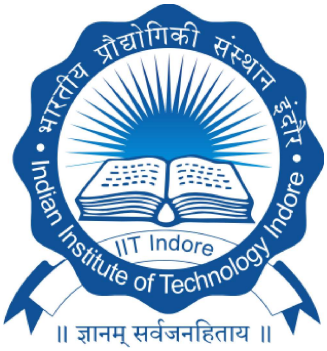
Structural Insights into HCV Glycoprotein E1-E2 Interactions: A Biomolecular Modelling Approach

M.Sc. Thesis

By

Dibyanka Dalai

(Roll No. 2303171004)



DEPARTMENT OF BIOSCIENCES AND BIOMEDICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

MAY 2025

Structural Insights into HCV Glycoprotein E1-E2 Interactions: A Biomolecular Modelling Approach

A THESIS

Submitted in partial fulfillment of the
requirements for the award of the degree
of
Master of Science

by

Dibyanka Dalai

DEPARTMENT OF BIOSCIENCES AND BIOMEDICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
MAY 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **"Structural Insights into HCV Glycoprotein E1-E2 Interactions: A Biomolecular Modelling Approach"** in the partial fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE** and submitted in the **DEPARTMENT OF BIOSCIENCES AND BIOMEDICAL ENGINEERING, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from July 2024 to May 2025 under the supervision of **Dr. Parimal Kar**, Associate Professor, Department of Biosciences and Biomedical Engineering, Indian Institute of Technology Indore.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Dibyanka Dalai 6/5/25.
Signature of the student with date
(Dibyanka Dalai)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Parimal Kar
Signature of the Supervisor
(Dr. Parimal Kar)

Dibyanka Dalai has successfully given her M.Sc. Oral Examination held on **6th May 2025**.

Parimal Kar
Signature of Supervisor of MSc Thesis
Date: 06/05/2025

P.V. Kodgire
Convener, DPGC
Date: 06/05/2025

Signature of PSPC Member
Date:

Signature of PSPC Member
Date:

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the Head of the Department and my supervisor, **Dr. Parimal Kar**, for his invaluable guidance, support, and encouragement throughout the course of this research. His expertise, mentorship, and constructive feedback have been instrumental in shaping this thesis and my academic journey.

I am also thankful to the DPGC Convenor, **Dr. Prashant Kodgire**, and the faculty members at the Department of Biosciences and Biomedical Engineering at the Indian Institute of Technology Indore for their academic insights and continuous support.

I am grateful to all my colleagues at the Computational Biophysics Lab**, Ms. Subhasmita Mahapatra, Mr. Suman Koirala, Mr. Kapil Ursal, Mr. Sunanda Samanta, Mr. Uday Kumar, Ms. Anamika Shukla, Mr. Pradeep Kumar and Ms. Ahana Chakraborty**. With their constant understanding and support, the first tryst with research was fruitful and intellectually stimulating.

My heartfelt appreciation goes to my parents, **Mr. Niranjan Dalai** and **Mrs. Amarabati Dalai** for their unwavering love, understanding, and encouragement. Their encouragement and belief in my abilities have been a constant source of strength and motivation.

I would like to extend my gratitude to my friends for their companionship, encouragement, and occasional distractions during moments of stress. Finally, I am grateful to IIT Indore for providing the necessary resources, facilities, and conducive environment for conducting this research. Thank you to everyone who has contributed in any way to this endeavour. Your support and encouragement have been deeply appreciated.

Yours truly,

Dibyanka Dalai

Dedicated to my Family

Abstract

The Hepatitis C virus is a leading contributor to various liver related diseases such as cirrhosis and liver cancer. It creates considerable challenges for treatment due to its extensive genetic variability. As a result, the virus is able to mutate rapidly and evade the immune system of the host's body, complicating the formulation of effective vaccines. The E1-E2 heterodimer complex is a potential target for therapeutic development as they contain several critical regions that are essential for the viral infection process. However, the dynamic behaviour of the glycoprotein complex is not yet completely understood. In the present study titled as the "**Structural Insights into HCV Glycoprotein E1-E2 Interactions: A Biomolecular Modelling Approach**", the main objective is to better understand how the hepatitis C virus (HCV) envelope glycoproteins E1 and E2 interact with each other at the molecular level using advanced biomolecular modeling techniques. In our research, we explored the dynamic behaviour of these two glycoproteins along with their critical regions. Here, we investigated the structural dynamics of E1 and E2 through molecular simulations of two distinct systems: an apo form consisting solely of the proteins, and a complex form containing the proteins along with two N-linked glycans positioned at their interface. These specific glycans were included based on their known roles in promoting glycoprotein binding, enhancing structural stability, and supporting proper folding. Gaussian accelerated molecular dynamics (GaMD) was employed for 1 microsecond in triplicate to observe the conformational variation in both apo and complex structures. By comparing the simulation outcomes of both systems, we aim to uncover the structural and dynamic changes induced by the presence of these glycans, providing deeper insight into their role in stabilizing the E1-E2 interaction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AMBER	Assisted Model Building with Energy Refinement
CLDN1	Claudin-1
CTR	C-terminal loop Region
DAAs	Direct-acting antivirals
DCCM	Dynamic Cross Correlation Matrix
EGF	Epidermal growth factor
HCV	Hepatitis C virus
GAFF2	General AMBER Force Field 2
GalNac	N-acetylglucosamine
GaMD	Gaussian accelerated molecular dynamics
GBPs	Glycan binding proteins
IRES	Internal Ribosome Entry Site
MD	Molecular Dynamics
NSP	Non-structural Proteins
NTD	N-terminal domain
OCLN	Occludin
PCA	Principal Component Analysis
PCR	putative fusion peptide (pFP) containing region
PDB	Protein Data Bank
PME	Particle Mesh Ewald
ROG	Radius of Gyration
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SASA	Solvent-accessible Surface Area
SR-BI	Scavenger receptor class B type I
TIP3P	Transferable Intermolecular Potential with 3 points
TGF-β	Transforming growth factor-beta
VR	Variable Regions
WHO	World Health Organization

CHAPTER 1

1. Hepatitis C Virus

1.1 Introduction and background

Hepatitis is the inflammatory condition of liver, can be caused due to a number of factors that includes viruses, genetic disorders, alcohol, drug and chemicals [1]. Hepatitis is primarily caused by viral infections, with several distinct viruses and most common etiological agents including hepatitis A, B, C, D, and E. In this study, we will focus on hepatitis C, which is a major viral causes of liver inflammation. **Hepatitis C** is mainly caused by the hepatitis C virus (HCV), and our goal is to explore the glycoprotein E1-E2 heterodimer complex that is present on the surface envelope of the virus. HCV exhibits substantial genetic variability, currently classified into eight major genotypes and 86 distinct subtypes.

1.1.1 Prevalence of HCV in India and the Global Context

Globally, Hepatitis C virus remains a significant public health concern. The World Health Organization (WHO), gives an estimation of around **50 million** people worldwide with chronic HCV infection. The global prevalence is around **2.5%** of the population [2]. The **Eastern Mediterranean Region** has the highest burden of chronic Hepatitis C, with 12 million cases. Hepatitis C leads to around 399,000 deaths annually, primarily due to complications such as cirrhosis and liver cancer (hepatocellular carcinoma) [3]. The other regions like **South-East Asia** and **Europe** both have 9 million each, **Africa** has 8 million, and **America** has 5 million cases. **Egypt** is commended by WHO for becoming the first country to achieve "**gold tier**" status in eliminating hepatitis C and meeting the criteria to lower new infections and deaths, that can eradicate the epidemic [4].

In **India**, the prevalence of hepatitis C has been estimated around 3.23% [5]. However, the prevalence of this dreadful disease varies across India. The higher rates are observed in the north-eastern region, Punjab, and tribal populations, and lower rates are found in eastern and western parts of the country. Following China, India has the second-highest number of hepatitis B and C cases according to the WHO's 2024 Global Hepatitis Report [6]. In India, genotype 3 is the most prevalent, accounting for approximately 63.85% of cases, with genotype 1 being the second most common at 25.72% [7].

1.2 Comprehensive Overview of HCV

Hepatitis C virus (HCV) is a small positive single stranded RNA virus that specifically affects the liver, and cause liver damage that can progress to cirrhosis and potentially lead to the development of hepatocellular carcinoma.

Figure 1.1 shows HCV is a spherical enveloped virus (55-65 nm in size), a member of **Hepacivirus** genus classified within the **Flaviviridae** family, with a lipid bilayer derived from the host cell membrane [8].

It contains E1 and E2 glycoproteins that are crucial for attachment and fusion with the host cell membrane. Beneath the envelope, the virus has a nucleocapsid composed of the core protein, which forms a protective shell around the viral RNA genome, possessing a 9.6 kb single-stranded positive sense RNA genome.

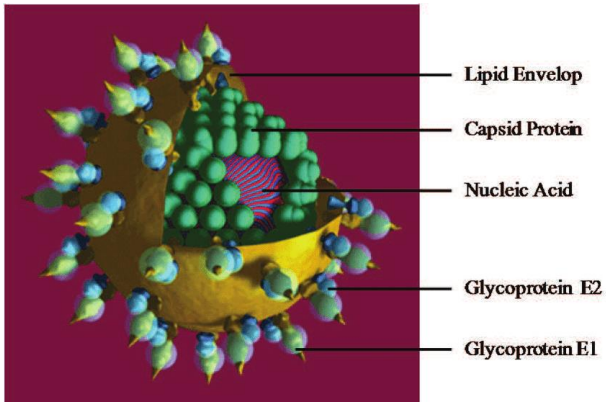


Figure 1.1: Morphological structure of Hepatitis C virus [9].

This genome in **Figure 1.2** contains two most conserved regions - 5' UTR and 3' UTR (Untranslated Region). The 5′ UTR is of approximately 341 nucleotides and the 3′ UTR ranges from about 200 to 235 nucleotides in length. The 5' UTR contains an IRES (Internal Ribosome Entry Site) that allows the virus to translate its RNA without a 5' cap. It is useful for genotype identification. The 3' UTR is involved in packaging the viral genome into new infectious particles, a process known as encapsidation. Additionally, the 3′ UTR influences the stability of viral RNA and modulates its translation efficiency. The genome also has a single long open reading frame encodes a polyprotein of 3,010 amino acids which is cleaved, either during or after translation, into structural proteins (core, E1, and E2) and non-structural proteins (p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B). The core forms the capsid protein of the virus. E1 facilitates membrane fusion and E2 is a major receptor binding protein which interacts with host cell receptor during entry into cell. p7 is an ion channel protein and plays a vital role in the assembly, envelopment and secretion of viral particles.NS2 is important for assembly of virion. NS3 contains protease and helicase. NS4A is a cofactor for NS3 protease activity. NS4B induces the formation of membranous web which is a site for viral replication. NS5A is a phosphoprotein. NS5B is an RNA dependent RNA polymerase which replicates the RNA [10].

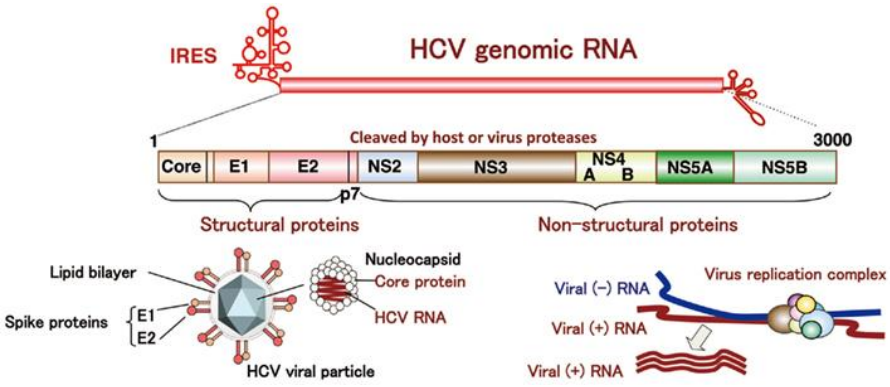


Figure 1.2: Single strand RNA (+) genome (9.6 kb) of HCV [11].

Table 1.1: Constituent proteins of HCV with molecular weight and amino acid residues.

Protein	No. of Amino acids	Molecular weight (kDa)
Core	**177**	**21**
E1	**192**	**35**
E2	**363**	**70**
p7	**63**	**7**
NS2	**217**	**21**
NS3	**631**	**70**
NS4A	**54**	**4**
NS4B	**261**	**27**
NS5A	**448**	**56**
NS5B	**591**	**66**

1.3 Life Cycle of Hepatitis C Virus (HCV)

The Hepatitis C Virus (HCV) life cycle is composed of several interconnected processes that are essential for viral infection, replication, and propagation within the host, as shown below in **Figure 1.3.** These processes include:

Attachment and Entry

HCV initiates infection by binding to the basolateral surface of hepatocytes. The virus interacts with several host cell receptors, including CD81, Claudin-1 (CLDN1), Occludin (OCLN), and scavenger receptor class B type I (SR-BI). This multistep attachment process facilitates the virus's internalization via clathrin-mediated endocytosis. Upon acidification within the endosome, fusion between the viral envelope and the endosomal membrane occurs, leading to the release of the viral genome into the cytoplasm.[12]

Uncoating and Translation

The released positive-sense single-stranded RNA genome serves as a template for translation. The viral RNA contains an internal ribosome entry site (IRES), enabling cap-independent translation of the viral polyprotein [13]. This polyprotein undergoes cleavage by both host and viral proteases, resulting in the formation of 10 structural and non-structural proteins essential for viral replication and assembly.

Replication

Non-structural proteins (NSP) orchestrate the replication of the viral genome. These proteins recruit host cell membranes from the endoplasmic reticulum, forming a specialized structure known as the membranous web. Within this environment, NS5B (RNA-dependent RNA polymerase) synthesizes negative-strand RNA templates from the positive-strand genome, which subsequently serve as templates for the production of new positive-strand genomic RNA.

Assembly and Release

Newly synthesized viral RNA genomes are encapsidated into nucleocapsids near lipid droplets. These nucleocapsids associate with the endoplasmic reticulum, where they acquire their envelope through the secretory pathway and released from the hepatocyte, glycoproteins, E1 and E2. The mature virions are then transported completing the viral life cycle.

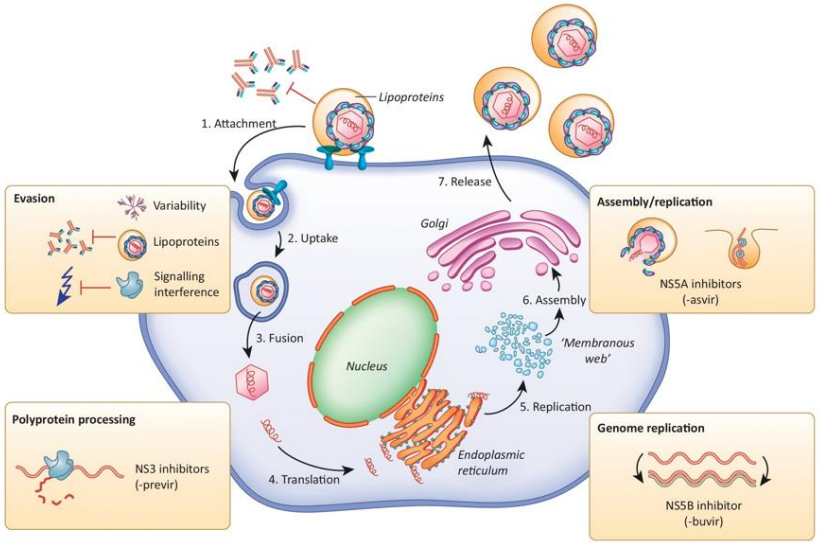


Figure 1.3: Infection Process and Replication Mechanism of HCV [14].

1.4 HCV infection

Acute infection occurs within six months of exposure, typically 2 to 24 weeks after infection. They have symptoms like jaundice, nausea, and abdominal pain in some patients. In most cases, many remain asymptomatic whereas people having symptoms usually recover in 2 to 12 weeks. Acute infection frequently develops into chronic infection.

Chronic infection is a long-term infection that follows the acute phase, with symptoms such as jaundice, easy bruising and bleeding, and dark-colored urine. Over time, chronic HCV can result in liver damage, cirrhosis, liver failure, and may even progress to liver cancer [15].

The progression from a healthy liver to cirrhosis and ultimately to hepatocellular carcinoma (HCC) given in **Figure 1.4.**

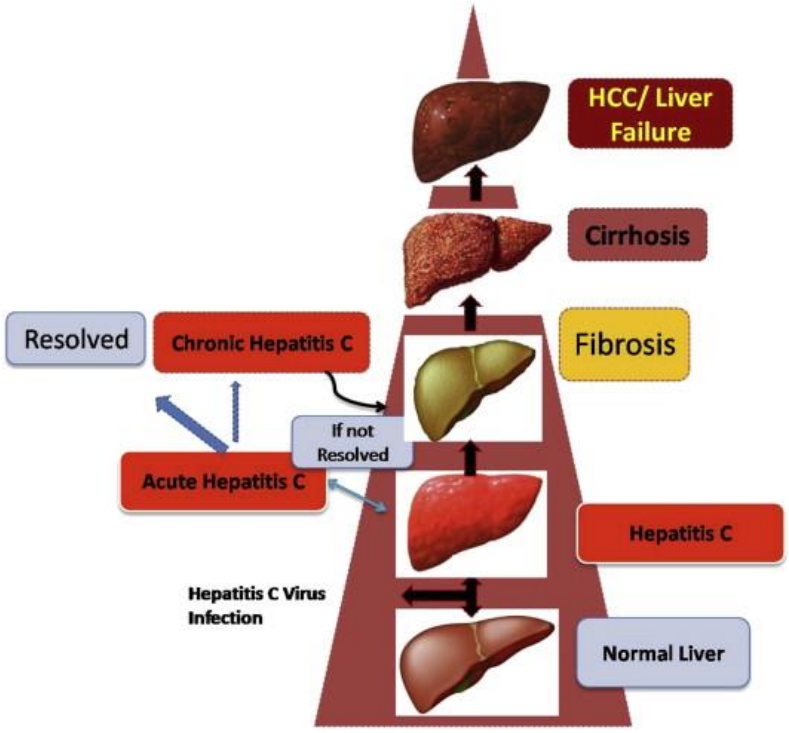


Figure 1.4: Course of illness with Hepatitis C [16].

1.4.1 HCV Transmission and Diagnosis

HCV is primarily transmitted through contact with infectious blood and body fluids. The most common routes include sharing needles, unsafe medical practices, needle stick injury, through improper sterilization techniques like tattooing dyes, piercing and transfusions with contaminated blood and organ transplantation. It can also be transmitted from mother to child during childbirth and through sexual contact which is less frequent. HCV diagnosis includes the antibody test to detect earlier encounter, the HCV RNA test to measure the amount of viral genetic material in acute infection, and HCV genotype test to identify the specific strain of the virus. A liver biopsy may be used to assess liver damage from chronic infection [17].

In many cases, the hepatitis C virus is naturally eliminated by the body's immune system, particularly through robust innate and adaptive immune responses. However, in individuals with compromised immune systems,

the production of antibodies against HCV may be insufficient or delayed. This can result in negative outcomes on both anti-HCV antibody tests and HCV RNA tests conducted via polymerase chain reaction (PCR), potentially leading to undetected infections .

The main treatment for Hepatitis C is direct-acting antivirals (DAAs), which are administered orally for 8-12 weeks, give a cure rate of more than 95% and have few side effects regardless of HCV genotype [18]. These are medications that directly inhibit the replication of the hepatitis C virus (HCV) by targeting specific proteins essential for its life cycle. DAAs of generic versions have made treatment more affordable. Commonly used DAA combinations include:

- **Sofosbuvir/Velpatasvir** (Epclusa)
- **Sofosbuvir/Ledipasvir** (Harvoni)
- **Glecaprevir/Pibrentasvir** (Mavyret)
- **Elbasvir/Grazoprevir** (Zepatier)
- **Sofosbuvir/velpatasvir/voxilaprevir** (Vosevi)

Although no vaccine is available yet, many are in the development process. The main challenges in creating a vaccine for hepatitis include HCV's genetic diversity, the tendency to mutate its envelope protein, and its ability to avoid the immune system.

1.5 E1 and E2 Glycoprotein

The envelope glycoproteins E1 and E2 of HCV are cleaved from the viral polyprotein precursor by cellular peptidases of both host and virus within the endoplasmic reticulum. These proteins are extensively N-glycosylated and are type I transmembrane proteins. They form a stable and noncovalent heterodimeric complex with their C-terminal transmembrane domains, which is important for viral entry, virulence, and evasion from the host immune response. E1 helps the virus to attach to the cell membrane whereas E2 interacts with cellular receptors **[19]**.

The E1 has two key structural components: the stem region and the core domain. The stem region helps anchor the E1 protein in the viral envelope and supports the structural integrity of the heterodimer complex. The core domain has multiple functions in viral entry, assembly of virus and fusion of its membrane with the host cell. The core protein of E1 contains several key regions: the N-terminal domain (NTD), the C-terminal loop Region (CTR), and the PCR [putative fusion peptide (pFP) containing region]. The NTD region is involved in proper folding of E1 and its interaction with E2. The PCR plays an important role in the fusion of the viral envelope and membrane of the endosome during virus entry into the host cell. The CTR connects the PCR with the stem region.

The E2 consists of three major subdomains: the head, the stem, and the transmembrane domain (TMD). The stem and the TMD regions are involved in anchoring the virus to the host membrane. The fusion process is thus facilitated, which makes the E2 region important for viral infectivity. The stem region connects the base with the TMD region and plays a critical role in membrane fusion and viral entry into host cell. The E2 head domain contains a central β-sandwich core which forms the backbone of the head domain, CD81 binding site required for binding to the CD81 receptor on the host cell membrane and initiates the infection process, hypervariable regions (HVR1) help virus to evade the immune system of host, Front and back layers contribute to the overall stability and structure of head region, Variable Regions –VR2 and VR3 whose variability allow the virus to escape recognition by the host's immune system, and base region which is an extended loop interrupted by an antiparallel β-sheet in the E2 head. **Figure 1.5** presents the key intradomain regions of the HCV glycoproteins E1 and E2, as visualized using ChimeraX **[20]**.

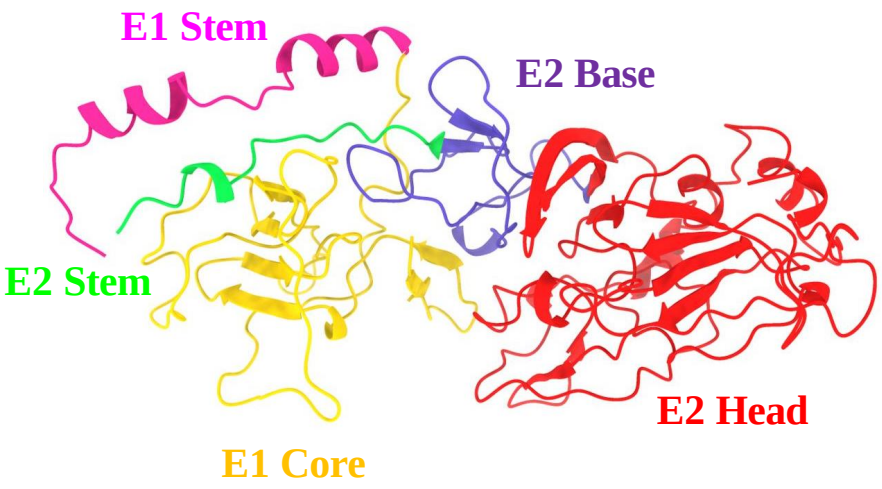


Figure 1.5: Intra-domain regions of E1 and E2 in HCV.

1.6 Glycans

Glycans are chains of sugar molecules that are covalently attached to biomolecules such as proteins (forming glycoproteins) or lipids (forming glycolipids). The Golgi apparatus is the main site within the cell where glycoproteins and glycolipids are synthesized.

Glycoproteins are proteins with sugar chains, called glycans, covalently attached to their amino acids through a process known as glycosylation. There are two main types: N-linked (where glycans attach to asparagine through the nitrogen atom) and O-linked (where they attach to serine or threonine via the oxygen atom). These glycans can be linked to a single site or multiple sites on the protein. These glycoproteins play vital roles in many cellular functions, such as maintaining cell structure, facilitating communication between cells, triggering immune responses, and regulating hormones. In viruses, glycoproteins are crucial for the virus to attach to and enter host cells, making them important targets in understanding infection mechanisms. Glycoproteins are more hydrophilic than regular proteins because of the sugar -OH groups, so they are more drawn towards water [21].

Glycolipids are lipids with attached sugar chains, mostly found on the outer surface of cell membranes. They help with cell recognition, signaling, membrane stability, and immune responses. Glycolipids consist of polar oligosaccharide chains covalently linked to hydrophobic lipid components via glycosidic bonds, rendering them amphiphilic in nature.[22]

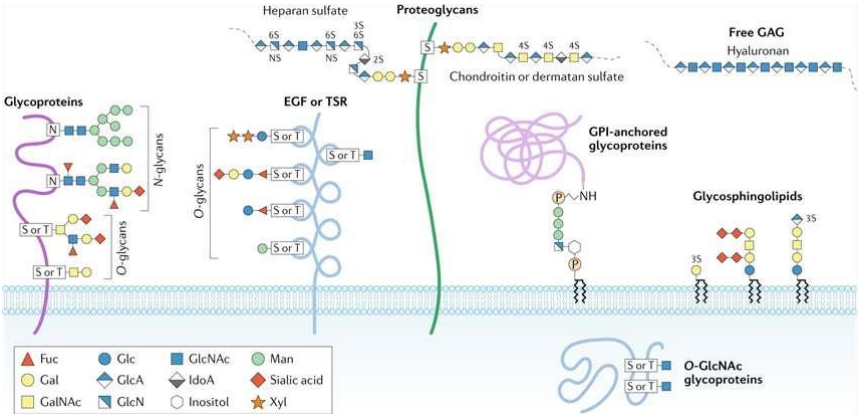


Figure 1.6: Diverse glycan structures [23].

1.6.1 Physiological functions of glycans

Glycans play a wide array of critical roles in the human body, encompassing structural support, metabolic activity, and molecular recognition. These complex carbohydrate structures are essential to various physiological processes, including tissue organization, immune defense, and cellular communication. Their functions can be broadly categorized into three main areas: (1) Structural support- they contribute to the formation of cell walls and extracellular matrices, and assist in protein folding and stability, affecting protein function and interactions. (2) Energy metabolism- involved in energy storage and supply, fueling various metabolic processes, and (3) Information Carriers- function as molecular signals, conveying information through their interactions with glycan-binding proteins (GBPs). The GBPs can be subdivided into two groups: (i) Intrinsic GBPs, which recognize glycans within the same organism, mediating processes such as cell–cell communication, trafficking, and immune signaling, (ii) Extrinsic GBPs- bind to glycans

from different organisms, playing roles in host-pathogen interactions, including microbial adhesion, invasion, and immune evasion.

Pathogens often exploit glycan recognition mechanisms for host attachment and invasion. Some engage in molecular mimicry by displaying host-like glycan structures to evade immune detection, while others actively modulate host immunity using glycan-based strategies.

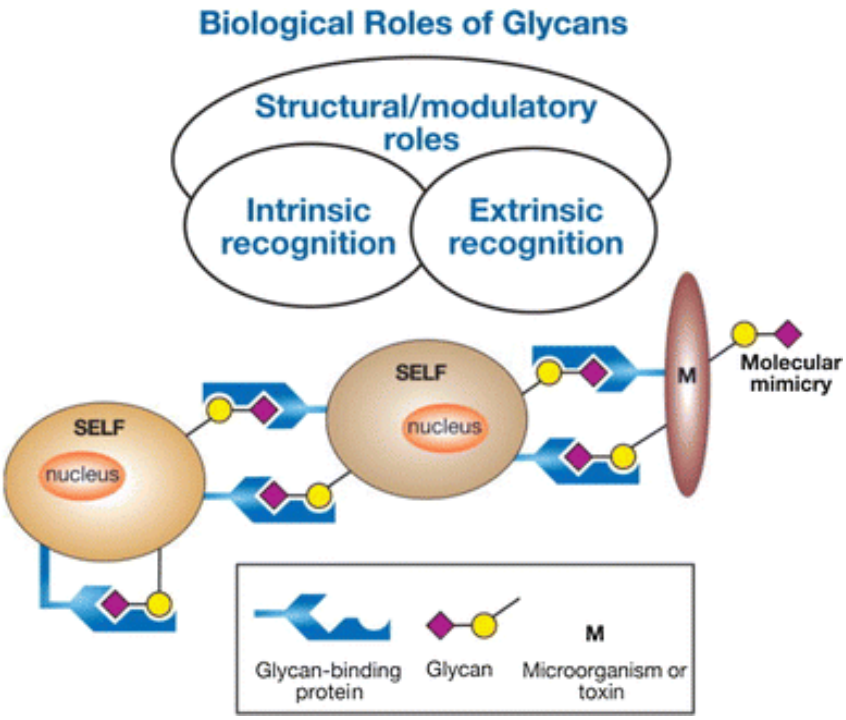


Figure 1.7: Glycan-Driven Biological Interactions [24].

Cell surface and secreted proteins are synthesized within the lumen of the ER, where they enter the secretory pathway. During this process, many proteins undergo co-translational or post-translational glycosylation, which begins in the ER. From there, the partially glycosylated proteins move to the Golgi apparatus, where their sugar structures are further modified and extended The coordinated activity of glycosidases and glycosyltransferases help to generate diverse and complex glycan patterns. Once fully glycosylated, the proteins are sorted and directed to their appropriate cellular compartments or

secreted to the extracellular environment, where they perform a wide range of structural and signaling functions, can be seen in **Figure 1.8**.

Glycans and glycoproteins at the cell surface play crucial roles in a wide range of cellular activities. They act as receptors for signals such as epidermal growth factor EGF and TGF-β, initiating key signaling pathways that regulate cellular responses. Glycans also mediate communication between cells and their surrounding environment, facilitating signal exchange and coordination. Through their influence on integrin function, glycosylation affects cell attachment and motility, which are vital for migration and tissue remodeling. Also, glycoproteins contribute to cell-cell adhesion, allowing cells to recognize and adhere to one another for tissue integrity and immune response coordination. Some enzymes are glycosylated as well, and this modification can impact their stability, activity, and localization. Glycan structures further contribute to cell immunogenicity by modulating how immune cells recognize self and non-self, either triggering or evading immune responses. Moreover, external agents such as viruses, bacteria, and toxins often exploit host glycan structures to enter into cells or interfere with cellular functions, using these sugars as specific recognition targets.

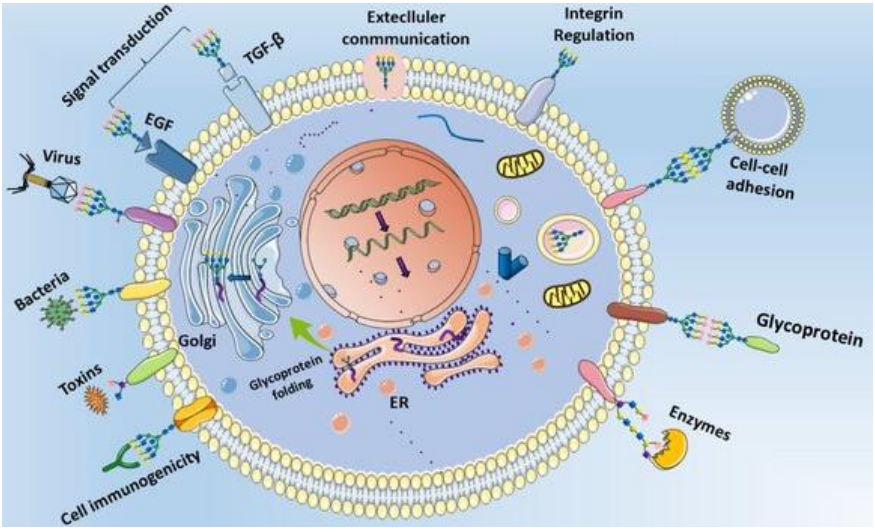


Figure 1.8: Roles of glycans in cellular mechanisms [25].

1.6.2 Glycan-Protein interaction

Protein–glycan interactions are essential for various biological processes such as cellular recognition, immune response, and pathogen adhesion. These interactions are mediated by glycan-binding proteins such as lectins and antibodies, which engage specific carbohydrate structures through non-covalent forces like hydrogen bonds, vander Waals interactions, CH–π stacking, and water-mediated contacts.

Due to the inherently weak affinity of individual glycan–protein interactions (dissociation constants in the μM–mM range), multivalency is often employed to enhance binding strength and specificity. Glycans exhibit high structural flexibility; however, upon binding, this flexibility is reduced, resulting in an unfavourable entropy change. Additionally, the hydrophilic nature of carbohydrates results in an enthalpic cost due to desolvation. As a result, glycan–protein binding is generally characterized by enthalpy–entropy compensation [26].

In N-glycosylated proteins, glycan–protein interactions often act synergistically with protein–protein interactions, further increasing binding affinity and biological specificity.

1.6.3 E1-E2 Glycoprotein Complex with glycans

Glycans were attached to the E1–E2 glycoprotein complex at residue positions 196 and 305 of the E1 subunit using the **GLYCAM** web server, which allows for the automated modelling of carbohydrate structures and their integration into protein systems in **Figure 1.9.** These specific glycosylation sites were chosen based on their proximity to the E1–E2 interface and their potential functional relevance in modulating inter-subunit interactions. Glycosylation at N196 and N305 has been shown to be essential for the formation of the E1–E2 heterodimer and for the infectivity of the hepatitis C virus (HCV). Specifically, glycosylation at N196 is critical for E1 folding and its incorporation into HCV particles, while glycosylation at N305 influences the formation of

disulfide bonds and modulates the immunogenicity of the E1 protein. The modified structure is illustrated in the figure below, where the glycans are attached at the heterodimer interface. To investigate the conformational dynamics and evaluate the influence of glycosylation on intermolecular interactions, we performed GaMD simulations for 1 microsecond in three independent replicates. This enhanced sampling technique was selected for its ability to capture rare conformational transitions and to provide a more comprehensive view of the protein's dynamic landscape compared to classical MD. The simulations aimed to explore how glycosylation influences the conformational flexibility, stability, and interaction patterns of the E1–E2 complex, particularly at the glycan-modified interface. The resulting trajectories were subjected to extensive structural and energetic analyses to assess the role of glycosylation in modulating protein–protein interactions and potential implications for viral fusion or immune evasion.

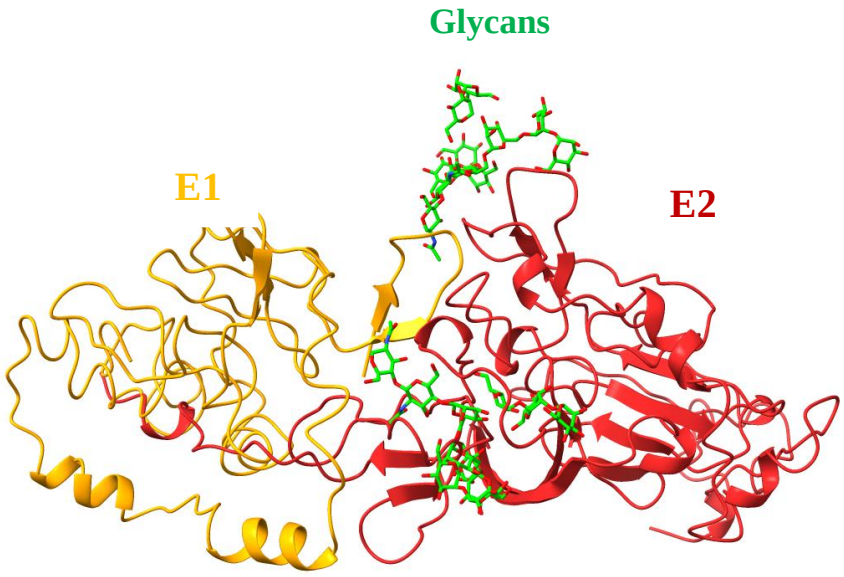
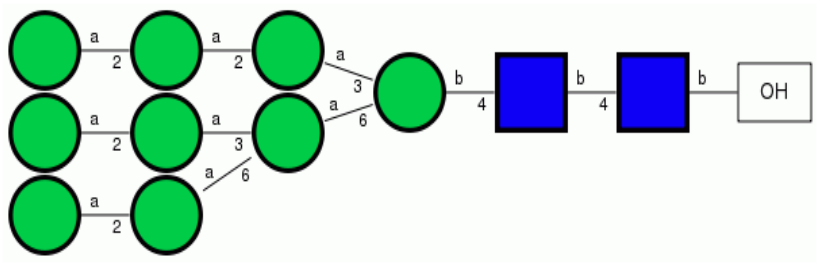


Figure 1.9: E1-E2 heterodimer complex following glycosylation

The glycan attached at the interface of the complex is shown below. This glycan is $Man_9GlcNAc_2$ which comprises of 9 mannose (Man) sugar units attached to the core of 2 molecules of N-acetyl glucosamine (GlcNAc) residues and a terminal hydroxyl group (OH) forming a

highly branched oligosaccharide. High-mannose glycans such as $Man_9GlcNAc_2$ are critical for proper protein folding, ER-associated degradation, and quality control through interactions with lectin chaperones. The mannose-rich branches serve as recognition sites for enzymes and lectins, aiding in protein trafficking and immune signaling. One of the glycans is integrated within the E1E2 heterodimer, potentially contributing to the stability or conformation of the complex. In contrast, another glycan is positioned away from the interface of the heterodimer.



Sequence:
DManpa1-2DManpa1-6[DManpa1-2DManpa1-3]DManpa1-6[DManpa1-2DManpa1-2DManpa1-3]DManpb1-4DGlcpNAcb1-4DGlcpNAcb1-OH

Figure 1.10: Structure of a high-mannose N-glycan ($Man_9GlcNAc_2$).

Figure 1.10 is a representative structure of a **highly branched mannose rich** N-linked **glycan** synthesized in the endoplasmic reticulum, playing a crucial role in protein folding, quality control, and trafficking during glycoprotein maturation.

CHAPTER 2

2. Theoretical Framework

2.1 MD Simulations

Molecular Dynamics (MD) is a powerful computer simulation technique based on Newton's laws of motion and interatomic potentials, widely used to study biological molecules such as proteins and nucleic acids. It allows us to observe how atoms and molecules move and interact over time, offering a dynamic view of molecular systems. By simulating the physical movements of these atoms, MD helps us understand how the structure of biomolecules changes, providing atomic-level insights into their behaviour. They enable the investigation of processes such as protein folding and unfolding, conformational transitions, stability assessments, molecular interactions, and recognition mechanisms.

In MD simulations, the behaviour of a system of particles (such as atoms or molecules) is modelled over time by numerically solving Newton's equations of motion. This approach allows for the simulation of atomic and molecular interactions over time, providing insights into the system's dynamic behaviour.

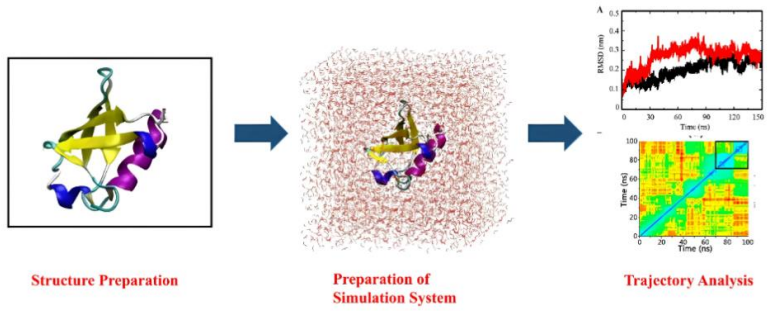


Figure 2.1: Schematic representation of the molecular dynamics simulation workflow. The process is illustrated in three main stages: (i) system setup involving protein-ligand complex construction, (ii) execution of molecular dynamics simulations, and (iii) post-simulation trajectory analysis

The fundamental equation used is Newton's second law:

$$m_i \frac{d^2r}{dt^2} = F_i = -\nabla_{ri} U(\mathbf{r}_1, \mathbf{r}_2, \ldots\ldots \mathbf{r}_N) \qquad (2.1)$$

Here, m_i is the mass of particle i, $\mathbf{r}_i$ is its position vector, F_i is the force acting on it, and U is the potential energy function dependent on the positions of all N particles in the system.

In recent years, progress in biological and medical sciences has increasingly depended on modeling and simulation, enabled by advancements in computing technology. This integration allows for accurate, tractable representations of complex biological systems across multiple scales, enhancing our understanding of their functions. While simulations cannot replace experiments, they provide valuable insights that aid in interpreting results and optimizing experimental design.

2.2 Force Fields

In molecular dynamics simulations, a **force field** is a collection of mathematical functions and parameters that define the potential energy of a molecular system based on atomic positions. These functions model both bonded interactions (such as bond stretching, angle bending, and torsional rotations) and non-bonded interactions (including van der Waals forces and electrostatic interactions). Force fields are essential for simulating the behaviour of molecules, particularly in complex biological systems.

Molecular modeling force fields are typically characterized by four main components representing both inter- and intramolecular forces. In molecular mechanics, specific functional forms are employed to model energy variations due to bond rotations and interactions between non-bonded atoms. The total potential energy of a macromolecular system, denoted as **V(r) Total**, is generally partitioned into internal interactions (e.g., bonded terms) and external interactions (e.g., non-bonded terms), shown in Figure 12.

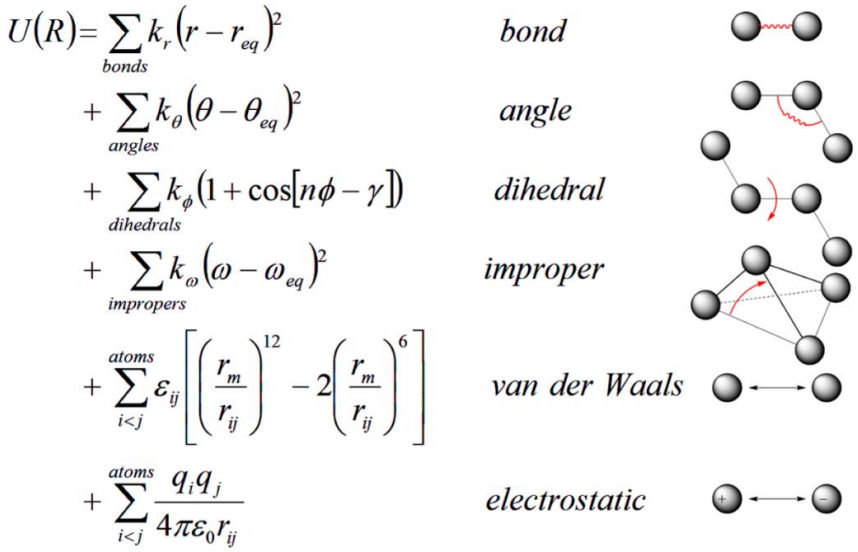


Figure 2.2: Components Breakdown of potential energy in force field-based simulations [27].

In molecular mechanics, the **potential energy function** models the total energy of a molecular system by using a a bonded term (V_{bonded}) for covalent interactions such as bond stretching, angle bending, and torsional rotations, and a non-bonded term ($V_{non-bonded}$) for long ranged electrostatic and short ranged VanderWaals forces.

$$V_{total} = V_{bonded} + V_{non-bonded} \quad (2.2)$$

$$V_{bonded} = V_{bond} + V_{angle} + V_{dihedral} \quad (2.3)$$

$$V_{non-bonded} = V_{electrostatic} + V_{van\ der\ Waals} \quad (2.4)$$

Due to variations in the bonding patterns and primary structures of proteins and carbohydrates, different force fields are utilized for accurate molecular dynamics simulations. These specialized force fields are incorporated into simulation packages such as AMBER (Assisted Model Building and Energy Refinement) [28], CHARMM (Chemistry at HARvard Macromolecular Mechanics) [29], OPLS (Optimized Potentials for Liquid Simulations) [30], and GROMOS (GROningen MOlecular Simulation) [31].

2.2.1 Protein force field

The most widely used families of protein force fields include AMBER, CHARMM, and OPLS. The AMBER (Assisted Model Building with Energy Refinement) software package includes a variety of force fields for biomolecular modelling, particularly for proteins. Force fields such as ff14SB [32], ff19SB [33], and CHARMM36m [34] are commonly used for modelling protein systems. Among all biological macromolecules, proteins are the most extensively studied, and AMBER's force fields are widely used in protein simulations. However, a key limitation of AMBER force fields is the use of fixed atomic charges, which can reduce accuracy compared to polarizable force fields that better account for electronic redistribution. The ff19SB force field is the most recent development in AMBER's protein force field and is optimized for use with the high-accuracy OPC water model, enhancing

the reliability of protein simulations. Of these, ff14SB is one of the most widely applied due to its reliability and efficiency. In our work, we employed the ff14SB force field for protein modelling and updated generalized Amber force field (GAFF2) [35] for small molecules or inhibitors. The ff14SB force field was designed to be used in combination with the TIP3P water model. Its backbone parameters were derived primarily from simulations of alanine and glycine residues.

2.2.2 Carbohydrate force field

Carbohydrate-specific force fields play a crucial role in accurately simulating the structure and motion of saccharides, which are now widely recognized for their involvement in processes such as cellular communication and pathogen recognition. However, building accurate carbohydrate force fields is particularly difficult due to the diverse and flexible nature of sugar structures and the limited experimental data available for validating or refining force field parameters. Currently, four major carbohydrate-specific force fields are commonly used: CHARMM36, GROMOS, GLYCAM family, and OPLS-AA [36]. In our study, we employed force fields from the GLYCAM family. The GLYCAM_06 [37] series represents one of the most widely utilized families of carbohydrate force fields. This series includes several versions, such as GLYCAM_06a, 06b, 06e, 06EP, and 06j, among others. Of these, GLYCAM_06j [38] is the most recent and commonly adopted variant, offering improved accuracy for modelling a wide range of glycan structures. In our study, we have used GLYCAM_06j-1 force field for the glycans.

2.3 Integration Algorithms in MD

These equations predict how atoms move over time by updating their positions and velocities, which are recorded in trajectory files. Since the atomic positions depend on potential energy, and this function lacks an exact solution for complex systems, numerical methods are used. Common integration algorithms include the Verlet [39], Velocity Verlet,

Leapfrog [40] and Beeman's algorithm. [41]. These methods apply Taylor series expansions to estimate atomic positions, velocities, and accelerations. Each has its own strengths and limitations and is selected based on the specific requirements of the simulation.

2.3.1 Verlet algorithm

The **Verlet algorithm** is a widely used numerical method for integrating Newton's equations of motion in molecular dynamics simulations. It predicts the new positions of particles in a molecular system by utilizing their positions at the current and previous time steps, along with the accelerations computed from the forces acting at the current step. As a two-step integration method, the algorithm relies on positional data from two distinct time points, making it both computationally efficient and numerically stable for molecular dynamics simulations over long simulation times.

It is simple, efficient and requires minimal memory (only current and previous positions). However, since it does not explicitly compute velocities and it is less accurate for systems requiring precise velocity-dependent properties.

$$v(t) = \frac{r(t + \Delta t) - r(t - \Delta t)}{2\Delta t} \qquad (2.5)$$

where, r is the position at time $t + \Delta t$ and $t - \Delta t$

2.3.2 Velocity Verlet algorithm

The Velocity Verlet algorithm is an enhancement of the basic Verlet integration method, where the velocity is calculated as step $n + \frac{1}{2}$ and then the coordinates at step n + 1. This algorithm calculates the positions, velocities and the acceleration simultaneously at time $(t + \Delta t)$.

$$r(t + \Delta t) = r(t) + \Delta t\, v\Delta(t) + \frac{1}{2}\Delta t^2\, a(t) \quad (2.6)$$

$$v(t + \Delta t) = v(t) + \frac{1}{2}\Delta t[a(t) + a(t + \Delta t)] \quad (2.7)$$

2.3.3. Leapfrog algorithm

The Leapfrog algorithm is a numerical integration method used to solve Newton's equations of motion, particularly in molecular dynamics simulations which is a modification of the Verlet algorithm where the velocities are calculated for the time $t + \frac{1}{2}\Delta t$, then positions are estimated at the $t + \Delta t$**.** This interleaving means that velocities "leap over" positions and vice versa, hence the name "Leapfrog". It provides explicit velocity information without the need for additional position data, making it advantageous over the basic Verlet method in many simulation scenarios.

$$r(t + \Delta t) = r(t) + v(t + \frac{1}{2}\Delta t)\Delta t \quad (2.8)$$

$$v\,(t + \frac{1}{2}\,\Delta t) = v\,(t - \frac{1}{2}\,\Delta t) + a(t)\Delta t \quad (2.9)$$

2.4 Simulation time-step

In molecular dynamics (MD) simulations, selecting an appropriate timestep (Δt) is crucial for ensuring both the accuracy and stability of the simulation. The timestep determines how frequently the simulation updates the positions and velocities of atoms, and it must be small enough to resolve the fastest motions within the system. In biological macromolecules, these rapid motions are typically associated with bond vibrations involving hydrogen atoms, which occur on the femtosecond timescale ($\sim 10^{-13}$ seconds).

To achieve stable integration of the equations of motion and accurate energy conservation, the timestep is generally chosen to be significantly shorter than the fastest motion in the system. However, using extremely small timesteps can substantially increase computational demands. To overcome this limitation, algorithms such as SHAKE [42] and LINCS [43] are employed to constrain the motion of bonds involving hydrogen atoms. By eliminating the need to explicitly simulate these high-frequency vibrations, these algorithms permit the use of larger timesteps, typically in the range of 1–2 femtoseconds (fs), without compromising the accuracy or stability of the simulation which is widely accepted standard in all-atom simulations of biological macromolecules.

2.5 Periodic boundary conditions

Periodic boundary conditions (PBC) are widely used in molecular dynamics simulations for approximating the bulk behaviour of a system while minimizing edge effects that arise due to the finite size of the simulation box. In real biological systems, biomolecules are surrounded by a vast number of solvent molecules but simulating such an environment with infinite solvent molecules is computationally infeasible. PBC helps overcome this limitation by replicating the simulation box in all three Cartesian dimensions, effectively creating the illusion of an infinite system.

With PBC, when a particle moves out of the primary simulation box, its image simultaneously re-enters from the opposite side, preserving the overall number of particles and maintaining equilibrium. This continuous exchange prevents the occurrence of surface effects and ensures that all particles, including those near the edges, experience forces similar to those in the system's interior.

To reduce computational load, a cutoff radius (rcut) is applied to limit the calculation of non-bonded interactions. This cutoff is usually set to a value less than or equal to half the length of the simulation box to avoid

interactions being counted multiple times across the periodic images. For long-range interactions, especially electrostatic forces that extend beyond the cutoff distance, specialized algorithms like the Particle Mesh Ewald (PME) [44] method are used to maintain accuracy.

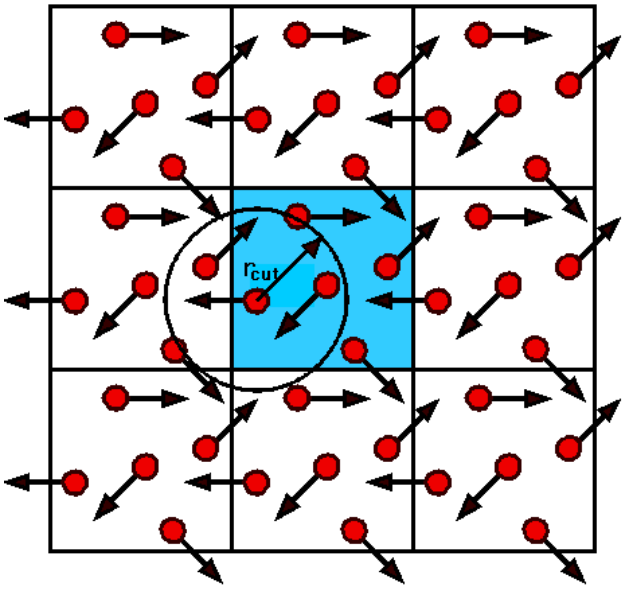


Figure 2.3: Two-dimensional representation of periodic boundary conditions (PBC). $\mathbf{r_{cut}}$, or the **cutoff radius** is applied when calculating the force between two atoms [45].

2.6 Long-range interactions

In molecular dynamics simulations, interactions between atoms are categorized into bonded and non-bonded types. Bonded interactions, involving atoms connected through covalent bonds, encompass bond stretching, angle bending, and torsional rotations. These interactions are limited in number and remain constant during simulations, making their computation relatively straightforward and less resource-intensive.

Non-bonded interactions, encompassing electrostatic forces, van der Waals interactions, hydrogen bonds, and salt bridges, occur between all pairs of atoms not directly bonded, leading to a computational cost that scales quadratically with the number of atoms and are fundamental to

the structural stability and functional dynamics of proteins. These interactions, though individually weaker than covalent bonds, collectively contribute significantly to the maintenance of a protein's tertiary and quaternary structures.

2.7 Thermostats

Thermostats are algorithms designed to regulate the system's temperature by modifying the Newtonian equations of motion, which inherently conserve energy. While thermostats are essential for maintaining a desired temperature during the equilibration phase, they can interfere with the accurate calculation of dynamical properties, such as diffusion coefficients. Several thermostat algorithms are commonly employed in MD simulations: Gaussian [46], Berendsen [47], Bussi-Donadio-Parrinello [48],Andersen [49], and Langevin . In our research, we have used Langevin thermostat [50].

2.7.1 Langevin Thermostat

The Langevin thermostat integrates the principles of microcanonical ensemble dynamics with aspects of Brownian motion to model the behavior of particles in a viscous medium. It uses a general equation of the form,

$$F = F_{interaction} + F_{friction} + F_{random} \qquad (2.10)$$

Where $F_{interaction}$ is the standard interactions calculated during the simulation, $F_{friction}$ is acting on particles, effectively tuning the "viscosity" of the implicit solvent or heat bath, and F_{random} effectively gives random collisions with the solvent molecules. The frictional and random forces are coupled through a user-defined friction damping parameter.

2.8 Barostats

In molecular dynamics simulations, replicating laboratory conditions—typically constant temperature and pressure is achieved using the isothermal-isobaric (NPT) ensemble. This ensemble maintains a constant number of particles (N), pressure (P), and temperature (T), allowing the simulation box to adjust its volume in response to pressure fluctuations.

Barostats adjust the system's volume to maintain the desired pressure, and they are often used in conjunction with thermostats to achieve the NPT ensemble. Several barostat algorithms are commonly utilized in MD simulations: **Berendsen** [51], **Andersen** [49], **Parrinello - Rahman** [52], and **Martyna-Tuckerman-Tobias-Klein** [53]. In our study we have used the Berendsen barostat to control the pressure.

2.8.1 Berendsen barostat

The Berendsen barostat regulates pressure in molecular dynamics simulations by uniformly scaling the system's volume based on the difference between the current and target pressures. This method introduces a correction term to the equations of motion, facilitating rapid pressure equilibration. However, it does not accurately reproduce the pressure fluctuations characteristic of the NPT ensemble, making it unsuitable for production runs.

$$\frac{dP}{dt} = \frac{P_0 - P}{\mathcal{J}_P} \tag{2.11}$$

Where, P_0 is the reference pressure, i.e. the pressure of the external pressure "bath", and P is the instantaneous pressure and $\mathcal{J}_P$ is a time constant.

2.9 Molecular Dynamics Simulation Protocol

The core workflow of molecular dynamics (MD) simulations includes a series of well-defined steps that includes:

2.9.1 System Preparation

System preparation is a vital first step in molecular dynamics (MD) simulations, as errors at this stage can impact the reliability of the results. It begins with selecting molecular components—such as proteins, ligands, or glycans—whose structures are sourced from experimental methods (e.g., X-ray crystallography [54], NMR [55], Cryo-EM [56]) or databases like the RCSB PDB. For glycans, tools like GLYCAM may be used to build structures.

The initial model is then validated for completeness, proper protonation states, and charge neutrality. An appropriate force field is assigned to define atomic interactions, including bonded and non-bonded terms. To simulate a realistic environment, the system is solvated in a water box and neutralized with counter ions (such as Na^+ or Cl^-). The solvent box size is chosen to minimize boundary effects and ensure accurate long-range interactions.

2.9.2 Solvation

Since biological reactions occur in aqueous environments, it is essential to solvate systems in MD simulations. This can be achieved using **implicit** models, which simulate water as a continuous field, or **explicit** models that place individual water molecules around the solute. While implicit models are faster and less resource-intensive, explicit models represent individual water molecules, offering greater accuracy at higher computational cost.

Among explicit models, **TIP3P** [57] which is a 3-site model is the most commonly used due to its balance between efficiency and compatibility with most force fields. The other 3-site models are SPC, SPC/E and

TIPS. In our study, TIP3P was selected for solvating glycan and protein-glycan systems.

2.9.3 Minimization

Energy minimization is a crucial initial step in molecular dynamics simulations, aimed at stabilizing the system's initial structure by reducing potential energy and resolving any steric clashes. This is critical to prevent simulation instability during subsequent heating phases. Minimization adjusts atomic coordinates to locate a local minimum on the potential energy surface, typically using algorithms such as steepest descent, conjugate gradient, or Newton-Raphson.

Minimization is usually carried out in two stages: the first involves restraining the solute to allow the solvent to relax, while the second relaxes the entire system without restraints. This step ensures structural integrity and helps avoid distortions like bad contacts that can arise from high-energy interactions between solute and solvent. By ensuring a stable starting point, energy minimization lays the foundation for reliable molecular simulations.

2.9.4 Heating

Following energy minimization, the heating step is performed to gradually introduce kinetic energy into the system, bringing it from 0 K to the target simulation temperature. To prepare it for simulation, we need to gradually heat it up to the desired temperature. This step increases the atoms' velocities over time, helping the system reach thermal equilibrium without becoming unstable. This is typically achieved using the NVT ensemble, which maintains constant volume and allows for the safe addition of energy via velocity rescaling or thermostats based on the Maxwell-Boltzmann distribution.

Gradual heating over a defined timeframe ensures smooth thermal equilibration and reduces the risk of sudden atomic displacements. The

NVT ensemble is preferred over NVE and NPT during this stage, as NVE does not permit energy input, and NPT could lead to unwanted volume changes due to pressure coupling. Controlled heating prepares the system for stable dynamics and helps the system adjust gently, reducing the risk of it “blowing up” and ensuring it’s ready for the next stage of the simulation.

2.9.5 Equilibration

Equilibration is a critical phase following the heating step, allowing the system to reach a stable thermodynamic state before entering the production run. This step ensures that the temperature, pressure, and density stabilize under the desired simulation conditions. Initially, equilibration is typically performed under the NVT ensemble to allow the system’s kinetic and potential energies to balance. During this phase, the thermal energy introduced during heating is distributed evenly across all degrees of freedom.

As the production run is usually carried out in the NPT ensemble, a buffer period is introduced to transition smoothly from NVT to NPT, during which the solvent density and other properties adjust accordingly. Throughout equilibration, key thermodynamic parameters like temperature, pressure, and potential energy are monitored until they plateau, indicating the system has achieved equilibrium. Once fluctuations in energy and other properties become minimal, the system is considered equilibrated and ready for production simulations.

2.9.6 Production Run

After the completion of the equilibration, the simulation enters the production phase. This is the stage where the system runs steadily for a longer period, allowing it to generate the trajectory data used in the analysis of structural, dynamic, and thermodynamic properties of the system. The production run typically maintains the same simulation

parameters as the equilibration phase, except that data is now actively collected. Ensembles such as NPT, NVT, or NVE may be employed depending on the objectives of the simulation. Atomic positions, velocities, and other relevant information are saved at defined time intervals to capture the system's behaviour over time.

Simulation lengths in the production phase can range from nanoseconds to microseconds, depending on the complexity of the molecular system and the desired resolution of the analysis. With advancements in high-performance computing, especially the use of GPUs, longer simulations at microsecond scales have become increasingly accessible. [58]

2.9.7 Analysis

Trajectory analysis was carried out using the **Cpptraj** [59] module included in **AmberTools19** [60]. To minimize the impact of initial fluctuations, the first 200 ns of each trajectory were discarded. The remaining segments from the three independent replicates were combined and analyzed to explore the dynamic behaviour of the systems. Initial assessments of structural stability and flexibility were performed by calculating the **root mean square deviation (RMSD)** (both proteins and glycans) and **root mean square fluctuation (RMSF)** relative to well-equilibrated reference conformations. The **radius of gyration** was calculated to assess the overall compactness of the protein complex throughout the simulations. In addition, hydrogen bond analysis was performed using a distance cutoff of **≤ 3.0 Å** and an occupancy threshold of **20%** to identify interactions. LigPlot analysis was performed to visualize key intermolecular interactions, including hydrogen bonds and hydrophobic contacts, particularly at the E1E2 interface. **Dynamic cross-correlation matrix (DCCM)** were generated to investigate coordinated movements between residue pairs. **Principal component analysis (PCA)** [61] was performed to capture the dominant motions and explore conformational transitions. Furthermore, **protein structure network (PSN)** analysis was performed to explore

residue-residue interaction networks, enabling identification of key communication hubs and pathways potentially relevant to allosteric regulation and complex stability.

CHAPTER 3

3. Objectives

- **To identify the structural and functional interaction interface between E1 and E2 using Molecular Dynamics simulations.**

  This objective focuses on a comprehensive structural and dynamic analysis of our target protein. The hepatitis C virus (HCV) envelope glycoproteins E1 and E2 form a noncovalent heterodimer essential for viral entry. Understanding the specific regions and interactions that facilitate this heterodimerization is crucial. The process begins with retrieving the protein's three-dimensional structure from the Protein Data Bank (PDB). Given that PDB entries often have missing residues due to limitations in experimental techniques, it's essential to identify and model these absent segments to ensure a complete and accurate structure using Modeller. Studies have shown that both the ectodomain and transmembrane domains of E1 and E2 contribute to their interaction, with certain conserved motifs playing pivotal roles in maintaining the structural integrity of the complex. Elucidating these interfaces can provide insights into the mechanisms of viral assembly and entry.

- **To investigate the role of E1-E2 Interaction in the Viral Life Cycle, including implications for entry and fusion**

  Upon surveying the literature, it is evident that the E1-E2 heterodimer is not only structural but also functional in mediating HCV entry into host cells. While E2 is primarily responsible for receptor binding, E1 is believed to facilitate membrane fusion. Recent studies suggest that E1 contains a

putative fusion peptide and can form trimers, characteristics typical of fusion proteins. The coordinated action of E1 and E2 is essential for the conformational changes required during the fusion process, highlighting the importance of their interaction in the viral life cycle. Analyzing the simulation trajectories will allow us to identify key conformational changes- interactions between E1 and E2, and potential fusion intermediates. This computational approach aims to provide detailed insights into the structural transitions and interactions that facilitate HCV membrane fusion.

- **To understand and explore the structural dynamics resulting from glycan interactions with the E1–E2 heterodimer.**

  This objective aims to investigate how glycosylation affects the conformation and function of the hepatitis C virus (HCV) envelope glycoproteins E1 and E2. Both E1 and E2 are heavily glycosylated, with E1 possessing up to five N-linked glycosylation sites and E2 up to eleven, depending on the genotype. These glycans are crucial for proper protein folding, stability, and the formation of the E1–E2 heterodimer, which is essential for viral entry into host cells. Glycosylation also influences the immunogenicity of the virus, with certain glycan modifications enhancing the virus's ability to evade the host immune response.

  GaMD simulations can elucidate how specific glycan modifications, such as the removal or addition of particular N-linked glycans, influence the structural integrity and functional properties of the E1–E2 complex. This approach helps to identify critical glycosylation sites that are essential for maintaining the heterodimer's stability and functionality, providing potential targets for therapeutic intervention.

CHAPTER 4

4. Methodology

4.1 Protein structure preparation

In the current investigation, we utilized the crystal structure of the HCV glycoprotein E1-E2 heterodimer complex (PDB ID: 7T6X) [62]. To model all of the missing regions, the Modeller [63] web server in UCSF Chimera [64] was used. Specific regions or domains of interest within the E1 and E2 proteins were identified based on known functional motifs or regions critical to the protein's activity. We prepared the Apo system in which both the E1 (Chain E) and E2 (Chain U) glycoproteins of HCV were included. The structure has a resolution of 3.83 Å.

Then, we developed a second system, referred to as the glycosylated complex, by incorporating specific N-linked glycans at positions N196 and N305 of the E1 glycoprotein at the interface of E1–E2 heterodimer based on an extensive literature review. For our molecular dynamics simulations, we selected high-mannose-type glycans, such as $Man_9GlcNAc_2$, which is commonly associated with HCV envelope glycoproteins. The molecular formula of the glycan is $C_{61}H_{111}NO_{46}$ and molecular weight is 1940.7 g/mol.

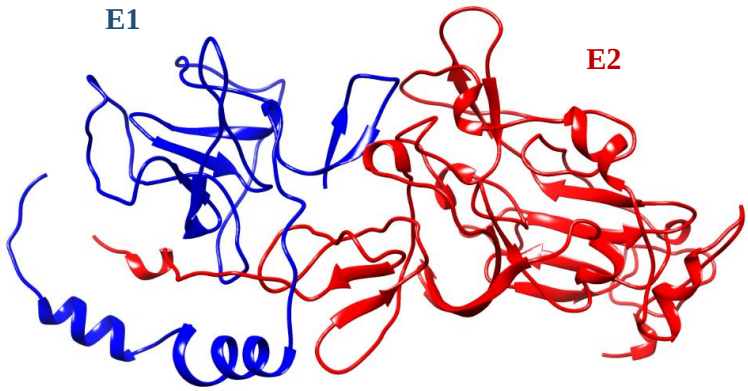


Figure 4.1: Apo Structure of HCV

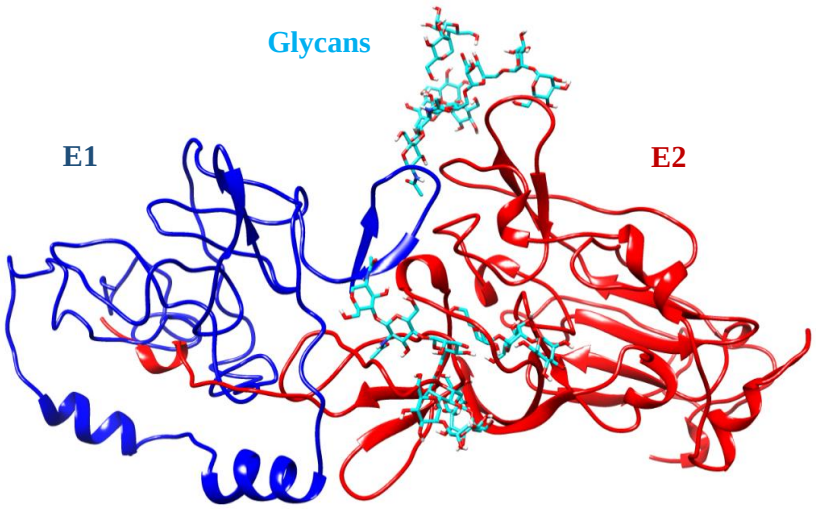


Figure 4.2: Complex structure (with attached glycans) of HCV.

4.2 Simulation Protocol

First, we performed standard Molecular Dynamic simulations using *pmemd.cuda* module of AMBER 18. Then we generated Force field parameters in AMBER prior to simulations utilizing the LEap Module. The force field that is used for protein is ff14SB. We used the TIP3P water model to solvate each system in an octahedral box, maintaining a 12 Å gap between the solute and the box boundary. We added 103 Na^+ and 108 Cl^- ions to neutralize the system. The SHAKE algorithm was used to restrict the lengths of all the hydrogen bonds and cause vibrational motion of other atoms. The method used to manage non-bonded electrostatic interactions was Particle Mesh Ewald (PME) , with a threshold set at 12 Å. We kept a constant timestep of 2 fs during the simulation. A clear step by step processes of minimization, heating, and equilibration was carefully followed before starting the production simulation. For the solvated complexes, two stages of energy minimization were carried out. A weak harmonic constraint of 2 $kcal\ mol^{-1}$ $Å^{-2}$ was included in the first energy minimization stage. Then, the second minimization stage was carried out without any constraints. The steepest descent approach was used for 500 steps in each minimization stage, and then the conjugate gradient algorithm was used

for another 500 steps. Following the minimization steps, the systems were heated to 300 K from 0 K in the systemic manner of the NVT ensemble. The Langevin thermostat and Berendsen barostat, having a collision frequency of 2 ps^{-1}, are used to maintain a constant temperature and pressure. Each system went through 1 ns of equilibration,8 ns of Conventional MD simulation was run for 2 ns timestep and later we did the GaMD Equilibration with total and dihedral boost potential for 64 ns in each run. Finally, the GaMD production run was calculated for 1 μs in each run for apo and complex systems.

4.3 Gaussian Accelerated Molecular Dynamics (GaMD) Simulations

Gaussian Accelerated Molecular Dynamics (GaMD) [65] is an advanced augmented sampling technique that introduces a non-negative harmonic boost potential to the system's initial potential energy surface. This approach, which utilizes a Gaussian distribution for the boost potential, effectively reduces energy barriers, thereby accelerating the exploration of conformational space. In contrast to the previously used Accelerated Molecular Dynamics (aMD) [66] method, GaMD solves the problem of statistical noise that commonly occurs in large biomolecular systems during reweighting. A distinct advantage of GaMD is that it does not need the definition of collective variables (CVs) or specific reaction coordinates, making it particularly well-suited to study the dynamic behavior of biological systems without requiring pre-defined CVs.

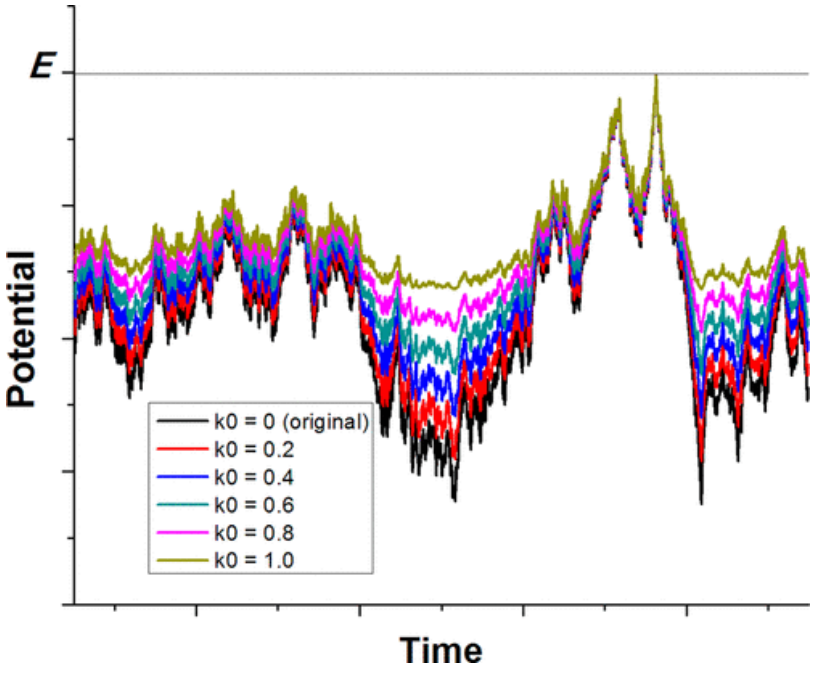


Figure 4.3: Schematic representation of Gaussian accelerated molecular dynamics [65].

A harmonic boost potential is applied to smooth the system's potential energy surface when the threshold energy is set to the maximum potential (E = Vmax), facilitating enhanced sampling by reducing energy barriers. The parameter k_o (ranging from 0 to 1) controls the magnitude of the boost; higher values of k_o correspond to greater smoothing and improved exploration of biomolecular conformations.

4.3.1 Boost Potential Formulation

If we consider a system with N atoms at positions $=\{r1, r2 \cdots rN\}$, a boost potential is added when the system potential $V(r)$ is lower than a threshold energy E :

$$\Delta \mathrm{V}(r) = \frac{1}{2}\, k\{E - V(r)\}^2 \quad , \qquad V(r) < E \qquad (4.1)$$

Here, k is the harmonic force constant.

The modified system potential, $V^*(r) = (r)+\Delta V(r)$ is given by:

$$V^*(r)= V(r) + \frac{1}{2}\, k\{E - V(r)\}^2 \quad , \quad V(r) < E \qquad (4.2)$$

Otherwise, when the system potential is above the threshold energy, i.e., $V(\mathrm{r}) \geq E$, the boost potential is set to zero and $V^*(r) = V(r)$.

For any two arbitrary potential energy $V_1(r)$, $V_2(r)$ found on the original energy surface; where $V_1(r) < V_2(r)$ and the ΔV potentials satisfy $V_1^*(r) > V_2^*(r)$, then the equation can be expressed as follows:

$$\mathrm{E} < \frac{1}{2}\{V_1(r) + V_2(r)\} + \frac{1}{k} \qquad (4.3)$$

and if $V_1(r) < V_2(r)$ and the difference in modified potential energy surface should be smaller than the original energy surface, that is $V_2^*(r) - V_1^*(r) < V_2(r) - V_1(r)$ and the equation can be modified as:

$$\mathrm{E} > \frac{1}{2}\{V_1(r) + V_2(r)\} \qquad (4.4)$$

Combining both the equations (4.3) and (4.4) and using the relationship, $V_{min} \leq V_1(r) < V_2(r) \leq V_{max}$, the threshold energy E follows the range given below:

$$V_{max} \leq \mathrm{E} \leq V_{min} + \frac{1}{k} \qquad (4.5)$$

Here, V_{max} and V_{min} are the maximum and minimum potential energies.

We employed the dual potential boost for GaMD modeling in our research. The dual boost parameter was determined utlizing the first 8 ns of conventional MD simulations. This was followed by applying the boost potential during 56 ns of GaMD simulations. Then a 1 μs GaMD simulation was performed within the NVT ensemble, with coordinates recorded every 10 ps to generate 100,000 conformations in a single run.

In the conducted Gaussian accelerated molecular dynamics (GaMD) simulations, both the apo and complex systems underwent three independent runs each. Each run generated 100,000 conformations, resulting in a total of 300,000 conformations for the apo system and 300,000 for the complex system.

4.4 Trajectory analysis techniques

Molecular dynamics simulations produce highly intricate datasets by capturing every atom's Cartesian coordinate in the system, which may include thousands or even millions of atoms, at each time step along the trajectory. These simulations can cover thousands to millions of time steps. As a result, advanced analytical techniques are necessary to extract valuable information from the data. This section introduces various analytical approaches aimed at studying conformational changes in typical short- and long-term biomolecular simulations.

4.4.1 Stability and flexibility analyses

The structural stability of biomolecular simulation is mainly defined by its root mean square deviation **(RMSD).** RMSD is a statistical measure of finding similarities between two sets of values in superimposed structures using algorithms like the Kabsch algorithm. RMSD measures the target coordinate's deviation from the reference coordinates. It calculates the average distance between the reference structure and selected atoms. A lower Root Mean Square Deviation (RMSD) signifies that the structure is more closely aligned with the reference conformation, reflecting higher structural similarity, while a higher value indicates a greater structural difference between the compared conformations. This suggests that the structure under analysis deviates more significantly from the reference, reflecting lower structural similarity. Such divergence can result from conformational changes or flexibility. A plateau in the RMSD plot suggests that the system has reached equilibrium, while significant fluctuations may indicate conformational changes or instability RMSD calculations can be

performed on all atoms or specific subsets, such as backbone or Cα atoms. In molecular dynamics simulations, the Root Mean Square Deviation (RMSD) is plotted over time to assess the structural stability and conformational changes of biomolecules, such as proteins. It is defined as:

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}(r_i(1)-r_i(2))^2}{N}} \qquad (4.6)$$

Where, N is the number of atoms whose positions are being compared and $r_i(1), r_i(2)$ are the position of atom *i* in each molecule.

Another important quantity is root-mean-squared-fluctuations **(RMSF)** to explore residual flexibility. RMSF indicates the positional differences for the entire structure over time. RMSF (Root Mean Square Fluctuation) is a measure of the average deviation of atomic positions relative to their mean positions throughout a molecular dynamic simulation used to quantify the flexibility of individual atoms or residues within a protein over time. It provides information about the flexibility and dynamic behavior of a protein structure. RMSF analysis is often applied to backbone or alpha-carbon atoms. Higher RMSF values indicate greater atomic mobility, often observed in flexible regions such as loops or terminal residues. Conversely, lower RMSF values suggest limited movement, typically associated with more rigid structural elements like α-helices and β-sheets. The RMSF for atom *i* is calculated using the formula:

$$\rho_i = \sqrt{[(x_i - \langle x_i \rangle^2]} \qquad (4.7)$$

Here, x_i represents the position of atom *i* at a given time, and $\langle x_i \rangle$ denotes the average position of atom *i* over the simulation period. This calculation yields the standard deviation of the atom's position, reflecting its mobility.

The B-factor, also known as the Debye-Waller factor or temperature factor, is derived from X-ray crystallography experiments and reflects the atomic displacement or thermal motion within the crystal structure of a protein. These values are included in Protein Data Bank (PDB) files and offer experimental insight into the flexibility of different regions within the protein. Comparing RMSF values from MD simulations with B-factors from crystallographic data can validate the simulation results. A strong correlation between high RMSF regions and high B-factor regions suggests that the simulation accurately captures the flexible regions of the protein, such as loops or terminal residues. The B-factor is defined as:

$$B = \frac{8}{3N}\pi^2(\mathrm{RMSF})^2 \qquad (4.8)$$

In protein crystallography, the B-factor (also known as the temperature factor or Debye–Waller factor) quantifies the mean square displacement of atoms from their average positions. Higher B-factor values indicate greater atomic mobility or flexibility, often corresponding to regions such as loops or terminal in proteins.

We also measured the compactness of the simulated systems by the radius of gyration **(Rg).** The radius of gyration is a measure that reflects the distribution of a protein's atoms relative to its center of mass, providing an indication of the overall spatial spread of the protein's structure. Mathematically, it represents the root-mean-square distance of the protein's atoms from its center of mass, providing insight into how tightly the protein is folded. A lower Rg value indicates a more compact, well-folded protein conformation indicating stable structure, whereas a higher Rg suggests a more extended or unfolded structure. The radius of gyration Rg for a protein can be represented as a collection of $\boldsymbol{N}$ atoms and calculated using the following formula:

$$Rg = \sqrt{\frac{1}{N}\sum_{i=1}^{N} r_i^2} \qquad (4.9)$$

Here, Rg is the radius of gyration, ***N*** is the number of atoms in the protein and r_i is the distance of each atom from the center of mass of the protein.

Solvent Accessible surface area (SASA) was originally introduced by Lee and Richards in 1971 and is often referred to as the Lee-Richards molecular surface. Later, in 1973, Shrake and Rupley developed the widely used 'rolling ball' method to calculate ASA, where a sphere representing a solvent molecule rolls over the surface of the structure to map accessible regions. We also measured the Solvent Accessible Surface Area **(SASA)** to analyze the exposure of a biomolecule's surface to the solvent. It provides information about the structural changes and it helps to identify regions of a protein that are exposed to the surrounding solvent, providing critical insights into how the protein folds, maintains its stability, and interacts with other molecules. A higher SASA value means that a greater portion of the protein's surface is exposed to the solvent, often leading to increased flexibility and a higher potential for interactions with other molecules suggesting a more expanded or diffused protein structure, while a lower SASA value indicates a more compact and tightly folded structure. Here, the equation is:

$$\nabla . [\varepsilon(r)\nabla\varphi(r)] - k' \sinh[\varphi(r)] = -4\pi\rho(r) \qquad (4.10)$$

4.4.2 Dynamic cross-correlation matrix (DCCM)

The degree of correlation within a system can be assessed by examining the cross-correlation coefficients between pairs of atoms. This information is typically presented graphically in a matrix format known as the dynamical cross-correlation matrix (DCCM). DCCM analysis is widely employed to measure the correlated motions among atoms. It is a widely used technique for studying the movement patterns in molecular dynamics (MD) simulation trajectories.

$$DCC(i,j) = \frac{\langle \Delta \mathbf{r}_i(t) \cdot \Delta \mathbf{r}_j(t) \rangle_t}{\sqrt{\langle \|\Delta \mathbf{r}_i(t)\|^2 \rangle_t} \sqrt{\langle \|\Delta \mathbf{r}_j(t)\|^2 \rangle_t}}, \quad (4.11)$$

In this method, $r_i(t)$ represents the position of atom ***i*** over time ***t***, and $\Delta r_i(t)$ shows how much the atom's position changes compared to its average position. The DCCM produces an N×N heatmap, where ***N*** is the number of atoms (usually alpha carbons), and each point shows how two atoms move in relation to each other.

The correlation value ranges from -1 to +1. A value of +1 means the atoms move together (complete correlation), -1 means they move in opposite directions (complete anti-correlation), and 0 means no connection in their movements (no correlation). Movements that are fully correlated happen at the same time and in the same way, while anti-correlated movements happen at the same time but in opposite ways. The high diagonal value occurs when $i = j$, i.e., $DCC\ (i,j) = 1.00$.

A strong correlation appears along the diagonal of the matrix because each atom is perfectly correlated with itself. Positive values near the diagonal show that nearby residues move together, while off-diagonal values indicate movement between atoms that are farther apart in the structure.

4.4.3 Principal component analysis (PCA)

Principal Component Analysis (PCA) is a technique used for reducing the dimensionality of data. It calculates the principal components, which are eigenvectors associated with large eigenvalues, based on atomic coordinates from molecular dynamics (MD) trajectories. The eigenvectors indicate the direction of motion, while the eigenvalues represent the extent of these movements. PCA is used to analyze the trajectory data and focus on the main modes of motion in a system, reducing it to a few degrees of freedom.

To apply PCA to MD data, the first step is to remove the overall rotational and translational movements using a least-squares fitting procedure. Then, a covariance matrix is created based on the Cartesian coordinates of the atoms. This matrix shows how the movements of atoms are related to each other. The matrix is typically 3N×3N, where N is the number of atoms in the system. After diagonalizing the covariance matrix, a set of eigenvectors and their associated eigenvalues are obtained. The eigenvectors represent the directions of motion, while the eigenvalues indicate the extent of these movements.

PCA helps to identify new axes along which the data is spread out the most:

1. **First Principal Component (PC1)**: This is the direction that captures the maximum variance in the data (the most spread).
2. **Second Principal Component (PC2)**: This is the next direction that captures the next largest variance, and is perpendicular to PC1, and so on.

Consider a covariance matrix C and the elements C_{ij} of the matrix is defined as:

$$C_{ij} = \langle (x_i - \langle x_i \rangle)\,(x_j - \langle x_j \rangle) \rangle \qquad (4.12)$$

Where x_i and x_j are coordinates of the i^{th} or j^{th} atom, $\langle x_i \rangle$ and $\langle x_j \rangle$ are the mean average coordinates of the i^{th} or j^{th} atom. For three dimensions, the covariance matrix for (x, y), (x,z) and (y, z) coordinates are carried out, and a covariance matrix C generates the matrix of 3N × 3N, where N denotes the number of atoms.

The covariance matrix is then diagonalized to get the eigenvalues:

$$A^T\ C\ A = \lambda \qquad (4.13)$$

Where A is the eigenvectors and λ is the eigenvalues.

The eigenvectors are ranked according to their eigenvalues in descending order. The first principal component corresponds to the eigenvector with the largest eigenvalue, which represents the dominant motion in the system. Additional principal components follow this order, describing less significant motions.

PCA allows for the reduction of the system's dimensionality by focusing on the first few principal components. These components capture the majority of the system's movement, and only a small number of them are typically needed to accurately represent the dynamics of the system.

To put it simply, the purpose of PCA is to reduce the complexity of a dataset by decreasing the number of variables, while retaining as much information as possible.

4.4.4 Hydrogen Bond analysis

Hydrogen bonds are formed through electrostatic interactions between hydrogen donor and acceptor groups. These interactions are facilitated by the partial positive charge on hydrogen and the electronegative atoms, such as oxygen or nitrogen, on the receptor. The geometry and strength of these hydrogen bonds influence glycan-protein interactions, impacting both binding kinetics and thermodynamics. Intramolecular hydrogen bonds help stabilize protein structures, especially in α-helices and β-sheets. Intermolecular hydrogen bonds play a key role in facilitating specific interactions between proteins and ligands, proteins and DNA, as well as other biological macromolecules.

Hydrogen bonding is essential in glycan-protein interactions, particularly in determining binding specificity and affinity. Glycans, which are abundant in hydroxyl groups, readily form hydrogen bonds with protein side chains and water molecules, creating a complex network of interactions. These bonds are crucial for the recognition and binding of specific glycans, and the strength and nature of these interactions can influence protein binding and function. Protein residues like aspartate, glutamine, arginine, and histidine are often involved in glycan recognition because they can both donate and accept hydrogen bonds. Additionally, water molecules can serve as bridges, enhancing the hydrogen bonding network. The number, type, and location of these hydrogen bonds play a key role in determining the specificity and affinity of the glycan-protein interactions.

Specificity and Selectivity: Hydrogen bonds play a crucial role in conferring specificity and selectivity to ligand-receptor interactions. They allow for precise geometric complementarity and recognition between binding partners. Structural analyses demonstrate how hydrogen bonds create a detailed network that governs molecular recognition with high precision.

Affinity and Binding Kinetics: The establishment of hydrogen bonds contributes significantly to the overall binding affinity of glycan-protein complexes, affecting both the rates of association and dissociation. Molecular dynamics simulations and experimental kinetics provide insights into the dynamic nature of hydrogen bond-driven interactions and their influence on binding energetics.

4.4.5 LigPlot analysis

LigPlot analysis is a method that creates 2D diagrams showing how a ligand interacts with a protein, based on a PDB file. It automatically identifies and illustrates key interactions like hydrogen bonds and hydrophobic contacts, which are important for the stability and binding

of the ligand to the protein. It can be used to study a single complex or compare different complexes to better understand binding site specificity and selectivity. The process begins by using a PDB file as input, where LigPlot+ detects the hydrogen bonds and hydrophobic interactions between the protein and ligand. It then generates a clear 2D plot showing these connections, placing the ligand at the centre and surrounding it with the interacting protein residues. This analysis is widely used in drug discovery and structural biology to understand how molecules interact and to assist in the design of new therapeutics.

In this study, we used LigPlot to analyze protein-protein interactions. LigPlot automatically generates 2D diagrams showing key interactions such as hydrogen bonds and hydrophobic contacts between two proteins. Although LigPlot is commonly used for protein-ligand studies, it can also effectively highlight important contact points in protein-protein complexes. By using PDB files as input, the software identifies interacting residues and represents them clearly in a 2D format, helping us visualize and understand how the two proteins are connected. This analysis provided valuable insights into the binding interfaces and the nature of interactions stabilizing the protein complex.

4.4.6 Protein structure network (PSN) analysis

For visualization of protein structures beyond just their secondary structure and fold, we can use network representations to highlight interactions between residues. These networks provide valuable insights into the structure-function relationship. In this study, we used the WebPSN server **[67]**. NAPS **[68]** tool also can be used to create protein networks, which allow interactive visualization of inter-residual interactions from both modeled proteins and MD simulation trajectories.

In these networks, amino acids are represented as nodes, and the connections between them are shown as edges. An edge is created between two nodes if their Cα-Cα distance is within approximately 7 Å.

NAPS allows users to analyze nodes based on centrality, physiochemical properties, and clusters of connected residues, helping to identify functional or coevolving residues and predict protein-protein interactions.

The degree of a node indicates the number of direct connections it has and hubs are key nodes with four or more connections, and they are considered dynamically stable if they appear as hubs in over 50% of MD simulation snapshots are considered dynamically stable and are often referred to as "hot spots" due to their significant role in preserving structural integrity and mediating allosteric communication. Changes in hub residue positions between glycan-bound and unbound states reflect structural alterations in glycoproteins.

Some residues are intra-linked, meaning they are connected to each other in a way that indicates structural rigidity. Communities are groups of closely connected nodes linked by common interactions. They help spread structural rigidity throughout the protein network. By comparing individual nodes and their communities, small conformational changes that affect the protein’s rigidity and flexibility can be identified. Communities within the network consist of residues that interact closely, and these communities help communication within the protein.

Overall, network analysis of protein structures provides valuable insights into the intricate interplay between residues, highlighting regions critical for structural stability and functional dynamics.

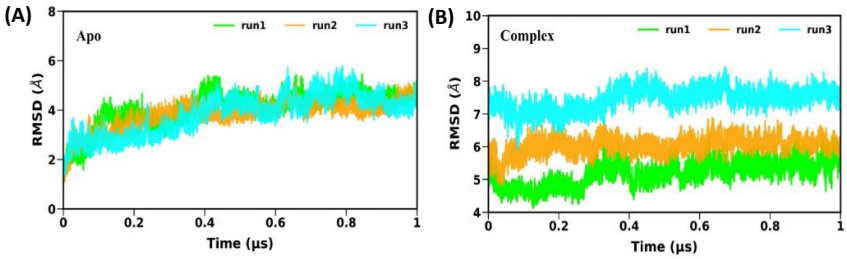
CHAPTER 5

5. Results and Discussion

5.1 Stability and convergence analysis of the protein systems

In the investigation of the E1-E2 structural dynamics, extensive Gaussian Accelerated Molecular Dynamics (GaMD) simulations were performed for the E1-E2 heterodimer for the apo and glycosylated (complex) systems, covering a time span of 1 μs in triplicate. During the 1 μs production simulations, the E1-E2 complex exhibited stability shown in the root-mean-squared deviations (RMSD) from the initial structure. The (Figure 17) shows the time evolution of the root mean square deviation (RMSD) for the backbone atoms in each system concerning the initial configurations.

As shown in panel A of **Figure 5.1**, the apo form exhibits a gradual increase in RMSD, stabilizing around 4–5 Å after the first 0.4 μs, indicating moderate flexibility and convergence across all the three runs. In contrast, the complex in panel B of **Figure 5.1** form displays higher and more variable RMSD values ranging from 5 to 8 Å, with each run stabilizing at different levels indicates differing convergence behaviour. Higher RMSD values indicate larger structural fluctuations suggesting that the complex undergoes larger conformational rearrangements.



Time evolution of Root-mean-squared deviations (RMSD)

Figure 5.1: (A) Time evolution of the root-mean-square deviation (RMSD) of the E1-E2 complex structure in Apo system. (B) Time evolution of the root-mean-square deviation (RMSD) of the E1-E2 complex structure in Complex system.

5.2 Structural Stability Analysis of E1-E2 complex

To further investigate structural fluctuations, RMSD-based probability density plots were generated for both apo and complex systems as shown in **Figure 5.2**, Panel A. The apo form displays a broader distribution with three distinguishable peaks centered around ~4.5 Å, ~5.6 Å, and ~6.8 Å, suggesting transitions between multiple conformational states. In contrast, the complex showed narrower, more sharply defined peaks at higher RMSD values (~5.7 Å and ~7.5 Å), indicating fewer but more distinct conformational states. Although the complex shows higher RMSD values, this does not necessarily mean it is more flexible. Instead, it appears to adopt fewer and well-defined conformational states, implying a more conformationally restricted but stable structure, likely maintained by stabilizing interactions formed during binding. In contrast, the apo state explores a wider range of conformations, indicating higher structural variability.

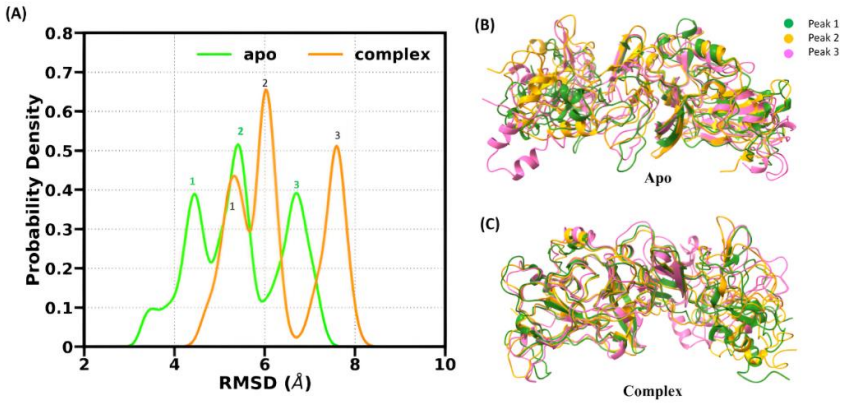


Figure 5.2: (A) Probability density plot of backbone RMSD for Apo (green) and Complex (orange) systems over the simulation period. (B) Structural overlays of representative frames from the three RMSD peaks in the apo system, colored by peak: Peak 1 (green), Peak 2 (orange), and Peak 3 (pink). (C) Structural overlays of representative frames from the complex system.

5.3 Conformational Dynamics of Glycans

Figure 5.3 shown below illustrates the distinct conformational behaviors of two glycans attached to the protein and a free glycan. The panel A displays the RMSD probability density distributions, revealing that glycan 2 (orange) has a sharp and narrow RMSD peak**,** suggesting that it remains in a more stable and restricted conformation during the simulation, likely due to spatial constraints or stronger interactions at its attachment site (N305). In contrast, glycan 1 (attached at N196) and the free glycan exhibit broader distributions, indicating they undergo greater conformational fluctuations and are more flexible.

The Panel B further supports these observations by showing the structural positioning of the glycans on the protein surface, where glycan 2 appears more embedded within the protein interface, potentially contributing to its limited mobility. The glycan 2 experiences reduced conformational dynamics relative to glycan 1 and the free glycan.

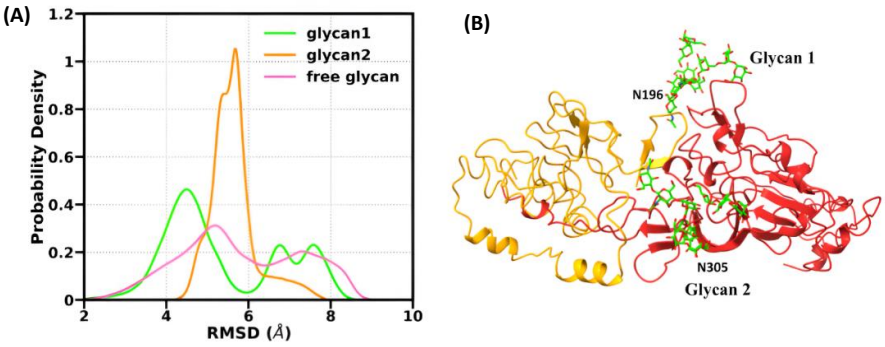


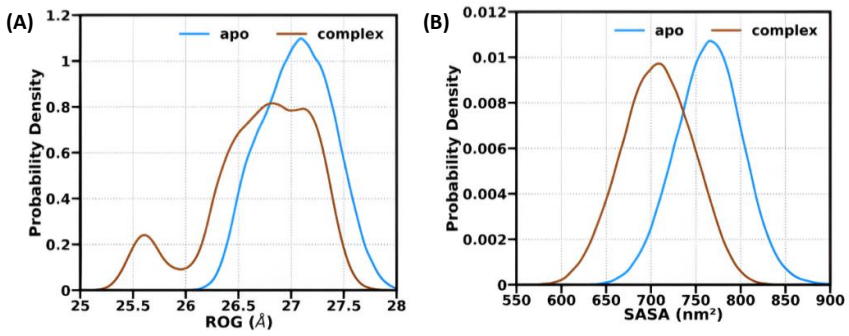
Figure 5.3: (A) Probability density distribution of RMSD for glycan 1 (green), glycan 2 (orange), and the free glycan (pink) throughout the simulation. (B) Structural representation of the glycoprotein complex highlighting the positions of glycan 1 (N196 site) and glycan 2 (N305 site), with glycans shown in green sticks.

5.4 Solvent accessibility and protein compactness

Figure 5.4 illustrates comparative analyses of the radius of gyration (RoG) and solvent-accessible surface area (SASA) between the apo and complex systems.

The panel A shows the RoG distributions, where the complex (brown line) exhibits a broader and more variable profile, suggesting a slightly more compact and structurally diverse arrangement compared to the apo system (blue line), which shows a sharp peak around 27 Å.

The panel B displays SASA distributions, where the complex form has lower solvent exposure than the apo form, indicating reduced surface accessibility upon complex formation. The apo form (blue) shows a broader distribution (650–900 nm^2), while the complex (brown) shows with a narrower range of 600–830 nm^2. This reduction in solvent exposure is typically attributed to the formation of intermolecular interactions upon glycan attachment, which results in burying previously exposed hydrophobic residues within the protein-protein interface. These results indicate that complex formation results in a structurally more compact configuration.



RoG: Radius of gyration**; SASA:** Solvent accessible surface area

Figure 5.4: (A) Probability density distribution of the radius of gyration (RoG) for the Apo (blue) and Complex (brown) forms of the protein. (B) Probability density distribution of the solvent accessible surface area (SASA) for the Apo and Complex systems.

5.5 Conformational Stability of E1 and E2

Separate plots were generated for E1 and E2 for both apo and complex systems, displaying the RMSD probability distribution in **Figure 5.5**.

For E1 (Panel A), the apo (green) system exhibits a broader distribution with multiple peaks across a wider RMSD range (~3–9 Å), suggesting higher conformational flexibility. In contrast, the complex form shows narrower and more defined peaks clustered around higher RMSDs (~6–11 Å), indicating a reduction in structural variability upon complex formation, but those conformations are somewhat more deviated from a reference structure.

For E2 in the apo form displays broader distribution with peaks starting at around 2.5 Å and extending up to ~5.8 Å, indicating higher flexibility and structural diversity, while the complex form has sharper and more compact peaks between ~2–4.5 Å, indicating greater stability and reduced conformational variability upon complexation.

It can be interpreted that the E1 protein becomes more structurally stable (less flexible) but shows a greater conformational shift when part of the complex while, the E2 protein becomes more conformationally stable in the complex while maintaining a structure closer to its reference. Overall, both proteins exhibit greater structural stability in the

complexed state compared to their unbound forms.

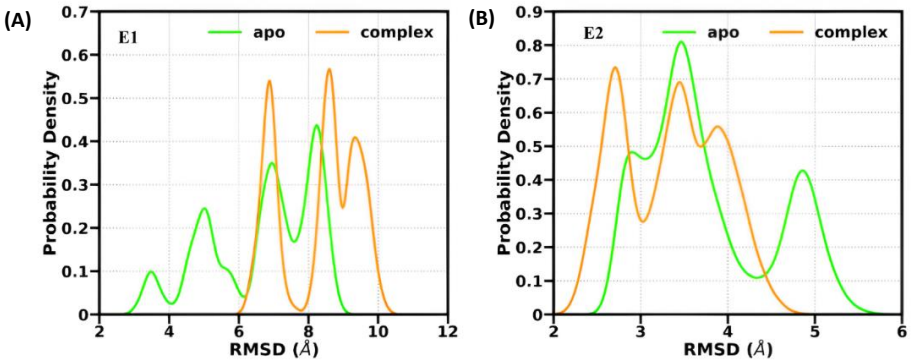


Figure 5.5: (A) Probability density plot of backbone RMSD of E1 for Apo (green) and Complex (orange) systems across the simulation timeframe. (B) Probability density plot of backbone RMSD of E2 for Apo and Complex systems across the simulation timeframe.

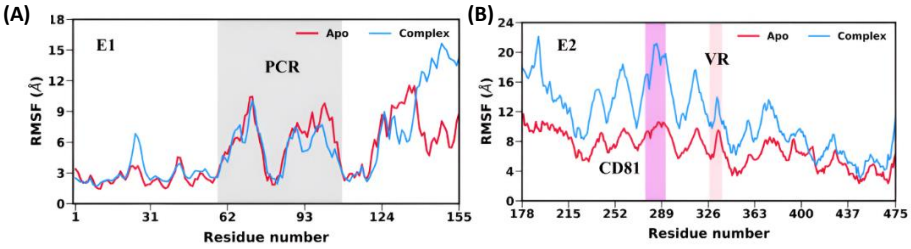
5.6 Residual Flexibility Analysis of E1 and E2 glycoproteins

Figure 5.6 presents the Root Mean Square Fluctuation (RMSF) profiles for the E1 and E2 proteins in their apo (red) and complex (blue) forms, indicating the residue-wise flexibility across the sequences.

For E1, the RMSF values are generally comparable between the apo and complex states across most of the sequence. However, increased fluctuations are observed in the complex form, particularly in the stem region (after residue 124), suggesting enhanced flexibility in this region upon complex formation. Within the PCR (Pfp-containing region), both forms show similar fluctuation patterns, indicating that this region remains relatively stable in both states.

In the case of E2, there is a marked increase in flexibility in the complex form across several regions, especially between residues ~250 and 370. Notably, the CD81 binding region and the variable region (VR) show significantly higher RMSF values in the complex, indicating increased mobility. In contrast, the apo form maintains lower and more consistent flexibility.

This can be interpreted as the E2 protein becoming more dynamic upon complex formation, particularly in functionally important regions, while E1 shows only limited flexibility changes, suggesting differing roles or structural responses of the two proteins upon binding.



RMSF: Root-mean-squared fluctuation

Figure 5.6: RMSF profiles of E1 and E2 proteins in apo (red) and complex (blue) systems. (A) Residue-wise fluctuations of E1. (B) Residue-wise fluctuations of E2.

5.7 Analysis of Intra Domain Regions of E1

5.7.1 PCR Region

The PCR region (residues 249 to 299) located in the core of E1 plays an important role in the fusion process of viral membrane with endosomal membrane to release the RNA genome into the host cell. Additionally, it may contribute to the assembly and structural organization of the viral particle. We have plotted the probability distribution of the PCR region in **Figure 5.7** to get an insight into its structural stability.

The panel A presents the RMSD distribution of the E1 putative contact region (PCR) in both apo (green) and complex (orange) systems, reflecting its structural deviation over time. The apo form displays a broader and bimodal RMSD distribution at around 3.8 and 8 Å, suggesting that the PCR is more conformationally flexible in the absence of binding. In contrast, the complex form exhibits a narrow and more defined distribution centered around higher RMSD values (~8–9 Å), indicating a more shifted but more stable and consistent conformation. This suggests that upon complex formation, the PCR becomes

structurally more restrained and adopts a specific conformation, which may be important for its interaction role within the E1-E2 assembly.

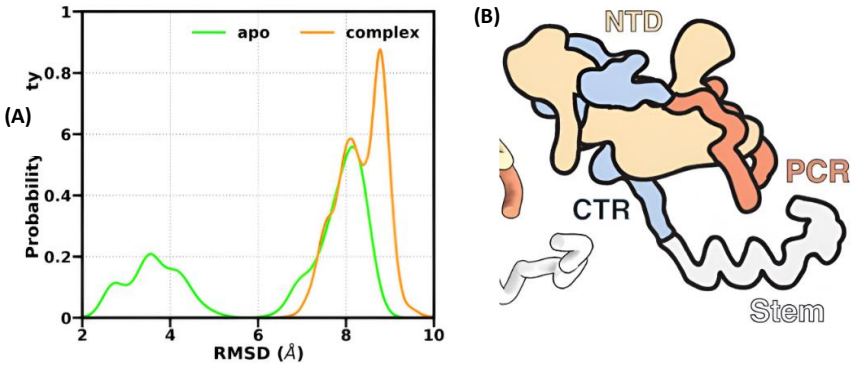


Figure 5.7: (A) Probability distribution plot of RMSD for the PCR region in E1 in apo (green) and complex (orange) systems. (B) Schematic representation highlighting the PCR within the E1 domain organization, including the N-terminal domain (NTD), C-terminal loop region (CTR), and stem **[62]**.

5.8 Analysis of Intra Domain Regions of E2

5.8.1 CD81 binding site

The CD81 binding site (amino acids 518 to 534) located in the head of E2 is essential for binding to the CD81 receptor on the host cell membrane to initiate the infection process.

The RMSD (Root Mean Square Deviation) probability density plot for the E2-CD81 region in **Figure 5.8** illustrates the structural flexibility of this domain in the apo system (green) compared to the glycan-bound complex system (orange).

In the apo state, this region shows a sharp, high-intensity peak at low RMSD values (~0.8 Å), indicating high structural stability and minimal conformational deviation, with a tail extending upto ~5 Å and a smaller secondary peak at around 3.5 Å reflects a low-probability population of transient or rare conformations that deviate notably from the reference structure. These deviations likely arise due to the loop structure of this region.

However, upon glycan binding, the RMSD distribution shifts, and two broader peaks emerge at higher RMSD values (~1.8 Å and ~4.5 Å), reflecting increased conformational variability. The shift in RMSD suggests that glycan interaction promotes dynamic structural rearrangement or flexibility in this region upon binding CD81.

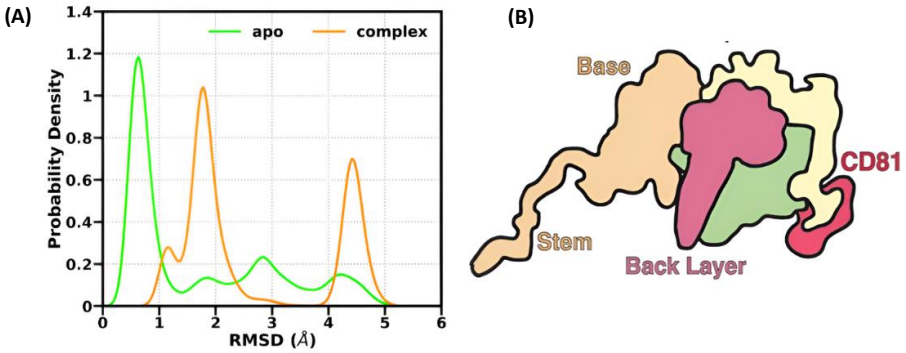


Figure 5.8: (A) RMSD probability density plot for the CD81 binding region of E2 in apo (green) and complex (orange) systems. (B) Schematic representation of the E2 subdomains highlighting the CD81 binding site **[62]**.

5.8.2 Variable Regions – VR2, VR3

The Variable regions (VR2, residues 459 to 483; VR3, residues 569 to 579) located in the head of E2 allows the virus to escape host's immune system recognition and aids in immune evasion.

Figure 5.9 illustrates how glycan binding influences the conformational flexibility of the E2 glycoprotein's variable regions VR2 and VR3 in hepatitis C virus. Probability density plots of RMSD values are shown for each region in the apo (green) and glycan-bound (orange) states.

The RMSD probability density plot for VR2 (panel A) indicates that the glycan-bound distribution shows two sharp peaks at lower RMSD values (~0.9 Å and ~1.7 Å), indicating stable conformations, whereas the apo state (green) exhibits a broader distribution with peaks at around (~1.9 Å and ~12.6 Å), suggesting greater structural flexibility.

The RMSD probability density plot for VR3 (panel B) shows that glycan

binding reduces the flexibility of this region. In the apo state (green), VR3 displays a broader distribution with a peak around 1.8 Å, indicating higher structural variability. In contrast, the glycan-bound state (orange) shows a shift toward lower RMSD values, with a sharper peak near 1.4 Å, suggesting that VR3 adopts more stable conformations when glycans are present.

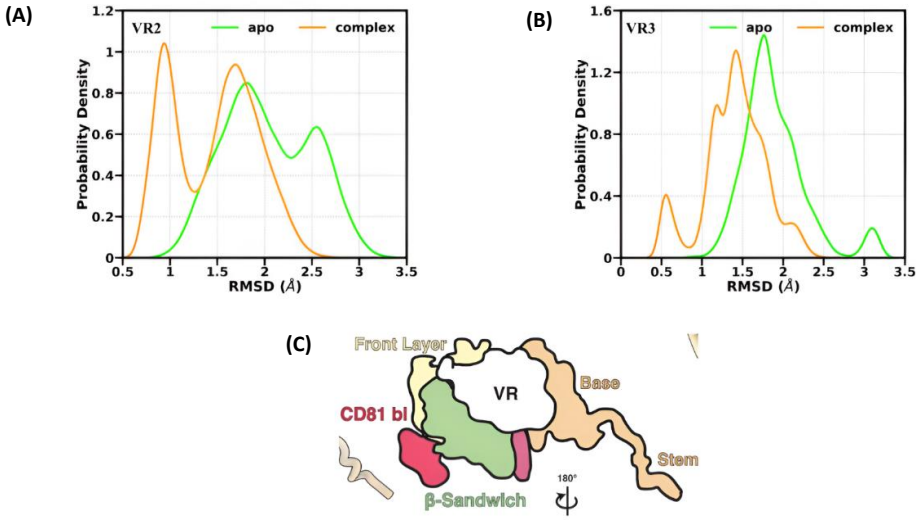


Figure 5.9: (A) Probability density plot of RMSD distribution of E2 variable region VR2 in apo (green) and glycan-bound complex (orange). (B) Probability density plot of RMSD distribution of VR3 in apo and glycan-bound states. (C) Schematic representation of the E2 domain organization highlighting variable regions (VR) along with, CD81 binding loop (CD81 bl), and structural domains including the front layer, β-sandwich, base, and stem **[62]**.

5.9 E1–E2 binding interaction

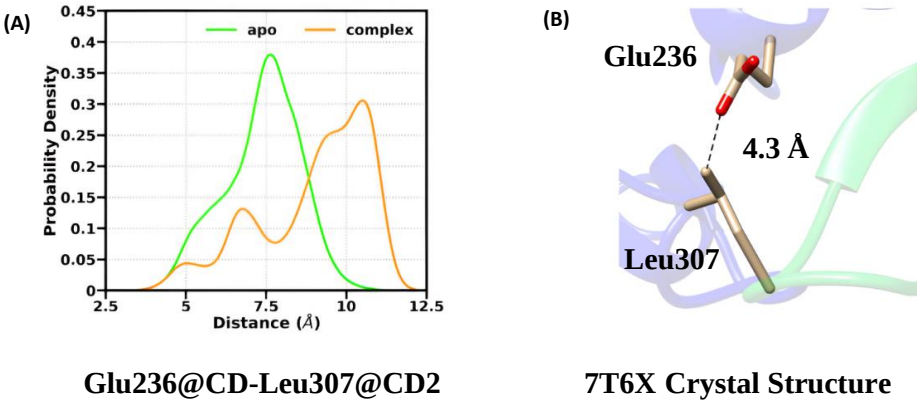
Earlier structural studies, including cryo-EM, have captured E1–E2 in static conformations, but molecular simulations reveal their dynamic nature and offer a better understanding of how the complex behaves in different states. **Figure 5.10** illustrates the structural impact of glycan binding on the E1–E2 interface.

The panel A shows the probability density plot of the inter-residue distance between Glu236 (E2) and Leu307 (E1), revealing a rightward shift in the distance distribution upon glycan binding. Specifically, the average distance increases from approximately 7.2 Å in the apo state to 8.1 Å in the glycan-bound complex can be seen in Panel C, indicating a weakening of the E1–E2 connection after glycan binding.

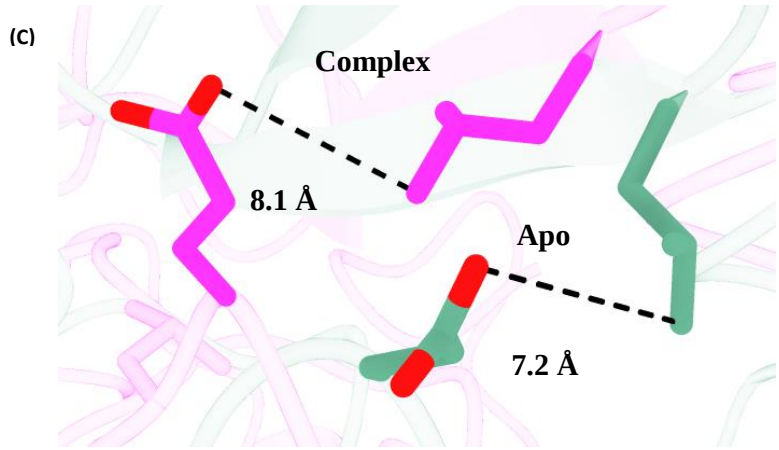
The panel B presents the crystal structure (PDB ID: 7T6X), where Glu655 (E2) and Leu200 (E1) are in close proximity (4.3 Å), highlighting a native E1–E2 contact at the interface.

Panel C further supports the simulation-based observation, showing that glycan interaction leads to increased spatial separation between E1 and E2 residues compared to the apo system.

Collectively, these results suggest that glycan engagement induces a conformational rearrangement at the E1–E2 interface, which may modulate the structural integrity or dynamics of the heterodimer.



Glu236@CD-Leu307@CD2 **7T6X Crystal Structure**



Post-Simulation

Figure 5.10: (A) Probability density plot showing the distribution of Glu236–Leu307 inter-residue distances in apo and glycan-bound (complex) systems. (B) Interaction observed between Glu655 and Leu200 in the E1–E2 crystal structure (PDB: 7T6X). (C) Post-simulation structural comparison of E1–E2 distance highlighting changes in the distance of apo and complex systems.

5.10 Hydrophobic Interactions between E1 and E2

The RMSD distributions shown in panels A and B of **Figure 5.11** reveals distinct differences in the conformational dynamics of hydrophobic residues between the apo and complex states.

For interaction study we took residues Phe167, Pro171, Phe260, Thr262, Leu270 of hydrophobic cavity of base in E2 interacts with residues Tyr299, Val310, Tyr308, Val301 of E1. In panel A, the complex state (orange) exhibits a sharp peak around 1 Å, indicating a more rigid and structurally conserved conformation compared to the apo state (green), which shows a broader distribution with a peak closer to 1.8 Å. This suggests that complex formation significantly stabilizes the structure of the hydrophobic region under investigation.

We then took residues Leu235, Tyr282, Tyr284 of E2 stem interacting with residues Ile415, Tyr416, Pro417, His419, and Met425 of E1. The panel B shows a wider RMSD range overall, with both apo and complex states displaying multiple peaks. The apo state has broader distrbution

with peaks at around 3 Å, 3.8 Å and 5.5 Å, while the complex state shows narrower peaks at ~3 Å and ~5 Å accompanying a small peak at around ~2.5 Å. Although some flexibility remains, the more defined peaks in the complex state indicate partial structural stabilization. This implies partial stabilization upon complex formation, but with retained flexibility in certain regions.

Together, these data highlight that complex formation generally leads to a reduction in conformational variability of key hydrophobic residues, likely contributing to the structural integrity and functional relevance of the protein interface.

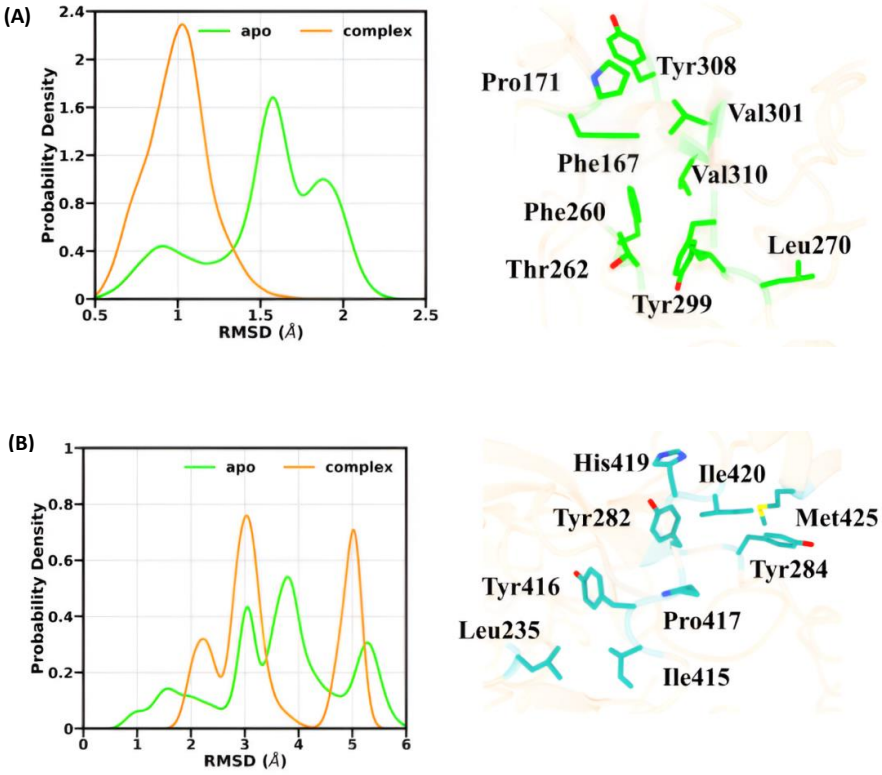


Figure 5.11: Probability density plots showing RMSD distributions of hydrophobic interface residues in E1 and E2 for apo (green) and complex (orange) systems obtained for two specific set of residues identified in earlier structural studies. **Right:** Visualization of two specific set of key hydrophobic residues at the E1 and E2 interface contributing to inter-subunit stabilization are highlighted and labelled.

5.11 E1-E2 interaction profile in apo structure

In the apo system of the E1-E2 heterodimer complex as shown below in **Figure 5.12**, critical interactions, such as electrostatic and hydrophobic interactions, play a key role in binding and stabilizing the two proteins together. This figure illustrates the interaction landscape between the E1 and E2 glycoproteins in the apo state, generated using LigPlot. The analysis reveals a complex network of non-covalent interactions that stabilize the E1–E2 interface, including hydrogen bonds, hydrophobic contacts, and electrostatic interactions. Each residue involved is annotated, and the types of interactions are depicted using standardized LigPlot symbols, allowing for a residue-level understanding of the interface architecture. Several residues from both E1 and E2 engage in stabilizing contacts, such as salt bridges and polar interactions, which are critical for maintaining the native structural integrity of the complex. The presence of recurring polar residues and charged side chains at the interface indicates a prominent role for electrostatic complementarity in mediating E1–E2 association. For example, side chains such as Arg, Glu, Asp, and Lys engage in salt bridges and hydrogen bonding that span across the interface, contributing significantly to structural integrity. Additionally, hydrophobic patches involving residues such as Val, Leu, Ile, and Phe may contribute to van der Waals interactions that further stabilize the complex.

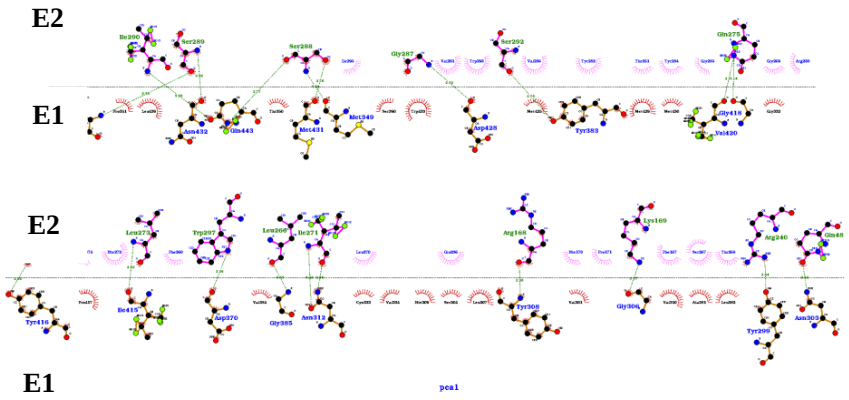


Figure 5.12: 2D Interaction map depicting residue-level contacts between E1 and E2 at the interface in the apo conformation, highlighting electrostatic interactions (hydrogen bonds), and hydrophobic interactions. In the figure, red regions represent hydrophobic interactions, while green dotted lines highlight electrostatic interactions.

5.12 E1-E2 interaction profile in complex structure

Figure 5.13 presents a detailed interaction profile between the E1 and E2 glycoproteins in the complex state, generated using LigPlot. It captures the array of inter-residue contacts formed upon complex formation, providing a two-dimensional visualization of the molecular interface. Compared to the apo structure, a marked shift in interaction character is evident that hydrophobic contacts are notably more abundant and widespread across the interface. Residues such as Leu, Val, Ala, and Phe are observed clustering together, forming a hydrophobic core that likely enhances the stability of the E1–E2 association in the complex state. Additionally, hydrogen bonds continue to contribute to the interface, but the prominence of van der Waals and hydrophobic interactions suggests a restructuring of molecular forces upon complex formation. This rearrangement may reflect a conformational stabilization that takes place when E1 and E2 associate under the influence of external elements, such as glycan interactions. The Lig-Plot visualization offers an effective representation of these changes, illustrating how the molecular contacts between E1 and E2 are highly dependent on the surrounding environment. Such insights are

crucial for understanding the dynamic nature of viral envelope assembly and could have implications for therapeutic strategies aimed at disrupting E1–E2 interactions.

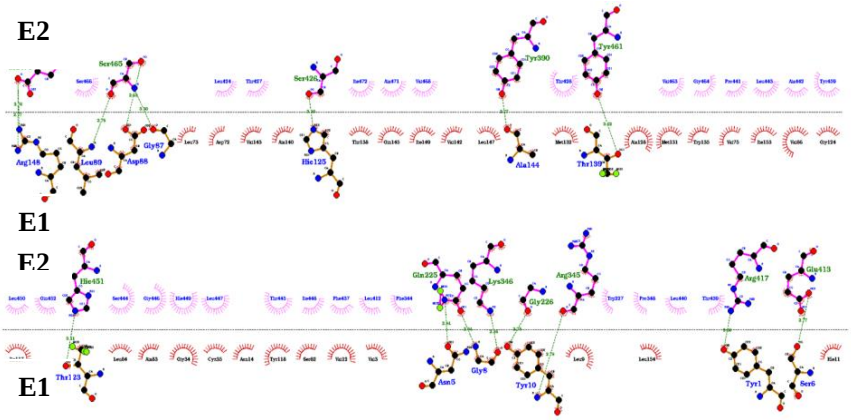


Figure 5.13: 2D Interaction map depicting residue-level contacts between E1 and E2 at the interface in the complex conformation, highlighting electrostatic interactions (hydrogen bonds), and hydrophobic interactions. In the figure, red regions represent hydrophobic interactions, while green dotted lines highlight electrostatic interactions.

5.13 Hydrogen bonds in protein-glycan interaction

Figure 5.14 illustrates the variation in the number of hydrogen bonds formed by Glycan1 (orange) and Glycan2 (green) throughout three separate molecular dynamics simulation runs (Run1, Run2, and Run3). And the background colors (pink, blue, and purple) separate the three simulation runs. The horizontal axis denotes the simulation time in microseconds (μs), while the vertical axis shows the corresponding number of hydrogen bonds formed during the simulations.

The data clearly show that Glycan 2 consistently forms a significantly higher number of hydrogen bonds compared to Glycan1 throughout all simulation runs, indicating greater interaction with the E1 and E2 heterodimer complex. Glycan 2 maintains a stable range of 15–20 hydrogen bonds, suggesting persistent and strong interactions with its environment. In contrast, Glycan 1 forms fewer hydrogen bonds, ranging mostly between 2–10, and exhibits greater fluctuation,

indicating weaker or less consistent interactions. This consistent trend across all three runs supports the conclusion that Glycan 2 plays a more prominent role in stabilizing the protein structure through hydrogen bonding. These observations suggest that Glycan2 may contribute more significantly to the conformational stability and functional dynamics of the glycoprotein complex.

The reason behind this might be the positional difference between the two glycans. Glycan2 (linked at N305) is situated closer to the interfacial region between the two protein domains, placing it in a more confined and interaction-rich environment. This positioning enables Glycan2 to establish a greater number of hydrogen bonds with nearby residues, contributing to its enhanced stability, as reflected in the hydrogen bond analysis. In contrast, Glycan1 (linked at N196) is located further away from the interface, in a more exposed and flexible region of the protein. This spatial orientation limits its ability to form stable interactions with surrounding residues, resulting in fewer hydrogen bonds.

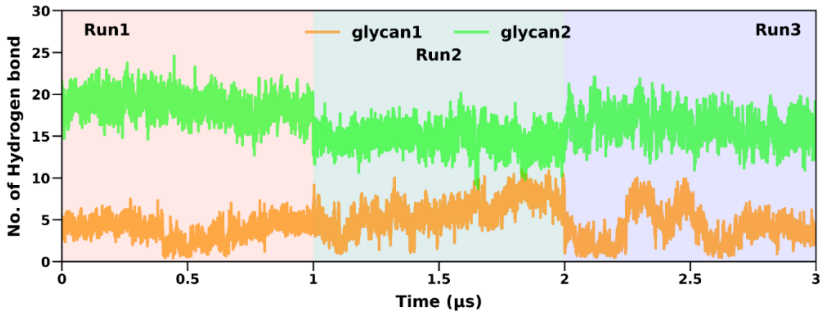


Figure 5.14: Hydrogen bond analysis of Glycan1 and Glycan2 across three simulation runs (Run1–Run3), showing the number of hydrogen bonds formed over time (µs).

The Table 2 below presents the hydrogen bond interactions between a protein and two glycans—Glycan 1 and Glycan 2, based on molecular dynamics simulation data. Each interaction is characterized by the donor and acceptor atoms involved, the occupancy percentage (indicating how frequently the hydrogen bond exists during the simulation during the simulation), and the average bond distance in angstroms (>3.0 Å).

Glycan 1 has only 2 hydrogen bonds with relatively lower occupancies (49.69% and 30.66%), whereas Glycan 2 has 10 hydrogen bond interactions, including the highest occupancy of 88.22% between THR_405@O and 0MA_177@O2, indicating a strong, stable interaction.

Table 5.1: Occupancy of hydrogen bonds between the glycan and protein over the course of the MD simulation.

Binding couple		**Molecular Dynamics**	
Acceptor	**Donor**	**Occupancy (%)**	**Distance (Å)**
Glycan 1			
4YB_156@O2N	ASN_5@ND2	49.69	2.84
GLN_225@OE1	4YB_156@O6	30.66	2.70
Glycan 2			
THR_405@O	0MA_177@O2	88.22	2.75
LEU_450@O	4YB_168@O3	57.90	2.71
ILE_454@O	VMB_169@O4	39.35	2.75
4YB_168@O3	ARG_415@NH2	38.59	2.85
4YB_167@O2N	ASN_114@ND2	36.22	2.84
0MA_177@O5	GLU_408@N	33.47	2.84
4YB_168@O2N	ARG_415@NE	27.77	2.83
VMA_170@O5	VAL_455@N	26.51	2.90
2MA_175@O6	ARG_409@N	24.95	2.86
4YB_168@O6	ARG_415@NH2	22.53	2.85

5.14 Dynamic Cross-Correlation Matrix (DCCM) analysis

The dynamic cross-correlation analysis in **Figure 5.15** highlights distinct differences in the motion of protein regions between the apo and complex systems. The PCR region (R1), associated with the N-terminal domain of E1, exhibits strong anti-correlated motion in the apo form, whereas it transitions to positively correlated motion upon complex formation. This shift suggests a stabilization and coordination of movement within the PCR region in the presence of the glycan or binding partner. Similarly, the CD81 binding region (R2) shows a clear contrast: in the apo system, this region demonstrates positive correlation, indicating synchronized movement with surrounding residues, but it shifts to negative correlation in the complex, implying a reversal in the direction of motion likely due to altered interaction dynamics. Interestingly, the VR region (R3) maintains a similar residual correlation pattern in both apo and complex forms, suggesting its dynamics remain relatively unaffected by complexation. Collectively, these observations point to glycan-induced modulation of the internal dynamics of specific protein regions, which may be critical for functional conformational transitions.

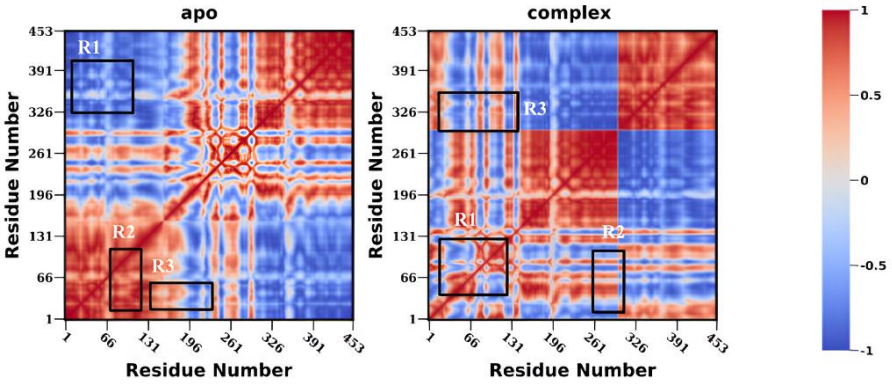


Figure 5.15: Dynamic Cross-Correlation Matrix (DCCM) analysis of protein residues in apo and complex forms.

5.15 Principal Component Analysis of E1-E2 Complex

In the free energy landscape (FEL) analysis based on principal component analysis (PCA), **Figure 5.16** illustrates the dominant motions of the protein in both apo and complex states using PC1 and PC2 axes. The color gradient reflects the free energy distribution, with dark red denoting the most stable (lowest energy) regions and blue indicating higher-energy, less stable conformations. In the apo state (Panel A), three distinct low-energy basins were observed, suggesting that the unbound form explores a wider range of conformational states, indicative of greater flexibility. In contrast, the complex state (Panel B) exhibited only two main energy basins, with a narrower and more confined energy surface, implying restricted dynamics and increased structural stability upon binding. The lower panels present representative structures from each energy minimum, where the apo conformations show greater divergence, while those from the complex state appear more compact and similar. These findings suggest that binding limits structural variability, stabilizing the protein in fewer, energetically favorable states.

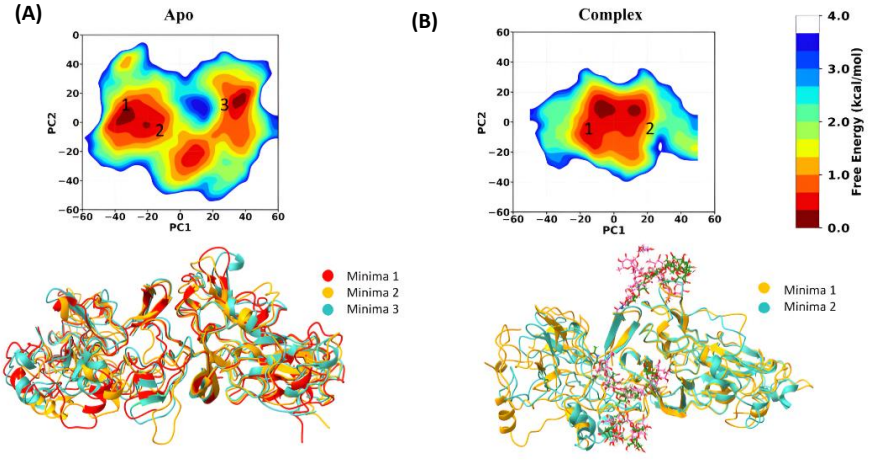


Figure 5.16: Principal Component Analysis (PCA) and Free Energy Surface (FES) of the protein in apo and complex systems.

5.16 Protein Structure Network Analysis

Figure 5.17 presents a comparative analysis of **Residue Connectivity via Protein Structure Network (PSN)** between the **Apo** and **Complex** states of the protein. Each color represents a distinct community of residues, with the most to least populous communities labeled from red to light orange, respectively. In the **Apo form**, residues surrounding the E2 region are grouped into several smaller, relatively localized communities (e.g., red, green, and orange), with limited inter-community communication, suggesting less coordinated structural behavior. However, upon complex formation, a noticeable shift in network organization is observed. In the **Complex**, the red community (the most populous) near the E2 region becomes more densely connected, indicating stronger communication among residues in this region. Also, in the Complex, the residue network undergoes significant reorganization, with the glycan-associated region integrating into the 2nd most populous community (green), indicating its role in enhancing structural stability through strengthened interactions. Additional residue groups also reorganize or emerge, particularly around regions E1 and E2, implying a dynamic restructuring of the interaction network. These changes signify increased inter-residue connectivity, likely driven by glycan interactions, which in turn could enhance the overall structural stability and promote functional coordination within the protein.

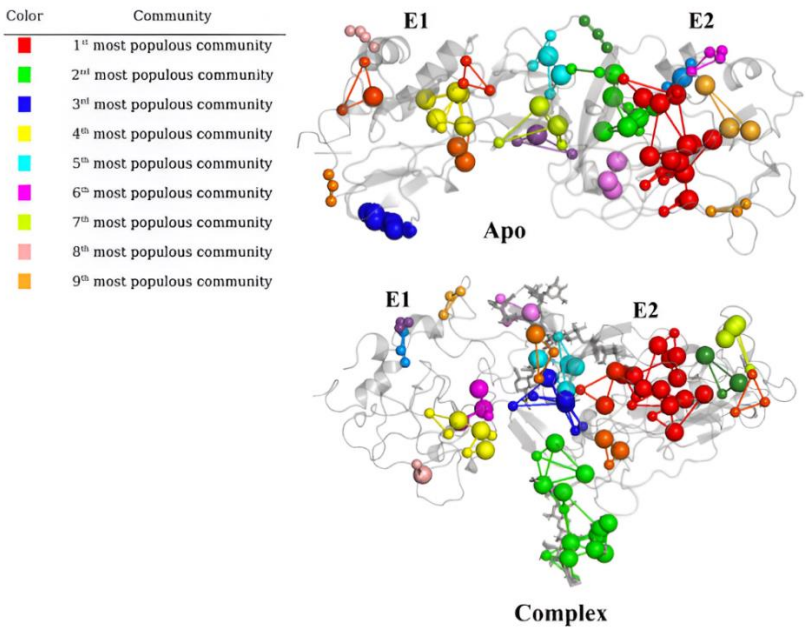


Figure 5.17: Residue Connectivity via Protein Structure Network for apo and complex systems depicting hubs, links and communities.

Table 5.2: Comparison of network properties between Apo and Complex systems.

Network Properties	**Apo**	**Complex**
I_{min}	3.69	4.16
Number of Linked Nodes	420	447
Number of Links	484	509
Number of Hubs	67	75
Number of Links mediated by Hubs	251	289
Number of Communities	18	17
Number of Nodes involved in Communities	92	91
Number of Links involved in Communities	118	124

CHAPTER 6

6. Conclusions and scope for future work

6.1 Conclusions

Our findings demonstrate that glycosylation induces significant conformational stabilization in the protein complex, with a notable difference in stability between the two domains. Specifically, the E2 domain exhibits better structural stability than E1, as reflected by more defined residue communities and reduced flexibility. Despite this, key intra-domain regions within both E1 and E2 maintain dynamic behavior, suggesting localized flexibility important for function. Our RMSD analysis showed that hydrophobic interactions become more structurally stable upon complex formation, indicating reduced flexibility and tighter packing, particularly in the complex state, thus playing a crucial role in mediating the E1–E2 complex formation, supporting their contribution to interface stability. In particular, E2 displays more tightly interconnected residue communities in both apo and complex forms, further emphasizing its stabilizing role. Moreover, the embedded glycan at the domain interface enhances connectivity, likely contributing to the overall structural integrity of the complex. The principal component analysis reveals that the complex form exhibits reduced conformational flexibility compared to the apo form, indicating a more stable and compact structure upon binding. The dynamic cross-correlation analysis further supports this, exhibiting a shift in motion patterns where the PCR region displays reduced anti-correlation and the CD81 region transitions from positive correlation in the apo to negative correlation in the complex. Additionally, protein structure network analysis shows tighter and more interconnected residue communities in the complex, particularly around the glycan-associated region and E2, reinforcing the role of glycan interactions in enhancing structural stability and modulating functional dynamics.

6.2 Future Work

In this study, we investigated the structural dynamics and interaction landscape of a glycosylated E1–E2 protein complex, where two glycans were binded to the E1 (at residues N196 and N305) at the interface of the heterodimer complex.

For future studies, it is essential to examine all glycosylation sites in E1 (5) and E2 (11) to better understand their influence on complex dynamics. Additionally, introducing site-specific glycan mutations could help elucidate the role of individual glycans in modulating E1–E2 conformational behavior and stability. Molecular dynamics simulations will be extended to longer timescales to capture slow conformational changes and provide a more comprehensive understanding of glycan flexibility and its influence on the dynamic behavior of the E1-E2 complex. Glycan-mediated shielding effects will be systematically investigated to determine how specific glycosylation patterns obscure antigenic epitopes and contribute to immune evasion by the virus. The effect of glycan modifications on the receptor-binding affinity of the E1-E2 complex will be thoroughly investigated to elucidate how specific glycosylation patterns influence viral attachment and entry into host cells. Then, advanced trajectory analysis tools will be employed to monitor local and global structural rearrangements induced by glycan dynamics.

This will allow a deeper understanding of how glycosylation influences protein behaviour at both the structural and functional levels, potentially informing strategies for therapeutic targeting or vaccine design.

References

[1] P. Mehta, L. M. Grant, and A. K. R. Reddivari, “Viral Hepatitis,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: Apr. 30, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK554549/

[2] A. Petruzziello, S. Marigliano, G. Loquercio, A. Cozzolino, and C. Cacciapuoti, “Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes,” *World J. Gastroenterol.*, vol. 22, no. 34, p. 7824, 2016, doi: 10.3748/wjg.v22.i34.7824.

[3] “Hepatitis C: Practice Essentials, Background, Pathophysiology,” Mar. 2025, Accessed: Apr. 30, 2025. [Online]. Available: https://emedicine.medscape.com/article/177792-overview?form=fpf

[4] “WHO commends Egypt for its progress on the path to eliminate hepatitis C.” Accessed: Apr. 30, 2025. [Online]. Available: https://www.who.int/news/item/09-10-2023-who-commends-egypt-for-its-progress-on-the-path-to-eliminate-hepatitis-c

[5] M. Chandra, A. A. Paray, and K. Arora, “Prevalence of hepatitis C virus infection in India: a systematic review,” *Int. J. Res. Med. Sci.*, vol. 12, no. 7, pp. 2529–2536, Jun. 2024, doi: 10.18203/2320-6012.ijrms20241906.

[6] “India has second-most hepatitis B, C cases after China: WHO report - The Hindu.” Accessed: Nov. 29, 2024. [Online]. Available: https://www.thehindu.com/sci-tech/health/who-sounds-alarm-on-viral-hepatitis-infections-claiming-3500-lives-every-day/article68048999.ece

[7] J. Christdas, J. Sivakumar, J. David, H. Daniel, S. Raghuraman, and P. Abraham, “Genotypes of hepatitis C virus in the Indian sub-continent: A decade-long experience from a tertiary care hospital in South India,” *Indian J. Med. Microbiol.*, vol. 31, no. 4, pp. 349–353, Oct. 2013, doi: 10.4103/0255-0857.118875.

[8] D. R. Taylor, S. T. Shi, and M. M. C. Lai, “Hepatitis C virus and interferon resistance,” *Microbes Infect.*, vol. 2, no. 14, pp. 1743–1756, Nov. 2000, doi: 10.1016/S1286-4579(00)01329-0.

[9] M. A. Maqbool, “Impact of Hepatitis C Virus NS5A Genetic Variability on Liver Pathogenesis and Viral Replication,” Jan. 2012.

[10] L. B. Dustin, B. Bartolini, M. R. Capobianchi, and M. Pistello, “Hepatitis C virus: life cycle in cells, infection and host response, and analysis of molecular markers influencing the outcome of infection and response to therapy,” *Clin. Microbiol. Infect.*, vol. 22, no. 10, pp. 826–832, Oct. 2016, doi: 10.1016/j.cmi.2016.08.025.

[11] “Matched Antigen Pairs for HCV Serology Test - Creative Diagnostics.” Accessed: Nov. 29, 2024. [Online]. Available: https://www.creative-diagnostics.com/news-matched-antigen-pairs-for-hcv-serology-test-127.htm

[12] H.-C. Li, C.-H. Yang, and S.-Y. Lo, “Cellular factors involved in the hepatitis C virus life cycle,” *World J. Gastroenterol.*, vol. 27, no. 28, pp. 4555–4581, Jul. 2021, doi: 10.3748/wjg.v27.i28.4555.

[13] S. Modrow, D. Falke, U. Truyen, and H. Schätzl, “Viruses with Single-Stranded, Positive-Sense RNA Genomes,” in *Molecular Virology*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 185–349. doi: 10.1007/978-3-642-20718-1_14.

[14] T. Pietschmann and R. J. P. Brown, “Hepatitis C Virus,” *Trends Microbiol.*, vol. 27, no. 4, pp. 379–380, Apr. 2019, doi: 10.1016/j.tim.2019.01.001.

[15] A. C. Araujo, I. V. Astrakhantseva, H. A. Fields, and S. Kamili, “Distinguishing Acute from Chronic Hepatitis C Virus (HCV) Infection Based on Antibody Reactivities to Specific HCV Structural and Nonstructural Proteins,” *J. Clin. Microbiol.*, vol. 49, no. 1, pp. 54–57, Jan. 2011, doi: 10.1128/JCM.01064-10.

[16] A. Nawaz, S. F. Zaidi, K. Usmanghani, and I. Ahmad, “Concise Review on the Insight of Hepatitis C.,” *J. Taiba Univ. Med. Sci.*, vol. 10, pp. 1–8, Dec. 2014, doi: 10.1016/j.jtumed.2014.08.004.

[17] T. Stroffolini and G. Stroffolini, “Prevalence and Modes of Transmission of Hepatitis C Virus Infection: A Historical Worldwide Review,” *Viruses*, vol. 16, no. 7, p. 1115, Jul. 2024, doi: 10.3390/v16071115.

[18] A. Geddawy, Y. F. Ibrahim, N. M. Elbahie, and M. A. Ibrahim, “Direct acting anti-hepatitis C virus drugs: Clinical pharmacology and future direction,” *J. Transl. Intern. Med.*, vol. 5, no. 1, pp. 8–17, Mar. 2017, doi: 10.1515/jtim-2017-0007.

[19] A. Torrents De La Peña *et al.*, “Structure of the hepatitis C virus E1E2 glycoprotein complex,” *Science*, vol. 378, no. 6617, pp. 263–269, Oct. 2022, doi: 10.1126/science.abn9884.

[20] T. D. Goddard *et al.*, “UCSF ChimeraX: Meeting modern challenges in visualization and analysis,” *Protein Sci.*, vol. 27, no. 1, pp. 14–25, 2018, doi: 10.1002/pro.3235.

[21] M. Zhang *et al.*, “Isolation, structures and biological activities of medicinal glycoproteins from natural resources: A review,” *Int. J. Biol. Macromol.*, vol. 244, p. 125406, Jul. 2023, doi: 10.1016/j.ijbiomac.2023.125406.

[22] K. Brandenburg and O. Holst, “Glycolipids: Distribution and Biological Function,” in *Encyclopedia of Life Sciences*, 1st ed., Wiley, 2015, pp. 1–10. doi: 10.1002/9780470015902.a0001427.pub3.

[23] “Glycosylation in health and disease | Nature Reviews Nephrology.” Accessed: Apr. 30, 2025. [Online]. Available: https://www.nature.com/articles/s41581-019-0129-4

[24] A. Corfield, “Eukaryotic protein glycosylation: a primer for histochemists and cell biologists,” *Histochem. Cell Biol.*, vol. 147, no. 2, pp. 119–147, Feb. 2017, doi: 10.1007/s00418-016-1526-4.

[25] Z. Yue *et al.*, “Advances in protein glycosylation and its role in tissue repair and regeneration,” *Glycoconj. J.*, vol. 40, pp. 1–19, Apr. 2023, doi: 10.1007/s10719-023-10117-8.

[26] A. W. Barb, A. J. Borgert, M. Liu, G. Barany, and D. Live, “Intramolecular Glycan–Protein Interactions in Glycoproteins,” in

Methods in Enzymology, vol. 478, Elsevier, 2010, pp. 365–388. doi: 10.1016/S0076-6879(10)78018-6.

[27] C. Chang, Y. Huang, L. Mueller, and W. You, "Investigation of Structural Dynamics of Enzymes and Protonation States of Substrates Using Computational Tools," *Catalysts*, vol. 6, no. 6, p. 82, May 2016, doi: 10.3390/catal6060082.

[28] D. A. Case *et al.*, "AmberTools," *J. Chem. Inf. Model.*, vol. 63, no. 20, pp. 6183–6191, Oct. 2023, doi: 10.1021/acs.jcim.3c01153.

[29] B. R. Brooks *et al.*, "CHARMM: The biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–1614, Jul. 2009, doi: 10.1002/jcc.21287.

[30] W. L. Jorgensen and J. Tirado-Rives, "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin," *J. Am. Chem. Soc.*, vol. 110, no. 6, pp. 1657–1666, Mar. 1988, doi: 10.1021/ja00214a001.

[31] M. J. Abraham *et al.*, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1–2, pp. 19–25, Sep. 2015, doi: 10.1016/j.softx.2015.06.001.

[32] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB," *J. Chem. Theory Comput.*, vol. 11, no. 8, pp. 3696–3713, Aug. 2015, doi: 10.1021/acs.jctc.5b00255.

[33] C. Tian *et al.*, "ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution," *J. Chem. Theory Comput.*, vol. 16, no. 1, pp. 528–552, Jan. 2020, doi: 10.1021/acs.jctc.9b00591.

[34] "CHARMM36m: an improved force field for folded and intrinsically disordered proteins | Nature Methods." Accessed: Apr. 30, 2025. [Online]. Available: https://www.nature.com/articles/nmeth.4067

[35] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, “Development and testing of a general amber force field,” *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, Jul. 2004, doi: 10.1002/jcc.20035.

[36] L. S. Dodda, J. Z. Vilseck, J. Tirado-Rives, and W. L. Jorgensen, “1.14*CM1A-LBCC: Localized Bond-Charge Corrected CM1A Charges for Condensed-Phase Simulations,” *J. Phys. Chem. B*, vol. 121, no. 15, pp. 3864–3870, Apr. 2017, doi: 10.1021/acs.jpcb.7b00272.

[37] K. N. Kirschner *et al.*, “GLYCAM06: A generalizable biomolecular force field. Carbohydrates,” *J. Comput. Chem.*, vol. 29, no. 4, pp. 622–655, Mar. 2008, doi: 10.1002/jcc.20820.

[38] E. J. Denning, U. D. Priyakumar, L. Nilsson, and A. D. Mackerell, “Impact of 2′-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA,” *J. Comput. Chem.*, vol. 32, no. 9, pp. 1929–1943, Jul. 2011, doi: 10.1002/jcc.21777.

[39] H. Grubmüller, H. Heller, A. Windemuth, and K. Schulten, “Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-range Interactions,” *Mol. Simul.*, vol. 6, no. 1–3, pp. 121–142, Mar. 1991, doi: 10.1080/08927029108022142.

[40] W. F. Van Gunsteren and H. J. C. and Berendsen, “A Leap-frog Algorithm for Stochastic Dynamics,” *Mol. Simul.*, vol. 1, no. 3, pp. 173–185, Mar. 1988, doi: 10.1080/08927028808080941.

[41] D. Beeman, “Some multistep methods for use in molecular dynamics calculations,” *J. Comput. Phys.*, vol. 20, no. 2, pp. 130–139, Feb. 1976, doi: 10.1016/0021-9991(76)90059-0.

[42] J.-P. Ryckaert, 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.*, vol. 23, no. 3, pp. 327–341, Mar. 1977, doi: 10.1016/0021-9991(77)90098-5.

[43] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, “LINCS: A linear constraint solver for molecular simulations,” *J.*

Comput. Chem., vol. 18, no. 12, pp. 1463–1472, Sep. 1997, doi: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[44] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, Jun. 1993, doi: 10.1063/1.464397.

[45] "Democritus: Periodic Boundary Condition." Accessed: Apr. 30, 2025. [Online]. Available: https://people.bath.ac.uk/chsscp/teach/md.bho/Theory/pbc-mi.html

[46] D. J. Evans and A. Baranyai, "The Gaussian thermostat, phase space compression and the conjugate pairing rule," *Mol. Phys.*, vol. 77, no. 6, pp. 1209–1216, Dec. 1992, doi: 10.1080/00268979200103081.

[47] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, Oct. 1984, doi: 10.1063/1.448118.

[48] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *J. Chem. Phys.*, vol. 126, no. 1, p. 014101, Jan. 2007, doi: 10.1063/1.2408420.

[49] H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature," *J. Chem. Phys.*, vol. 72, no. 4, pp. 2384–2393, Feb. 1980, doi: 10.1063/1.439486.

[50] R. W. Pastor, B. R. Brooks, and A. Szabo, "An analysis of the accuracy of Langevin and molecular dynamics algorithms," *Mol. Phys.*, vol. 65, no. 6, pp. 1409–1419, Dec. 1988, doi: 10.1080/00268978800101881.

[51] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, Oct. 1984, doi: 10.1063/1.448118.

[52] M. Parrinello and A. Rahman, “Polymorphic transitions in single crystals: A new molecular dynamics method,” *J. Appl. Phys.*, vol. 52, no. 12, pp. 7182–7190, Dec. 1981, doi: 10.1063/1.328693.

[53] G. J. Martyna, D. J. Tobias, and M. L. Klein, “Constant pressure molecular dynamics algorithms,” *J. Chem. Phys.*, vol. 101, no. 5, pp. 4177–4189, Sep. 1994, doi: 10.1063/1.467468.

[54] Y. Shi, “A Glimpse of Structural Biology through X-Ray Crystallography,” *Cell*, vol. 159, no. 5, pp. 995–1014, Nov. 2014, doi: 10.1016/j.cell.2014.10.051.

[55] P. R. L. Markwick, T. Malliavin, and M. Nilges, “Structural Biology by NMR: Structure, Dynamics, and Interactions,” *PLoS Comput. Biol.*, vol. 4, no. 9, p. e1000168, Sep. 2008, doi: 10.1371/journal.pcbi.1000168.

[56] J.-P. Renaud *et al.*, “Cryo-EM in drug discovery: achievements, limitations and prospects,” *Nat. Rev. Drug Discov.*, vol. 17, no. 7, pp. 471–492, Jul. 2018, doi: 10.1038/nrd.2018.77.

[57] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, “Comparison of simple potential functions for simulating liquid water,” *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, Jul. 1983, doi: 10.1063/1.445869.

[58] S. A. Hollingsworth and R. O. Dror, “Molecular Dynamics Simulation for All,” *Neuron*, vol. 99, no. 6, pp. 1129–1143, Sep. 2018, doi: 10.1016/j.neuron.2018.08.011.

[59] D. R. Roe and T. E. Cheatham, “PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data,” *J. Chem. Theory Comput.*, vol. 9, no. 7, pp. 3084–3095, Jul. 2013, doi: 10.1021/ct400341p.

[60] D. R. Roe and T. E. I. Cheatham, “PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data,” *J. Chem. Theory Comput.*, vol. 9, no. 7, pp. 3084–3095, Jul. 2013, doi: 10.1021/ct400341p.

[61] L. Skjaerven, A. Martinez, and N. Reuter, “Principal component and normal mode analysis of proteins; a quantitative comparison

using the GroEL subunit," *Proteins Struct. Funct. Bioinforma.*, vol. 79, no. 1, pp. 232–243, Jan. 2011, doi: 10.1002/prot.22875.

[62] A. Torrents De La Peña *et al.*, "Structure of the hepatitis C virus E1E2 glycoprotein complex," *Science*, vol. 378, no. 6617, pp. 263–269, Oct. 2022, doi: 10.1126/science.abn9884.

[63] N. Eswar, D. Eramian, B. Webb, M.-Y. Shen, and A. Sali, "Protein Structure Modeling with MODELLER," in *Structural Proteomics: High-Throughput Methods*, B. Kobe, M. Guss, and T. Huber, Eds., Totowa, NJ: Humana Press, 2008, pp. 145–159. doi: 10.1007/978-1-60327-058-8_8.

[64] E. F. Pettersen *et al.*, "UCSF Chimera—A visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004, doi: 10.1002/jcc.20084.

[65] Y. Miao, V. A. Feher, and J. A. McCammon, "Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation," *J. Chem. Theory Comput.*, vol. 11, no. 8, pp. 3584–3595, Aug. 2015, doi: 10.1021/acs.jctc.5b00436.

[66] D. Hamelberg, J. Mongan, and J. A. McCammon, "Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules," *J. Chem. Phys.*, vol. 120, no. 24, pp. 11919–11929, Jun. 2004, doi: 10.1063/1.1755656.

[67] A. Felline, M. Seeber, and F. Fanelli, "webPSN v2.0: a webserver to infer fingerprints of structural communication in biomacromolecules," *Nucleic Acids Res.*, vol. 48, no. W1, pp. W94–W103, Jul. 2020, doi: 10.1093/nar/gkaa397.

[68] B. Chakrabarty, V. Naganathan, K. Garg, Y. Agarwal, and N. Parekh, "NAPS update: network analysis of molecular dynamics data and protein–nucleic acid complexes," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W462–W470, Jul. 2019, doi: 10.1093/nar/gkz399.