Computational Modelling of Mucin-Nanoparticle Interactions

Adam Morrissey

UCD ID: 19359553

Thesis submitted to the faculty of the School of Physics at

University College Dublin in partial fulfilment

of the requirements for the degree of

MASTER OF SCIENCE

In

Computational Physics

Supervisor:

Assoc. Prof. Nicolae-Viorel Buchete

August 2024

Acknowledgements

I would like to extend my deepest gratitude to Assoc. Prof. Nicolae-Viorel Buchete for his invaluable mentorship and insightful guidance throughout this research project. My appreciation also goes to UCD Sonic for providing essential high-performance computational resources. I am thankful to the members of the Theoretical and Computational Biophysics group, especially Vigneshwari Annapoorani, for their technical support and engaging discussions. Above all, I am profoundly grateful to my family and partner for their enduring support and encouragement, without which this endeavour would not have been possible.

Statement of Original Authorship

I hereby certify that the submitted work is my own work, unless otherwise referenced, is entirely my own. The work was completed while registered as a candidate for the degree of Master of Science, and I have not obtained a degree elsewhere on the basis of the research presented in this submitted work.

Adam Morrissey

Table of Contents

Abstract

This study investigates the structural dynamics and interactions of MUC5AC mucin, a key component of mucus, in simulated environments with a focus on the effects of NaCl ions and silica nanoparticles (SiNPs). MUC5AC is integral to the formation of mucosal gels, which are essential for protecting and lubricating epithelial surfaces in the respiratory and gastrointestinal tracts. To understand the impact of environmental factors on MUC5AC's structural properties, we conducted molecular dynamics (MD) simulations that tracked global variables such as RMSD, Radius of Gyration, SASA, hydrogen bonds, and surface charge.

Our results indicate that MUC5AC maintains a stable core structure while its peripheral regions, including the tails and specific helical segments, exhibit notable conformational flexibility. This flexibility is essential for MUC5AC’s functional role in mucosal gel formation. Markov State Modelling (MSM) was employed to further explore MUC5AC’s conformational landscape, revealing discrete metastable states and providing insights into its adaptability to various physiological conditions.

Additionally, molecular docking and MD simulations assessed MUC5AC’s interactions with SiNPs of varying sizes. Larger SiNPs were found to exhibit stronger binding interactions with MUC5AC, influencing the protein's conformational stability and flexibility. These findings lay the groundwork for future studies aimed at extending simulation times to capture additional conformational states and exploring MUC5AC’s interactions with other biomolecules in the presence of NPs. This research contributes to the broader understanding of MUC5AC's role in biological systems and its response to environmental factors, with potential implications for drug delivery and nanomaterial safety.

List of Abbreviations

Abbreviation	Full Form
CHARMM	Chemistry at HARvard Macromolecular Mechanics (Web Server)
DSSP	Dictionary of Secondary Structure of Proteins
ETED	End-to-End Distance
HBOND	Hydrogen Bond
MD	Molecular Dynamics
MSM	Markov State Modelling
MUC5AC	Mucin-5AC
MUC5B	Mucin-5B
NAMD	Nanoscale Molecular Dynamics (Software)
NP	Nanoparticle
NS	Nanosphere
PBC	Periodic Boundary Conditions
PCA	Principal Component Analysis
PC	Protein Corona
PDB	Protein Data Bank
PME	Particle Mesh Ewald
RCSB	Research Collaboratory for Structural Bioinformatics
Rg	Radius of Gyration
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SASA	Solvent Accessible Surface Area
SASA+	Positive Solvent Accessible Surface Area
SASA-	Negative Solvent Accessible Surface Area
SASAH	Hydrophobic Solvent Accessible Surface Area
SEM	Standard Error of the Mean
TCL	Tool Command Language
TICA	Time-lagged Independent Component Analysis
VAMP	Variational Approach for Markov Processes
VMD	Visual Molecular Dynamics (Software)

List of Figures

List of Tables

1. Introduction

Nanoparticles (NPs) stand at the forefront of advanced material science, exhibiting unique properties that bridge the gaps between various disciplines. Their application spans from enhancing electronic devices to revolutionizing biomedical imaging and improving drug delivery systems. NPs are particularly praised for their potential in targeted drug delivery, which leverages their small size and surface modifiability to interact precisely at molecular and cellular levels. This capability allows for the development of highly specialized therapeutic agents that can directly target diseased cells while minimizing side effects to healthy tissues.

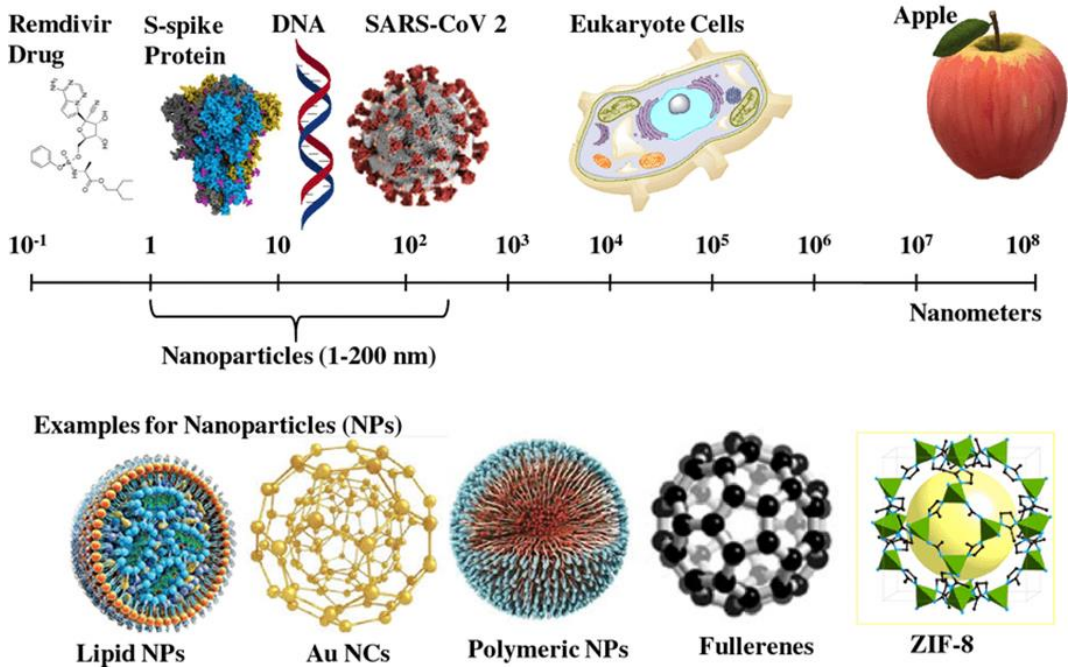


Figure 1: Schematic of Nanoparticle Sizes and Types

This diagram adapted from [1] illustrates the relative sizes of various NPs, biological molecules and objects, providing a visual comparison that shows the scale at which NPs operate

Upon introduction into biological systems, NPs encounter a complex milieu of biomolecules that rapidly adsorb onto their surfaces, forming what is known as a protein corona (PC). This corona fundamentally alters the NPs' surface characteristics and, consequently, their interactions within the body. The dynamic equilibrium of proteins binding and unbinding to

the NP surface affects not only how these particles are recognized by cellular receptors but also their subsequent cellular uptake, biodistribution, and overall pharmacokinetics.

The widespread application of silica NPs (SiNPs) across various industries, including biomedicine and electronics, necessitates a thorough understanding of their potential toxicological impacts on human health. SiNPs commonly enter the human body through inhalation, ingestion, and skin contact, dispersing broadly across different biological systems. Such pervasive exposure routes have been substantiated by studies like those of [2] which highlight the ease with which these NPs can distribute throughout the body, potentially accumulating in vital organs.

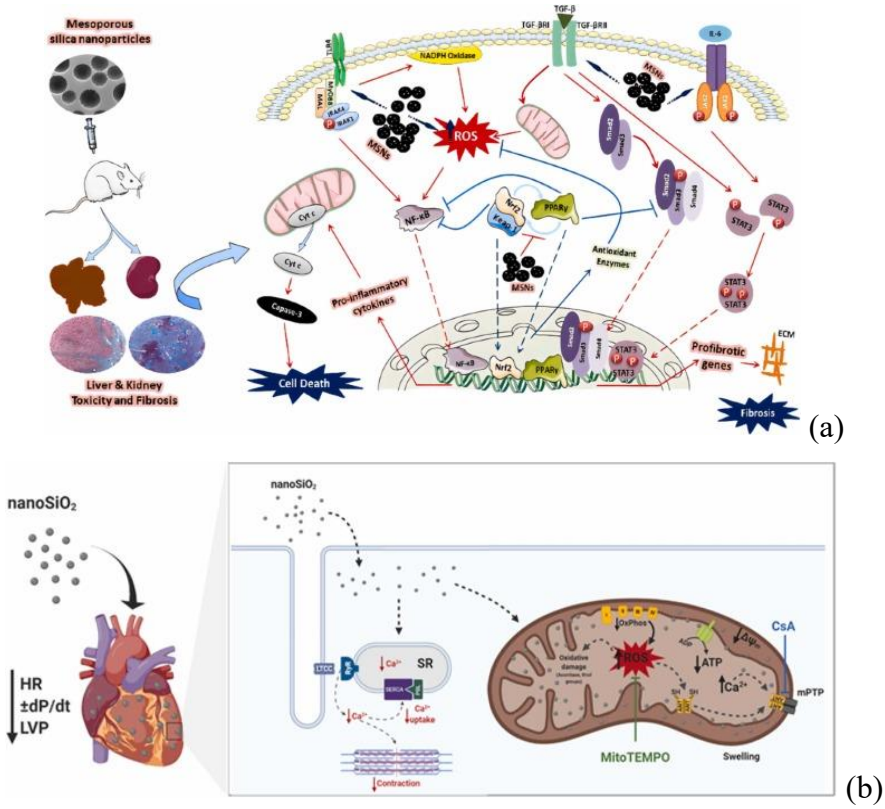


Figure 2: Pathways of SiO_2 NP-Induced Toxicity in Organ Systems and Mitochondrial Dysfunction

a. Organ System Toxicity [3]: Depicts the impact of mesoporous silica nanoparticles (SiNPs) on cellular pathways, highlighting the production of reactive oxygen species (ROS) and subsequent pro-inflammatory responses, cell death, and fibrosis in liver and kidney tissues.
b. Mitochondrial Dysfunction [4]: Shows nano silica's effects on cardiac function and mitochondrial integrity, detailing the biochemical cascade from SiNP exposure to altered calcium dynamics, increased oxidative stress, and mitochondrial swelling affecting cardiac health.

Upon entry into the respiratory system, SiNPs are known to induce a spectrum of adverse effects, including significant inflammatory responses and oxidative stress. For example, exposure to these NPs can trigger cytokine production in bronchial cells, potentially exacerbating pulmonary diseases [5]. Furthermore, SiNPs have been shown to cause oxidative stress and autophagy in lung fibroblasts, processes that can lead to cell death and tissue damage if unchecked [6]. Research also indicates that smaller SiNPs are particularly potent in generating reactive oxygen species (ROS), enhancing their potential for cellular damage [7]. This body of evidence underscores the critical need for targeted studies and regulatory measures to address and mitigate the health risks associated with NP exposure.

The illustrations provided in Figure 2 depict the cellular pathways affected by SiNPs, highlighting their impact on cell death and fibrosis due to ROS and their influence on heart function and mitochondrial dynamics, which further demonstrate the depth and complexity of SiNP interactions within biological systems

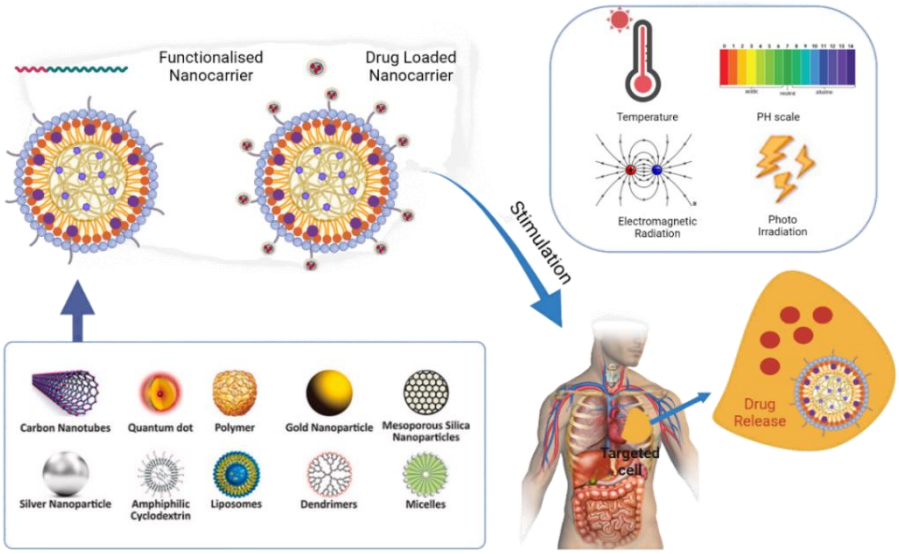


Figure 3: Targeted Drug Delivery Systems

Highlights the functionalization of nanocarriers that respond to specific stimuli such as pH changes and temperature, facilitating controlled drug release at targeted disease sites.

A focus of nanoparticle-protein corona (NP-PC) research within biomedical contexts is their interaction with mucins, such as MUC5AC, which are prevalent in the respiratory tract. Mucins

are high-molecular-weight glycoproteins that form a viscous barrier protecting epithelial cells from pathogens and particles. The interaction between NPs and mucins like MUC5AC is crucial for designing nano-carriers designed for targeted drug delivery in the respiratory system. Understanding these interactions at the molecular level, particularly their conformations and binding affinities, is vital for predicting and optimizing the therapeutic efficacy of NP-based treatments.

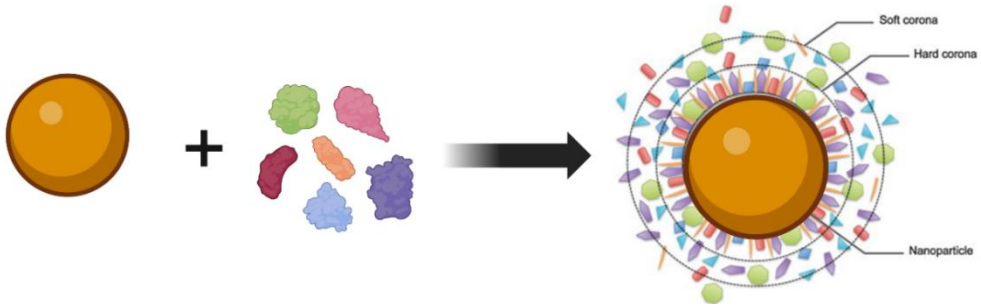


Figure 4: Formation of the Protein Corona

Shows the progressive formation of a protein corona around a NP, demonstrating how biomolecules from the surrounding environment adsorb to form structured layers.

Recent advances in molecular dynamics simulations and high-resolution structural biology have allowed for unprecedented insights into the complex structure and behaviour of mucins. For example, detailed studies on MUC5B [8] have elucidated the significant impact of glycosylation on mucin structure, revealing how these post-translational modifications contribute to the characteristic bottlebrush architecture of mucins.

As the field of NP research advances, the synergy between detailed molecular dynamics simulations and robust experimental data becomes central for the development of safe and effective nanotherapeutics. By investigating the complex interactions at the nano-bio interface, scientists can more adeptly engineer NPs that successfully navigate the intricate biological environment, thereby maximizing their therapeutic efficacy and minimizing potential adverse effects. This general approach is vital for fostering the growth and acceptance of nanotechnology across medical and other critical applications. Specifically, this study will

utilize a combination of molecular dynamics simulations and molecular docking studies to offer a detailed understanding of how NPs interact with mucin structures and functions. This investigation is particularly pertinent for intravenous nano-carriers, as mucins can significantly impact the stability and functionality of these carriers within the bloodstream and lungs, influencing their overall effectiveness in clinical settings.

1.1 Objectives

This project aims to explore the dynamics and interactions of MUC5AC mucin in various simulated environments using advanced molecular dynamics simulations and Markov state modelling. Key objectives include:

- **Mucin Dynamics**: Analyse MUC5AC behaviour in aqueous environments with ions to understand its conformational dynamics.
- **NP Interactions**: Investigate interactions between MUC5AC mucin and silica NPs, focusing on binding energies, structural adaptability, and implications for nanotoxicology.
- **Cytotoxicity Assessment**: Evaluate the cytotoxic effects of NPs on respiratory cells and mucin to discern patterns that might influence cellular responses.
- **Molecular Interactions and Environmental Effects**: Examine molecular-level interactions and environmental influences on MUC5AC's behaviour with NPs, emphasizing Markov state modelling to capture dynamics across multiple states.
- **Protein Corona and Immune Modulation**: Study the role of the protein corona in modulating the immune response to NPs, particularly how it affects mucin's recognition and clearance.

These goals aim to deepen our understanding of nano-bio interactions, providing valuable insights for drug delivery systems and safety assessments of nanomaterials.

2. Methods

2.1 Proteins

Proteins are intricate, essential biomolecules that perform a wide array of functions critical to biological systems. Comprising sequences of amino acids, proteins assume specific three-dimensional shapes that are important for their diverse roles. These functions include structural support as seen with collagen in connective tissues, catalytic activity through enzymes such as amylase in saliva, and transport duties demonstrated by haemoglobin carrying oxygen in the bloodstream. Proteins are also key players in the immune system, forming antibodies, and are integral to cellular signalling, with hormones like insulin regulating glucose metabolism. Furthermore, proteins such as actin and myosin are vital for muscular movement [9].

The structure of proteins is traditionally categorized into four levels: primary, secondary, tertiary and quaternary. The primary structure simply refers to the sequence of amino acids in a polypeptide chain, linked by peptide bonds. The secondary structure involves local arrangements of the chain, primarily alpha-helices and beta sheets, which are stabilized by hydrogen bonds between backbone atoms. The tertiary structure represents the overall three-dimensional configuration of the protein, which results from interactions among various side chains and segments of the polypeptide. The quaternary structure pertains to the assembly of multiple polypeptide chains into a functional protein complex [10]. This complex structure determines a protein's functionality and its interactions within the cell. These four levels of protein structure are illustrated schematically Figure 5:

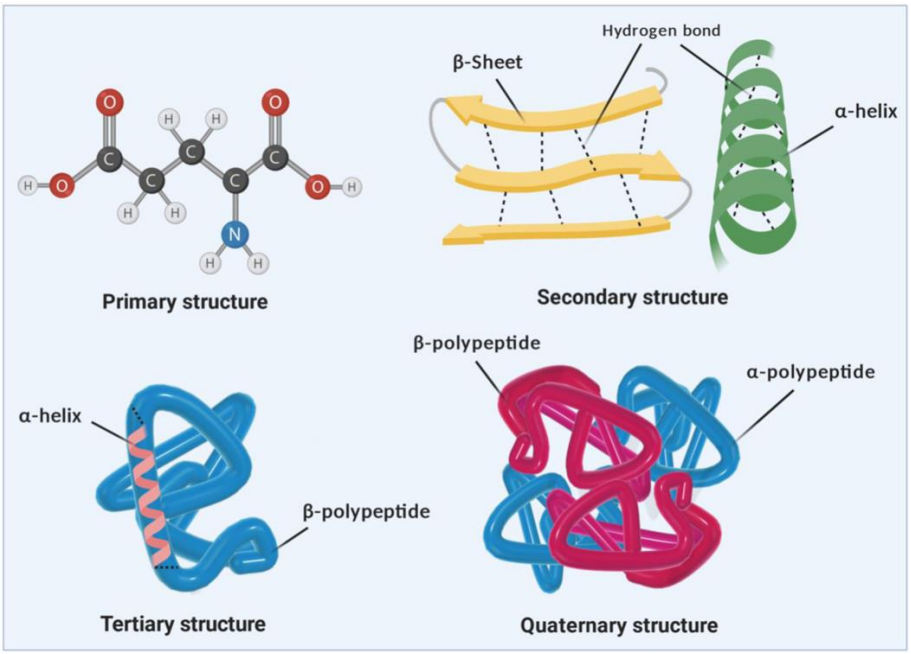


Figure 5: A schematic depiction of primary, secondary, tertiary, and quaternary protein structures, adapted from [11].

For the detailed analysis and visualization of protein structures, tools like VMD (Visual Molecular Dynamics) and PyMol are invaluable. These tools help researchers analyse the protein's mass and charge distribution, key for understanding interaction dynamics within cells. These molecular insights are important for various applications, including drug design and synthesis and therapeutic agents, giving a deeper understanding of disease mechanisms and potential treatments [12, 13]. This structural data can guide the development of interventions that mimic or modulate protein functions, potentially leading to innovate treatments for a range of diseases [14].

2.2 Mucins

Mucins are large, highly glycosylated proteins that form a critical component of mucus, which helps in protecting and lubricating epithelial surfaces, such as those in the respiratory, gastrointestinal and urogenital tracts. This mucus layer is a viscoelastic gel that is formed by the crosslinking and entanglement of mucin fibres, which exhibit hydrogel-like properties [15]. Understanding the structural and functional complexities of mucins is essential for exploring

their role in health and disease, as well as their interactions with nanomaterials for therapeutic applications.

The mucin family is divided into two main subgroups: secreted mucins and membrane-associated mucins, each playing distinct roles in the formation and function of mucus.

1. **Secreted Mucins**: These mucins are released into the extracellular space and contribute to the formation of mucus gels. They are primarily responsible for the viscosity and gel-like characteristics of mucus. Examples include MUC2, MUC5AC, MUC5B, and MUC6.
2. **Membrane-Associated Mucins**: These mucins are attached to the cell surface and play roles in cell signalling, adhesion and forming protective barriers Examples include MUC1, MUC4, and MUC16.

The polymerization of these mucins is facilitated by disulfide bonds, with further stabilization provided by non-glycosylated and non-covalent interactions. Hydrophobic interactions within cysteine-rich domains (CysD), and electrostatic repulsion between charged saccharide side chains also contribute to the stability and functionality of the mucin network [16, 17]. In the healthy respiratory tract, MUC5AC and MUC5B are the predominant secretory mucins. Under pathological conditions such as rhinitis, an increased expression of MUC5AC significantly alters the properties of the mucus gel [17].

MUC5AC is a prominent secreted mucin found in the airway, nasal cavity and stomach, playing a significant role in the respiratory system's mucus layer. It features a large central domain with both repetitive and non-repetitive sequences, interspersed with cysteine-rich domains that are important for the formation and stability of the mucin network. Recent advancements have made the structure of MUC5AC more accessible to researchers. The RCSB protein data bank now includes an x-ray crystallographic structure of MUC5AC (PDB code: 8ov0), which

provides valuable insights into its molecular configuration. This structure is particularly useful for understanding how MUC5AC contributes to the biophysical properties of mucus and its interactions with other molecules.

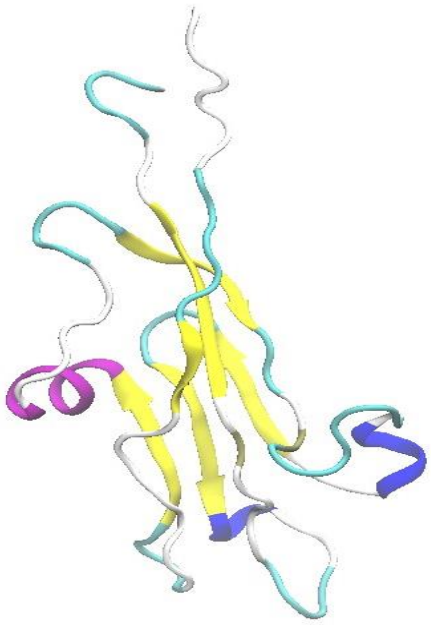


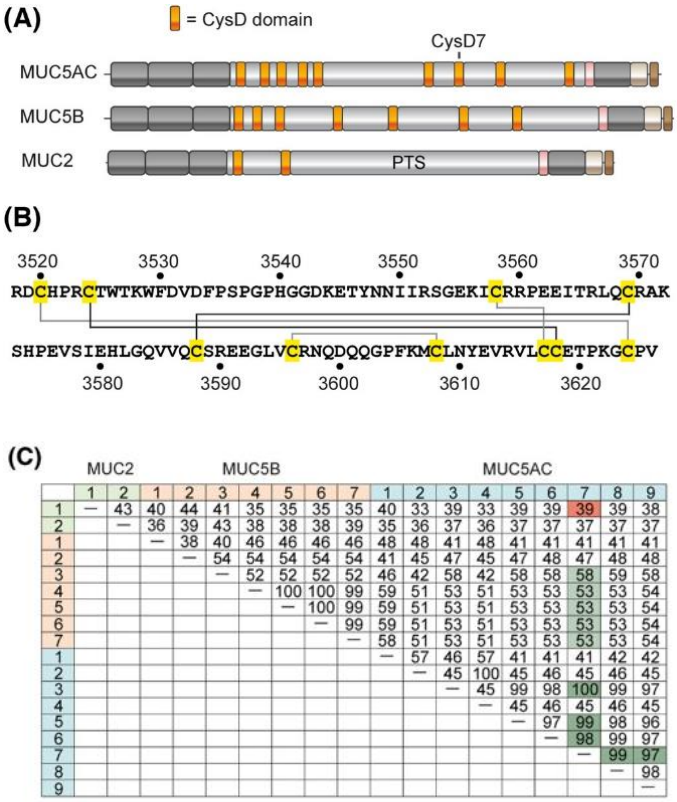
Figure 6: Pre-minimization of MUC5AC CysD7 domain.

Given the complex gel properties of mucus, it is vital to investigate how nanocarriers interact with the mucosal barrier to create efficient drug delivery systems [18]. The mucin network extensively captures nanoscale carriers, and these interactions are influenced by the characteristics of the NPs. Positively charged ligands such as chitosan (COS) bind to negatively charged mucus through electrostatic interactions, which make them ideal for mucoadhesive delivery systems [19]. In contrast, NPs coated with polyethylene glycol (PEG) offer a hydrophilic, muco-inert surface, improving particle movement and penetration through the mucus barrier [20]. Although both strategies are promising, the swift turnover of mucus in the respiratory tract means that NPs must quickly reach the underlying epithelium to prevent being cleared away [21, 22]. Understanding the detailed interaction dynamics between nanocarriers and mucins is vital for the rational development of more effective drug delivery systems.

MD simulations have been extensively employed to investigate the mechanisms and processes involved in drug delivery, providing significant insights into the interactions between carrier molecules and biomolecules [23-25]. These simulations replicate specific physiological environments, such as pH, ionic strength, and volume, with results that can be validated through biological experiments [24, 26]. This offers a microscopic view of the interactions between ligand molecules and proteins, enhancing our understanding of the nanocarrier delivery process. For mucin simulations, coarse-grained (CG) simulations are often utilized to model the complex and dynamic structure of mucin networks, studying the effects of particle shape [27], rigidity [28] and temperature [29] on mucosal drug delivery.

In this study, we focus on the interactions between MUC5AC and silica (SiO_2) NPs. SiNPs are widely used in biomedical applications due to their biocompatibility and ease of functionalization. MUC5AC, which is prevalent in the airway, nasal cavity and the stomach is encoded by one of the largest genes in the human genome and features a central domain containing both repetitive and non-repetitive sequences interspersed with nine cysteine-rich domains (CysD), key for mucin networks formation and stability.

As mentioned earlier, the newly available structure of MUC5AC on the RCSB [30] will be used for our MD simulations. This structure, which includes 111 amino acids and the CysD domain, is essential for accurately simulating interactions within the mucin gel network.



	1	2	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	9
1	—	43	40	44	41	35	35	35	35	40	33	39	33	39	39	39	39	38
2		—	36	39	43	38	38	38	39	35	36	37	36	37	37	37	37	37
1			—	38	40	46	46	46	46	48	48	41	48	41	41	41	41	41
2				—	54	54	54	54	54	41	45	47	45	47	48	47	48	48
3					—	52	52	52	52	46	42	58	42	58	58	58	59	58
4						—	100	100	99	59	51	53	51	53	53	53	53	54
5							—	100	99	59	51	53	51	53	53	53	53	54
6								—	99	59	51	53	51	53	53	53	53	54
7									—	58	51	53	51	53	53	53	53	54
1										—	57	46	57	41	41	41	42	42
2											—	45	100	45	46	45	46	45
3												—	45	99	98	100	99	97
4													—	45	46	45	46	45
5														—	97	99	98	96
6															—	98	99	97
7																—	99	97
8																	—	98
9																		—

Figure 7: Cysteine-Rich domains in Human Gel Forming Mucins, adapted from Ref [30].

Figure 7 provides a detailed view of the structure of human-gel forming mucins, particularly MUC5AC, MUC5B, and MUC2. MUC5AC, which consists of 5654 amino acids (UniProt P98088), features several CysD domains. These orange-highlighted domains support the formation of disulfide bones that stabilize the mucin network essential for mucus's gel-forming properties. This figure also includes various functional domains such as Von Willebrand Factor Type C domain sets which contribute to cellular signalling and barrier functions, and C-terminal Cystine Knot-like domains that facilitate protein-protein interactions within the mucus. Disulfide bonds within MUC5AC CysD7 illustrate the structural stability critical under physiological conditions.

An accompanying table displays the amino acid sequence identities among these domains, highlighting significant variances that impact mucin's structural and functional diversity. High

sequence identities suggest strong conservation, which is essential for maintaining mucin's protective functions in different physiological environments.

Understanding these structures is beneficial for designing NPs for drug delivery, targeting specific mucin interactions to enhance penetration or adhesion to the mucus barrier. Investigating how SiNPs impact the structural integrity of the mucin network can provide insights into their potential as drug carriers and how they might be optimized for better performance.

2.3 Modelling Protein-Nanoparticle Interactions using Molecular Docking

To investigate the interactions between SINPs and the mucin MUC5AC, molecular docking simulations were performed using Patchdock [31], which was implemented on a High-Performance Computing (HPC) system. This analysis focused on understanding the geometric complementarity between NPs and the protein, essential for studying potential non-covalent binding configurations.

PatchDock is ideal for this study as it segments the molecular surfaces of both the NPs and MUC5AC into smaller, manageable patches. These patches are analysed for their shape and potential docking sites, facilitating the exploration of optimal binding orientations. The docking process, therefore, highlights possible molecular conformations and interactions that will help in understanding how SINPs can effectively bind to the protein's structure.

The algorithm's scoring function is rooted in shape complementarity, supplemented by evaluations of Lennard-Jones van der Waals interactions and electrostatic forces [32-34]. This approach allows PatchDock to assess the fit between the NP and protein surfaces robustly, ensuring that the most stable and feasible interactions are identified. Additionally, PatchDock provides insights into the Atomic Charge Energies (ACE), which gauge the solvation-free

energy changes involved in forming protein-protein and protein-solvent contacts. This metric helps determine the stability and feasibility of the docked complexes.

The docking process in PatchDock unfolds in three steps:

1) **Shape Representation:** Initially, the surfaces of the molecules are computed to identify distinct geometric patches. These patches are filtered to retain those with high interaction potential, referred to as “hotspots”.
2) **Patch Matching:** In this phase, the algorithm aligns the identified patches between the protein and the NP, using geometric and local similarity measures. This step employs pose-clustering algorithms to refine the alignment and produce a set of potential complexes, which are ranked on their geometric fit and Root Mean Square Deviation (RMSD) values.
3) **Complex Evaluation:** The final selection of complexes undergoes a rigorous check for steric clashes and unnatural overlaps, discarding any configurations with significant intermolecular penetrations. The remaining viable complexes are scored and ranked, focusing on their molecular surface interactions to find the most promising binding models.

Conducted on a HPC framework, this docking analysis is computationally intensive but important for accurately modelling the interactions across a range of NP sizes. This approach not only enhances the reliability of the docking predictions but also provides a comprehensive view of the potential interactions that could inform further experiments and guide the development of novel therapeutic agents.

2.4 Atomistic Molecular Modelling

Molecular dynamics (MD) simulations are a powerful computational technique used to predict the evolution of atomic and molecular systems over time. By numerically integrating Newton’

equations of motion, MD simulations generate detailed atomic trajectories based on specified interatomic potentials, initial conditions, and boundary conditions [35]. This method provides unique insights into the structure, energetics, and dynamics of biomolecules, allowing us to investigate across broad spatial and temporal scales – from atomic motions on quantum energy surfaces to protein binding in macromolecular assembles.

2.4.1. Basic Principles

At the core of MD simulations is Newton's second law of motion:

$$F_i = m_i a_i = m_i \ddot{r}_i \quad (1)$$

Where F is the force acting on an atom i, m is the mass of the atom, and a is the acceleration of the atom [36]. The force is related to the potential energy function $U(\vec{x})$ by:

$$\vec{\boldsymbol{F}}(\vec{\boldsymbol{x}}) = -\nabla U(\vec{\boldsymbol{x}}) \quad (2)$$

Where $\vec{\boldsymbol{x}}$ represents the positions of all atoms in the system. The equations of motion for an atom can be written as:

$$\frac{d\vec{\boldsymbol{v}}}{dt} = \vec{\boldsymbol{a}}(\vec{\boldsymbol{x}}) \quad (3)$$

$$\frac{d\vec{\boldsymbol{x}}}{dt} = \vec{\boldsymbol{v}} \quad (4)$$

Here, $\vec{\boldsymbol{v}}$ is the velocity of the atom, and $a(\vec{\boldsymbol{x}})$ is its acceleration as a function of position.

To accurately describe atomic behaviour, it is essential to solve the time-dependent Schrodinger equation. However, the computational demands of solving the equations of motion for the molecular wavefunction of systems with more than just a few electrons and nuclei become prohibitively expensive [37]. This complexity arises because the number of interactions and quantum mechanical variables increase exponentially with the addition of each particle, making detailed quantum mechanical calculations impractical for larger molecules or extensive

systems. Consequently, researchers often resort to approximations or simplified models to simulate the behaviour of such complex systems efficiently.

2.4.2. Force Fields

The potential energy $U(\vec{x})$ in MD simulations is typically expressed using a classical force field. For classical MD, electrons are not treated as particles, but their influence is incorporated into bond interactions. Force field parameters are derived from spectroscopic measurements and quantum mechanical calculations. The most commonly used force fields are Amber [38], CHARMM [39], Gromos [40], and OPLS [41]. This force field includes five terms representing bond stretching, angle bending, torsional rotation, and non-bonded interactions such as electrostatics and van der Waals forces, Lennard-Jones potential. The principal terms in the general form of the potential energy function for simple organic molecules and biological macromolecules are:

$$\boldsymbol{U(\vec{x})} = \underbrace{\sum_{bonds} k_i^{bond}(r_i - r_0)^2}_{U_{bond}} + \underbrace{\sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2}_{U_{angle}} +$$

$$\underbrace{\sum_{dihedrals} k_i^{dihe}[1 + \cos(n_i\phi_i + \delta_i)]}_{U_{dihedral}} + \underbrace{\sum_i \sum_{j \neq i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}}_{U_{nonbond}} \quad (5)$$

Where the following term in the equation signifies:

- U_{bond}: The first sum represents the potential energy from bond stretching, where k_i^{bond} are the force constants for bonds, r_i are the current bond length, r_0 are the equilibrium bond length. This term accounts for the energy cost associated with deviations from ideal bond lengths.
- U_{angle}: Accounts for potential energy from angle bending, with k_i^{angle} as force constants for angles, θ_i current bond angles, and θ_0 the equilibrium angels. This reflects the energy required to alter the angle between tow bonds connected by a common atom.

- U_{dihedral}: Addresses the energy from torsional rotations around bond angles, or dihedrals. $k_i^{dihedral}$ are the force constants, n_i are the multiplicities of the dihedral angles, ϕ_i the actual dihedral angels, and δ_i are phase shifts. This term describes how the energy varies with the rotation around the axis formed by the middle bond in a sequence of three consecutive bonds.
- $U_{\text{non-bonded}}$: The last two sums cover non-bonded interactions. The Lennard-Jones potential, given by the terms involving ϵ_{ij} (depth of the potential well) and σ_{ij} (distance at which the potential is zero), models the van der Waals forces, describing both repulsion at short ranges (proportional to r^{-12}) and attraction at longer ranges (proportional to r^{-6}). The electrostatic term, involving q_i and q_j (partial charges on atoms i and j), and ϵ (dielectric permittivity), calculates the Coulomb forces between charged particles, which are inversely proportional to the distance between them.

Each of these components model different aspects of molecular interactions, together giving a comprehensive picture of the forces that influence the dynamics and stability of molecular systems for simulations.

In MD, the behaviour of atoms is dictated by Newtonian mechanics and the defining feature of these simulations, as mentioned, is the force field, which outlines the potential energy landscape of the system. This force field captures key information about atomic interactions through equation (5). By mathematically modelling these interactions, the force field not only simplifies complex molecular behaviours into quantifiable terms but also enhances our ability to predict how these interactions influence the overall dynamics of the molecular system, as detailed in MD simulation tools such as NAMD [42].

2.4.3. Simulation Workflow

The workflow of a MD simulation involves several steps that ensure accurate modelling of biomolecular systems. Here is a detailed explanation:

System Preparation:

The initial step involves selecting and obtaining the structure of interest, such as the mucin protein, from the Protein Data Bank (PDB) or constructing it using molecular modelling tools.

Parameterization:

The selected structure is parameterized using appropriate force fields, like CHARMM36m, which accurately represent proteins, lipids, and carbohydrates.

Simulation setup:

There are several sub-steps designed to prepare the system for the main simulation run:

1. **Minimization**: Energy minimization is performed to alleviate any steric clashes or unrealistic geometries in the initial structure. NAMD uses conjugate gradient to find a local minimum in the potential energy landscape. This ensures that the starting configuration is physically plausible and stable.
2. **Equilibration**: The system is gradually heated to the desired temperature and equilibrated. During this phase, the solvent and ions surrounding the biomolecule are allowed to adjust, ensuing that the system is at equilibrium before the main simulation begins. Typically, an NVT (constant Number of particles, Volume, and Temperature) ensemble is used to maintain constant temperature and volume at this phase.
3. **Constraints**: Throughout the simulation, constraints may be applied or removed as necessary. These can be used to fix certain atoms or groups of atoms in space, allowing the rest of the system to equilibrate or simulate more accurately. This ensures that specific parts of the molecule remain stable while other parts equilibrate, which can help maintain the structural integrity of complex biomolecules.

4. **Production Run:** After equilibration, the main simulation, known as the production run, is conducted. This phase uses an NPT (allows system to evolve at constant Number of Particles, Pressure, and Temperature) ensemble. This setup mimics the natural and physiological conditions more accurately.

Analysis:

Following the simulation, extensive analysis is performed on the resulting data:

- **Trajectory Analysis:** The trajectories generated during the simulation are examined to extract structural and dynamic properties of the biomolecule. Key properties such as RMSD, radius of gyration, end-to-end distance and solvation effects are investigated. These analyses provide information into the conformational stability and flexibility of the mucin.
- **Binding Interactions:** For a system involving interactions between mucins and NPs, molecular docking and further MD simulations are used to investigate binding sites and interactions dynamics. This helps in understanding how NPs interact with and influence the behaviour of MUC5AC.

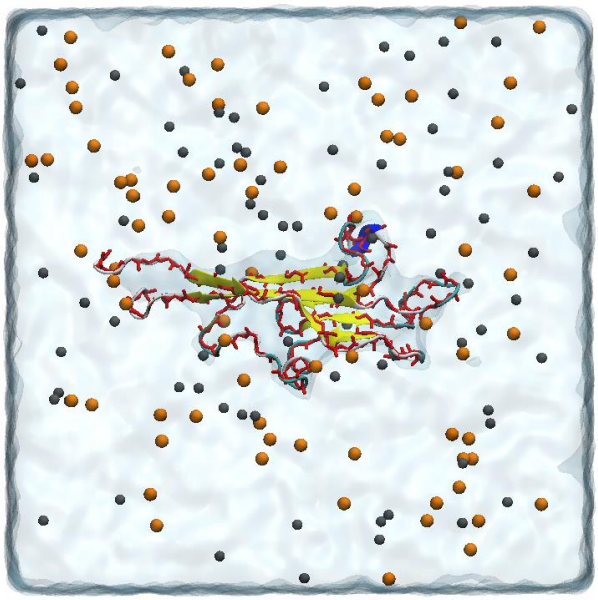


Figure 8: Snapshot of MUC5AC solvated in water and neutralized.

Protein backbones are coloured in dark red and Na+ and Cl-ions are shown in orange and grey spheres.

2.4.4. Thermodynamic and Kinetic Properties

MD simulations provide detailed descriptions of atomic motions and can be used to compute various thermodynamic and kinetic properties. These properties include free energy landscapes, diffusion coefficients, and rates of conformational changes. By simulating the system under different conditions and analysing the trajectories, we can get insights into the molecular mechanisms underlying biomolecular functions and interactions.

2.4.5. Coulomb Interaction

In molecular simulations, the electrostatic contributions are modelled by assigning specific partial charges to atoms, which approximates their electronegativities. These charges are utilized in calculating the electrostatic potential energy between atoms using Coulomb's law, defined by the equation:

$$U_{coulomb} = \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \qquad (6)$$

Here, q_i and q_j, represent the partial charges on atoms i and j respectively, ϵ_0 is the permittivity of free space, and r_{ij} is the distance between the atoms. The notation $\sum_{i<j}$ ensures that the summation is carried over each unique pair of atoms without any repetitions, thereby avoiding double counting.

Given the vast number of pairwise interactions in typical molecular systems, direct computation of long-range electrostatics can be computationally expensive. To manage this, the Particle Mesh Ewald (PME) method is often employed [43]. This technique efficiently computes electrostatic forces by dividing them into short-range interactions, which are computed directly as per above equation, and long-range interactions, which are handled in Fourier space.

The PME method significantly reduces computational demands, yielding a time complexity of $O(N \log N)$ [43], where N is the number of particles in the system. This balance allows for

more accurate and efficient simulations of molecular interactions, especially in systems with many particles.

2.4.6. Lennard-Jones Potential

The Lennard-Jones potential [44] is a mathematical model used to describe the interactions between particles through van der Waals forces. These forces encompass Keesom interactions (permanent dipoles), Debye forces (induced dipoles), and London dispersion forces (instantaneous dipoles). The Lennard-Jones potential is crucial for simulating these interactions and is characterized by two primary components:

1. **Long-range attractive force:** This is modelled by a term that decreases as the inverse twelfth power of distance between particles, r_{ij}^{-12}. This term reflects the attraction between fluctuating dipoles as their distance increases.
2. **Short-range repulsive force:** This is described by a term that decreases as the inverse sixth power of the distance, r_{ij}^{-6}. This force arises due to the Pauli exclusion principle, which states that no two fermions (such as electrons) can occupy the same quantum state simultaneously within a quantum system. This principle effectively generates a "pressure" that keeps the particles apart when they come too close to each other.

The combined Lennard-Jones potential is given by the equation:

$$U_{LJ} = \sum_{i<j} 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \qquad (7)$$

Here ϵ represents the depth of the potential well, indicating the strength of the interactions and σ is the finite distance at which the inter-particle potential is zero and typically represents the size of the particles.

The Lennard-Jones potential is applied to pairs of atoms that are either part of different molecules or are separated by at least two covalent bonds. The modelling of these interactions

helps simulate realistic physical behaviours of molecular systems, providing details into the structural and dynamical properties of materials and biological molecules. The attractive component of the potential contributes to the stabilization of molecular structures, while the repulsive part prevents the collapse of molecules by maintaining structural integrity under various conditions.

2.4.7. Cutoff and Neighbour List

The evaluation of non-bonded potentials is a critical and computationally demanding aspect of MD simulations. In these simulations, the interaction energies are calculated pairwise for all atoms that are separated by at least two covalent bonds. Given a system with N non-bonded atoms, the number of pairwise interactions that must be considered is $\frac{N(N-1)}{2}$. This computation is required regardless of the system's volume, making it on of the most intensive components of MD simulations.

To manage and reduce the computational load, two primary modifications are often implemented:

1. Cutoff Distance: This is used to limit the range over which non-bonded interactions are calculated. Interactions beyond this cutoff are considered negligible and are thus ignored. This approach is based on the observation that the potential energy contribution from distance atom pairs is minimal due to the rapid decay of the potential function with distance.

2. Smoothing Functions: Near the cutoff distance, a smoothening function is introduced to the Lennard-Jones potential. This function smoothly transitions the potential to zero at the cutoff distance, ensuring that there are no discontinuities in the force field that might lead to non-physical artifacts in the simulation. This modification does not significantly affect the system's behaviour because the potential naturally

converges to zero at long distances, and significant energy contributions are preserved, as seen in Figure 9.

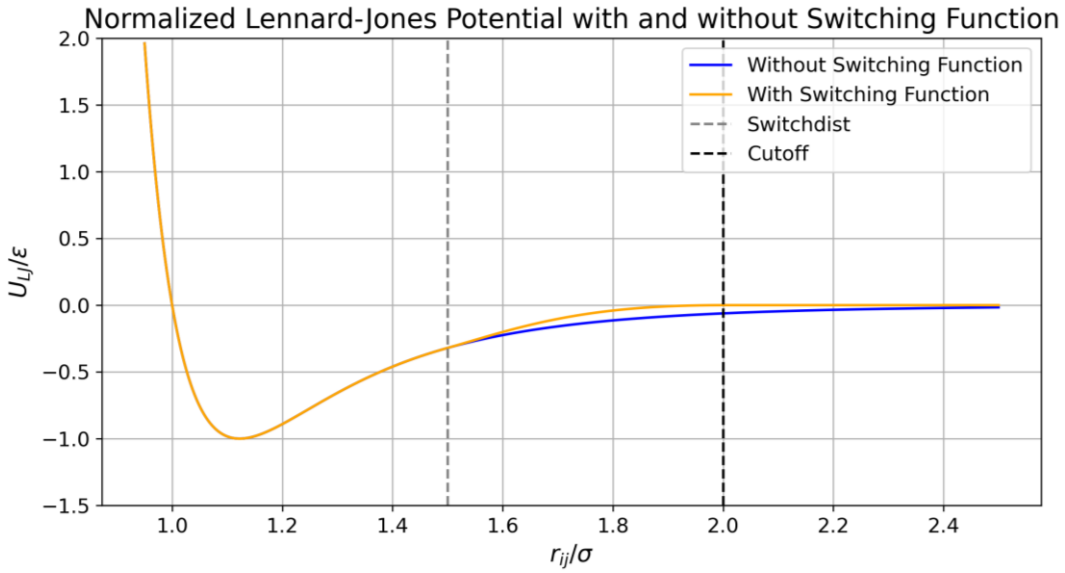


Figure 9: Normalized Lennard-Jones potential as a function of interparticle distance.

Represented in two scenarios: with and without a switching function. The blue curve shows the traditional potential, while the orange curve depicts the potential with a switching function that smoothly reduces to zero at the 'Cutoff' distance, beginning from the 'Switchdist'. This modification helps maintain computational efficiency and simulation accuracy.

In addition to these modifications, the use of neighbour lists [45] significantly optimizes the computation. These lists, often referred to as Verlet lists, contain atoms within the cutoff distance plus a buffer distance and are used to catalogue the interactions:

- **Verlet List Management**: Each atom's neighbour list includes those within the cutoff distance r_c plus a buffer distance Δr. This list tracks which pairwise interactions need to be computed and is vital between atoms that do not significantly relocate over a single time step. Thus, a pairwise potential between them can be assumed relatively constant.
- **List Update Strategy**: Every few iterations, typically every 20, the list is reviewed. Atoms that have moved beyond $r_c + \frac{\Delta r}{2}$ are removed from the list, and the list is refreshed to include new atoms that have come within range. This periodic updating

ensures that the list remains accurate without the need of recalculating interactions at every simulation step, striking a balance between computational efficiency and accuracy.

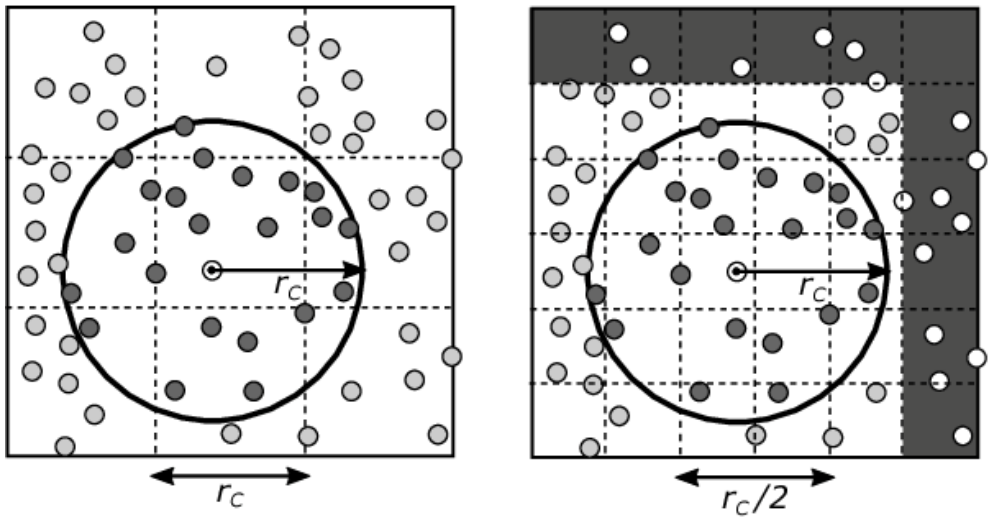


Figure 10: Two-Dimensional Verlet Neighbour List.

This diagram adapted from Ref [46] demonstrates a two-dimensional Verlet neighbour list system, showing atoms within the cutoff radius r_c, and a buffer distance $\Delta r/2$. The left panel displays atoms within r_c of the centre atom, marked for inclusion in the Verlet list. The right panel demonstrates the reduced search area to $r_c/2$, highlighting the area where the potential updates occur, typically every 20 steps or when an atom moves beyond $r_c + \Delta r/2$.

For electrostatic interactions, the cutoff distance is strategically set to balance computational load and simulation fidelity. This designated distance marks the point at which the potential expression for interactions shifts from being computed in short-range directly between nearby particle pairs to being addressed through long-range effects in Fourier space. For short-range interactions, direct computations are feasible as they involve only nearby pairs, reducing the overall computational effort.

For interactions that extend beyond this cutoff distance, the simulation adopts Fourier space methods to handle the long-range potentials. These are efficiently computed using the fast Fourier transform (FFT), a numerical technique [43] that transforms spatial data into frequency components. By employing FFT, long-range electrostatic forces are calculated with greater

speed due to the method's ability to efficiently manage convolution operations that are more computationally intensive in real space. This combination of direct short-range calculation and Fourier-based long-range computation improves the efficiency and scalability of MD simulations, enabling accurate modelling of complex systems with many interacting particles.

2.4.8. Periodic Boundary Conditions for MD

In MD simulations, the concept of a simulation cell is central to modelling the behaviour of biomolecules and solvents in a defined three-dimensional space. Typically represented as a polyhedral volume, often a cube, this cell contains the biomolecules and solvents under study. When simulations are conducted in a vacuum, the system exhibits non-periodic characteristics, leading to phenomena such as solvent evaporation and the formation of water-vacuum interfaces which are not representative of bulk conditions.

To mitigate these effects and emulate a more realistic infinite system, periodic boundary conditions (PBCs) are employed. This technique involves replicating the simulation cell in all directions, effectively surrounding the original cell with its own images. This setup creates a seemingly infinite tessellation, where molecules and atoms that exit one side of the primary simulation cell re-enter through the opposite side, maintaining their velocities and trajectories. This continuity ensures that the properties and dynamics of the system do not suffer from edge effects, and the simulation more closely represents the behaviour of a bulk material. For example, Figure 10 shows a 2D representation of how molecules are placed inside a cubic box. In this representation each atom retains its positional coordinates and velocities, ensuring consistent physical behaviour across all cells.

PBCs are particularly vital in simulations explicitly modelling solvation with solvent molecules, as they permit a relatively small number of solvent molecules to simulate a bulk environment effectively. This is achieved by treating interactions across the cell boundaries,

ensuring that molecules exiting one side of the primary cell are reintroduced on the opposite side as though surrounded by a continuous expanse of similar cells. The mathematical implementation of PBCS involves straightforward adjustments to the coordinates of particles crossing the boundaries, thus maintaining constant particle number and system density.

While often cubic for simplicity, the shape of the simulation cell can vary based on specific simulation needs. Alternative geometries like the truncated octahedron or rhombic dodecahedron may be used to reduce the number of particles required and better approximate spherical volumes around key molecules, optimizing computational resources.

However, the adoption of PBCs is not without challenges. The main issue arises from the artificial periodicity introduced, particularly significant in the presence of long-interactions that can span the cell dimensions, potentially causing a molecule to interact with its own images [47]. This necessitates careful consideration of cell size relative to the interaction range, ensuring that simulation outcomes remain physically valid and free from artifacts induced by the periodic setup.

2.4.9. Initial Conditions for MD

In MD simulations, the initial three-dimensional atomic positions of a molecule are typically specified using a Protein Data Bank (PDB) file. These structure are derived experimentally through methods such as x-ray crystallography [48] , nuclear magnetic resonance (NMR) [49], and cryogenic electron microscopy (cryo-EM) [50]. To initiate the simulation, initial velocities for the atoms are assigned randomly based on the Maxwell-Boltzmann distribution. This statistical distribution describes the expected speed distribution of particles in a gas in thermal equilibrium at a given temperature and is expressed mathematically as:

$$P(v_i) = 4\pi v_i^2 \left(\frac{m_i}{2\pi k_B T}\right)^{\frac{3}{2}} e^{-\frac{m v_i^2}{2k_B T}} \qquad (8)$$

Here, v_i represents the magnitude of the velocity vector of the i^{th} atom, calculated as $v_i = \sqrt{v_{x,i}^2 + v_{y,i}^2 + v_{z,i}^2}$, where T is the temperature, $k_B$is Boltzmann's constant, and m_i is the mass of the atom. This approach ensures that the simulated system begins in a state representative of thermal equilibrium, setting the stage for the dynamic evolution of the system under study.

2.4.10. Numerical Integration

In systems containing more than two atoms, analytical solutions to describe the interactions and movements become infeasible due to their complexity. Consequently, numerical methods are essential to solve the dynamics of such systems. In MD simulations the N second order differential equation Eq. (1), which describe the motion of atoms, are converted into 2N first-order differential equations. These equations are represented as:

$$\frac{d\vec{v_i}}{dt} = \vec{a_i} = \frac{\vec{F_i}}{m_i} \tag{9}$$

$$\frac{d\vec{r_i}}{dt} = \vec{v_i} \tag{10}$$

Where $d\vec{v_i}/dt$ and $d\vec{r_i}/dt$ denote the rate of change of velocity and position, respectively with $\vec{F_i}$ representing the force applied to the atom, and m_i, its mass. The numerical integration in MD involves discretizing these equations over small time intervals Δt, crucial for the accurate depiction of particle trajectories.

The discretization must preserve the thermodynamic properties of the molecular system while balancing computational efficiency and numerical stability. The stochastic nature of MD makes the convergence of these numerical methods particularly challenging. Due to the chaotic behaviour of macromolecular systems, even minor variations in initial conditions can lead to

significantly divergent trajectories, a phenomenon exacerbated by the propagation of rounding errors in calculations. This necessitates a high degree of accuracy in the simulation of thermodynamic and dynamic properties and adequate sampling of phase space to ensure reliable results. [51]

2.4.11. Integration timestep

The choice of timestep Δt is an essential parameter in numerical integration, influencing computational speed, accuracy and stability. Smaller timesteps can reduce truncation errors, thus enhancing the precision of the simulation but at the cost of increased computational load. Conversely, larger timesteps enhance the sampling of conformational space, potentially increasing the speed of simulations but may lead to numerical instability and less accurate results. The timestep must be sufficiently small to accurately sample the fastest motions in the system, such as vibrations in hydrogen bonds, which typically have a period of around 13 femtoseconds [52]. Recommended timesteps are about 1 fs for simulations involving non-rigid bonds and 2 fs for those with rigid bonds. [53]

2.4.12. Verlet Algorithm

The Verlet algorithm is a widely used method in MD for solving Newton's equations of motion. Employed by software such as NAMD, the Verlet method is favoured for its simplicity and efficiency. It exhibits low global errors, which are proportional to Δt^2, and is both time-reversible and symplectic, meaning it conserves the phases space volume [54]. The Verlet algorithm can be derived using a Taylor series expansion of Newton's second law to the second derivative, providing a framework that effectively handles the computation of particle trajectories while omitting direct velocity calculations, as demonstrated in the following equations:

$$r(t \pm \Delta t) = r(t) \pm \dot{r}(t)\Delta t + \frac{\ddot{r}(t)\Delta t^2}{2} \qquad (11)$$

By summing the positions $r(t + \Delta t)$ and $r(t - \Delta t)$, and substituting the acceleration from Eq. (2), the dependence on velocity can be removed, simplifying the calculations and enhancing the algorithm's stability and accuracy.

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) - \frac{\nabla U(r)}{m}\Delta t^2 \tag{12}$$

2.5 Modelling Steps to Build Nanoparticle

For this project, CHARMM-GUI Nanomaterial Modeler [55] was used to construct silica NPs of 4nm and 11 nm α-quartz structures, see Figure 11, and to combine it with the mucin system with CHARMM-GUI multicomponent assembler [56]. The system is constructed by first defining the NP size and material properties in CHARMM-GUI. The system is then parameterized with appropriate force fields, solvated, and ions are added to simulate physiological conditions. Following energy minimization to stabilize the initial configurations, MD simulations are conducted to observe hydrophobic and electrostatic interactions between the NP and protein.

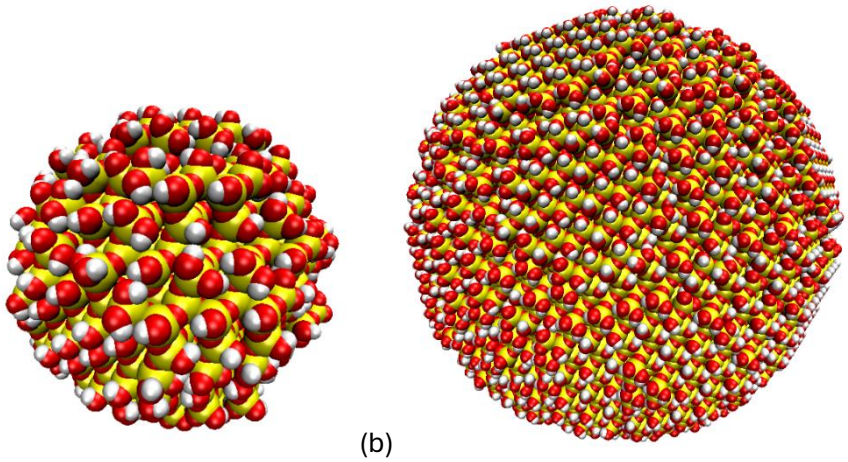


Figure 11: Nanoparticle (NP) of Silicon Dioxide (SiO_2) (a) 4nm and (b) 11nm in diameter.

2.6 Global Collective Variables

In MD simulations, global collective variables are quantities derived from the positions and velocities of a system's constituent particles that provide insight into its overall behaviour and

state. These variables are important for understanding complex dynamics, guiding the simulation towards specific states, and analysing the behaviour of large biomolecular systems. This section will cover and breakdown the variables which will be investigated in this project.

2.6.1. Root Mean Squared Deviation (RMSD)

The Root Mean Squared Deviation (RMSD) is a widely used measure in MD to quantify the structural similarity between two conformations of a molecule, typically between a reference structure and a given configuration during simulation. It provides a numerical value that reflects the average distance between the atoms of proteins, indicating the degree of conformational change over time.

RMSD is expressed as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left\| \vec{r_i} - \vec{r_i'} \right\|^2} \qquad (13)$$

Where N is the number of atoms considered, r_i represents the position vectors of atoms in the reference structure, and r_i' corresponds to the position vectors of the same atoms in the structure being compared. This calculation helps assess the consistency of protein conformations through various stages of the simulation, giving information in the dynamic behaviour and stability of the molecule.

The computation of RMSD is typically performed using Visual Molecular Dynamics (VMD), which aligns the structures to minimize the RMSD value, therefore offering a clear visual and understanding of molecular motions and transformations.

2.6.2. Radius of Gyration (R_g)

The radius of gyration is a metric used to quantify the size of a molecule by measuring the distribution of its atoms around its centre of mass. Mathematically, it is expressed as the root

mean square distance of the atoms from their collective centre of gravity, represented by the following equation:

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \overrightarrow{r_i^2}} \qquad (14)$$

This calculation provides information into the overall spatial configuration of the molecule, reflecting how compact or extended its structure is. The estimation of the R_g was carried out using scripts provided by VMD [12].

2.6.3. End-to-End distance

The End-to-End distance measures the linear separation between two designated points, typically the terminal atoms of a molecule. It is a key metric in molecular dynamics (MD) simulations, offering insights into the conformational size and spatial configuration of the molecule in its solvated state. The distance is calculated as:

$$End - to - End\ Distance = \|\overrightarrow{r_N} - \overrightarrow{r_1}\| \qquad (15)$$

Where $\overrightarrow{r_N}$ and $\overrightarrow{r_1}$ are the position vectors of the terminal atoms. This measurement is crucial for understanding how a molecule stretches or compacts under various simulation conditions. Tracking the End-to-End distance provides valuable information on the molecule's dynamic behaviour, aiding in the correlation of structural changes with biological functions, environmental effects, and interactions with potential drug molecules.

2.6.4. Hydrogen bond Analysis

Hydrogen bond analysis is essential for understanding the molecular interactions in protein-ligand complexes and other biomolecular systems within MD simulations. Despite being classified as weak, hydrogen bonds help the structural integrity and functional capabilities of biological molecules.

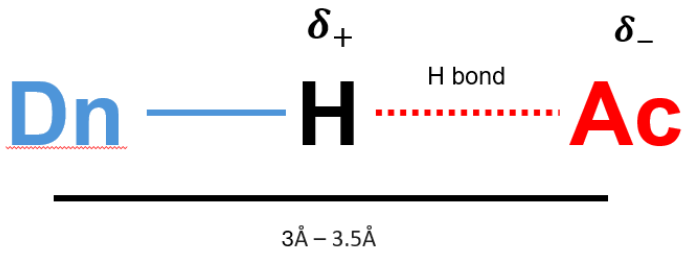


Figure 12: Hydrogen Bonding between Donor and Acceptor Atoms

A hydrogen bond is formed between a donor (Dn) and an acceptor (Ac). The donor is typically an electronegative atom such as fluorine, oxygen, or nitrogen, covalently bonded to a hydrogen atom, which develops a partial positive charge (δ+). The acceptor, also an electronegative atom, does not share a covalent bond with the hydrogen but attracts the hydrogen's partial positive charge due to its partial negative charge (δ-). This interaction is usually observed within about 3 Å to 3.5 Å as illustrated in Figure 12.

The 'Hbonds' plugin in VMD is employed to analyse these hydrogen bonds [12]. This is done by selecting the molecular system and then configuring parameters that define what constitutes a hydrogen bond, focusing on distance and angle thresholds. Following configuration, the plugin detects and records these bonds throughout the simulation.

2.6.5. Solvent Accessible Surface Area (SASA) as a Hydrophobic Descriptor

Hydrophobicity describes how substances repel water, influencing molecular interactions and stability in biological systems. This property is valuable in understanding how molecules like proteins interact with their surrounding environment, particularly in terms of their accessibility to solvent molecules.

The Solvent Accessible Surface Area (SASA) is a widely used measure to quantify the contact area between a molecule and the surrounding solvent rather than itself [57]. SASA is determined by rolling a rigid variable radii sphere over a molecular van der Waals surface. According to the Lee and Richards technique [58], refines this process by segmenting the surface into equidistant points and manoeuvring a spherical probe with a radius of 1.4 Å over

the exposed regions, thereby representing the solvent. SASA is expressed in standard units of $Å^2$ [59, 60]. Although there is no direct correlation between SASA and free energy changes, it offers insights into the conformational shifts and comparative analysis of molecular structures. The SASA metric helps to understand how solvent exposure varies with molecular conformation and interaction dynamics.

Other similar methods include the Shake and Rupley algorithm [61], and the power diagram method [62]. The precision of SASA measurements can be enhanced by adjusting the resolution of the surface analysis. This is particularly useful in scenarios requiring detailed surface gradients or when employing approximations for specific scientific or industrial applications [63, 64]. By refining the resolution, accurate representations of the solvent-accessible surface can be achieved. Figure 13 depicts how SASA is calculated with the spherical probe.

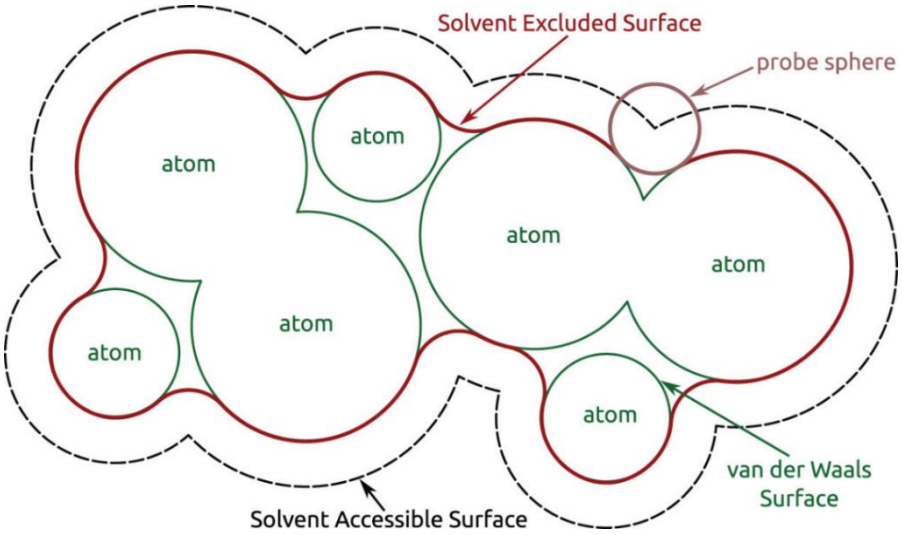


Figure 13: Visualization of Solvent Accessible Surface Area (SASA) and Solvent Excluded Surface Area (SESA) adapted from Ref [65].

This diagram depicts a spherical probe, representing a solvent molecule, rolling over the van der Waals surface of atoms to determine regions accessible and excluded to the solvent. This method is crucial for understanding molecular interactions at the atomic level.

One of the prominent tools for calculating SASA is FreeSASA [66]. It employs a spherical probe, akin to a solvent molecule, which traverses over the molecule's surface. The SASA is then defined by the pathway accessible to the centre of this probe [58]. FreeSASA calculates

both the total SASA as well as the polar SASA, which were used to calculate $SASA_H$ as seen in Eqn (16):

$$SASA_H = \frac{SASA_{hydrophobic}}{SASA} \tag{16}$$

FreeSASA ensures accuracy through multiple verification steps. The correctness of the SASA calculation is verified through visual inspection, analytical methods for simple systems, and cross-verification using two independent computational algorithms [66].

2.6.6. Electrostatic Descriptors

Understanding the electrostatic properties of both mucin proteins and NPs is essential for characterizing their interactions and subsequent formation of the protein corona (PC). The formation of a PC is influenced by the inherent properties of NPs, such as size and surface charge, as well as the characteristics of the proteins they encounter.

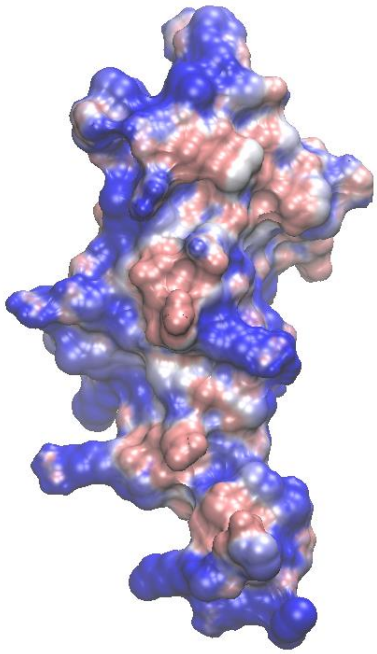


Figure 14: Electrostatic Potential Map of MUC5AC Protein

This figure shows the electrostatic potential map of the MUC5AC protein, coloured by potential values: blue for negative and red for positive regions. The map was generated using the PDB2PQR and APBS servers [67]. PDB2PQR assigned atomic charges and radii, converting the PDB file to a PQR file. APBS calculated the electrostatic potential from the

PQR file. This map highlights the charged regions, essential for understanding MUC5AC interactions with NPs.

Mucins, like MUC5AC, are heavily glycosylated proteins with significant negative charge due to the presence of sialic acid and sulphate groups [68]. This negative charge influences how mucins interact with NPs, especially those with positively charged surfaces. According to recent findings, reduced sialylation of mucins, such as MUC5B—a closely related mucin—can lead to a decreased negative charge, thereby affecting the electrostatic properties essential for mucociliary transport. These changes contribute to a less expanded and more compact mucin conformation, potentially hindering effective interactions with positively charged or hydrophilic regions of NPs. Mapping the electrostatic potential of mucins can identify regions of high negative charge, critical for predicting mucin behaviour in the presence of NP.

Understanding these charge distributions is important for predicting mucin behaviour in the presence of NPs. Figure 14 shows the electrostatic potential map of MUC5AC, highlighting the distribution of negative (blue) and positive (red) regions on the protein surface. This visualization illustrates the extensive negative surface charge of MUC5AC, particularly in regions with sialic acid and sulphate groups.

NPs such as silicon dioxide (SiO_2) can be engineered with various surface charges. The surface charge of SiNPs can be functionalized to have either positive or negative surface charges. The electrostatic optional of NPs is typically mapped using tools like Surface Racer, which calculates accessible surface areas, molecular surface area, and the curvature of molecular surfaces, providing a comprehensive profile of the NP's properties [69].

When mucins interact with NPs, the resulting PC formation is governed by the electrostatic compatibility between the two entities. Positively charged regions of the NP are likely to attract the negatively charged regions of the mucin, leading to a stable adsorption layer. This interaction can be quantitatively analysed by calculating the charged surface areas, and the

electrostatic potential using Surface Racer. By assessing these properties, we can predict the stability and behaviour of the mucin-NP complex.

2.7 Markov State Modelling

Markov State Models (MSM) are a sophisticated statistical framework used to understand the long timescale dynamical behaviours of complex molecular systems. These models offer a way to analyse and predict system dynamics over extended periods, which are often inaccessible through direct MD simulations due to computational constraints [70].

At its core, MSM is built upon the principles of the Master Equation, which describes the evolution of probabilities across a discretized state space. In practical terms, the vast configuration space of a molecular system is partitioned into a finite number of states. This partitioning can be based on geometric or kinetic criteria, often employing methods like Voronoi tessellation. Here, conformations of molecules are grouped into the same state if the transitions between them occur swiftly and frequently compared to transitions to other stages. This method ensures that each state ideally represents a local energy basin without significant internal barriers, adhering to the Markovian assumption – that the future state depends only on the current state and not the path taken to reach it.

The challenge in defining states lies in the balance between maximizing the number of conformations per state for robust statistical sampling and ensuring that no state spans multiple free energy barriers. If a state were to contain such barriers, it would violate the Markov property, as the direction from which the state was entered could influence future transitions, skewing the predictive power of the model.

MSMs leverage data from an ensemble of trajectories, capturing the conformation of each molecule at set intervals known as the ‘lag time.’ This lag time is important; it must be sufficiently long enough to ensure that the memory from a prior state is lost (satisfies the

Markov property), yet short to capture the essential dynamics and transitions of the system. At each interval, the state of the molecule is recorded, contributing to a transition count matrix N, where N_{ij} represents the number of observed transitions from state i to state j.

To enforce equilibrium conditions – where the inflow and outflow of each state are balanced – the matrix is often symmetrized. The symmetrized version N_{sym} adjusts the raw counts to reflect a balanced flow by averaging the forward and reverse transitions, $N_{ij}^{sym} = \frac{N_{ij}+N_{ji}}{2}$.

The estimation of transition rates between states is a key step for capturing the system's dynamics accurately. The branching probability, denoted as δ_{ji}, represents the likelihood of transitioning from state i to state j. It is calculated from the symmetrized transition count matrix N_{sym} as follows:

$$\delta_{ji} = \frac{N_{ji}^{sym}}{\sum_{l=1}^{N} N_{li}^{sym}} \tag{17}$$

The denominator sums all transitions originating from state i to any other state, which ensures normalization of probabilities. The average time spent in state i, denoted as T_i, gives information on the temporal aspects of state occupancy and dynamics within the model. The transition rate from state i to state j can be constructed using the branching probability and the average residence time in state i:

$$k_{ji} = \frac{\delta_{ji}}{T_i} \tag{18}$$

This formulation links the probability of transition of the rate at which transitions occur, thereby providing a dynamic view of state changes over time. It effectively captures how quickly or slowly the system moves out of state i into state j.

The lifetime-based method of estimating the transition rate matrix [71], K, utilizes the above relationships. Each non-diagonal element k_{ji} of matrix $K$is calculated using the normalized

branching probabilities divided by the average residence time. This approach ensures that the matrix K accurately reflects the dynamics observed across multiple trajectories or simulation runs.

In maintaining equilibrium and ensuring detailed balance, the diagonal elements of the transition rate matrix, k_{ii} are particularly important. They are calculated to ensure that the total rate of leaving state i equals the sum of the rates entering state i from all other states. This is represented as:

$$k_{ii} = -\frac{1}{T_i} = \sum_{j \neq i, j=1} k_{ji} \qquad (19)$$

This Eqn (19) reflects that the sum of the probabilities of leaving state i for all other states j must be normalized by the negative inverse of the average time spend in state i. This ensures that over time, the probability of being in any given state reaches a steady state, fulfilling the Markov property.

2.8 Master Equations

The dynamics for a protein in a solution are often explored through the Master Equation, a powerful mathematical tool that offers a comprehensive statistical description of state transitions with such complex systems. This equation is a first-order differential equation that effectively quantifies the flow of population among various states within the system.

The configuration space for a molecular system is typically divided into a set of discrete, non-overlapping metastable states. This division into states is based on clustering techniques such as Voronoi tessellation, which might be informed either geometrically or kinetically. Molecules or configurations are grouped into the same state if transitions between them occur quickly compared to transitions to other states. This ensures that each state representing a region of the conformational space where the system behaves similarly. The goal is to balance the granularity

of these states – maximizing the number of configurations per state while avoiding internal free energy barriers that would compromise the Markovian nature of the system. Essentially, the presence of significant energy barriers within a state would violate the Markovian assumption, as the system's future behaviour would depend on its paths into the state, not just the state itself.

The transition between states is governed by transition rates, which are derived from the probability of the system moving from one state to another. These rates are crucial aspect of the Master Equation and are analogous to elements of a rate matrix in the Mori-Zwanzig formalism [72], a theoretical framework that separates relevant system dynamics from those that can be considered background noise using projection operators.

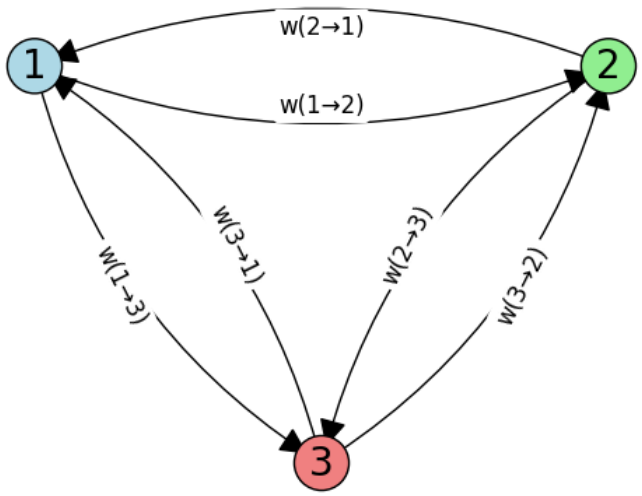


Figure 15: A directed graph illustrating a Markov chain with three distinct states, represented by coloured circles.

The arrows between the states indicate the possible transitions, with the rates $w(i \to j)$ specifying the transition rate from state i to state j.

The Master Equation can be derived starting from a simple two-state system, denoted as state 1 and state 2. Let us define $P_i(t)$ as the population of state i at time t, and $w(i \to j)$ as the transition rate from state i to state j.

Consider the population of state 1 at a slightly later time $t + dt$. This new population $P_1(t + dt)$ is influenced by two main factors:

(a) The portion of the population that remains in state 1, having not transitioned to state 2 within the small-time interval *dt.*

(b) The portion of the population that transitions from state 2 to state 1 within this same interval.

Mathematically, this can be represented as:

$$P_1(t+dt) = -P_1(t)\big(1 - w(1 \to 2)\big)dt + P_2(t)w(2 \to 1)dt \qquad (20)$$

Here, $P_1(t)w(1 \to 2)dt$ quantifies the expected decrease in P_1 due to transitions to state 2, and $P_2(t)w(2 \to 1)dt$ represents the influx into state 1 from state 2. This framework can be generalized to a system with N states, where the rate of change of the population in any state *i*, $\frac{dP_i}{dt}$ is given by summing the contributions from all other states j (where $j \neq i$):

$$\frac{dP_i}{dt} = \sum_{j \neq i}^{N} [-w(i \to j)P_i(t) + w(j \to i)P_j(t)] \qquad (21)$$

This sum includes the loss of population from state *i* to all other states *j*, and the gain from all states *j* transitioning into state *i*. As *dt* approaches zero, the likelihood of multiple transitions occurring within this infinitesimally small interval decreases with an order of $O(dt^2)$, indicating that such events become exceedingly rare. The direct connections between states are not necessary for transitions as they can occur through a sequence of intermediate states $i_1, i_2, \dots, i_m$ if there exists a non-zero probability pathway connecting these states, thus allowing for indirect transitions.

Transforming the population states into a vector representation offers a powerful approach to analysing systems with multiple states. By representing the populations as a vector $\mathbf{P} = (P_1, \dots, P_N)$, the dynamics of the entire systems can be captured into a single equation. This is facilitated by encoding the transition rates between states into a matrix form.

Expressing the rate of change of the vector **P** in terms of time *t gives* us the following matrix equation:

$$\frac{d\vec{P}}{dt} = \mathrm{K}\vec{P}(t) \tag{22}$$

Here, K is an $N \times N$ matrix referred to as the transition rate matrix, where each element k_{ij} of K is defined by the transition rates between states. Specifically, the off-diagonal elements k_{ij} (where $i \neq j$) are set as:

$$k_{ij} = w(j \to i), \text{ if } i \neq j \tag{23}$$

$$k_{ii} = -\sum_{j \neq i} w(j \to i) \tag{24}$$

As seen in Eqn (24), the diagonal elements k_{ii} of matrix K are determined by the negative sum of all transition rates into state i from all other states j, ensuring the conservation of total population. This condition ensures that the total rate of leaving any given state i is balanced by the sum of the rates of entering i from all other states, maintaining the overall stability of the system. The arrangement of the matrix K ensures that each column sums to zero, except for the diagonal elements. This property is crucial as it reflects the principle that the total probability within the system is conserved over time. The transition rate $w(j \to i)$ and their reverse $w(i \to j)$ can vary independently and are not confined to the interval $[0, 1]$. This flexibility allows the modelling of a wide range of real-world systems where transitions might not necessarily be probabilistic or symmetric.

The effectiveness of the Master Equation in modelling the dynamics of a system hinges on its adherence to the Markov property, which posits that the future state of a system is dependent solely on its present state, not on the path taken to reach that state. Eqn (25) illustrates this principle:

$$P(X_{n+1} = i \,|X_n = j, X_{n-1}, \dots, X_1) = P(X_{n+1} = i|\, X_n = j) \tag{25}$$

Even if the system exhibits non-Markovian behaviour over short time scales, Eqn (22), which describes the system dynamics through a differential matrix equation, remains valid. For systems with extended non-Markovian characteristics, a generalized Master Equation is typically more appropriate to accurately describe the dynamics.

In a state space where all states are positive recurrent, the system will eventually reach a unique, stationary equilibrium distribution P_{eq}. At equilibrium, the net flow of population between states ceases, expressed as:

$$\frac{d\boldsymbol{P}}{dt} = 0 \tag{26}$$

The equilibrium condition is defined as follows:

$$KP_{eq} \approx 0 \tag{27}$$

This implies that P_{eq} is a non-zero vector normalized such that the sum of its entries equals one, ensuring that it a probability distribution:

$$P_{eq}(i) > 0 \text{ and } \sum_{i=1}^{N} P_{eq}(i) = 1 \tag{28}$$

Under these conditions, the detailed balance condition, which states that at equilibrium the rate of transition into any state equals the rate of transition out of that state, can be applied:

$$k_{ij}P_{eq}(j) = k_{ji}P_{eq}(i) \tag{29}$$

This condition leads to the concept of a symmetric transition rate matrix, which be constructed to reflect this balance. The symmetric matrix K_{ij}^{sym} is derived by adjusting the original matrix K with the equilibrium probabilities:

$$k_{ij}^{sym} = P_{eq}^{-\frac{1}{2}}(i)k_{ij}P_{eq}^{\frac{1}{2}}(j) \tag{30}$$

The matrix notation is presented in Eqn (31), where P_{eq} is given as a diagonal matrix with the equilibrium probabilities along its diagonal, and $P_{eq}^{-1/2}$and $P_{eq}^{1/2}$ are matrices formed by taking the reciprocal square root and the square root, respectively, of the diagonal entries of P_{eq}.

$$\mathbf{K}^{sym} = \mathbf{P}_{eq}^{-\frac{1}{2}}\mathbf{K}\mathbf{P}_{eq}^{\frac{1}{2}} \text{ [71]} \tag{31}$$

This transformation ensures that the matrix K^{sym}, which adjusts K by the square roots of the probabilities, reflects the symmetry implied by detailed balance.

$$(K^{sym})^2 = KK \tag{32}$$

Thus, K^{sym}effectively represents the probability flows within the system under equilibrium conditions, emphasizing the self-regulating natures of the dynamics in a system that adheres to Markovian principles but may need to accommodate more complex interactions or memory effects in a non-Markovian framework.

The dynamics of a system represented by the transition rate matrix K can be understood by examining its eigenpairs. By constructing K^{sym}, one can extract information about the system's dynamics through its eigenvalues and eigenvectors. Both K andK^{sym} share the same eigenvalues, indicating that the transformation to a symmetric matrix does not alter the spectrum but simplifies the interpretation of the dynamics:

$$K^{sym}\phi_n = \lambda_n\phi_n \tag{33}$$

Where ϕ_n are the orthonormal eigenvectors and λ_n are the corresponding eigenvalues of K^{sym}. The first eigenvalue, λ_1, is zero, which corresponds to the equilibrium mode of the system where no net changes occur. The subsequent eigenvalues are real, negative and organized in decreasing magnitude: $\lambda_1 = 0 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_N$.

The left and right eigenvectors, $\psi^{\boldsymbol{L}_n}$ and $\psi^{\boldsymbol{R}_n}$ respectively, are bi-orthonormal:

$$\psi^{\boldsymbol{L}_n}\psi^{\boldsymbol{R}_n} = \delta_{ij} \tag{34}$$

There are computed as:

$$\psi_n^L K = \lambda_n \psi_n^L \text{ and } \quad K\psi_n^R = \lambda_n \psi_n^R \tag{35}$$

The left and right eigenvectors of K^{sym} are related to those of K by the equilibrium probabilities:

$$\psi_n^L(i) = P_{eq}^{-\frac{1}{2}}(i)\phi_n(i)$$

$$\psi_n^R(i) = P_{eq}^{\frac{1}{2}}(i)\phi_n(i) \tag{36}$$

Notably, the equilibrium population corresponding to $\lambda_1 = 0$ ensures that:

$$\psi_1^{\boldsymbol{R}_i} = P_{eq}(i)$$

As well:

$$\psi_1^{\boldsymbol{L}_i} = P_{eq}^{-\frac{1}{2}}(i)P_{eq}^{\frac{1}{2}}(i) = 1$$

The orthonormality conditions of K^{sym}'s eigenvectors maintain the systems integrity and help in understanding the interactions within:

$$\phi_n \cdot \phi_m = \sum_{i=1}^{N} \phi_n(i)\phi_m(i) = \delta_{nm} \tag{37}$$

Where δ_{nm} is the Kronecker delta.

$$= \sum_{i=1}^{N} \psi_n^L(i)\psi_m^R(i) \tag{38}$$

$$= \sum_{i=1}^{N} \psi_n^L(i)\psi_m^L(i)P_{eq}(i) \tag{39}$$

$$= \sum_{i=1}^{N} \psi_n^R(i)\psi_m^R(i)P_{eq}^{-1} \tag{40}$$

We can observe that for $m = 1$, the sum of the right eigenvector components $\psi_m^R(i)$ for the equilibrium state directly equals the sum of the equilibrium probabilities weighted by the left eigenvector components of ψ_m^L:

$$\sum_{i=1}^{N} \psi_m^R(i) = \sum_{i=1}^{N} P_{eq}(i)\psi_m^L = \delta_{1m} \tag{41}$$

This condition emphasizes that:

- For $m = 1$: $\psi_m^R(i) = P_{eq}(i)$, which sums to 1, upholding the conservation of total probability,
- For $m \neq 1$: These components sum to zero, reflecting modes that deviate from equilibrium.

The dynamics of the system can be further be explored through its implied timescales (ITS), which relate to the eigenvalues of the transition rate matrix:

$$t_i = -\frac{\tau}{ln|\lambda_i(\tau)|} \tag{42}$$

These timescales estimate the decorrelation time of the dynamic process within the system providing information into how quickly the system transitions away from or towards equilibrium. By understanding these properties, we can get an understanding of the stability, transitions and long-term behaviour of the system.

To implement Markov state analysis, we can use the `PyEMMA` package [73], a Python toolkit designed to specifically construct and analyse Markov state models from MD simulations. `PyEMMA` simplifies the workflow by automating tasks such as clustering trajectory data into discrete states, estimating transition matrices that describe state-to-state transitions and calculating implied timescales that may help reveal the long-term dynamics of the system. This makes it essential for understanding the metastable states and their transitions for the mucin system

3. Modelling Results and Discussion

This study leverages atomistic MD simulations conducted on high-performance computing (HPC) resources to explore the conformational dynamics and interactions of Mucin 5AC (MUC5AC) within aqueous environments supplemented with silicon dioxide (SiNPs). The simulations, detailed in Table 1, were performed in a production NPT ensemble at a constant temperature of 310.15 K and pressure of 1 atm. The primary structure analysed was the MUC5AC D7 domain (amino acids 3518-3626) from the RCSB Protein Data Bank (PDB ID: 8ov0).[30].

***Table 1**: Overview of Simulations throughout Project*

Simulation ID	Description	System Components	System Size	Duration	Purpose
Sim1	Long simulation of solvated MUC5AC	MUC5AC, H2O (TIP3), NaCl ions	96 Å x 96 Å x 96 Å	1.8 µs	Detailed system behaviour analysis over extended period
Sim2	Swarm of short simulations of MUC5AC for MSM	MUC5AC, H2O (TIP3), NaCl ions	96 Å x 96 Å x 96 Å	15 runs, each 100 ns	Markov State Modelling for dynamic state analysis
Sim3	Simulation of SiO_2 NP with MUC5AC	SiO_2 Np (11 nm), MUC5AC, H2O (TIP3), NaCl ions	127 Å x 127 Å x 127 Å	300 ns	Interaction study between NP and mucin in ionic solution
Sim4	Simulation of smaller SiO_2 NP with MUC5AC	SiO_2 NP (4 nm), MUC5AC, H2O (TIP3), NaCl ions	90 Å x 90 Å x 90 Å	300 ns	Comparative study to Sim3 focusing on size effects of NPs

Each simulation employed NAMD 3.0 with CUDA GPU acceleration to ensure computational efficiency and utilized the CHARMM36m force field for accurate interaction potentials. The simulation workflow included stages of minimization, heating, equilibration in an NVT ensemble, and production in an NPT ensemble, carefully capturing the dynamics and interactions at the molecular level.

***Table 2**: Data of atomistic MD. Number of atoms in protein, silica NPs and total number of atoms (atoms and water)*

System	Number of Atoms	Total number of atoms (atoms + water + *mucin)
MUC5AC	1713	83629
SiO_2 NP 4nm	3192	67678*
SiO_2 NP 11nm	59622	178508*

3.1 Molecular dynamics Results

The results from this section will cover the MD simulations of MUC5AC that were conducted to analyse its conformational dynamics when solvated in water and NaCl ions. The initial phase involved energy minimization for 10000 timesteps to relax the system into a stable configuration, reducing potential energy (PE) and preparing the molecular structure for dynamic interactions, as shown in Figure 16 (a). Following this, the system was heated and equilibrated under an NVT ensemble for 10 ns. This phase allowed the protein to reach a consistent thermal equilibrium (at 310.15 K), with just the PE displaying regular fluctuations indicative of a stable system, as shown in Figure 16 (b).

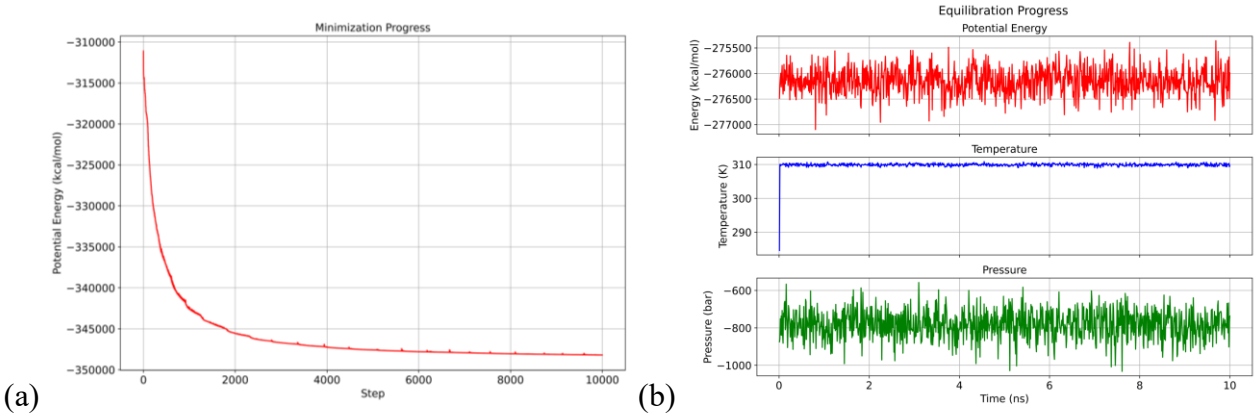


Figure 16: Minimization and Equilibration Phases in MD MUC5AC Simulation

(a) **Minimization Phase**: displays the potential energy (PE) as a function of the step number. The rapid decrease in PE indicates effective convergence towards a lower energy configuration, stabilizing around a minimum value.

(b) **Equilibration Phase:** Shows the time evolution of PE, temperature, and pressure. After initial fluctuations, the system achieves stability, demonstrating effective equilibration. The PE plot shows minor fluctuations around a stable mean, while temperature and pressure stabilize within target ranges, confirming the system's equilibration at desired conditions.

In the production phase, conducted under an NPT ensemble for a total duration of 1.8 µs, the system exhibited consistent stability, as evidenced by the steady PE, temperature, and pressure (Figure 17). Throughout this phase, the PE remained stable with only minor fluctuations, showing a well-maintained energy state. Similarly, temperature and pressure showed minimal deviations, staying within acceptable ranges, which highlights the effective regulation of the system's thermodynamic properties.

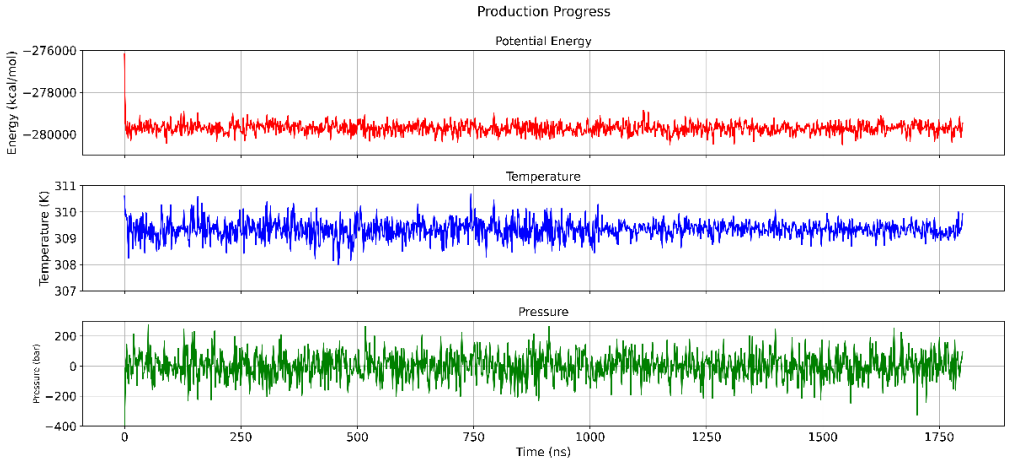


Figure 17: Production Phase in MD MUC5AC Simulation

Production Progress: The potential energy remains stable with minor fluctuations, indicative of a consistent energy state. Temperature and pressure show minimal deviations and acceptable fluctuations, respectively, demonstrating effective regulation and stability throughout the production phase. This ensures realistic simulation conditions are maintained.

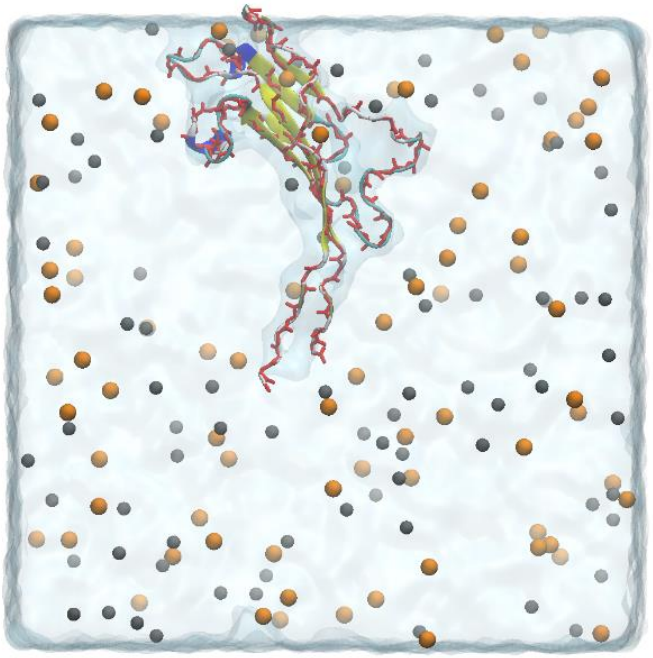


Figure 18: Snapshot of final frame of MUC5AC solvated in water and NaCl ions.

These results confirm the successful maintenance of realistic simulation conditions, providing a solid foundation for further analysis of MUC5AC dynamics and interactions.

3.1.1. RMSD

The RMSD trajectory of the MUC5AC MD simulation is depicted in Figure 19. The x-axis represents the simulation time in nanoseconds (ns), while the y-axis shows the RMSD values in angstroms (Å). The blue curve corresponds to the RMSD calculated with respect to the initial frame, while the orange curve represents the RMSD relative to the average structure over the simulation.

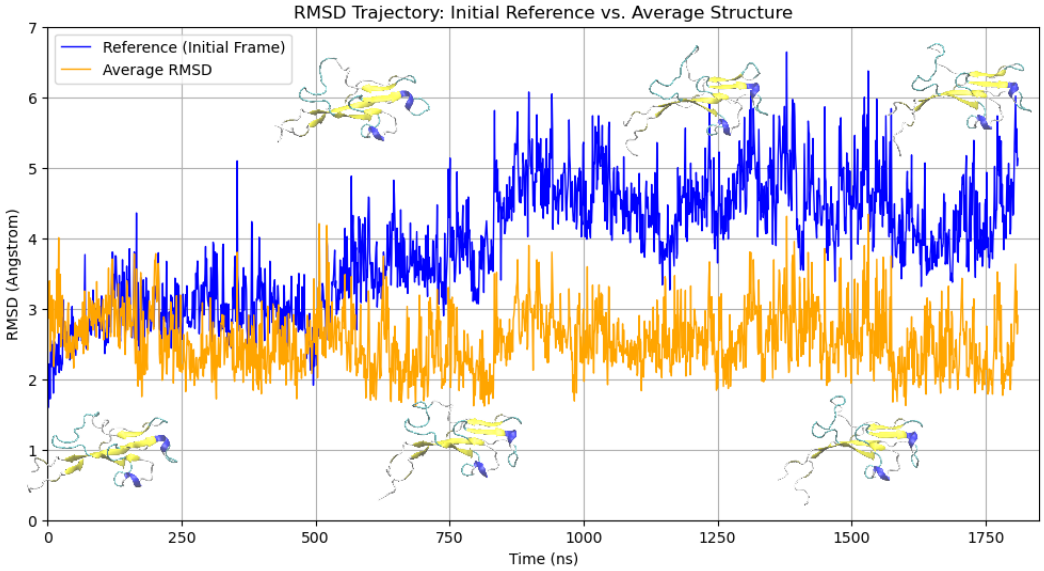


Figure 19: RMSD Analysis of Mucin Over Time Relative to Initial Frame and Average Structure

This plot shows the Root Mean Square Deviation (RMSD0 of mucin over time, comparing the structure to both the initial frame (blue) and the average structure (orange). Significant deviations occur after 500ns, indicating conformational changes. The RMSD stabilizes against the average structure, suggesting that the protein reaches a consistent conformation. Structural snapshots illustrate these changes.

For majority of the simulation, MUC5AC remains relatively stable, as expected, with only minor fluctuations in the overall structure. This stability is evident from the consistency of the orange curve, which indicates the RMSD relative to the average structure. The most notable structural changes occur in the tails of the protein, which exhibit fluctuations throughout the simulation. Additionally, the unfolding of a helical region can be observed, particularly around

the 750 ns mark, where the blue curve rises, suggesting a conformational shift. These observations suggest that while the core of the mucin remains stable, the dynamic behaviour is primarily seen in the protein's tails and specific secondary structures like helices.

3.1.2. Radius of Gyration results

The time evolution of the MUC5AC's Radius of Gyration (R_g) in Figure 20 shows the overall compactness during the MD simulation. Throughout the 1.8 μs the protein fluctuates within a range of approximately 16.0 to 17.4 Å, indicating periodic expansions and contractions of the protein structure. These fluctuations can reflect the dynamic nature of mucins, known for their flexible and extended structures that allow them to perform various biological functions. The average R_g, indicated by the green dashed line, shows that the protein remains stable around 16.6 Å, with occasional significant deviations (marked by red points), suggesting localized unfolding events. Such deviations are often associated with the flexibility of mucin's extended domains, which allow it to adapt to different environmental conditions, such as pH changes or the presence of binding molecules.

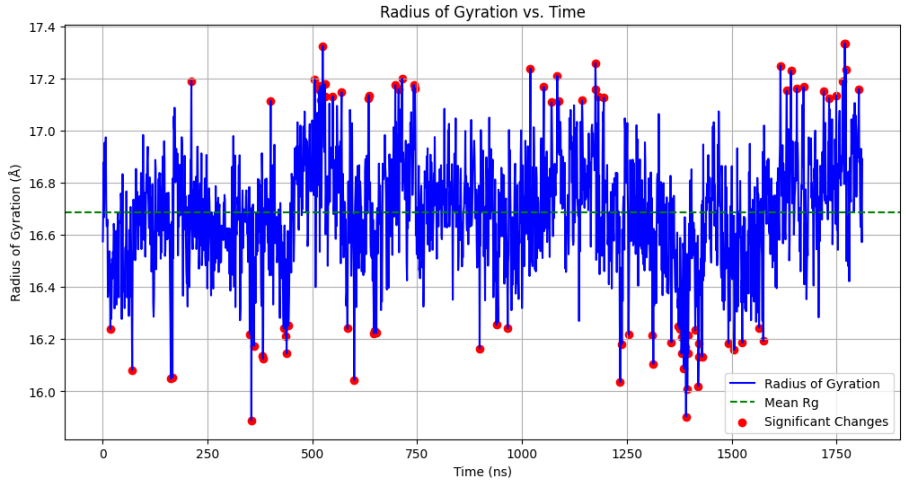


Figure 20: Radius of Gyration R_g vs Time (ns)

This plot shows the time evolution of R_g of the MUC5Ac protein over the course of the MD simulation. The Rg fluctuates around a mean value, with significant deviations marked in red. These fluctuations suggest the structural changes in the protein, particularly in the flexible regions such as tails and helices.

The Probability Density of Rg $P(Rg)$ plot in Figure 21 provides a statistical view of the structural fluctuations in MUC5AC. The distribution peaks around 16.6 Å, highlighting that

the mucin predominately adopts a compact configuration during the simulation, which aligns with the functional role of mucins in creating dense, protective barriers on epithelial surfaces. This distribution reflects the balance between the protein's intrinsic flexibility and the stabilizing forces in its environment, such as electrostatic interactions with ion like NaCl present in the simulation. This is typical for mucins, which need to maintain a certain degree of compactness to effectively perform their gel-forming functions while still allowing for the flexibility required for interacting with various biological molecules.

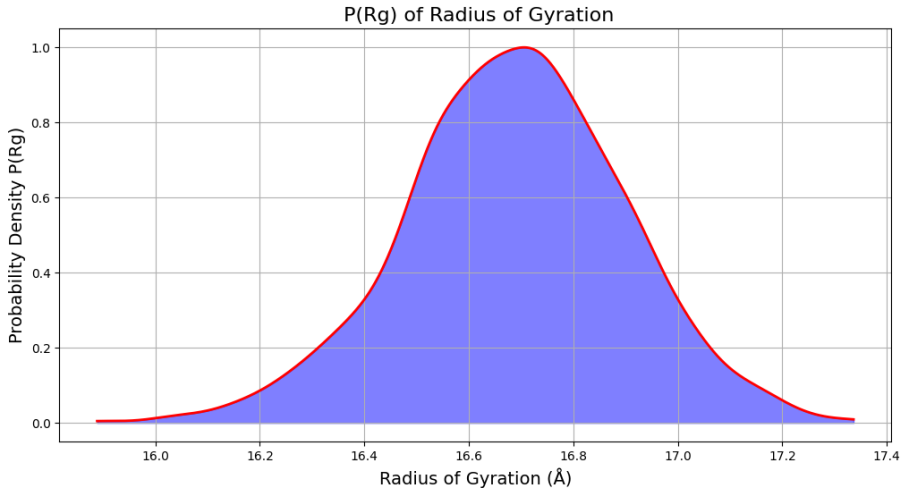


Figure 21: Probability Density of Radius of Gyration $P(R_g)$

This figure shows the probability distribution of the Rg for the MUC5AC protein. The distribution is centred around the mean Rg value of approximately 16.6 Å, indicating that the protein predominately maintains this compactness during the simulation.

3.1.3. End-to-End distance

The End-to-End distance (ETED) of the terminal atoms in MUC5AC were measured and show in Figure 22. The fluctuations seen in the plot indicate the flexible and dynamic nature of the mucin in the simulation. The range of values, between approximately 35 and 70 Å, reflects the changes in conformation, which may correspond to stretching and compaction events that occur as the molecule explores its conformational space. The general trend of the ETED stabilizing suggests that the molecule achieves a relatively consistent conformational state throughout the simulation.

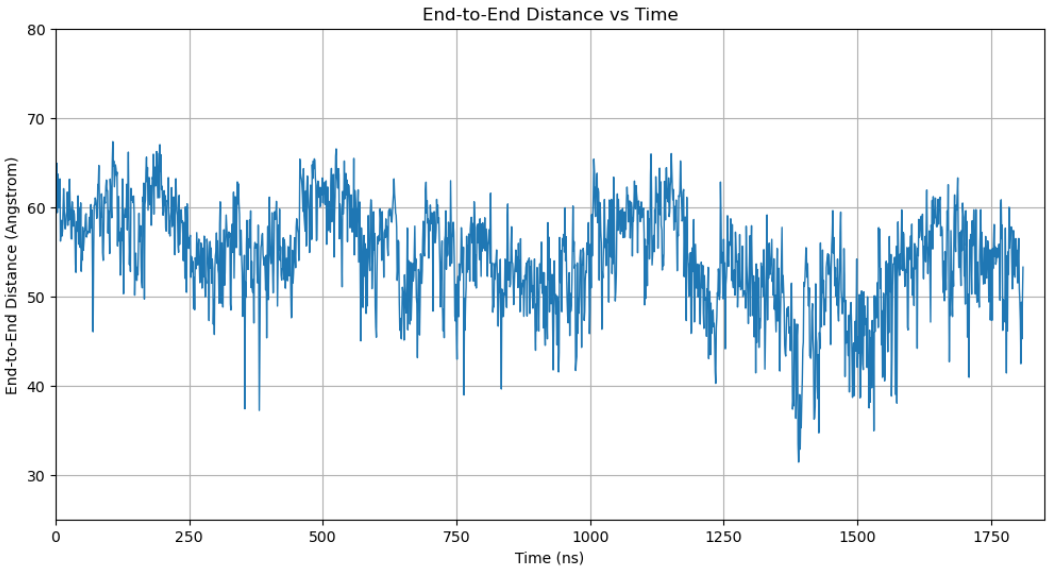


Figure 22: End-to-End Distance (Å) vs Time (ns) for MUC5AC

Plot displays the ETED of MUC5AC, showing the fluctuations in distance between terminal residues throughout the simulation. The dynamic changes in distance indicate conformational variations in the molecule, reflecting folding, stretching and other structural adjustments over the course of the simulation.

3.1.4. Hydrogen Bond

An analysis of the HBONDs within MUC5AC protein structure and solvent molecules provides details on the stability and dynamics of the molecule during the simulation. Figure 23 shows the time evolution of the HBONDs for different donor-acceptor distance (3.0 Å, 3.5 Å, and 4.0 Å) at two angular cutoffs: 20 degrees (top plot) and 30 degrees (bottom plot). The results show that as the distance increases, the number of HBONDs also increases, which is consistent with expectations because a longer cutoff distance encompasses more potential HBOND pairs. However, these bonds are generally weaker compared to those formed at short distances.

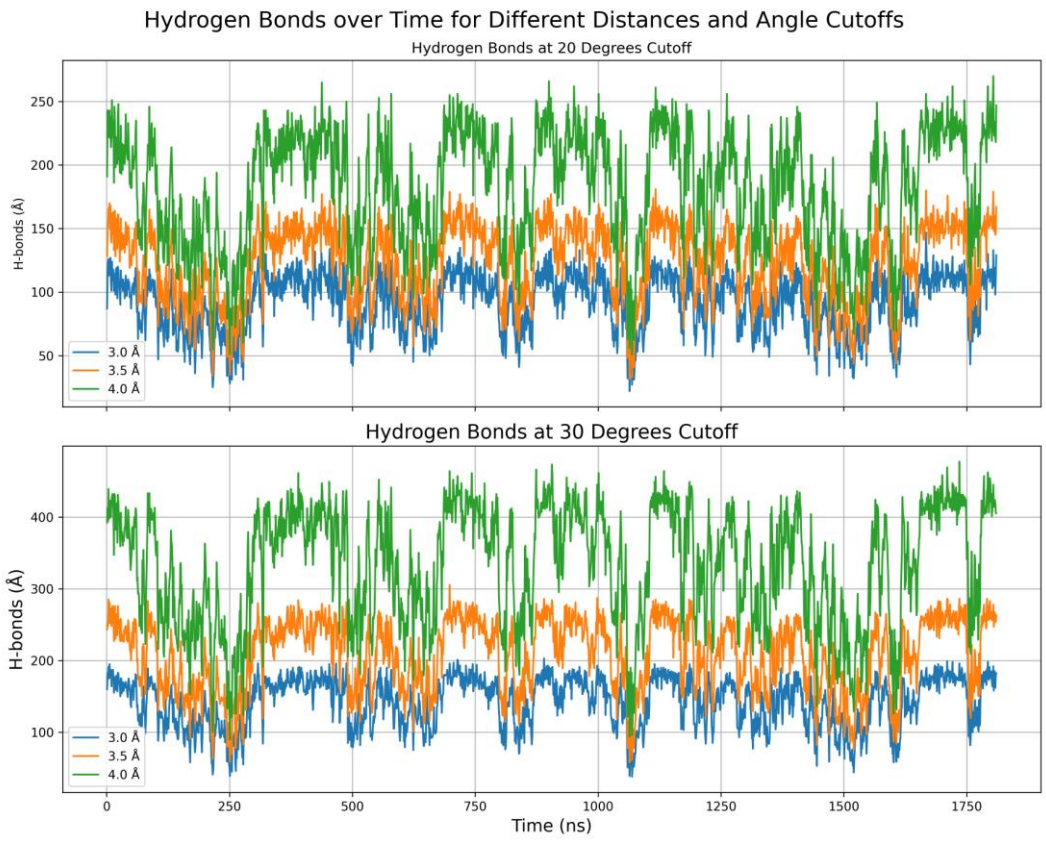


Figure 23: Hydrogen Bonds over Time for Different Distances and Angle Cutoffs

This figure displays the hydrogen bond count over time in the solvated MUC5AC system for two angle cutoffs (20 and 30 degrees) and three donor-acceptor distances (3.0 Å, 3.5 Å, and 4.0 Å). The top plot shows results for the 20-degree cutoff, and the bottom plot shows results for the 30-degree cutoff.

Both 20-degree and 30-degree angle cutoffs show similar trends: more hydrogen bonds are detected as the simulation progresses, particularly at the larger 4.0 Å cutoff. This suggests that the solvated mucin undergoes dynamic structural changes, with solvent molecules participating in hydrogen bonding interactions with the protein. This suggests that the interactions are looking to stabilize the protein's structure and maintain its functional conformations during the simulation.

3.1.5. SASA

The SASA of MUC5AC was calculated over the course of the simulation using a TCL script. This showed how much the protein's surface was exposed to the solvent, which can be used as an indicator of the protein's folding state and the interactions with its environment.

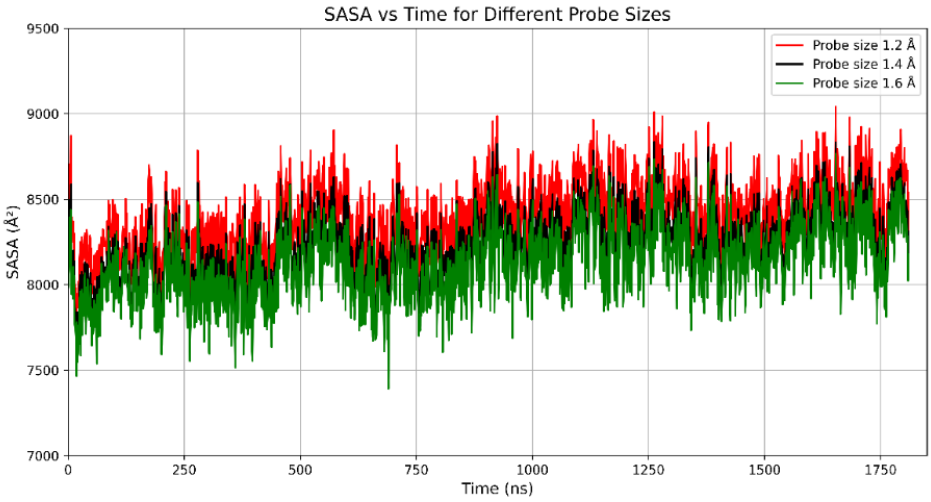


Figure 24: SASA vs Time for Different Probe Sizes in MUC5AC Simulation

This figure illustrates the SASA over time during the 1.8 µs MUC5AC simulation, evaluated using three different solvent probe radii: 1.2 Å (red), 1.4 Å (black), and 1.6 Å (green). The varying probe sizes help capture differences in surface accessibility, revealing the protein's interaction with the solvent across multiple spatial scales.

To explore potential differences in the hydrophobic solvent accessible surface area (SASAH), three different probe radii were tested: 1.2 Å, 1.4 Å (default for water), and 1.6 Å. These variations allowed for an examination of how different solvent probe sizes impact the SASA values and the hydrophobic regions of the protein exposed to the solvent.

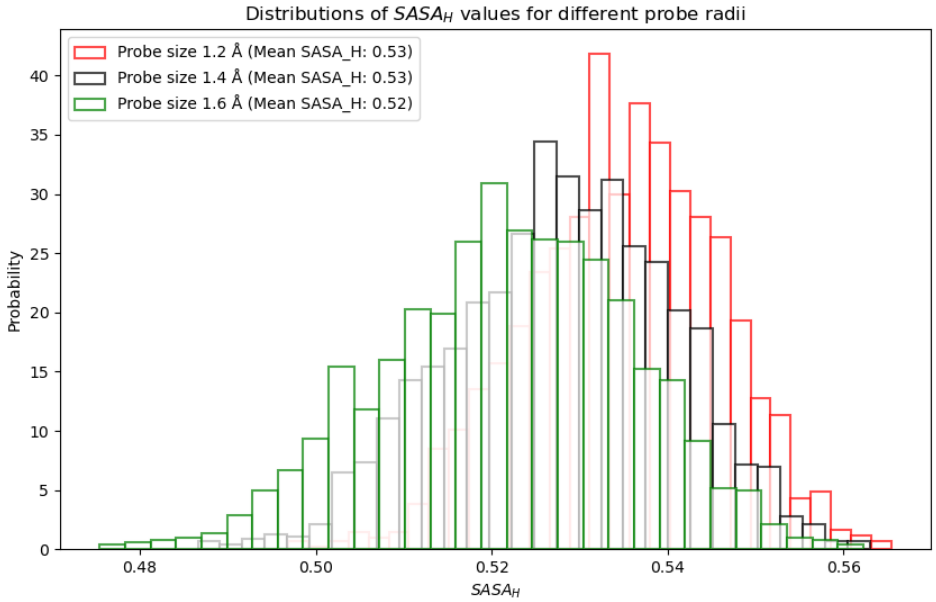


Figure 25: Distributions of SASA_H Values for Different Probe Radii

This figure shows the distributions of hydrophobic solvent-accessible surface area (SASA_H) values for MUC5AC, calculated with probe sizes of 1.2 Å, 1.4 Å, and 1.6 Å. Each distribution highlights how different probe radii affect the solvent interaction with hydrophobic surface regions.

The distribution of hydrophobic solvent accessible surface area ($SASA_H$) values seen here is centred around 0.52 to 0.54, which represents roughly half of the total accessible surface area. This is consistent with the expected behaviour for proteins, where a significant portion of the surface is composed of hydrophobic regions. These regions, while buried in a folded state, may become accessible to solvent under specific conformational changes. The similarity across different probe radii suggests stable hydrophobic exposure in this mucin protein, a common trait for proteins with a mixture of hydrophobic and hydrophilic surface residues.

3.1.6. Surface Charge of MUC5AC Protein

Using Surface Racer, the SASA of the MUC5AC protein was also calculated, but focusing on its charged regions. The analysis showed that the positively charged surface area (SASA+) constitutes approximately 13.56% of the total SASA, while the negatively charged surface area (SASA-) accounts for about 8.17%. This distribution suggests a predominance of positively charged regions on the protein's surface, which could significantly impact its electrostatic interactions with other molecules, such as NPs or other biological entities.

3.2 Residue Secondary Structures

The secondary structure composition was analysed to determine the structural characteristics of MUC5AC from the 15 series of 100 ns MD trajectories. This analysis focused on identifying the percentage of time each residue spent in specific secondary structures, including β-strands, α-helices, and coils. The secondary structures were determined using DSSP (Dictionary of Secondary Structure of Proteins), a standard tool for assigning protein secondary structures from atomic-resolution coordinates [74].

To ensure robustness of the results, the standard error of the mean (SEM) was calculated by sampling the trajectories multiple times. This approach provides a reliable estimate of the variability in secondary structure content across different regions of the protein.

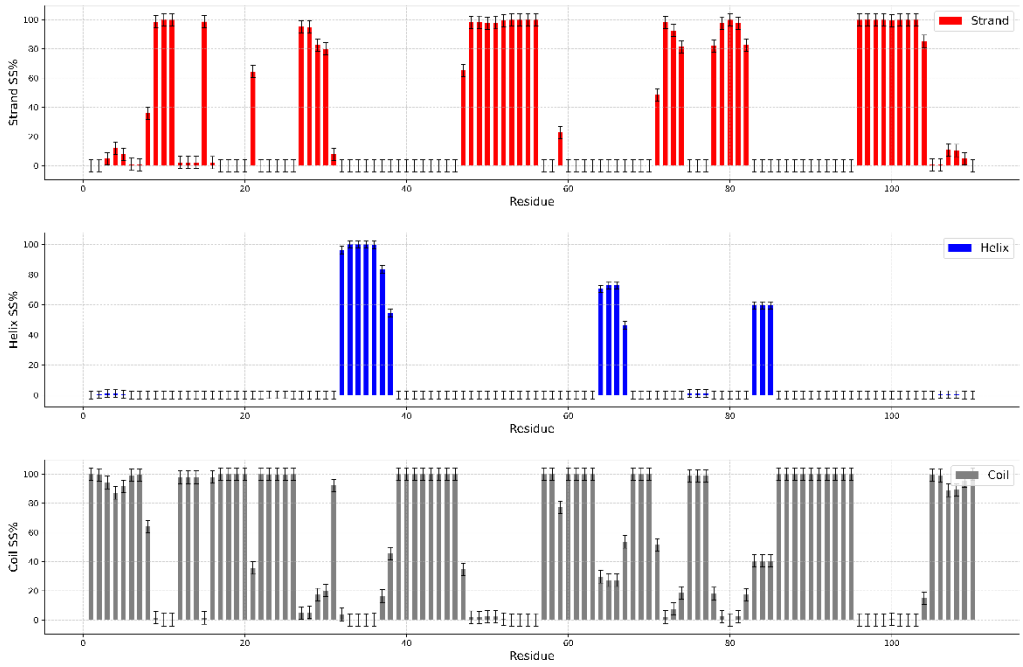


Figure 26: Comparison of Secondary Structure Percentages per Residue in MUC5AC

The figure shows percentage of time each residue in MUC5AC spent in different secondary structures. The top panel represents the strand structures, the middle panel shows the helix content, and the bottom panel indicates the coil structures. The error bars represent the standard error of the mean (SEM) across the sampled trajectories.

Figure 26 shows the comparison of the percentage of time spent in each secondary structure per residue for the MUC5AC protein. The protein showed a diverse range of secondary structures, with certain regions predominantly adopting coils configuration, while others displayed a high β-strand or helix content. The most prevalent secondary structure in MUC5AC is the coil configuration, observed across most of the residues. Coils provide the flexibility needed for the protein's function in forming mucus, allowing MUC5AC to expand, contract, and adapt to the dynamic environment of the mucus layer.

β-strand structures are observed in specific regions, notably around residues 20-40 and 60-80. These regions likely contribute to forming stable, sheet-like structures within the mucus, which could play a role in cross-linking interactions with other mucins or mucus components [75]. These β-strand regions may contribute to the gel-like properties of mucus, providing structural integrity and resistance to mechanical stress. The central region of MUC5AC displays

noticeable α-helix content, particularly around residues 30-40 and parts of 60-90. Helices in mucins are often associated with maintaining the protein's elongated structure, which is essential for the formation of the mucus matrix [76]. These helical regions may also facilitate interactions with other proteins and contribute to the overall stability of the mucin network.

The prevalence of coil structures suggests that MUC5AC is highly flexible, which is essential for its role in maintaining the fluid and adaptive properties of mucus. Meanwhile, the β-strand and helix regions likely contribute to the structural stability and organization within the mucus matrix.

3.3 Markov State Modelling

The `PyEMMA` Python package, along with `Deeptime` (due to some deprecated packages in `PyEMMA`), was utilized to perform Markov state modelling and analysis on the trajectories of MUC5AC. Backbone torsion angles, specifically the ϕ (phi) and ψ (psi) dihedral angles representing rotations around the C-N and C-C bonds, were chosen as the primary features for analysing the protein's conformational space. These torsion angles initially showed a significant advantage over CA distances and contacts at smaller lag times, though this gap diminished as lag times increased and fast kinetic processes decayed. Given their superior ability to represent the kinetics of MUC5AC mucin, backbone torsion angles were selected as the main feature for subsequent analysis.

The conformational space of MUC5AC was originally characterized by 436 backbone torsion angles. To facilitate more effective analysis and visualization, methods such as Principal Component Analysis (PCA), Time-lagged Independent Component Analysis (TICA), and Variational Approach for Markov Processes (VAMP) were employed to reduce the dimensionality of this data.

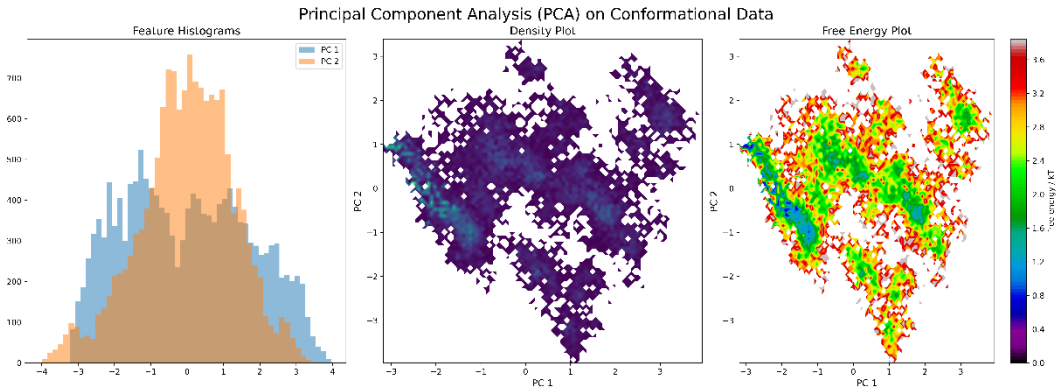


Figure 27: Principal Component Analysis (PCA) on Conformational Data

This figure presents the results of PCA applied to the conformational data of MUC5Ac. The left panel shows the histograms of the two principal components (PC1 and PC2). The middle panel displays the density plot of the data in the reduced PC space, while the right panel illustrates the free energy landscape, highlighting regions of low and high free energy across the principal components.

PCA was applied to capture the principal modes of variation in the data, resulting in a reduction to two principal components (PCs). TICA, which focuses on capturing the largest variance in the data, is particularly useful for identifying slow dynamic processes [77]. It does so by building linear combinations of features that maximize their autocorrelation over a specific lag time. This means that TICA produces reaction coordinates that are most relevant for slow transitions in the system, which can help describe long-timescale behaviours such as protein unfolding, conformational changes or ligand binding. TICA was used to reduce the dimensionality to two independent components (ICs) and partition the conformational space into kinetically meaningful states. Finally, VAMP was employed to explore the kinetic landscape of the protein by focusing on the slowest dynamical processes.

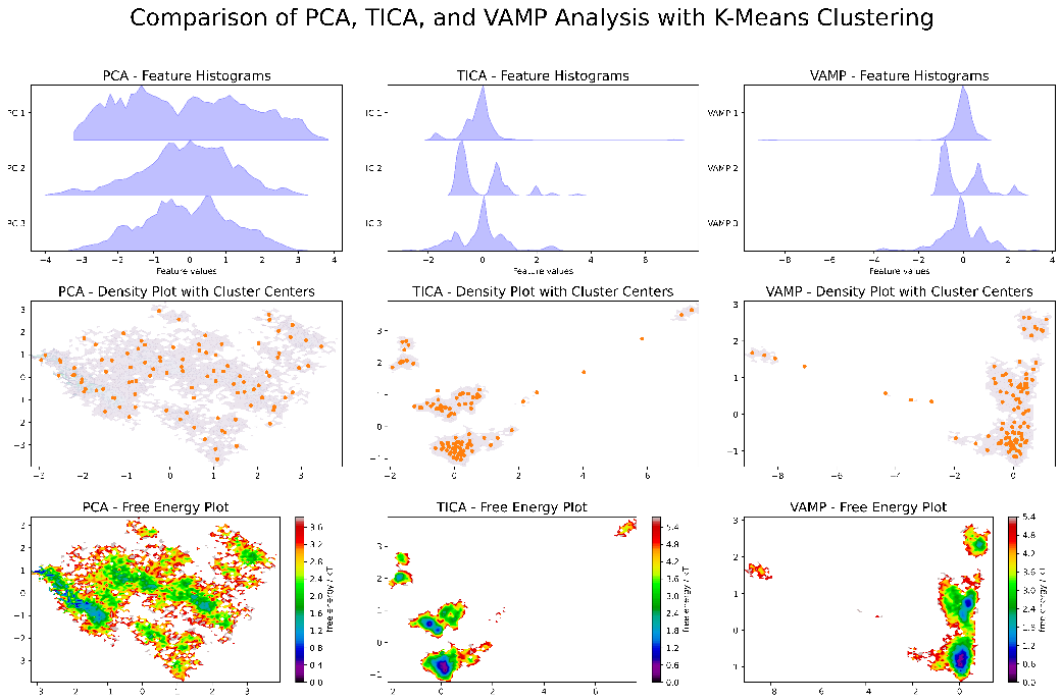


Figure 28: Comparison of PCA, TICA, and VAMP Analyses with K-Means Clustering

This figure compares PCA, TICA, and VAMP analyses of MUC5AC conformational data. The top row shows feature histograms for each method, illustrating the distribution of principal components (PCA) and independent components (TICA, VAMP). The middle row presents density plots with K-means cluster centres overlaid, identifying key conformational states. The bottom row depicts the corresponding free energy plots, revealing the energetic landscape and stability of the identified states.

To discretize the conformational space, the K-means clustering algorithm [78] was applied to the reduced data. K-means is an unsupervised learning method that partitions the data into predefined number of clusters through minimizing the variance within each cluster. Therefore, grouping these data points will give cluster, providing a discretized representation of the conformational space, which is essential for building MSMs. The clustering centres represent key conformational states, and each frame of the trajectory is assigned to is nearest cluster centre, forming discrete trajectories for kinetic analysis.

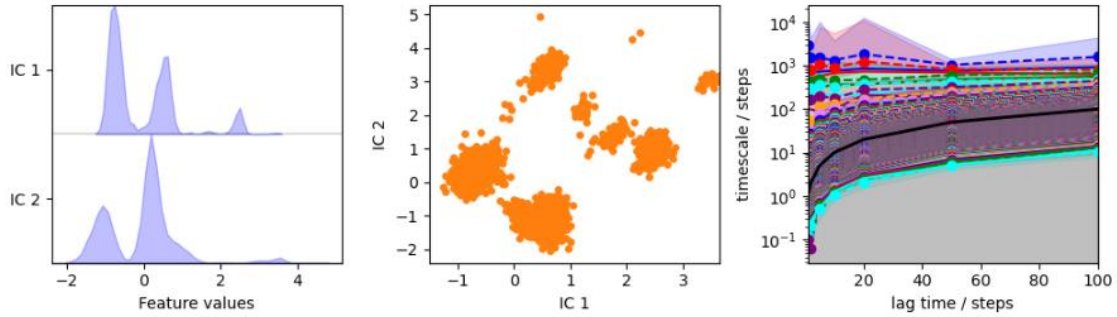


Figure 29: Time-lagged Independent Component Analysis (TICA) and Implied Timescales for MUC5AC Conformational Data

TICA results for MUC5AC conformational data: The left panel shows the feature value histograms for the first two independent components (IC 1 and IC 2). The middle panel illustrates the K-means clustering results in TICA space, with each orange dot representing a cluster centre. The right panel presents the implied timescales across various lag times, highlighting the slow dynamical processes and confirming the stability of the identified metastable states.

The discretized trajectories were subsequently used to estimate a Markov State Model (MSM) with varying lag times. The implied timescales (ITS) of the MSM were analysed to assess the timescale separation between the metastable states identified in the TICA-transformed space. The results show that the ITS was relatively resolved after ten steps, therefore the selected lag time of 10 (corresponding to 100 picoseconds or 0.1 nanoseconds) was appropriate for capturing the slowest dynamical processes in the system, as evidenced by the convergence of the implied timescales.

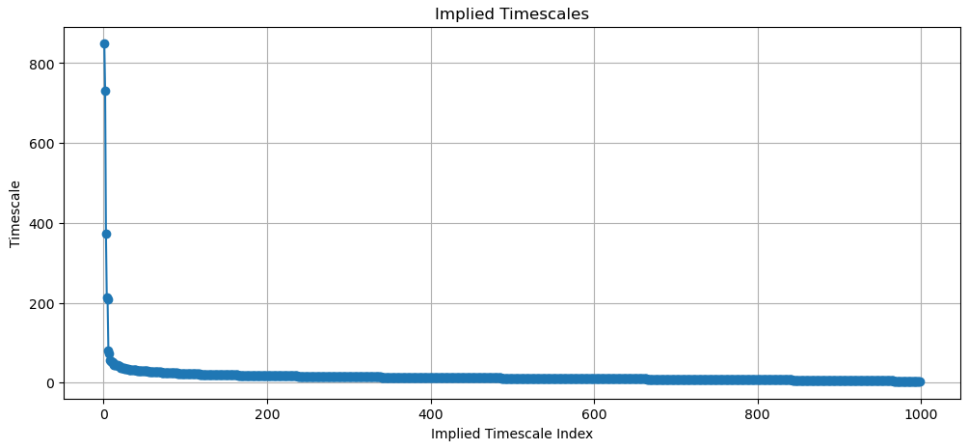


Figure 30: Implied Timescales for MUC5AC

This figure displays the implied timescales for MUC5AC as a function of the implied timescale index. The rapid decay in timescales indicates the presence of a few dominant slow processes, with the remaining timescales corresponding to faster dynamics.

Figure 30, reveals three distinct timescales, each corresponding to a transition between different metastable states. This indicates that the global conformational landscape of MUC5AC can be effectively partitioned into four discrete states. To ensure the reliability of the estimated MSM, a Chapman-Kolmogorov test was conducted as seen in Figure 31. This test compares the transition probabilities predicted by the MSM over multiple lag times (kτ) with those observed directly from the data, thereby verifying the model's assumption of Markovian dynamics. If the model successfully passes this test, it confirms that the underlying dynamics are indeed Markovian, making the MSM a valid representation of the system's kinetics. Following this validation, the metastable states of MUC5AC were identified using Perron Cluster Analysis (PCCA+), which successfully delineated the system into four distinct states.

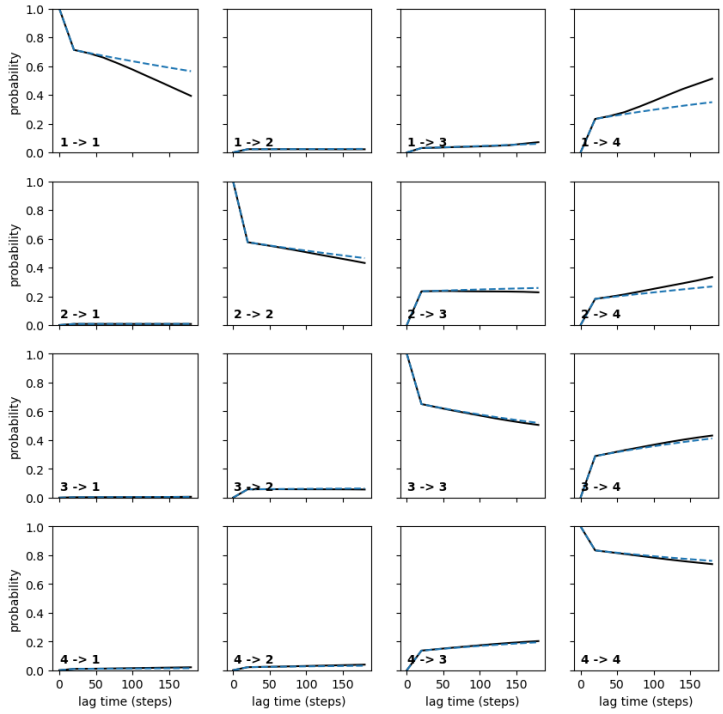


Figure 31: Chapman-Kolmogorov Test for the Four-State MSM

This figure presents the results of the Chapman-Kolmogorov test for the four-state MSM. The test evaluates the consistency of the MSM by comparing predicted transition probabilities over increasing lag times with those observed in the trajectory data. The good agreement between predicted and observed probabilities validates the robustness of the MSM.

Finally, the mean first passage times (MFPT) between the metastable states were calculated to quantify the kinetics of transitions between the different conformational states.

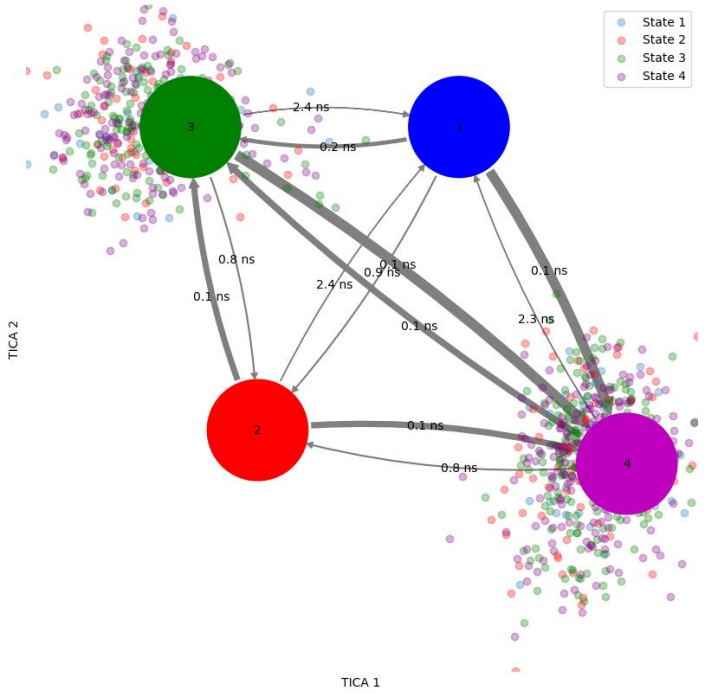


Figure 32: Transition Network of Metastable States in TICA Space

Visualization of the transition network between the four metastable states identified in the MUC5AC system, overlaid on the TICA-transformed space. The arrows indicate the transition probabilities between states, with thicker arrows representing more frequent transitions. The states are color-coded and positioned according to their spatial distribution in the TICA space.

The transition network of metastable states identified in MSM analysis of MUC5AC provides key details into the protein's conformational dynamics, especially in understanding it's role in mucous gels. The visualization in Figure 32 demonstrates the transitions between four distinct states within the TIC space. This network suggests that certain states are more stable and frequently visited, while others represent transitional conformations. As well, the low Mean First Passage Times (MFPT) between these states suggest that there are minimal differences between them, implying that these states may represent highly similar or even identical conformations of MUC5AC.

This finding suggests that the metastable states identified are not significantly distinct from each other, which might reflect the intrinsic flexibility of MUC5AC, allowing it to rapidly switch between conformations without substantial energetic barriers. Such flexibility could be necessary for the protein's function in maintaining the viscoelastic properties of mucus. The

identified states may relate to how MUC5AC adapts to various physiological conditions, such as changes in pH, mechanical stress, or interactions with other molecules, all of which are known to influence mucin structure and function. Understanding these conformational states and the transitions between them could offer new insights into therapeutic targets for modulating mucus properties in disease contexts, particularly in conditions like chronic obstructive pulmonary disease (COPD), where MUC5AC overproduction or misfolding is a significant factor [79, 80].

3.4 Docking Results:

Molecular docking was employed to predict the binding interactions between SINP and the glycoprotein MUC5AC, with the NPs serving as the ligand and MUC5AC as the receptor. The primary objective was to investigate non-covalent binding conformations and binding affinities between these two structures. The docking simulations were carried out using PatchDock, which was executed on the UCD Sonic supercomputer due to the computational intensity of the process.

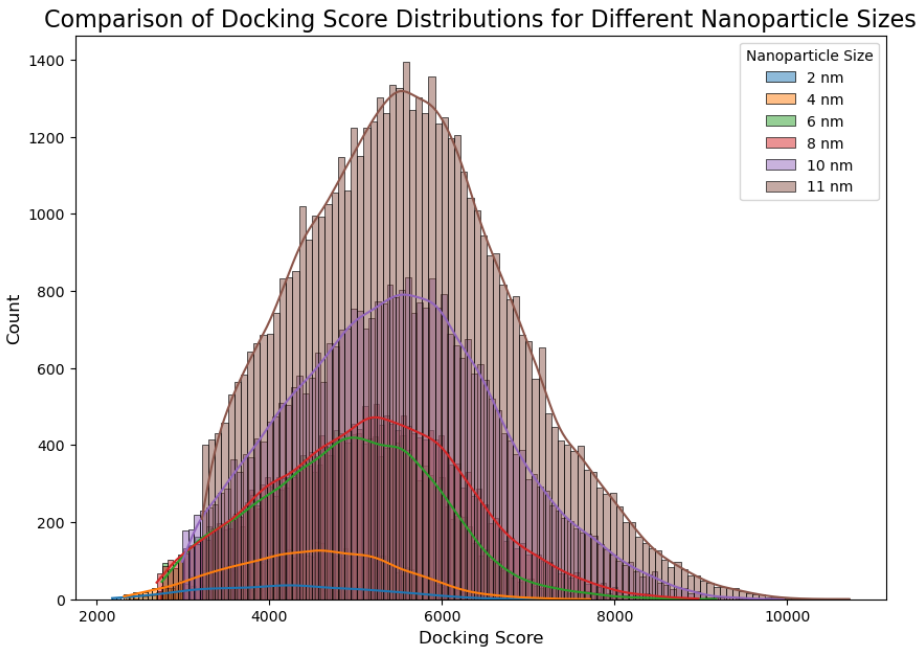


Figure 33: Comparison of Docking Score Distributions for Different Nanoparticle Sizes

This figure shows the distribution of docking scores for various Silicon Dioxide NP sizes (2 nm to 11 nm) interacting with MUC5AC. The histogram visualizes the frequency of docking scores, indicating the variability in interaction quality across different NP sizes.

PatchDock produces a variety of docking configurations, leveraging the geometric complementarity of the interacting molecules. The analysis revealed that most docking configurations resulted in lower-quality scores, with only a small proportion indicating strong and favourable binding interactions.

The results showed that NP size significantly influenced docking performance. The comparison of docking score distributions for various NP sizes, from 2 nm to 11 nm, highlights the variability in binding affinity. Larger NPs generally presented higher docking scores, suggesting stronger interactions, although there was a trade-off in terms of geometric fit and interface area.

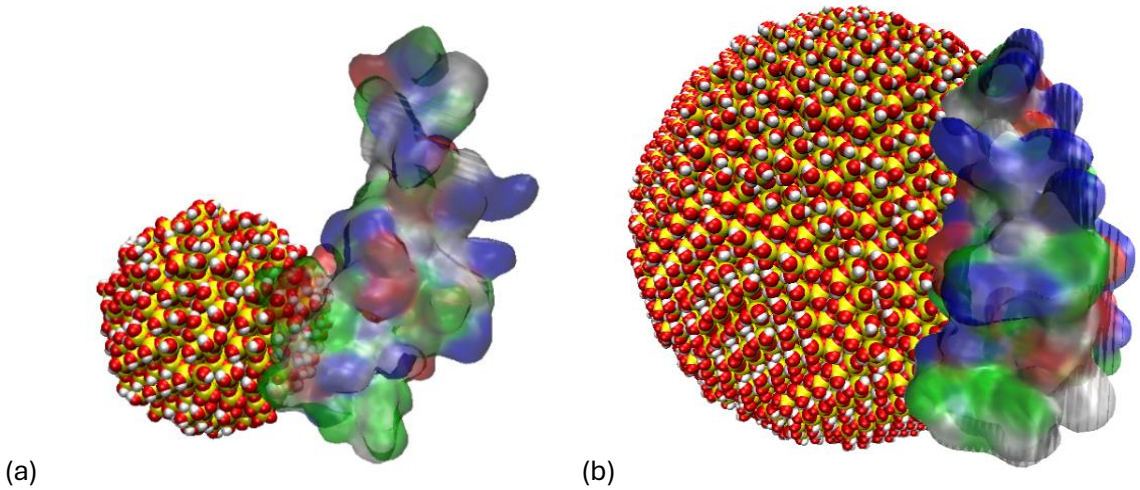


Figure 34: 3D Visualization of MUC5AC Protein Docked with 4 nm and 11nm SINP

A 3D visualization of the docking interaction between MUC5AC protein (right) and (a) 4 nm (b) 11nm SINP (left). The image highlights the complex formed through molecular docking simulations, showcasing the binding conformation.

The images of the docked complexes in Figure 34 provide a visual representation of these interactions, where the MUC5AC protein is seen interacting with different sizes of SiO_2 NPs. Additionally, heatmaps (see Figure 35) were generated to illustrate the spatial distribution of docking scores for the 4nm-MUC5AC simulation, which indicate the areas on the protein surface that are more favourable for NP binding.

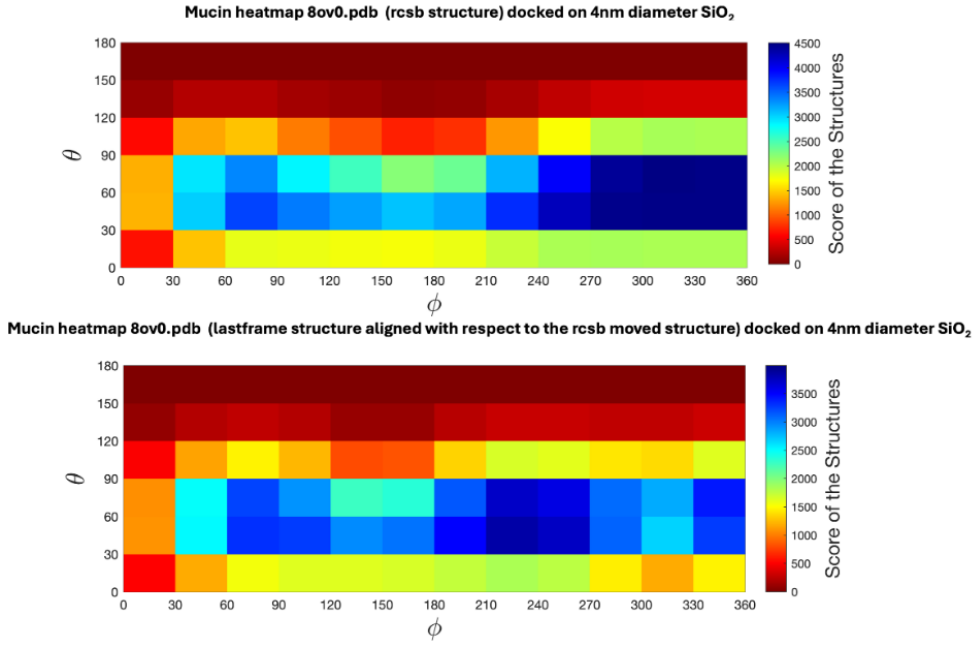


Figure 35: Heatmap of Docking Scores for MUC5AC Protein with 4 nm SINP

Heatmaps illustrating the distribution of docking scores across the surface of MUC5AC when docked with a 4 nm SiNP. The top panel shows the interaction with the RCSB structure, while the bottom panel shows the interaction after aligning the final frame of the simulation. The colour gradient represents the docking score intensity, with higher scores indicating stronger binding regions.

The heatmaps depict the docking scores of the mucin structure on the 4nm SiO_2 NP, with different orientation represented by θ and ϕ angles. The top heatmap shows the original RCSB structure, where higher scores (in blue) are concentrated in specific regions, indicating strong, favourable docking orientations. These clusters suggest that the mucin has well-defined preferred interactions with the NP at certain orientations.

In contrast, the bottom heatmap, which represents the aligned structure from the last frame of the simulation of MUC5AC and NP, still has an evenly distributed pattern of docking scores but with fewer high-scoring regions. This suggests that the aligned altered the interaction landscape, leading to a broader range of possible docking orientations but with generally weaker interactions compared to the original structure.

3.5 Combined Molecular Dynamics Simulation of SiNP-MUC5AC Interactions

MD simulations were conducted to investigate the interactions between MUC5AC protein and SINPs of two different sizes 4 nm and 11 nm. The SINPs were randomly placed near the mucin

protein using CHARMM-GUI multicomponent assembler, and the system was subjected to energy minimization, followed by a 20 ns equilibration, then to a 300 ns production simulation to observe the structural dynamics and stability of the protein in the presence of the SiO_2 NPs. The key parameters evaluated were the RMSD, RMSF and SASA.

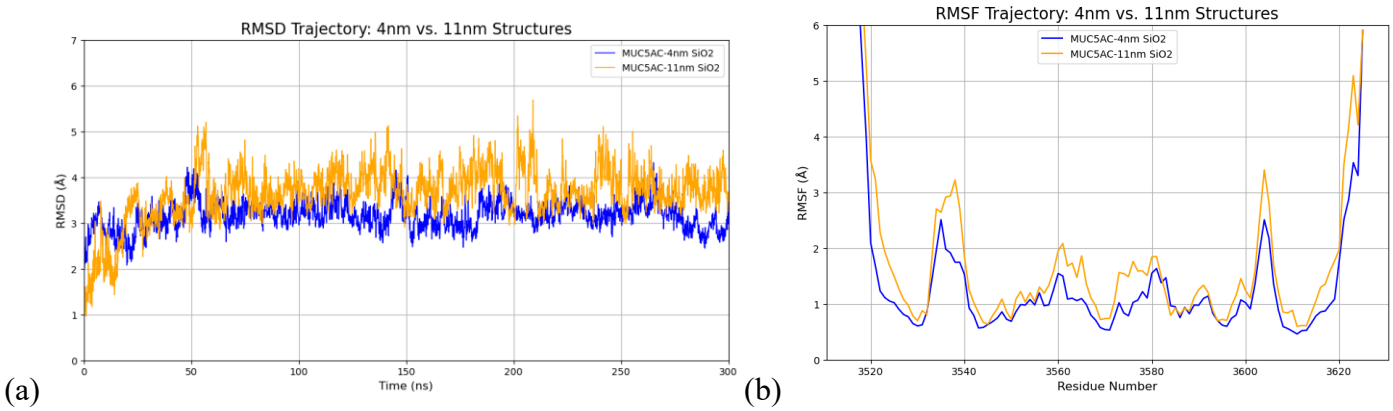


Figure 36 Comparative Analysis of RMSD and RMSF for MUC5AC in 4 nm and 11 nm Systems

(a) RMSD Trajectory: Comparison of Root Mean Square Deviation (RMSD) over time for MUC5AC in 4 nm and 11 nm systems. The RMSD values indicate the deviations of the MUC5AC structure from its initial configuration, highlighting system stability.

(b) RMSF Trajectory: Root Mean Square Fluctuation (RMSF) across MUC5AC residues in 4 nm and 11 nm systems collected across simulation. The graph displays local fluctuations per residue, illustrating regions of flexibility or rigidity within the molecule.

As illustrated in Figure 36 (a) and (b) the RMSD and RMSF were used to characterize the stability of MUC5AC in the presence of SINPs. The RMSD quantifies the average deviation of the MUC5AC backbone atoms from their initial positions throughout the simulation. This meteric is inversely correlated with the system's overall stability. Higher RMSD values suggest greater deviations from the initial configuration, indicating less stability. Initially both systems exhibit higher RMSD values, reflecting the equilibration process as the protein adapts to its environment. As the simulation progresses, these values stabilize between 2 and 5 Å, indicating that both systems reach a steady equilirium without significant ongoing conformational changes.

Throughout the simulation, the RMSD values for the 4 nm and 11 nm systems are closely aligned, though the 11 nm system occasionally shows slightly higher peaks. This could suggest

that the larger system size allows for more conformational flexibility or is influenced by interactions with the surrounding environment or boundary conditions. Despite these fluctuations, the overall similarity in RMSD profiles indicate that the structural integrity of MUC5AC is robust across different system sizes.

Conversely, RMSF assesses the flexibility of individual amino acid residues along the MUC5AC chain, highlighting regions with significant structural fluctuations [81]. The RMSF values from Figure 36 (b) show some sharp peaks, particularly at the ends of the residue range, suggest regions with high flexibility. This could be indicative of terminal residues that often exhibit more movement than those tightly packed with the protein core. The fluctuations and correlations between 4 nm and 11 nm systems, particularly in the middle range of resiudes, might suggest similar stability or conformational behaviors influenced by the surrounding enivroment or interaciton with NPs.

Based on the RMSF data, resiudes that exhibit high flexibility might include turn, coil and bend conformations, which are susceptible to disturbances. These flexible regions may influence how MUC5Ac folds and interacts with external molecules like NPs. The interaction dynamics, such as non-covalent linkages at specific residues could explain variations in flexibility and stability observed in RMSF data. For instance, regions where silica forms hydrogen bonds or where it could generate salt bridges could exhibit decreased flexibility due to increased rigidity from these interactions.

The SASA plot in Figure 37 shows how the MUC5AC protein's surface area exposed to solvent changes over time in both the 4 nm and 11 nm systems. The SASA values for both systems fluctuate throughout the 300 ns simulation, indicating dynamic changes in the protein's interaction with its solvent enviroment. Notably, the SASA for the 4 nm system is more variable, particularly during the first 100 ns, suggesting that the protein undergoes

conformational changes in this period, potentially exposing or hiding different parts of its surface. In contrast, the 11 nm system displays a more stable SASA profile after the initial phase, which could indicate that the protein has reached a more stable conformation with less fluctuation in surface exposure. The generally higher SASA values in the 11 nm system might suggest a slighly more open or unfolded conformation, exposing more the protein's surface to the solvent.

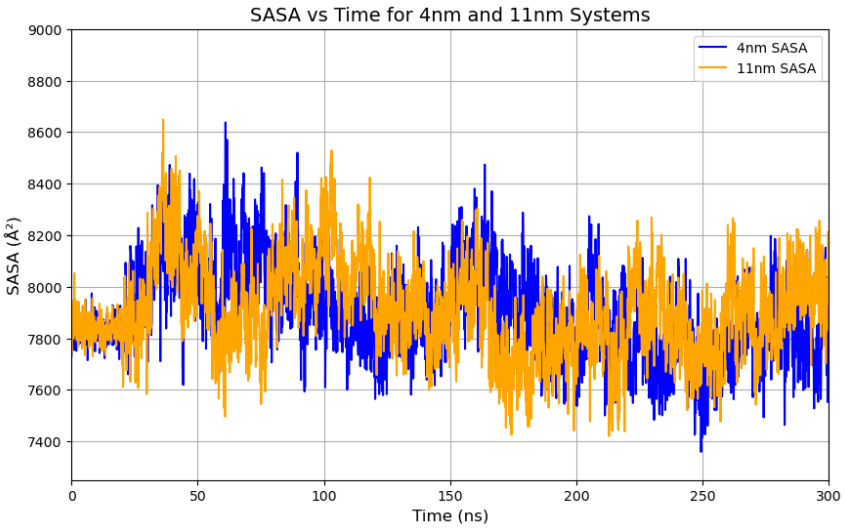


Figure 37: SASA vs. Time for 4 nm and 11 nm MUC5AC-NP Systems

The plot illustrates the Solvent Accessible Surface Area (SASA) of MUC5Ac over 300 ns for both 4 nm and 11 nm systems, highlighting dynamic changes in the protein's surface exposure to the solvent environment.

These SASA fluctuation reflect dynamic shifts between more compact (lower SASA) and more open (higher SASA) conformations, which correlate with the protein's hydrophillic and hydrophobic interations. The higher variability in SASA for the 4 nm system suggests that the protein may be more sensitve to enviromental factors in this smaller system, possibly due to stronger interactions with the boundary or other simulation components. Conversely, the relatively higher and more stable SASA in the 11 nm system implies a more expanded and potentially more stable conformation. This behavior indicates that the protein's hydrophilic and hydrophobic properties influence its behaviour differently in theese two nanoscale environments, potentially affecting its interactions with other molecules or its function in biological systems.

4. Conclusions

This project sought to study the structural dynamics and interactions of MUC5AC mucin, a key component of mucus, within various simulated environments. MUC5AC was selected due to its essential role in forming mucosal gels, which serve to protect and lubricate the epithelial surfaces in the respiratory and gastrointestinal tracts. Understanding MUC5AC's behaviour in the presence of environmental factors like NaCl ions and silica nanoparticles (SiNPs) is vital for uncovering the mechanisms behind its biological functions and responses to external stimuli. SiNPs were chosen due to their widespread use in various industries and their potential for both beneficial and adverse interactions with biological systems. The study was motivated by the growing interest in nano-bio interactions, particularly in how NPs influence protein structures and functions, which is important for advancing drug delivery systems and assessing the safety of nanomaterials in biomedical and environmental applications.

In the simulations of MUC5AC in water with NaCl ions, key conformational characteristics were observed, offering insights into the mucin's structural dynamics. Global variables such as RMSD, Radius of Gyration (Rg), SASA, ETED, hydrogen bonds, and surface charge were tracked during the 1.8 μs production phase, revealing the stability, flexibility, and interaction potential of mucin in an ionic environment.

The RMSD results indicated that while MUC5AC maintains a relatively stable core structure, there are notable fluctuations in the tails and special helical regions suggesting dynamic behaviour in these areas (as shown in Figure 19). The Rg analysis further supported this, with the protein exhibiting periodic expansions and contractions, reflecting its flexible nature (Figure 20). The SASA results show that MUC5AC has a stable exposure to the solvent, with the hydrophobic regions being particularly consistent across different probe size, which highlights the protein’s balanced interaction with its environment (Figure 24 - Figure 25). The ETED measurements revealed a dynamic range of the mucin’s conformations, while the

HBOND analysis suggested that the protein's structural integrity is maintained through consistent hydrogen bonding with the surrounding solvent molecules (Figure 23). Additionally, the surface charge analysis showed a predominance of positively charged regions, which could influence the protein's electrostatic interaction with other molecules, including NPs (as seen in the SASA charge distribution analysis).

The findings on MUC5AC's conformational dynamics in an aqueous ionic environment provide a foundation for future studies on its interactions with NPs. The stability of its core structure and flexibility of its peripheral regions likely plays key roles in its biological functions, particularly in mucosal gel formation and maintenance.

The analysis of MUC5AC's secondary structures across the MD trajectories revealed a varied range of conformations, with a predominant presence of coil configurations indicating the protein's inherent flexibility. This flexibility is essential for MUC5AC's role in mucus formation, allowing it to adapt and maintain the viscoelastic properties necessary for its protection functions. Specific regions exhibited a higher content of β-strands, likely contributing to the structural integrity of the mucus gel, while α-helices were concentrated in the central regions, supporting the protein's elongated structure essential for forming the mucus matrix. These findings demonstrate the balance between structural stability and adaptability in MUC5AC, highlighting its complex functional role in the mucus layer.

In the Markov State Modelling (MSM) analysis, MUC5AC's conformational dynamics were investigated, focusing on backbone torsion angles (ϕ and ψ dihedral angles) to capture key kinetic features. These angles offered a more precise representation of the protein's conformational space compared to other metrics. The high-dimensional data was initially reduced using Principal Component Analysis (PCA), Time-lagged Independent Component Analysis (TICA), and the Variational Approach for Markov Processes (VAMP), simplifying

the analysis and identifying kinetically meaningful states. K-means clustering then discretized the conformational space into key states, forming the basis for constructing the MSM.

The implied timescale analysis revealed three distinct timescales, corresponding to transitions between different metastable states, confirming that MUC5AC's conformational landscape could be partitioned into four discrete states. The MSM was validated using the Chapman-Kolmogorov test, affirming its accuracy in representing the protein's kinetics. The transition network analysis showed that the low Mean First Passage Times (MFPT) suggest minimal differences between these states, indicating that they may represent similar or even identical conformations, highlighting MUC5AC's inherent flexibility. This flexibility likely plays a fundamental role in maintaining mucus's viscoelastic properties, enabling MUC5AC to adapt to various physiological conditions. These MSM results also provide a foundation for future studies, particularly in drug discovery, where the identified metastable states could be targeted for testing small drug molecules. Extending simulation lengths could uncover additional conformations, enhancing our understanding of MUC5AC's full conformational landscape.

Molecular docking simulations predicted the interactions between MUC5AC and silicon dioxide NPs (SiNPs), focusing on identifying non-covalent binding conformations and affinities. Using PatchDock, a variety of docking configurations were generated, revealing that NP size significantly influenced binding interactions with MUC5AC. Larger NPs generally exhibited higher docking scores, suggesting stronger interactions, though smaller NPs showed more variability in their docking performance. Visualizations highlighted the binding conformations for 4 nm and 11 nm SiNPs with MUC5AC, showing how NP size affects the binding interface. Heatmaps of docking scores for the 4 nm SiNP-MUC5AC interaction further illustrated the distribution of docking preferences across the protein surface, revealing regions with stronger or weaker binding interactions.

Building upon the docking results, molecular dynamics (MD) simulations examined the structural dynamics of MUC5AC in the presence of SiNPs of two distinct sizes—4 nm and 11 nm—in an aqueous NaCl environment. These simulations were crucial for understanding the stability and conformational flexibility of MUC5AC when interacting with NPs. Key metrics such as RMSD, RMSF, and SASA assessed the stability and structural integrity of the protein. The RMSD analysis indicated that both NP sizes influenced MUC5AC's stability, with the protein structure stabilizing after initial fluctuations. RMSF data revealed regions of higher flexibility in the presence of the larger 11 nm SiNPs, suggesting NP size affects the protein's conformational behaviour. SASA analysis showed dynamic changes in the protein's surface exposure to the solvent, with the 11 nm system displaying a more stable and slightly more expanded conformation. These findings highlight the complex interplay between NP size and protein structure, providing valuable insights for future studies on mucin-NP interactions in biological systems.

Future studies could explore longer simulation times to capture additional conformational states and transitions that were not observed in this study. The MSM results provide a foundation for testing small drug molecules against identified metastable states, offering potential pathways for therapeutic intervention. Additionally, investigating protein-protein interactions and how these influence the behaviour of MUC5AC in the presence of NPs would further enhance the understanding of its role in various biological processes and its response to environmental factors.

5. References

1. Abdelhamid, H.N. and G. Badr, *Nanobiotechnology as a platform for the diagnosis of COVID-19: a review.* Nanotechnology for Environmental Engineering, 2021. **6**(1): p. 19.
2. Murugadoss, S., et al., *Toxicology of silica nanoparticles: an update.* Arch Toxicol, 2017. **91**(9): p. 2967-3010.
3. Mahmoud, A.M., et al., *Mesoporous Silica Nanoparticles Trigger Liver and Kidney Injury and Fibrosis Via Altering TLR4/NF-κB, JAK2/STAT3 and Nrf2/HO-1 Signaling in Rats.* Biomolecules, 2019. **9**(10): p. 528.
4. Lozano, O., et al. *Amorphous SiO2 nanoparticles promote cardiac dysfunction via the opening of the mitochondrial permeability transition pore in rat heart and human cardiomyocytes*. Particle and fibre toxicology, 2020. **17**, 15 DOI: 10.1186/s12989-020-00346-2.
5. Låg, M., et al., *Silica Nanoparticle-induced Cytokine Responses in BEAS-2B and HBEC3-KT Cells: Significance of Particle Size and Signalling Pathways in Different Lung Cell Cultures.* Basic & Clinical Pharmacology & Toxicology, 2018. **122**(6): p. 620-632.
6. Petrache Voicu, S.N., et al., *Silica Nanoparticles Induce Oxidative Stress and Autophagy but Not Apoptosis in the MRC-5 Cell Line.* Int J Mol Sci, 2015. **16**(12): p. 29398-416.
7. Refsnes, M., et al., *Concentration-dependent cytokine responses of silica nanoparticles and role of ROS in human lung epithelial cells.* Basic & Clinical Pharmacology & Toxicology, 2019. **125**(3): p. 304-314.
8. Kearns, F., M. Rosenfeld, and R. Amaro, *Breaking Down the Bottlebrush: Atomically-Detailed Structural Dynamics of Mucins*. 2024.
9. Alberts, B., *Molecular biology of the cell*. Seventh edition. ed. 2022, New York: W. W. Norton & Company. pages cm.
10. Nelson, D.L., M.M. Cox, and A.A. Hoskins, *Lehninger principles of biochemistry*. Eighth edition. ed. 2021, Austin: Macmillan Learning. 1 volume (various pagings).
11. Delfi, M., et al., *Self-assembled peptide and protein nanostructures for anti-cancer therapy: Targeted delivery, stimuli-responsive devices and immunotherapy.* Nano Today, 2021. **38**.
12. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.
13. Schrödinger, L.L.C., *The PyMOL Molecular Graphics System, Version 1.8*. 2015.
14. Jorgensen, W.L., *The many roles of computation in drug discovery.* Science, 2004. **303**(5665): p. 1813-8.
15. Wagner, C., K. Wheeler, and K. Ribbeck, *Mucins and Their Role in Shaping the Functions of Mucus Barriers.* Annual Review of Cell and Developmental Biology, 2018. **34**: p. 189-215.
16. Ridley, C., et al., *Biosynthesis of the polymeric gel-forming mucin MUC5B.* Am J Physiol Lung Cell Mol Physiol, 2016. **310**(10): p. L993-L1002.
17. Thornton, D.J., K. Rousseau, and M.A. McGuckin, *Structure and function of the polymeric mucins in airways mucus.* Annu Rev Physiol, 2008. **70**: p. 459-86.
18. García-Díaz, M., et al., *The role of mucus as an invisible cloak to transepithelial drug delivery by nanoparticles.* Adv Drug Deliv Rev, 2018. **124**: p. 107-124.
19. Schattling, P., et al., *A Polymer Chemistry Point of View on Mucoadhesion and Mucopenetration.* Macromol Biosci, 2017. **17**(9).
20. Huckaby, J.T. and S.K. Lai, *PEGylation for enhancing nanoparticle diffusion in mucus.* Adv Drug Deliv Rev, 2018. **124**: p. 125-139.
21. Wagner, C.E., K.M. Wheeler, and K. Ribbeck, *Mucins and Their Role in Shaping the Functions of Mucus Barriers.* Annu Rev Cell Dev Biol, 2018. **34**: p. 189-215.
22. Witten, J., T. Samad, and K. Ribbeck, *Selective permeability of mucus barriers.* Curr Opin Biotechnol, 2018. **52**: p. 124-133.
23. Mollazadeh, S., et al., *Nano drug delivery systems: Molecular dynamic simulation.* Journal of Molecular Liquids, 2021. **332**: p. 115823.
24. Pai, R.V., J.D. Monpara, and P.R. Vavia, *Exploring molecular dynamics simulation to predict binding with ocular mucin: An in silico approach for screening mucoadhesive materials for ocular retentive delivery systems.* J Control Release, 2019. **309**: p. 190-202.
25. Liu, H., et al., *Understanding Functional Group and Assembly Dynamics in Temperature Responsive Systems Leads to Design Principles for Enzyme Responsive Assemblies.* Nanoscale, 2021.
26. Gupta, K.M., S. Das, and P. Chow, *Molecular Dynamics Simulations to Elucidate Translocation and Permeation of Active from Lipid Nanoparticle to Skin: Complemented with Experiments.* Nanoscale, 2021. **13**.
27. Yu, M., et al., *Rapid transport of deformation-tuned nanoparticles across biological hydrogels and cellular barriers.* Nat Commun, 2018. **9**(1): p. 2607.

28. Bao, C., et al., *Enhanced Transport of Shape and Rigidity-Tuned α-Lactalbumin Nanotubes across Intestinal Mucus and Cellular Barriers.* Nano Lett, 2020. **20**(2): p. 1352-1361.
29. Yu, M., et al., *Temperature- and rigidity-mediated rapid transport of lipid nanovesicles in hydrogels.* Proc Natl Acad Sci U S A, 2019. **116**(12): p. 5362-5369.
30. Khmelnitsky, L., et al., *Diversity of CysD domains in gel-forming mucins.* Febs j, 2023. **290**(21): p. 5196-5203.
31. Schneidman-Duhovny, D., et al., *PatchDock and SymmDock: servers for rigid and symmetric docking.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W363-7.
32. Katchalski-Katzir, E., et al., *Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.* Proc Natl Acad Sci U S A, 1992. **89**(6): p. 2195-9.
33. Vakser, I.A., *Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex.* Proteins, 1997. **Suppl 1**: p. 226-30.
34. Norel, R., et al., *Examination of shape complementarity in docking of unbound proteins.* Proteins, 1999. **36**(3): p. 307-17.
35. Mohanty, D., et al., *Kinetics of peptide folding: computer simulations of SYPFDV and peptide variants in water.* J Mol Biol, 1997. **272**(3): p. 423-42.
36. Saiz, E. and M.P. Tarazona, *Molecular Dynamics and the Water Molecule: An Introduction to Molecular Dynamics for Physical Chemistry Students.* Journal of Chemical Education, 1997. **74**(11): p. 1350.
37. Ollitrault, P.J., A. Miessen, and I. Tavernelli, *Molecular Quantum Dynamics: A Quantum Computing Perspective.* Accounts of Chemical Research, 2021. **54**(23): p. 4229-4238.
38. Ponder, J.W. and D.A. Case, *Force fields for protein simulations.* Adv Protein Chem, 2003. **66**: p. 27-85.
39. Brooks, B.R., et al., *CHARMM: the biomolecular simulation program.* J Comput Chem, 2009. **30**(10): p. 1545-614.
40. Schmid, N., et al., *Definition and testing of the GROMOS force-field versions 54A7 and 54B7.* Eur Biophys J, 2011. **40**(7): p. 843-56.
41. Jorgensen, W.L., D.S. Maxwell, and J. Tirado-Rives, *Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids.* Journal of the American Chemical Society, 1996. **118**(45): p. 11225-11236.
42. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* J Comput Chem, 2005. **26**(16): p. 1781-802.
43. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
44. Rahman, A., *Correlations in the Motion of Atoms in Liquid Argon.* Physical Review, 1964. **136**(2A): p. A405-A411.
45. Yao, Z., et al., *Improved neighbor list algorithm in molecular simulations using cell decomposition and data sorting method.* Computer Physics Communications, 2004. **161**: p. 27-35.
46. Ohno, K., T. Nitta, and H. Nakai, *SPH-based Fluid Simulation on GPU Using Verlet List and Subdivided Cell-Linked List.* 2017 Fifth International Symposium on Computing and Networking (CANDAR), 2017: p. 132-138.
47. Brasiello, A., et al., *Molecular dynamics of triglycerides: atomistic and coarse-grained approaches.* 2006.
48. Garman, E.F., *Developments in x-ray crystallographic structure determination of biological macromolecules.* Science, 2014. **343**(6175): p. 1102-8.
49. Hu, Y., et al., *NMR-Based Methods for Protein Analysis.* Analytical Chemistry, 2021. **93**(4): p. 1866-1879.
50. Binshtein, E. and M.D. Ohi, *Cryo-electron microscopy and the amazing race to atomic resolution.* Biochemistry, 2015. **54**(20): p. 3133-41.
51. Allen, M.P. and D.J. Tildesley, *Computer simulation of liquids*. Second edition. ed. 2017, Oxford, United Kingdom: Oxford University Press. xiv, 626 pages.
52. Schlick, T., *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Vol. 21. 2010.
53. Frenkel, D. and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*. 2001. p. 664.
54. Haile, J.M., *Molecular dynamics simulation: elementary methods*. 1992: John Wiley & Sons, Inc.
55. Choi, Y.K., et al., *CHARMM-GUI Nanomaterial Modeler for Modeling and Simulation of Nanomaterial Systems.* Journal of Chemical Theory and Computation, 2022. **18**(1): p. 479-493.
56. Kern, N.R., et al., *CHARMM-GUI Multicomponent Assembler for modeling and simulation of complex multicomponent systems.* Nature Communications, 2024. **15**(1): p. 5459.
57. Drozdov, A.N., A. Grossfield, and R.V. Pappu, *Role of solvent in determining conformational preferences of alanine dipeptide in water.* J Am Chem Soc, 2004. **126**(8): p. 2574-81.

58. Lee, B. and F.M. Richards, *The interpretation of protein structures: estimation of static accessibility.* J Mol Biol, 1971. **55**(3): p. 379-400.
59. Krack, M. and K. Jug, *Molecular electrostatic potentials for large systems*, in *Theoretical and Computational Chemistry*, J.S. Murray and K. Sen, Editors. 1996, Elsevier. p. 297-331.
60. Shumilina, A. *A Fast Method for Determination of Solvent-Exposed Atoms and Its Possible Applications for Implicit Solvent Models*. in *Computational Science and Its Applications – ICCSA 2005*. 2005. Berlin, Heidelberg: Springer Berlin Heidelberg.
61. Shrake, A. and J.A. Rupley, *Environment and exposure to solvent of protein atoms. Lysozyme and insulin.* J Mol Biol, 1973. **79**(2): p. 351-71.
62. Klenin, K.V., et al., *Derivatives of molecular surface area and volume: simple and exact analytical formulas.* J Comput Chem, 2011. **32**(12): p. 2647-53.
63. Dorosh, L. and M. Stepanova, *Probing Oligomerization of Amyloid Beta Peptide in Silico.* Mol. BioSyst., 2016. **13**.
64. Fraczkiewicz, R. and W. Braun, *Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules.* Journal of Computational Chemistry, 1998. **19**(3): p. 319-333.
65. Daberdaku, S. and C. Ferrari, *Computing voxelised representations of macromolecular surfaces:A parallel approach.* The International Journal of High Performance Computing Applications, 2018. **32**(3): p. 407-432.
66. Mitternacht, S., *FreeSASA: An open source C library for solvent accessible surface area calculations.* F1000Res, 2016. **5**: p. 189.
67. Jurrus, E., et al., *Improvements to the APBS biomolecular solvation software suite.* Protein Sci, 2018. **27**(1): p. 112-128.
68. Harris, E.S., et al., *Reduced sialylation of airway mucin impairs mucus transport by altering the biophysical properties of mucin.* Scientific Reports, 2024. **14**(1): p. 16568.
69. Tsodikov, O.V., M.T. Record, Jr., and Y.V. Sergeev, *Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature.* J Comput Chem, 2002. **23**(6): p. 600-9.
70. Bowman, G., V. Pande, and F. Noé, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Vol. 797. 2014.
71. Buchete, N.V. and G. Hummer, *Coarse master equations for peptide folding dynamics.* J Phys Chem B, 2008. **112**(19): p. 6057-69.
72. Garrido, L., *Systems far from equilibrium / Sitges Conference on Statistical Mechanics, June 1980, Sitges, Barcelona/Spain ; edited by L. Garrido*. Lecture notes in physics ; 132. 1980, Berlin ;: Springer-Verlag.
73. Scherer, M.K., et al., *PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models.* Journal of Chemical Theory and Computation, 2015. **11**(11): p. 5525-5542.
74. Abrami, M., et al., *Mucus Structure, Viscoelastic Properties, and Composition in Chronic Respiratory Diseases.* International Journal of Molecular Sciences, 2024. **25**: p. 1933.
75. Bansil, R. and B. Turner, *Mucin structure, aggregation, physiological functions and biomedical applications.* Current Opinion in Colloid & Interface Science, 2006. **11**: p. 164-170.
76. Kirkham, S., et al., *Heterogeneity of airways mucus: variations in the amounts and glycoforms of the major oligomeric mucins MUC5AC and MUC5B.* Biochemical Journal, 2002. **361**(3): p. 537-546.
77. Pérez-Hernández, G., et al., *Identification of slow molecular order parameters for Markov model construction.* The Journal of chemical physics, 2013. **139**: p. 015102.
78. Jain, A.K., *Data clustering: 50 years beyond K-means.* Pattern Recognition Letters, 2010. **31**(8): p. 651-666.
79. Li, J. and Z. Ye, *The Potential Role and Regulatory Mechanisms of MUC5AC in Chronic Obstructive Pulmonary Disease.* Molecules, 2020. **25**(19): p. 4437.
80. Warfield, B.M. and P.C. Anderson, *Molecular simulations and Markov state modeling reveal the structural diversity and dynamics of a theophylline-binding RNA aptamer in its unbound state.* PLOS ONE, 2017. **12**(4): p. e0176229.
81. Gowers, R.J., et al. *MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations*. 2019. United States.