond them with lengths 1 to 5. (d) Time series of the length of water bridges and the phosphate oxygen atoms' minimum distance during the simulation [22]. Image is from [22], a study presented in this thesis.

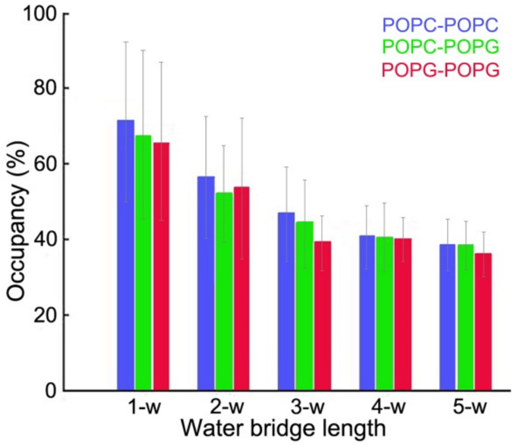


Figure 4.9: Water bridge H-bonds dynamics for POPC–POPC, POPC–POPG, and POPG–POPG. Results are presented for water bridge lengths ranging from 1 to 5. Analysis was conducted for the final 100 ps of 5 independent 1 ns simulations in *NVE*, the mean occupancy and typical deviation is shown. An occupancy cutoff of 30% is applied [22]. Image is from [22], a study presented in this thesis.

4.1.4 Summary

In this study, two hydrated bilayers composed of 4:1 and 5:1 POPC: POPG lipids, were modeled and simulated [22]. We implemented computational approaches to examine and visualize the H-bonding dynamics at the interface of these two bilayers. We discovered that the membrane interface is marked by a complicated and dynamic H-bonding network whereby lipid phosphate groups bridge via water molecules; these dynamic interactions can give rise to clusters of lipids whose headgroups link via direct H bonds or via water molecules and ion bridges in the picosecond timeframe (Figure 4.1 – Figure 4.9). Our results are mostly comparable for the two systems.

Water bridging is the most common interaction found in our simulations. About 50% of lipids in the extracellular and cytoplasmic leaflet participate in dynamic H bonding mediated by one-water bridges forming linear paths of various lengths during the simulation (Figure 4.2, Figure 4.6, Figure 4.7). Permitting 2-5 H-bonded waters that interconnect lipid headgroups elevates the percentage of water-bridged lipids to 85–95% (Figure 4.4).

Extended water H-bonded bridges between lipids, as reported here for POPC: POPG bilayers may lead to lateral proton transfer along the membrane plane [35]. Like in proteins, lipid headgroups enhance the lifetime of waters at the bilayer interface relative to the bulk participating in H-bond bridges and clusters with waters (Figure 4.8). It is indicated that protein groups such as in the PsbO subunit of Photosystem II or the SecA protein, can be joined via dynamic H-bonded water molecules, causing a longer lifetime of water molecules, and diminishing their mobility near the protein surface [220-222]. Their H bonds break and are reconstructed on the picosecond–nanosecond timeframe [31].

We showed that anionic lipids create ephemeral clusters of 2–3 POPG lipids by direct or water/ion-mediated bridges (Figure 4.3). In the two simulations presented in this thesis, ~ 50% of the anionic lipids bind with H bonds with other anionic lipids. Altogether, the average lifetimes of the one-water phosphate bridges and sodium-mediated clusters suggest dynamic interactions, while sodium-mediated lipid clusters can also be sustained over time (Figure 4.5).

Adding calcium ions can increase the lifespan of water molecules that form H-bond bridges between lipid headgroups. This is because the movement of the lipids is dependent on the dynamics of the H-bonded water chains that connect two lipid phosphates (Figure 4.8), the decreased self-diffusion of lipids following calcium binding [32] could be attributed to the slower dynamics of the phosphate/water H-bond networks.

Lipid clustering via direct, water- or sodium-mediated interactions might offer a platform on which proteins or pharmaceutical molecules can attach to cationic surfaces.for protein or medicinal molecule binding to cationic surfaces. Our simulations indicate that the interfaces between anionic lipid headgroups in membranes contain dynamic clusters. Within these clusters, a few lipid molecules interact with each other through transient H-bonded water chains (Figure 4.1– Figure 4.9). In lipid/water bridges with and without POPG, the dynamics of these bridges seem to be substantially comparable (Figure 4.9). When a protein or therapeutic chemical binds to the membrane interface, it is likely that the size and dynamics of the lipid clusters will change.We propose that the lipid headgroup interface's network of H bonds (Figure 4.2) may play a role in modifying the dynamics of lipid molecules at a distance away from the protein binding site.

We employed computational techniques to compute, analyze, and visualize the complex and dynamic lipid H bonding at the membrane interface. We could expand the range of our analysis of linear connections to include pathways that lead to more intricate and non-linear structures. This could entail examining the pathways that connect phosphate bridges via longer H-bonded water chains, as well as pathways that consider the potential effects of ion binding on the dynamics of the water/lipid H-bonded networks. The studies presented in the next chapters of this thesis present theoretical and computational approaches to address those challenges using graph theory and algorithm implementation.

4.2 Algorithm to Catalogue Topologies of Dynamic Lipid Hydrogen-Bond Networks

The study presented in this chapter is available in the publication:

Karathanou, K. and Bondar, A.N., 2022. Algorithm to catalogue topologies of dynamic lipid hydrogen-bond networks. Biochimica et Biophysica Acta (BBA)-Biomembranes, p.183859.

Figures and tables originally published in the Journal Biochimica et Biophysica Acta (BBA) - Biomembranes have been reproduced/reprinted with permission from BBA-Biomembranes 1864, no. 4 (2022): 183859. Copyright © 2022 Elsevier B.V.

The codes are published in Mendeley repository: "Karathanou, Konstantina (2022), "Graph-based algorithm for common topologies of dynamic lipid clusters", Mendeley Data, V2, DOI: 10.17632/9c7f9vbymh.2"

and as a GitLab repository: "Karathanou, Konstantina (2022), "Graph-based algorithm for common topologies of dynamic lipid clusters", https://gitlab.com/kkarathanou/algorithm-for-lipid-cluster-topologies".

4.2.1 Introduction

Recent technological advances, bioinformatics, and computer engineering have provided efficient methods for the analyses of big data derived from computer simulations and the prediction of biophysical phenomena.

Graph theory and the concept of networks have been introduced over the past years for the analyses, visualization, and prediction of interactions in complex biological systems like proteins [102, 223-231], and to a lesser extent in lipid membrane systems [22, 216, 231-233]. In graph theory, we perceive a complex system as a network of interacting sites (nodes) connected by lines (edges). Nodes usually are functional groups (e.g., protein amino acids, lipid headgroups), and the edges represent the type of interactions between nodes (e.g., distances, H-bonding connections, ion interactions, conformational transitions) [234]. Edges can be weighted (e.g., according to H-bond occupancies, average distances, etc.) or have a direction from one node to another.

Networks also contain hubs; nodes with a higher degree of importance in the graph (e.g., nodes with many neighbors, nodes as part of significant biological pathways) [102] (Figure 4.10).

To cluster nodes that interact closely with other nodes in the graph, data clustering techniques have been proposed and used in a wide range of fields, such as pattern recognition, machine learning, image analysis, and data/web mining [235-237]. Data clustering techniques describe the process of grouping data with a high degree of similarity into classes while separating dissimilar data into different classes. Clustering algorithms based on graph theory regard nodes as data points and edges as the relation among them. Nodes are grouped by using the concept of graph minimum weight division [238] or the minimum spanning tree [239]. Additionally, in graph theory, the connected component analysis is used to find clusters of connected nodes in the graph (subgraphs) using either the Breadth-first or the Depth-first search algorithms [180]. Visualization tools have been developed for the detection of clusters in biological networks like Cytoscape [240], MDAnalysis [241, 242] for analyses of dynamical H bonds and Bridge [231], a graph-based algorithm to efficiently analyze dynamic H-bond networks in proteins or lipid membrane models.

Computations of H-bond lipid clusters are performed using geometric criteria or graph-theory approaches. MD simulation studies in bilayers containing phosphatidylcholine (PC), phosphoglycerol (PG), phosphoserine (PS), phosphoethanolamine (PE), and phosphatidic acid (PA) lipids [22, 40, 216, 232, 243] [244] reveal a highly dynamic and “mosaic” H-bond network of direct or mediated by water bridges lipid pairs. Water H bonds between lipids reduce the mobility of head groups and therefore stabilize the bilayer structure [216].

The study by [245], shows that H bonds internal to a cluster of PE lipids remain stable for longer than the cluster size and that clusters of lipids may diffuse and reorient as groups. The study by [27] shows that ion inter-lipid bridges and strong molecular H bonding play a vital role in the attractive interactions between POPG lipids, overcoming the electrostatic repulsion between negative charges of PG headgroups.

Seven distinct lipid species, including the glycolipid GM3 in the outer leaflet and the anionic lipid phosphatidylinositol 4,5-bisphophate (PIP2) in the inner leaflet, are included in the complex asymmetric plasma membrane model that was simulated by [28], revealing clustering of the GM3 and to a smaller extent of the anionic PIP2. Cholesterol, PIP2, and GM3 interact preferentially with transmembrane proteins when they are inserted. Lipid nanodomains and membrane geometry may be related, as evidenced by the finding that membrane curvature is correlated with the local lipid composition.

A graph-theory approach is followed in the study by [233] to describe the interfacial properties of hydrated phospholipid and mono galactolipid bilayers and computes properties like the H-bond lipid clusters, their sizes, the number of network bridges, etc. from weighted undirected graphs.

In our study, we model bilayers that comprise of zwitterionic lipids, 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine (POPE), and anionic POPG, key components of the inner bacterial membrane [204], POPS lipids and complex membrane mixture of PE and PG lipids that contain a cyclic moiety and imitates the diverse lipids population within the *E. coli* cytoplasmic membrane [205].

To identify H-bonded lipid/water clusters and characterize their dynamics, we implemented a set of algorithmic tools that allow us to compute and visualize networks of interactions during the simulation time used in our analyses. We introduce a new concept in lipid clusters, the topology, as the spatial distribution of lipids in membranes, we catalog types of topological schemes like linear or circular connected lipid clusters, and we present our findings.

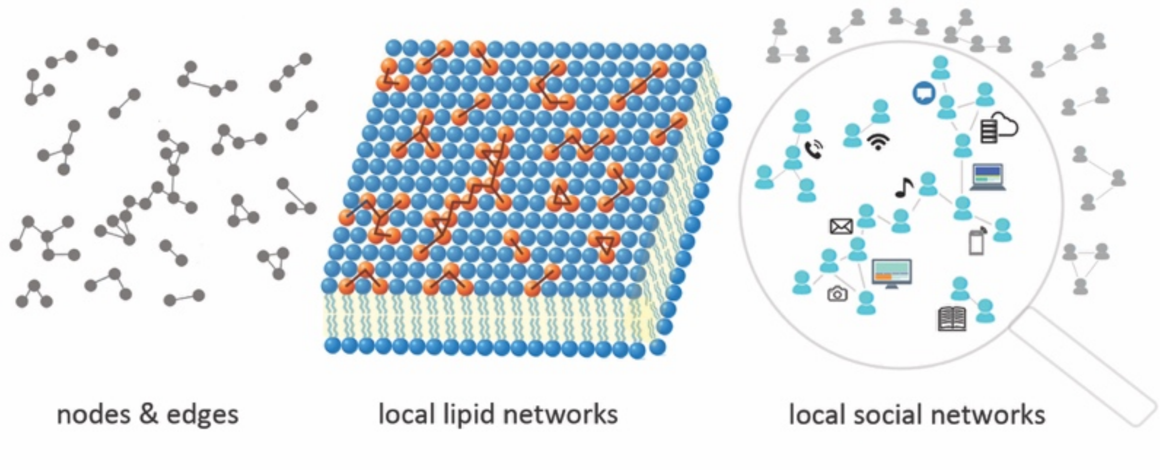


Figure 4.10: Inspired by graph-theory and social networks, algorithms to cluster and catalogue lipid molecules in bilayers are built. Image represents a local lipid network, the corresponding local social network and nodes & edges connections based in graph-theory [187]. Image is the graphical abstract of [187], a study presented in this thesis.

4.2.2 Coordinates set-up & MD Simulations

Coordinates for all membrane models were generated using CHARMM-GUI [206, 207] (Table 3). Bilayers consisting of POPE lipids, 3:1 and 5:1 POPE:POPG, and POPS lipids were constructed in solution at a neutral charge, or in the presence of 0.15M KCl (Table 3). To model the *E. coli,* we used the lipid types and composition of the membrane denoted as Top6 in ref. [205]. The Top6 membrane consists of lipids with monounsaturated acyl chains: POPE, 1-oleoyl-2-palmitoleoyl-snglycero-3-phosphoethanolamine (OSPE), 1-palmitoyl-2-palmitoleoyl-snglycero-3-phosphoglycerol (PSPG), and lipids with ring acyl chains: 1-palmitoyl-2-cis-9,10-methylene-hexadecanoic-acid-sn-glycero-3-phosphoethanolamine (PMPE), 1-

palmitoyl-2-cis-9,10-methylene-hexadecanoicacidglycero-sn-3-phosphoglycerol (PMPG), and 1-pentadecanoyl-2-cis-9,10-methylene-hexadecanoic-acid-snglycero-3-phosphoethanolamine (QMPE) having a PE: PG composition of 4:2:1 [205]. According to the latest version of CHARMM-GUI, we selected 3-palmitoleoyl-2-oleoyl-d-glycero-1-phosphatidylethanolamine (YOPE), and 1-hexadecanoyl-2-(9Z-hexadecenoyl)-glycero-3-phospho-(1'-sn-glycerol) (PYPG) lipids instead of OSPE and PSPG lipids used in [205] (Figure 4.11) and we doubled the size of the bilayer (Table 3). After constructing the system of membrane models, we followed the MD simulation protocol as described in Section 4.1.2. Each production run was extended to 200ns in *NPT*, in total 1.6 μs, for the whole set of simulations. The average values were calculated by considering the last 100 ns of each *NPT* simulation, unless stated otherwise.

Table 3: Collectively all the performed MD simulations. Results are provided for each simulation using the final 100ns from [187], a study presented in this thesis.

Membrane	**#Lipids/ leaflet**	**#Atoms**	**dP-P (Å)**
POPE	138	65922	42.7 ± 0.4
3:1 POPE:POPG	102 POPE, 34 POPG	65545	41.6 ± 0.5
5:1 POPE:POPG	115 POPE, 23 POPG	66387	41.9 ± 0.5
3:1 POPE:POPG(I)a)	102 POPE, 34 POPG	65660	41.7 ± 0.4
5:1 POPE:POPG(I)a)	115 POPE, 23 POPG	66499	42.2 ± 0.4
POPS	135	65646	41.8 ± 0.5
POPS(I)		65731	42.1 ± 0.5
E. coli Top 6	74 PMPE, 20 POPE, 12 QMPE,12 YOPE,16 PMPG, 14 PYPG	81738	38.8 ± 0.4

a)0.15M NaCl. b) Number of lipids in each leaflet.

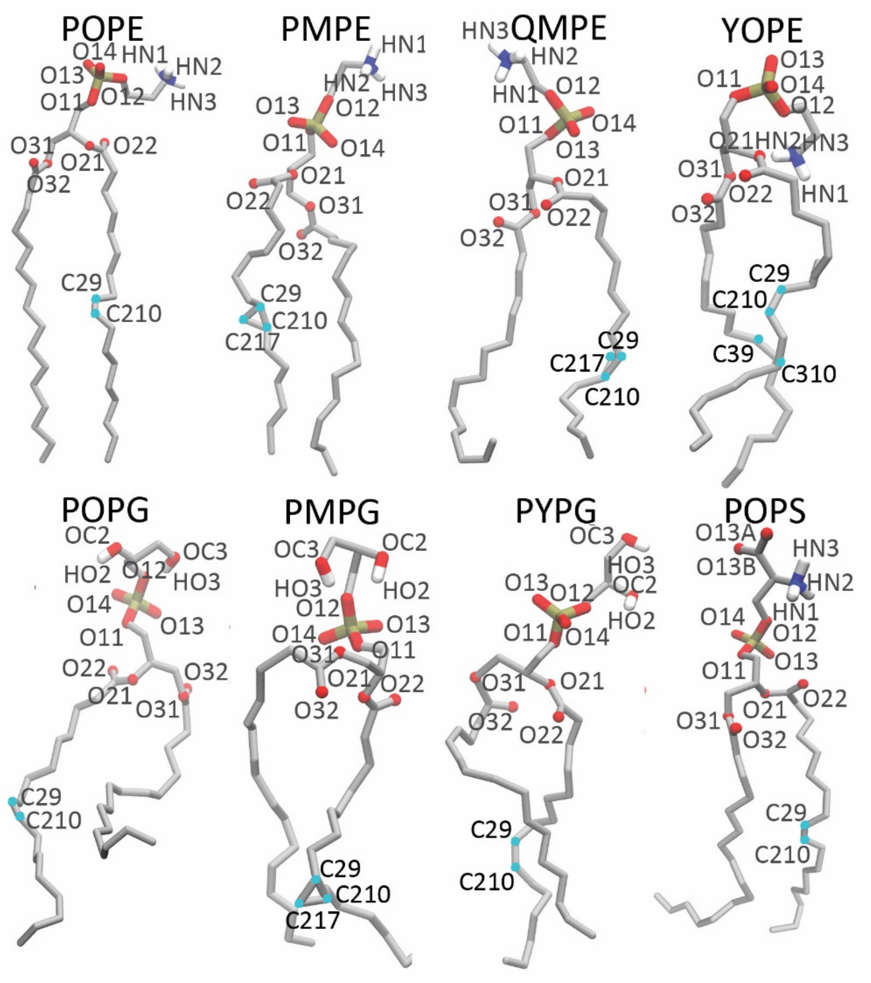


Figure 4.11: Illustration of the lipid molecules examined in this study [187]. All the visuals derive from coordinate snapshots of the simulations using VMD [23]. Image is from [187], a study presented in this thesis.

4.2.3 Results and Discussion

4.2.3.1 Bilayer thickness

The POPE bilayer thickness is close to that of the 3:1 and 5:1 POPE: POPG bilayers (dP-P = ~41.0 - 42.0 Å, Table 3). On the contrary, as previously reported [205], the Top6 *E. coli* membrane is approximately 3 Å thinner, with dP-P = 38.8 ± 0.3 Å (Table 3). Its lesser thickness is owing to the existence of lipids that contain shorter alkyl chains and lipids that include a cyclopropane ring (Figure 4.11). The POPS membranes are almost 42 Å thick (Table 3). The thickness values of POPE, POPE: POPG, and

POPS membranes from the present study (Table 3) are similar to the results from previous simulations [246-248].

From the *NPT* simulations, we calculated the bilayer thickness (Table 3) as the average separation between the peaks of the distribution of phosphorous atoms in the two bilayer leaflets using the MEMBPLUGIN extension [249] in VMD [23].

4.2.3.2 H-bonding dynamics

Each POPE lipid has around 1.7 H bonds with other lipids in a pure POPE bilayer. The amount of POPE-POPE H bonds reduces to approximately 1.3–1.5 when POPE lipids form H bonds with POPG in 3:1 POPE: POPG (Figure 4.12) or 5:1 POPE: POPG. Similarly, the *E. coli* membrane contains around 1.3 H bonds between PE lipids. In the *E. coli* membrane and two POPE: POPG bilayers, each PE lipid forms approximately 0.1-0.4 H bonds with PG lipids, while each PG lipid forms approximately 0.8-1.3 H bonds with PE lipids, on average (Figure 4.12). This aligns with previous calculations, indicating approximately 0.4 POPE-POPG H bonds per POPE lipid and 1.3 per POPG lipid [204]. POPS lipids form H bonds with other POPS lipids in the bilayer. The ammonium group serves as an H-bond donor, whereas the carboxylate, phosphate, and ester group as H-bond acceptor (Figure 4.11). We find on the average ~2.7-2.8 H bonds per POPS lipid (Figure 4.12). A summary of the average number of H bonds per lipid in membrane models is given in Table 4.

For all lipid types, we observed an average of 4.3 to 4.7 water H bonds per lipid phosphate group. This value is consistent with previous reports for POPC and POPG membranes [22]. Table 5 presents the average amount of water H bonds for each phosphate group of different types of lipids in membrane models.

Table 4: Average number of H bonds per lipid as computed from the *NPT* simulations from [187], a study presented in this thesis.

Simulation	**Lipid**	**# H bonds**
POPE	PE---PE / PE	1.7 ± 0.1
3:1 POPE:POPG	*i)* PE---PE / PE *ii)* PE---PG / PE *iii)* PE---PG / PG *iv)* PG---PG / PG	*i)* 1.3 ± 0.1 *ii)* 0.3 ± 0.1 *iii)* 1.0 ± 0.2 *iv)* 0.2 ± 0.1
3:1 POPE:POPG (I)		*i)* 1.2 ± 0.2 *ii)* 0.3 ± 0.1 *iii)* 1.0 ± 0.2

		iv) 0.2 ± 0.1
5:1 POPE:POPG		*i)* 1.5 ± 0.1 *ii)* 0.2 ± 0.1 *iii)* 1.1 ± 0.2 *iv)* 0.1 ± 0.1
5:1 POPE:POPG(I)		*i)* 1.4 ± 0.1 *ii)* 0.2 ± 0.1 *iii)* 1.1 ± 0.2 *iv)* 0.2 ± 0.1
Top6		*i)* 1.3 ± 0.1 *ii)* 0.3 ± 0.1 *iii)* 1.1 ± 0.2 *iv)* 0.1 ± 0.1
POPS	PS---PS / PS	2.7 ± 0.1
POPS(I)		2.8 ± 0.2

Table 5: Average number of water H bonds per lipid phosphate group as computed from the *NVE* simulations from [187], a study presented in this thesis.

Lipid	**Simulation**	**#phosphate/water H bonds**
POPE	POPE	*4.3 ± 0.1*
	3:1 POPE:POPG	*4.5 ± 0.1*
	3:1 POPE:POPG (I)	*4.5 ± 0.1*
	5:1 POPE:POPG	*4.5 ± 0.1*
	5:1 POPE:POPG(I)	*4.4 ± 0.1*
	Top6	*4.4 ± 0.2*
POPG	3:1 POPE:POPG	*4.7 ± 0.2*
	3:1 POPE:POPG(I)	*4.7 ± 0.1*
	5:1 POPE:POPG	*4.5 ± 0.2*
	5:1 POPE:POPG(I)	*4.6 ± 0.2*
POPS	POPS	*4.3± 0.2*
	POPS(I)	*4.5± 0.1*

PMPE	Top6	*4.4 ± 0.1*
QMPE		*4.7 ± 0.3*
YMPE		*4.6 ± 0.2*
PMPG		*4.7 ± 0.2*
PYPG		*4.7 ± 0.3*

The list of donors and acceptors from each membrane model is presented in Table 6 below and atoms are labeled in Figure 4.11. The list of donor and acceptor atoms is editable in the code released with this manuscript.

Table 6: Donors and acceptors for H-bonding calculations as presented for each membrane model used in our analyses from [187], a study presented in this thesis.

System	**Donors**	**Acceptors**
POPE	N *(HN1, HN2, HN3)*	O11, O12, O13, O14, O21, O22, O31, O32
3:1 POPE:POPG	N *(HN1, HN2, HN3)*, OC2 *(HO2),* OC3 *(HO3)*	O11, O12, O13, O14, O21, O22, O31, O32
5:1 POPE:POPG		
3:1 POPE:POPG (I)		
5:1 POPE:POPG (I)		
Top6		
POPS	N *(HN1, HN2, HN3)*	O13A, O13B, O11, O12, O13, O14, O21, O22, O31, O32
POPS(I)		

We compared the distance-based H-bond criterion (2.5 Å between H and acceptor heavy atoms) and the distance & angle H-bond criterion (3.5 Å between heavy atoms and 60° H-bond angle). For direct H bonds between lipids for one leaflet of pure POPE lipid membrane, we report the average number of unique H bonds as calculated from 10 coordinate sets from the simulation in *NPT*. Each leaflet consists of 138 lipid molecules. Using the distance & angle criterion we calculated 101 ± 2 average H bonds, while the distance-only criterion resulted in 105 ± 3 H bonds on average. Results are very similar.

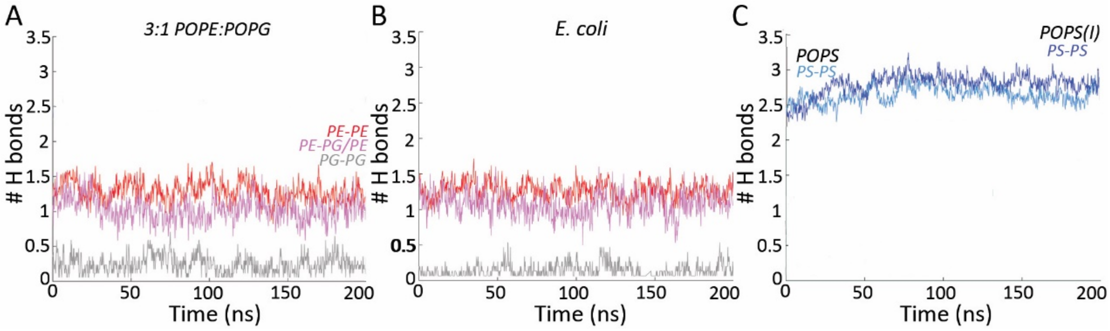


Figure 4.12: Time series of the average number of H bonds/lipid. Average number of lipid H bonds from simulations of the (A) 3:1 POPE: POPG bilayer, and (B) *E. coli* Top6 membrane. (A-B) Time series for the average number of POPE-POPE H bonds/POPE lipid, POPE-POPG H bonds per POPE lipid, and POPG-POPG H bonds per POPG lipid molecule are depicted as red, light red, and gray, respectively. (C) Time series of the average number of POPS-POPS H bonds per lipid in POPS membranes with and without salt with dark blue representing the first and light blue the latter [187]. MATLAB R2017b was used for plotting the results [186]. Image is from [187], a study presented in this thesis.

4.2.3.3 H-bond cluster dynamics

The comparable numbers of H bonds per lipid mentioned above for POPE, POPE: POPG, and *E. coli* Top6 membranes agree with the average number of clusters mediated by direct H bonds among lipids. Approximately 24 clusters per leaflet were obtained for each of these membranes, with around 4-5 lipids per cluster (Table 7).

Temporary clusters made up of anionic POPG or PG-type lipids are observed in both the POPE: POPG and Top6 membranes, respectively (Figure 4.13). At any given time, up to four clusters in each bilayer leaflet can be visited, with an occurrence of 10-35 % (Figure 4.13). The incidence of POPG clusters is lower in 5:1 POPE:POPG membranes compared to 3:1 POPE:POPG membranes. As expected, given the ratio of POPE to POPG lipids, the former typically displays a single POPG cluster (Figure 4.13). The POPG and PE membranes each contain small clusters of anionic lipids, typically consisting of two to three lipids per cluster.

POPS lipids form H bonds with other POPS lipids in such a way that each lipid molecule has, on average, 2.7-2.8 POPS-POPS H bonds (Figure 4.12). Clusters containing POPS lipids are larger, around 6 per cluster, compared to those containing POPE and/or POPG. POPS cluster-forming lipids account for a larger fraction of lipids [250] (87-90%) compared to POPE or POPE:POPG (67-80%) (Table 7).

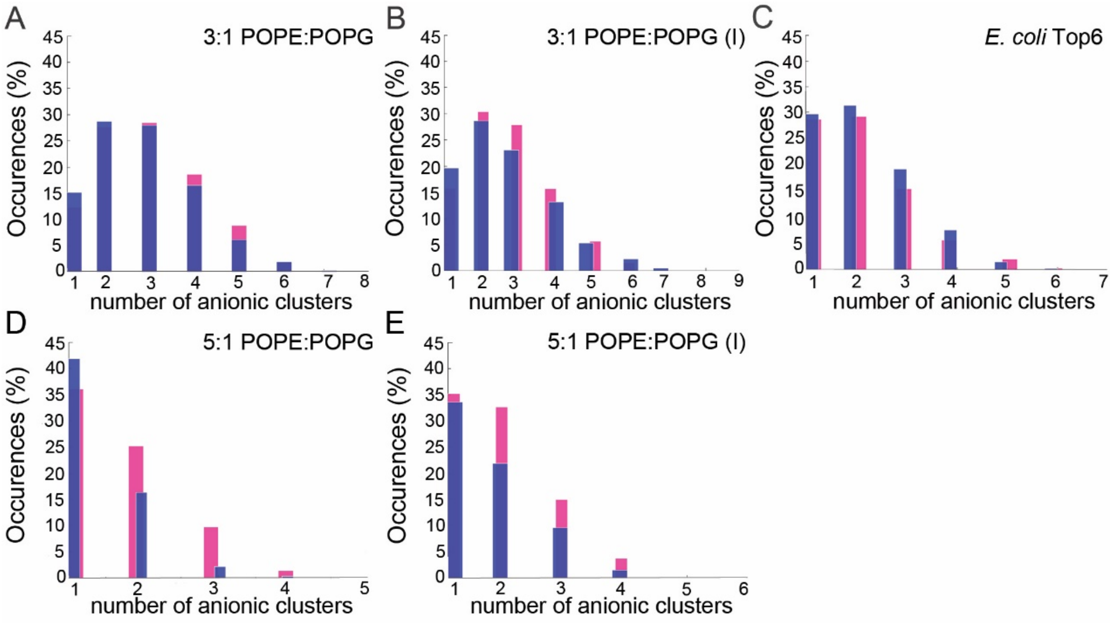


Figure 4.13: Histogram of the number of direct H-bonded PG clusters in (A) 3:1 POPE: POPG, (B) 3:1 POPE: POPG(I), (C) *E. coli* Top6, (D) 5:1 POPE: POPG, and (E) 5:1 POPE:POPG(I). The purple and pink bars represent average values estimated for the extracellular and the cytoplasmic side of the bilayers, respectively. Analyses are from the last 100 ns of the *NPT* simulations [187]. Image is from [187], a study presented in this thesis.

Table 7: Lipid H-bond clusters. We report the average number of clusters (ANC), the occurrence of ANC denoted ANCO, the average number of lipids within each cluster (ALC), and the percentage of average total lipids included in lipid clusters (ALEC) for each lipid bilayer. Table is from [187], a study presented in this thesis.

Clusters	Membrane	ANC	ANCO (%)	ALC	ALEC (%)
Direct H bonds	POPE	24± 3	12	5± 1.0	80
	3:1 POPE:POPG	24± 3	13	4± 1.0	76
	5:1 POPE:POPG	24± 3	11	5± 1.0	79
	3:1 POPE:POPG(I)	24± 3	12	4± 1.0	67
	5:1 POPE:POPG(I)	24± 3	12	5± 1.0	78
	POPS	19± 3	15	6± 1.0	88
	POPS(I)	20± 3	11	6± 1.0	90
	E. coli Top6	29± 3	11	4± 0.5	78
1-water bridges	POPE	27± 3	13	3± 0.2	52
	3:1 POPE:POPG	27± 3	14	3± 0.2	53
	5:1 POPE:POPG	27± 3	14	3± 0.2	52
	3:1 POPE:POPG(I)	25± 4	10	3± 0.2	48
	5:1 POPE:POPG(I)	27± 3	13	3± 0.2	53
	POPS	26± 3	14	3± 0.2	50
	POPS(I)	28± 3	14	3± 0.3	55
	E. coli Top6	29± 3	12	3± 0.2	49
Ion-mediated bridges	POPE	-	-	-	-
	3:1 POPE:POPG	2± 1	28	2± 0.7	3
	5:1 POPE:POPG	1± 1	40	1± 1	1
	3:1 POPE:POPG(I)	1± 1	36	1± 1	2

	5:1 POPE:POPG(I)	1± 1	34	1± 1	2
	POPS	12± 2	16	2± 0.2	20
	POPS(I)	9± 2	18	2± 0.1	14
	E. coli Top6	1± 1	32	1± 1	2

The sodium ions bridge pairs of lipids, and the typical cluster size for ion-mediated interactions is around 2. When compared to clusters with direct or one-water-mediated interactions, ion-mediated clusters are inclined to have rather higher occurrences (Table 7), since ions may stay in the proximity of lipids for longer [22].

4.2.3.4 Lipid H-bond topologies

Our algorithms were used to characterize the size and structure of sampled clusters as well as the frequency at which certain cluster types may be found in each of the membrane simulations that were conducted.

Direct H bonds may facilitate POPS lipid complex formation, while linear configurations are typically formed by one-water mediated bridges (Figure 4.15, Figure 4.14, Figure 4.19).

The number of direct H bonds can be determined by the length λ of a linear path of lipids that are bonded to each other via H bonds. The typical linear path has a value of λ = 1 (Figures 4.16A, 4.17A, Table 9). This suggests that a direct H bond is often formed by only two lipids. However, there are other pathways containing linear clusters of 3-4 lipids with a value of λ equal to 2 or 3. (Figures 4.16A, 4.17A, Tables 8, 9).

When mediated by direct H bonding between lipids, star and linear paths are sampled often (Table 8), and usually, they have γ = 4 (Figure 4.16B, Table 9). Circular paths with lengths greater than the minimum length σ = 3 are uncommon (Figure 4.16C, Table 9).

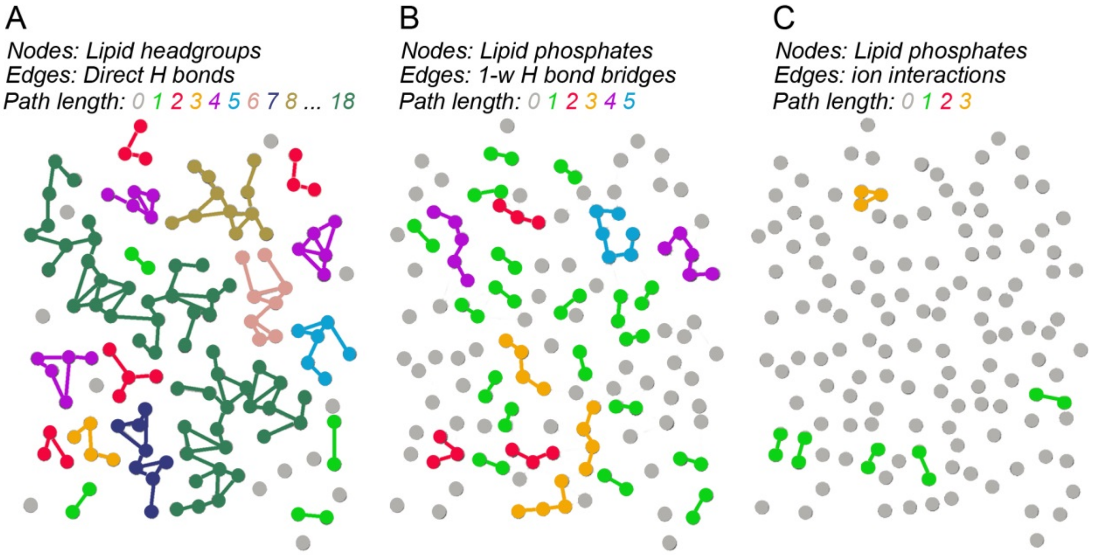


Figure 4.14: POPS lipid clusters. The graphics were created using the POPS simulation's last coordinate image. Nodes in the graphs stand for head groups or phosphate groups, while edges represent H bonds between lipids. (A) Direct H-bond paths of distinct lengths with colors based on the length value. (B) One-water-mediated paths found between lipid phosphates. (C) Ion-mediated paths are small, including only 2-3 lipids [187]. Image is from [187], a study presented in this thesis.

A combined linear, circular, and star graph was the most complicated type of topology found in our analysis. The preferred length of the complex graph mediated by direct lipid H bonds is $\xi = 5$ - 6 (Figure 4.16D).

Bridges formed by one water molecule between lipids create linear and star-shaped patterns. Optimal path lengths are $\lambda = 1$ and $\gamma = 3$-4, respectively (Figures 4.17A, B). Similarly, direct and one-water-mediated circular pathways have a preferable length of $\sigma = 3$ (Figure 4.17C, Figure 4.16C). Four bridges, which are mediated by water, are not sampled on circular pathways, unlike paths involving direct H bonds (Figures 4.16C, 4.17C). For lipids interconnected by one-water H-bond chains, we found for the combined linear, circular, and star graph, a smaller length on average, $\xi = 3$ (Figure 4.17C).

After obtaining the median path length value for each type of path from the simulation (Figure 4.16), the corresponding occupancy was calculated (Table 9, Figure 4.17). There is a considerable occupancy of small paths with $\lambda = 1$ for both direct and water-mediated interactions between lipids (Figures 4.17). The path length's occupancy at median value, for all types of paths, is lower for water mediated-bridges, as opposed to direct lipid H bonds (Figure 4.17). The reason for this is likely due to the fact that water-phosphate H bonds have short lifetimes.

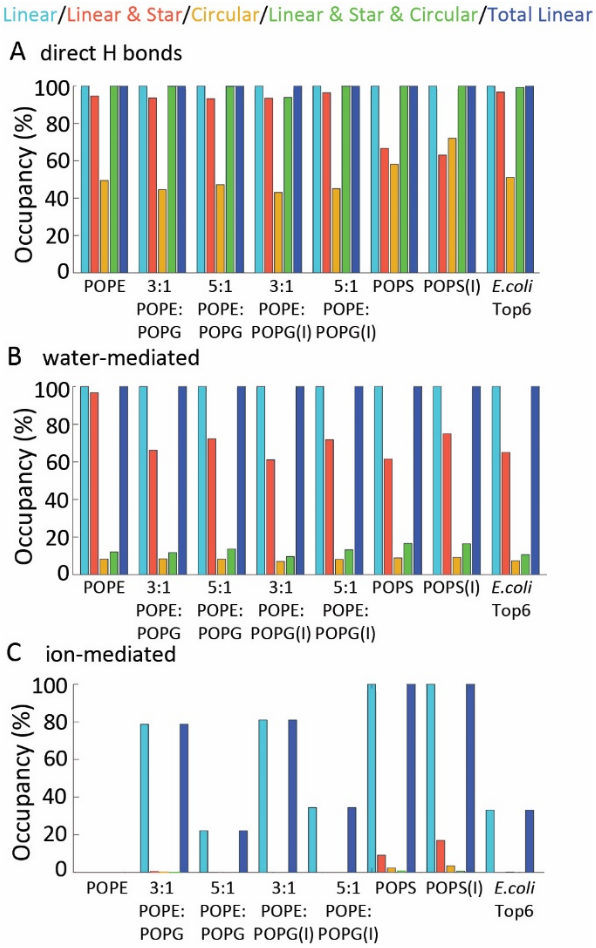


Figure 4.15: Paths' occupancies overview. (A) Overall occupancies of all paths mediated by direct H bonding between lipids. (B) Overall occupancies of all paths occupied by one-water bridges between lipid phosphates. (C) Overall occupancies of all paths mediated by lipid phosphate group ion bridging.[187]. MATLAB R2017b was used to create the plots [186]. Image is from [187], a study presented in this thesis.

We analyzed the H-bonded paths of POPS lipids in a POPS membrane (Figure 4.18), considering their strong tendency for H bonding even in complex and lengthy paths with $\xi = 6$ (Table 8, Figure 4.17D). One-water bridges form 15-25 linear pathways connecting two lipids ($\lambda = 1$ in Figures 4.18), at any given moment during the simulation. POPS have high occupancies (~53-71%) of circular paths with $\sigma = 3$ and of complex star & linear & circular with $\xi = 6$ (Table 9, Figure 4.17).

As the path length increases, the average number of water-bridged linear paths decreases. For instance, when $\lambda = 2$, there are approximately 5 to 10 linear paths, but only one path exists when $\lambda = 6$ (Figures 4.18A,B). During the whole simulation, circular path with $\sigma = 3$ is sampled not frequently, while a path with $\sigma = 4$ is only detected one time (Figure 4.18C).

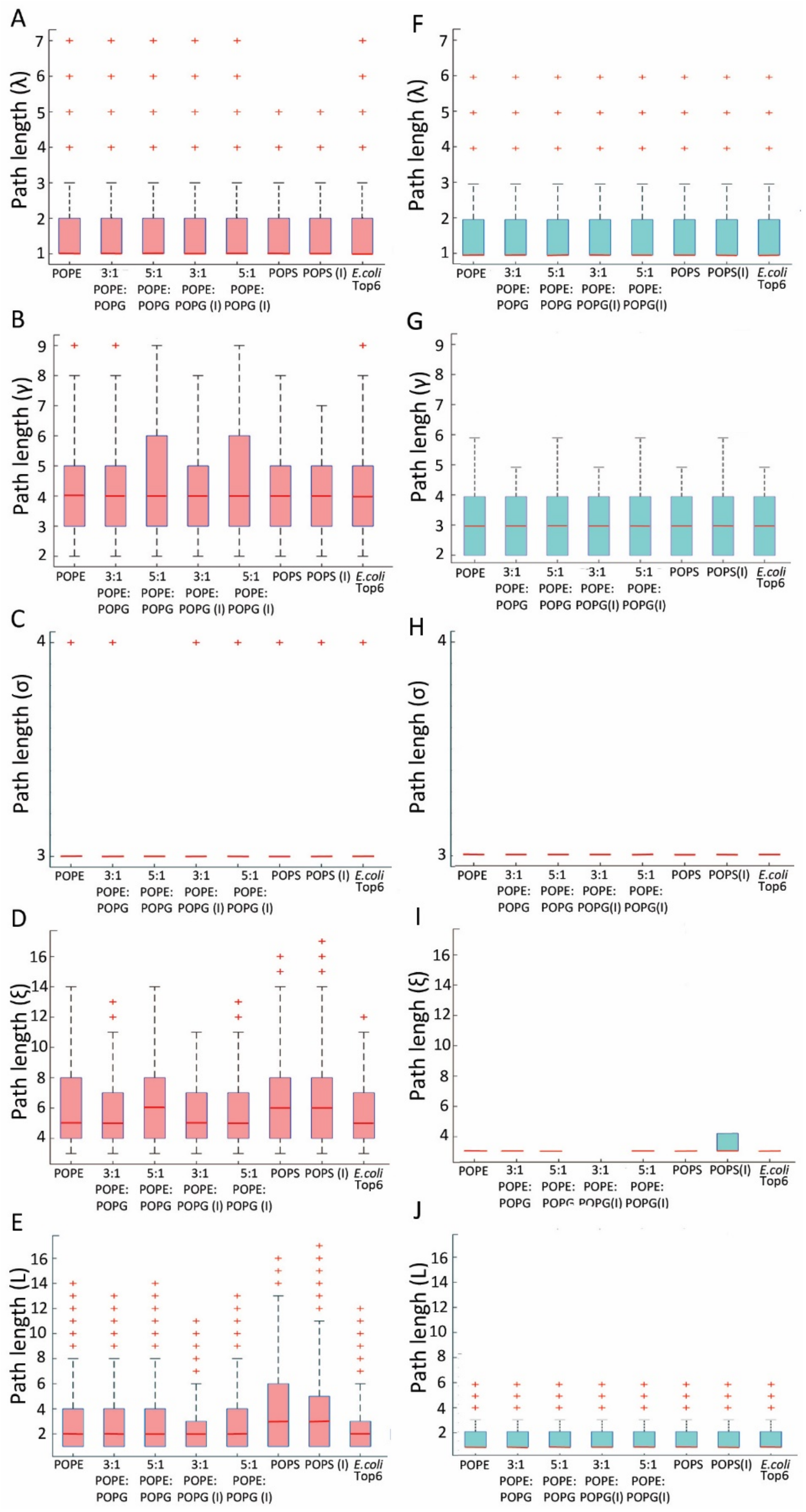


Figure 4.16: Clusters mediated by direct H bonds formed between lipids or by short one-water bridges. (A-E) Path length distribution for directly H-bonded lipids that sample linear paths (panel A), star & linear paths (panel B), circular paths (panel C), star & circular & linear paths (panel D), and overall linear paths (panel E). (F-I) Path length distribution for one-water mediated bridges formed between lipids, for linear paths (panel F), star and linear paths (panel

G), circular (panel H), star and linear and circular (panel I), as well as any linear path, either of a separate linear path or a linear path section of star and linear or star and circular and linear paths (panel J). MATLAB R2017b [186] was used to create box plots. The central sign, on each box, represents the median, whereas outliers are represented using the '+' character [187]. Image is from [187], a study presented in this thesis.

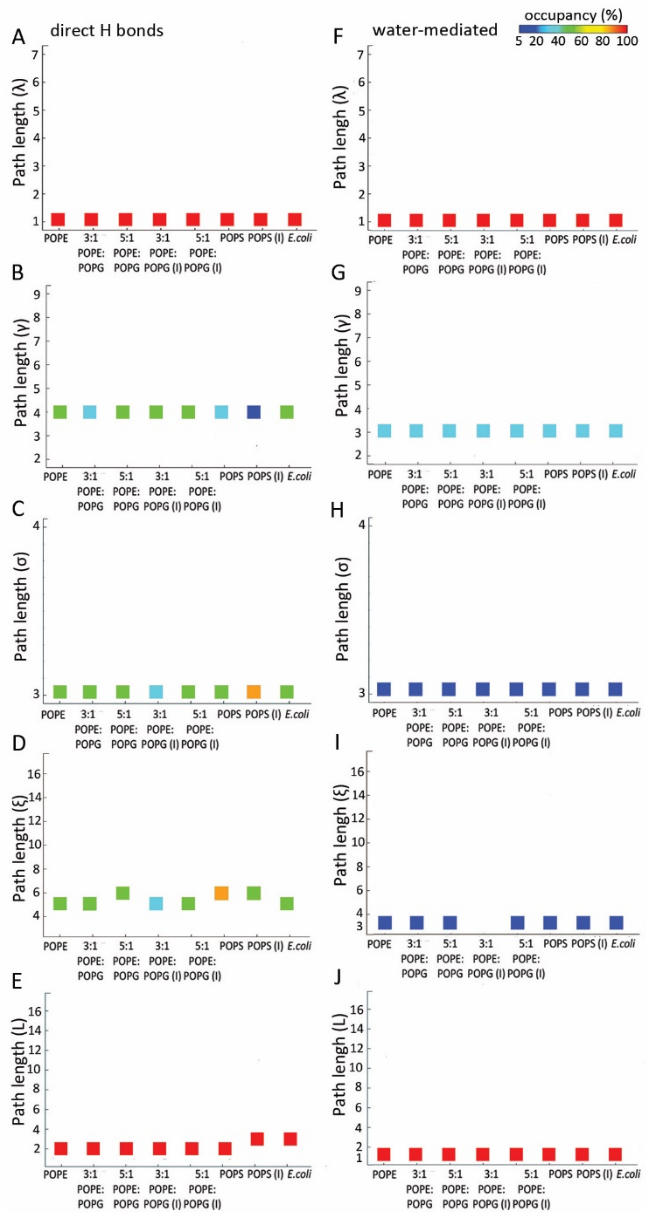


Figure 4.17: Median path length occupancies for direct and one-water mediated lipid clusters. (A-E) The median value of the occupancy for lipid paths that contain solely direct H bonds between lipids, for linear paths with lengths λ (panel A), for star and linear paths with length γ (panel B), for circular paths with length σ (panel C), for star, circular and linear paths with length ξ (panel D), and for all linear path sections with length L (panel D). (F-J) Median value of the occupancy for lipid paths that bridge two lipids via one water molecule, for linear paths with path lengths λ (panel F), star and linear with path length γ (panel G), circular paths with

length σ (panel H), star &and circular and linear with length ξ (panel I), and all linear path sections with length L (panel J) [187]. MATLAB R2017b [186] was used to create the plots. Image is from [187], a study presented in this thesis.

Table 8: Overview of H-bonded paths detected by the DFS method. We provide the occupancy of each type of lipid path type calculated for the final 100ns. Table is from [187], a study presented in this thesis.

Membrane	Total occurrence of H-bonded paths (%)				
	Linear	Star & Linear	Circular	Star & Circular & Linear	Total linear
Paths with direct lipid H bonds					
POPE	100	94.7	49.5	100	100
3:1 POPE:POPG		93.7	44.5	99.9	
5:1 POPE:POPG		93.3	47.2	99.9	
3:1 POPE:POPG(I)		93.5	43.1	94.0	
5:1 POPE:POPG(I)		96.5	45.1	99.9	
POPS		66.7	58.1	100	
POPS(I)		63.1	72.1	100	
E. coli Top 6		96.9	51.1	99.4	
Paths mediated by one-water bridges between two lipids					
POPE	100	69.8	8.3	12.1	100
3:1 POPE:POPG		66.2	8.4	11.7	
5:1 POPE:POPG		72.3	8.3	13.6	
3:1 POPE:POPG(I)		61.1	7.0	9.6	
5:1 POPE:POPG(I)		71.8	8.1	13.2	
POPS		61.5	8.9	16.6	
POPS(I)		74.9	9.2	16.4	
E. coli Top 6		65.0	7.4	10.6	
Paths with ion-mediated bridges between lipids					
POPE	-	-	-	-	-
3:1 POPE:POPG	78.7	<5	<5	<5	78.7
5:1 POPE:POPG	22.1	<5	-	-	22.1
3:1 POPE:POPG(I)	81.0	-	-	-	81.0
5:1 POPE:POPG(I)	34.4	-	-	-	34.4
POPS	100	9.1	<5	<5	100
POPS(I)	100	17.0	<5	<5	100
E. coli Top 6	33.1	-	<5	-	33.1

Table 9: Overview of the paths detected in membrane simulations. We provide the occupancy of each lipid topology type with a length identical with the median value indicated in the plots of Figure 4.17. In the case of circular graphs, we additionally provide the occurrence of paths with σ= 4, when sampled. The paths with occupancies less than 5% are excluded. Table is from [187], a study presented in this thesis.

Membrane	Median value of path length/ Occurence of medium-value path length (%)				
	Linear, λ	Star & Linear, γ	Circular, σ	Star & Circular & Linear, ξ	Total linear, L = λ+ γ + ξ

Paths mediated by direct lipid H bonding					
POPE	1 / 100	4 / 40.6	3 / 46.0 4 / 6.4	5 / 53.0	2 / 99.8
3:1 POPE:POPG		4 / 39.0	3 / 40.7 4 / 5.1	5 / 47.9	2 / 99.2
5:1 POPE:POPG		4 / 40.2	3 / 44.4	6 / 42.6	2 / 99.4
3:1 POPE:POPG(I)		4 / 41.1	3 / 38.9 4 / 5.8	5 / 36.6	2 / 99.2
5:1 POPE:POPG(I)		4 / 43.3	3 / 40.7 4 / 6.6	5 / 55.6	2 / 99.1
POPS	1 / 98.8	4 / 20.4	3 / 53.9 4 / 12.1	6 / 64.4	3 / 92.5
POPS(I)	1 / 99.3	4 / 19.4	3 / 71.0 4 / 5.9	6 / 59.1	3 / 94.2
E. coli Top 6	1 / 100	4 / 44.6	3 / 47.3 4 / 6.4	5 / 46.5	2 / 99.8
Paths mediated by one-water bridges between lipid phosphate groups					
POPE	1 / 100	3 / 29.0	3 / 6.0	3 / 5.0	1 / 100
3:1 POPE:POPG		3 / 26.5	3 / 6.8	3 / 5.0	1 / 100
5:1 POPE:POPG		3 / 28.6	3 / 6.3	3 / 5.0	1 / 100
3:1 POPE:POPG(I)		3 / 23.9	3 / 5.4	3 / <5	1 / 100
5:1 POPE:POPG(I)		3 / 29.6	3 / 6.1	3 / 5.0	1 / 100
POPS		3 / 21.1	3 / 8.1	3 / 9.3	1 / 100
POPS(I)		3 / 34.4	3 / 8.0	3 / 6.4	1 / 100
E. coli Top 6		3 / 27.0	3 / 6.0	3 / 5.0	1 / 100
Paths mediated by ions bridging lipid phosphate groups					
POPE	-	-	-	-	-
3:1 POPE:POPG	1 / 72.4	-	-	-	1 / 72.4
5:1 POPE:POPG	1 / 20.5	-	-	-	
3:1 POPE:POPG(I)	1 / 48.2	-	-	-	
5:1 POPE:POPG(I)	1 / 33.9	-	-	-	
POPS	1 / 100	2 / 7.4	-	-	1 / 100
POPS(I)	1 / 100	2 / 15			1 / 100
E. coli Top 6	1/ 32.1	-	-	-	1 / 32.1

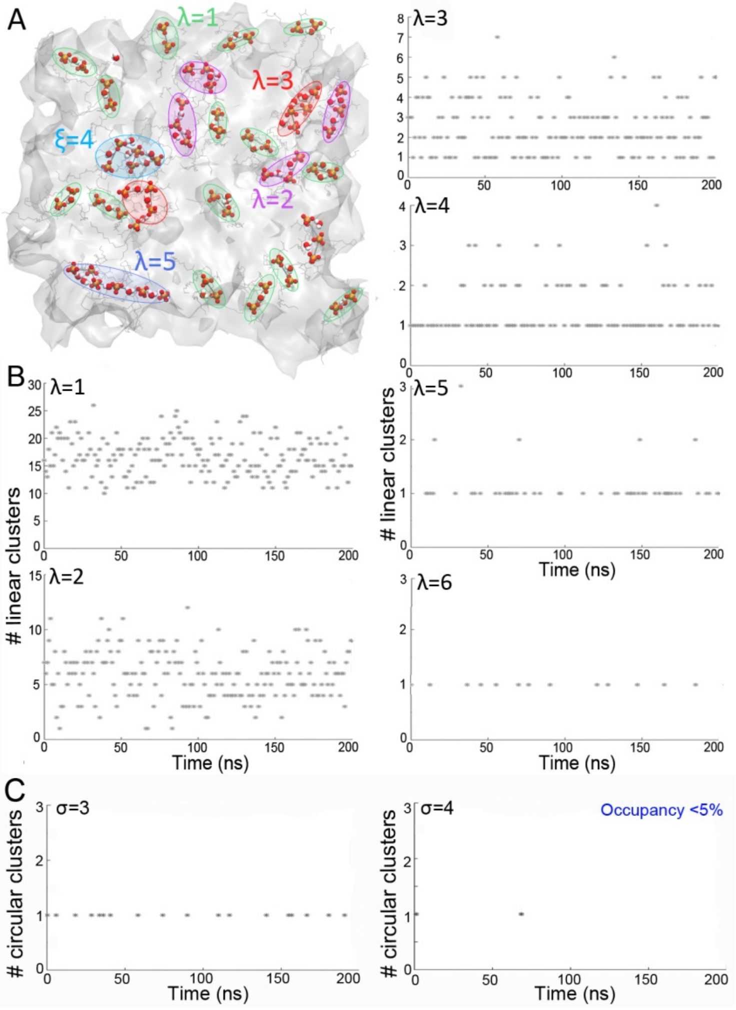


Figure 4.18: One-water mediated bridges in the POPS membrane. (A) A coordinate image of water-mediated paths sampled in one of the membrane leaflets. To avoid confusion, lipid phosphate groups are displayed as orange spheres, while water molecules as van der Waals spheres. Paths are color-coded based on their length. (B) Time series of the number of paths, with lengths ranging from 0 to 6. (C) Time series showing the number of paths with σ=3 and σ= 4 throughout the simulation [187]. We used VMD [23] for the molecular graphics and Matlab [186]for the plots. Image is from [187], a study presented in this thesis.

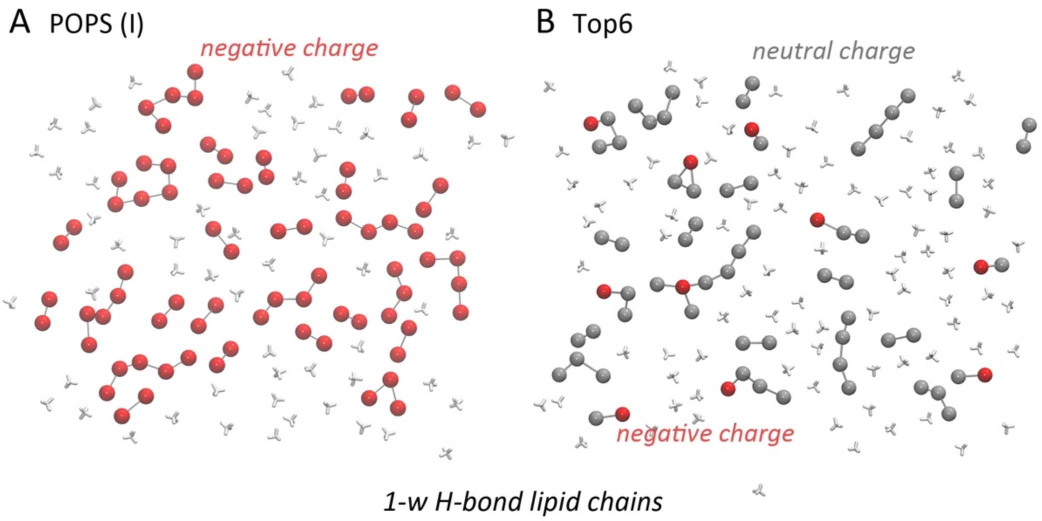


Figure 4.19: Topology illustration based on the last coordinate snapshot from the A) POPS (I) and B) Top6 showing one-water mediated H-bond pathways in the extracellular side of the bilayer. In the graph, nodes represent phosphate groups and edges H bonds that interconnect lipid molecules. Neutral and negatively charged lipids are illustrated as silver and red graphic spheres, respectively. Molecular graphics were generated with VMD [23]. Image is generated for the scope of this thesis.

4.2.3.5 DFS convergence tests for H-bond lipid clusters

To determine if our cluster analyses are dependent on the trajectory length utilized for calculations, we performed separate calculations for H-bond clusters for the complete 200ns vs. the first 100ns and compared them with our results presented in the previous sections from the last 100ns.

In the following Table 10, the comparison is done for all membrane models for the calculation of ANC, ANCO (%), ALC, ALEC (%). We report very close values for the H-bond clusters, direct or water-mediated, and ion bridges, based on the three parts of the trajectories (Table 10). Based on our extended tests, we suggest that our algorithm gives converged results on the simulation timescale we use for our calculations.

Table 10: Covergence tests for H-bond cluster computations. Table is from [187], a study presented in this thesis.

Clusters	**Membrane**	**ANC**	**ANCO (%)**	**ALC**	**ALEC (%)**
Direct H bonds	POPE (last 100ns)	24± 3	12	5± 1.0	80
	POPE (first 100ns)	23 ± 3	12	5± 1.0	80

	POPE (complete 200ns)	23 ± 3	12	5± 1.0	80
	3:1 POPE:POPG (last 100ns)	24± 3	13	4± 1.0	76
	3:1 POPE:POPG (first 100ns)	24± 3	12	5± 1.0	78
	3:1 POPE:POPG (complete 200ns)	24± 3	13	4± 1.0	77
	5:1 POPE:POPG (last 100ns)	24± 3	11	5± 1.0	79
	5:1 POPE:POPG (first 100ns)	24± 3	12	5± 1.0	79
	5:1 POPE:POPG (complete 200ns)	24± 3	12	5± 1.0	79
	3:1 POPE:POPG(I) (last 100ns)	24± 3	12	4± 1.0	67
	3:1 POPE:POPG(I) (first 100ns)	25± 3	12	4± 1.0	76
	3:1 POPE:POPG(I) (complete 200ns)	24± 3	12	4± 1.0	72
	5:1 POPE:POPG(I) (last 100ns)	24± 3	12	5± 1.0	78
	5:1 POPE:POPG(I) (first 100ns)	24± 3	11	5± 1.0	80
	5:1 POPE:POPG(I) (complete 200ns)	24± 3	11	5± 1.0	78
	POPS (last 100ns)	19± 3	15	6± 1.0	88
	POPS (first 100ns)	17± 3	14	7± 1.0	87
	POPS (complete 200ns)	18± 3	14	7± 1.0	88
	POPS(I) (last 100ns)	20± 3	11	6± 1.0	90
	POPS(I) (first 100ns)	19± 3	15	6± 1.0	88
	POPS(I) (complete 200ns)	19± 3	13	6± 1.0	89
	E. coli Top6 (last 100ns)	29± 3	11	4± 0.5	**78**
	E. coli Top6 (first 100ns)	29± 3	11	4± 0.5	79
	E. coli Top6 (complete 200ns)	29± 3	11	4± 0.5	79
1-water bridges	POPE (last 100ns)	27± 3	13	3± 0.2	52
	POPE (first 100ns)	28± 3	14	3± 0.2	55
	POPE (complete 200ns)	27± 3	13	3± 0.2	53
	3:1 POPE:POPG (last 100ns)	27± 3	14	3± 0.2	53
	3:1 POPE:POPG (first 100ns)	28± 3	13	3± 0.2	54
	3:1 POPE:POPG (complete 200ns)	27± 3	13	3± 0.2	53
	5:1 POPE:POPG (last 100ns)	27± 3	14	3± 0.2	52

	5:1 POPE:POPG (first 100ns)	28± 3	13	3± 0.2	55
	5:1 POPE:POPG (complete 200ns)	28± 3	13	3± 0.2	53
	3:1 POPE:POPG(I) (last 100ns)	25± 4	10	3± 0.2	48
	3:1 POPE:POPG(I) (first 100ns)	27± 3	14	3± 0.2	54
	3:1 POPE:POPG(I) (complete 200ns)	26± 4	11	3± 0.2	51
	5:1 POPE:POPG(I) (last 100ns)	27± 3	13	3± 0.2	53
	5:1 POPE:POPG(I) (first 100ns)	28± 3	13	3± 0.2	55
	5:1 POPE:POPG(I) (complete 200ns)	28± 3	13	3± 0.2	54
	POPS (last 100ns)	26± 3	14	3± 0.2	50
	POPS (first 100ns)	27± 3	14	3± 0.2	52
	POPS (complete 200ns)	26± 3	14	3± 0.2	51
	POPS(I) (last 100ns)	28± 3	14	3± 0.3	55
	POPS(I) (first 100ns)	27± 3	13	3± 0.2	55
	POPS(I) (complete 200ns)	27± 3	13	3± 0.2	55
	E. coli Top6 (last 100ns)	29± 3	12	3± 0.2	49
	E. coli Top6 (first 100ns)	31± 3	12	3± 0.2	54
	E. coli Top6 (complete 200ns)	30± 3	12	3± 0.2	53
Ion-mediated bridges	POPE	-	-	-	-
	3:1 POPE:POPG (last 100ns)	2± 1	28	2± 0.7	3
	3:1 POPE:POPG (first 100ns)	2± 1	29	2± 0.7	3
	3:1 POPE:POPG (complete 200ns)	2± 1	28	2± 0.7	3
	5:1 POPE:POPG (last 100ns)	1± 1	40	1± 1	1
	5:1 POPE:POPG (first 100ns)	1± 1	39	1± 1	1
	5:1 POPE:POPG (complete 200ns)	1± 1	39	1± 1	1
	3:1 POPE:POPG(I) (last 100ns)	1± 1	36	1± 1	2
	3:1 POPE:POPG(I) (first 100ns)	1± 1	34	2± 1	2
	3:1 POPE:POPG(I) (complete 200ns)	1± 1	35	2± 1	2

	5:1 POPE:POPG(I) (last 100ns)	1± 1	34	1± 1	2
	5:1 POPE:POPG(I) (first 100ns)	1± 1	41	1± 1	2
	5:1 POPE:POPG(I) (complete 200ns)	1± 1	38	1± 1	2
	POPS (last 100ns)	12± 2	16	2± 0.2	20
	POPS (first 100ns)	11 ± 3	15	2± 0.2	19
	POPS (complete 200ns)	12 ± 3	15	2± 0.2	19
	POPS(I) (last 100ns)	9± 2	18	2± 0.1	14
	POPS(I) (first 100ns)	9± 2	17	2± 0.1	15
	POPS(I) (complete 200ns)	9± 2	18	2± 0.1	14
	E. coli Top6 (last 100ns)	1± 1	32	1± 1	2
	E. coli Top6 (first 100ns)	1± 1	38	1± 1	2
	E. coli Top6 (complete 200ns)	1± 1	35	1± 1	2

4.2.4 Summary

We implemented an algorithm to detect H-bond lipid clusters, describe their dynamics, and identify the geometrical topological arrangement of lipids in H-bonded paths via DFS searches. The method was used to detect clusters sampled in POPE, three distinct *E. coli* membrane models - the 3:1 and 5:1 POPE: POPG models, POPS, as well as the Top6 *E. coli* model. Separate simulations of lipid bilayers were conducted with a total sampling duration of 1.6μs to analyze H-bond clusters in solution.

Lipids can interact with each other through ion bridges, direct, or one water H bonding. The application of the algorithm can achieve the identification of four types of graphs: linear, star, circular, as well as their combination (Figure 3.8). Our algorithm detects the clusters per bilayer leaflet, provides ths ids of the lipids and waters per simulation time, detects the topology, the occupancy, and the longest linear length of each cluster. Our implementation allows the graphical representation of the H-bond network, the clusters and the topologies distinguished by their type or length. We show convergence on our results compared outcomes from different parts of the simulation and efficiency in computing time.

We demonstrated that all membranes contain small linear pathways that connect two lipids through H bonds or one water molecule. The star & linear graph of five lipids, when facilitated by water bridges, or four lipids, when the lipids themselves establish

H bonds with one another, is the second most common lipid topology. When three or four lipids H bond directly to one another, circular graphs are visited quite frequently. Nevertheless, circular routes mediated by water molecules are uncommon, probably due to the limited lifespan of H bonds between lipids and water. More intricate path topologies, in which a minimum of four lipids connect in a hybrid circular, star, and linear graph, are infrequently visited (Figure 4.16, Figure 4.17).

Several research studies have been conducted on PS lipids, as they are an essential component of eukaryotic cell membranes [31, 219, 251-253]. PS lipids' attractive interactions, such as H bonding, could be the reason for smaller surface areas in the absence of salt compared to PC [37]. During simulations, networks of direct H bonds were observed between POPS lipids. This suggests that sections of POPS lipids connected by serine interactions might function as potential binding sites for external partners. In our study, we showed that POPS lipids have a strong tendency to H bond to one another.

For the *E. coli* membrane, 3:1 POPE: POPG and 5:1 POPE: POPG constitute the principal models [204, 205], with the latest and most precise being the Top6 involving three types of PE and PG lipids each [205]. Top6 membrane is approximately 3 Å thinner than POPE: POPG membranes (Table 3), implying that it is a better model for the *E. coli* membrane. The H-bond clusters sampled in the *E. coli* Top6 and POPE: POPG membrane are similar, implying that simulations, for example, events of biomolecule interactions to lipid membranes, will provide similar outcomes.

The length of the simulation, the force field applied to all molecules in our system, and the size of the membrane patch may impact the H-bond clusters. Our algorithm is generic and based on the above modifications will provide results accordingly. However, for the 200ns of the simulations we performed, our tests of computing H-bond clusters from different time lengths of each simulation revealed very similar results for all membrane models and showed that our algorithm provides converged results. We reported the efficiency of our algorithm, less than one second is needed to detect the H-bond clusters for 10 coordinate snapshots. This will enable studies in the future to efficiently report results from larger membrane patches or prolonged simulations.

Our algorithm can be applied to a wide range of biological models such as lipid bilayers of various compositions and proteins. I contributed to the study by [254] as a co-author to the implementation of graph-based computational tools to compute and visualize H-bond lipid clusters and characterize their topologies during MD simulations. The study presents an application of our graph algorithm for a model bilayer consisting of POPS and cholesterol with a concentration of 10%. Cholesterol tends to form dimers [255], typically two molecules in a cluster. As we previously showed in our study [187], POPS samples dynamic direct, water-mediated and ion-mediated H-bond clusters with a linear topology more frequently than more complex topologies. However, complex and extended networks of direct H bonds and water-mediated bridges between POPS headgroups are sampled more frequently in POPS:POPS than in POPS:cholesterol clusters [254]. It should be noted, though, that the length of the MD simulations as well as the size of the lipid membrane patch may have an impact on the number and the size of the clusters sampled.

Our implementation could be used as a guide for experimental studies relevant to lipid H-bond clustering, such as protein-membrane and lipid-mediated protein interactions [256-258], or lateral proton transfer events in membranes [259-261].

4.3 Role of water and dynamic H-bond networks in the reaction coordinate of a molecular motor

The study presented in this chapter is available in the following publications:

Karathanou, K. and Bondar, A.N., 2019. Using graphs of dynamic hydrogen-bond networks to dissect conformational coupling in a protein motor. Journal of chemical information and modeling, 59(5), pp.1882-1896.

Krishnamurthy, S., Eleftheriadis, N., Karathanou, K., Smit, J.H., Portaliou, A.G., Chatzi, K.E., Karamanou, S., Bondar, A.N., Gouridis, G. and Economou, A., 2021. A nexus of intrinsic dynamics underlies translocase priming. Structure. 29(8):846-858.e7. doi: 10.1016/j.str.2021.03.015

Krishnamurthy, S., Sardis, M.F., Eleftheriadis, N., Chatzi, K.E., Smit, J.H., Karathanou, K., Gouridis, G., Portaliou, A.G., Bondar, A.N., Karamanou, S. and Economou, A., 2022. Preproteins couple the intrinsic dynamics of SecA to its ATPase cycle to translocate via a catch and release mechanism. Cell Reports, 38(6), p.110346.

4.3.1 Introduction

Supporting the secretion process, SecA protein transports secretory proteins through the SecYEG translocon using energy from ATP hydrolysis (Section 1.1.4- Protein translocation in bacteria, Section 1.1.4.1– SecA protein). It belongs to the DExD/H motif (Asp-Glu-x-Asp/His, where x denotes any amino acid residue) family [262-265]. There are distinct functional domains (NBD1, NBD2, HSD, HWD, PBD – Section 1.1.4.1) of the protein, and we show in our study that H bonds between domains create paths of connected amino acids participating in long-distance allosteric coupling (Figure 1.13) [79, 188, 266-268]. We implemented algorithms to visualize the H-bond networks of proteins and investigate SecA's response to mutations or changes in nucleotide-binding state [188].

Crystal structures of SecA can be obtained by different bacteria such as *Bacillus subtilis* [87, 89, 269], *Escherichia coli* [90], Mycobacterium *tuberculosis* [91], *Thermotoga maritima* [93, 270], and *Thermus thermophilus* [271] and describe the protein in its apo or ADP-bound states. The study by [90] provides coordinates for the ATP but not for the magnesium ion and regions of PBD. The study by [267] solves the structure of SecA with NMR experiments and provides the PBD and a signal peptide, but not the nucleotide coordinates.

In our study, we model *B. subtilis* ADP-bound SecA and we study four mutations near the nucleotide-binding site. We perform MD simulations for the mutations T107N, E208Q, R489K, and R517A as they are identified to influence the function of SecA. More specifically, the *B. subtilis* T107 increases the ATPase activity compared to the wild type [272], while the *B. subtilis* E208 and R489 inhibit the ATP activity [272]. E208Q prevents the coupling between SecA and the SecY translocon [273], and the R517A mutation inhibits the translocation of the pre-protein [274]. There is a salt bridge between D215 and R517 in *B. subtilis* SecA which is denoted as Gate1, and it is vital for the coupling between the PBD and the NBD region where the nucleotide binds [274].

In our study, we performed MD simulations of the wild-type and mutations near the NBD region. We implemented algorithms to map the H-bonding connections on the protein structures and detect long-distance pathways that connect different regions of the protein. We used graph theory to characterize the nodes (amino acids) and their edges (H bonding) by their importance in the graph and their local clustering. We showed that changes in the nucleotide-binding site were associated with changes in H-bond dynamics in the PBD region. We found that the role of water is essential to the coupling between distant functional domains contributing to inter-domain H bonds with high occupancies during the simulations.

At the end of this chapter, we present atomistic MD simulations and H-bond graphs of the *E. coli* SecA monomer, as well as two independent simulations of SecA dimers from our studies [99, 100].

4.3.2 Coordinates set-up & MD simulations

We modeled the *B. subtilis* ADP-bound SecA, PDB: 1M74 [87]. We prepared the coordinates for the protein (amino acids M1 to G802), the ADP, waters within 6 Å of the protein surface, and sodium ions for charge neutrality with CHARMM [208]. Our system included a total of 277,124 atoms (Table 11). We also modeled the ADP-bound SecA, PDB: 1TF2 [269] with a magnesium ion, waters, and sodium ions using CHARMM. Protein consists of amino acids from M1 to I780. Our system included a total of 196,703 atoms (Table 11).

We included four mutations (T107N, E208Q, R489K, and R517A) (Table 11) starting from the wild-type and replacing amino acids. For E208 we replaced the Oε1 or Oε2 with -NH2 concluding in two simulations, the E208Q1, and E208Q2, respectively.

To model the *B. subtilis* ATP-bound SecA, we used the crystal structure 2FSG [90] and overlap it to our 1M74 [87] and 1TF2 [269] replacing ADP with ATP and keeping the magnesium ion from the previous systems. The overlap was performed using Coot [275]. The reason for not using directly the 2FSG [90] is because lacks coordinates for most of the PBD. Significant geometric collisions were not observed between the ATP and the protein coordinates.

We used the CHARMM force field [276-279] and TIP3P model to describe water molecules [211]. We performed our simulations using NAMD [214, 215]. The MD simulation protocol is the same as described in Section 4.1.2 and Section 4.2.2. Each production run in *NPT* was prolonged to 200ns saving coordinates every 10ps. All average values were calculated from the final 100ns of each *NPT* simulation unless stated otherwise. *NVE* simulations of 1ns, saving coordinates every 10fs, were performed for each system to investigate the dynamics of water. As a reference simulation we used Sim1 (Table 11) beginning from the crystal structure from [87].

We took a coordinate snapshot from [95] NMR structure to examine the SecA monomer dynamics. By attaching the SecA monomer's structure separately onto dimers of *B. subtilis* and *M. tuberculosis* SecA (PDB ID: 1M6N and 1NL3_1, respectively), two of the dimeric states suggested as physiologically related [280, 281], respectively, ecSecA1M6N and ecSecA1NL3_1 were created. The total number of atoms in the simulations for the SecA monomers and dimers is 345.330 (2VDA), 635.979 (1M6N), and 666.629 (1NL3).

To better understand the energetics of clamp motions we performed a 325 ns MD simulation in monomeric *E.coli* SecA starting with the the Open clamp and 200ns of the two dimeric SecA systems. The simulation protocol we followed is described above.

Table 11: Performed *NPT* simulations in our study [187]. The prime sign indicates the 1ns *NVE* simulations started from each *NPT* simulation. Table is from [102], a study presented in this thesis.

Sim	Protein		Nucleotide binding pocket	Length (ns)	Rmsd (Å)
	Structure	Type			
SecA					
Sim1, 1'	1M74	Wild type	ADP, Mg^{2+}	203.8	1.46
Sim2, 2'		T107N		203.6	1.71
Sim3, 3'		$E208Q_1$		204.5	1.67
Sim4, 4'		$E208Q_2$		200.6	1.51
Sim5, 5'		R489K		199.6	1.53
Sim6, 6'		R517A		203.9	1.64
Sim7, 7'		Wild type	ATP, Mg^{2+}	203.8	1.33
Sim8	1TF2	Wild type	ADP, Mg^{2+}	176.0	1.00
Sim9	1TF2	Wild type	ATP, Mg^{2+}	222.0	1.01

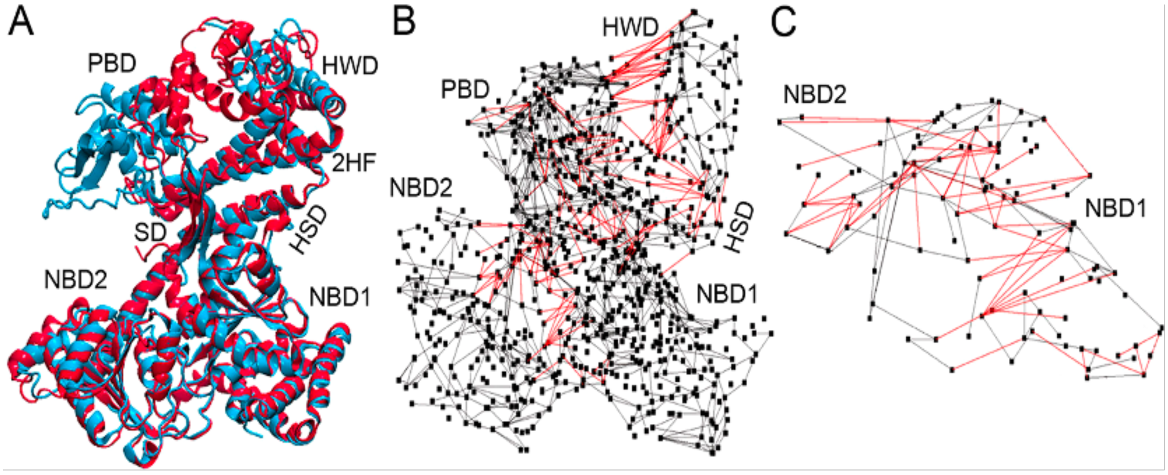


Figure 4.20: SecA structure and dynamics. (A) Overlap of ADP-bound *B.subtilis* SecA crystal structures with PDB id: 1M74 (red) [Sim1] and PDB id: 1TF2 (cyan) [Sim8]. The different orientation of the PBD region is shown in the two structures. (B) Inter- and intra- domain H bonds are shown as red and gray edges, respectively. (C) Network of H bonds between NBD1 and NBD2. Graphics are from [102]. Image is from [102], a study presented in this thesis.

4.3.3 Results & Discussion

4.3.3.1 RMSD profiles

We computed the Cα root-mean-square distances (RMSD) from the last 50ns of Sim1 to Sim10 using the structured regions of NBD1, NBD2, HSD and more specifically the SF and 2HF, structured PBD and HWD. The time series of RMSDs are shown in (Figure 4.20- Figure 4.22)[102]. Cα RMSDs reveal high stability of SecA structures in all simulations. We observe relatively low (<2Å) values for the structured domains and greater structural alterations for the loop and termini regions (<3Å). The NBDs exhibit low values, suggesting that these domains resemble the initial crystal structures (Figure 4.20, Figure 4.21) [102].

4.3.3.2 H bonds between SecA functional domains

Computations of H bonds (algorithm illustrated in Figure 3.13) showed a rich network of interactions across the whole protein (Figure 4.20B). On average, we computed ~200 intra-domain and ~60 inter-domain H bonds during Sim1. NBD1 possesses many intra-domain H bonds, around 20% having occupancies of at least 75% in ADP-bound wild-type SecA (Figure 4.23). The fact that NBD1 groups form multiple H bonds (Figure 4.23) is compatible with NBD1 possessing a structure that is stable as shown in RMSD analysis. On the contrary, PBD is dynamic as forms fewer intra-domain H bonds and as a result, it is a more flexible region (Figure 4.23).

Obtaining inter-domain Hbonds of NBDs with HSD, we conclude interactions of high occupancies (Figures 4.24). The PBD has rather limited stable inter-domain H bonds (Figure 4.25) for instance the PBD-HWD and PBD-HSD H bonds. We speculate that those interactions assist in maintaining the PBD's orientation throughout Sim1 with respect to the HWD.

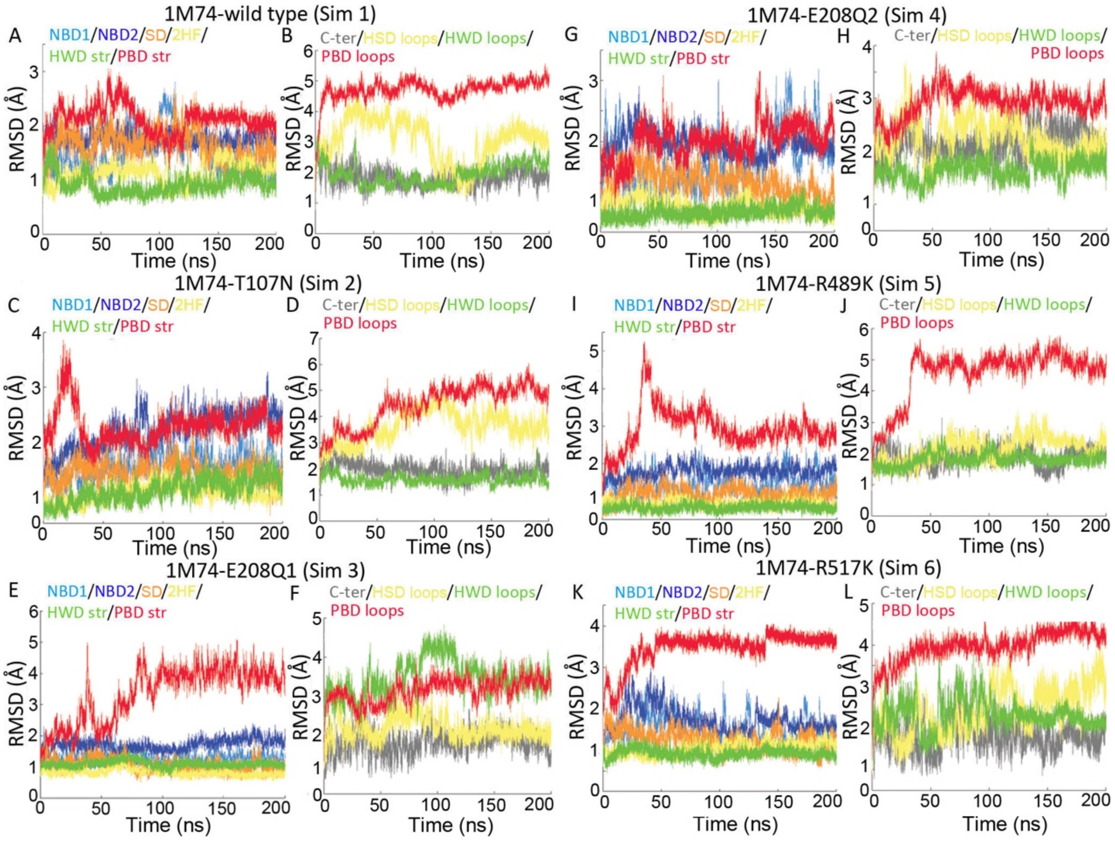


Figure 4.21: Cα RMSD profiles for SecA computed from Sim1-Sim6 (Table 11). We show the structured regions of SecA and the unstructured loops and termini for each Sim in *NPT*. Results are shown for the entire length of each simulation [187]. Image is from [102], a study presented in this thesis.

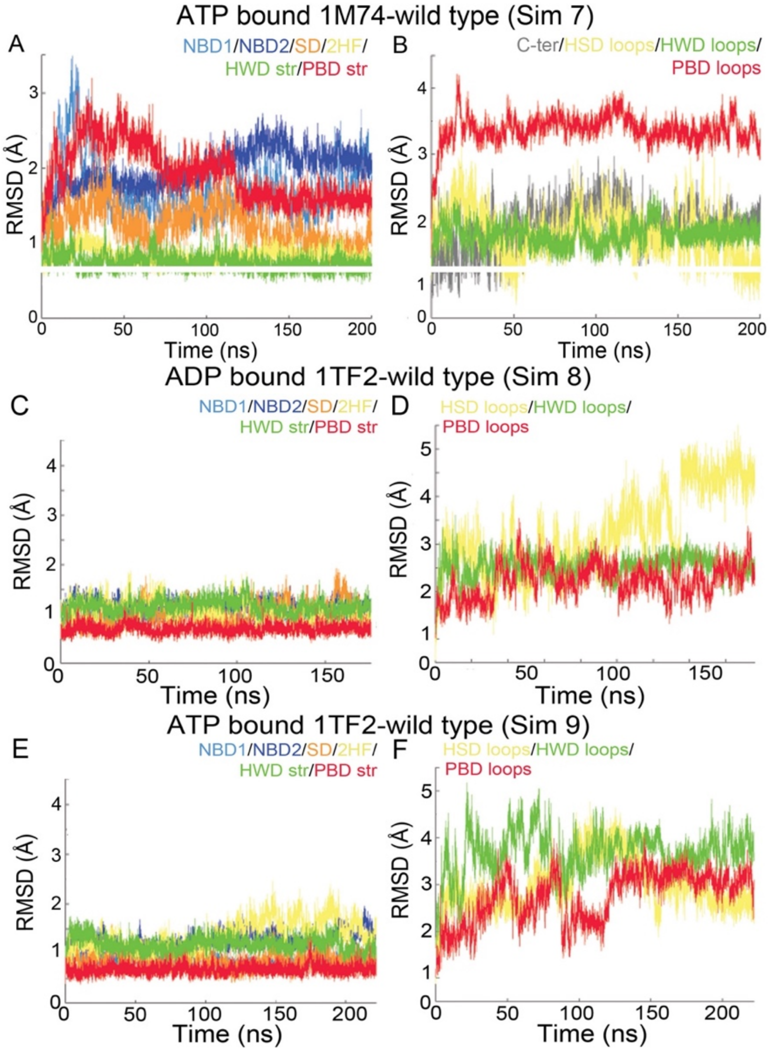


Figure 4.22: Cα RMSD profiles for SecA computed from Sim7-Sim9 (Table 11). We show the structured regions of SecA and the unstructured loops and termini for each Sim in *NPT*. Results are shown for the entire length of each simulation [187]. Image is from [102], a study presented in this thesis.

The distinct orientation of the PBD compared to the HWD in the initial crystal structures is a significant difference between Sim8 (ATP bound SecA) and Sim1 (ADP bound SecA) (Figure 4.20A). As a result, H bonds are not sampled between the PBD and the HWD in Sim8, and amino acids from the PBD form H bonds temporarily with those in NBD2 (Figures 4.20, Figure 4.26). Figure 4.26 shows the minimum distances between PBD-HWD, PBD-NBD2 and NBD2-HWD for Sim 1 and Sims 7-9. In Sim8, PBD is more distant from HWD during the simulation time and closer to NBD2 where it interacts temporally via H bonds.

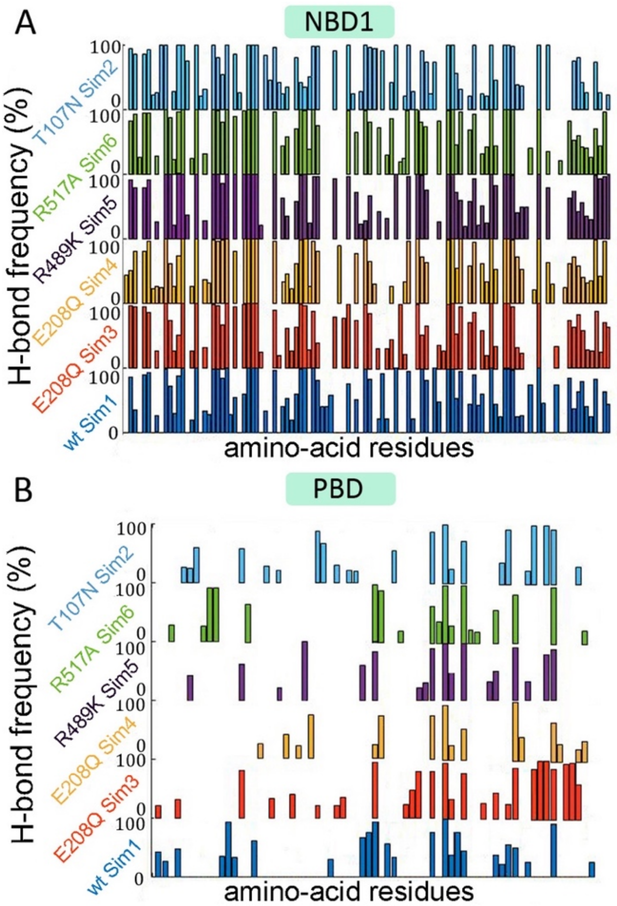


Figure 4.23: Intra-domain H bonds of (A) NBD1 and (B) PBD region. H-bond frequencies (%) are from the last 100ns of Sim1-Sim6 using 20% occupancy threshold [187]. Image is from [102], a study presented in this thesis.

4.3.3.3 Comparison of H-bond networks of wild-type and mutant SecA

To describe the way in which SecA responds to H-bonding changes, as a result of mutations, we generate graph visualizations of the H bonds (Figure 4.27). Our implementation is vital to understanding the correspondence of the protein to mutations and changes close to the nucleotide-binding site.

Several inter-domain H bonds in wild-type SecA tend to be particularly vulnerable to mutations at the NBD1/NBD2 interface, despite not being located close to the mutation location. For instance, dynamic H bonds between the PBD and the HWD, Y252-G656, Y252-K653, and K653-T251, are broken in all four mutants (Figure 4.27). Furthermore, in four mutations (Sim2,3,5) (Figure 4.27C), the H bond between HSD-E765 and PBD-K255, which is absent in wild-type SecA (Sim1, Figure 4.27C), exhibits significant occupancies ranging from 82% to 95%. Figure 4.27 shows the network of inter-domain H bonds for ADP-bound wild-type SecA (Sim1) and a network comparison between wild-type and mutations. For example, Figure 4.27B shows H bonds color-coded based on the number of mutations present. Figure 4.27C depicts H bonds present in wild type but absent in all other mutations and H bonds present in all simulation systems.

We conclude that mutations close to NBD1 and NBD2 regions affect H bonding at the PBD, a distant region relative to the DEAD motor, suggesting long-distance conformational coupling.

The Gate-1 groups D215 and R517 in Sim1 belong to a large H-bond network that contains all protein domains (Sim1, Figure 4.27). In T107N, E208Q, and R489K (Figure 4.27B), the Gate-1 between D215 and R517 has high occupancies, while mutations break additional H bonds between NBD1 and NBD2. For example, N107 (Sim 2) samples more H bonds than T107 in wild type. R489 is the center of many H bonds with low occupancy compared to mutation K489 (Sim 5). E208 H bonds with 100% occupancy to R489 while Q208 (Sim3, Sim4) H bonds with a low occupancy to R489. K489 (Sim 5) shows also low occupancy H bonding to E208 compared to Sim1. In R517A mutation (Sim6) the network of H bonds close to mutations shows lower occupancies compared to all other simulations (Figure 4.28).

4.3.3.4 PBD altered dynamics in ATP-bound SecA

We employed a graph model of SecA's inter-domain H bonds to investigate whether alterations in SecA's nucleotide-binding state affect the PBD's inter-domain H bonds. A comparison made between ADP-bound (Sim1) vs. ATP-bound SecA (Sim7) demonstrates that altering the ADP to ATP influences numerous H bonds at PBD-HWD distant region from the altered nucleotide (Figures 4.29A).

In Sims 8 and 9, we observe a modified H bonding pattern between the HSD and the PBD, as well as between the HSD and the HWD. In these simulations, the PBD and the HWD are almost completely unconnected due to the open conformation of the PBD (Figure 4.29B). In other words, our findings indicate that the alteration to the nucleotide state results in changes to NBD1, HSD and PBD H-bond dynamics (Figure 4.29B).

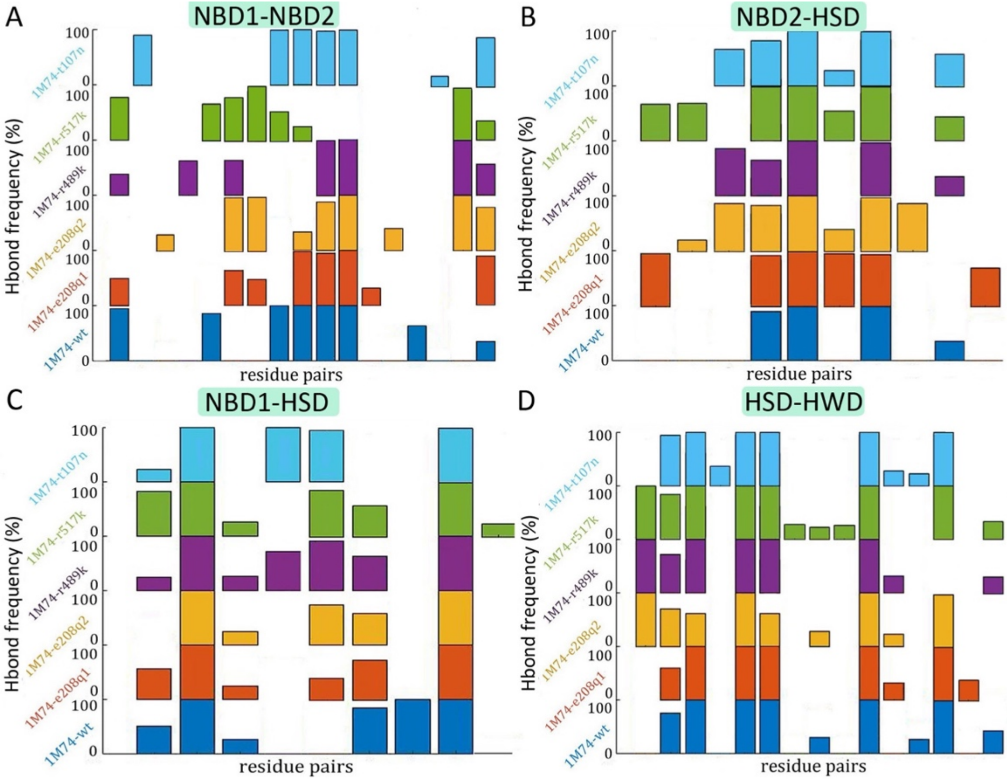


Figure 4.24: Inter-domain H bonds between domains including NBD1, NBD2, HSD, HWD. We used an H-bond frequency (%) cutoff of 20%. Results are from the last 100ns of Sim1 to Sim6 [187]. Image is from [102], a study presented in this thesis.

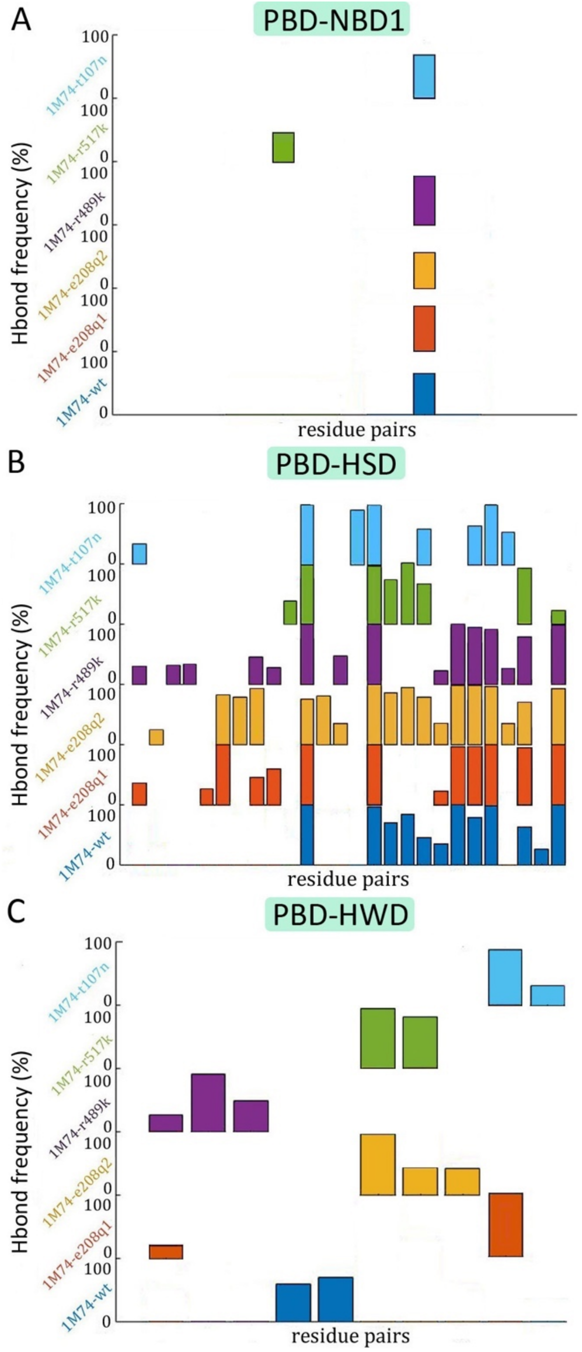


Figure 4.25: Inter-domain H bonds including the PBD with NBD1, HSD, and HWD. We used an H-bond frequency (%) cutoff of 20%. Results are from the last 100ns of Sim1 to Sim6 [187]. Image is from [102], a study presented in this thesis.

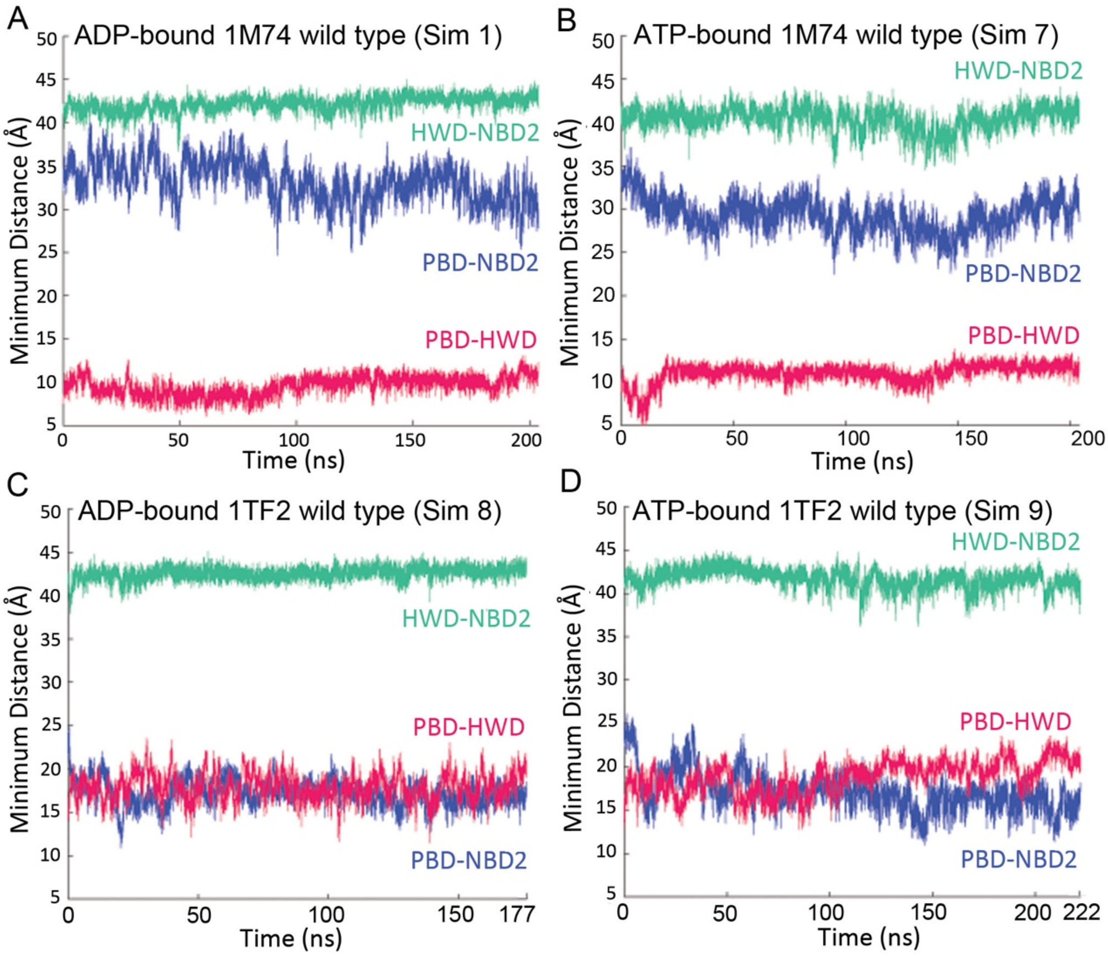


Figure 4.26: Calculation of minimum distances between PBD, NBD2 and HWD of SecA. We computed all possible distances between amino acids of the three functional regions and we kept the minimum distance for each time step of the simulations. Plots are from [102]. Image is from [102], a study presented in this thesis.

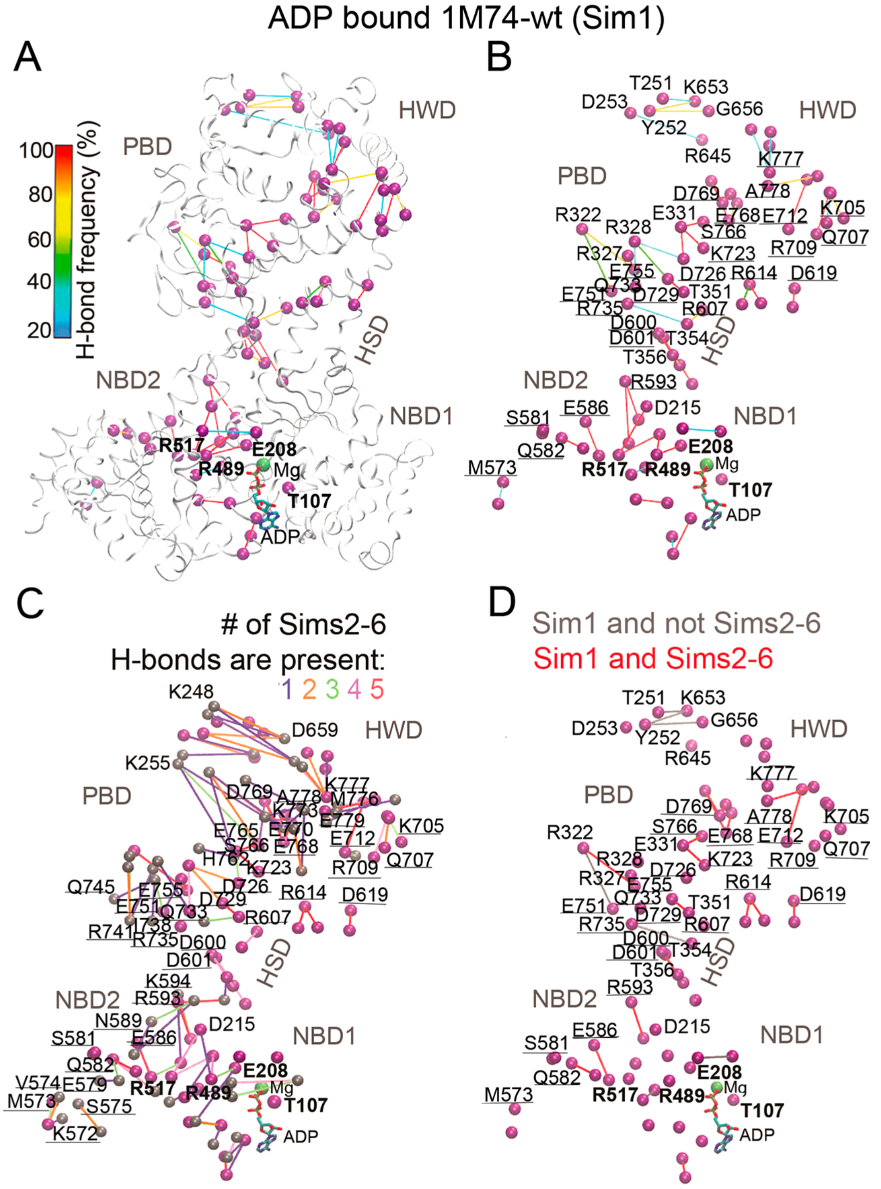


Figure 4.27: SecA mutants disrupt inter-domain H-bond networks. (A, B) Sim1 Cα atoms of the amino acid residues implicated in inter-domain H bonding are depicted as purple spheres. Edges represent H bonding between amino acids and are colored based on occupancy (%). The same protein groups and inter-domain connections illustrated in panel A are depicted in panel B, but this time they are labeled. (C) Sim2-Sim6 inter-domain H bonds. Gray spheres represent groups with inter-domain H bonding in minimum one of the mutations but not in the wild type Sim1. Lines are color-coded based on the number of mutation Sims in which a particular inter-domain H bond is present. (D) Mutation affects inter-domain H bonding. Grey lines show H bonds that are present in Sim1 but not in Sims 2-6; red lines indicate H bonds that are present in Sims 1-6. The last 100ns of the corresponding Sim were used for the analyses, and the H bonds have a minimum occupancy of 20% [187]. Image is from [102], a study presented in this thesis.

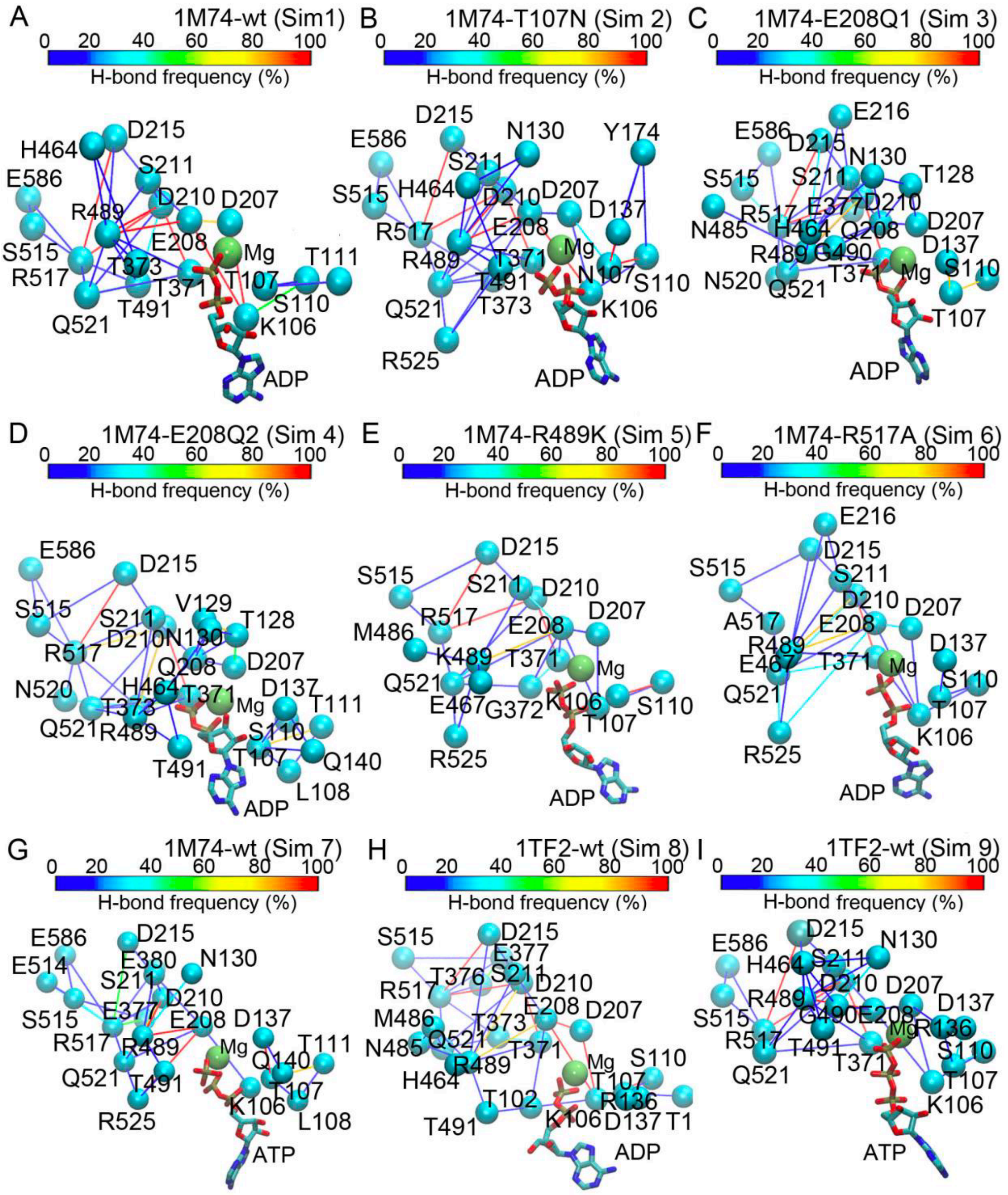


Figure 4.28: Close view of the H-bond networks at the nucleotide-binding site in Sim1-Sim9 [102]. Image is from [102], a study presented in this thesis.

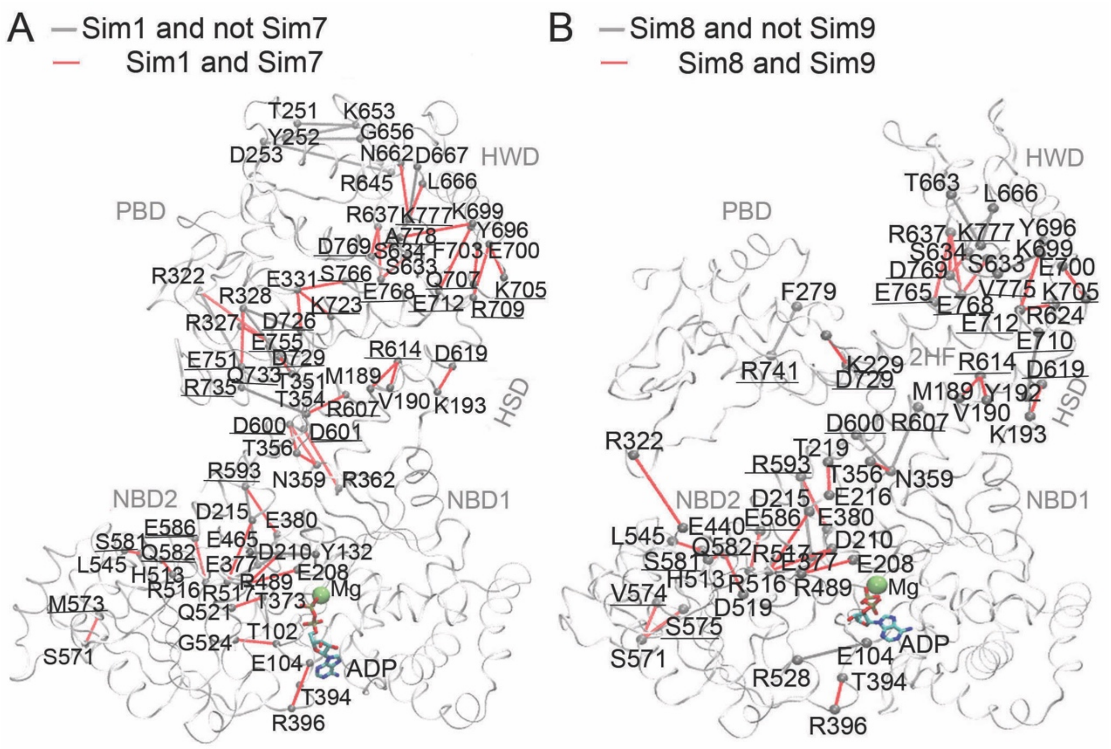


Figure 4.29: Inter-domain H bonds in SecA are influenced by the nucleotide-binding state. Protein Cα atoms are depicted as small grey spheres. Grey lines illustrate H bonds that are found in ADP-bound SecA but not in ATP-bound SecA, while red lines indicate H bonds that are found in both ADP-bound and ATP-bound SecA. (A) ADP- vs. ATP- bound 1M74 (Sim1 versus Sim7). (B) ADP- vs. ATP- bound 1TF2 (Sim8 versus Sim9). We utilized the last coordinate image of the Sim1 to locate the Cα atoms with H-bond occupancies ≥20% as calculated from the last 100ns of each simulation [102]. Image is from [102], a study presented in this thesis.

4.3.3.5 H-bond pathways coupling NBD to PBD region

Our simulation analyses presented demonstrate that changing nucleotide-binding state and therefore H bonds at the NBD1-NBD2 interface are related to changed dynamics at distant locations of SecA, comprising the PBD domain. We employed the graph model of SecA's H bonds and tried to find uninterrupted pathways that that promote structural and dynamical changes from the NBD interface to the PBD via direct H bonds between protein amino acids, or H-bond water bridges.

We used our algorithm implementation and Dijkstra's algorithm to search for H-bond pathways giving an initial (source) and end node without any intermediate node. As a source node, we selected T107 which is in the NBD1 region as it is close to the nucleotide, and as an end node, we selected K248 located in the PBD-HWD interface. That region is found to show altered dynamics in mutant simulations and SecA with ATP nucleotide (Figure 4.30A).

The search for H-bond shortest-distance pathway in Sim1 resulted to a connection of 16 direct protein H bonds including 17 nodes and a total length of ~27Å. The path starts from T107 to K248 with a branch to K257 which is connected both to K248 and D659 from the HWD. The pathway also contains DEAD motif's E208 and Gate-1's D215 making it crucial for NBD-PBD long-distance connection. The distance from each H bond is an average of the distance between each protein groups through the simulation time (Figure 4.30B).

We also used our algorithm to compute pathways with the highest occupancy (Figure 4.30C). The connections are similar with high ocuupancy between NBD1 and HSD region so that the pathway remains continuous. Separate sections of the paths have been examined, and we concluded that mutations or alterations in the nucleotide-binding state can modify the dynamics of these sections (Figures 4.30D, 4.30E).

The SecA is a soluble protein having the ability to interact with water. We extended our algorithm to allow the contribution of water molecules to our calculations. We started from *NVE* Sim1', we computed again the highest-occupancy H-bond paths and allowed water chains of maximum length three to participate to H bonding. We concluded on a similar result with rather high occupancies for most pathways (Figure 4.30F) implying the important role of water and its contribution to conformational coupling between distant domains of SecA protein.

We followed the same procedure for each Sim providing similar results. Long-distance pathways were found in each simulation system. We further remark that, considering the dense network of H bonds (Figure 4.30A), our implementation gives the power for various sampled H-bond pathways to be computed and illustrated setting different initial and end nodes and examining different protein regions.

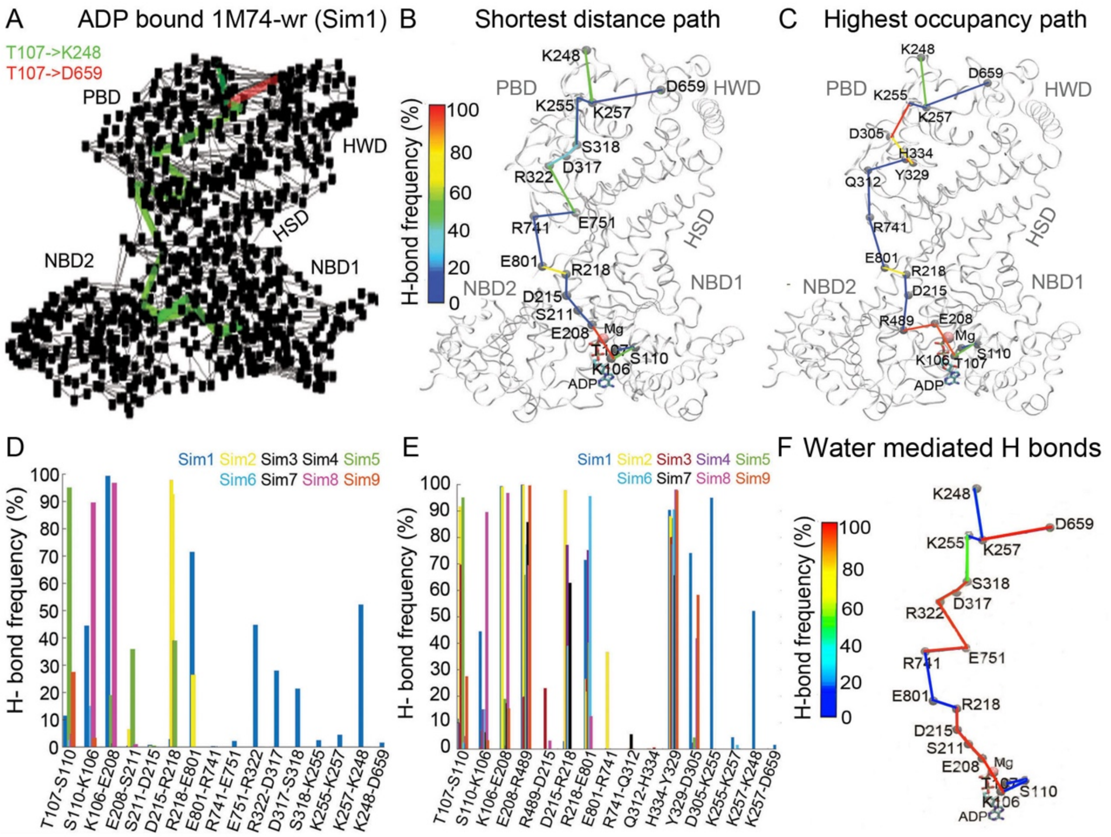


Figure 4.30: H-bond connections between NBD nad PBD interface. (A) Graph representation of H bonds sampled during Sim1. Black dots represent nodes and gray edges represent H bonds between the nodes. The green line shows the shortest-distance path between T107 and K248 or D659 (red line for the branching connection). A close-up of the (B) shortest-distance pathway showing each node and the corresponding occupancies and (C) the highest-occupancy pathway, respectively. (D) Sims1-9 occupancy values for the path illustrated in panel B and (E) panel C. (F) Water mediated H-bond pathway connecting NBD1 to PBD [102]. The graphics were generated in MATLAB R2017b [186]. Image is from [102], a study presented in this thesis.

4.3.3.6 Protein groups act as hubs in H-bond communication pathways

BC values could be utilized to determine nodes that belong to the shortest paths of other pairs of nodes in the network. When a node is structurally central, it links several amino acid residues via shortest paths and consequently may link remote areas in the graph. Any mutation /change to central nodes may disrupt the overall connections in the network.

We computed the normalized *BC* for each amino acid of SecA (Figure 4.31A) and we created a sub-graph that comprises nodes with elevated normalized *BC* values (Figure

4.31B). We can see that all protein domains include amino acids with high *BC* especially, NBD regions. This interface is structurally stable containing many H bonds as we showed in Figure 4.23, and Figure 4.24. A deeper examination of Figure 4.31B reveals protein groups that are involved to the long-distance H-bond pathway we showed in Figure 4.30. Examples of amino acids are T107, R517, E208, and R489, which are linked to high *BC* nodes. Other examples are Q800 of the HSD region, and K255, K257 of the PBD (see the long-distance path in Figures 4.30B, 4.30C).

The *DC* is a measurement of the number of distinct H bonds that are sampled at a single node. If a node has low *DC* during the simulation means that it has limited changes to the H-bond partner, it binds. On the other hand, high *DC* implies dynamic H bonds with diverse protein groups.

We mapped the *DC* values computed during the simulation on the protein structure (Figure 4.31C) and we showed a sub-graph of nodes with $DC \geq 7$ (Figure 4.31D). We detect that PBD is a rich region of high *DC* nodes. This is compatible with our findings that PBD is flexible and can reorient forming dynamic and various H bonds while NBD1 and NBD2 is more stable structurally maintaining specific H bonds during the simulation.

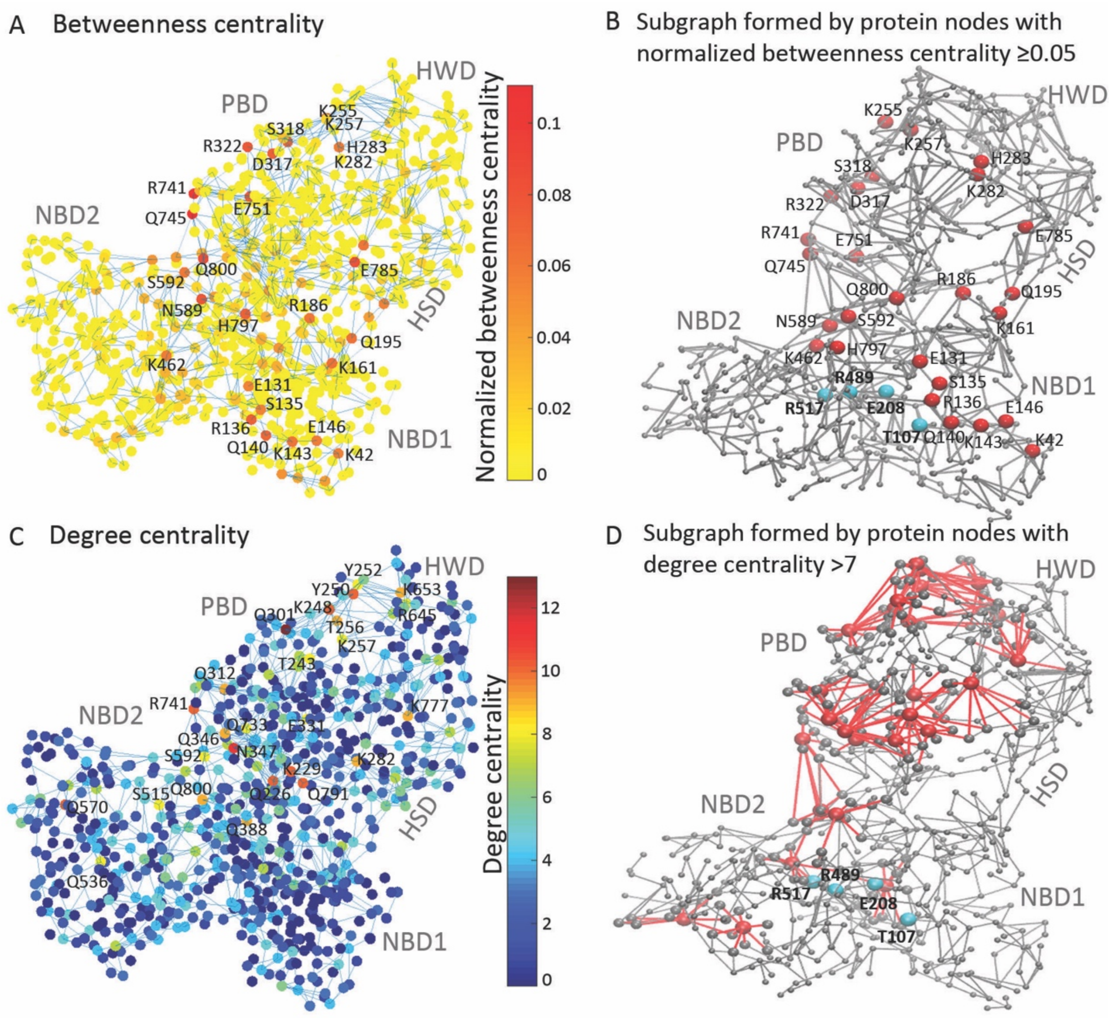


Figure 4.31: Protein groups that act as hubs in H-bond graphs. (A) *BC* values mapped in SecA (Sim1). (B) Molecular graphics showing nodes (red spheres) with higher normalized *BC* values. (C) *DC* values mapped in SecA (Sim1). (D) Molecular graphics showing nodes (red spheres and edges) with higher normalized *DC* values. Panels A and C were generated in Matlab [186]. Panels B and D were generated In VMD [23]. Image is from our study [102]. Image is from [102], a study presented in this thesis.

4.3.3.7 Inter-domain H-bond water bridges

We computed the residence times of waters within 4 Å to each SecA amino acid residue from the *NVE* simulation, Sim1' (Table 11). Our initial results show that ~ 90% of protein groups are enclosed by waters with a lifetime of <50ps (Figure 4.32). Our results identified waters that steadily remain close to the protein. These long-lived waters are primarily found at the NBD1-NBD2 interface, with a few exceptions at the interface PBD-HSD (Figure 4.32).

In the first step of our study, we implemented scripts to compute and visualize the H-bonded water chains that connect separate domains of SecA with a maximum of three H-bonded waters (L ≤3) in the bridge. Our results indicate that SecA domains possess wide and dynamic networks of protein/water H-bond bridges (Figure 4.33A). Connections are dynamic as most protein sites have low residence times (Figure 4.32B).

In the second step of the study, we wanted to investigate the way quite stable water-protein sites with residence times ≥50ps contribute to H-bond bridges between different domains of the protein. This time, the L value was less or equal to five (Figure 4.33C, Figure 4.33D).

Our results show that most water bridges are long (L=5) having low occupancies during Sim1' (Figure 4.33C, Figure 4.33D). We highlight 4 high occupancy water bridges between PBD and HSD, 2 between NBD1 and HSD, and six between NBD1 and NBD2. The latter involves stable connections even if they are long implying structural stability.

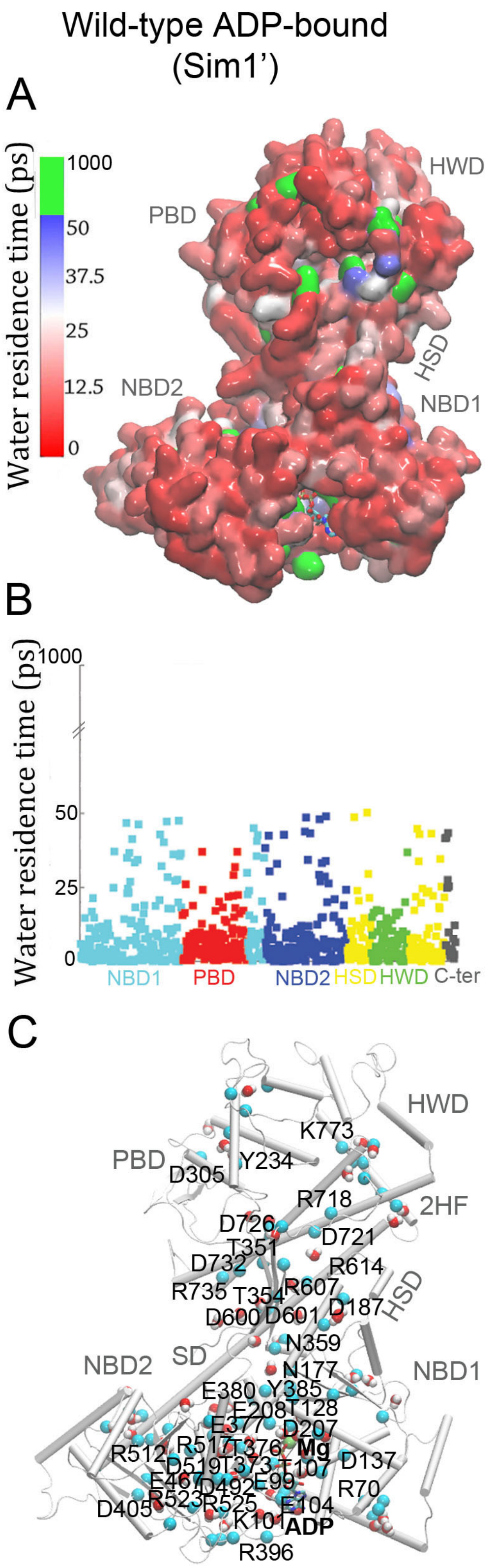


Figure 4.32: Illustration of water dynamics at the SecA's interface computed from Sim1'. (A) The first hydration shell of SecA is used to color its surface based on the average residence times of water molecules. (B) Water molecules' residence times within the first hydration shell

of each SecA group while excepting stable waters. (C) Depiction of amino-acid residues with high computed water residence times. Image is from [102], a study presented in this thesis.

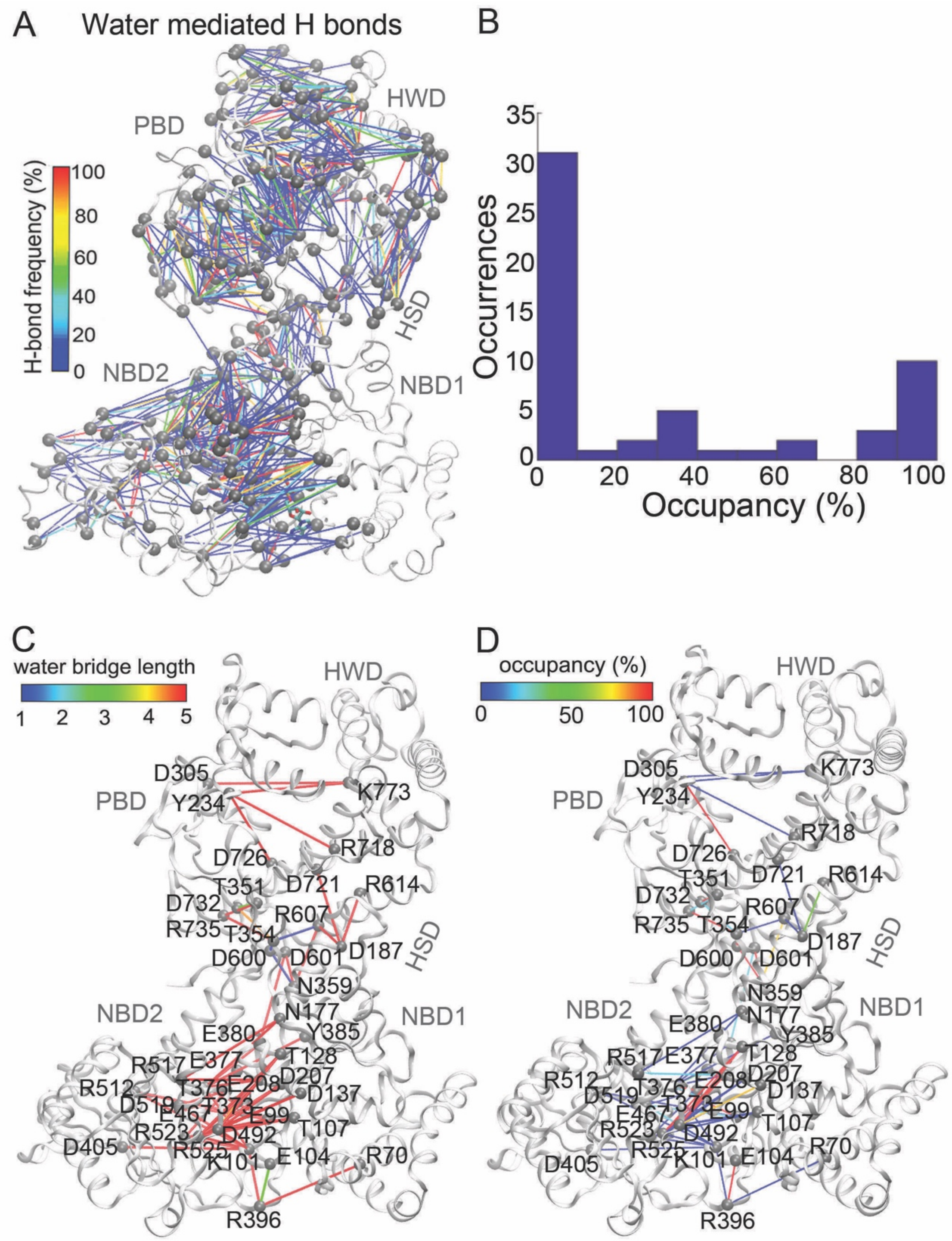


Figure 4.33: H-bonded water bridges between functional domains computed from Sim1' (A) Network representation of water bridges of length, L ≤3 showing the H-bond frequency of each interaction. (B) Histogram of water-bridges occupancies with residence times of ≥ 50ps. Inter-domain H bonds of stable water bridges (residence times of ≥ 50ps) based on (C) the length (1

to 5) of the water chain and (D) the bridge occupancy. Image is from [102], a study presented in this thesis.

4.3.3.8 Simulations of *E.coli* SecA

In the simulation of the monomeric SecA the clamp moved completely from the Open to semi-closed within ~60ns and from the Open to closed state within ~150ns. Clamp motion is shaped by interactions between charged group clusters on the PBD, scaffold and NBD2, while NBD1 undergoes only minor rearrangements. The PBD moves towards and salt bridges with NBD2 to form the closed state.

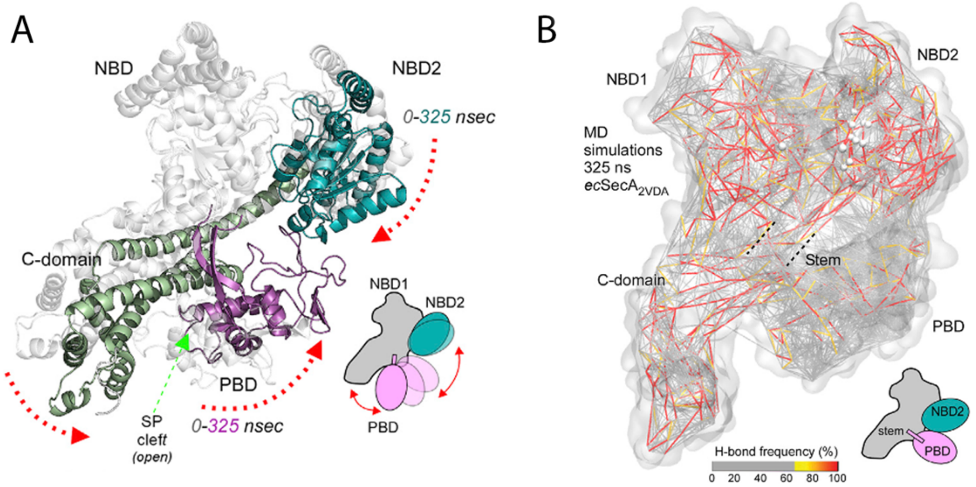


Figure 4.34: (A) Based on NBD1, we present an overlap of two coordinate snapsots from 0ns (protein in grey) to 325ns of ecSecA2VDA to illustrate the movement of PBD towards NBD2. (B) Water mediated H-bond network in ecSecA2VDA. The network's lines are color coded according to how frequently they appeared in the simulation and each one of them represents a water bridge in the graph. Image is adapted from [99], a study presented in this thesis.

Clamp closing appeared largely driven by H bonding. Since we imposed no constrains, the simulated PBD rearrangement informs on the likely reaction coordinates of PBD reorientations. These appear to be accompanied by large changes in the local hydration at the PBD/NBD2 interface. Three coordinate snapshots, approximating the Open, Semi-closed and Closed clamp states, were analyzed. Intra-domain H bonds are markedly decreased as PBD moves from the Open to the Semi-closed and Closed states [99].

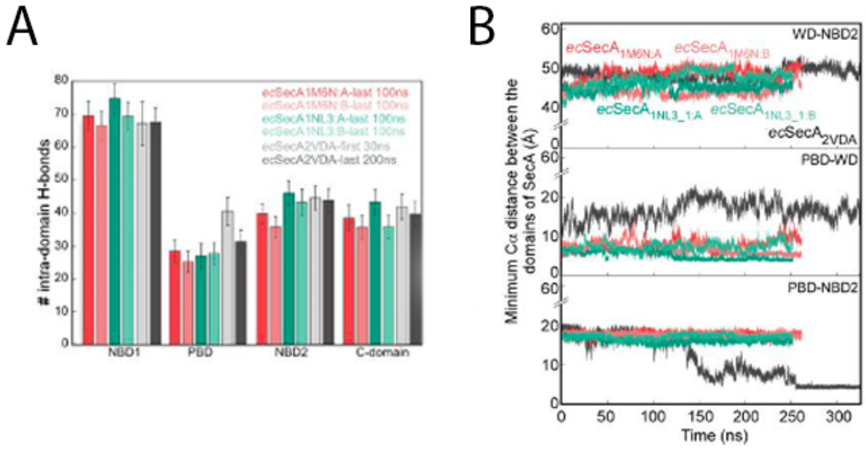


Figure 4.35: Dimeric and monomeric SecA analysis. (A) Average number of intra-domain H bonds for distinct protein domains for three simulation systems, monomeric SecA (ecSecA2VDA) and dimeric SecA (1M6N), (1NL3_1). We use light shade for the first 30ns and dark for the last 200ns of the monomer. Similarly, red and green shades are used for the last 100ns of each dimeric protomer. (B) Time series of the minimum Cα distance between WS-NBD2, PBD-WD, and PBD-NBD2 for each simulation system. Image is adapted from [99], a study presented in this thesis.

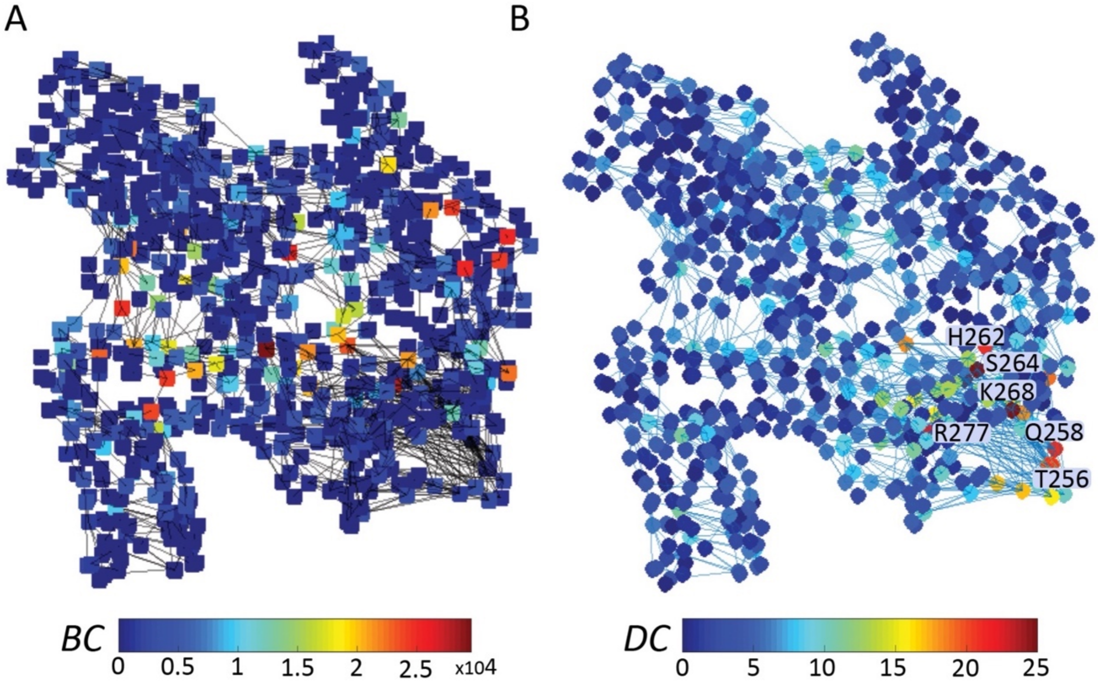


Figure 4.36: High-centrality groups in H-bond networks of the ecSecA2VDA. Protein amino acid residues are visualized as spheres colored according to the *BC* and *DC* values, respectively. Map of (A) *BC* and (B) *DC* values computed from the last200ns of the simulation. H-bond connections shown in graphs account for all unique H bonds that are sampled during the last 200 ns of Sim1; each of these unique H bonds has a different occupancy. Centrality values were computed and prepared with MATLAB R2017b. Image is generated for the scope of this thesis.

As we can see in Figure 4.35A, the PBD of monomeric SecA shows the greatest variation in H bonds between 30ns and the last 200ns of the simulation. Intra-domain H bonds in the two forms of dimeric SecA do not change during simulation, however, there are small variations between the two protomers.

In Figure 4.36, we present our analysis of *BC* and *DC* for ecSecA2VDA computed from the 200ns of the simulation. High *BC* nodes are the crossroads of shortest-distance and continuous pathways in the network of sampled H bonds. High average *DC* nodes are in an environment that allows them to bind to different H-bond partners during the simulation indicating structural stability or not. Panel B shows that PBD has the highest *DC* values compatible with our simulation as this domain reorients and moves towards NBD2. It forms and breaks numerous H bonds with different amino acids during the simulation. The difference between the monomer and the two dimers is that *DC* values of PBD are lower in dimers as this region does not diverge from NBD2. This suggests that extra contacts provided by the dimers are what account for the PBD's stability in the Wide-open state. The same observation is shown in Figure 4.36B. We measure minimum Cα distances between protein domains. Only in monomeric SecA, the distance between PBD and NBD2 decreases. A salt bridge between R342PBD and E460NBD2 is sampled only in ecSecA2VDA [99].

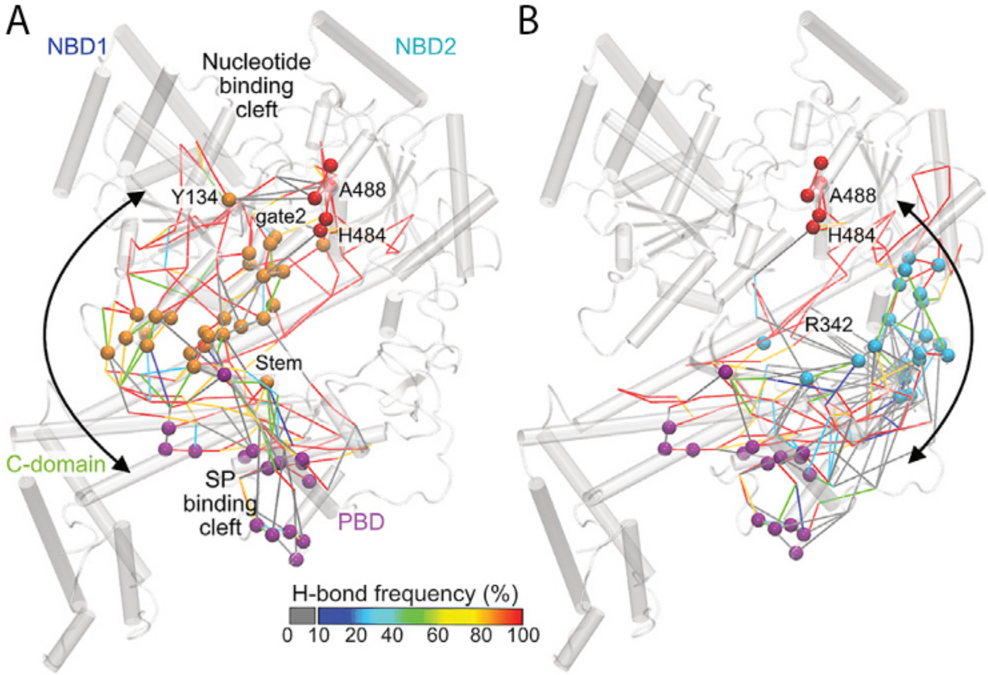


Figure 4.37: Graph analysis of ecSecA2VDA simulation revealed two pathways connecting gate2 (red spheres) to the PBD through the stem (orange spheres) or through NBD2/PBD interface (cyan spheres). Lines show the H-bond frequency of each sampled H bond. Image is adapted from [100], a study presented in this thesis.

It is shown that pre-protein binding alters the dynamics of gate2 (H484 and A488 in motif Iva), powers the ADP release and restarts the ATPase cycle. Additionally, it is anticipated that clamp dynamics and PBD movements can be regulated by stem dynamics. Stem seems to be a critical checkpoint of allosteric networks [100]. It is an anti-parallel b sheet with location shown in Figure 4.34B.

We identified H-bonded water bridges connecting two allosteric regions to ascertain how PBD closure would let the signal peptide binding cleft to communicate with gate2. The shortest or most common H-bond paths that may be changed along the protein reaction coordinate were identified via graph analysis. There were two suggested routes. One uses the stem interface which interconnects PBD to gate2 (Figure 4.37A) and the other the closed NBD2/PBD interface leading to motif IVA of gate2 (Figure 4.37B). Our results are of importance as they are confirmed by observations of experimental studies [99, 100].

4.3.4 Summary

Our algorithm implementations to compute H bonds from simulations and create graphs of sampled interactions revealed a large H-bond network in the SecA protein motor (Figure 4.20) [78, 79, 99, 100, 103]. Our findings are correlated with the functional role of each protein domain and the overall conformational dynamics. NBD1 and NBD2 are stable regions forming less H bonds with different protein groups whereas PBD is more dynamic and can reorient. Our observations are relative to RMSD profiles, intra- and inter-domain H bonds for each region and between regions, centrality calculations (*BC* and *DC*), and H-bond water-bridges dynamics.

Our graph theory calculations transformed the protein to nodes and edges of H bonds revealing that the area where ATP binds to the NBD1-NBD2 region couples to the protein binding domain and mutations at the NBD modify the dynamics of PBD. This observation suggests a long-distance conformational coupling between functional domains of the protein. Transient and continuous H-bond pathways connecting distant regions are sampled (Figure 4.30). The existence of a signal peptide at the PBD affects the movements of the NBD2, according to SecA's prior studies [79], which are consistent with this long-distance conformational coupling.

The role of water dynamics in the soluble SecA protein is essential. We computed the water residence times at the first hydration shell of the protein and we found less than 50ps for most of the protein sites (Figure 4.32B). There are also stable and long-lived waters many of these found at the NBD1-NBD2 interface. This observation is compatible with the structural stability of the region.

Water H bonds connecting protein groups enhance the coupling of different regions (Figure 4.30, Figure 4.33). Figure 4.32A shows that the surface of SecA is diverse and complex regarding the dynamics of the adjacent water molecules, which is consistent with calculations on the PsbO component of photosystem II, a different soluble protein [220].

Our centrality analyses revealed important results for protein sites that might act as hubs to the network being part of many pathways (Figure 4.31). That means that one node is essential for the connectivity because if we remove it from our calculations, the graph will be disrupted. We measure *BC* for hubs through shortest-distance pathways and *DC* for local connectivities. Results are compatible with the protein structure of SecA, the stability of NBD and the flexibility of PBD. We also highlighted important nodes for the whole connectivity with some of them being part to long-distance pathways from NBD to PBD region.

We extended our MD simulation studies to *E.coli* SecA structures, one monomer and two dimers (ecSecA$_{2VDA}$, ecSecA$_{1M6N}$ and ecSecA$_{1NL3\_1}$). In the monomeric SecA the PBD moved completely from the Open to closed state within ~150ns. We detected a significant movement and reorientation of PBD towards NBD2 (Figure 4.34). Clamp closure appears thermodynamically favourable, and largely driven by H-bond clusters and intra-domain H-bonding shifts (Figure 4.35); the latter propagating changes to the rest of the protein (Figure 4.37). Nevertheless, the three experimentally detected stable clamp states (Open to the Semi-closed and Closed) suggested the existence of energetic barriers and troughs, presumably imposed/regulated by intra-protomeric elements and ligands. On the other hand, the two dimers of SecA start from the Wide-Open state and remain to that state during ~200ns of the simulation. We analyzed intra- and inter- H bond-graphs, minimum distances between functional domains, salt bridges, water dynamics, H-bond graph theory algorithms including shortest-distance pathways and centrality measures [99, 100]. Our calculations show the plasticity of PBD region and the long-distance communication between PBD to NBD2 and the impact of changes in PBD to the whole graph of interactions in the protein. We found two pathways connecting gate2 to the PBD through the stem or through NBD2/PBD interface (Figure 4.37).

Our algorithms developed here are robust and can analyze complex H-bond networks of other protein complexes. Studies on dynamic H-bond networks, such as the one shown here, are particularly interesting for proteins whose function includes protonation state variations [282, 283]. In myosin, the reaction coordinate for ATP hydrolysis includes many transition states and proton exchanges [284]. The ATP-bound models described here provide a foundation for future research on the ATP hydrolysis mechanism and allosteric conformational coupling of SecA.

4.4 Applications of Graph-based algorithms to spike protein S of SARS-CoV-2, VASA and Channelrhodopsin *C1C2*

Our algorithms are designed for solving not only a particular model but applied to a wide range of complex bio-molecular structures. We present our algorithm applications to the spike protein S or SARS-CoV-2, the DEAD-box protein VASA Channelrhodopsin C1C2.

In 4.4.1, figures and tables originally published in the Journal of Structural Biology have been reproduced with permission from J. Struct. Biol. 212, no. 2 (2020): 107617. Copyright © 2020 The Authors. Published by Elsevier Inc.
In 4.4.2, figures and tables originally published in the Journal of Revue Roumaine de Chimie have been reproduced with permission from Rev. Roum. Chim 66, no. 10-11 (2021): 845-853.
In 4.4.3, figures and tables originally published in the Journal Chemical Theory and Computation have been reproduced with permission from J. Chem. Theory Comput. 15, no. 12 (2019): 6781-6798. Copyright © 2019, American Chemical Society.

4.4.1 Spike protein S

A novel coronavirus caused an outbreak of pneumonia-like illness in December 2019. Severe Acute Respiratory Syndrome (SARS-CoV-2) is a homotrimer and its structure is adorned with many membrane-bound spike proteins S anchored to Angiotensin Converting Enzyme 2 (ACE2) host cell receptor [285-288]. More specifically, the Receptor Binding Domain (RBD) of the protein S [287, 289] binds to ACE2 }[286, 290].

Cryo-electron microscopy (cryo-EM) categorized protein S into three states: open, closed and pre-fusion [291, 292]. To gain understanding into the potential impact of H bonds on the conformational dynamics of the protein, we computed H-bond graphs for the three states and for ACE2 bound to a segment of RBD and utilized this approach to find interaction networks that might be important in deciding which protein conformations bind to the receptor. Those interactions are essential for virus invasion. We employed centrality metrics to pinpoint amino acid residues that play a pivotal role in the connectivity of nearby H-bond clusters [194].
Our results from the closed conformation indicated a wide central cluster of H bonds which is symmetric to the three protomers and same H-bond clusters for each protomer close to the ACE2 receptor. That symmetry of H bonds is significantly disrupted in the open and to a greater degree, in the pre-fusion conformation. We also found four H-

bond clusters in the PBD-ACE2 complex. An explanation of the high binding affinity between S and ACE2 is a deep-lying H-bond network. In conclusion, our findings imply that the H-bond rearragements result to the loss of each protomer H-bond clusters symmetry in open and pre-fusion contrary to the closed conformation.
As starting coordinates, we used PBD:6VYB, PBD:6VXX [291] and PBD:6VSB [292] for the open, closed and the pre-fusion conformation [194].

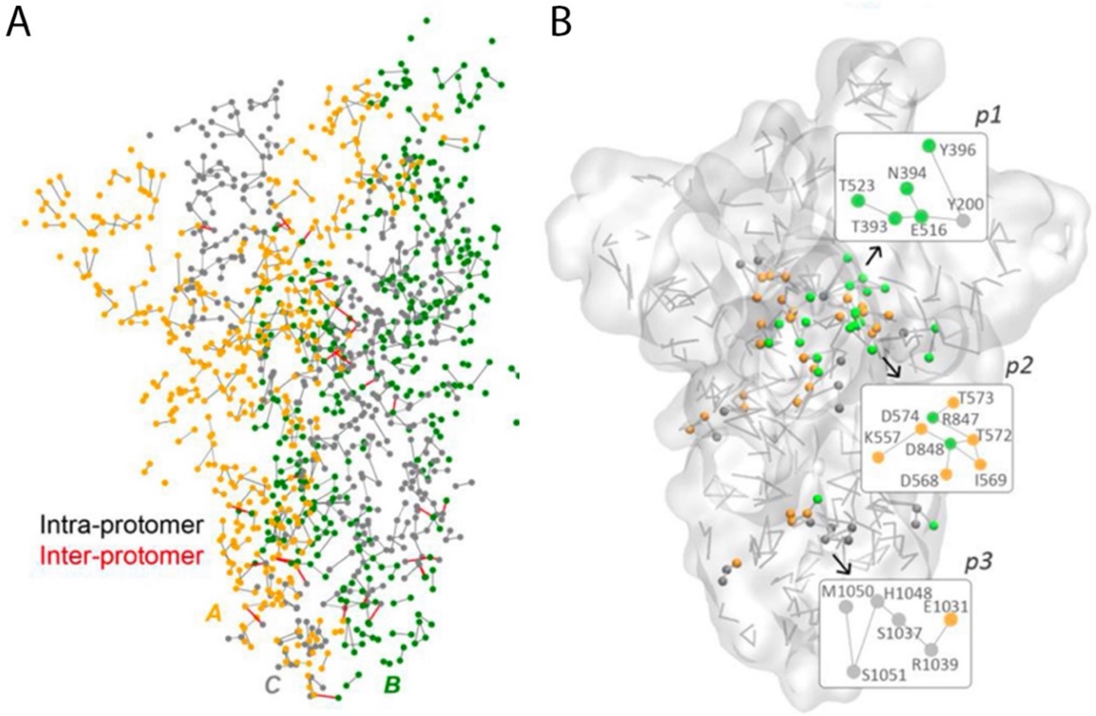


Figure 4.38: H-bond network of protein S in the pre-fusion conformation. (A) Intra- and inter-protomer H bonds. Lines represent H bonds and the protein is color coded based on the protein chain. (B) Illustration of clusters with a size of σ ≥ 6 that comprise both intra- and inter-protomer H bonding. We denote as p the three calculated clusters. Image is adapted from [194], a study presented in this thesis.

For every conformation of protein S, there are approximately 798-902 H bonds overall (Figure 4.38A). In Figure 4.38B, we show H-bond clusters of cluster size at least 6. There are two clusters (p1,p2) in the central region and one cluster (p3) at the stem of the protein structure. A similar picture is observed in the open and closed conformation.

Since it is difficult to describe the stryctural dynamics of the protein due to its large size, we wondered if H-bond networks could help by pointing out locations where interactions are altered. Our approach involved ranking H-bonding groups based on their participation in H bonds using centrality measures and examining the impact of high-centrality nodes to the total connectivity in the graph.

We found higher *BC* values for the closed compared to the open and pre-fusion conformation suggesting a larger number of extensive H-bond paths in the closed state (Figure 4.39, Figure 4.40). Most nodes have *DC*<2, and only a small number of groups have three or five H bonds. Let's examine amino acid N437 both having high *BC* and *DC* (Figure 4.39). While in the pre-fusion conformation, the *DC* of N437 is 5, in the closed conformation the *DC* is equal to 1. *BC* values remain high in both states. This suggests N437's local environment underwent structural rearrangements.

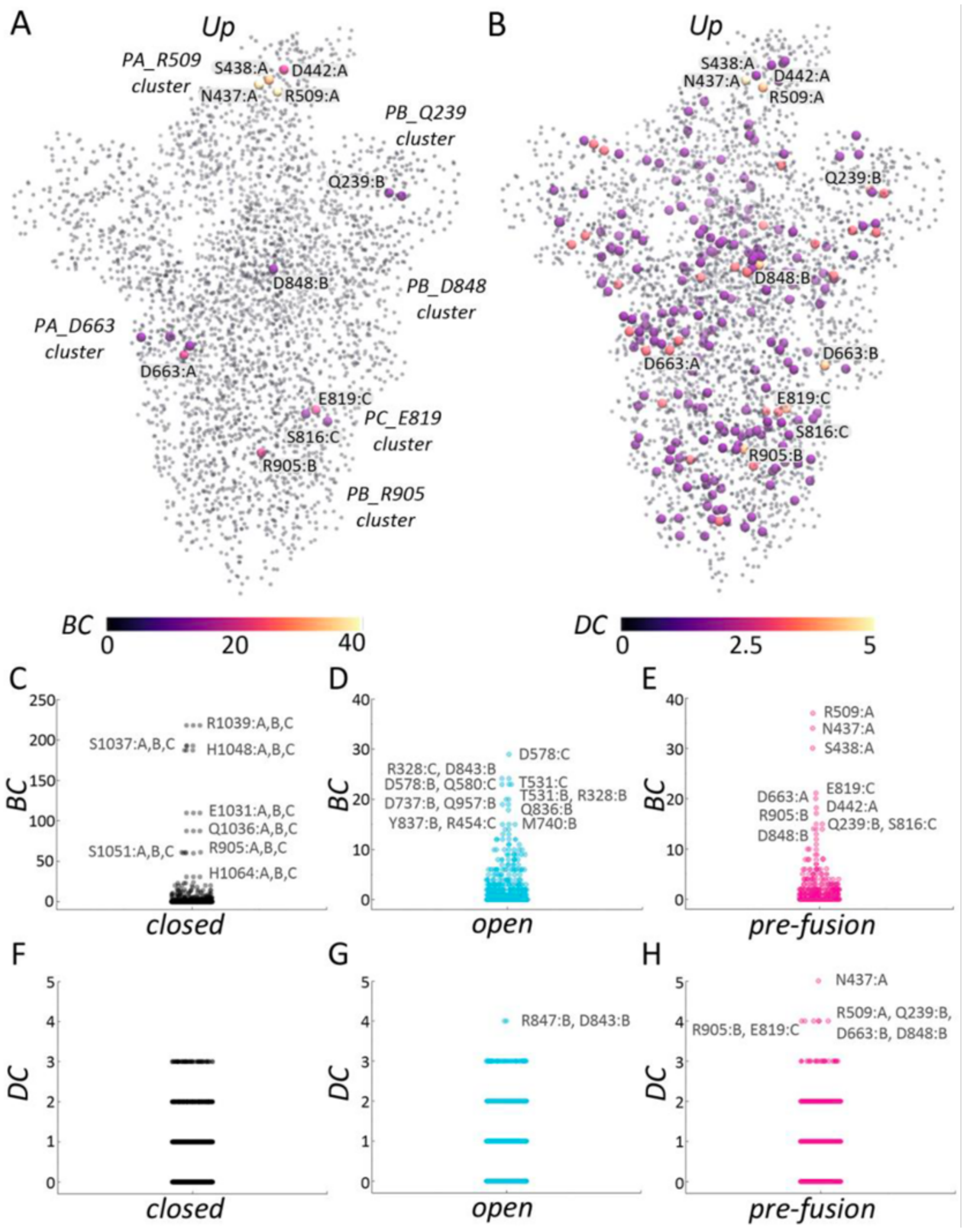


Figure 4.39: Centrality analyses in protein S. (A-B) Protein S molecular images with Cα atoms depicted as colored spheres based on *BC* and *DC* values. (C–H) Plots of the scattering of *BC* (panels C-E) and *DC* values (panels F-H). We label specific amino acids with high centrality values. Image is from [194], a study presented in this thesis.

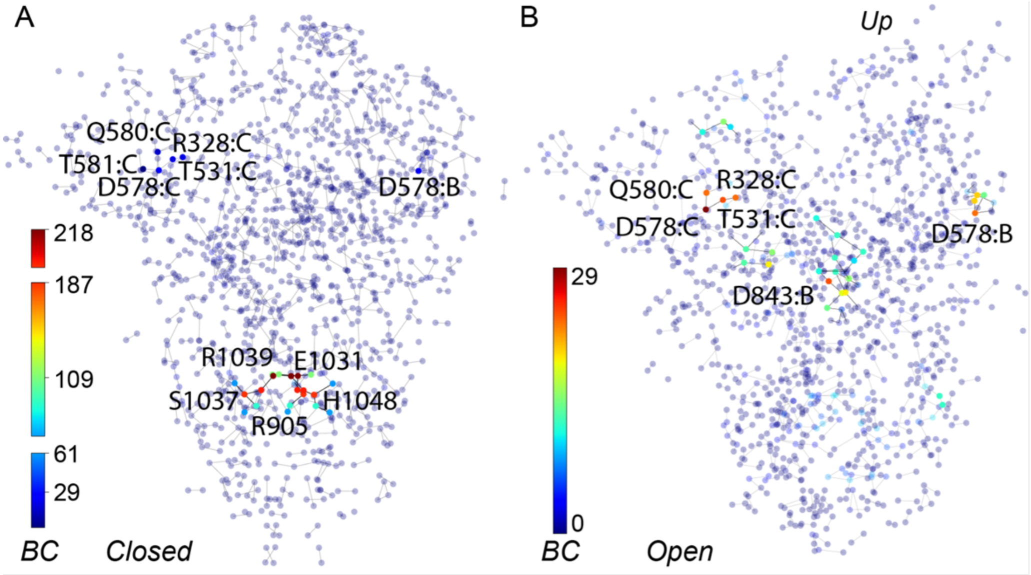


Figure 4.40: *BC* centrality analyses in protein S. Amino acids are represented as spheres color-coded based on *BC* values from closed (panel A) and open (panel b) conformation of spike protein S. Image is from [194], a study presented in this thesis.

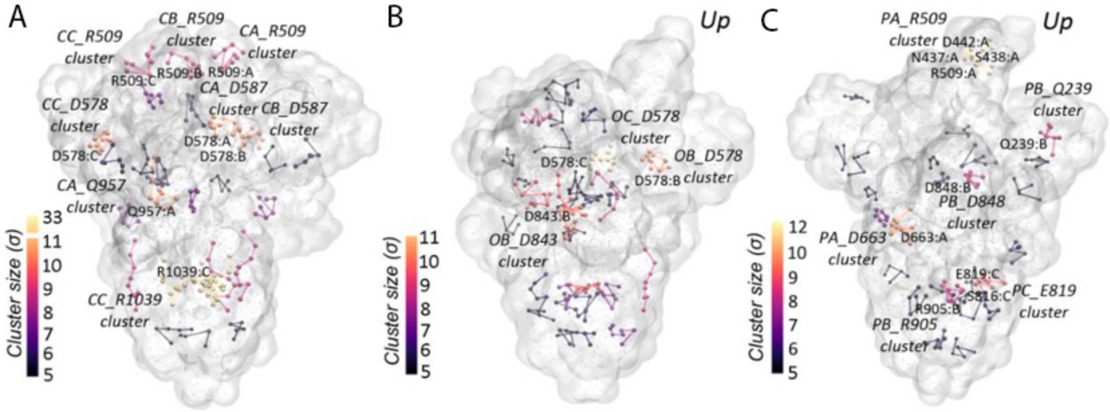


Figure 4.41: H-bond cluster size analyses in protein S. For clarity, we display H-bond clusters with σ ≥ 5 in the closed (panel A), open (panel B), and pre-fusion (panel C) conformation. Image is adapted from [194], a study presented in this thesis.

A centrally located group in a cluster with many H-bonding groups (large cluster size σ) can be a crossroad of numerous shortest-distance pathways and therefore have high *BC*. Clusters with relatively large σ can also show low *BC* values, including *BC* = 0. These are H-bonding groups that are situated at the cluster's periphery (Figure 3.24). An important cluster here, is R509. In SARS-CoV-2, mutating R509 to alanine reduces the ability of the protein to bind ACE2 [293].Our findings are as follows: in the closed conformation, R509 is a component of threefold symmetrical clusters; in the open conformation, two protomers contain symmetric R509 H-bond clusters; in the pre-fusion phase, there is no symmetry of the H bond, and an extensive H-bond cluster with high centrality values is only visible for the RBD up (Figure 4.39A, Figure 4.41). This may be crucial for the binding or RBD to ACE2 of the host cell.

In the closed conformation, we found a remarkable cluster of 33 symmetrical groups (11 of each protomer). The cluster is smaller to the open conformation but changed more to the pre-fusion suggesting that reaction coordinate of spike protein S involves a reorganization of H-bond networks and a possible mechanism of interaction between the protein and the receptor [194]. H-bond networks between RBD and ACE2 are analytically presented in [194]. The H bonds and clusters found close to receptor are compatible with the electrostatic potential of the surfaces [194]. An overview of our findings are illustrated below (Figure 4.42).

Another key finding of our study is a group of four clusters with high centrality values between the receptor-binding domain. One of the clusters include N501, which was mutated to TYR in the new SARS-CoV-2 501Y.V2 lineage [294].

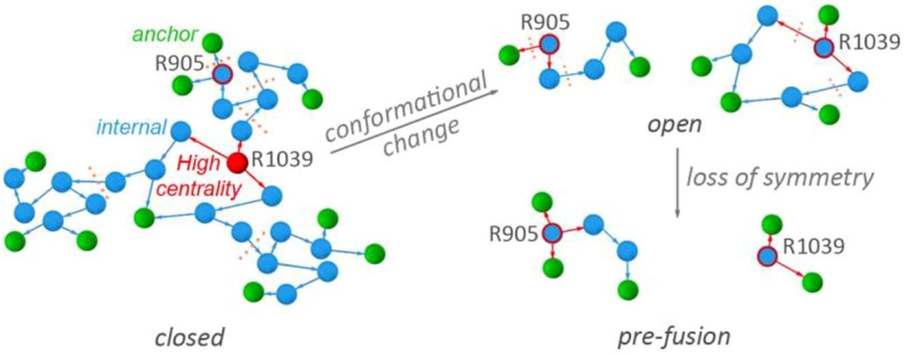


Figure 4.42: Illustration showing conformational change in spike protein S. Two important and extensive clusters found in the closed conformation (R1039, R905) are reorganized and changed in size and therefore centrality. We found a loss of symmetry as structure goes to its pre-fusion state. Image is from [194], a study presented in this thesis.

4.4.2 DEAD-box protein VASA

RNA helicases that couple binding and hydrolysis of ATP with the binding and modification of the RNA are called DEAD(Asp-Glu-Ala-Asp)-box motor enzymes [295]. SecA belongs to the DEAD-motor helicases and serves as a pre-protein motor necessary for the bacteria protein export mechanism [262, 296]. Coronaviruses, such as SARS-CoV-2, may also utilize host DEAD motors for the viral entry to the host cell [297]. HIV-1 is another example as it depends on host cell proteins to assure the expression of its genes since it lacks the ability to encode its own helicase. Through their effects on post-transcriptional phases of HIV-1 replication, DEAD-box proteins may have an impact on the preservation of viral latency [298]. Therefore, it is important to describe the response mechanism of DEAD-box enzymes since it may help in the development of medicines.

Here, we study the *Drosophila melanogaster* VASA, a DEAD-box enzyme involved in the cell cycle with ATP- and RNA-binding sites. We use MD simulations to investigate dynamic H bonds that may promote the conformational coupling between the two binding sites [299]. Using graph theory algorithms in SecA, a DEAD-motor, we have shown that dynamic H-bond networks guides the nucleotide binding and ensures long-distance conformational coupling to the binding domain [99, 100, 102, 300]. By querying the H-bond graph, transient H-bond subgraphs connecting different protein regions can be obtained.

We find that VASA is home to a large network of H bonds, some of which are rather small and suggestive of strong interactions. Water plays a vital role in H-bonding clusters. Protein changes following ATP binding and hydrolysis may be relayed to the protein through networks of dynamic protein-water H bonds. These networks include amino acid residues found in the DEAD box [301].

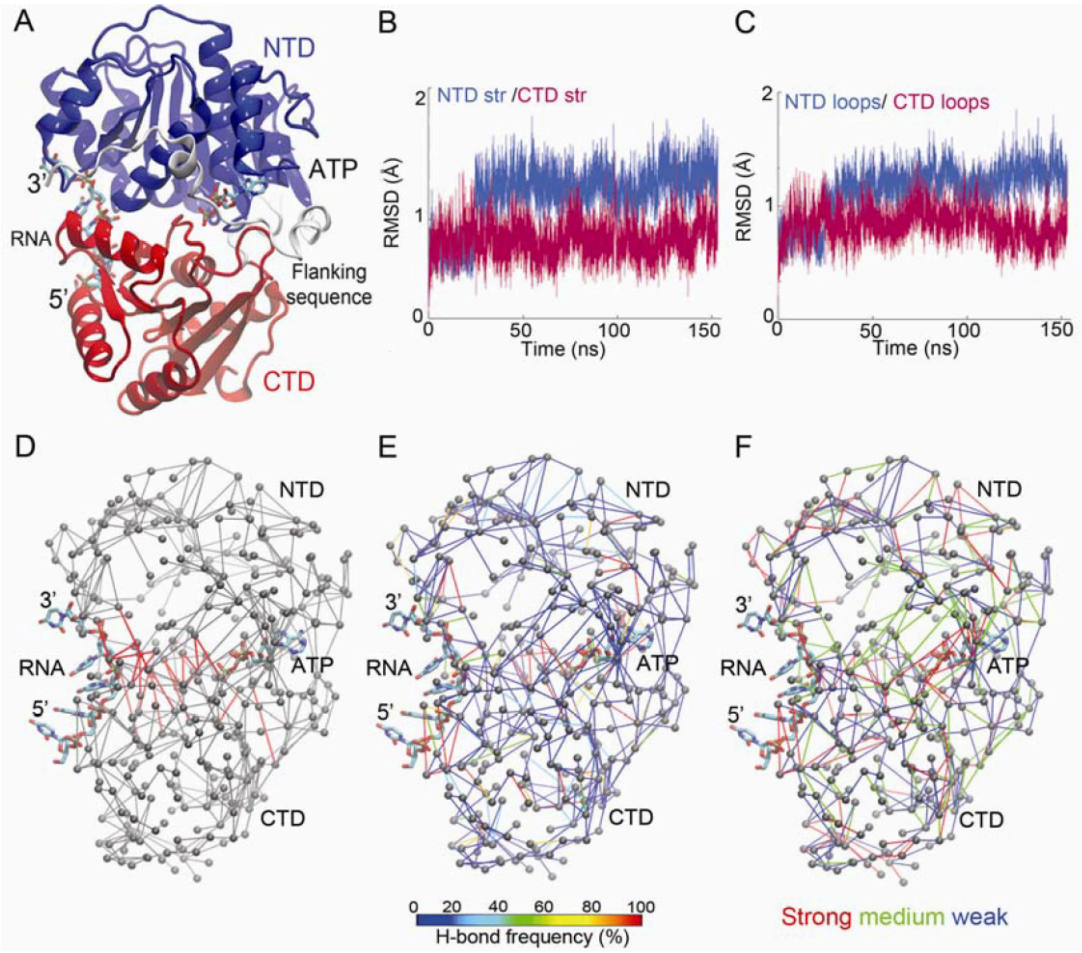


Figure 4.43: H-bond dynamics of *D. melanogaster* VASA. (A) Protein functional domains are depicted in its molecular graphics. (B, C) Cα RMSD profiles of structured (panel A) and loops and termini (panel B) regions of VASA were produced during the simulations. (D-F) H bonds calculated throughout the simulations of ATP-bound VASA. (D) H-bond interactions are illustrated as grey lines while H bonds between NTD and CTD are depicted as red lines. (E) H-bond interactions are shown as colored lines that range from red (100% occupancy) to blue depending on the H-bonds occupancy. (F) H-bond network displaying the sampled H-bonds' strength during the simulation. Image is from [301], a study presented in this thesis.

The Cα RMSD results are within 2Å for both structural and loop regions of the protein (Figure 4.43). The structural stability of the protein is influenced by the formation of H bonds between protein sidechains. During simulations, an average of approximately 140 H bonds are present at any given time, with around 20 occurring between NTD and CTD. In Figure 4.43 panels E and F, we see that the H-bond connections tend to have small occupancies and their strength is mostly medium and weak. We also observe sampled H bonds with high occupancy and strong strength meaning short distances. Specific amino acids being part of H-bond clusters and water dynamics are shown in [301] suggesting a potential functional role of these protein groups. Additionally, mutations may affect the long-distance conformational coupling and hydrolysis of ATP can be disrupted from RNA unwinding. Examples of such mutations investigated in [301] are Q333A and R551A. We suggest that our algorithms can give a versatile tool to compute networks and detect changes while mutations are inserted.

4.4.3 Channelrhodopsin chimera *C1C2*

I collaborated with Malte Siemers who developed the Bridge analysis algorithm package in Python programming language and available as a PyMol plugin[231] for the detection and visualization of H-bond networks and water wires under the guidance and supervision of Prof. Dr. Ana-Nicoleta Bondar. I contributed to the scientific discussions and the application of Bridge to channelrhodopsin-1–channelrhodopsin-2 chimæra (C1C2) dimer (PDB ID: 3UG9) to investigate lipid-protein H-bond networks mediated by water molecules from ~200ns of MD simulations in *NPT.* The C1C2 dimer was placed within a hydrated bilayer of POPC lipids. The simulation protocol is presented in [231].

Light-gated cation channels called Channelrhodopsins are a component of the process that regulates how microalgae move in response to light. Optogenetics, the study of neural activity in tissues and living organisms, makes extensive use of their function due to their expressibility in a variety of host cells and the highly controlled generation of photocurrents.

C1C2 has seven transmembrane helices, five (TM1-TM5) from channelrhodopsin-1 (ChR1) and two (TM6-TM7) from channelrhodopsin-2 (ChR2). The structure (PDB: 3UG9) at 2.3 Å resolution includes the retinal-binding pocket and the cation-conducting pathway, key components of the ChR's molecular architecture [302].

To investigate how lipid-protein H bonds link the membrane to the intracellular H-bond network, the developed algorithm illustrated in Figure 4.44 computes: i) Protein amino acids H bond directly to lipid phosphates, ii) Protein amino acids mediated by direct phosphate H bonds, iii) Lipid phosphates interconnect amino acids by 1 one-water H-bond pathway, iv) Lipid phosphates interconnect amino acids by >1 one-water H-bond pathways.

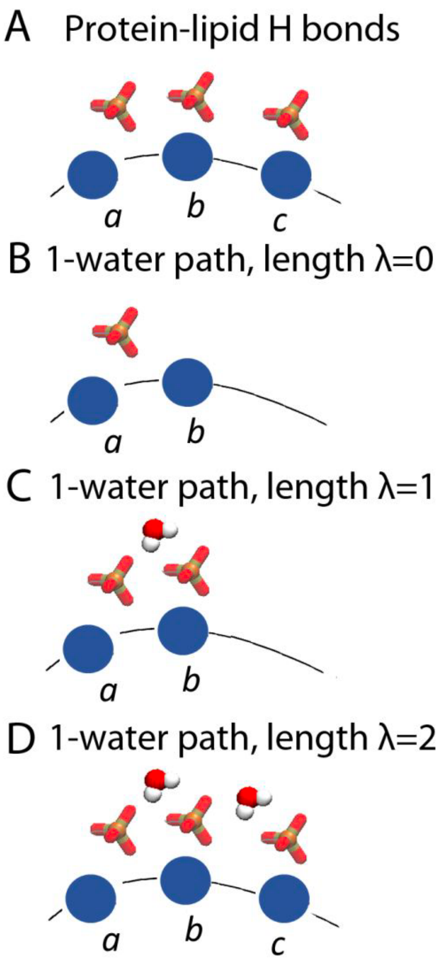

A Protein-lipid H bonds
a
b
c
B 1-water path, length λ=0
a
b
C 1-water path, length λ=1
a
b
D 1-water path, length λ=2
a
b
c

Figure 4.44: Illustration of the lipid-protein H bonds mediated by water molecules as developed in Bridge. Protein groups are depicted as blue spheres. The length λ represents the number of unique phosphate bridges that connect protein groups in presence of one water molecule. (A) Protein groups (a, b, c) are directly H bonded to lipid phosphates. (B) Nodes a and b are mediated by direct lipid H bonds. Here, the length λ is equal to 0. (C) Nodes a and b are interconnected by lipid phosphates mediated by one 1-water H-bond bridge. Here, the length λ is equal to one. (D) Nodes a, b and c are interconnected by lipid phosphates mediated by two 1-water H-bond bridges. Here, the length λ is equal to two. Image is from [231], a study presented in this thesis.

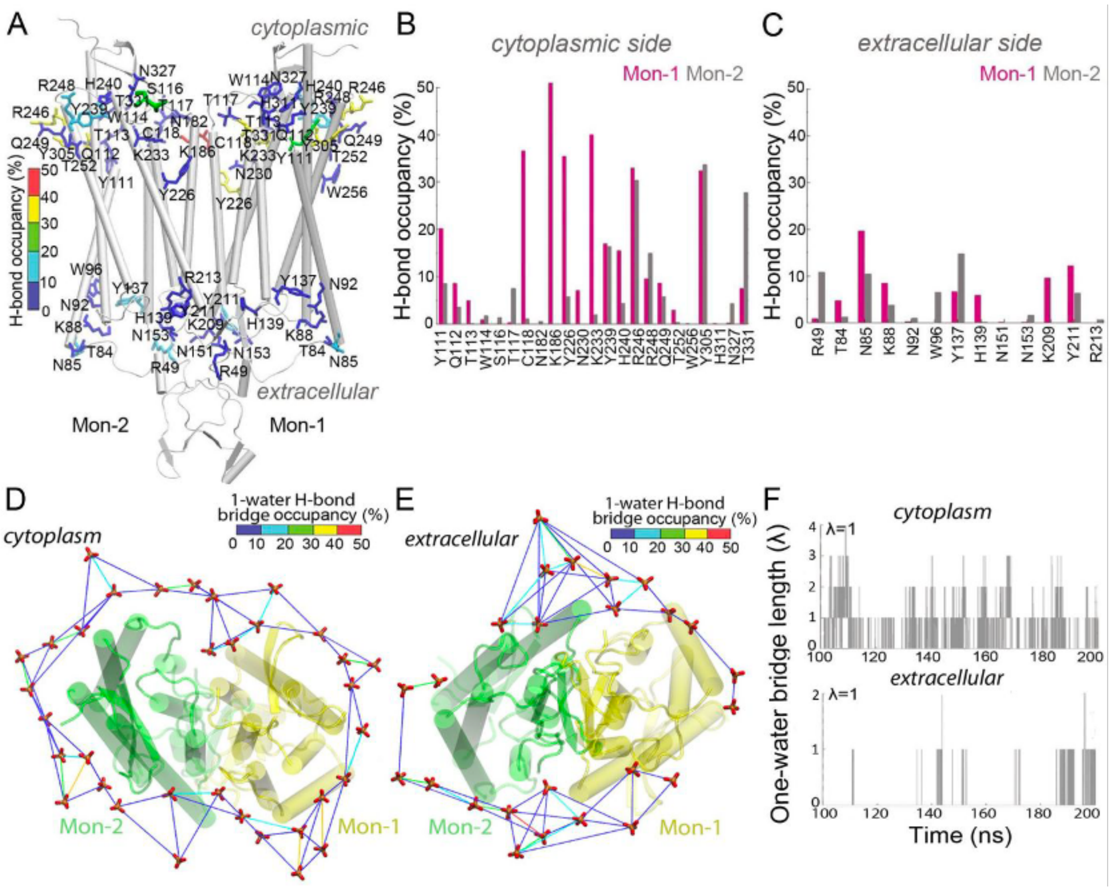


Figure 4.45: C1C2 direct protein-lipid interactions and mediated by water as computed with Bridge. (A) Direct H bonds between protein and lipid phosphate groups. The protein is illustrated as a white cartoon and selected amino acids are color-coded based on the occupancy of the H bonds. Each protein-lipid pair can form multiple H bonds at any given time during the simulation. Here, we show the H bond of each pair with the highest occupancy. (B-C) Histogram representation of the lipid-protein H-bond occupancies of the amino acids shown in panel A both at the cytoplasmic and the extracellular side (panels B and C, respectively). (D-E) 1-water H-bond wires connecting lipids and protein amino acids of C1C2. Lines represent H-bond water bridging colored based on occupancies. (F) Time series of the length λ of 1-water bridges joining lipids directly H bonded to the protein's cytoplasmic and extracellular side. Time series of lengths λ equal to 2 and 3 are depicted in Figure 4.46. Image is from [231], a study presented in this thesis.

Direct H bonding between protein amino acids and lipid phosphates were found for both the cytoplasmic and the extracellular side of the protein (Figure 4.45A-C). Most H bonds are rather dynamic with occupancies<30%. Protein is encircled by 'belts' of lipid phospahtes mediated by water networks on the cytoplasm and extracellular. Water bridges are transient with most of the interaction occupancies <20% (Figure 4.45D-E). There are ~1-2 protein-phosphate bridges with λ=1. Pathways of length λ=2-3 are barely sampled. It is noted that on the cytoplasmic side of the membrane as opposed to the extracellular side, these short protein-phosphate water-mediated bridges are sampled more frequently (Figure 4.45F, Figure 4.46).

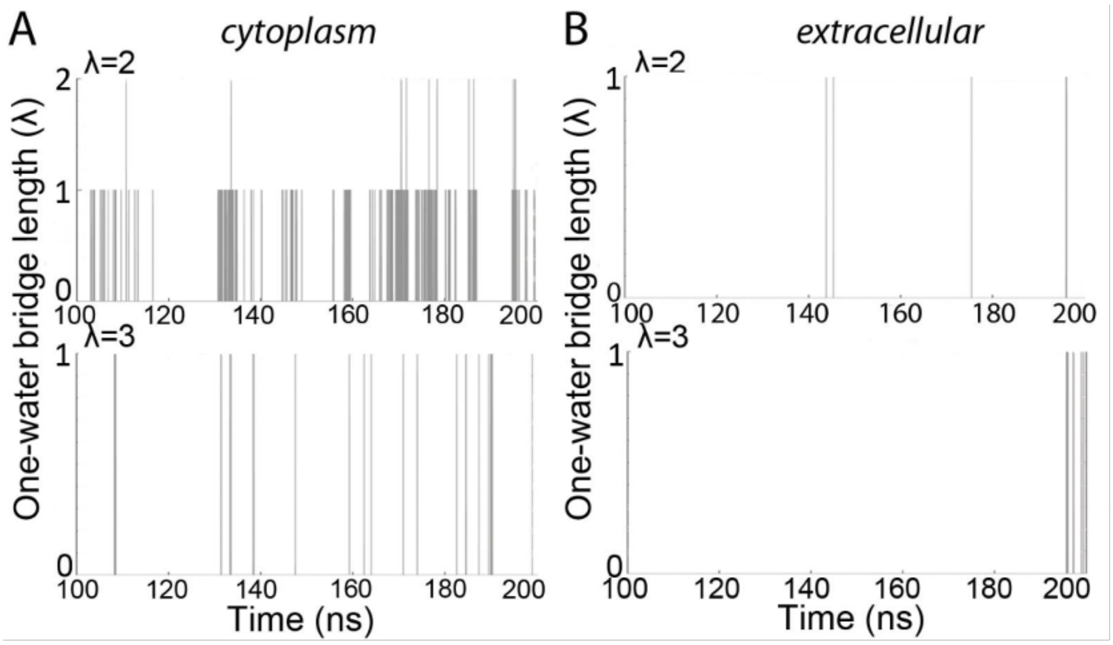


Figure 4.46: C1C2 protein-lipid interactions mediated by water as computed with Bridge. (A-B) Time series of length λ=2-3 of water bridges interconnecting lipids directly H bonded to the protein for both the cytoplasmic (panel A) and the extracellular side (panel B) of C1C2. Image is from [231], a study presented in this thesis.

Protein groups that bind directly to lipids (Figure 4.45A-C) can connect the bilayer to the inner H-bond network of C1C2. At the extracellular side, we note Y137 and H139 directly bound to phosphates (Figure 4.45A, C) which are close to E136, part of a long-distance pathway that connects Schiff base to the bulk via water bridges. Additionaly, E136 is part of a dynamic long pathway that connects N85 of Monomer 1 to the N85 of Monomer 2 via carboxylate groups. Simillarly, at the cytoplasmic side, Y226 (Figure 4.45A, B) of Monomer 1 is linked to Y226 of Monomer 2 via extensive water mediated pathways. The mentioned paths are illustrated in [231].

We suggest that the investigation of protein dynamics using algorithms inspired by graph-theory could shed light to the proton transfer mechanism by detecting dynamic and extensive H-bond interactions between the Schiff base to extracellular carboxylate groups.

Chapter 5

“Each problem that I solved became a rule, which served afterwards to solve other problems.”
Rene Descartes

5 Summary of Developed Methodologies

In this Chapter, I summarize the algorithmic methods that I developed to conduct my research studies. Firstly, I present a methodology for analyzing lipid H-bond dynamics using Tcl [185] within VMD [23] to visualize H-bond networks (Section 3.1). In this methodology, the lipid molecules are considered nodes in the network, and edges are created by direct H bonds between lipids, water-mediated bridges, or ion-mediated bridges. During each simulation step, the pairs of molecules that are H bonded and the distance between them is monitored. Based on this information, binary adjacency or connection matrices are generated, which indicate whether a connection exists between each pair of groups in the system. The adjacency matrices are used to create graphs of lipid headgroups connected by edges of H bonds that are projecting in VMD [23].

To identify dynamic lipid clusters, I introduce the Network Components algorithm based on the Depth-First Search (DFS) algorithm. A cluster of interconnected lipids is composed of a subgroup of nodes and edges and the cluster size is the number of nodes interconnected by H bonds or ion interactions in the case of ion-mediated bridges. The steps of the algorithm are presented in Section 3.2. In summary, the algorithm begins by selecting a starting/root node in the graph. It then proceeds to identify all H-bond paths that originate from that node. To locate all components, present in the graph, the algorithm launches a new search from any node that was not included in a previously discovered component. This process is repeated until all nodes in the graph have been reported as visited. As a result, we create paths of linked nodes in our graphs. From these paths, we count the number of clusters that we discover and list the H-bonded nodes. These results provide us with valuable insights into the behavior of each cluster. To visualize our networks, planar and circular connectivity graphs and illustrations in VMD are employed [23].

To evaluate the dynamics of water H-bonded bridges among lipid headgroups, I propose a computational approach in Section 3.3. This method examines bridges that contain 1 to 5 water molecules, isolating unique water chains of the shortest length. The H-bond occupancy of each bridge is calculated to determine the dynamics of these water bridges. Additionally, I present my implemented method to compute the average lifetime of water molecules H bonded to proteins, lipids, or waters at the first hydration layer of a protein. The calculation uses the residence time correlation function explained in Section 3.4. The correlation function is calculated with a moving time origin for each time span selected for the calculation, giving a computational cost to overcome together with the complexity of interactions during the simulation. The data structures used in our algorithm reduce the computational complexity. I provide a visualization script to illustrate the residence time to the structure of the protein in VMD [23].

In Sections 3.5-3.7, I extend the calculations of H-bond clusters to the H-bond topology pathways, a DFS computational method to describe the geometric configuration of the components of the graphs, defining how the interactions are established between the nodes. Topologies are classified into three main types: linear, star, circular, and more complicated topological schemes derived from those three types. All paths are categorized by length, measured as the linear longest number of edges between a start and an end node. I provide a visualization script to illustrate the H-bond network of lipids colored according to graph topology in VMD [23]. The algorithm's computational efficiency was tested, and the results indicate that scripts can handle larger membrane patches and simulation lengths (Section 3.15).

I introduce the centrality measures, which are vital tools from graph theory to determine the importance of any given node in the network. In addition, I have designed an algorithm (Section 3.8) to derive graphical network representations of H bonds between amino acids. This algorithm also enables the characterization of the strength of the connections between each node and its centrality in the network. Betweenness (*BC*) and degree centralities (*DC*) are employed in our studies. *BC* expresses the extent to which a node is in shortest-distance paths between a pair of nodes regulating the information flow across the network. *DC* is a local measure since it is determined only by the number of H-bonded ‘neighbors’ in the graph. The shortest distance paths are computed by Dijkstra's algorithm, setting a source or an intermediate and an end node. A correlation between cluster size and shape with centrality measures is provided in Section 3.14.

I proposed an algorithm (Section 3.11) that combines water bridges and residence times. The script selects specific amino acid residues based on the duration for which water molecules remain close to them during a simulation and whether they form water H-bond bridges between two groups from distinct protein regions. The algorithm is applied to the SecA protein to analyze H-bond pathways and study its long-distance conformational coupling.

In Section 3.12, I present an algorithm to compare H-bond graphs between proteins of different organisms, wild-type and mutations, or mutations within one network. In addition, another script allows for a graphical comparison of crystal structures and their corresponding MD simulations (Section 3.13). This comparison is crucial, and when combined with centrality measurements, it highlights the importance of MD simulations in providing detailed information about the interactions and conformational dynamics of our systems.

Chapter 6

"Every science begins as philosophy and ends as art."
Will Durant

6 Conclusions and Outlook

The advent of modern computers, capable of running and storing results of MD simulations, has paved the way for new scientific research. Scientists can use simulations to create a theoretical model of a real biomolecular system and gain insight into its behavior. The use of MD simulations can serve a dual purpose in scientific research, by connecting experimental data with theoretical hypotheses and impacting the theory's advancement. However, the use of MD simulations can present significant challenges related to managing big data that need to be addressed.

My research aimed to answer scientific questions derived from extensive datasets obtained from atomistic MD simulations of lipid membrane models and proteins. For that direction, I developed a set of computational tools applying concepts based on graph theory during my doctoral studies. During simulations of a given structure, it is possible to gain important insights into the underlying molecular picture by projecting biomolecular interactions onto a plane, where each molecular group is represented as a dot (node or vertex), and each interaction between pairs of groups is represented as a line (edge) on a graph. This allows for a visual representation of the molecular interactions and their relationships, which can be highly informative in understanding the function of the structure and essential in interpreting data obtained from experiments and computations.

Under physiological conditions, neutral phospholipids compose the outer leaflet of the plasma membrane,while negatively charged phospholipids constitute the inner leaflet [13]. The exposure of the outer leaflet to anionic phospholipids like PS has been associated to several human diseases, including cancer [13, 16, 24, 25].Therefore, negatively charged lipids with different concentrations are useful as model systems as they provide a potential target for cationic proteins or drug molecules that could differentiate between neutral non-cancer cells and negatively charged cancer membranes.

The main objective of the studies of lipid membrane models presented in this thesis [22, 187] is to develop efficient algorithms to investigate the dynamics of the lipid H-bond networks allowing for water and ion interactions and determine their topological configurations during MD simulations.

Our research [22] presented in this thesis, reveals that the surface of the bilayers composed of 4:1 and 5:1 POPC: POPG lipids is marked by a complex and constantly changing network of H bonds between the lipids using our implementation of H-bond graphs. We found that lipids form clusters as shown in [26-28], utilized the Network Component algorithm, and their headgroups connect through direct H bonds or through the mediation of water and ion bridges within a picosecond timeframe.

Anionic POPG lipids can form clusters made up of 2-3 lipids which are held together by direct H bonds and sodium-mediated bridges (Figure 4.3) [22]. These findings are consistent with a lipid membrane comprising of DPPC and DPPS, as reported in [219]. During our simulations, we found that water bridging played a key role in the interactions between lipids. About 50% of the extracellular and cytoplasmic leaflet lipids engaged in dynamic H bonding through one-water bridges (Figure 4.2). About 20% of the POPG lipids connect via one-water mediated H bonds. The one-water phosphate bridges have a short lifetime, consistent with previous findings demonstrating dynamic H-bond water bridges between DMPC lipids [40]. It is shown that H bonding mediated by one-water bridges forms linear paths of lengths of 1 or 2 bridges with a considerably lower probability of longer linear paths during the simulation (Figure 4.6, Figure 4.7).

We suggest that the dense and dynamic network of water H-bond bridges between the lipid phosphates [220-222], reveal that water molecules near the surface of proteins exhibit longer H-bond lifetimes than those found in bulk water. Future studies of extensive H-bonded bridges between lipids may reveal lateral proton transfer along the membrane [35].

The presence of calcium ions can extend the lifetime of water H-bond bridges between lipid headgroups. Therefore, the slower dynamics of the phosphate/water H-bond networks may be the reason for the decreased self-diffusion of lipids after calcium binding, as shown in [303].

It is possible for lipids to cluster together and create a surface that is conducive to binding proteins to cationic surfaces. Our proposal is that the network of H bonds present at the interface of the lipid headgroup could potentially alter the movement of lipid molecules. This could happen even at a distance away from the site where the protein binds and could be an area of exploration in future studies.

Our algorithms are extended to determine topological configurations of H-bond clusters through DFS exhaustive searches during MD simulations. Clusters are not only depicted but analyzed regarding the number of lipids in each cluster, their geometrical configuration (topology) and the length of each arrangement in both the extracellular and the cytoplasmic side of the membrane models.

The study [187], presented in this thesis, describes the above methodological approach to analyze membrane models containing POPE, POPS, 3:1 and 5:1 POPE:POPG models, as well as the Top6 *E. coli* model as described in [205]. Our research revealed that small linear paths, which connect two lipids through direct H bonds or one-water H-bond bridges, are present in all membrane models. The second most common arrangement of lipids is in the form of a star and linear graph consisting of five and four lipids in the case of direct and water-mediated H bonding, respectively. Circular graphs are often observed when three or four lipids form a direct H bond. However, circular water-mediated pathways are rare, likely due to the limited lifespan of lipid-water H bonds. More complex topologies of lipid arrangements are infrequently encountered where at least four lipids link in a complex of circular, star, and linear networks (Figure 4.16, Figure 4.17).

Eukaryotic cell membranes consist of various key components, one of which is PS lipids [31, 219, 251-253]. In simulations, networks of direct H bonds between POPS lipids have been detected forming clusters of inter-connected lipid molecules as shown in study [187] presented in this thesis. This indicates that regions of POPS lipids coupled via serine interactions could serve as substrates to which an external partner could potentially attach.

Several factors can affect the specific formation of H-bond clusters in a system, such as the size of the membrane patch, the force field that describes the molecules in the system, and the simulation length. However, these factors do not compromise the accuracy of our system's results. We tested our approach to compute H-bond clusters with varying simulation times and discovered that it delivers consistent results for all membrane models. Moreover, our algorithm performs impressively, as extracting H-bond clusters for ten coordinate snapshots takes less than a second. This suggests that our technique can be utilized in future studies to provide precise and efficient findings, even for longer simulations or larger membrane patches.

Our approach characterizes dynamically how the lipids are distributed in clusters and their lengths, and could serve as a model for future experimental investigations into lipid H-bond clustering including interactions between proteins and membranes, lipid-mediated protein interactions [256-258], and lateral proton transfer events in membranes [259-261].

We introduce centrality measures; *BC* expresses the extent to which a node is in pathways between a pair of vertices. Such measures are important in networks where a given node can regulate the information flow across the graph. *DC* gives a picture of the local connectivity of nodes. Our algorithms are extended to study the conformational dynamics of the SecA protein motor [102], presented in this thesis.

A key finding enabled by the graph-based approach developed as part of this doctoral thesis is that there are long-distance pathways that interconnect the nucleotide-binding pocket and the pre-protein binding site, and mutations at the NBD affect the dynamics of PBD. Previous research by SecA [79] suggests that the movements of NBD2 are influenced by the presence of a signal peptide at the PBD, which aligns with this long-range conformational coupling. Water H bonds connecting protein groups facilitate the communication between different regions (as depicted in Figure 4.30, Figure 4.33). Figure 4.32A illustrates that the surface of SecA is heterogeneous and intricate concerning the movements of the surrounding water molecules, which aligns with the findings from research on the PsbO component of photosystem II, another soluble protein [220].

Our centrality analysis revealed significant results regarding protein sites that potentially act as hubs in the network by being involved in multiple pathways (refer to Figure 4.31). This means that the connectivity of the entire network depends on these vital nodes. The results align with the protein structure of SecA, the stability of NBD, and the flexibility of PBD. Additionally, we identified essential nodes that contribute to the entire network's connectivity, with some of them being involved in long-distance pathways from NBD to the PBD region.

Our highly robust algorithms can analyze the dense H-bond networks present in various protein complexes. The study of dynamic H-bond graphs is particularly interesting for proteins that involve changes in protonation states [282, 283]. In the future, the ATP-bound models for *B. subtilis* SecA described in our study [102] can be used to investigate the ATP hydrolysis process and allosteric conformational coupling of SecA.

We expanded our focus to *E. coli* SecA structures, including MD simulations of one monomer and two dimers presented in [99, 100], studies presented in this thesis. In the monomeric SecA, we noticed a significant movement and reorientation of PBD towards NBD2 (as shown in Figure 4.34). During the simulation, the two dimers of SecA remained in the Wide-Open state. Our calculations revealed the plasticity of the PBD region and the long-distance communication between the PBD and NBD2. The changes occurring in the PBD is shown to have a profound impact on the entire interaction network in the protein. We also found two pathways connecting gate2 to the PBD, one through the stem and another through the NBD2/PBD interface (as shown in Figure 4.37).

Our algorithms are applied to the SARS-COV-2 protein S crystal structures [194].Our findings suggest that when the spike protein S transitions from a closed to an open or pre-fusion state, H bonds are rearranged, resulting in a conformational change in the open and loss of symmetry in the pre-fusion state. The strong affinity between the spike protein S and ACE2 is due to extensive H-bond clustering in the PBD-ACE2 complex. Our research has identified N501 as a critical residue in the H-bond network that connects the spike protein S to ACE2. In a new variant of COVID-19, this network has mutated.

Our methodologies are also applied to protein VASA [301], a DEAD-box enzyme involved in the cell cycle with ATP- and RNA-binding sites and explore the conformational coupling between the two binding sites. We suggest that our algorithms can give a versatile tool for future studies to compute networks and detect changes while mutations are inserted.

Lastly, my contribution to [231] Channelrhodopsin's C1C2 lipid-protein H-bond molecular dynamics reveals that protein is surrounded by lipids mediated by short and dynamic water wires on the cytoplasm and extracellular. We suggest that the investigation of protein dynamics using algorithms inspired by graph theory could shed light on the proton transfer mechanism by detecting dynamic and extensive H-bond interactions between the Schiff base and extracellular carboxylate groups.

Our software tools are designed to generate and analyze graphs for multiple types of membrane models and protein systems. They are not limited to studying only H-bond dynamics, but can also investigate other interactions, such as hydrophobic interactions. The applied methodologies and visualizations of H bonds in molecular structures provide a powerful tool for understanding the functionality of complex biological networks derived from simulations.

Bibliography

1. Urban, S., J.R. Lee, and M. Freeman, *Drosophila rhomboid-1 defines a family of putative intramembrane serine proteases.* Cell, 2001. **107**(2): p. 173-182.
2. Bordi, F., C. Cametti, and A. Naglieri, *Ionic transport in lipid bilayer membranes.* Biophysical journal, 1998. **74**(3): p. 1358-1370.
3. Bankaitis, V.A., C.J. Mousley, and G. Schaaf, *The Sec14 superfamiliy and mechanisms for crosstalk between lipid metabolism and lipid signaling.* Trends Biochem. Sci., 2009. **35**: p. 150-160.
4. van Klompenburg, W., et al., *Anionic phospholipids are determinants of membrane protein topology.* EMBO J., 1997. **16**: p. 4261-4266.
5. Alami, M., et al., *Nanodiscs reveal the interaction between the SecYEG channel and its cytosolic partner SecA.* EMBO J., 2007. **26**: p. 1995-2004.
6. König, S., et al., *Hydration dependence of chain dynamics and local diffusion in L-alpha-dipalmitoylphosphtidylcholine multilayers studied by incoherent quasi-elastic neutron scattering.* Biophysical journal, 1995. **68**(5): p. 1871-1880.
7. Watson, H., *Biological membranes.* Essays in biochemistry, 2015. **59**: p. 43-69.
8. *Parts of the cell.* . September 5, 2021; Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cell.
9. Singer, S.J. and G.L. Nicolson, *The Fluid Mosaic Model of the Structure of Cell Membranes: Cell membranes are viewed as two-dimensional solutions of oriented globular proteins and lipids.* Science, 1972. **175**(4023): p. 720-731.
10. Lee, A.G., *Lipid-protein interactions in biological membranes: a structural perspective.* Biochim. Biophys. Acta, 2003. **1612**: p. 1-40.
11. Petrov, A.G., *The lyotropic state of matter: molecular physics and living matter physics*. 1999: CRC Press.
12. Petrov, A., *Flexoelectricity of lyotropics and biomembranes.* Il Nuovo Cimento D, 1984. **3**(1): p. 174-192.
13. Bevers, E., P. Comfurius, and R. Zwaal, *Regulatory mechanisms in maintenance and modulation of transmembrane lipid asymmetry: pathophysiological implications.* Lupus, 1996. **5**(5): p. 480-487.
14. Bondar, A.-N., *Biophysical mechanism of rhomboid proteolysis: setting a foundation for therapeutics.* Seminars Cell Dev. Biol., 2016. **60**: p. 46-51.
15. Chaurio, R.A., et al., *Phospholipids: key players in apoptosis and immune regulation.* Molecules, 2009. **14**(12): p. 4892-4914.
16. Zwaal, R.F.A., P. Comfurius, and E.M. Bevers, *Surface exposure of phosphatidylserine in pathological cells.* Cell. Mol. Life Sci., 2005. **62**: p. 971-988.

17. *Fluid mosaic model of a cell membrane.* . September 4, 2021; Available from: https://en.wikipedia.org/wiki/Fluid_mosaic_model.

18. Chen, W., et al., *Determination of the main phase transition temperature of phospholipids by nanoplasmonic sensing.* Scientific reports, 2018. **8**(1): p. 1-11.

19. Balali-Mood, K., T.A. Horroun, and J.P. Bradshaw, *Molecular dynamics simulations of a mixed DOPC/DOPG bilayer.* European Physics Journal E, 2003. **12**: p. S135-S140.

20. Bradshaw, J.P., S.M. Davies, and T. Hauss, *Interaction of substance P with phospholipid bilayers: a neutron diffraction study.* Biophysical journal, 1998. **75**(2): p. 889-895.

21. Chiu, S.-W., et al., *Combined Monte Carlo and molecular dynamics simulation of fully hydrated dioleyl and palmitoyl-oleyl phosphatidylcholine lipid bilayers.* Biophysical Journal, 1999. **77**(5): p. 2462-2469.

22. Karathanou, K. and A.-N. Bondar, *Dynamic water hydrogen-bond networks at the interface of a lipid membrane containing palmitoyl-oleoyl phosphatidylglycerol.* Journal of Membrane Biology, 2018.

23. Humphrey, W., W. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J. Mol. Graph., 1996. **14**: p. 33-38.

24. Ran, S., A. Downes, and P.E. Thorpe, *Increased exposure of anionic phospholipids on the surface of tumor blood vessels.* Cancer Res., 2002. **62**(21): p. 6132-40.

25. Riedl, S., et al., *In search of a novel target - phosphatidylserine exposed by non-apoptotic tumor cells and metastases of malignacies with poor treatment efficacy.* Biochim. Biophys. Acta, 2011. **1808**: p. 2638-2645.

26. Dicko, A., H. Bourque, and M. Pézolet, *Study by infrared spectroscopy of the conformation of dipalmitoylphospatidylglycerol monolayers at the air-water interface and transferred on solid substrates.* Chem. Phys. Lipids, 1998. **96**: p. 125-139.

27. Zhao, W., et al., *Atomic-scale structure and electrostatics of anionic palmitoyloleoylphosphatidylglycerol lipid bilayers with Na^+ counterions.* Biophys. J., 2007. **92**: p. 1114-1124.

28. Koldsø, H., et al., *Lipid clustering correlates with membrane curvature as revealed by moleculad simulations of complex lipid bilayers.* PLoS Computational Biology, 2014. **10**: p. e1003911.

29. Berkowitz, M.L., D.L. Bostik, and S. Pandit, *Aqueous solution next to phospholipid membrane interfaces: Insights from simulations.* Chem. Rev., 2006. **106**: p. 1527-1539.

30. Hübner, W. and A. Blume, *Interactions at the lipid-water interface.* Chemistry and Physics of Lipids, 1998. **96**: p. 99-123.

31. Mukhopadhyay, P., L. Monticelli, and D.P. Tieleman, *Molecular dynamics simulations of a palmitoyl-oleoyl phosphatidylserine bilayer with Na^+ counterions and NaCl.* Biophys. J., 2004. **86**: p. 1601-1609.

32. Nagle, J.F. and S. Tristram-Nagle, *Structure of lipid bilayers.* Biochim. Biophys. Acta, 2000. **1429**: p. 159-195.

33. Smondyrev, A.M. and M.L. Berkowitz, *Structure of dipalmidoylphosphatidylcholine/cholesterol bilayer at low and high cholesterol concentrations: molecular dynamics simulation.* Biophys. J., 1999. **77**: p. 2075-2089.

34. Wiener, M.C. and S.H. White, *Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of X-ray and neutron diffraction data. III. Complete structure.* Biophys. J., 1992. **61**: p. 434-447.
35. Lopez, C.F., et al., *Hydrogen binding structure and dynamics of water at the dimyristoylphosphatidylcholine lipid bilayer surface from a molecular dynamics simulation.* J. Phys. Chem. B, 2004. **108**: p. 6603-6610.
36. Bhide, S.Y. and M.L. Berkowitz, *Structure and dynamics of water at the interface with phospholipid bilayers.* J. Chem. Phys., 2005. **123**: p. 224702.
37. Petrache, H.I., et al., *Structure and fluctuations of charged phosphatidylserine bilayers in the absence of salt.* Biophys. J., 2004. **86**: p. 1574-1586.
38. Marrink, S.J., M.L. Berkowitz, and H.J.C. Berendsen, *Molecular dynamics simulation of a membrane/water interface: The ordering of water and its relation to the hydration force.* Langmuir, 1993. **9**: p. 3122-3131.
39. Pasenkiewicz-Gierula, M., et al., *Computer modelling studies of the bilayer/water interface.* Biochim. Biophys. Acta, 2016. **1858**: p. 2305-2321.
40. Pasenkiewicz-Gierula, M., et al., *Hydrogen bonding of water to phosphatidylcholine in the membrane as studied by a molecular dynamics simulation: location, geometry, and lipid-bridging via hydrogen-bonded water.* J. Phys. Chem. A, 1997. **101**: p. 3677-3691.
41. Tielrooij, K.J., et al., *Dielectric relaxation dynamics of water in model membranes probed by terahertz spectroscopy.* Biophys. J., 2009. **97**: p. 2484-2492.
42. Volkov, V.V., D.J. Palmer, and R. Righini, *Heterogeneity of water at the phospholipid membrane interface.* J. Phys. Chem. B, 2007. **111**: p. 1377-1383.
43. *The general structure of an α-amino acid.* September 5, 2021; Available from: https://socratic.org/questions/how-does-ph-affect-amino-acid-structure
44. Smith, A., *Nucleic acids to amino acids: DNA specifies protein.* Nature Education, 2008. **1**(1): p. 126.
45. Akashi, H. and T. Gojobori, *Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis.* Proceedings of the National Academy of Sciences, 2002. **99**(6): p. 3695-3700.
46. Cobb, M., *60 years ago, Francis Crick changed the logic of biology.* PLoS biology, 2017. **15**(9): p. e2003243.
47. Alberts, B., et al., *Protein function*, in *Molecular Biology of the Cell. 4th edition*. 2002, Garland Science.
48. Rehman, I., C.C. Kerndt, and S. Botelho, *Biochemistry, Tertiary Protein Structure.* 2017.
49. *Central dogma of biology.* May 15, 2022; Available from: https://byjus.com/biology/central-dogma-inheritance-mechanism/.
50. *The peptide bond formation.* September 5, 2021; Available from: https://en.wikipedia.org/wiki/Peptide_bond.
51. Eisenberg, D., *The discovery of the α-helix and β-sheet, the principal structural features of proteins.* Proceedings of the National Academy of Sciences, 2003. **100**(20): p. 11207-11210.
52. Egli, M., *Diffraction techniques in structural biology.* Current protocols in nucleic acid chemistry, 2016. **65**(1): p. 7.13. 1-7.13. 41.
53. Zwanzig, R., A. Szabo, and B. Bagchi, *Levinthal's paradox.* Proceedings of the National Academy of Sciences, 1992. **89**(1): p. 20-22.

54. Moult, J., et al., *A large-scale experiment to assess protein structure prediction methods*. 1995, Wiley Online Library. p. ii-iv.
55. Tunyasuvunakool, K., et al., *Highly accurate protein structure prediction for the human proteome.* Nature, 2021. **596**(7873): p. 590-596.
56. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
57. Enciso, M., *Hydrogen bond models for the simulation of protein folding and aggregation.* arXiv preprint arXiv:1208.2177, 2012.
58. Von Heijne, G., *The membrane protein universe: what's out there and why bother?* Journal of internal medicine, 2007. **261**(6): p. 543-557.
59. Lodish, H., et al., *Intracellular ion environment and membrane electric potential*, in *Molecular Cell Biology. 4th edition*. 2000, WH Freeman.
60. Jones, D.T., *Improving the accuracy of transmembrane protein topology prediction using evolutionary information.* Bioinformatics, 2007. **23**(5): p. 538-544.
61. Viklund, H. and A. Elofsson, *OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar.* Bioinformatics, 2008. **24**(15): p. 1662-1668.
62. Nugent, T. and D.T. Jones, *Transmembrane protein topology prediction using support vector machines.* BMC bioinformatics, 2009. **10**(1): p. 1-11.
63. Hegedűs, T., et al., *AlphaFold2 transmembrane protein structure prediction shines.* bioRxiv, 2021.
64. Alberts, B., et al., *Principles of membrane transport*, in *Molecular Biology of the Cell. 4th edition*. 2002, Garland Science.
65. *Types of membrane proteins*. September 6, 2021; Available from: https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3A_Biochemistry_Free_For_All_(Ahern_Rajagopal_and_Tan)/03%3A_Membranes/3.01%3A_Basic_Concepts_in_Membranes.
66. Puthenveetil, R. and O. Vinogradova, *Solution NMR: A powerful tool for structural and functional studies of membrane proteins in reconstituted environments.* Journal of Biological Chemistry, 2019. **294**(44): p. 15914-15931.
67. Holland, I.B., *Translocation of bacterial proteins—an overview.* Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 2004. **1694**(1-3): p. 5-16.
68. von Heijne, G., *The signal peptide.* The Journal of membrane biology, 1990. **115**(3): p. 195-201.
69. Park, E. and T.A. Rapoport, *Mechanisms of Sec61/SecY-mediated protein translocation across membranes.* Ann. Rev. Biophys., 2012. **41**: p. 21-40.
70. Chatzi, K.E., et al., *SecA-mediated targeting and translocation of secretory proteins.* Biochim. Biophys. Acta, 2014. **1843**: p. 1466-1474.
71. Zhang, X. and S.-o. Shan, *Fidelity of cotranslational protein targeting by the signal recognition particle.* Annual review of biophysics, 2014. **43**: p. 381-408.
72. *Bacteria protein secretion through SecYE translocon*. September 6, 2021; Available from: https://basicmedicalkey.com/posttranslational-targeting-of-proteins/.
73. Nyathi, Y., B.M. Wilkinson, and M.R. Pool, *Co-translational targeting and translocation of proteins to the endoplasmic reticulum.* Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 2013. **1833**(11): p. 2392-2402.

74. Nithianantham, S. and B.H. Shilton, *Analysis of the isolated SecA DEAD motor suggests a mechanism for chemical-mechanical coupling.* J. Mol. Biol., 2008. **383**: p. 380-389.
75. Warshel, A., *Computer simulations of enzyme catalysis: methods, progress, and insights.* Annual review of biophysics and biomolecular structure, 2003. **32**(1): p. 425-443.
76. Tsirigotaki, A., et al., *Protein export through the bacterial Sec pathway.* Nature Reviews Microbiology, 2017. **15**: p. 21-36.
77. Papanikou, E., et al., *Identification of the preprotein binding domain of SecA.* J. Biol. Chem., 2005. **280**: p. 43209-43217.
78. Milenkovic, S. and A.-N. Bondar, *Mechanism of conformational coupling in SecA: key role of hydrogen-bonding networks and water interactions.* Biochim. Biophys. Acta, 2016. **1858**: p. 374-385.
79. Milenkovic, S. and A.-N. Bondar, *Motions of the SecA protein motor bound to signal peptide: Insights from molecular dynamics simulations.* Biochim. Biophys. Acta, 2018. **1860**: p. 416-427.
80. Karathanou, K. and A.-N. Bondar, *Dynamic hydrogen-bond networks in bacterial protein secretion.* FEMS microbiology letters, 2018. **365**(13): p. fny124.
81. Erlandson, K.J., et al., *A role for the two-helix finger of the SecA ATPase in protein translocation.* Nature, 2008. **455**: p. 984-987.
82. Zimmer, J., Y. Nam, and T.A. Rapoport, *Structure of a complex of the ATPase SecA and the protein-translocation channel.* Nature, 2008. **455**: p. 936-943.
83. Whitehouse, S., et al., *Mobility of the SecA 2-helix-finger is not essential for polypeptide translocation via the SecYEG complex.* J. Cell Biol., 2012. **199**: p. 919-929.
84. Bauer, B.W., et al., *A "Push and slide" mechanism allows sequence-insensitive translocation of secretory proteins by the SecA ATPase.* Cell, 2014. **157**: p. 1416-1429.
85. Allen, W.A., et al., *Two-way communication between SecY and SecA suggests a Brownian ratchet mechanism for protein translocation.* eLife, 2016. **5**: p. e15598.
86. Das, S., et al., *The variable subdomain of Escherichia coli SecA functions to regulate SecA ATPase activity and ADP release.* J. Bacteriol., 2012. **194**: p. 2205-2213.
87. Hunt, J.F., et al., *Nucleotide control of interdomain interactions in the conformational reaction cycle of SecA.* Science, 2002. **297**(5589): p. 2018-2026.
88. Osborne, A.R., W.A. Clemons Jr., and T.A. Rapoport, *A large conformational change of the translocation ATPase SecA.* Proc. Natl. Acad. Sci. USA, 2004. **101**: p. 10937-10942.
89. Zimmer, J., W. Li, and T.A. Rapoport, *A novel dimer interface and conformational changes revealed by an X-ray structure of B. subtilis SecA.* J. Mol. Biol., 2006. **364**: p. 259-265.
90. Papanikolau, Y., et al., *Structure of dimeric SecA, the Escherichia coli preprotein translocase motor.* J. Mol. Biol., 2007. **366**: p. 1545-1557.
91. Sharma, V., et al., *Crystal structure of Mycobacterium tuberculosis SecA, a preprotein translocating ATPase.* Proc. Natl. Acad. Sci. USA, 2003. **1000**: p. 2243-2248.

92. Zimmer, J. and T.A. Rapoport, *Conformational flexibility and peptide interaction of the translocation ATPase SecA.* J. Mol. Biol., 2009. **394**: p. 606-612.
93. Chen, Y., et al., *Conformational changes of the clamp of the protein translocation ATPase SecA.* J. Mol. Biol., 2015. **427**: p. 2348-2359.
94. Vassylyev, D.G., et al., *Crystal structure of the translocation ATPase SecA from Thermus thermophilus reveals a parallel, head-to-head dimer.* J. Mol. Biol., 2006. **364**: p. 248-258.
95. Gelis, I., et al., *Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR.* Cell, 2007. **131**: p. 756-769.
96. Schiebel, E., et al., *$\Delta\mu_H$+ and ATP function at different steps of the catalytic cycle of preprotein translocase.* Cell, 1991. **64**: p. 927-939.
97. van der Wolk, J.P.W., J.G. de Wit, and A.J.M. Driessen, *The catalytic cycle of the Escherichia coli SecA ATPase comprises two distinct preprotein translocation events.* EMBO J., 1997. **16**: p. 7297-7304.
98. Berg, B.V.D., et al., *X-ray structure of a protein-conducting channel.* nature, 2004. **427**(6969): p. 36-44.
99. Krishnamurthy, S., et al., *A nexus of intrinsic dynamics underlies translocase priming.* Structure, 2021. **29**(8): p. 846-858. e7.
100. Krishnamurthy, S., et al., *Preproteins couple the intrinsic dynamics of SecA to its ATPase cycle to translocate via a catch and release mechanism.* Cell Reports, 2022. **38**(6): p. 110346.
101. Gold, V.A.M., et al., *The dynamic action of SecA during the initiation of protein translocation.* Biochem. J., 2013. **449**: p. 695-705.
102. Karathanou, K. and A.N. Bondar, *Using graphs of dynamic hydrogen-bond networks to dissect conformational coupling in a protein motor.* J. Chem. Inf. Model., 2019. **59**: p. 1882-1896.
103. Karathanou, K. and A.-N. Bondar, *Dynamic hydrogen bonds in bacterial protein secretion.* FEMS Microbiol. Lett., 2018. **365**: p. fny124.
104. Arunan, E., et al., *Defining the hydrogen bond: An account (IUPAC Technical Report).* Pure and Applied Chemistry, 2011. **83**(8): p. 1619-1636.
105. Arunan, E., et al., *Definition of the hydrogen bond (IUPAC Recommendations 2011).* Pure and applied chemistry, 2011. **83**(8): p. 1637-1641.
106. Pimentel, G., *The Hydrogen Bond Franklin Classics*. 2018.
107. Jeffrey, G.A. and G.A. Jeffrey, *An introduction to hydrogen bonding*. Vol. 12. 1997: Oxford university press New York.
108. Jeffrey, G.A. and W. Saenger, *Hydrogen bonding in biological structures*. 2012: Springer Science & Business Media.
109. McNaught, A. and A. Wilkinson, *IUPAC Compendium of Chemical Terminology, 2nd edn.(the "Gold Book") Blackwell Scientific Publications*. 1997, Oxford.
110. Hubbard, R.E. and M.K. Haider, *Hydrogen bonds in proteins: role and strength.* eLS, 2010.
111. Espinosa, E., et al., *Topological analysis of the electron density in hydrogen bonds.* Acta Crystallographica Section B: Structural Science, 1999. **55**(4): p. 563-572.
112. Karle, I., *Hydrogen bonds in molecular assemblies of natural, synthetic and 'designer'peptides.* Journal of molecular structure, 1999. **474**(1-3): p. 103-112.

113. *Hydrogen bonds in secondary protein structure*. May 17, 2022; Available from: https://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/73-translation/protein-structure.html.

114. DW, H., *Computer simulation methods in theoretical physics*. 1986, Springer-Verlag.

115. Alder, B.J. and T.E. Wainwright, *Phase transition for a hard sphere system.* The Journal of chemical physics, 1957. **27**(5): p. 1208-1209.

116. Alder, B.J. and T.E. Wainwright, *Studies in molecular dynamics. I. General method.* The Journal of Chemical Physics, 1959. **31**(2): p. 459-466.

117. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of folded proteins.* nature, 1977. **267**(5612): p. 585-590.

118. Mortier, J., et al., *The impact of molecular dynamics on drug design: applications for the characterization of ligand–macromolecule complexes.* Drug Discovery Today, 2015. **20**(6): p. 686-702.

119. Dror, R.O., et al., *Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations.* Journal of General Physiology, 2010. **135**(6): p. 555-562.

120. Pierce, L.C., et al., *Routine access to millisecond time scale events with accelerated molecular dynamics.* Journal of chemical theory and computation, 2012. **8**(9): p. 2997-3002.

121. Ferruz, N., et al., *Multibody cofactor and substrate molecular recognition in the myo-inositol monophosphatase enzyme.* Scientific reports, 2016. **6**(1): p. 1-10.

122. Bowman, G.R., X. Huang, and V.S. Pande, *Using generalized ensemble simulations and Markov state models to identify conformational states.* Methods, 2009. **49**(2): p. 197-201.

123. Noé, F. and S. Fischer, *Transition networks for modeling the kinetics of conformational change in macromolecules.* Current opinion in structural biology, 2008. **18**(2): p. 154-162.

124. Harvey, M.J., G. Giupponi, and G.D. Fabritiis, *ACEMD: accelerating biomolecular dynamics in the microsecond time scale.* Journal of chemical theory and computation, 2009. **5**(6): p. 1632-1639.

125. Páll, S. and B. Hess, *A flexible algorithm for calculating pair interactions on SIMD architectures.* Computer Physics Communications, 2013. **184**(12): p. 2641-2650.

126. Müller, M., K. Katsov, and M. Schick, *Coarse-grained models and collective phenomena in membranes: Computer simulation of membrane fusion.* Journal of Polymer Science Part B: Polymer Physics, 2003. **41**(13): p. 1441-1450.

127. Shelley, J.C. and M.Y. Shelley, *Computer simulation of surfactant solutions.* Current opinion in colloid & interface science, 2000. **5**(1-2): p. 101-110.

128. Waidyasooriya, H.M., M. Hariyama, and K. Kasahara, *An FPGA Accelerator for Molecular Dynamics Simulation Using OpenCL.* Int. J. Networked Distributed Comput., 2017. **5**(1): p. 52-61.

129. Cornell, W.D., et al., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.* Journal of the American Chemical Society, 1995. **117**(19): p. 5179-5197.

130. Wang, J., P. Cieplak, and P.A. Kollman, *How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?* Journal of computational chemistry, 2000. **21**(12): p. 1049-1074.

131. Cramer, C.J., *Essentials of computational chemistry: theories and models*. 2013: John Wiley & Sons.

132. Sikorska, C. and N. Gaston, *Modified Lennard-Jones potentials for nanoscale atoms.* Journal of Computational Chemistry, 2020. **41**(22): p. 1985-2000.

133. Allen, M.P. and D.J. Tildesley, *Computer simulation of liquids*. 2017: Oxford university press.

134. Ewald, P.P., *Die Berechnung optischer und elektrostatischer Gitterpotentiale.* Annalen der physik, 1921. **369**(3): p. 253-287.

135. Leach, A.R. and A.R. Leach, *Molecular modelling: principles and applications*. 2001: Pearson education.

136. Frenkel, D. and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*. 2nd ed. 2002: Academic Press, Inc.

137. Katiyar, R.S. and P.K. Jha, *Molecular simulations in drug delivery: Opportunities and challenges.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2018. **8**(4): p. e1358.

138. Verlet, L., *Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules.* Physical review, 1967. **159**(1): p. 98.

139. Swope, W.C., et al., *A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters.* The Journal of chemical physics, 1982. **76**(1): p. 637-649.

140. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* J. Comput. Chem, 2005. **26**: p. 1781-1802.

141. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: an N x log(N) method for Ewald sums in large systems.* J. Chem. Phys., 1993. **98**: p. 10089-10092.

142. Grubmüller, H., et al., *Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions.* Mol. Simul., 1991. **6**: p. 121-142.

143. Tuckermann, M., B.J. Berne, and G.J. Martyna, *Reversible multiple time scale molecular dynamics.* J. Chem. Phys., 1992. **97**: p. 1990-2001.

144. Nosé, S., *A unified formulation of the constant temperature molecular dynamics methods.* The Journal of chemical physics, 1984. **81**(1): p. 511-519.

145. Andersen, H.C., *Molecular dynamics simulations at constant pressure and/or temperature.* The Journal of chemical physics, 1980. **72**(4): p. 2384-2393.

146. Ryckaert, J.-P., G. Ciccotti, and H.J. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of computational physics, 1977. **23**(3): p. 327-341.

147. Hoover, W.G. and B.L. Holian, *Kinetic moments method for the canonical ensemble distribution.* Physics Letters A, 1996. **211**(5): p. 253-257.

148. Alder, B., W. Hoover, and D. Young, *Studies in molecular dynamics. V. High-density equation of state and entropy for hard disks and spheres.* The Journal of Chemical Physics, 1968. **49**(8): p. 3688-3696.

149. Martyna, G.J., D.J. Tobias, and M.L. Klein, *Constant-pressure molecular-dynamics algorithms.* J. Chem. Phys., 1994. **101**: p. 4177-4189.

150. Feller, S.E., et al., *Constant pressure molecular dynamics simulation: The Langevin piston method.* J. Chem. Phys. , 1995. **103**: p. 4613-4621.

151. Shields, R., *Cultural topology: The seven bridges of Königsburg, 1736.* Theory, Culture & Society, 2012. **29**(4-5): p. 43-57.

152. *A small example network with eight vertices and ten edges*. September 12, 2021; Available from: https://en.wikipedia.org/wiki/Network_theory.
153. Wilson, R.J., *Introduction to graph theory*. 1979: Pearson Education India.
154. Van Steen, M., *Graph theory and complex networks.* An introduction, 2010. **144**.
155. Gross, J.L., J. Yellen, and M. Anderson, *Graph theory and its applications*. 2018: Chapman and Hall/CRC.
156. Fionda, V. and L. Palopoli, *Biological network querying techniques: analysis and comparison.* Journal of Computational Biology, 2011. **18**(4): p. 595-625.
157. *Bipartite graph*. September 13, 2021; Available from: https://www.javatpoint.com/graph-theory-types-of-graphs.
158. *Path or linear graph*. September 12, 2021; Available from: https://mathworld.wolfram.com/PathGraph.html.
159. *Star graph*. September 13, 2021; Available from: https://www.tutorialspoint.com/graph_theory/types_of_graphs.htm.
160. *Cycle graph*. September 13, 2021; Available from: https://www.tutorialspoint.com/graph_theory/types_of_graphs.htm.
161. Zhang, C., et al., *Taxonomy-aware collaborative denoising autoencoder for personalized recommendation.* Applied Intelligence, 2019. **49**(6): p. 2101-2118.
162. Klunder, G. and H. Post, *The shortest path problem on large-scale real-road networks.* Networks: An International Journal, 2006. **48**(4): p. 182-194.
163. Shang, S., et al. *Finding the most accessible locations: reverse path nearest neighbor query in road networks*. in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2011.
164. Shin, S.-H., et al. *Efficient shortest path finding of k-nearest neighbor objects in road network databases*. in *Proceedings of the 2010 ACM Symposium on Applied Computing*. 2010.
165. Dijkstra, E.W., *A note on two problems in connexion with graphs.* Numerische mathematik, 1959. **1**(1): p. 269-271.
166. Russell, S. and P. Norvig, *Artificial intelligence: a modern approach.* 2002.
167. Vaira, G. and O. Kurasova, *Parallel bidirectional Dijkstra's shortest path algorithm.* Databases and Information Systems VI, Frontiers in Artificial Intelligence and Applications, 2011. **224**: p. 422-435.
168. Cherkassky, B.V., A.V. Goldberg, and T. Radzik, *Shortest paths algorithms: Theory and experimental evaluation.* Mathematical programming, 1996. **73**(2): p. 129-174.
169. *Shortest-distance path*. September 13, 2021; Available from: https://www.sci.unich.it/~francesc/teaching/network/geodesic.html.
170. Kumar, N., et al., *Geospatial school bus routing.* The International Journal of Engineering and Science (IJES) Vol, 2014. **3**: p. 80-84.
171. Jordan, C., *Sur les assemblages de lignes.* 1869.
172. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-42.
173. Ahn, W.-k., et al., *Causal status as a determinant of feature centrality.* Cognitive Psychology, 2000. **41**(4): p. 361-416.
174. Jackson, M.O., *Social and economic networks*. 2010: Princeton university press.

175. Lorenzen, M. and K.V. Andersen, *Centrality and creativity: Does Richard Florida's creative class offer new insights into urban hierarchy?* Economic Geography, 2009. **85**(4): p. 363-390.

176. Gomez, D., et al., *Centrality and power in social networks: a game theoretic approach.* Mathematical Social Sciences, 2003. **46**(1): p. 27-54.

177. Anthonisse, J.M., *The rush in a directed graph.* Stichting Mathematisch Centrum. Mathematische Besliskunde, 1971(BN 9/71).

178. Freeman, L.C.J.S., *A set of measures of centrality based on betweenness.* 1977: p. 35-41.

179. Freeman, L.C.J.S.n., *Centrality in social networks conceptual clarification.* 1978. **1**(3): p. 215-239.

180. Cormen, T.H., et al., *Introduction to algorithms, third edition.* Massachusetts Institute of Technology, 2009.

181. *Connected components in graphs*. September 13, 2021; Available from: https://en.wikipedia.org/wiki/Connectivity_(graph_theory).

182. Friedman, R., et al., *Understanding conformational dynamics of complex lipid mixtures relevant to biology.* The Journal of membrane biology, 2018. **251**(5): p. 609-631.

183. Sengupta, N., S. Jaud, and D.J. Tobias, *Hydration dynamics in a partially denatured ensemble of the globular protein human a-lactalbumin investigated with molecular dynamics simulations.* Biophys. J., 2008. **95**: p. 5257-5267.

184. *Time correlation function. September 14,2021; Available from: https://www.scribd.com/document/273349383/Chem860-09-L9-pdf*

185. Welch, B.B., K. Jones, and J. Hobbs, *Practical Programming in Tcl/Tk.* 2003: Prentice Hall Professional.

186. The MathWorks, I., *MATLAB.* Natick, Massachusetts, United States, 2017.

187. Karathanou, K. and A.-N. Bondar, *Algorithm to catalogue topologies of dynamic lipid hydrogen-bond networks.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2022: p. 183859.

188. Karathanou, K. and A.-N. Bondar, *Using graphs of dynamic hydrogen-bond networks to dissect conformational coupling in a protein motor.* Journal of Chemical Information and Modeling, 2019. **59**(5): p. 1882-1896.

189. *Graph Data Structure 4. Dijkstra's Shortest Path Algorithm*. June 19, 2022; Available from: https://www.youtube.com/watch?v=pVfj6mxhdMw.

190. Freeman, L.C., *A set of measures of centrality based on betweenness.* Sociometry, 1977: p. 35-41.

191. Brandes, U.J.o.m.s., *A faster algorithm for betweenness centrality.* 2001. **25**(2): p. 163-177.

192. Salavaty, A., M. Ramialison, and P.D. Currie, *Integrated value of influence: an integrative method for the identification of the most influential nodes within networks.* Patterns, 2020. **1**(5): p. 100052.

193. Lazaratos, M., K. Karathanou, and A.-N. Bondar, *Graphs of dynamic H-bond networks: from model proteins to protein complexes in cell signaling.* Current Opinion in Structural Biology, 2020. **64**: p. 79-87.

194. Karathanou, K., et al., *A graph-based approach identifies dynamic H-bond communication networks in spike protein S of SARS-CoV-2.* Journal of structural biology, 2020. **212**(2): p. 107617.

195. Urban, S. and M.S. Wolfe, *Reconstitution of intramembrane proteolysis in vitro reveals that pure rhomboid is sufficient for catalysis and specificity.* Proc. Natl. Acad. Sci. USA, 2005. **102**: p. 1883-1888.

196. Connor, J., C.C. Pak, and A.J. Schroit, *Exposure of phosphatidylserine in the outer leaflet of human red blood cells.* J. Biol. Chem., 1994. **269**: p. 2399-2404.

197. Zwaal, R.F.A. and A.J. Schroit, *Pathophysiological implications of membrane phospholipd asymmetry in blood cells.* The Journal of the American Society of Hemathology, 1997. **89**: p. 1121-1132.

198. Noble, J.M., T.H. Thomas, and G.A. Ford, *Effect of age on plasma membrane asymmetry and membrane fluidity in human leokocytes and platelets.* Journal of Gerontology: Medical Sciences, 1999. **1999**: p. M60-M606.

199. Peetla, C., A. Stine, and V. Labhasetwar, *Biophysical interactions with model lipid membranes: applications in drug discovery and drug delivery.* Mol. Pharm., 2009. **6**: p. 1264-1276.

200. Ran, S. and P.E. Thorpe, *Phosphatidylserine is a marker of tumor vasculature and a potential target for cancer imaging and therapy.* International Journal of Radiation Oncology* Biology* Physics, 2002. **54**(5): p. 1479-1484.

201. Janosi, L. and A.A. Gorfe, *Simulating POPC and POPC/POPG bilayers: conserved packing and altered surface reactivity.* J. Chem. Theor. Comput., 2010. **6**: p. 3267-3273.

202. Binder, H. and O. Zschörnig, *The effect of metal cations on the phase behavior and hydration characteristics of phospholipid membranes.* Chem. Phys. Lipids. , 2002. **115**(1-2): p. 39-61.

203. Pandit, S.A. and M.L. Berkowitz, *Molecular dynamics simulations of dipalmitoylphosphatidylserine bilayer with Na$^+$ counterions.* Biophys. J., 2002. **82**: p. 1818-1827.

204. Murzyn, K., T. Róg, and M. Pasenkiewicz-Gierula, *Phosphatidylethanolamine-phosphatidylglycerol bilayer as a model of the inner bacterial membrane.* Biophys. J., 2005. **88**: p. 1091-1103.

205. Pandit, K.R. and J.B. Klauda, *Membrane models of E. coli containing cyclic moieties in the aliphatic lipid chain.* Biochim. Biophys. Acta, 2012. **1818**: p. 1205-1210.

206. Jo, S., et al., *CHARMM-GUI: a web-based graphical user interface for CHARMM.* Journal of Computational Chemistry, 2008. **29**: p. 1859-1865.

207. Wu, E.L., et al., *CHARMM-GUI Membrane Builder toward realistic biological membrane simulations.* J. Comput. Chem, 2014. **35**: p. 1997-2004.

208. Brooks, B.R., et al., *CHARMM: a program for macromolecular energy, minimization, and dynamics calculations.* J. Comput. Chem, 1983. **4**: p. 187-217.

209. Feller, S.E. and A.D. MacKerell Jr., *An improved empirical potential energy function for molecular simulations of phospholipids.* J. Phys. Chem. B, 2000. **104**: p. 7510-7515.

210. Klauda, J.B., et al., *Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types.* J. Phys. Chem. B, 2010. **114**: p. 7830-7843.

211. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* J. Chem. Phys., 1983. **79**: p. 926-935.

212. Essmann, U., et al., *A smooth particle mesh Ewald method.* J. Chem. Phys., 1995. **103**: p. 8577-8593.

213. Ryckaert, J.-P., G. Ciccotti, and H.J.C. Berendsen, *Numerical integration of the Cartesian equations of motion of a system with constraints. Molecular dynamics of n-alkanes.* J. Comput. Phys., 1977. **23**: p. 327-341.

214. Kalé, L., et al., *NAMD2: greater scalability for parallel molecular dynamics.* J. Comput. Phys., 1999. **151**: p. 283-312.

215. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of computational chemistry, 2005. **26**(16): p. 1781-1802.

216. Pasenkiewicz-Gierula, M., et al., *Charge pairing of headgroups in phosphatidylcholine membranes: A molecular dynamics simulation study.* Biophys. J., 1999. **76**: p. 1228–1240.

217. Muegge, I. and E.-W. Knapp, *Residence times and lateral diffusion of water at protein surfaces: application to BPTI.* J. Phys. Chem., 1995. **99**: p. 1371-1374.

218. Elmore, D.E., *Molecular dynamics simulation of a phosphatidylglycerol membrane.* FEBS Lett., 2006. **580**: p. 144-148.

219. Pandit, S.A., D. Bostick, and M.L. Berkowitz, *Mixed bilayer containing dipalmitoylphosphatidylcholine and dipalmitoylphosphatidylserine: Lipid compensation, ion binding, and electrostatics.* Biophys. J., 2003. **85**: p. 3120-3131.

220. Lorch, S., et al., *Dynamic carboxylate/water networks on the surface of the PsbO subunit of photosystem II.* J. Phys. Chem. B, 2015. **119**: p. 12172-12181.

221. Ebbinghaus, S., et al., *An extended dynamical hydration shell around proteins.* Proc. Natl. Acad. Sci. USA, 2007. **104**: p. 20479-20752.

222. Makarov, V.A., et al., *Diffusion of solvent around biomolecular solutes: a molecular dynamics simulation study.* Biophys. J., 1998. **75**: p. 150-158.

223. Artymiuk, P.J., et al., *Structural resemblance between the families of bacterial signal-transduction proteins and of G proteins revealed by graph theoretical techniques.* Protein Eng. Des. Sel., 1990. **4**(1): p. 39-43.

224. Samudrala, R. and J. Moult, *A graph-theoretic algorithm for comparative modeling of protein structure.* J. Mol. Biol., 1998. **279**(1): p. 287-302.

225. Kannan, N. and S. Vishveshwara, *Identification of side-chain clusters in protein structures by a graph spectral method.* J. Mol. Biol., 1999. **292**(2): p. 441-64.

226. Vendruscolo, M., et al., *Small-world view of the amino acids that play a key role in protein folding.* Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 2002. **65**(6 Pt 1): p. 061910.

227. Aftabuddin, M. and S. Kundu, *Hydrophobic, hydrophilic, and charged amino acid networks within protein.* Biophys J., 2007. **93**(1): p. 225-31.

228. Bikadi, Z., L. Demko, and E. Hazai, *Functional and structural characterization of a protein based on analysis of its hydrogen bonding network by hydrogen bonding plot.* Arch. Biochem. Biophys., 2007. **461**(2): p. 225-34.

229. Jacobs, D.J., et al., *Protein Flexibility Predictions Using Graph Theory.* Proteins: Struct., Funct., Genet., 2001. **44**: p. 150-165.

230. Rylance, G.J., et al., *Topographical complexity of multidimensional energy landscapes.* Proc. Natl. Acad. Sci. U S A, 2006. **103**(49): p. 18551–18555.

231. Siemers, M., et al., *Bridge: A Graph-Based Algorithm to Analyze Dynamic H-Bond Networks in Membrane Proteins.* J. Chem. Theory Comput., 2019. **15**(12): p. 6781-6798.

232. Srivastava, A. and A. Debnath, *Hydration dynamics of a lipid membrane: Hydrogen bond networks and lipid-lipid associations.* J. Chem. Phys. , 2018. **148**(9): p. 094901.

233. Szczelina, R., et al., *Network of lipid interconnections at the interfaces of galactolipid and phospholipid bilayers.* J. Mol. Liq., 2019. **298**.

234. Böde, C., et al., *Network analysis of protein dynamics.* FEBS Lett., 2007. **581**(15): p. 2776-82.

235. Bhamare, D. and P. Suryawanshi, *Review on Reliable Pattern Recognition with Machine Learning Techniques.* Fuzzy Inf. Eng., 2018. **10**(3): p. 362–377.

236. Wazarkar, S. and B.N. Keshavamurthy, *A survey on image data analysis through clustering techniques for real world applications.* J. Vis. Commun. Image Represent., 2018. **55**: p. 596-626.

237. Runkler, T.A. and J.C. Bezdek, *Web mining with relational clustering.* INT. J. APPROX. REASON, 2003. **32**(2-3): p. 217-236.

238. Sharan, R. and R. Shamir, *CLICK: a clustering algorithm with applications to gene expression analysis.* Proc. Int. Conf. Intell. Syst. Mol. Biol., 2000. **8**: p. 307-16.

239. Liu, Q., et al., *Robust MST-Based Clustering Algorithm.* Neural. Comput., 2018. **30**(6): p. 1624-1646.

240. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res., 2003. **13**(11): p. 2498-504.

241. Gowers, R.J., et al., *MDAnalysis: A Phyton package for the rapid analysis of molecular dynamics simulations.* S. Benthall and S. Rostrup, Editors, Proceedings of the 15th Phyton in Science Conference, Austin, TX, 2016 SciPy, 2016: p. 102-109.

242. Michaud-Agrawal, N., E.J. Denning, and T.B. Woolf, *MDAnalysis: A toolkit for the analysis of molecular dynamics simulations.* J. Comput. Chem, 2011. **32**: p. 2319-2327.

243. Pyrkova, D.V., et al., *Dynamic clustering of lipids in hydrated two-component membranes: results of computer modeling and putative biological impact.* Journal of Biomolecular Structure and Dynamics, 2013. **31**(1): p. 87-95.

244. Zhuang, X., et al., *An extensive simulation study of lipid bilayer properties with different head groups, acyl chain lengths, and chain saturations.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2016. **1858**(12): p. 3093-3104.

245. Pitman, M.C., et al., *Molecular dynamics investigation of dynamical properties of phosphatidylethanolamine lipid bilayers.* The Journal of chemical physics, 2005. **122**(24): p. 244715.

246. Lyu, Y., et al., *Characterization of interactions between curcumin and different types of lipid bilayers by molecular dynamics simulation.* J. Phys. Chem. B, 2018. **122**: p. 2341-2354.

247. Bondar, A.-N., *Mechanisms by which lipids influence conformational dynamics of the GlpG intramembrane protease.* The journal of physical chemistry B, 2019. **123**(19): p. 4159-4172.

248. Shahane, G., et al., *Physical properties of model biological lipid bilayers: insights from all-atom molecular dynamics simulations.* Journal of molecular modeling, 2019. **25**(3): p. 1-13.

249. Guixà-González, R., et al., *MEMBPLUGIN: studying membrane complexity in VMD.* 2014. **30**(10): p. 1478-1480.

250. Pan, J., et al., *The molecular structure of a phosphatidylserine bilayer determined by scattering and molecular dynamics simulations.* Soft matter, 2014. **10**(21): p. 3716-3725.

251. Yeung, T., et al., *Membrane phosphatidylserine regulates surface charge and protein localization.* Science, 2008. **319**: p. 210-213.

252. Leventis, P.A. and S. Grinstein, *The distribution and function of phosphatidylserine in cellular membranes.* Annual review of biophysics, 2010. **39**(1): p. 407-427.

253. Venable, R.M., et al., *Simulations of anionic lipid membranes: development of interaction-specific ion parameters and validation using NMR data.* J. Phys. Chem. B, 2013. **117**: p. 10183-10192.

254. Jain, H., K. Karathanou, and A.-N. Bondar, *Graph-Based Analyses of Dynamic Water-Mediated Hydrogen-Bond Networks in Phosphatidylserine: Cholesterol Membranes.* Biomolecules, 2023. **13**(8): p. 1238.

255. Bandara, A., et al., *Exploring the structure and stability of cholesterol dimer formation in multicomponent lipid bilayers.* Journal of computational chemistry, 2017. **38**(16): p. 1479-1488.

256. Jost, P.C. and O.H. Griffith, *The lipid-protein interface in biological membranes.* Annals of the New York Academy of Sciences, 1980. **348**: p. 391-407.

257. Malhotra, K., et al., *Cardiolipin mediates membrane and channel interactions of the mitochondrial TIM23 protein import complex receptor Tim50.* Science advances, 2017. **3**(9): p. e1700532.

258. Seinen, A.-B., et al., *Cellular dynamics of the SecA ATPase at the single molecule level.* Scientific reports, 2021. **11**(1): p. 1-16.

259. Öjemyr, L.N., et al., *Functional interactions between membrane-bound transporters and membranes.* Proceedings of the National Academy of Sciences, 2010. **107**(36): p. 15763-15767.

260. Nilsson, T., et al., *Lipid-mediated protein-protein interactions modulate respiration-driven ATP synthesis.* Scientific Reports, 2016. **6**: p. 1-11.

261. Prats, M., J. Teissie, and J.F. Tocanne, *Lateral proton conduction at lipid-water interfaces and its implications for the chemiosmotic-coupling hypothesis.* Nature, 1986. **322**: p. 756-758.

262. Tanner, N.K. and P. Linder, *DExD/H box RNA helicases: from generic motors to specific dissociation functions.* Molecular cell, 2001. **8**(2): p. 251-262.

263. Henn, A., M.J. Bradley, and E.M. de la Cruz, *ATP utilization and RNA conformational rearrangement by DEAD-box proteins.* Annu. Rev. Biophys., 2012. **41**: p. 247-267.

264. Pyle, A.M., *Translocation and unwinding mechanisms of RNA and DNA helicases.* Annu. Rev. Biophys., 2008. **37**: p. 317-336.

265. Tsirigotaki, A., et al., *Protein export through the bacterial Sec pathway.* Nature Reviews Microbiology, 2017. **15**(1): p. 21-36.

266. Karamanou, S., et al., *A molecular switch in SecA protein couples ATP hydrolysis to protein translocation.* Mol. Microbiol., 1999. **34**: p. 1133-1145.

267. Gelis, I., et al., *Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR.* Cell, 2007. **131**(4): p. 756-769.

268. Milenkovic, S. and A.-N. Bondar, *Mechanism of conformational coupling in SecA: key role of hydrogen-bonding networks and water interactions.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2016. **1858**(2): p. 374-385.

269. Osborne, A.R., W.M. Clemons, and T.A. Rapoport, *A large conformational change of the translocation ATPase SecA.* Proceedings of the National Academy of Sciences, 2004. **101**(30): p. 10937-10942.

270. Zimmer, J. and T.A. Rapoport, *Conformational flexibility and peptide interaction of the translocation ATPase SecA.* Journal of molecular biology, 2009. **394**(4): p. 606-612.

271. Vassylyev, D.G., et al., *Crystal structure of the translocation ATPase SecA from Thermus thermophilus reveals a parallel, head-to-head dimer.* Journal of molecular biology, 2006. **364**(3): p. 248-258.

272. Mitchell, C. and D. Oliver, *Two distinct ATP-binding domains are needed to promote protein export by Escherichia coli SecA ATPase.* Mol. Microbiol., 1993. **10**: p. 483-497.

273. Robson, A., et al., *A large conformational change couples the ATP binding site of SecA to the SecY protein channel.* J. Mol. Biol., 2007. **374**: p. 965-976.

274. Karamanou, S., et al., *Preprotein-controlled catalysis in the helicase motor of SecA.* EMBO J., 2007. **26**: p. 2904-2914.

275. Emsley, P., et al., *Features and development of Coot.* Acta Crystallographica Section D: Biological Crystallography, 2010. **66**(4): p. 486-501.

276. MacKerell Jr., A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* J. Phys. Chem. B, 1998. **102**: p. 3586-3616.

277. MacKerell Jr., A.D. and N. Banavali, *All-atom empirical force field for nucleic acids: 2) Application to molecular dynamics simulations of DNA and RNA in solution.* J. Comput. Chem, 2000. **21**: p. 105-120.

278. Foloppe, N. and A.D. MacKerell Jr., *All-atom empirical force field for nucleic acids: 1) Parameter optimization based on small molecule and consdensed phase macromolecular target data.* J. Comput. Chem, 2000. **21**: p. 86-104.

279. Pavelites, J.J., et al., *A molecular mechanics force field for NAD+ NADH, and the pyrophosphate groups of nucleotides.* Journal of computational chemistry, 1997. **18**(2): p. 221-239.

280. Gouridis, G., et al., *Quaternary dynamics of the SecA motor drive translocase catalysis.* Mol Cell, 2013. **52**(5): p. 655-66.

281. Gouridis, G., et al., *Quaternary dynamics of the SecA motor drive translocase catalysis.* Molecular Cell, 2013. **52**: p. 655-666.

282. Guerra, F., et al., *Dynamics of long-distance hydrogen-bond networks in photosystem II.* J. Phys. Chem. B, 2018. **122**: p. 4625-4641.

283. del Val, C., et al., *Channelrhodopsins - a bioinformatics perspective.* Biochim. Biophys. Acta Bioenergetics, 2014. **1837**: p. 643-655.

284. Kiani, F.A. and S. Fischer, *Catalytic strategy used by the myosin motor to hydrolyze ATP.* Proc. Natl. Acad. Sci. USA, 2014. **111**: p. 2947-2956.

285. Hoffmann, M., et al., *SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor.* cell, 2020. **181**(2): p. 271-280. e8.

286. Li, W., et al., *Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus.* Nature, 2003. **426**(6965): p. 450-454.

287. Xiao, X., et al., *The SARS-CoV S glycoprotein: expression and functional characterization.* Biochemical and biophysical research communications, 2003. **312**(4): p. 1159-1164.

288. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* nature, 2020. **579**(7798): p. 270-273.

289. Babcock, G.J., et al., *Amino acids 270 to 510 of the severe acute respiratory syndrome coronavirus spike protein are required for interaction with receptor.* Journal of virology, 2004. **78**(9): p. 4552-4560.

290. Graham, R.L. and R.S. Baric, *Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission.* Journal of virology, 2010. **84**(7): p. 3134-3146.

291. Walls, A.C., et al., *Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein.* Cell, 2020. **181**(2): p. 281-292. e6.

292. Wrapp, D., et al., *Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.* Science, 2020. **367**(6483): p. 1260-1263.

293. Chakraborti, S., et al., *The SARS coronavirus S glycoprotein receptor binding domain: fine mapping and functional characterization.* Virology journal, 2005. **2**(1): p. 1-10.

294. Tegally, H., et al., *Detection of a SARS-CoV-2 variant of concern in South Africa.* Nature, 2021. **592**(7854): p. 438-443.

295. Linder, P., et al., *Birth of the DEAD box.* Nature, 1989. **337**(6203): p. 121-122.

296. Vrontou, E. and A. Economou, *Structure and function of SecA, the preprotein translocase nanomotor.* Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 2004. **1694**(1-3): p. 67-80.

297. Squeglia, F., et al., *Host DDX helicases as possible SARS-CoV-2 proviral factors: a structural overview of their hijacking through multiple viral proteins.* Frontiers in chemistry, 2020. **8**: p. 602162.

298. Rao, S. and T. Mahmoudi, *DEAD-ly affairs: the roles of DEAD-box proteins on HIV-1 viral RNA metabolism.* Frontiers in Cell and Developmental Biology, 2022. **10**: p. 917599.

299. Yajima, M. and G.M. Wessel, *The multiple hats of Vasa: its functions in the germline and in cell cycle progression.* Molecular reproduction and development, 2011. **78**(10-11): p. 861-867.

300. Bondar, A.-N., H. Mishima, and Y. Okamoto, *Molecular movie of nucleotide binding to a motor protein.* Biochimica et Biophysica Acta (BBA)-General Subjects, 2020. **1864**(10): p. 129654.

301. KARATHANOU, K. and A.-N. BONDAR, *CONFORMATIONAL COUPLING VIA HYDROGEN-BONDING IN THE DEAD-BOX PROTEIN VASA.* Rev. Roum. Chim, 2021. **66**(10-11): p. 845-853.

302. Kato, H.E., et al., *Crystal structure of the channelrhodopsin light-gated cation channel.* Nature, 2012. **482**: p. 369-374.

303. Böckmann, R.A. and H. Grubmüller, *Multistep binding of divalent cations to phospholipid bilayers: a molecular dynamics study.* Angew. Chem. Int. Ed., 2004. **43**: p. 1021-1024.

Selbstständigkeitserklärung

Name: Karathanou

Vorname: Konstantina

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht. Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Datum: 01.11.2023   Unterschrift: ________________________