Protein multiscale modelling workflows – from formulation development to protein engineering

Lanyu Fan

May 2025

Abstract

The activity of any given protein, whether in cell or *in vitro*, relies on a complex network of interactions among this protein and other molecules, such as other proteins, water, and small molecular cosolvents/excipients. These interactions occur at different spatial and temporal scales, spanning about 10 orders of magnitude in the space and 15 orders of magnitude in the time domain. As such, many different modelling techniques, each suitable for a particular spatiotemporal regime, are routinely used. However, a single process often spans more than a single time or space scale. Thus, the necessity arises for combining different modelling and simulation techniques in multiscale workflows.

In this work, structurally and functionally diverse proteins such as immunoglobulins, transcription factors, autophagy receptors and non-LTR retrotransposon, were treated using several different structure-based methods (multiscale molecular dynamics simulation, molecular docking, and cosolvent-based approaches) in order to study their stability and propensity to aggregate, misfold and small molecule allosteric binding. The central objective of this work was to evaluate applicability of those structure-based computational workflows in studies of protein "druggability" and development of formulations of protein-based therapeutics.

First, effects of cosolvents/excipients on stability and formulation of protein therapeutics have been studied. In this part of my dissertation, multiscale (all-atom and coarse grain) simulations of monoclonal antibodies (mAbs) were carried out in different concentrations of excipients such as amino acids, salt, and sugars to check the stabilizing conditions in different formulations. Another aim of this work was to assess whether the Fab fragment may be used as a representative for the full-length mAb protein in computational studies.

Next part of my work was to assess the applicability of cosolvent-based workflows, such as solvent-mapping and cosolvent MD simulations in finding the stabilizing "hotspots" that contribute to allosteric regulation of estrogen receptors (ER) by small molecules. Herein, solvent mapping was used to investigate the protein-protein interaction and ESR allosteric activation by methylimidazolium ionic liquids (MILs). This part of work has been carried out in collaboration liver toxicology group at Newcastle University, supporting their experimental findings on MILs acting as endocrine disruptors.

Finally, I have addressed the assembly of two regulatory coiled-coil proteins, namely LINE1 ORF1p and NDP52, which are regulated by redox environment, post-translational modifications, protein-protein interactions, and cosolvents such as divalent cations. This part of work included modelling of full-length proteins and studies of protein-protein interactions in different environments. The assembly model of NDP52, which supports experimental data has been proposed. The model of Orf1p regulation, recapitulating its modulation by site-specific phosphorylation, copper and Pin1 interactions has been developed. In addition, computational site-directed mutagenesis has been applied in order to validate the models.

Acknowledgment

In the past three years, I enjoyed work with my supervisors Dr Agnieszka Bronowska and Prof Jarka Glassey, and my colleagues, Dr Shangze Xu and Dr João Victor de Souza. They have taught me a lot during my study. This thesis cannot be completed without their support and encouragement. I would also like to say thank you to Prof Viktor Korolchuk, Prof Matthew Wright, and Dr Ruchi Shukla for their experimental support.

My family has been very supportive during my research and writing up period, thanks to their encouragement when I lost confidence.

This work received funding from EPSRC with grant number EPR51309X1.

Declaration

The work described in this thesis was carried out between September 2019 and June 2023 in the Computational Medicinal Chemistry Laboratories (School of Natural and Environmental Sciences, Bedson Building, Newcastle University, Newcastle Upon Tyne, NE1 7RU), the School of Engineering, Newcastle University, and Cancer Research Laboratories, Paul O'Gorman Building, Northern Institute for Cancer Research, Newcastle University, and Ageing Research Laboratories, Biosciences Institute, Newcastle University. The research was conducted in collaboration with Liver Toxicology Laboratories, Faculty of Medical Sciences, Newcastle University.

All of the research described in this thesis is original in context and does not incorporate material or ideas previously published or presented by other authors except where due reference is given in the text.

No part of this these has been previously submitted for a degree, diploma or any other qualification at any other University.

Table of Contents

List of Figures

List of Abbreviations

5-HT	5-hydroxytryptamine
AhR	aryl hydrocarbon receptors
APC	antigen presenting cell
B_{22}	second virial coefficient
CDR	complementary determining region
CG	Coarse-grained
Cryo-Em	Cryo-electron microscopy
CTD	C-terminal domain
DBD	DNA-binding domain
ER	estrogen receptor
ERE	estrogen response element
Fab	antigen binding fragment
Fc	crystallizable fragment
FF	force fields
FR	framework region
GdmCl	guanidinium chloride
GPCR	G-protein coupled receptor
HSCs	hematopoietic stem cells
IgG	Immunoglobulin G
ILs	Ionic liquids
LBD	ligand binding domain
LGIC	ligand-gated ion channels
LINE-1	long interspersed nuclear element 1
LIR	LC3-interation region
mAb	monoclonal antibody
MD	molecular dynamics
MHC	major histocompatibility complex

MM	Molecular mechanics
nAChRs	nicotinic acetylcholine receptors
NTD	N-terminal domain
PCA	principal component analysis
PPI	protein-protein interactions
QM	Quantum mechanics
RDF	radial distribution function
Rg	radius of gyration
RMSD	root-mean-square deviation
RMSF	root-mean-square fluctuation
RNP	ribonucleoprotein
ROS	reactive oxygen species
RRM	RNA recognition motif
SASA	solvent accessible surface area
SRC	steroid receptor coactivators
TAD	transactivation domain
TCR	T cell receptor

Chapter 1: Introduction

1.1 Proteins as molecular sensors

Starting from 1838, the word protein was created and used ever since.[1] Protein, made from amino acid residues coded by genes, plays a vital role in biological processes inside human body such as immune reactions towards pathogens or autophagy, and cell signalling. Also processes inside our cells such as cell replication, transcription and translation also involve proteins.

In 1905, Langley used the phrase 'receptive substance' when investigating the antagonistic effect of curare and nicotine on the muscles and his student, Hill expanded the receptor idea further using a quantitative method.[2, 3] From then on, the receptor theory gradually emerged and accepted by pharmacologists. And receptors become the major research area of pharmacology.

Because of their receptive nature and they are all around body, for example, myoglobin in muscles, haemoglobin in blood, collagen in skin, hair and bones, trypsin in small intestine, antibody in immune system and a lot of other proteins inside cells throughout the body. They can be common molecular sensors in mammals.

The subcellular localization of proteins within cells affects their functions. Most proteins studied in this project are intracellular (cytosolic, nuclear, or shuttling), and antibodies are secreted (extracellular). There are other types of proteins based on their subcellular localisation, for example membrane proteins.

G-protein coupled receptors (GPCRs) are the largest family of membrane protein

in human.[4] They can response to ions, hormones, neurotransmitters and environmental stimuli.[5] There is no sequence similarity in different GPCR families but they have a similar central core containing seven transmembrane helices, extracellular N-terminal and intracellular C-terminal.[6] When ligands bind to the receptor, G protein is activated and exert the corresponding response.

Another example of commonly known membrane protein is ligand-gated ion channels (LGICs) which are integral membrane proteins. They allow the flow of specific ions across plasma membrane by active transport. The channels are gated by neurotransmitters, examples of proteins in the LGIC superfamily are nicotinic acetylcholine receptors (nAChRs), 5-hydroxytryptamine (5-HT) and so on.[7]

Antibodies are extracellular sensors in our immune system. They are glycoproteins secreted by B lymphocytes when antigens entering the body that belong to the immunoglobulin superfamily. They can also activate the complement system to lysis pathogens and aid in the process of phagocytosis.

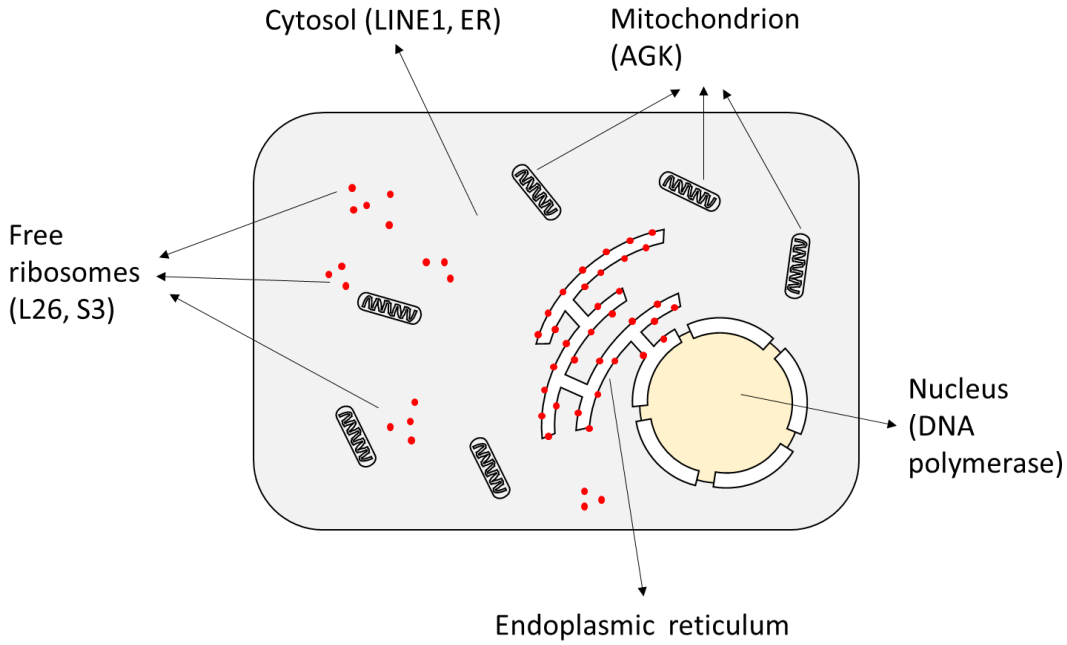


Figure 1: An illustrative diagram of cell containing the nucleus, cytosol, endoplasmic reticulum, ribosomes, and mitochondrion. Examples of proteins located in the nucleus: DNA/RNA polymerase; in the cytosol: LINE1, AhR, ER and NDP52; in the ribosomes: ribosomal protein L26 and S3; in the mitochondrion: Acylglycerol kinase (AGK).

1.2 Protein domains

Proteins are composed of domains and the vast majority of proteins contains at least two domains.[8] The function of proteins is also depending on how the domains interact with each other.

For examples, antibodies contain two antigen binding fragment domains and one crystallizable fragment domain. These domains function together to allow antibodies to recognise antigens.

LINE1 ORF1p contains coiled-coil domain, RNA recognition motif (RRM) domain and C-terminal domain (CTD). The coiled-coil domain is important for homotrimerization and the RRM and CTD are essential for RNA binding.

NDP52 contains SKICH domain, the LC3-interation region (LIR) domain and coiled-coil domain. The SKICH domain is crucial for selective autophagy.

The ER contains a modulating N-terminal domain (NTD), DNA-binding domain (DBD) and C-terminal ligand-binding domain (LBD). The DBD of ER normally binds to the ERE of genes to start transcription. The LBD is responsible for most of the ligand binding related functions.

Most proteins work by interacting with another protein or binding to ligands varied from small organic molecules, amino acids to ions. These interactions/bindings have been studied both experimentally and computationally. The interactions can be covalent such as disulfide bonds or non-covalent like hydrogen bonds and Van der Waals forces.

1.3 Proteins as therapeutics

Proteins are usually regarded as targets for therapeutics (small molecules or peptides). Yet, some proteins can also be therapeutics and exert their activity by interacting with other proteins.

Drug targets are macromolecules that drugs can bind specifically to give therapeutic effects in patients. There are four major macromolecule types that can interact with small molecules: carbohydrates, lipids, proteins, and nucleic acids. Highly effective and selective compounds binding to carbohydrates, lipids and nucleic acids are difficult to obtain which makes proteins the main drug targets for current drugs in the market. And nucleic acids are the second possible targets. Within the protein category, the most common drug targets are receptors such as G-protein coupled

receptors and nuclear hormone receptors, enzymes like protein kinases and ion channels.

Besides drug targets, proteins can also be therapeutic drugs, which is an important type of medicines treating a wide range of diseases including inflammation, cancers, nervous system disorders and so on.

We live in a world that contains pathogenic microbes and toxic substances that can affect our normal homeostasis. Our immune system uses protective mechanisms to eliminate these microbes and substances. The immune system contains the innate immune system and the adaptative immune system.

The innate immune system is the first line of defence to foreign substance, and it gives immediate responses. It involves non-specific defence mechanisms that are encoded by genes in the germ line of hosts, including physical barriers like skin and other epithelial layers, chemical barriers like lysozyme, gastric juice and saliva. Also, there are phagocytic cells that destroy microbes in our bodies. This response cannot recognize the same pathogen in the future.

The adaptative immune system is highly specific for its target antigens. The adaptive responses rely on antigen-specific receptors on the surfaces of specific white blood cells called lymphocytes. They provide destructive responses towards invading microbes and toxic substances, so it is vital that they can recognize what is foreign and not respond to original host molecules. Fail to recognize such molecules and destroy host molecules can lead to autoimmune disease.

There are two main classes of adaptive immune responses, cell-mediated immune

responses and antibody responses which are carried out by two classes of lymphocytes, T lymphocytes and B lymphocytes also called T cells and B cells.

There are three main types of T cells, the helper T cell, the cytotoxic T cell and regulatory T cell. T cells come from hematopoietic stem cells (HSCs) in bone marrow and migrate to thymus for maturation. They have T cell receptor (TCR) on the surface of membrane to recognize antigens. The presence of antigen presenting cell (APC) is also required for T cell to bind to specific antigen. The major histocompatibility complex (MHC) on the surfaces of APCs is involved in the recognition. There are class I and class II MHC molecules, cytotoxic T cell recognize class I bound antigens and helper T cells recognize class II bound antigens. When TCR binds to MHC molecule, the complexes are a bit unstable, so co-receptors are required.

Cytotoxic T cells also known as killer T cells have the co-receptor CD8, they need to be activated to give effector functions. The interaction with APCs leads to an intracellular pathway that stimulate TCRs on the T cell which then bind to class II MHCs and exert immune function by killing infected cells.

Helper T cells have the co-receptor CD4, they exert their effect by directing other cells to carry out their tasks. The activation is similar to cytotoxic T cells except the TCRs bind to class I MHCs. The activated cells release cytokines that influence immune responses from other white blood cells including B cells to produce T cell-dependent antibodies.

Regulatory T cells also have CD4 on the surface and they stop the immune responses when they are no longer needed to inhibit autoimmune processes.

B cells also come from HSCs in the bone marrow and leave the marrow when they are mature. They have unique antibodies on the cell surface to recognize antigens directly, so APCs are not required. When they are activated by antigens, they proliferate and differentiate into memory B cells. These cells release antibodies.

The earliest idea of antibody therapy originated from clinical studies by Behring and Kitasato in the 1890s.[9] They discovered diphtheria and tetanus antitoxin; the antitoxins were proved to be effective and selective in animals. Then large trial started by immunising sheep and horses before treating human patients. Behring also thought it was a safe method to use serum with mixture of diphtheria toxin and antitoxin to provide active immunity in human. But back to that time, anaphylactic reactions were common after giving horse serum and serum sickness always occurred.[10] Soon, Paul Ehrlich proposed the first antibody molecule model and then it took scientists more than half a century to study the mechanism behind the therapy.[11] With the understanding of antibodies, the first atomic structure of an antibody fragment was obtained in 1972.[12]

Although, the immune response to an antigen is always complicated and polyclonal in nature, two years later, the first murine monoclonal antibody (mAb) was generated by Kohler and Milstein in 1975, this signalling the start of novel therapeutic antibodies development.[13] In the 1980s, murine mAbs showed drawbacks in clinical stages: short half-life, weak binding, and poor effector function. To overcome the drawbacks, chimeric mouse-human antibodies using genetic engineering technique were first generated and molecules are around 65% human.[14] Then, to improve properties further, humanized mAbs with molecules are 95% human were generated.[15] In the 20th century, scientists are able to generate fully human mAbs. The properties of humanized and fully human mAbs are similar

to the endogenous IgGs. Nowadays, most of the FDA approved mAbs are humanized or fully human with few chimeric and rare murine mAbs.[10]

In comparison with polyclonal antibodies, mAbs are homogeneous and highly specific which makes them more effective when developing therapies. They can be used to inhibit alloimmune and autoimmune response, antitumour, antiviral and antiplatelet therapies.

Besides mAbs, other therapeutic proteins including recombinant enzymes like asparaginase, fusion proteins like etanercept and bispecific antibodies like Removab.

Therapeutic proteins represent a fast-growing sector in the biopharmaceutical industry over the past 20 years and their stability is crucial for clinical safety and commercial viability. As more and more therapeutic proteins are under development or on the market, the stability improvement will require more investigations in this field.

1.4 Challenges in protein studies

The three-dimensional structure of protein depends on its primary amino acids sequence and how they fold. Alpha helixes and beta sheets are common folded structures. Normally, the native state of the protein represents the most stable conformation with a lowest energy. The folding of protein is constrained by many factors such as covalent bond between neighbouring amino acids, disulphide linkage and weak interactions. Although folding is not well understood. Levinthal's paradox states that it is impossible to find the native conformation of a protein by a random search of the astronomical number of possible conformations in a

meaningful timescale, yet protein folds so rapidly. This suggests protein is not folding randomly, there should be a pathway to follow. To predict the structure of a protein purely from its amino acid sequence is a big challenge due to the number of possible conformations, recent improvements in machine learning like AlphaFold shows progress in theoretical computational modelling of protein.

Both thermodynamics and kinetics contribute to the folding-unfolding process of protein. Protein folding kinetics is quite complex, both lattice and off-lattice models have been used to represent the protein and mimic folding process.[16] Protein needs to be able to fold to a native state following a reasonable folding funnel.[17]

Protein misfolding is an intrinsic propensity of proteins when they folded following a wrong pathway and common in living cells. It is affected by the composition of amino acid and happens continuously. The environmental conditions such as temperature and pH can also affect the folding process.

The misfolding of protein can lead to aggregation and even folded protein can aggregate. Some of native state protein have large hydrophobic patches are more like to aggregate, high concentrations of protein can also increase the chance of protein-protein interaction and initiate aggregation. Although aggregation formation in proteins is a complex process with multiple pathways and involves different mechanisms. In general, it is believed aggregation follows the steps of nucleation, propagation and polymerization.[18] In the Lumry-Eyring Nucleated Polymerization model, a pre-nucleation stage was considered. The protein monomer goes through reversible conformational changes and leads to interactions between monomers, forming large reversible oligomers, then small aggregates start to form due to irreversible conformational rearrangement which then behaves as nuclei allowing more aggregation to occur into larger aggregates.[19] This model

shown that protein aggregation is controlled by colloidal stability and conformational stability.

Example of common aggregated protein such as amyloid conformation. Amyloids are very stable aggregates that contain linear fibrils. More than 20 plasma proteins can form amyloid and cause diseases such as Alzheimer's disease, Type II diabetes and so on.[20]

Colloidal stability involves the intermolecular forces between native state antibodies in solution. A good understanding of aggregation needs information about these intermolecular interactions. To assess the colloidal stability, thermodynamic parameters such as osmotic second virial coefficient (B_{22}) is used.[21, 22] B_{22} quantifies weak protein-protein interactions induced by the appearance of other compositions in the solutions such as excipients at molecular level. The value of B_{22} can be calculated using static light scattering (SLS) and advanced thermodynamics model such as mxDLVO model introduced by Herhut.[23] There is scientific research shown that B_{22} is directly related to protein aggregation. Positive B_{22} values suggest repulsive protein-protein interactions (PPI) are dominant due to favourable protein-solvent interactions and less chance to aggregate. Negative B_{22} values indicate attractive PPI are more favourable than protein-solvent interactions and increases the chance of aggregation.[24, 25]

Conformational stability involves the Gibbs free energy difference between native state antibodies and their denatured states. Differential scanning calorimetry (DSC) and Differential scanning fluorimetry (DSF) are used to evaluate the stability based on relevant thermodynamic parameters such as the Gibbs free energy difference and unfolding temperature.

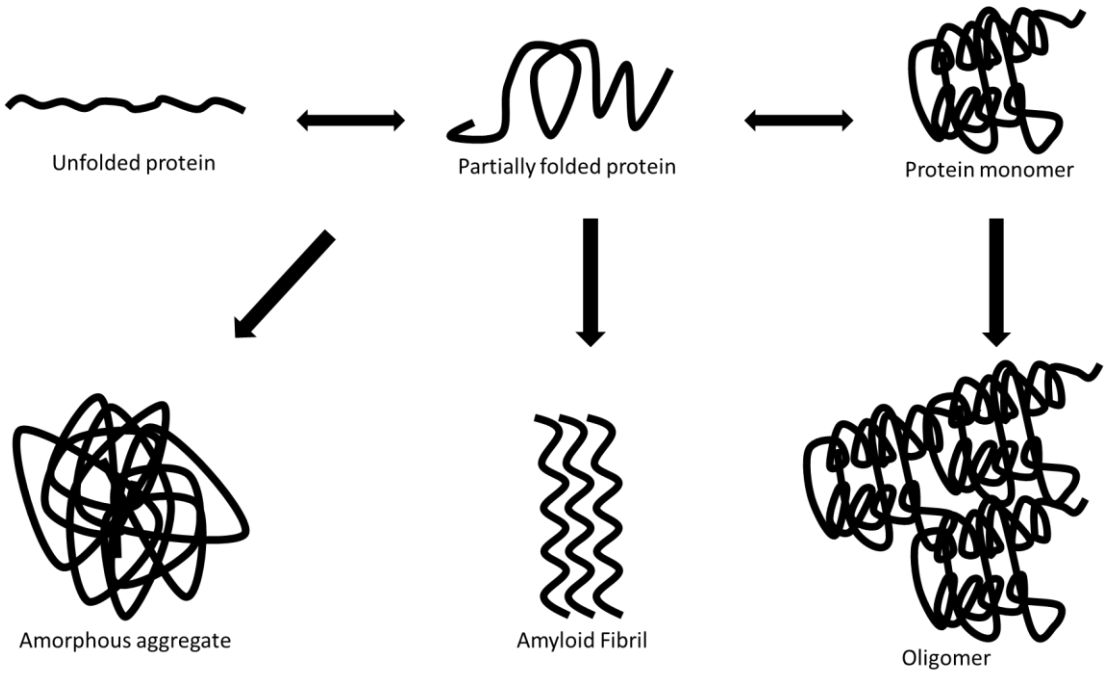


Figure 2: The schematic representation of protein folding and formation of aggregates. In normal process, the unfolded protein forms partially folded protein, then protein monomer and the whole process is reversible. Oligomer can be formed from protein monomer and this process is irreversible. From the partially folded protein, amorphous aggregate and amyloid fibril can be formed irreversibly.

1.5 Solvent effects: water and cosolvents

Protein needs to be hydrated in order to function properly. Solvent plays an important role in protein folding, maximizing the solvent entropy provides the free energy for folding. Burial of nonpolar side chains and hydrophobic effect can stabilize proteins. The interactions between solvent and protein, such as van der Waals, ionic and hydrogen bonding can also affect the stability.

It is important to select appropriate liquid formulation to increase the stability and mitigate aggregation. Excipient molecules are added to therapeutic products to improve their manufacturability, stability, and bioavailability. In the current market, amino acids, salts, sugars, and surfactants are commonly used.[26, 27]

Free amino acids such as histidine, arginine and amino acid salts show positive influence on protein stability.[28, 29] They prevent aggregation by stabilize the secondary structure and the charges on the side chain of some amino acids can stabilise the protein further. They can also be used as buffers to control the pH of formulations.[30] The surface charge of protein is altered when protein interact with amino acids, this causes electrostatic repulsion of particles. This alteration of surface charges can increase colloidal stability however it can have some drawbacks. The conformation of the protein might be slightly altered due to the increased charge and causing decrease in conformational stability.

L-histidine and L-histidine salts are commonly used in approved formulations. For example, Canakinumab, Efalizumab, Golimumab, Obinutuzumab, Ranibizumab and Ustekinumab all uses either L-histidine or the salt or most of the cases, both are used.[30]

Arginine is a fascinating case in formulation design, a "Gap Effect" mechanism has been proposed by Baynes and Trout.[31] When two protein molecules associate, the gap between them is too small for the excipient to penetrate but water molecules are small enough to enter. The free energy of the encounter complex increases due to the removal of water molecules from the gap and slows aggregation. They suggested additives like arginine as "neutral crowder", they are larger than water but does not affect the free energy of isolated protein molecules.[32, 33]

Sugars such as sucrose and trehalose can also increase protein stability and usually used in approved formulations. Sugar molecules stabilise protein by excluded volume effect, they are excluded from protein surfaces, since protein surfaces are surrounded by these molecules, they cannot penetrate this layer results in repulsive

PPI.[34] Scientists have done research on intermolecular interactions in sugar solutions to find the molecular origins of this effect. They can easily form hydrogen bonds between neighbouring sugar molecules or with water molecules, sucrose and trehalose are more likely to form hydrogen bonded networks than the other sugars.[35]

Salts slow aggregation based on ion-charge mechanism. As we know, salts are always dissociated into ions in solution. This dissociation allows the shielding of exposed charged areas of proteins by the ions and stop aggregation at these areas.[36] Sodium phosphate and sodium chloride are examples of salts used in commercial formulations, e.g., Abciximab (sodium phosphate; sodium chloride; polysorbate 80).

But when the ionic strength is too high, the aggregation increases due to the electric double layer is compressed.[37] There is less repulsive force between particles.

Quaternary amines such as choline can also stable proteins with a different choice of anions such as chloride, phosphate[38] and amino acids[39]. Choline chloride based deep eutectic solvents (DES) are also used as solvents, catalysts or chemical donors.[40]

Ionic liquids (ILs) are compounds composed of ions with a melting point below 100 °C. They are greener alternative for organic solvents. Proteins tends to be more soluble and stable in ILs and less self-aggregated.[41] They can also be used as solvents when extract bioactive compound and electrolyte materials.[42] Examples of ILs including ammonium-based ILs, Choline-based ILs and imidazolium-based ILs.[43, 44]

Surfactants such as polysorbate 20 and polysorbate 80 are present in most of the market formulations. The proposed working mechanism is these surfactants

covering hydrophobic areas of protein surfaces hence prevent aggregation. However, the degradation of these surfactants can form insoluble particles and result auto-immune response.[45]

Some solvents can destabilise proteins such as urea and GdmCl. The unfolding of protein starts with expanding of hydrophobic core, next the core is solvated by water and followed by urea. Urea can act as a denaturant by interacting with the protein directly to form hydrogen bonds with polar parts of the protein or change the solvent environment makes the solvation of hydrophobic residues more favourable.[46] The mechanism of how guanidinium chloride (GdmCl) denatures protein is not well understood, studies have been carried out to focus on the effect of GdmCl on hydrophobic interaction.[47]

To develop working formulations, the synergetic effect of excipient mixtures needs to be considered. For example, sugars will surely stabilize the tertiary structure of protein and slow aggregation by excluded volume effect but at the same time, viscosity of the solution is increased hence increases the chance of aggregation upon agitation. By adding salts to the solution, the viscosity increase can be mitigated.[48]

1.6 Binding events: thermodynamics and kinetics

Binding specificity and affinity are vital for molecular recognition. The thermodynamic properties can provide enthalpy and entropy for the ligand binding to its target receptor and the kinetics give the residence time a ligand can stay in contact with the pocket of the protein.

Regarding the thermodynamics of binding, binding affinity is defined by the Gibbs

free energy $\Delta G = \Delta H - T\Delta S$. The main factor affects the entropy is hydration effect.[49] Another unfavourable but important factor is translational degrees of freedom.[50] The binding enthalpy change is due to breaking/formation of hydrogen bonds, van der Waals interactions, salt bridges and so on.[51]

The equation for binding is $[Protein] + [Ligand] \leftrightarrow [Protein{\cdot}Ligand]$ and the binding affinity is defined by dissociation constant $K_d = \frac{[Protein][Ligand]}{[Protein{\cdot}Ligand]}$. When the system reaches equilibrium, the K_d is the value when half of the protein-ligand complexes are formed. The time required to reach equilibrium and maintaining in position in the pocket is rather important.

The value of K_d reflects the affinity between the protein and ligand therefore it's not directly affected by the time to reach equilibrium, however when it's too long to reach equilibrium, the value of K_d can be difficult to measure and leading to inaccuracy. For example, when the protein concentration is too high, the time takes for half of the protein to bind to the ligand might be too long and if the ligand binds tightly to the protein, eventually equilibrium can be reached and K_d is calculated, but if the ligand doesn't bind to the protein strong enough, the residence time is too short, system might not be able to reach equilibrium and lead to inaccurate measurements. Different to thermodynamics, kinetics affect binding affinity indirectly.

1.7 Experimental and theoretical structural techniques to obtain simulation models

To study protein using molecular simulation, a good model is needed as an input. There are experimental techniques such as x-ray crystallography and cryo-electron microscopy, theoretical techniques such as homology modelling and alpha fold.

X-ray crystallography is an experimental technique to obtain three-dimensional atomic structure of a crystal. A reliable source of protein is needed, purification is then taken place to gain high yield crystals. Then the crystal is exposed to a beam of X-ray, diffraction will occur, and the images are collected. The resolution needs to be sufficient to get the atomic details, unit cell dimensions can also be determined from the diffraction pattern. Date collected is processed mathematically to get the positions and intensity of diffraction spots. The intensity of diffraction spot is calculated using the amplitude of the reflection and phase angle of the waves. These data are then used to get the electron density map which gives the arrangement of the protein atoms. Finally, the built structure is exported as a PDB file and uploaded into online data bank.[52]

Cryo-electron microscopy (Cryo-EM) is another experimental technique to gain structural information of biological systems. The specimen is prepared by freezing the unstained biological sample to below -150 °C, trapped it in vitreous ice. Then the specimen is placed under electron microscope where a beam of electrons passes through it and images are formed. These 2D projections are then used to reconstruct 3D model.[53] Cryo-EM can determine the structure of large macromolecules in their native state in solution so it doesn't require crystallization of samples which is an advantage over X-ray crystallography. However, the resolution level of cryo-EM is typically below 3Å which is not high enough for a starting structure of molecular simulations.[54] In the past decade, advances in emission source, electron director, and image processing improve the resolution of cryo-EM greatly and it will keep improving in the future.[54, 55]

Homology modelling also known as comparative modelling is a computational method to predict the 3D structure of protein using its primary sequence based on

a template. Homology modelling is based on two observations, the first one is the 3D structure of a protein in more conserved than its primary sequence, so a few changes in amino acid sequence normally result in a small change in the 3D structure; Second one is homologous proteins have similar 3D structure.[56] There are several steps involved, 1) identify and select a template; 2) alignment of target and template sequences; 3) model building; 4) loop modelling; 5) addition of side-chains; 6) model optimization; 7) model validation.[57, 58] A well-known homology modelling server is SWISSMODEL.[59] It was the first automated homology modelling server and improved continuously during the last 30 years. Giving the primary sequence of the protein, it can provide models of homo- and heteromeric proteins.

Another computational modelling method is using AlphaFold. It is an artificial intelligence (AI) system to predict the protein's 3D structure using machine learning. It involves novel neural network architecture; the network directly predicts the coordinates of heavy atoms of proteins from the amino acid sequence. Different to homology modelling, it can predict protein 3D structure without any template. AlphaFold was built based on the learning from PDB data with minimal imposition of handcrafted features. It can produce accurate model even some physical contexts are missing.[60]

1.8 Hot spots

There are smaller regions known as hot spots within the binding sites of proteins that contribute most to the binding free energy, so they are vital to protein binding at these sites. This concept was originated in 1995 when residues of human growth hormone bound receptor (hGHbp) were replaced to alanine, the binding affinity was affected.[61] Based on this concept, a residue is viewed as a hot spot when a significant drop in binding affinity is observed due to its mutation to alanine. These

hot spots can also be identified experimentally when they bind to small organic molecules and the weak binding is detected by Nuclear Magnetic Resonance (NMR) or X-ray Crystallography.[62]

Hot spots are less sensitive to conformational changes and can be studied in apo protein as well as bound state. They are prime targets in drug design since they are energetically important areas of binding sites, these are the regions needed to be investigated when studying the protein-ligand interactions. They are also involved in the identification of new binding sites and the possible residues mediate protein-protein interactions.[63]

1.9 Cosolvent molecular dynamics simulations

Cosolvent simulations are MD simulations of a protein in explicit water and cosolvent molecules. Molecular Dynamics simulations with mixed solvents (MDmix) is an early stage of cosolvent simulation showing the use of organic solvents in MD simulations for hotspot detection.[64] Site Identification by Ligand Competitive Saturation (SILCS) method employs small aliphatic and aromatic molecules to find out the affinity pattern of proteins.[65] Mixed-solvent molecular dynamics (mixMD) is another method designed for prediction of hot spots[66] but it can also be used to identify orthosteric and allosteric binding sites on proteins.[67]

Cosolvent MD simulations become popular recently to detect hotspot[68, 69], predict and characterise cryptic binding sites[70, 71], study peptide aggregation[72], but currently there is no clear evidence of using these techniques in antibody formulations.

Although there are tools developed to analyse the trajectories such as Cosolvent Analysis Toolkit[73] and cosolvKit[74], cosolvent simulation still requires manual

inspection of relevant sites, and the other drawback is the hydrophobic sites tend to aggregate. The timescale limitation of normal MD simulation also implies to cosolvent simulations.

In this chapter, introduction starts on how protein can be molecular sensors followed by briefly explaining the importance of domains in protein function. The use of protein as therapeutics especially antibodies is also introduced. There are also challenges in protein studies such as aggregation, and ways to reduce aggregation such as solvent effect is mentioned. Thermodynamics and kinetics all counting for successful binding between protein and ligand. For reliable simulations, the initial structure used is vital, there are experimental and theoretical techniques to obtain the 3D model. Lastly, hot spots and cosolvent molecular dynamics simulation are introduced.

Chapter 2: Methodology

In this chapter, the methodology used in this project is introduced, including atomic and coarse-grained molecular dynamics, AMBER force field, software to run simulation and docking.

2.1 Computational Simulations

Quantum mechanics (QM) is a purely physics-based theory that describe the physical properties of matters in a microscopic scale of atoms and sub-atoms like electrons and nucleus.[75] Molecular mechanics (MM) uses the classical description of molecular systems. It has wide applications to different sized systems, varied from small molecules with a few carbons to large biomolecular complexes such as proteins with or without membranes. MM force fields are the methods used for protein simulations and vital to study the conformational flexibility.[76]

Due to the hierarchical nature of biological systems, multiscale approaches have been used to study biological processes.[77] Highly accurate quantum chemical calculations are used to study chemical reactions and the mechanisms driving them.[78] The computational cost for quantum calculations is too high so the calculations can only be done with a few atoms in a short time scale. The next level is molecular dynamics (MD) simulations, all-atom MD can give conformational dynamics of proteins and their interactions with ligands and other proteins. For biological processes happen in a longer timescale such as aggregation, coarse-grained simulations will be involved.[79]

In this project, mainly molecular dynamics simulations were used. MD simulations are the most common technique used to study protein folding, predict binding site, and estimate binding affinity.[80] Other calculations like homology modelling and molecular docking were also carried out.

2.1.1 Molecular Dynamics

MD is the computational method to simulate the time-dependant atomic motions of biological molecules based on the classical Newton's law of motion. First, we need to know the force exerted on the atoms from the potential energy of all interactions in the system. The force equals to the gradient of the potential energy:

$$\vec{F}_i = -\frac{\mathrm{dV}}{d\vec{r}_i} \tag{1}$$

Then according to Newton's second law of motion, we can write the derivative form of the equation:

$$a = \frac{d^2\vec{r}_i}{dt^2} = \frac{\vec{F}_i}{m_i} \tag{2}$$

In MD simulations, the Newton's equation is integrated in order to get the new coordinates **r** at time (**t**+**Δt**) with the known initial coordinates **r0** at time **t0**.

There are two main approximations in MD simulations, Born-Oppenheimer approximation and classical approximation. The Born-Oppenheimer approximation states that the nuclear motion and electron motion in a molecule can be considered separately since nucleus are much heavier than electrons (more than 1000 times). According to the approximation, only electron motion needs to be calculated when nucleus is thought to be frozen. Nucleus are regarded as point particles, and the motion is treated using classical Newton's second law.

In all-atom MD simulations, each atom is treated as a hard sphere, using a model of potential energy known as force fields (FF). Force fields are mathematical

parameters that describes the molecular energy potentials of molecules in molecular mechanics. These parameters were developed using experimental data and high-level quantum mechanical calculations and designed based on All-atom, United-Atom and Coarse-grained (CG) force fields. All-atom force fields are the most accurate one used for all the types of atoms in the system including hydrogen. United-Atom and Coarse-grained force fields treat the hydrogen and carbon atoms as one interaction centre. Coarse-grained force fields are normally used in long-time simulations of large proteins, from 500 nanosecond to microseconds.

2.1.2 AMBER force field

AMBER force field is one of the most common classes of FF used in biomolecular simulations. There are two main terms contribute to the potential energy, bonded potential and non-bonded potential shown below in the equation:

$$U = U_{bonded} + U_{non-bonded} \tag{3}$$

The bonded term can be calculated using equation:

$$U_{bonded} = \sum_{bonds} k_r(r_i - r_0)^2 + \sum_{angles} k_\theta(\theta_i - \theta_0)^2 + \sum_{dihedrals} k_c[1 + cos(n\phi_i - \phi_0)^2 \tag{4}$$

Where k_r is the force constant of the bond, r_i is the actual bond length and r_0 is the bond length at equilibrium. k_θ is the force constant of the angle, θ_i is the actual angle and $\theta 0$ is the angle at equilibrium. k_c is the force constant of the dihedral, ϕ_i is the actual dihedral, ϕ_0 is the dihedral at equilibrium and n is the multiplicity gives the number of energy minima.

The non-bonded term can be calculated using equation:

$$U_{non-bonded} = \sum_i \sum_j \frac{q_i q_j}{\epsilon r_{i,j}} + \sum_i \sum_j 4\epsilon_{i,j}[(\frac{\sigma_{i,j}}{r_{i,j}})^{12} - (\frac{\sigma_{i,j}}{r_{i,j}})^{6}] \quad (5)$$

Where ϵ is the dielectric constant of the solvent, $r_{i,j}$ is the distance between particle i and j, and q_i, q_j are partial atomic charges of particles. ϵ is the Lennard-Jones well depth and σ referred as van der Waals radius, is the distance when the intermolecular energy between two particles is zero.

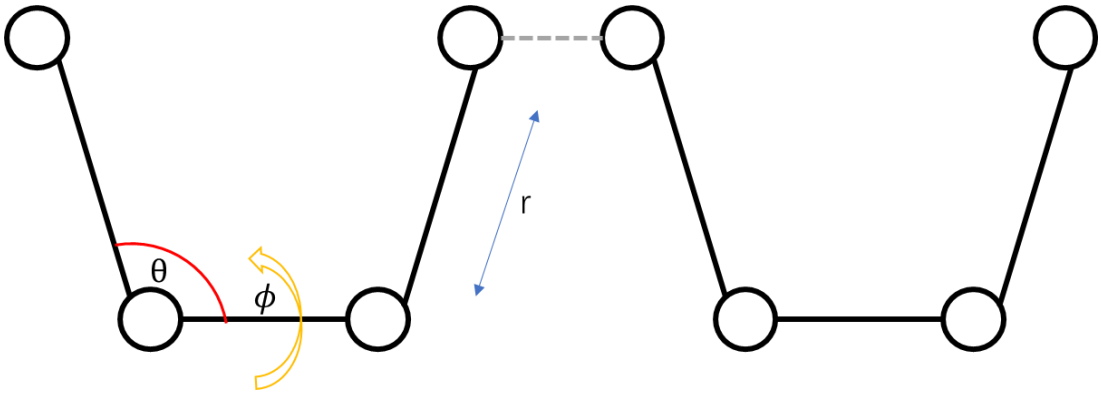


Figure 3: The schematic representation of molecules showing bonded and non-bonded interactions. Balls represent atoms and the lines represent the bond between them. r is the bond length that changes while the bond is stretching. θ is the bond angle that changes when the bonds are bending. ϕ is the torsional angle that changes when the bond is rotating. The dashed lines represent non-bonded interactions between neighbouring molecules.

By using molecular mechanical force field, one can calculate the potential energy of a system of choice (e.g., protein in water) and a set of molecular properties (e.g., flexibility, density, radius of gyration) which can be compared to experimental data. These simulations are powerful tools to study biomolecules since the speed of calculations are much quicker compared to real experiments, it is unlikely to make mistakes throughout the whole process and they can provide a lot of information, such as giving the positions of atoms of macromolecules at a femtosecond timescale which is difficult to achieve using experimental techniques and predict the response of biomolecules towards perturbations such as phosphorylation, mutation and effect of ligands which can be then compared with experimental results to validate the

findings.

There are also drawbacks of computational simulations. The main problem is the timescale. For computational simulation, it is mostly under microsecond because of the high computational cost, a lot of folding-unfolding processes, conformational changes cannot be simulated in this timescale. The AMBER FF99SBILDN is the improvement version of original 99SB force field with better side-chain torsion potentials, although it has better agreement with experimental data, force field calculation is still estimation, we can never say the data is hundred percent correct and that's why experimental results are always needed to validate computational results. It is highly depending on the input file, so a correct starting structure is crucial to initiate calculation. Processes involving formation and cleavage of covalent bonds cannot be simulated.

2.1.3 Steps of MD simulation

The general steps of MD simulation are structure preparation, setup of the simulation system, simulation and analysis. During structure preparation, the protein structure will be normally obtained from protein data bank or homology modelling. If the structure is from data bank, it can be checked using visualising software like Chimera or VMD. We can select the chains or regions that we want to simulate and make sure there is no missing loops. Mutations can also be introduced during this stage. Once we have the final pdb file that contains the atomic coordinates, we can proceed to the setup stage.

During the setup stage, the pdb file is converted to the correct file format required by the simulation software and using chosen force field to define the inter-atomic bonded and non-bonded interactions, topology file is also generated. Then the simulation box is built with one's choice of shape and size, boundary conditions are

also specified. Next, the box is solvated with water or other solvents of choices and neutralised or building up to certain molar concentration by adding ions. Finally, energy minimization is performed to give a good starting conformation for equilibration and then simulation.

There are two parts in the simulation stage, equilibration and production. During the equilibration stage, leap-frog integrator is used with 2 fs timestep, the initial velocities are assigned to the molecules randomly using a random seed in Maxwell-Boltzmann distribution at 298K and the system is heating up to the required temperature. Once the system reaches thermodynamic equilibrium with constant temperature and temperature, the production simulation can start.

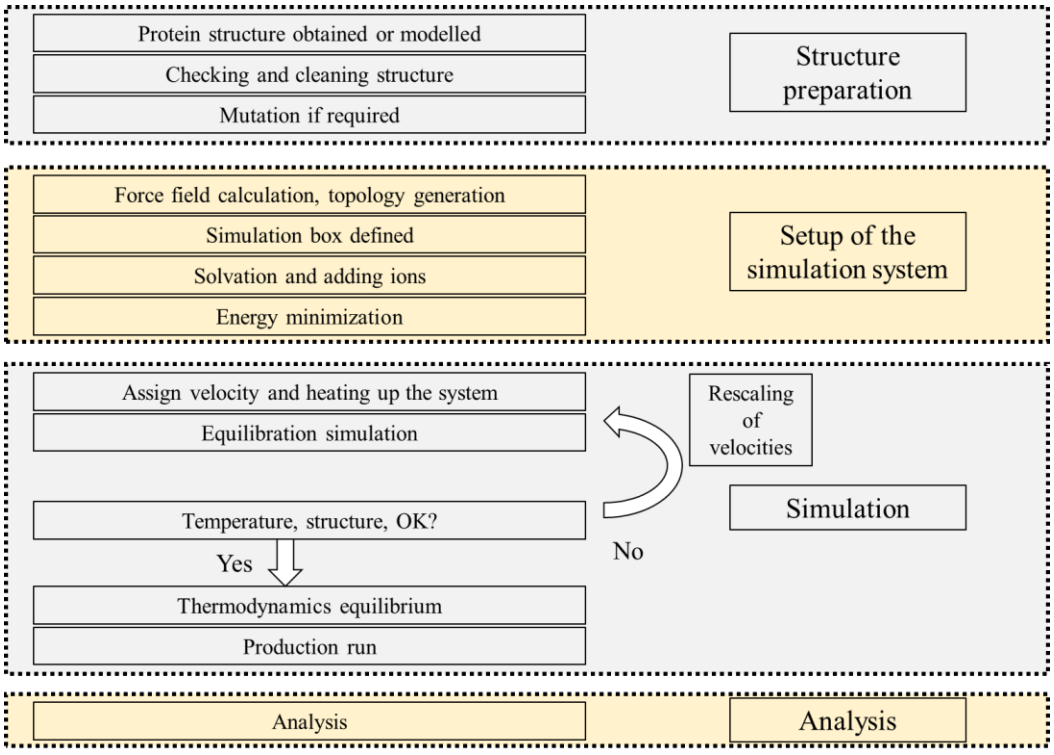


Figure 4: A flow chart showing the steps of general MD simulations including structure preparation, setup of the simulation, running the simulation and analysis. Structure preparation, setup and running of the simulation are further divided into smaller steps as described in the paragraphs.

The limitations of the simulations are the system can only be single molecular connectivity because of the force-filed model which stops us modelling any processes that break or make any covalent bonds and when very large systems are involved, the calculations for all atom MD simulations are enormous which slows down the simulations to a large degree, that's why coarse-grained (CG) simulations are also required to study large biological macromolecules and interactions that occur on a longer timescale from microseconds to milliseconds. In some cases, experimental studies of biomolecular aggregates are carried out using low to medium techniques such as cryo-electron microscopy (cryo-EM)[54] so there is no need to investigate them using atomically detailed simulations.[81]

2.2 Coarse-Grained simulation

The idea of CG simulation of proteins has been around since 1970s[82], and due to the improvement on computing powers it becomes more popular lately and being studied further.[83] CG models can simply the representation of full atomic models and keep the essential molecular properties of interests so the CG models have less degrees of freedom. This simplified representation of the system is why large biological molecules can be simulated and faster sampling is enabled due to these less degrees of freedom.[81]

The development of CG models needs to define the position of pseudoatoms that represent the group of atoms, then using energy function to define the interactions between pseudoatoms. This should reproduce the thermodynamic properties of the original system. Next, dynamical equations need to be defined to study the time-based movement of the CG model.[79] The methods to derive the energy function have been classified into structure-based, knowledge-based and dynamics-based. Structure-based method uses experimental data and usually defines the position of pseudoatoms based on Cα atoms. The energetics can then be defined by elastic

network models. Knowledge-based method also uses experimental data in order to parametrise the CG model with greater degree of transferability so it can be used in any system of interest. Dynamics-based method uses purely statistics data from full atomic simulations to derive mappings and energetics through systematic algorithms.[84]

2.2.1 MARTINI

MARTINI CG was used in this project since it is compatible with the GROMACS package. It was developed mainly for lipid model then extended to include other biomolecules.[85] There are on-lattice and off-lattice CG models, the on-lattice model is faster and the off-lattice model is more flexible and realistic. The MARTINI model is an off-lattice model that main based on a four-to-one mapping, which means four heavy atoms are represented by one bead. Hydrogen atoms are ignored because they are tiny in size. The number is not limited strictly, because sometimes it is more appropriate to include three or more atoms in one bead.[86] Ring-like molecules are exceptions, they need higher resolution, one-to-one or two-to one mapping.[87] There are mainly four types of beads: polar (P), nonpolar (N), apolar (C) and charged (Q). For early version, only type N and Q had subtypes and newer version all four main types have subtypes.[88] The subtypes are divided according to hydrogen-bonding capacities into none (0), doner (d), acceptor (a) and both (da) or based on the degree of polarity, from low polarity 1 to high polarity 5. Examples of the CG model of some molecules and types of beads are shown below in Figure 5:

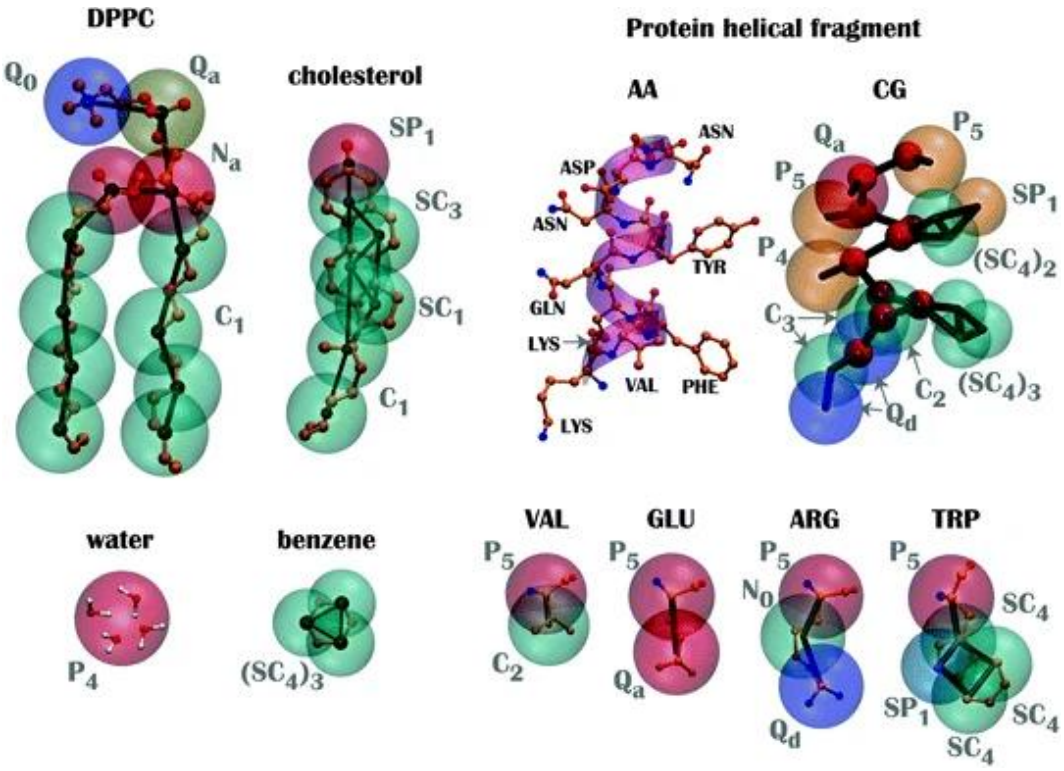


Figure 5: A graph showing the mapping between the atomic structure and the coarse-grained model of DPPC, cholesterol, protein helical fragment, water, benzene and four amino acids (VAL, GLU, ARG and TRP).[87] The transparent spheres are the coarse-grained beads, and the atomic structures are shown as balls and sticks.

The MARTINI model has been developed considered four important aspects: speed, accuracy, applicability, and versatility. Smooth potentials are used to allow large integration steps and only short-range interactions are included to optimize the speed aspect. CG results are matched to atomistic simulations as much as possible to maximise the accuracy. The building blocks of the force field are calibrated extensively against thermodynamic data. To enhance the applicability of the model, the force filed used is simple, standard interaction potentials are employed and only few parameters. There is no need to reparametrize the model every time.[88] Finally, the phase of the system is not limited and plenty of room is left in the force field to adjust structural details of molecule.[86]

The bonded interactions, i.e., bonds, angles and dihedrals are described by weak

harmonic potentials. Non-bonded interactions are described by a Lennard-Jones (LJ) potential, charged groups are described by a Coulombic energy function. These non-bonded interactions are shifted with a cut-off distance of 1.2nm. They are parameterized by comparing to experimental thermodynamic data of different CG beads such as the free energy of hydration, vaporization and the partitioning free energies between water and some organic phases.

Elastic network can be used in combination with MARTINI CG model to keep the secondary, tertiary, and quaternary structures together without affecting the dynamics of protein. Extra harmonic bonds between non-bonded beads are added to the martini topology with a distance cut-off. The force constants of elastic bond can be adjusted accordingly on protein behaviour.

After running MARTINI CG simulations, reverse coarse graining is carried out to generate a reasonable full atomic structure from the positions of the beads. The method used was "*backward"* that generated by Tsjerk Wassenaar.[89]

2.2.2 SIRAH force field

SIRAH force field is also used to run CG simulations. It is implemented in GROMACS and the pairwise Hamiltonian is used to calculate the interactions. The parameters are chosen to fit structural data. It includes parameters for protein, DNA, and water molecules. Water molecules are represented by four beads based on the tetrahedral structure in solution. The mapping of proteins keeps the nitrogen, alpha carbon and oxygen in position and side chains are modelled at a lower degree. Compared to backbone representations only considering one bead on the Cα, they might need constraints to construct the secondary structure. Backbone beads in SIRAH force field are chosen to form optimal shapes of compacted α-helices and the partial charges on every single bead can roughly describe the hydrogen-bond

like interactions, the formation of α-helices and β-sheets can be stabilised without constraints.[90]

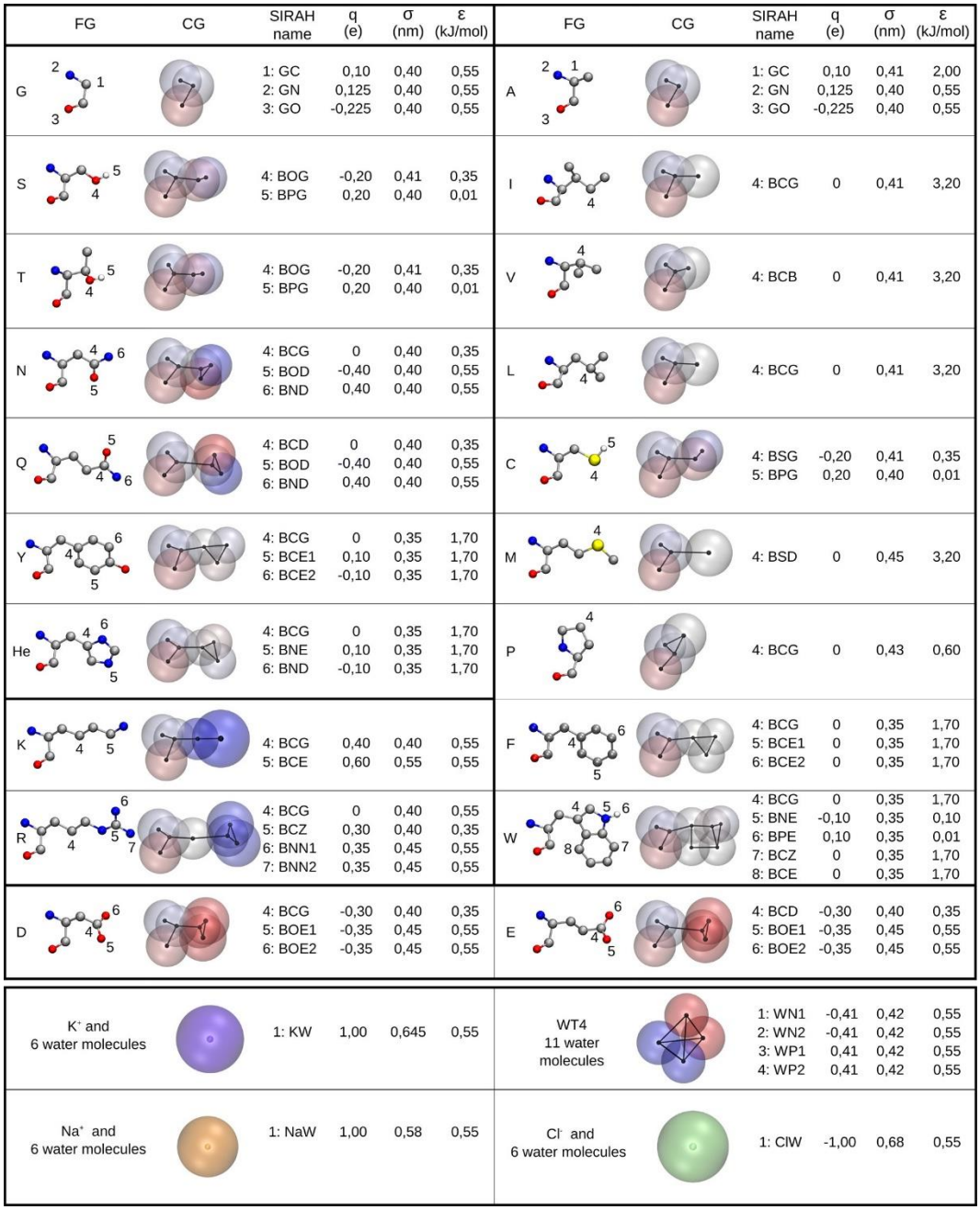


Figure 6: A graph showing the mapping between the atomic structure and the SIRAH coarse-grained model of 20 amino acids, water, K^+, NA^+ and Cl^-.[90]

2.3 Using GROMACS to run simulation and do analysis

In the project, GROMACS[91] molecular simulation software suite is used to prepare and run cosolvent MD simulations for all the proteins mentioned above, ER, AhR, NDP52, LINE-1 and mAb.

As mentioned before, AMBER FF is widely used in biomolecular simulations and specifically, AMBER99SB-ILDN was used in this work. Water molecules in the system were simulated using three-site transferable intermolecular potential (TIP3P) water model. The three-atom molecular H_2O molecule is described using experimental parameters.

Periodic boundary condition is applied to the simulation to avoid finite-size effect and keep the number of particles inside the simulation box constant. The protein of interest was placed in the centre of a cubic simulation box, and it is at least 1 nm distance from the edge of the box. It is surrounded by images of this box. Then solvation is required with water molecules and ions, cosolvent molecules can also be added to the simulation box. There are implicit and explicit solvent models, due to the stronger computational power in modern time, explicit solvent model was used to give a better accuracy. Ions such as sodium and chloride ions were added to neutralise the box and bring the solution to a 0.1 M NaCl to mimic the physiological condition.

Once the system was built, energy minimisation is required to avoid steric clash and therefore reduce the chance of crash during later simulations. The two commonly used minimising method are steepest descent algorithm and conjugate gradient algorithm. Steepest descent is very fast and travels in the direction of steepest

negative gradient. Once a new position is reached, a new minimisation is starting from that position and travels in the steepest gradient. The process will repeat until a reasonable minimum is reached. This uses only orthogonal gradient. Conjugate gradient starts with the same step as steepest gradient, then uses a combination of the gradient at the new position and the previous search direction, although less steps are needed to reach the minimum, it is slower in calculations.

After minimisation, equilibration runs of the system were carried out. There are two phases of equilibration, the first one is NVT ensemble and the second one is NPT ensemble. Initially, the system should have a temperature of zero Kelvin and specific conditions were given to increase the temperature to a desired value. NVT ensemble is also known as canonical ensemble, the number of particles, the volume of box and the temperature of the system are set as constants. A velocity rescaling method based on Berendsen thermostat was used during this phase, the system is coupled to a temperature bath and the velocities are rescaled at each step. The rescaling is randomised by adding a stochastic term.

After the 100ns NVT equilibration phase, the system needs to bring to a constant pressure of 1 bar. The NPT ensemble also known as isothermal-isobaric ensemble was involved, the number of particles, the pressure of the system and the temperature of the system are set as constants. Both thermostat and barostat are required to satisfy the conditions. Parrinello-Rahman barostat was used with a relaxation time of 2ps at the pressure of 1 bar. The NPT equilibration is also 100ns.

Finally, the well equilibrated system is ready for production run. The length of simulation varied from 50ns to 200ns.

After the simulations, trajectories need to be visualised and analysed. Software such

as chimera and VMD were used to visualise the runs. GROMCAS can be used to calculate root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF), radial distribution function (RDF), radius of gyration (Rg), solvent accessible surface area (SASA) and principal component analysis (PCA) for the running simulations.

Root-mean-square deviation (RMSD) is used to measure the stability of the protein during the simulation time compared to a reference structure and check if there are conformational changes. The equation is shown below:

$$RMSD(t) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i(t) - x_i^{ref}\right)^2 + \left(y_i(t) - y_i^{ref}\right)^2 + (z_i(t) - z_i^{ref})^2} \quad (6)$$

Where n is the number of particles, $x_i^{ref}, y_i^{ref}\ and\ z_i^{ref}$ are the reference positions of particle i, $x_i(t)$, $y_i(t)$ and $z_i(t)$ are the positions of particle i at time t after fitting to the reference structure.

Root-mean-square fluctuation (RMSF) is used to measure the atomic fluctuations of particle i and the reference position. The equation for RMSF is:

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{i=1}^{T}\left(x_i - x_i^{ref}\right)^2 + \left(y_i - y_i^{ref}\right)^2 + (z_i - z_i^{ref})^2} \quad (7)$$

Where T is time of the simulation that we want to use, $x_i^{ref}, y_i^{ref}\ and\ z_i^{ref}$ are the reference positions of particle i, x_i, y_i and z_i are the positions of particle i. The reference positions are normally time-averaged positions of particle i. RMSFs were mostly calculated for residues.

Radial distribution function (RDF) gives the probability of finding other particles

when given the positions of reference structure. It is defined:

$$gAB(r) = \frac{\langle \rho B(r) \rangle}{\langle \rho B \rangle local} \tag{8}$$

$$\langle \rho B(r) \rangle = \frac{1}{N_A} \sum_{i \in A}^{N_A} \sum_{j \in B}^{N_B} \frac{\delta(r_{ij} - r)}{4\pi r^2} \tag{9}$$

Where $\langle \rho B(r) \rangle$ is the particle density of particle B at distance r around particle A and it can be expressed as equation 7, N_A are N_B are the number of particles A and B respectively, r_{ij} is the distance between the ith and jth atoms.

Radius of gyration (Rg) is used to give a rough measure of compactness of the protein, the equation is:

$$Rg = \sqrt{\frac{\sum_i |r_i|^2 m_i}{\sum_i m_i}} \tag{10}$$

Where m_i is the mass of atom i and r_i is position of particle i with respect to the centre of mass of the molecule.

Principal component analysis (PCA) is used to give the correlated motions of a molecule during simulation. A symmetric 3Nx3N covariance matric C of atomic coordinates is calculated:

$$C = \langle (X_i - \bar{X_i})(X_j - \bar{X_j}) \rangle \tag{11}$$

The matrix C is diagonalised using orthonormal transformation matrix R:

$$R^T C R = diag(\lambda_1, \lambda_2, \ldots \lambda_{3N}) \tag{12}$$

The columns in R are the eigenvectors, also known as the principal components (PC) of trajectories which give the directions where the molecules showed a largest correlated motion and λ is the eigenvalue of matric C. The trajectories projects into the PCs p(t):

$$p(t) = R^T(x(t) - \bar{x}) \quad (13)$$

p is the position matrix.

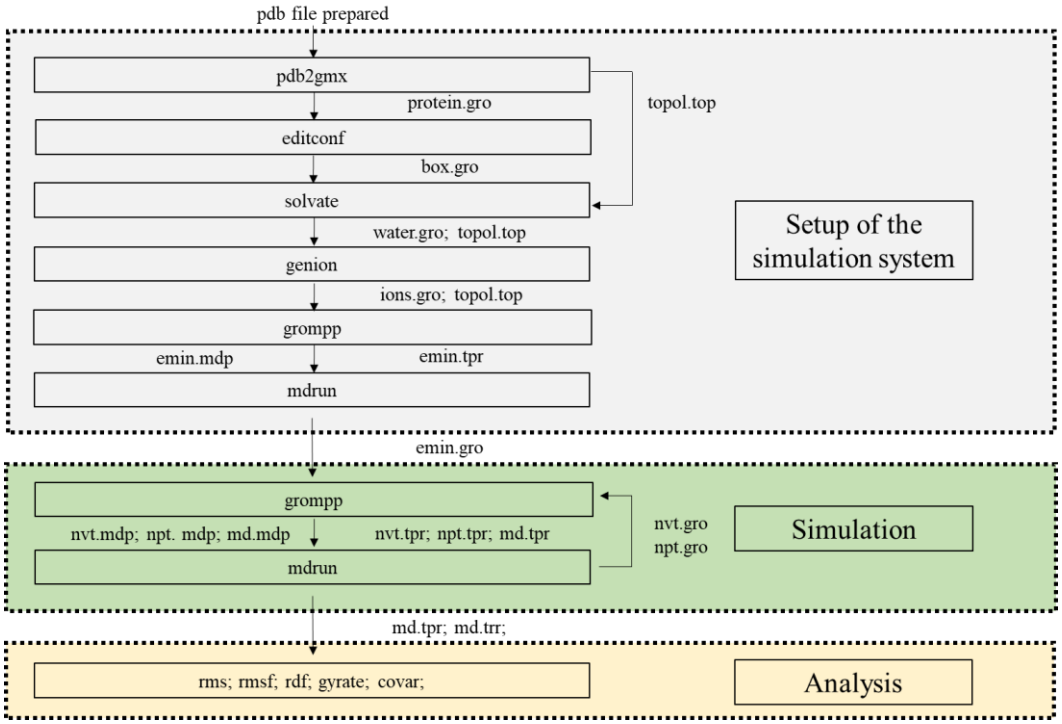


Figure 7: A flow chart showing the steps of running MD simulations using GROMACS including setup of the simulation, running the simulation and analysis. Each step has different command and input files that are required to carry out the process.

2.4 FTmap server

Experimental techniques need to solve several crystal structures which are expensive and time consuming. Computational solvent mapping is a fast and easy method to identify possible hot spots of proteins. FTmap server was used as a solvent mapping tool that using 16 small fragment-sized organic molecules as molecular probes and placing them around the surface of protein. These organic molecules are ethanol, isopropanol, isobutanol, acetone, acetaldehyde, dimethyl ether, cyclohexane, ethane, acetonitrile, urea, methylamine, phenol, benzaldehyde, benzene, acetamide and N, N-dimethylformamide. These molecules were selected with different shape, functional groups, and polarity, most of these probes contain one hydrophobic moiety and one or two polar groups. Hydrophobic and aromatic

compounds are also included. This set of 16 probes can define the hot spots effectively.[92, 93] These probes will accumulate at their preferable positions, FTmap finds these positions for every single type of probe, cluster the probes and ranks these clusters based on the average energy.

There are five steps of FTmap algorithm. The first step is rigid docking of those probe molecules to the protein and 2000 best poses with lowest energy are used for following step. The second step is energy minimisation and rescoring of the 2000 poses. The third step is clustering of the energy minimised structures and ranking based on their averaged free energies, 6 clusters with lowest free energies for every probe are kept for later. To cluster the probe conformations, the structure with lowest energy is selected and other structures within 3Å RMSD form the first cluster. The next lowest energy structure is then used to give the second cluster. The process will continue until all conformations are assigned to clusters. The fourth step is determining 'consensus' sites (CSs) and ranked based on the number of clusters within the sites. CS1 has a centre cluster that containing the maximum number of neighbours and the clusters within 4Å are the members of CS1. The final step is characterization of the binding site. The largest CS1 is considered to be the most important hot spot and forms the core of the binding site. And the binding site is expanded by adding other CSs that are within 7 A from the CSs already in the binding site.[92, 93]

The same set of hot spot residues of a protein may be responsible for mediating the interaction with small molecules like ligands and bigger molecules like peptides and proteins.[94, 95] For example, in the interaction of a neutralizing human IgG with HIV-1 gp120, hot spots have been found in the protein-protein interfaces.[96] The computational hot spots are not spread through the protein-protein interfaces randomly, they are more likely to form clusters. They are assembled within densely packed regions. Three or more hot spots that are close to and interact with each

other can form a hot region.[97] This highly packed nature of hot spots and hot regions ease the removal of water molecules when binding to other molecules so packing is an important factor in binding and protein folding.[98]

2.5 SeeSAR

Developed by BioSolveIT, SeeSAR is a great tool to use for protein-ligand docking and visualisation of binding poses. Using the docking function of the software, a protein model is required as an input, binding pocket needs to be defined to start docking process. Once the pocked is defined, ligand files can be imported to generate binding poses and providing binding affinities. SeeSAR uses a HYDE scoring function. It estimates binding free energy only based on hydrogen bonding and dehydration. HYDE is an atom-based scoring not calibrated based on experimental data so it can be used for all types of protein targets. The protein and ligand are placed in aqueous solutions unbounded, water molecules around the ligand and in the binding pocket are removed causing broken of hydrogen bonds leading to unfavourable energy changes. When hydrophobic groups are involved in the protein or ligand, removal of water from these sites will contribute favourable energy changes. The newly formed hydrogen bonds between protein and ligand lead to favourable energy changes. These energies contribute to the overall binding energy.[99, 100]

Starting with briefly mentioning QM and MM, stating that for this project all atomic MD was running with AMBER force field using GROMACS, steps involving running the simulations were listed, data analysis were also done using GROMACS, including RMSD, RMSF, RDF, Rg and PCA. Coarse grained simulations were prepared using MARTINI and SIRAH force field. Hot spots were identified using FTmap and docking was done using seeSAR.

Chapter 3: Development of the formulation stabilising mAbs using cosolvent simulations and CG

In this chapter, general structure of IgG antibody was given and one specific monoclonal antibody, NISTmAb was built and used to run all atomic and coarse-grained cosolvent simulations. Both the Fab region and the whole mAb were run, the results were analysed.

Immunoglobulin G (IgG) is one of the most common proteins found in human serum. The basic structure of IgG consists of two heavy and two light chains, arranging into two different regions based on their functions. There are two antigen binding fragment (Fab) domains which identify the antigen and one crystallizable fragment domain (Fc) which interacts with other components of the immune system to remove the antigen. Each Fab domain further dividing into two variable and two constant domains, the two variable domains forming the variable fragment which gives the specificity of the antibody.[12] There are three complementary determining regions (CDRs) and four framework regions (FR) in each variable domain, CDRs are the most important part involved in antigen recognition. The Fc domain only contains constant domains which are basically structural framework that are highly conserved in different antibodies. The modular nature of antibodies enables easier protein engineering.[101]

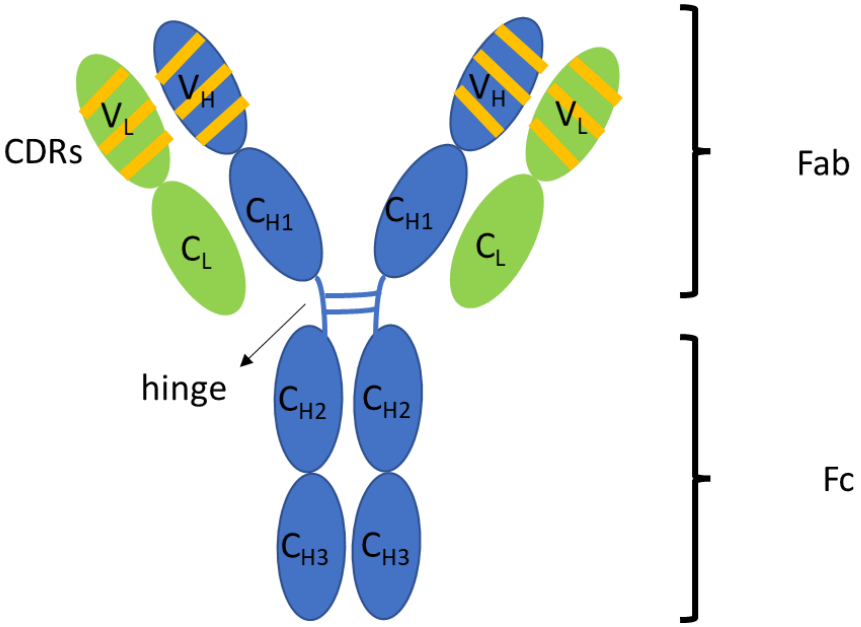


Figure 8: A schematic representation of IgG1. Variable domains in light and heavy domains, V_L and V_H respectively, constant domains in light and heavy chains, C_L and C_H respectively, complementary determining regions, CDRs and framework, FR. The antigen binding fragment, Fab domain, hinge region and fragment domain, Fc are labelled as well.

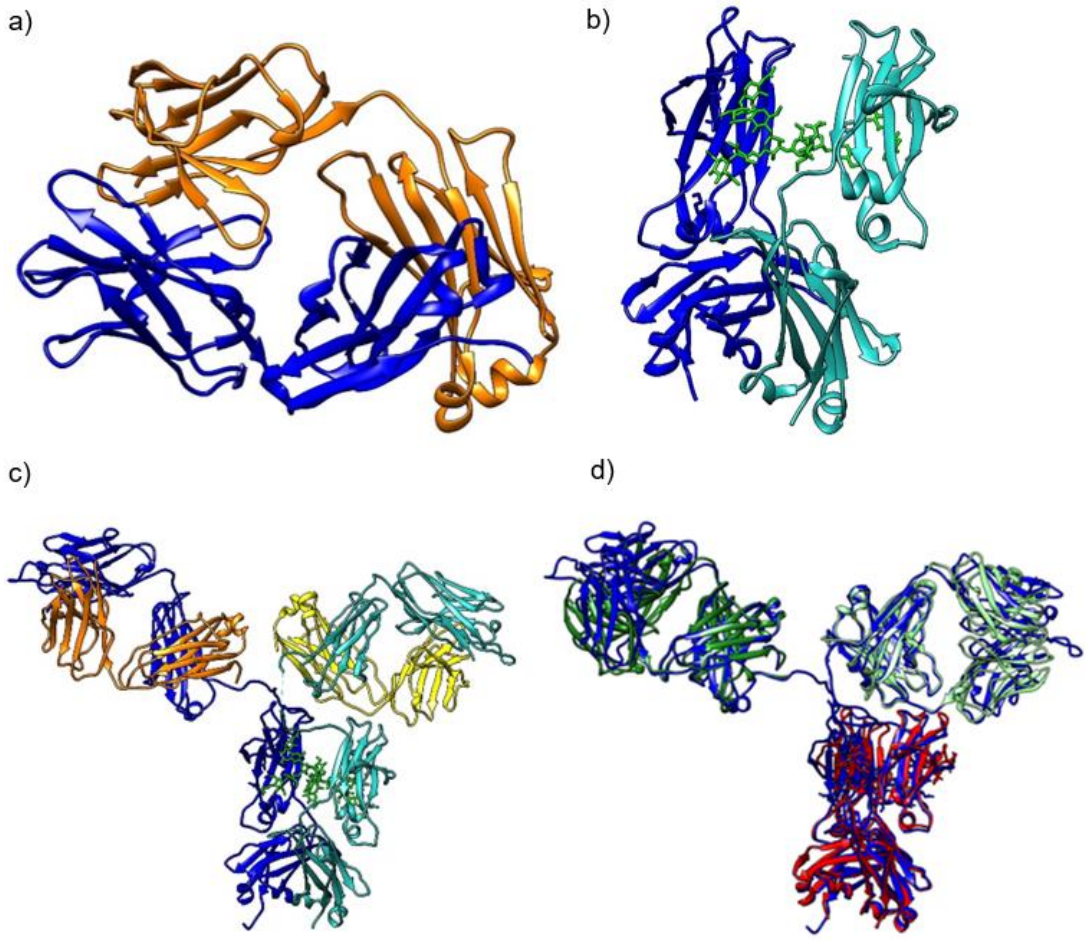
A series of all-atomic and CG simulations were carried out for the NISTmAb. NISTmAb Reference Material (RM) 8671 is a humanized IgG1 monoclonal used specifically for research purpose. It has been tested to fit the purpose for developing therapeutic protein characterisation technologies.[102] A paper by Xu and colleagues[103] gives some experimental information of NISTmAb in different cosolvents such as amino acids, sugars and salts. Especially the second osmotic viral coefficient (B_{22}) values were used as a main comparison to our simulated results.

The second osmotic coefficient, B_{22} measures the deviation of protein's behaviour inside a solution from ideal behaviour. It is often used to predict solubility, crystallization condition and aggregation propensity.[104] When protein molecules are inside aqueous solutions with a concentration, they interact with each other due to intermolecular forces. The weak protein-protein interactions induced by the

appearance of other compositions in the solutions such as excipients can be quantified using B_{22}. The value of B_{22} can be calculated experimentally using static light scattering (SLS), self-interaction chromatography (SIC), it is highly affected by concentration. The sign and values of B_{22} indicates the direction and strength of the net intermolecular forces, positive B_{22} values suggest repulsive protein-protein interactions (PPI) are induced by excipients and interactions between protein and solvent are more favoured so less chance to aggregate. Negative B_{22} values are due to attractive PPI and increases the chance of aggregation. Therefore, B_{22} is a good representation for protein aggregation propensity.[105] Since RDF gives the probability of finding other particles relating to the reference structure, the hypothesis is there could be a relationship between RDF and B_{22} values.

3.1 Model of NISTmAb

The crystal structure of the Fab region (PDB code: 5K8A) and Fc region (PDB code: 5VGP) of the NISTmAb were obtained from the PDB database. Since there is no published crystal structure for intact NISTmAb, the model of the whole mAb was built based on the crystal structure of another human IgG1, b12 (PDB code: 1HZH). SWISSMODEL and MATCHMAKER in Chimera was used to build the structure of NISTmAb. The sequence of the heavy chain was combined from 5K8A and 5VGP, by setting 1HZH as a template, SWISSMODEL is used for homology modelling. The modelled heavy chain, light chain, and b12 were opened in Chimera, by using b12 structure as a reference, light chains and heavy chains were matched to the b12 model (matched models shown below).


a)
b)
c)
d)

e)

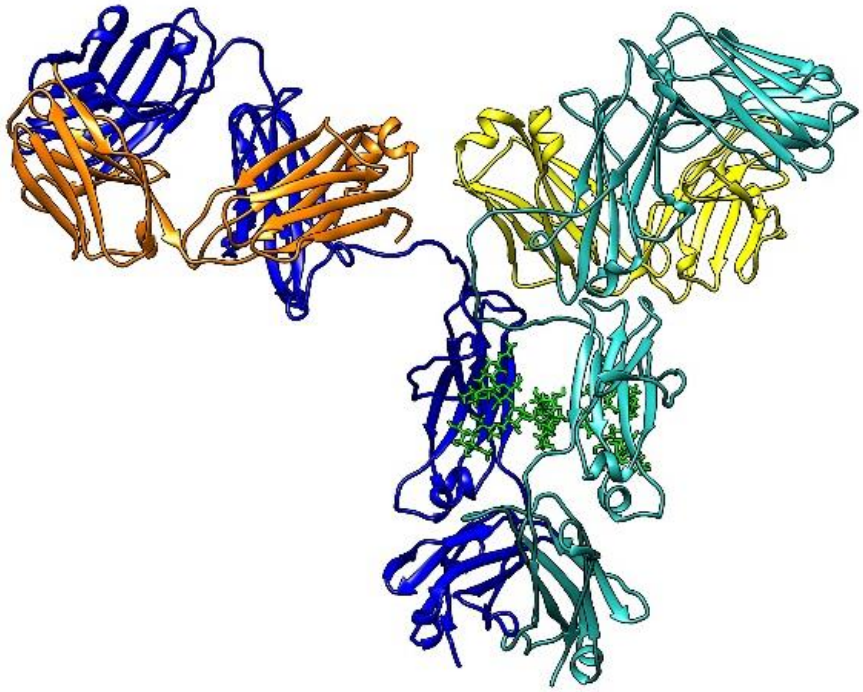


Figure 9: a) Fab region of NISTmAb (PDB:5K8A), the light chain is coloured orange, and heavy chain is coloured blue. b) Fc region of NISTmAb (PDB:5VGP), one of the heavy chains is coloured blue and the other one is coloured light blue, glycans are coloured green. c) Protein b12 (PDB: 1HZH) and d) Matched NISTmAb and Protein b12, green chains are 5K8A, red chains are 5VGP and blue chains are 1HZH e) Modelled NISTmAb, light chains are coloured orange and yellow; heavy chains are coloured blue and light blue, glycans are coloured green.

Accoding to the published paper on the determination of the sequence of the NISTmAb[106], the light chain of NISTmAb has the V_L region from residue 1 to 106 and C_L region from residue 107 to 213, the CDR1 is from 24 to 33, CDR2 is from 49 to 55 and CDR3 is from 88 to 96. The heavy chain of NISTmAb has the V_H region from residue 1 to 120, C_{H1} region from residue 121 to 223, C_{H2} region from 241 to 343 and C_{H3} region from 344 to 450. The CDR1 is from 31 to 37, CDR2 is from 52 to 67 and CDR3 is from 100 to 109.

3.2 Solvents

They were simulated using cosolvents mentioned in the paper[103] using different concentrations vary from the concentration used in the experiment which is 25 mM histidine with 171 mM – 300 mM other amino acids or sugars, 0.5% v/v, 2.5% v/v and 5% v/v. For a mixture of histidine with another solvent, the ratio is 1:1. These excipients are drawn in Chimera and charges are calculated using AMBER force field. All excipients were parametrized at a neutral charge except arginine has a +1 charge. They were added to the simulation box by using gmx insert-molecules function, number of molecules were specified.

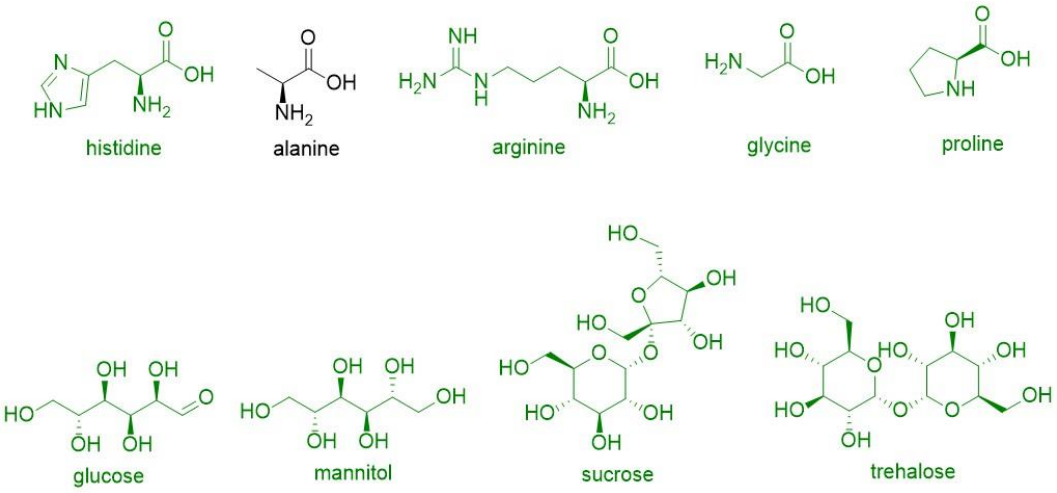


Figure 10: 2D structures of excipients and their names underneath. Most of them are shown in green as they can be used to stabilise protein, alanine is in black because there is not much evidence showing the effect on protein when it is used in formulation.

Histidine is often used as buffer in mAb formulations, it can decrease protein aggregation propensity and the proposed mechanism of the stabilisation is histidine can shield the hydrophobic region of the mAb surface.[107] Alanine is not commonly used in mAb formulations, its effect on stabilisation of mAb is unsure. Arginine is commonly used as a protecting excipient against aggregation, it tends to bind to the carboxylate group on the mAb surface and form cation-π interactions with aromatic

side chains.[29, 108] Glycine can stabilise mAb by interacting directly with charged side chains and partial charges of the backbone.[109] Proline can bind to aromatic residues and hydrophobic regions of mAb surface.[110, 111] Sugars in general have been used for years to stabilise macromolecules. A review on effect of trehalose gives three possible rationales, one is the Timasheff and Arakawa's preferential exclusion theory[112] which suggests the sugar molecules doesn't interact with proteins directly. These sugar molecules force the water molecules to be away from the protein, so its hydrated radius decreases, and compactness increases. Another rationale is the vitrification theory that trehalose shields the protein from abiotic stresses by forming a glassy matrix around the protein. The final rationale is water replacement theory, water molecules are substituted by trehalose.[113]

3.3 Simulations and analysis of Fab region

Initially, I have decided to run a set of 50ns all-atomic simulations in different excipients (histidine, alanine, arginine, glycine, proline, sucrose, glucose, trehalose, and mannitol) with a total concentration of 2.5% v/v for the fab region. The simulations were carried out as replicas of three. Most of the RMSD plots show there is no significant fluctuation during the simulation time. An example of one RMSD graph is shown below in Fig.11, from this plot, we can see there is a slightly increase towards the end of the simulation, so I decided to increase the simulation time to 100 ns then to 150 ns. From the whole set of 150 ns all-atomic simulations, although there are fluctuations up and down, the general RMSD values are between 0.15 nm - 0.2 nm.

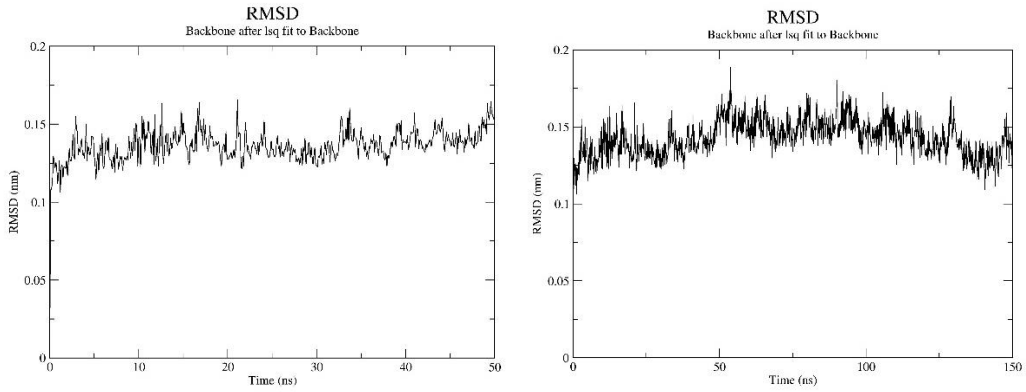


Figure 11: The average RMSD plots from three replicas for one fab region in 2.5% v/v histidine. On the left hand is the close look of the simulation in first 50 ns and on the right is the full 150ns simulation.

The RDF plots are also shown below, we can see that the distribution changes dramatically from the first 0 to 50 ns and then less increases occurred for the rest of 100 ns. This shows as the simulation progressing, some solvent molecules were moving closer to the protein and stays within a certain distance to the protein. The radial distribution function gives a probability of finding the excipient within the distances to protein. Once the distance is above 4 nm, the solvation cell is reached. For each excipient at every concentration, the maximum probability is used for further analysis.

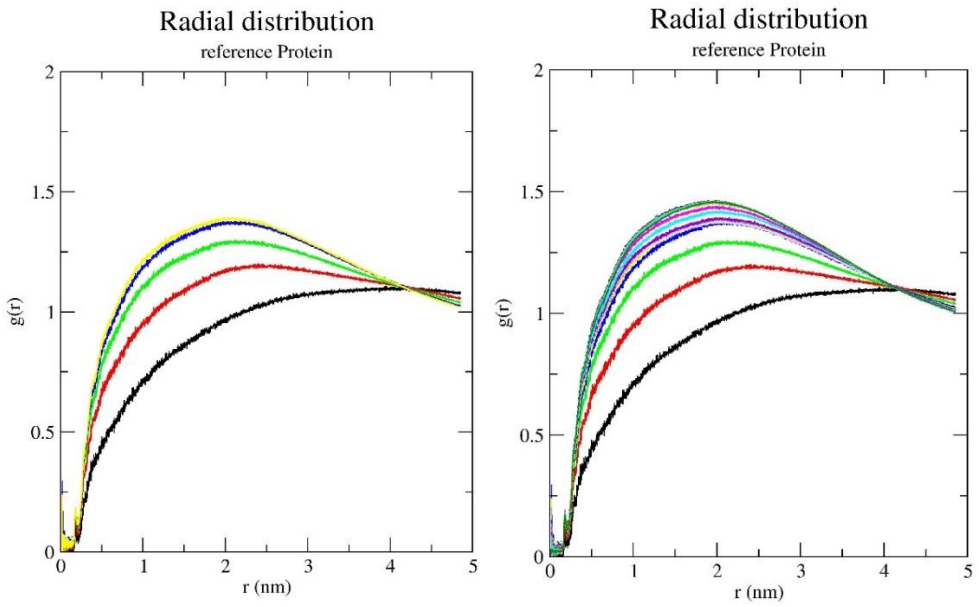


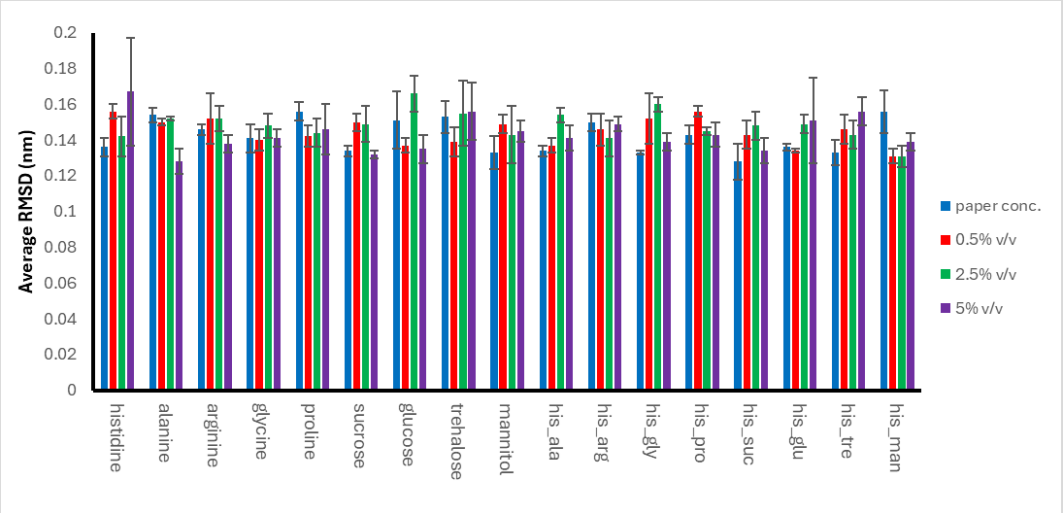
Figure 12: The RDF plots of one replica for one fab region in 2.5% v/v histidine, on the left-hand

side is first 50ns and on the right is the whole 150ns. Each colour represents one RDF plot at different time length from 10ns to 50ns/150ns at a 10ns interval.

Since most of RMSD graphs showed that the fab is rather stable throughout the simulation time and the RDF graphs normally converged to a reasonable extent in the first 50ns, and the general trend in different cosolvents remains similar so I decided to run all the other all-atomic simulations for fab region and the whole mAb as three replicas of 50ns. This is a reasonable time scale to run all the excipients at all concentrations and have a nice comparison at the end of the project. The RMSD, RMSF, SASA, Rg, PCA and RDF results calculated by GROMACS will be shown in the following pages.

3.3.1 RMSD

The discussion in the following pages is about the fab region (shown in previous Figure 9a). As mentioned previously, simulations were performed three times in every condition, so an average RMSD plot can be obtained and then the maximum value and average value of all the frames were calculated for every excipient at each concentration. The average and maximum values of the RMSD plots are shown below in Figure 13.



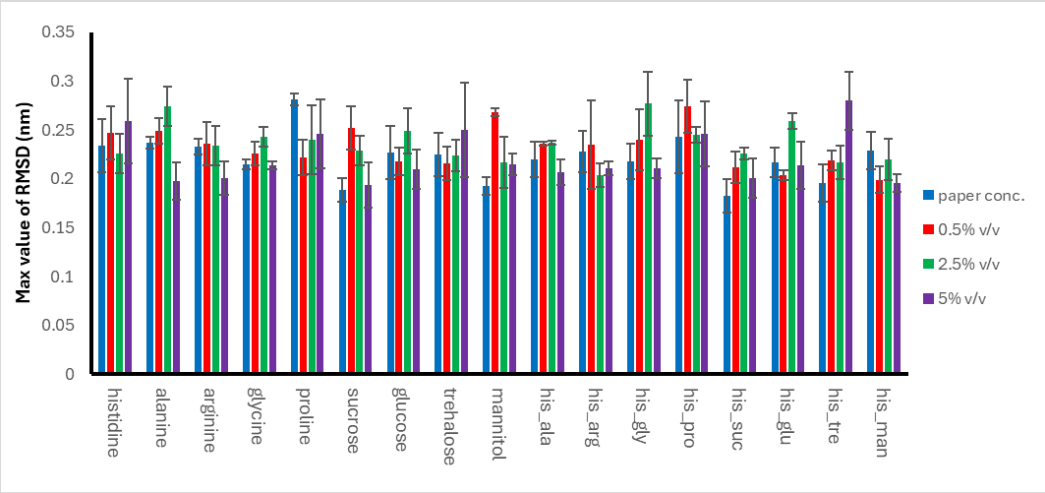


Figure 13: Bar charts for all the solvents in 4 different sets of concentrations. The top one lists the average RMSD values for all the replicas and the bottom one is the maximum values of the RMSD values with standard deviation shown as error bars.

The average values lie between 0.12 to 0.17 nm and the maximum values lie between 0.18 to 0.28 nm. The lowest average RMSD value is 0.128 nm occurs in alanine when the concentration is 5% v/v and in a mixture of histidine and sucrose when the concentration is 25 mM histidine and 300 mM sucrose. The lowest maximum RMSD value is 0.183 nm in the mixture of histidine and sucrose at the same concentration with the lowest average RMSD value. The highest average RMSD value is 0.167 nm occurs in histidine when the concentration is 5% v/v. The highest maximum RMSD value is 0.281 nm in proline when the concentration is 200 mM (1.2% v/v). From the RMSD bar charts, we can tell there is no significant changes between these plots and the simulations are generally stable in all the solvents in all concentrations. Taken the standard deviation into account, even within the replicas of the same cosolvent at certain concentration, the variation between runs sometimes is bigger than between formulations, there is no obvious instability in all simulations.

3.3.2 RMSF

The average RMSF per residue of three replicas were calculated and the flexibility

of the light and heavy chains of fab were examined. In most of the plots, the general trend is the same. Taken the RMSF plots for all the single excipient at paper concentration as an example, the plots are shown below in Figure 14.

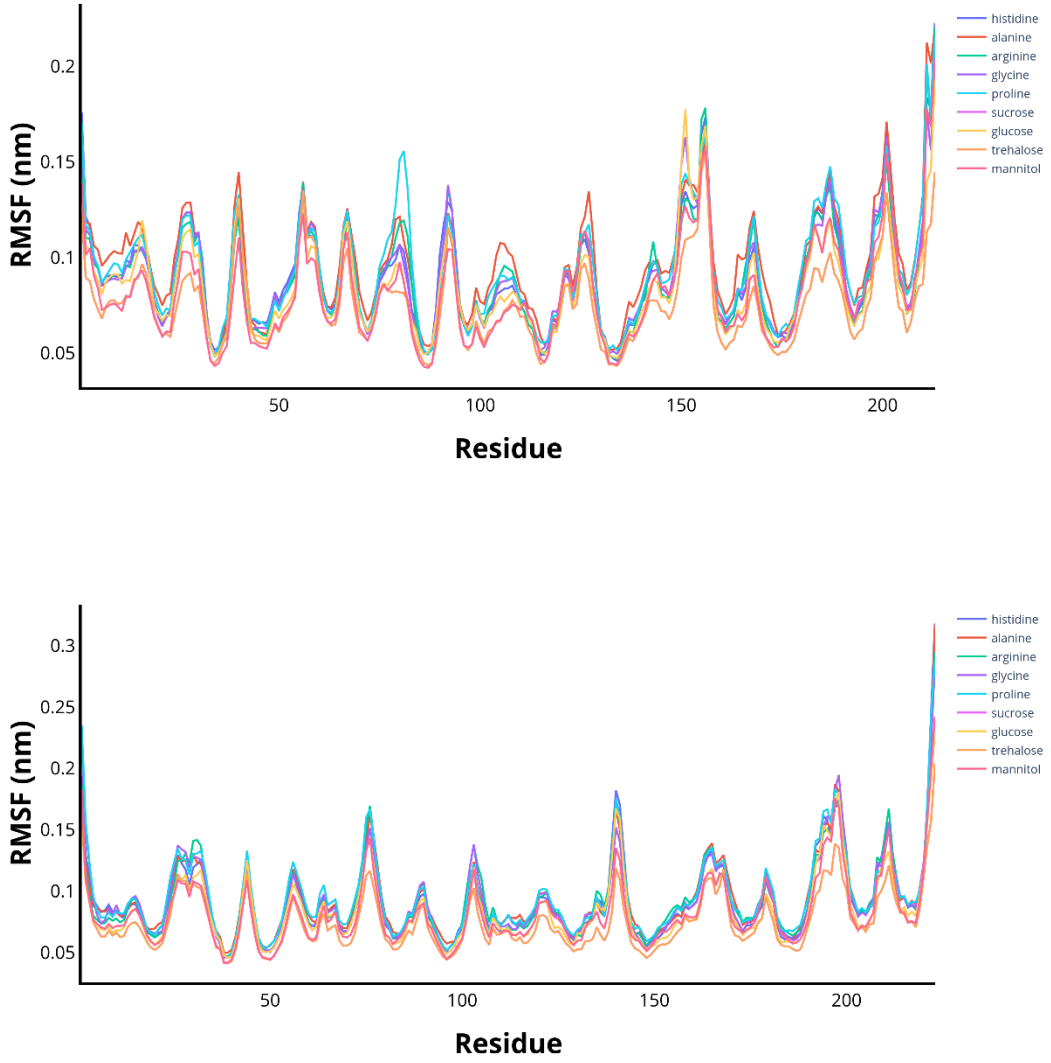


Figure 14: RMSF graphs showing the fluctuations of each residue of the light and heavy chains of the fab region in different solvents at paper concentrations. The top one is the light chain and the bottom one is the heavy chain.

From the RMSF graphs, we can see that the general trend for flexible residues is similar, but in some excipients the fluctuations are higher than the others. For example, the ASP 81 of light chain in proline has a higher peak compared to other

solvents and the ASN 151 of light chain in glucose has a higher peak.

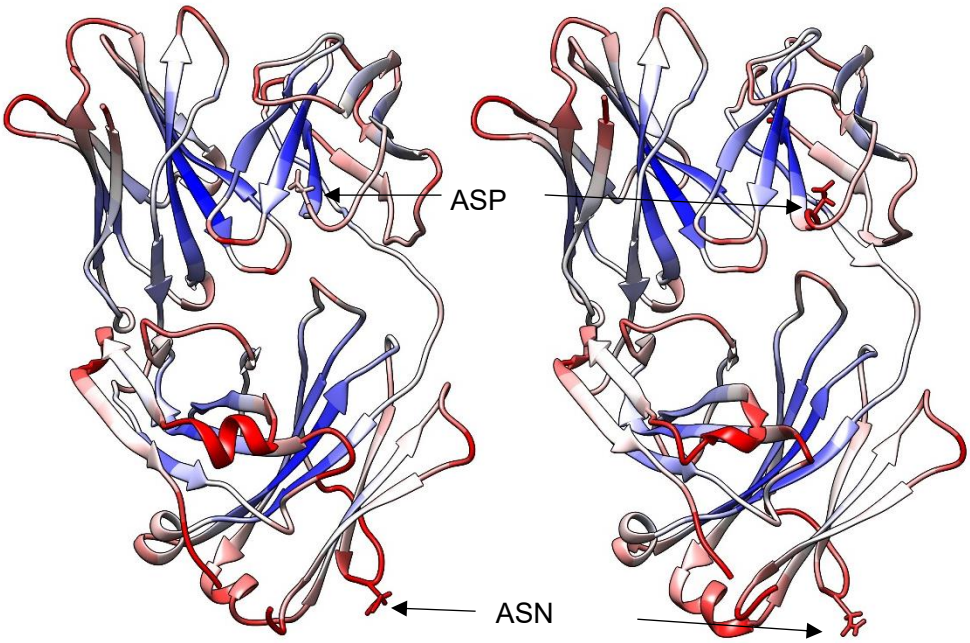


Figure 15: To visualise the RMSF per residue within the protein structure, last frames of the 50 ns simulations in 300mM glucose (left) and 200mM proline (right) were extracted using GROMACS trjconv tool and pictures were generated using chimera and rendered by their RMSF per residue values. Red coloured residues are most flexible and blue coloured residues are most stable ones. White coloured are in the middle range.

From the rendered figures, we can clearly tell the trend of the flexible parts in proline and glucose is similar and for ASP 81, the residue is darker in proline which suggests it is more flexible than in glucose. As for ASN 151, it is darker red in glucose and pink in glucose.

By examine all the RMSF plots for all the simulations, although there are slightly differences in the fluctuation of some residues, the overall stability of the fab in these excipients were not affected much by these fluctuations. The fab is confirmed to be relatively stable in all solvents in all the concentrations and agreed with the RMSD data.

3.3.3 SASA

The next analysis is the solvent accessible surface area (SASA) per residue of the fab region. There were not many differences in the SASA values in different conditions. Examples of SASA plots are shown below in Figure 16.

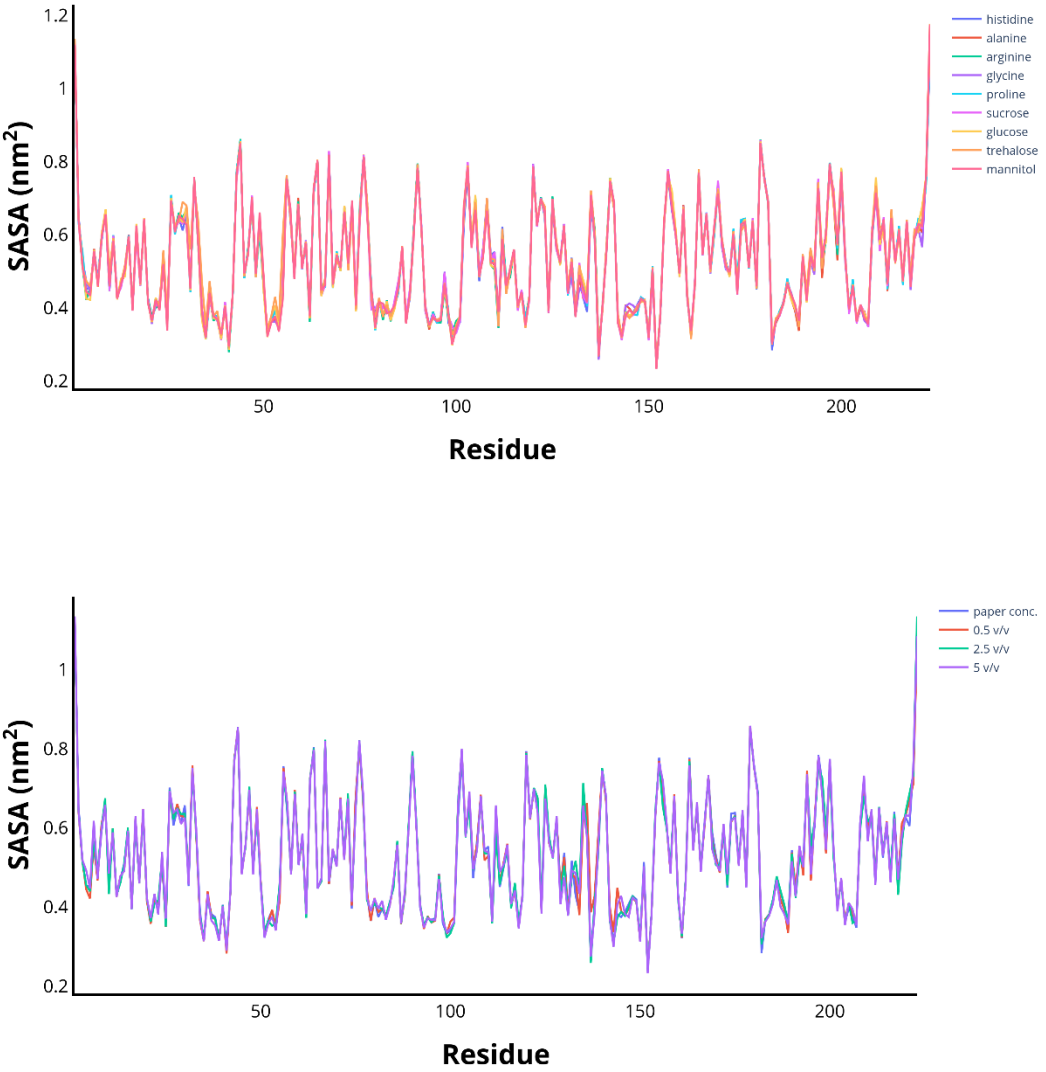


Figure 16: Top one showing SASA per residue of heavy chain in different solvents at the paper concentration, histidine buffer is 25Mm, other amino acids and sugars varied between 171 mM - 300 mM and the bottom one showing SASA per residue of heavy chain in histidine at four different concentrations.

From the SASA per residue graph, we can tell these values are not affected by the type and concentration of solvents. They remained a similar value ignoring the

conditions. This can be due to the immunoglobulin fold, there are local fluctuations but the overall is not changed much.

3.3.4 Radius of gyration

The values for radius of gyration are shown below in Figure 17. Again, this characterisation is not affected much by the type and concentration of solvents.

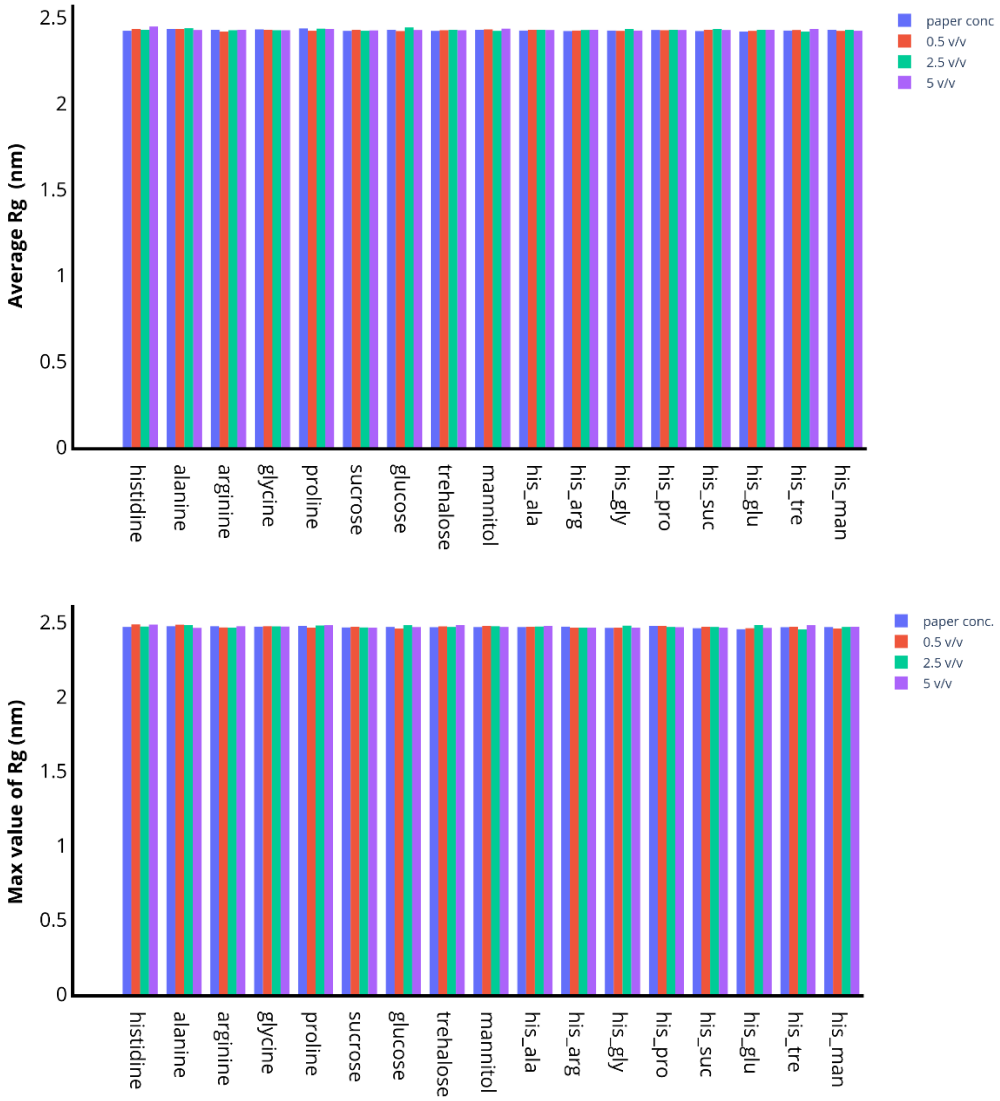
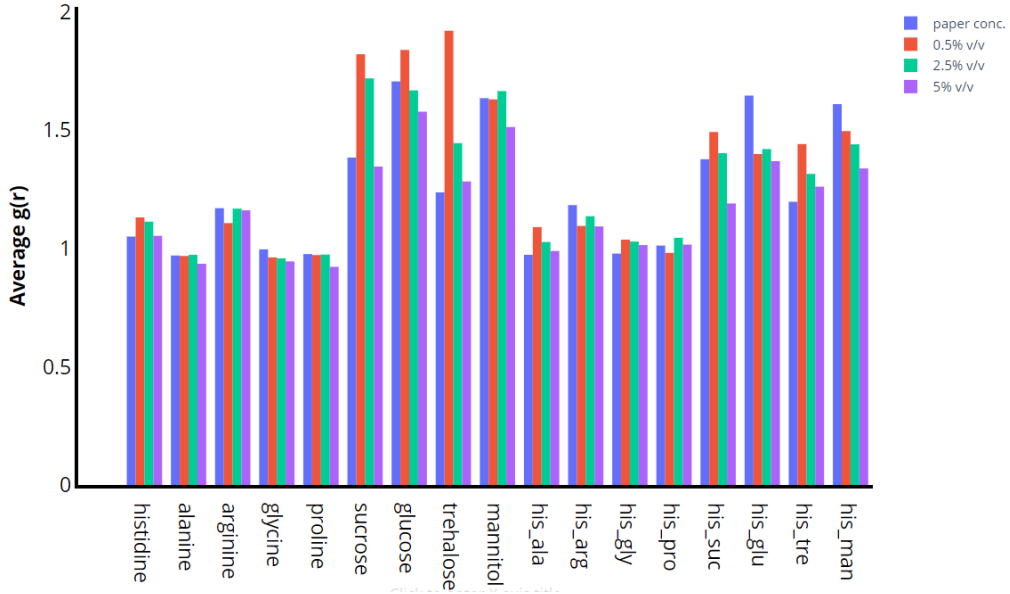


Figure 17: Bar charts for all the solvents in 4 different sets of concentrations. The top one lists the average radius of gyration values for all the replicas and the bottom one shows the maximum values of the radius of gyration values.

From the radius of gyration graphs, we can tell the Fab is quite stable and compact during the simulation ignoring the solvent and concentrations, the average and maximum values of Rg all lie between 2.4 to 2.5nm range.

3.3.5 Radial distribution function

The radial distribution function gives the normalised probability of finding the excipient within certain distances to the reference protein. The average and highest value of RDF is shown below in Figure 18.



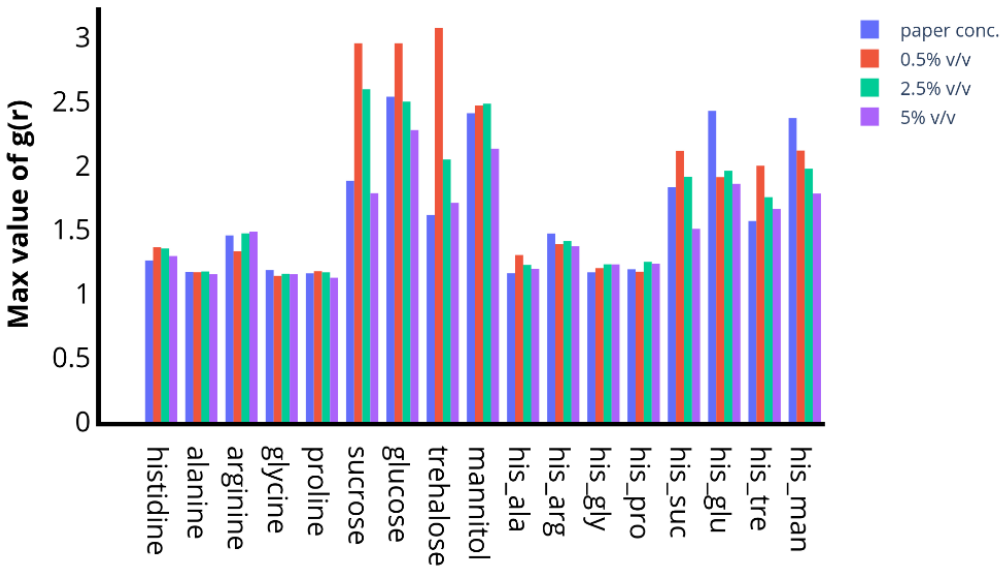


Figure 18: Bar charts for all the single and cosolvents in 4 different sets of concentrations. The top one lists the average RDF values for all the replicas and the bottom one is the average maximum RDF values of all the replicas.

From the RDF values, it is obvious that there is a higher chance of finding sugar molecules around the protein compared with amino acids. The RDF values are higher in single sugars at all the concentrations, particularly at 0.5% v/v, the maximum values reached the highest in single trehalose with a value of 3.08. When there is a mixture of histidine and sugars, the relative distribution value of excipients decreased because there is a less chance of finding the histidine excipient and a higher chance of finding the sugars, both contribute to the total RDF values.

3.3.6 Aggrescan3D

Taking 0.5% v/v histidine, trehalose, and mixture of both as examples, the last frames of the 50ns simulations are shown below, the protein surface is coloured based on their Aggrescan3D score, red areas are the aggregation prone areas.

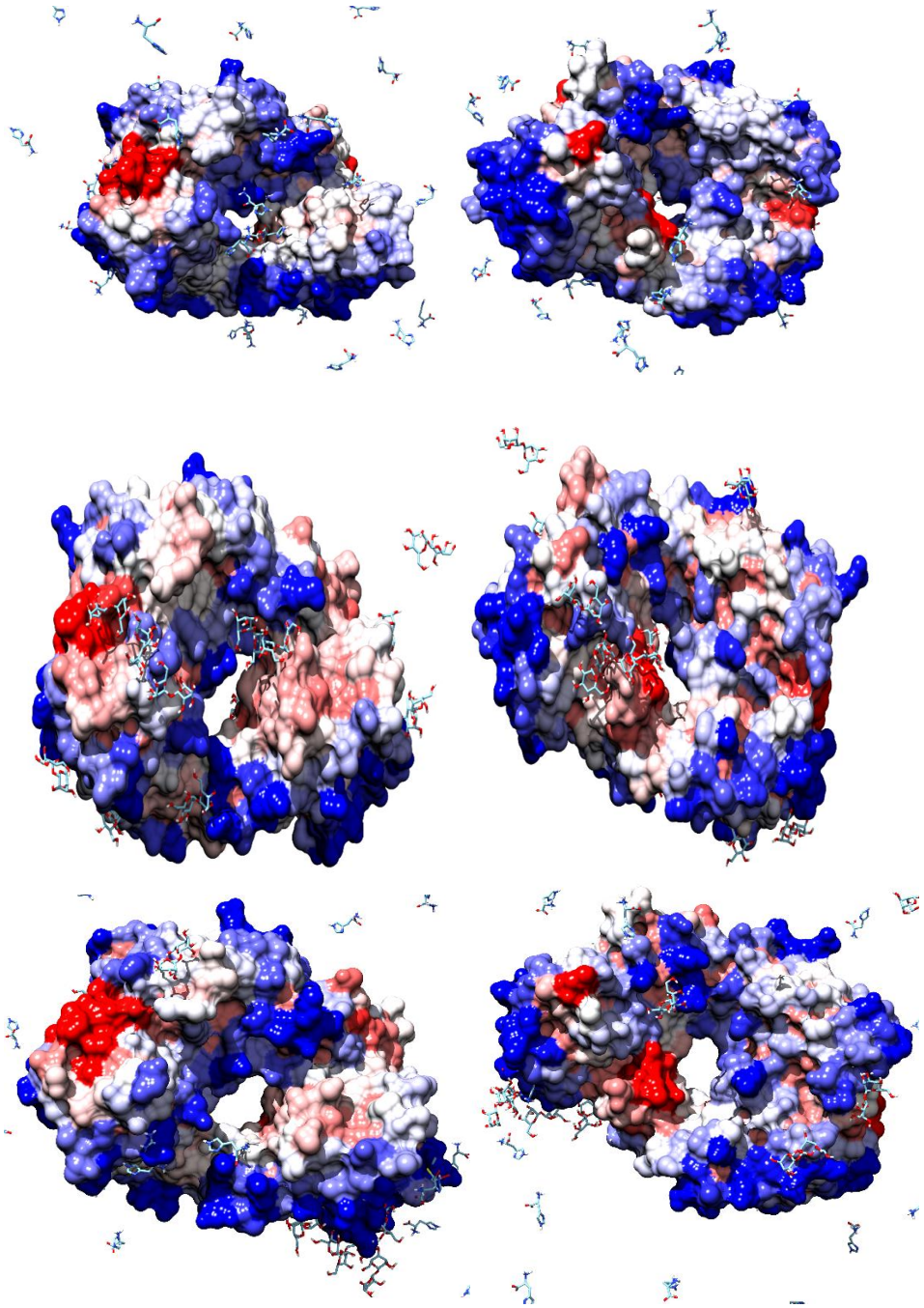


Figure 19: Top two pictures showing the fab in 0.5% v/v histidine; Two pictures in the middle showing the fab in 0.5% v/v trehalose and the bottom two pictures showing the fab in a mixture of histidine (0.25%) and trehalose (0.25%) adding up to 0.5% v/v excipients. The pictures were prepared using chimera and rendered according to the Aggrescan3D score.

From the above pictures, we can see that the predicted aggregation prone areas are similar in all three simulations, the residues in red are 101-106, 175-179 of heavy chain and 93-95 of light chain. The positions of the excipients reflect what was

calculated using RDF, histidine molecules tend to be in the outer region of the protein with only a small amount reached to a closer distance within the protein. On the other hand, most of the trehalose molecules are closer to the protein with only a small amount around the outer region of the protein. And from the mixture of histidine and trehalose, we can also see that histidine molecules are more likely to be a further distance away from the protein compared with trehalose molecules.

3.3.7 Principal component analysis

Principal component analysis (PCA) was also employed to analyse the global motion of all the simulations. First, trajectories of the protein backbone from all 3 replicas were concatenated together and computed using gmx covar. From the computed eigenvectors, PC1 and PC2 structures can be obtained and a 2d plot can also be generated. Examples of 2d plots are shown below in Figure 20 and 21.

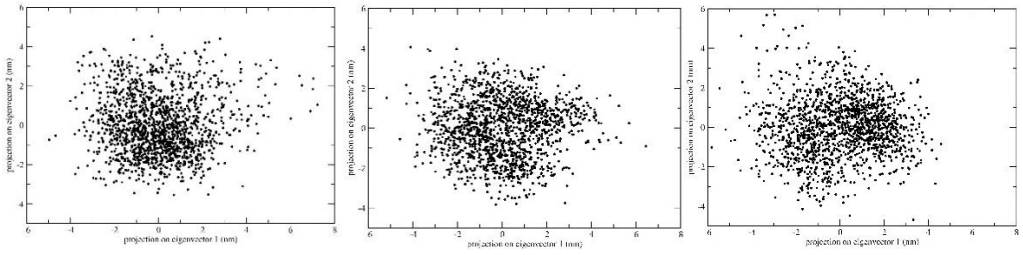


Figure 20: From left to right are 2d PCA plots for fab in histidine, trehalose and mixture of histidine and trehalose at 0.5% v/v.

From the 2d PCA plots above, we can tell this no obvious conformational change towards one direction in all three solvents at 0.5% v/v. However, in some 5% v/v solvents, there is a possible second cluster, for example alanine, which 2d plot and 30 frames from PC1 are shown in Figure 21.

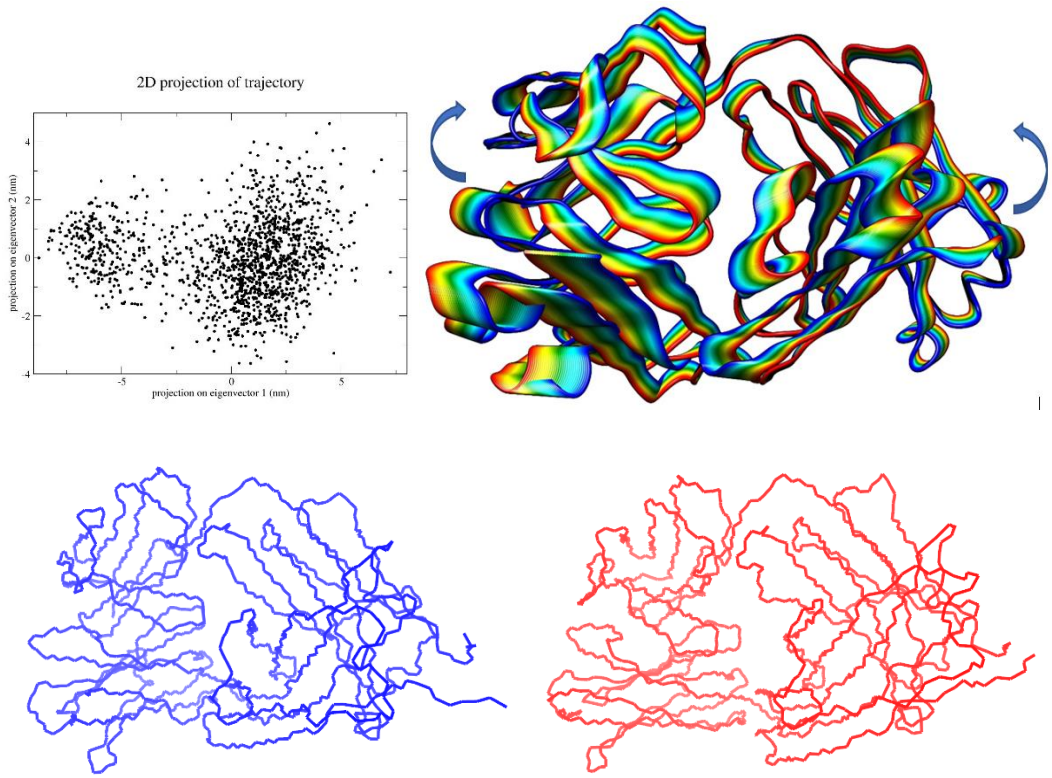


Figure 21: On the top left is the 2d PCA plots for fab in 0.5% v/v arginine and on the right shows 30 frames from PC1 coloured from blue to red showing the movement of this short trajectory. On the bottom shows the first and last frame of those 30 frames from PC1.

When visualising the trajectory of the PC1, the in and out of plane movement corresponding to the two extreme frames of this principal component. Whether this change will lead to further conformational changes or unfolding is not known, however it's unlikely this will cause unfolding of Fab as the secondary structures of the protein are not affected much when looking back the all-atomic simulation. The frames are shown below in Figure 22.

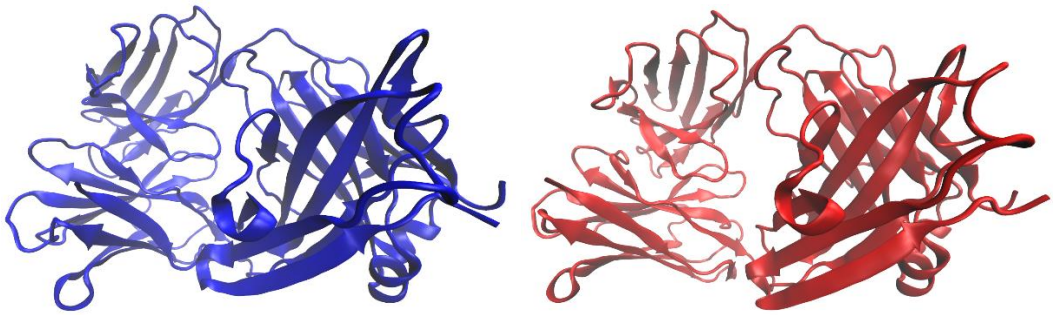


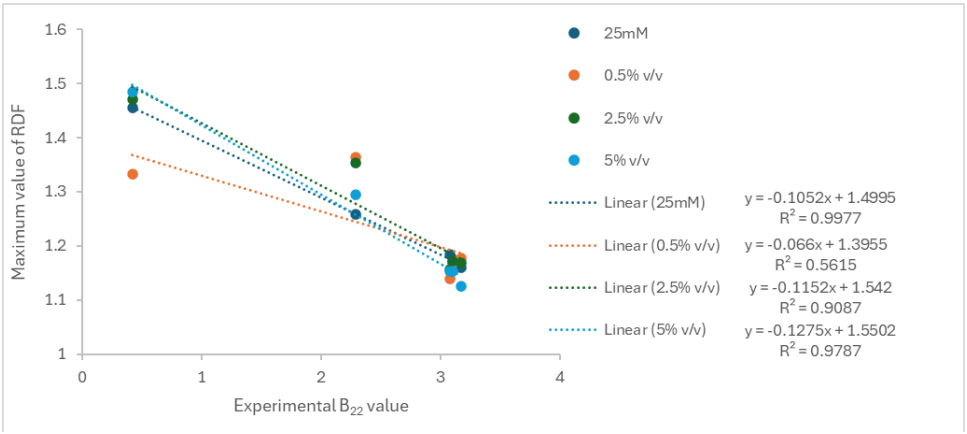
Figure 22: Same as backbone only PCA analysis, 30 frames from PC1 were generated. Then the first and last frame of this atomic analysis of the protein were shown in blue and red respectively.

3.3.8 RDF and B_{22}

From the above analysis, we can tell the Fab is quite stable and compact, it's hardly affected by different types of cosolvent at given concentrations. The only analysis that is worthy to study further is the RDF, so it is used to compared to the experimental B_{22} values. These values are shown below with the maximum RDF values in different concentrations.

	B_{22}(x10^{-4} mol ml/g^2)	Maximum value of RDF			
		25mMHIS	0.5% v/v	2.5% v/v	5% v/v
histidine	2.29±0.055	1.259	1.363	1.353	1.295
his_ala	3.10±0.052	1.161	1.303	1.225	1.194
his_arg	0.42±0.033	1.47	1.389	1.412	1.37
his_gly	3.08±0.104	1.168	1.2	1.23	1.229
his_pro	3.17±0.089	1.191	1.172	1.249	1.234
his_suc	3.86±0.056	1.831	2.115	1.912	1.507
his_glu	3.14±0.033	2.428	1.91	1.959	1.858
his_tre	3.24±0.046	1.569	1.999	1.754	1.662
his_man	2.95±0.061	2.372	2.118	1.976	1.783

	B_{22}(x10^{-4} mol ml/g^2)	Maximum value of RDF			
		25mMHIS	0.5% v/v	2.5% v/v	5% v/v
histidine	2.29±0.055	1.259	1.363	1.353	1.295
alanine	3.10±0.052	1.17	1.169	1.173	1.154
arginine	0.42±0.033	1.455	1.332	1.471	1.485
glycine	3.08±0.104	1.185	1.14	1.156	1.154
proline	3.17±0.089	1.16	1.177	1.168	1.125
sucrose	3.86±0.056	1.882	2.955	2.596	1.785
glucose	3.14±0.033	2.538	2.955	2.499	2.278
trehalose	3.24±0.046	1.616	3.075	2.048	1.71
mannitol	2.95±0.061	2.41	2.468	2.483	2.132



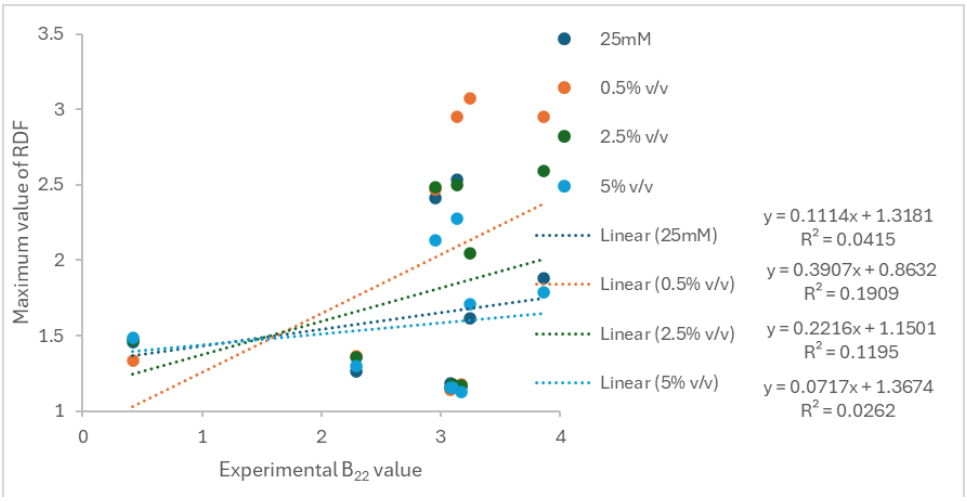


Figure 23: Tables showing the experimental second osmotic coefficient B22 and the maximum values of RDF at different concentrations. The top table is for the cosolvent mixtures, and the bottom table is for single excipients. Graphs showing the possible regression between B22 and maximum RDF. Top one uses only five amino acids, and the bottom one uses all nine excipients.

Looking at the table, for amino acids when there is a lower value of B_{22}, the RDF value is relatively higher. This can be clearly seen for arginine and histidine which have a B_{22} value around 0.42 and 2.29 respectively. When the Fab is in histidine and arginine mixtures, no matter about the concentrations, the maximum RDF value is always higher than in pure histidine. When the Fab in is pure arginine, the maximum RDF value is higher than in the histidine in the paper concentration, 2.5% v/v and 5% v/v. But in 0.5% v/v, the maximum RDF value in arginine is 1.332 which is slightly lower than the 1.363 in histidine. The difference is subtle. Because the value is an average value of three replica, I checked the values for

single replica, for arginine, the RDF values are quite stable with 1.327, 1.327 and 1.343. However, for histidine, the RDF values are in a larger range with 1.310, 1.243 and 1.535, so the 1.535 of the last replicas caused the average to become higher.

For other three amino acids, alanine, glycine and proline, it gets complicated to compare the results because of the similar values among them. The B_{22} range for alanine is 3.048-3.152, glycine is 2.976-3.184 and proline is 3.081-3.259, there is overlapping between their values makes difficult to give them a rank. As for RDF values, these three amino acids give very similar values in all the concentrations as well with a range of 1.13 to 1.18 in pure solvent and 1.16 to 1.3 in cosolvents.

Then considering the sugars, there is no obvious relationship between the experimental B_{22} values and the RDF values. This could be due to the clustering of sugar molecules around the protein surface ignoring the type of sugar molecules, they all have a stabilising effect on the protein.

When all the 9 solvents are considered together, the correlation between RDF and experimental B_{22} is rather weak, but when only 5 amino acids were considered, there seems to be a higher chance of correlation (shown in Figure 23). However, to do a regression analysis, a one in ten rule suggests there should be at least ten predictor parameters. The result will require more validations.

3.3.8 Solvent molecules close to the Fab region

To study the behaviour of cosolvent, I chose 11 frames from each 50 ns trajectory with a 5 ns interval including the first and last frame, and also selected cosolvent molecules within 5A distance of the protein. The purpose of this analysis is to see the movement of the cosolvents at a closer distance to the protein more clearly

and if any molecule stays at the same position throughout the simulation.

Examples of some frames for histidine at 0.5 % v/v are shown below in Figure 24.

And all 11 frames of this simulation are put together in Figure 25.

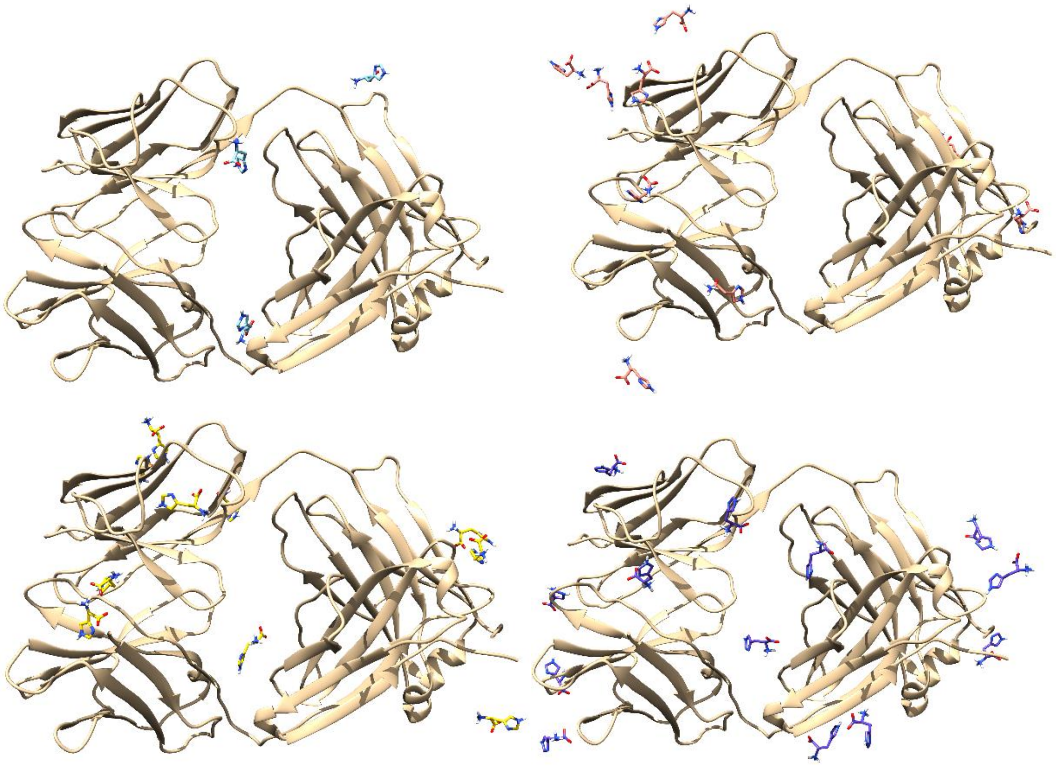
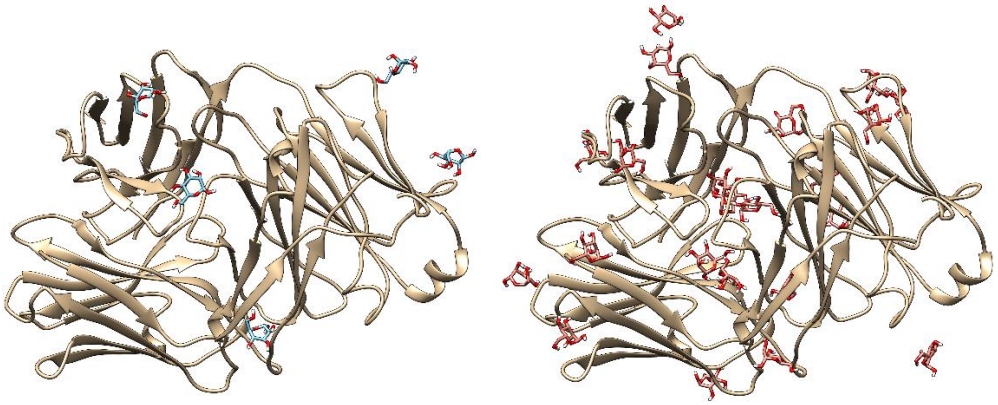


Figure 24: Four frames of the Fab region in replica 1 and the histidine molecules within 0.5nm distance of the protein at the same time are shown above including the Fab region at the beginning of the simulation, at 15ns, at 30ns and 45ns.

Figure 25: Last frame of Fab region in replica 1 and histidine molecules within 0.5nm distance of the protein at different timesteps from 0 to 50ns at a 5ns interval. Each colour represents a different timestep with a total of 11 colours.

From the trajectories and the frames obtained, these show that the movement of amino acid cosolvents were quite random, it's rare to see one amino acid molecule that stays at the same position throughout the simulation. Then examples of some frames for glucose at 0.5 % v/v are shown below in Figure 26. And all 11 frames of this simulation are put together in Figure 27.



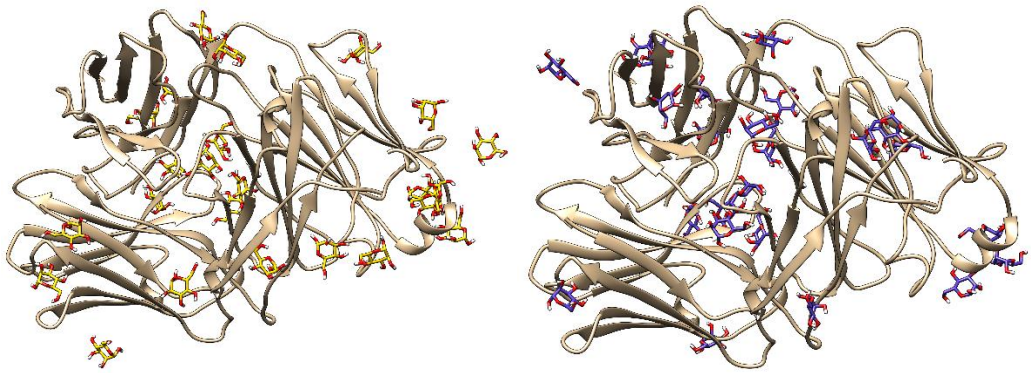


Figure 26: Frames of the Fab region in replica 1 and glucose molecules within 0.5nm distance of the protein at the same time are shown above including the Fab region at the beginning of the simulation, at 15ns, at 30ns and 45ns.

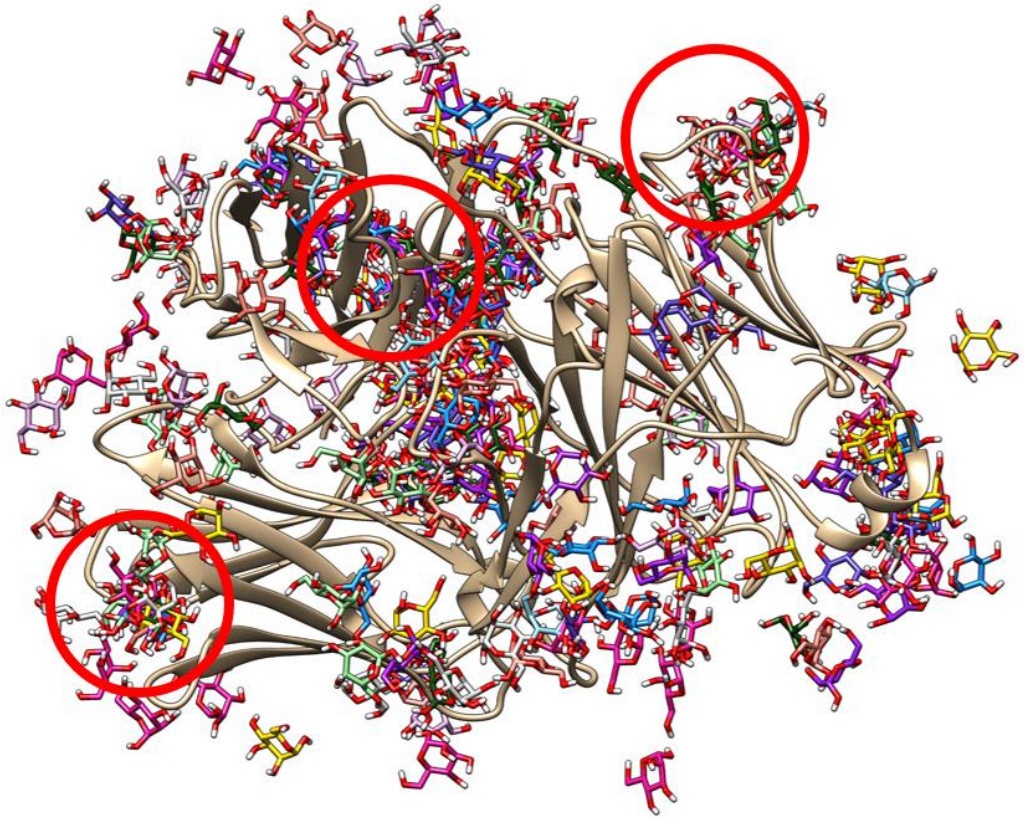


Figure 27: Last frame of Fab region in replica 1 and glucose molecules within 0.5nm distance of the protein at different timesteps from 0 to 50ns at 5ns interval. Each colour represents a different timestep with a total of 11 colours. Three red circles are examples of glucose stays at the same place for a longer time and these regions are enlarged in Figure 28.

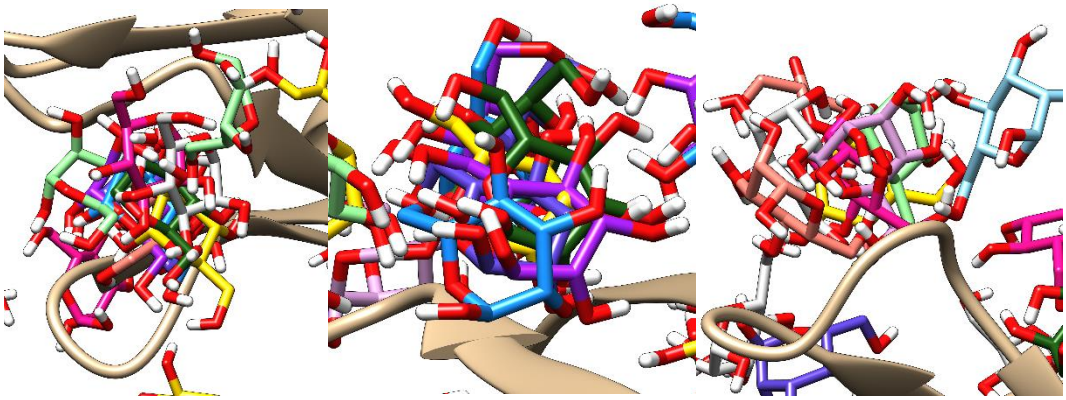


Figure 28: Sections from Figure 27 that showing examples of clusters of glucose molecules at different timesteps.

From the trajectories and the frames obtained, these show that the movement of sugar molecules were also random during the simulation, and sugar molecules seem to stay longer at one position throughout the simulation. But there is also a chance of moving away from that position.

Following the generated frames, the number of cosolvents within 0.5 nm of fab region are counted and put in the table below. The numbers at each time step are the average value of all three replicas.

Time (ns) / Excipient	0	5	10	15	20	25	30	35	40	45	50
histidine	4	8	11	11	9	11	11	8	12	14	13
alanine	4	14	11	11	11	11	11	14	10	12	11
arginine	5	7	8	10	9	8	9	9	12	7	11
glycine	8	12	9	11	11	12	13	12	10	12	9
proline	5	13	7	11	12	13	11	10	13	11	13
sucrose	1	8	9	10	13	14	14	14	15	15	15
glucose	7	16	18	21	20	23	18	21	23	24	25
trehalose	2	9	10	13	13	15	16	17	16	17	17
mannitol	5	11	14	17	18	17	18	19	21	20	18

Time (ns) / Excipient	0	5	10	15	20	25	30	35	40	45	50
his_ala(HIS)	2	6	8	7	6	6	6	6	8	10	6
	3	5	7	4	7	6	5	7	7	6	8
his_arg(HIS)	3	5	7	6	7	6	7	6	6	5	5
	1	5	6	4	6	3	5	6	5	7	5
his_gly(HIS)	2	4	5	5	6	6	7	4	7	5	5
	4	6	9	8	8	7	7	10	6	7	7
his_pro(HIS)	2	5	4	7	5	5	3	4	5	7	6
	2	5	7	4	3	4	4	5	3	3	5
his_suc(HIS)	1	4	4	4	6	5	6	5	9	6	6
	1	4	5	5	7	7	6	6	6	7	7
his_glu(HIS)	1	5	4	8	7	5	7	6	7	6	7
	2	5	7	9	12	11	10	8	12	11	11
his_tre(HIS)	5	4	6	5	8	5	6	8	7	6	7
	1	4	5	6	6	7	7	8	8	7	8
his_man(HIS)	2	6	8	6	8	7	6	5	7	7	5
	3	8	8	9	9	9	9	11	9	10	10

Figure 29: The table on the top is the number of different solvent molecules at 0.5% v/v pure solvent and the bottom is the number of different solvent molecules at 0.5% v/v of a mixture of histidine with other molecules in a 1:1 ratio (0.25% v/v of each).

For histidine, the number of solvent molecules increased in the first 10 ns then plateaued, at 20 ns and 35 ns, it decreased but quickly increased again and reached 14 at 45 ns. For alanine, the number increased from 4 to 14 in the first 5 ns, then decreased to 11 and plateaued until 30 ns, in the last 20 ns the number fluctuates. For arginine, the number increased from 5 to 10 in the first 15 ns, then plateaued with 9 molecules until 35 ns and increased to 12 at 40 ns. For glycine, the number increased from 8 to 12 in the first 5 ns, then fluctuated to a maximum of 13 at 30 ns and fluctuated again. For proline, it's similar in the first 5 ns, the number increased from 5 to 13, then fluctuated between 10 to 13 between 15 to 50 ns. In general, the number of amino acid molecules reached to a maximum value around 12 to 14 during the first 15 ns and then plateaued or fluctuated within a small degree.

For sucrose, the number of molecules increased to 15 at 40 ns and plateaued. For glucose, the number increased during the first 15 ns then fluctuated and reached to 25 at the end of the simulation. For trehalose, the number increased to 17 at 35 ns and then plateaued. For mannitol, the number increased to 18 during the first 20 ns then fluctuated and reached to 21 at 40 ns. For sugar molecules, the numbers of molecules are generally higher than amino acids. Sucrose and trehalose are

disaccharides, so the numbers are smaller compared to monosaccharides - glucose and mannitol. The number of glucose molecules are higher than mannitol molecules, which reason is not well understood.

For cosolvents, the number of amino acid molecules have a similar trend to single solvent. As the concentration is 0.25% of each solvent, the maximum number of molecules varied from 7 to 10 which is a reasonable range. The maximum number of sugar molecules within 5 Å of the protein in the mixture varied from 8 to 12, again glucose molecules are the highest among them.

When the concentration increases from 0.5% v/v to 2.5% v/v then to 5% v/v, the maximum number of solvent molecules within 5 Å of protein increased to 50s and 90s for amino acids and 70s and 100s for sugars, makes counting them more difficult, examples are in the Figure 30 and 31 below.

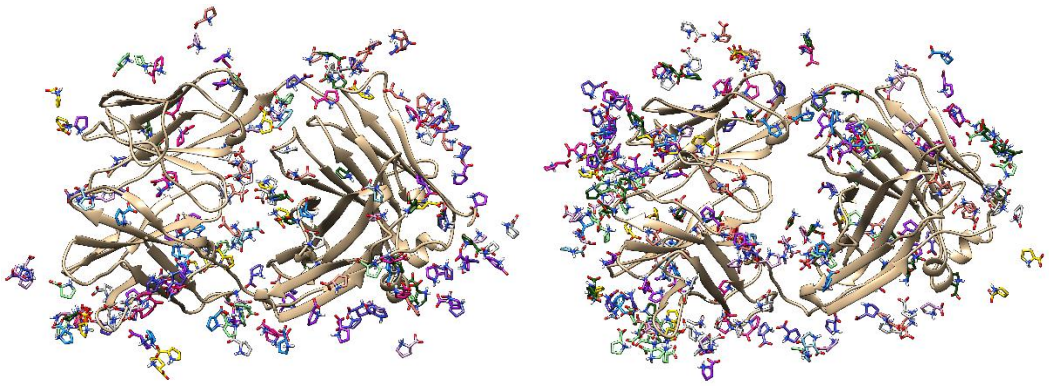


Figure 30: Last frame of fab region of replica 1 in 2.5% v/v and 5% v/v and proline molecules within 0.5 nm distance of the protein at different timesteps from 0 to 50ns at 5ns interval shown in different colours.

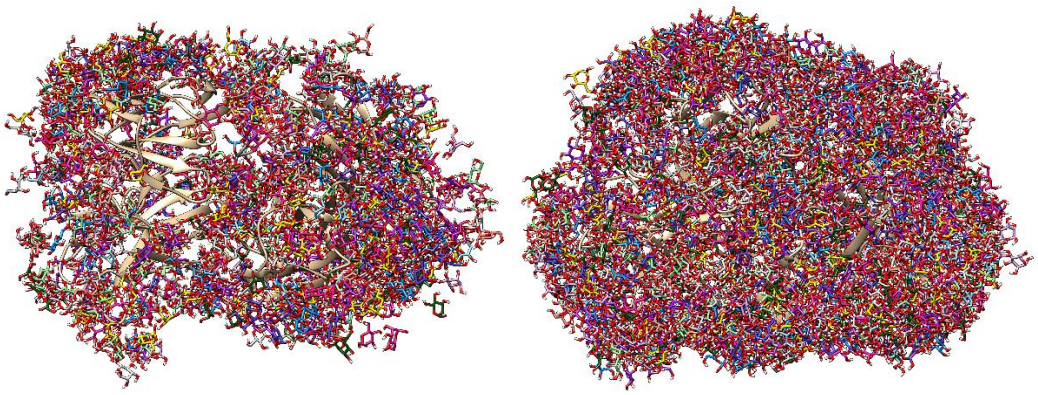


Figure 31: Last frame of fab region of replica 1 in 2.5% v/v and 5% v/v and glucose molecules within 0.5nm distance of the protein at different timesteps from 0 to 50ns at 5ns interval shown in different colours.

Even with the increased concentration, solvent molecules still moved around the Fab in a random way. From Figure 30, we can tell the number of proline molecules within 5 Å of the protein increases when the concentration increases at different timesteps and there are no preferential positions the molecules tend to stay throughout the simulation. From Figure 31, we can see the sugar molecules seem to cluster which is the same as 0.5% v/v and as the concentration increases, more and more molecules are clustered around the protein surface.

3.4 Simulations and analysis of whole mAb

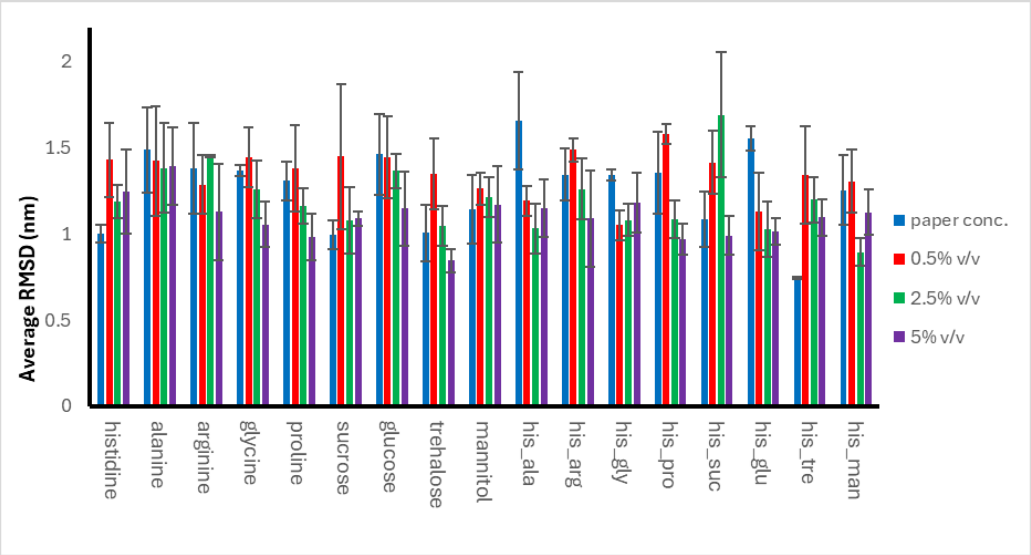
The constant region of IgG antibodies are highly conserved, only variable regions that bind to antigens are distinct for each antibody. By running the whole mAb region as a comparison, we want to check if the Fab region can be a reasonable model for the whole mAb. The sequence of whole mAb is much longer than the Fab region, time used to run simulation increases a lot. The stability of whole mAb should be slightly lower than the Fab region due to the flexibility of the hinge region so the RMSD values are expected to be higher.

Figure 32: The sequence alignment between NISTmAb and b12 protein. a) Sequence alignment of light chain. b) Sequence alignment of heavy chain.

3.4.1 RMSD

The discussion in the following pages is about the whole mAb (shown in previous figure 9c). Same as the fab region, simulations were performed three times in every condition, and the maximum and average value of all the frames were calculated

for every excipient at each concentration. The average and maximum values of the RMSD plots are shown below.



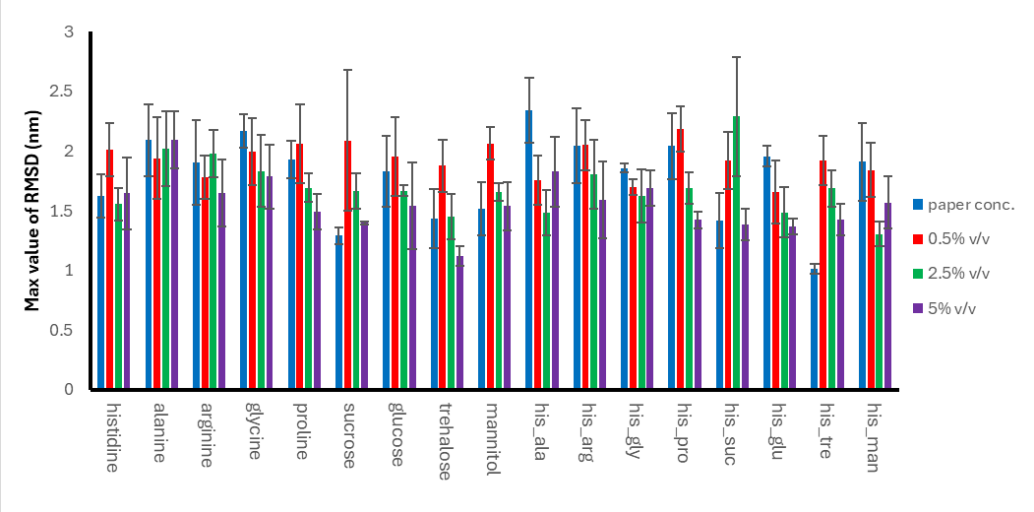


Figure 33: Bar charts for all the solvents in 4 different sets of concentrations. The top one lists the average RMSD values for all the replicas and the bottom one is the average maximum values of the RMSD for all the replicas with standard deviations shown as error bars.

The average values lie between 0.744 to 1.692 nm and the maximum values lie between 1.018 to 2.342 nm. Due to the flexibility of the hinge region, the RMSD values are a bit high, but they are generally quite stable. The lowest average RMSD value is 0.744 nm occurs in a mixture of histidine and trehalose when the concentration is 25 mM histidine and 300 mM trehalose. The lowest maximum

RMSD value is 1.018 nm in the same cosolvent as the lowest average RMSD value. The highest average RMSD value is 1.692 nm occurs in a mixture of histidine and sucrose when the concentration is 2.5% v/v. The highest maximum RMSD value is 2.342 nm in a mixture of histidine and alanine when the concentration is 25 mM histidine and 200 mM alanine. From the RMSD bar charts, we call see when the concentration is 5% v/v, the average RMSD values are lower in arginine, glycine, proline, glucose, trehalose, histidine with arginine, histidine with proline, histidine with sucrose and histidine with glucose compared to other concentrations even when error bars are considered. The average RMSD values of alanine and mannitol are not affected much by the concentrations, the error bars are also in a similar range. For some cosolvents, the error bars are quite big and for most of the cosolvents at each concentration, the RMSD values and error bars are larger than Fab region only.

3.4.2 RMSF

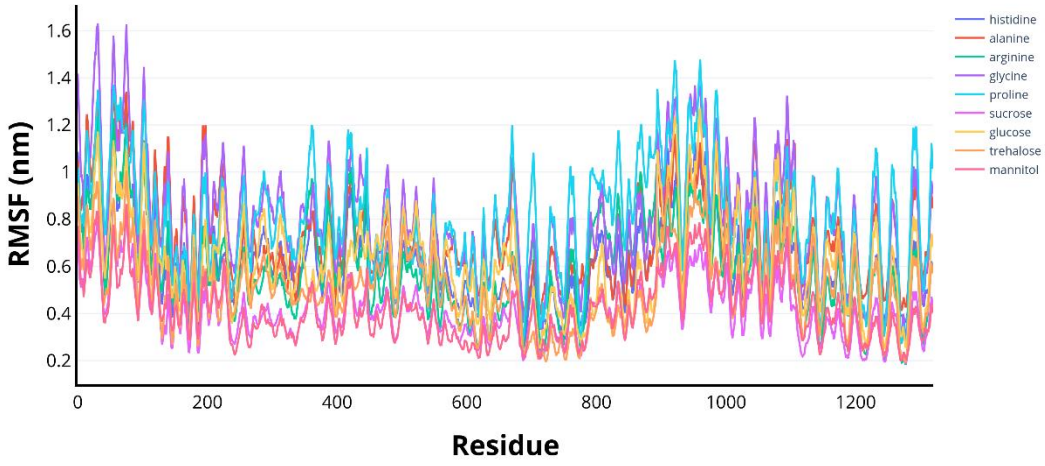


Figure 34: RMSF graph showing the fluctuations of each residue of the whole mAb in different solvents at paper concentration, 25mM histidine buffer and 171mM – 300 mM for other amino acids and sugars.

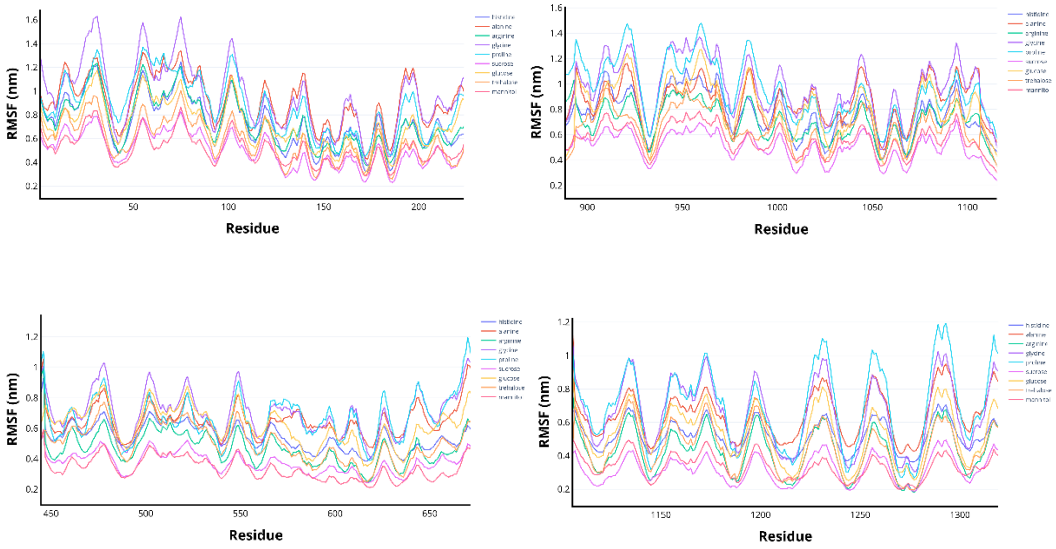


Figure 35: A close look of the RMSF plot shown in Fig.34. Top two plots showing one of the Fab regions and bottom two showing the other one with heavy chain on the left and light chain on the right.

From the RMSF graphs, we can see that the general trend for flexible residues is similar, but in some excipients the fluctuations are higher than the others. For example, in proline, some residues in the light chain and the Fc region have higher peaks.

In proline, the RMSF range varies from 0.27 to 1.48, and for one fab region the range is 0.45 to 1.48 and the other is 0.27 to 1.19. Residue SER921 (27), ARG922 (28), VAL 923 (29), GLY959 (65), SER960 (66), GLY984 (90), SER985 (91) and GLY 986 (92) of light chain have higher RMSF values in one fab region in proline. LEU1231 (124), LYS1232 (125), SER1233 (126), VAL1256 (149), ASP1257 (150), ASN1258 (151), LYS1289 (152), ALA1290 (153), ASP1291 (154), TYR1292 (155), GLU1293 (156) and LYS1294 (157) are flexible in the other fab region in proline.

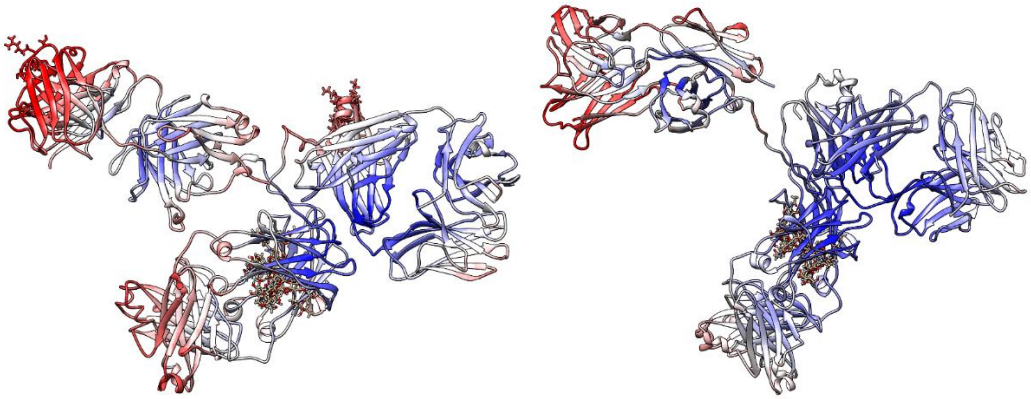


Figure 36: To visualise the RMSF per residue within the protein structure, last frames of the 50 ns simulations in 200mM proline (left) and 300mM glucose (right) were extracted using GROMACS trjconv tool and pictures were generated using chimera and rendered by their RMSF values. Red coloured residues are most flexible and blue coloured residues are most stable ones. White coloured ones are in the middle range.

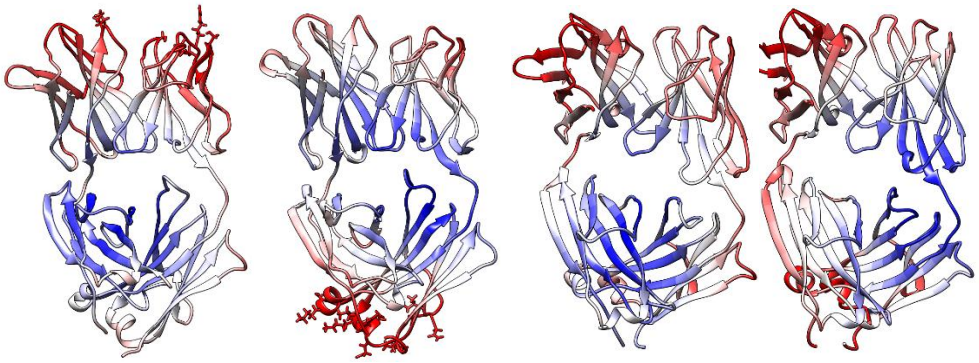


Figure 37: Closer look of the two Fab regions of the last frames of the simulations in proline (left) and sucrose (right) in Figure 36.

In sucrose, the RMSF range varies from 0.19 to 0.83, and for one fab region the range is 0.23 to 0.83 and the other is 0.19 to 0.52. We can tell from the figures that the general trend in both sucrose and proline is similar. But the degree of fluctuation is smaller in sucrose.

3.4.3 SASA

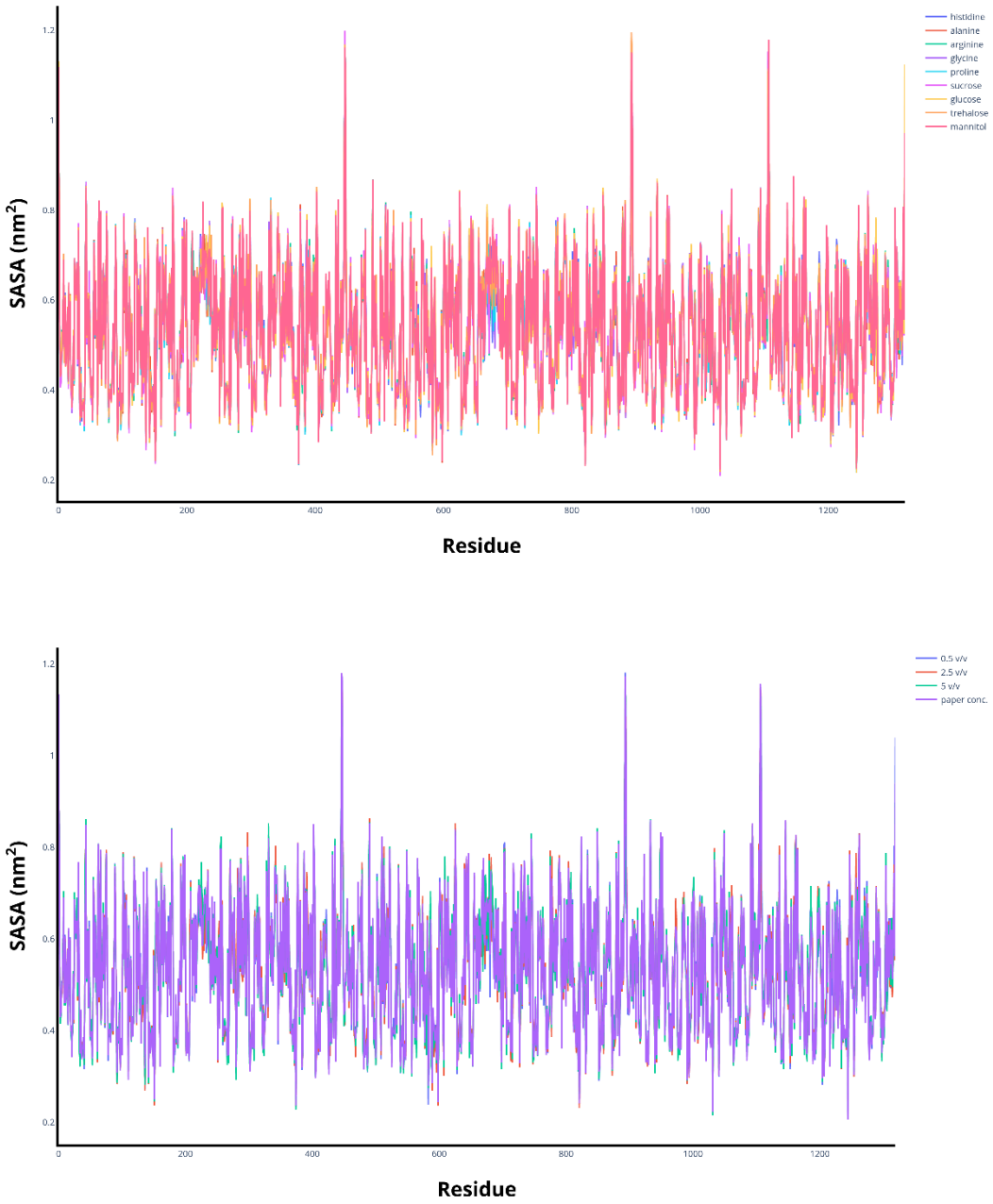


Figure 38: Top one showing SASA per residue of the whole mAb in different solvents at the same concentration and the bottom one showing SASA per residue of the mAb in a mixture of histidine and trehalose in four different concentrations.

From the SASA per residue graph, we can tell these values are again not affected by the type and concentration of solvents, same to the SASA for Fab region only. They remained a similar value ignoring the conditions.

3.4.4 Radius of gyration

The radius of gyration values for the whole mAb are shown below in Fig.39. These values are not affected much by the type and concentrations of excipients. In the paper by Xu and colleagues, the Rg values are also quite similar between different excipients. The values lie between 4.5 to 5 nm.

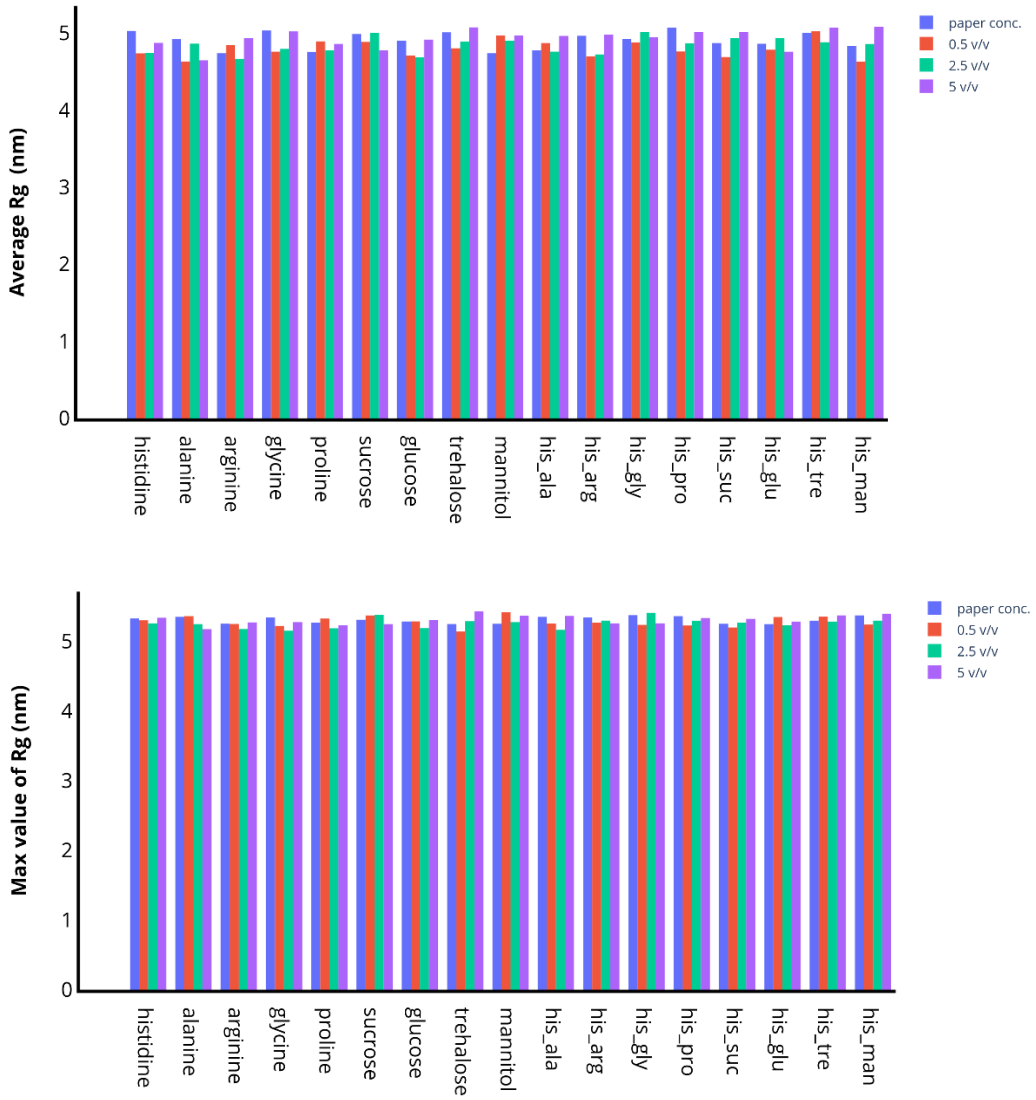


Figure 39: Bar charts for all the solvents in 4 different sets of concentrations. The top one lists the average radius of gyration values for all the replicas and the bottom one is the average maximum values of the radius of gyration for all the replicas.

From the radius of gyration graphs, we can tell the mAb is quite stable and compact

during the 50 ns simulation. The average values of Rg lie between 4.6 to 5 nm and maximum values of Rg lie between 5.1 to 5.5 nm range.

3.4.5 Radial distribution function

The average and highest value of RDF for the whole mAb is shown below in Fig.40. The general trend is similar to the fab region only with 0.5% v/v gives the highest RDF values in single sugar excipients. And sugars in general lead to higher RDF values when they are in a mixture with histidine residues.

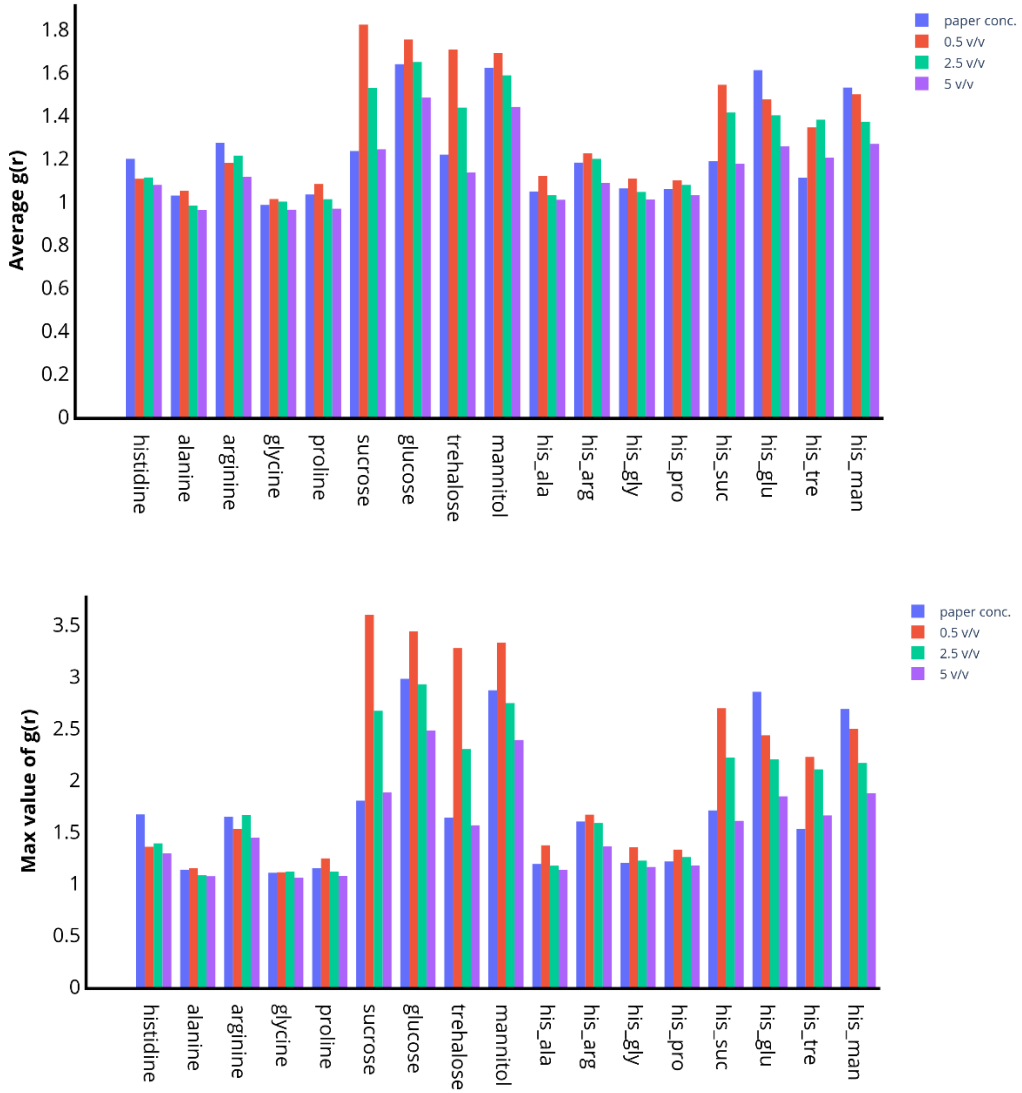


Figure 40: Bar charts for all the solvents in 4 different sets of concentrations. The top one lists the average RDF values for all the replicas and the bottom one is the average maximum values of the RDF for all the replicas.

From the RDF values, it is obvious that there is a higher chance of finding sugar molecules around the whole mAb compared with amino acids. The RDF values are higher in single sugars at all the concentrations, particularly at 0.5% v/v, the maximum values reached the highest in single sucrose with a value of 3.5. When there is a mixture of histidine and sugars, the relative distribution value of excipients also decreased.

3.4.6 Principal component analysis

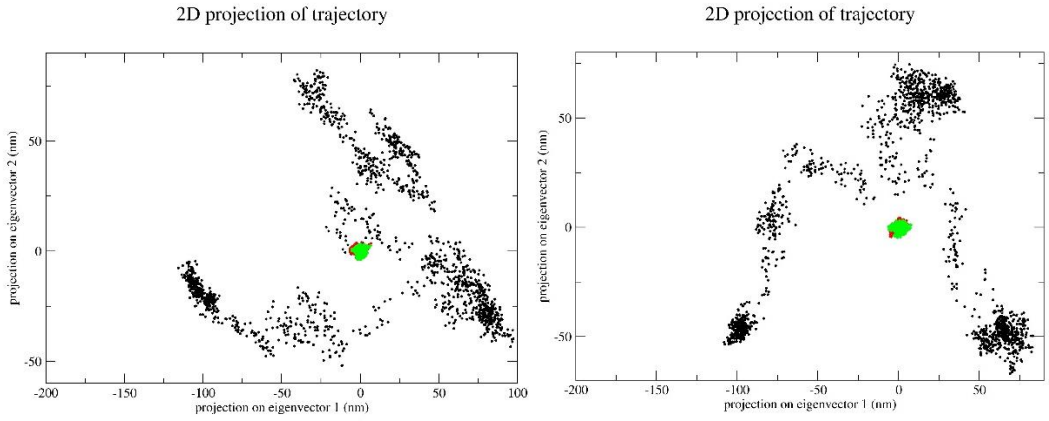


Figure 41: From left to right are 2d PCA plots for the mAb in 0.5% v/v histidine and glucose. Black dots are frames of the whole mAb, green and red dots are frames of the two Fab regions.

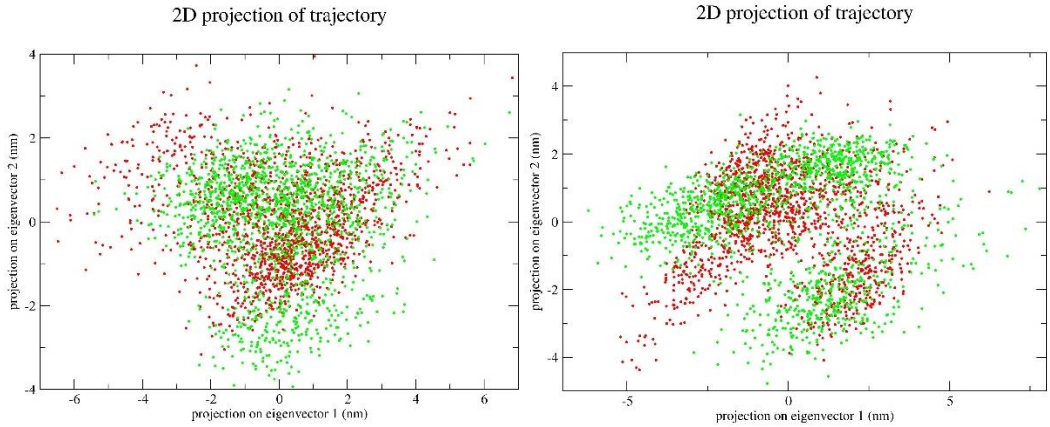


Figure 42: From left to right are 2d PCA plots for the Fab regions in 0.5% v/v histidine and glucose. Green dots represent one Fab region and red dots represent another Fab region.

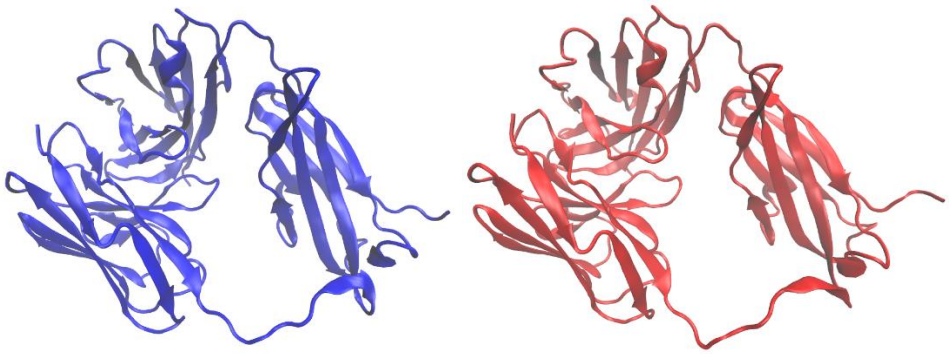


Figure 43: From the eigenvectors, 30 frames were generated. The first and last frame from PC1 of one of the Fab regions of those 30 frames in 0.5% v/v histidine are shown above in blue and red respectively.

From the PCA plots, we can tell that when the mAb was considered as a whole, the projections spread to a large area and when two Fab regions were considered separately, the points are more concentrated in the middle. From two frames from PC1 of one Fab region, we can see the Fab is quite stable and there is no obvious conformational change throughout the simulation.

3.4.7 Aggrescan3D

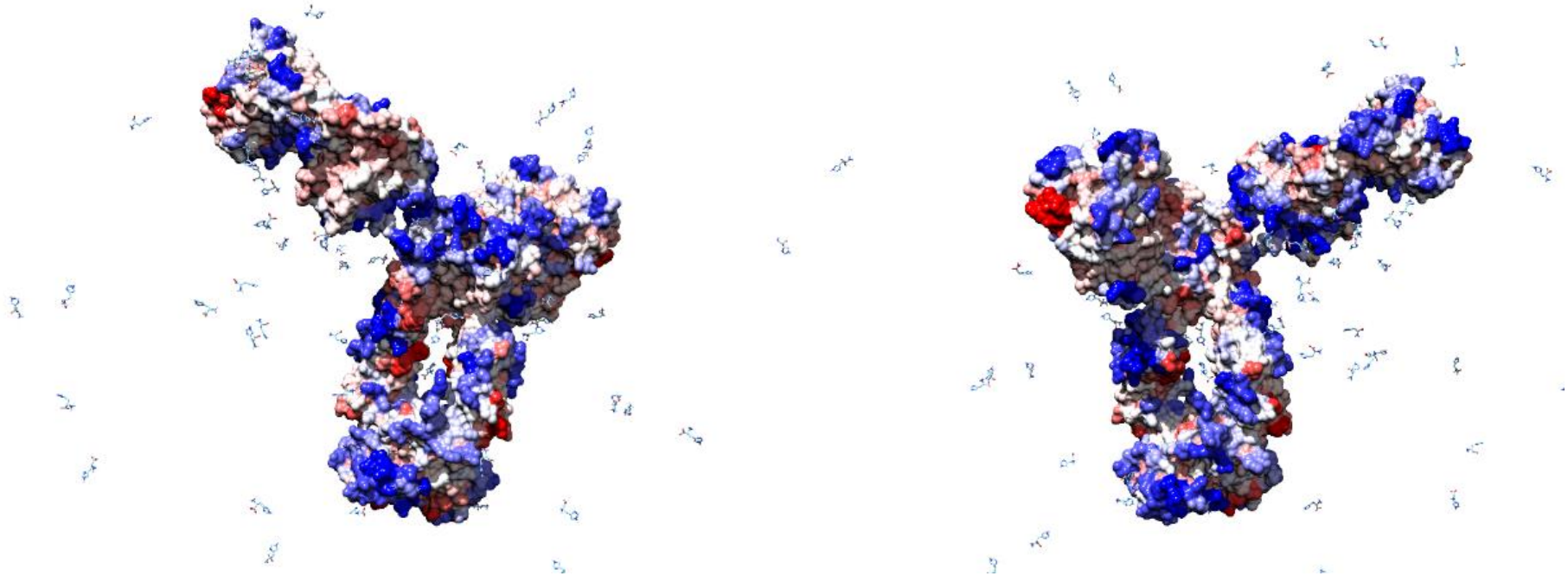


Figure 44: Two pictures showing the mAb in 0.5% v/v histidine with surface shown and rendered by Aggrescan score. The one on the right is obtained from rotating 180° of the structure on the left.

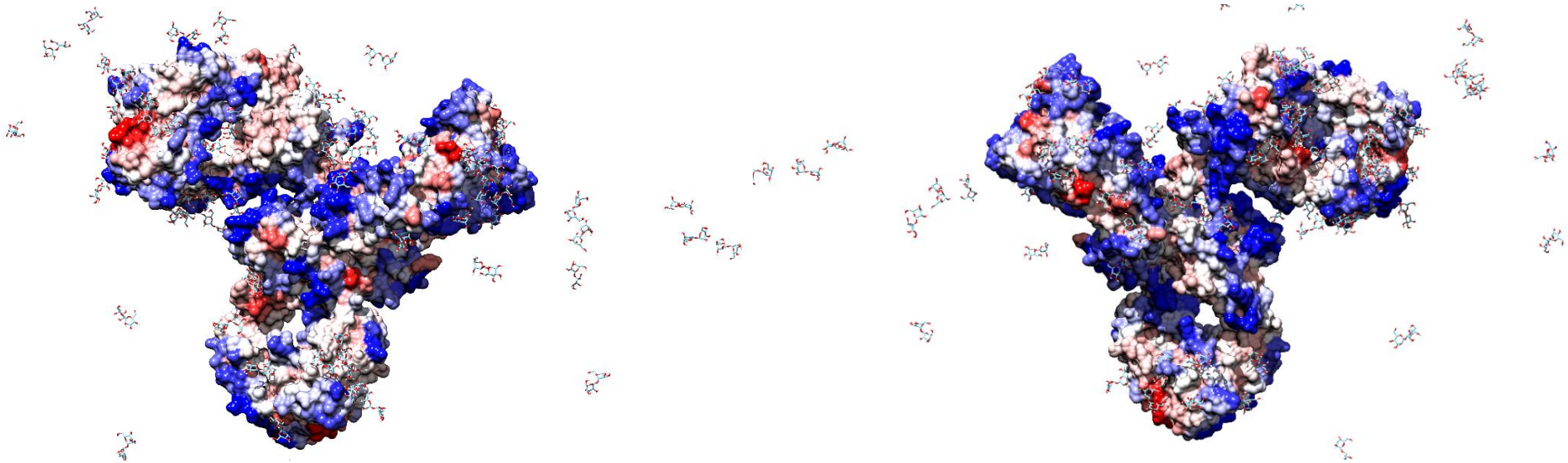


Figure 45: Two pictures showing the mAb in 0.5% v/v trehalose with surface shown and rendered by Aggrescan score. The one on the right is obtained from rotating 180° of the structure on the left.

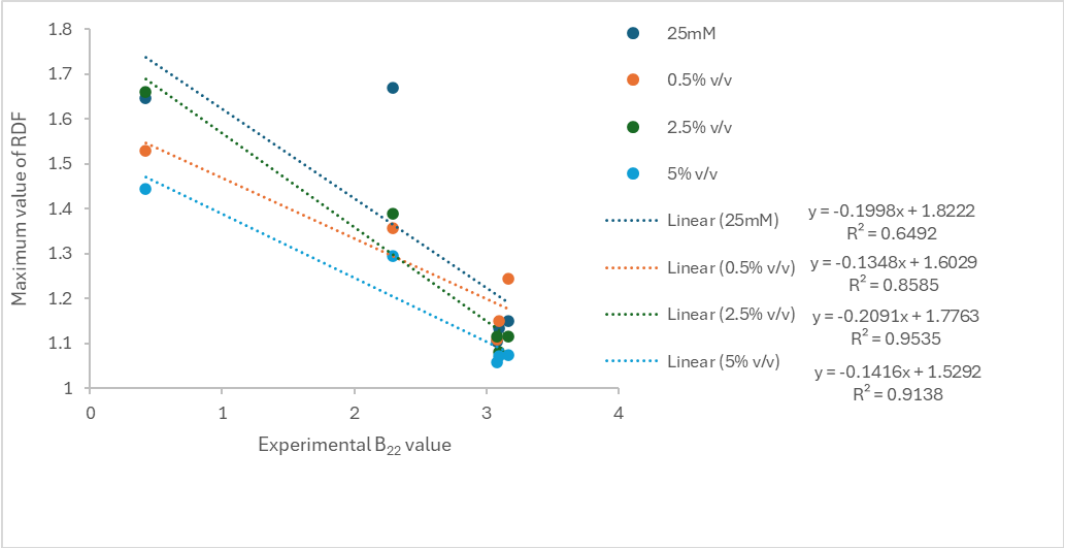
Same as the results of the fab region, histidine molecules are more likely floating around the outer region of the protein with only a small amount reached to a closer distance within the protein. On the other hand, most of the trehalose molecules are closer to the protein with only a small amount around the outer region of the protein. And from the mixture of histidine and trehalose, we can also see that histidine molecules are more likely to be a further distance away from the protein compared with trehalose molecules.

3.4.8 RDF and B_{22}

The experimental B_{22} and the maximum RDF values in different concentrations are shown below in Figure 46.

		25mMHIS	0.5% v/v	2.5% v/v	5% v/v
histidine	2.29	1.669	1.356	1.389	1.294
alanine	3.1	1.133	1.149	1.081	1.071
arginine	0.42	1.646	1.53	1.661	1.443
glycine	3.08	1.104	1.109	1.114	1.057
proline	3.17	1.149	1.245	1.115	1.074
sucrose	3.86	1.802	3.598	2.671	1.88
glucose	3.14	2.981	3.442	2.928	2.481
trehalose	3.24	1.637	3.278	2.301	1.563
mannitol	2.95	2.87	3.329	2.745	2.388

	$B_{22}(x10^{-4}$ mol ml/$g^2)$	Maximum value of RDF			
		25mMHIS	0.5% v/v	2.5% v/v	5% v/v
his_ala	3.1	1.189	1.37	1.173	1.132
his_arg	0.42	1.6	1.666	1.586	1.361
his_gly	3.08	1.202	1.351	1.222	1.161
his_pro	3.17	1.214	1.329	1.256	1.174
his_suc	3.86	1.708	2.696	2.218	1.606
his_glu	3.14	2.855	2.434	2.202	1.843
his_tre	3.24	1.529	2.224	2.104	1.659
his_man	2.95	2.688	2.497	2.165	1.876



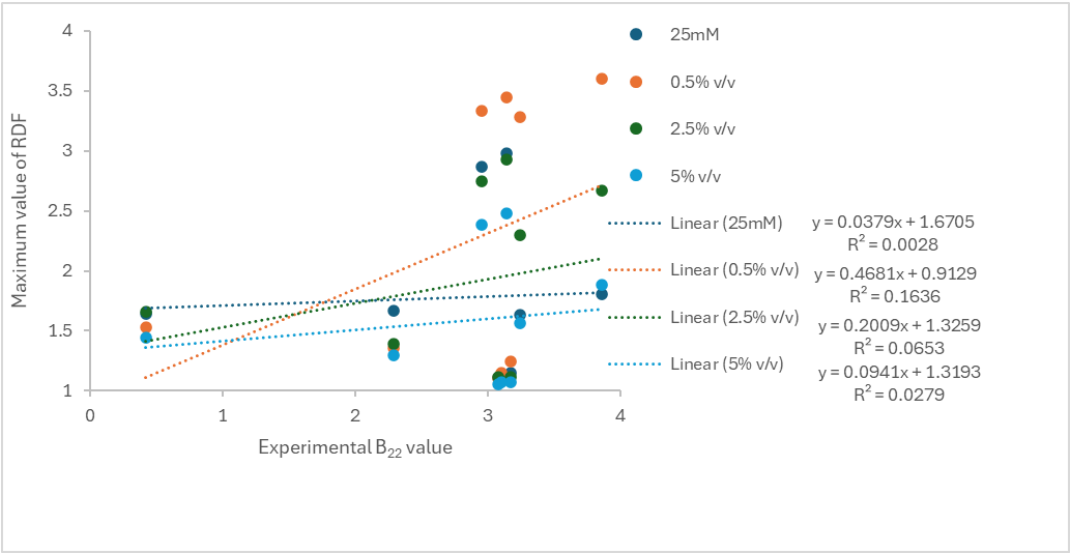


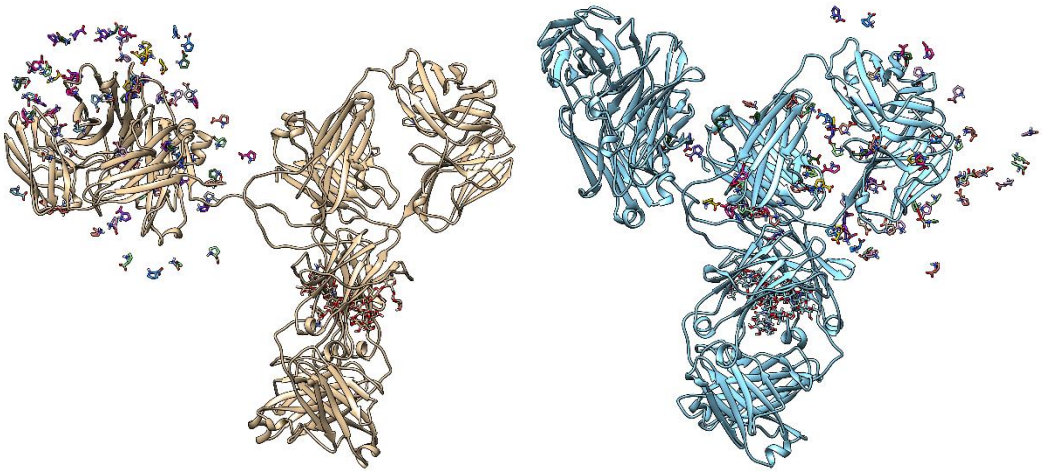
Figure 46: Tables showing the experimental second osmotic coefficient and the maximum values of RDF at different concentrations. The top table is for the cosolvent mixtures, and the bottom table is for single excipients. Graphs showing the possible regression between B22 and maximum RDF. Top one uses only five amino acids, and the bottom one uses all nine excipients

Looking at the table, for amino acids when there is a lower value of B_{22}, the RDF value is relatively higher. This can be clearly seen for arginine and histidine which have a B_{22} value around 0.42 and 2.29 respectively. When the mAb in is pure arginine, and the mixture of histidine and arginine, the maximum RDF value is higher than in the histidine at 0.5% v/v, 2.5% v/v and 5% v/v. But in 25Mm

histidine buffer, the maximum RDF value in arginine is 1.6 which is slightly lower than the 1.669 in histidine. The difference is subtle. Because the value is an average value of three replica, I checked the values for single replica, for arginine, the RDF values are 1.642, 1.674 and 1.485. For histidine, the RDF values are in a larger range with 1.645, 1.396 and 1.965, so the 1.965 of the last replicas caused the average to become higher.

The B_{22} value of arginine is a lot lower than the other amino acids gives a near zero value indicates a close to neutral interaction between proteins. The weak interaction might be due to arginine's charge. Although arginine can stabilize through electrostatic interactions, it can also affect the charge on protein surface causing greater chance of aggregation and the colloidal stability of native proteins are disturbed.[114] Arginine needs to be considered case by case.

For other three amino acids, alanine, glycine and proline, it also gets complicated to compare the results because of the similar values among them. And for the sugars, there is no obvious relationship between the experimental B_{22} values and the RDF values.



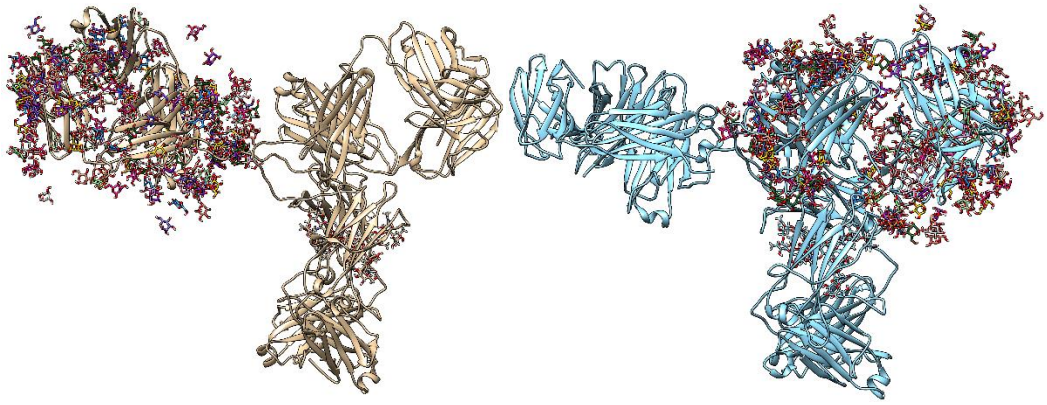


Figure 47: Top section is last frame of the whole mAb in replica 1 of 0.5 % v/v proline and proline molecules within 0.5 nm distance of two Fab regions shown separately at different timesteps from 0 to 50 ns at 5 ns interval. Bottom section is in 0.5 % v/v glucose.

3.4.8 Solvent molecules around the mAb

The movement of amino acid cosolvents were quite random and it's rare to see one amino acid molecule that stays at the same position throughout the simulation. For sugar molecules, there are more molecules accumulate around the protein surface and tends to cluster more than amino acids.

Time (ns) / Excipient	0	5	10	15	20	25	30	35	40	45	50
histidine	3	7	10	13	11	14	13	15	15	16	17
alanine	6	9	14	13	15	12	14	15	17	13	14
arginine	2	6	9	8	13	7	11	14	9	10	12
glycine	10	12	18	15	12	14	16	15	15	18	15
proline	5	11	10	11	10	10	9	14	12	11	15
sucrose	2	6	10	11	13	15	18	18	20	22	23
glucose	4	10	16	22	25	26	27	27	30	30	28
trehalose	2	8	11	13	14	16	16	19	19	21	26
mannitol	5	18	17	22	23	26	26	26	28	27	27

Time (ns) / Excipient	0	5	10	15	20	25	30	35	40	45	50
his_ala(HIS)	4	6	5	4	6	6	10	8	8	6	6
	2	6	5	5	7	7	6	7	7	8	6
his_arg(HIS)	3	2	6	7	5	7	8	6	6	7	7
	2	4	8	9	7	7	8	8	8	10	10
his_gly(HIS)	1	4	5	10	7	7	8	9	9	8	7
	7	7	7	8	6	10	8	9	11	10	8
his_pro(HIS)	2	4	5	4	5	7	5	8	6	7	4
	3	4	5	3	3	6	5	6	6	9	5
his_suc(HIS)	2	5	6	9	8	7	6	5	7	6	6
	2	4	5	8	10	12	12	10	11	12	14
his_glu(HIS)	2	3	6	7	7	5	7	6	5	8	6
	1	8	8	7	10	12	14	12	14	16	16
his_tre(HIS)	3	5	7	6	6	6	5	6	6	6	6
	1	7	7	7	9	9	11	12	12	13	13
his_man(HIS)	3	4	5	7	8	7	7	6	5	6	4
	1	5	9	9	11	12	13	14	18	17	16

Figure 48: For Fab1 from the whole mAb, the table on the top is the number of different solvent molecules at 0.5% v/v pure solvent and the bottom is the number of different solvent molecules at 0.5% v/v of a mixture of histidine with other molecules in a 1:1 ratio (0.25% v/v of each).

For histidine, the number of solvent molecules increased in the first 15 ns then decreased at 20 ns and 30 ns, but quickly increased again and reached 17 at 50 ns. For alanine, the number increased from 6 to 14 in the first 10 ns, then fluctuated, reached 17 at 40n ns and decreased slightly during last 10 ns. For arginine, the number increased from 2 to 13 in the first 20 ns, then fluctuated and reached a maximum of 14 at 35 ns. For glycine, the number increased from 10 to 18 in the first 15 ns, then fluctuated and reached 18 again at 45 ns. For proline, in the first 5 ns, the number increased from 5 to 11, then reached a maximum of 15 at 50 ns. In general, the number of amino acid molecules reached to a maximum value around 14 to 18 and generally fluctuated within a small degree.

For sucrose, the number of molecules increased from 2 to 23 at the end. For glucose, the number increased from 4 to 30 at 45 ns and then decreased to 28 at 50 ns. For trehalose, the number increased from 2 to 26 at the end. For mannitol, the number increased from 5 to 18 during the first 5 ns then decreased to 17 in the next 5 ns and quickly increased again reached to 28 at 40 ns. For sugar molecules, the numbers of molecules are again generally higher than amino acids. The number of glucose molecules are the highest amongst all the sugar molecules.

For cosolvents, the number of amino acid molecules have a similar trend to single solvent. As the concentration is 0.25% of each solvent, the maximum number of molecules varied from 8 to 11 which is a reasonable range. The maximum number of sugar molecules within 5 Å of the protein in the mixture varied from 13 to 18.

Time (ns) / Excipient	0	5	10	15	20	25	30	35	40	45	50
histidine	5	9	10	14	13	12	12	12	14	11	12
alanine	7	11	15	11	13	14	14	12	12	11	13
arginine	4	7	9	12	13	12	16	10	12	14	13
glycine	8	12	19	16	13	14	18	18	15	17	17
proline	4	8	10	10	12	13	8	10	11	10	10
sucrose	3	5	8	10	13	16	18	19	20	24	22
glucose	6	13	18	25	26	27	29	30	34	30	31
trehalose	0	4	8	10	12	13	14	15	16	18	19
mannitol	2	12	12	18	22	23	23	26	29	25	26

Time (ns) / Excipient	0	5	10	15	20	25	30	35	40	45	50
his_ala(HIS)	1	4	4	7	8	9	9	11	9	6	8
	3	6	8	7	8	7	7	4	8	7	7
his_arg(HIS)	2	3	6	7	5	6	6	7	6	6	8
	3	4	5	6	7	7	8	6	6	8	7
his_gly(HIS)	2	4	3	4	6	8	7	7	6	6	4
	2	5	9	6	9	8	7	6	11	13	10
his_pro(HIS)	2	3	6	4	6	8	8	7	6	8	6
	2	5	5	5	6	8	9	7	6	7	4
his_suc(HIS)	2	7	7	6	7	8	8	7	10	10	11
	2	5	4	6	7	7	9	9	9	9	10
his_glu(HIS)	1	5	6	6	6	5	7	9	6	8	6
	1	10	11	14	13	15	17	14	14	15	20
his_tre(HIS)	1	3	4	4	3	3	3	4	5	7	8
	0	2	2	4	3	7	4	5	6	7	7
his_man(HIS)	1	3	6	4	5	5	6	8	8	9	8
	1	8	7	8	9	10	11	13	15	15	16

Figure 49: For Fab2 from the whole mAb, the table on the top is the number of different solvent molecules at 0.5% v/v pure solvent and the bottom is the number of different solvent molecules at 0.5% v/v of a mixture of histidine with other molecules in a 1:1 ratio.

For the other Fab region of the whole mAb, for histidine, the number of solvent molecules increased from 5 to 14 in the first 15 ns then decreased to 12 and plateaued, increased to 14 again at 40 ns. For alanine, the number increased from 7 to 15 in the first 10 ns, then fluctuated within a range of 11-14. For arginine, the number increased from 4 to 16 at 30 ns, then decreased and fluctuated within a range of 10-14. For glycine, the number increased from 8 to 19 in the first 15 ns, then fluctuated between 13 to 18. For proline, the number increased from 4 to 13

at 25 ns, then decreased to 8 in the next 5 ns and then plateaued. In general, the number of amino acid molecules reached to a maximum value around 13 to 19 and generally fluctuated within a small degree.

For sucrose, the number of molecules increased from 3 to 24 at 45 ns and then decreased to 22 at 50 ns. For glucose, the number increased from 6 to 34 at 40 ns and then decreased to 31 at 50 ns. For trehalose, the number increased from 0 to 19 at the end. For mannitol, the number increased from 2 to 29 at 40 ns and then decreased to 26 at the end. For sugar molecules, the numbers of molecules are again generally higher than amino acids. The number of glucose molecules are the highest amongst all the sugar molecules.

For cosolvents, the number of amino acid molecules still have a similar trend to single solvent. The maximum number of molecules varied from 8 to 11 which is a reasonable range. The maximum number of sugar molecules within 5 Å of the protein in the mixture varied from 7 to 17. Both Fab regions behave in a similar way when considering the number of molecules within each region.

Looking back of all the analysis for simulations of the Fab region and for the whole mAb in the same set of excipients at same range of concentrations, RMSD, RMSF, SASA, Rg, PCA and RDF, there is a higher chance that by running fab region can give a reasonable representation of the whole mAb.

3.4.9 mAb in water

The paper by Xu and colleagues used histidine as the buffer, I also tried to use water to compare how mAb behave differently in water and see if there are any conformational changes at a longer timescale, so the simulation is 1 μs. The RMSD and RMSF graphs are shown below in Figure 50 and 52.

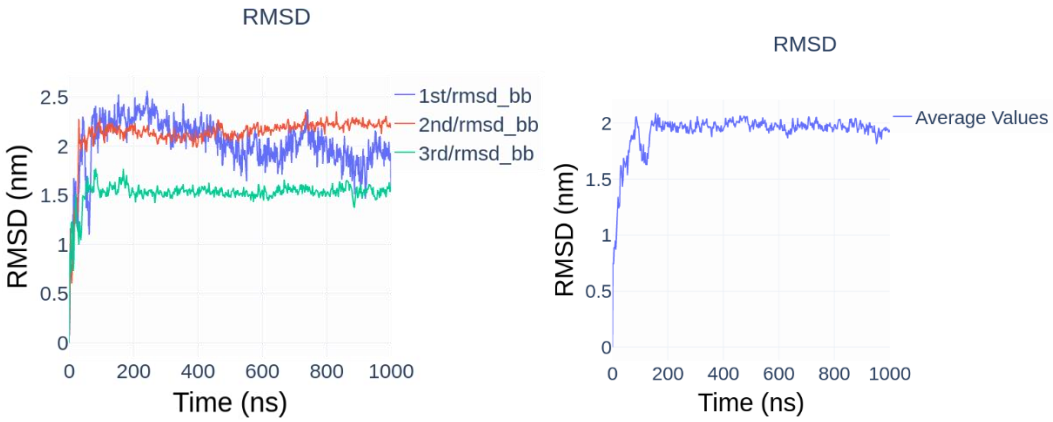


Figure 50: RMSD plot for all atomic simulation of NISTmAb in water. On the left is for the single replicas and on the right is the average RMSD from 3 replicas.

From the RMSD plots of three replicas and the average plot, we can see that the mAb in stable in water. There is a big jump at the beginning of the simulation for all the replicas. For replica 1, there are larger fluctuations for the rest of the simulation between 1.6 to 2.5. For replica 2 and 3, they are quite stabilised for the rest of the simulation.

The jump at the beginning corresponding to the movement of the hinge region which is the same scenario in different cosolvents. The beginning and last frame of the 1 μs simulation for replica 2 are shown below in Figure 51.

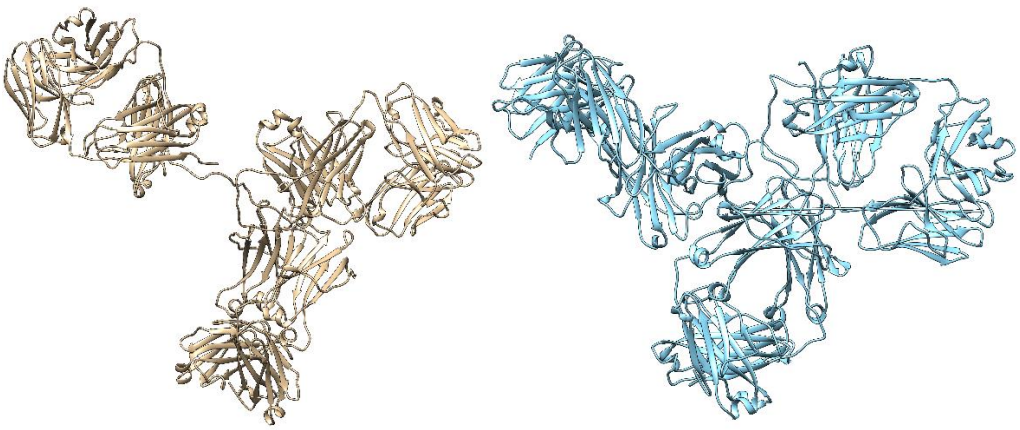


Figure 51: The first and last frames of the 1 μs atomic simulation in water.

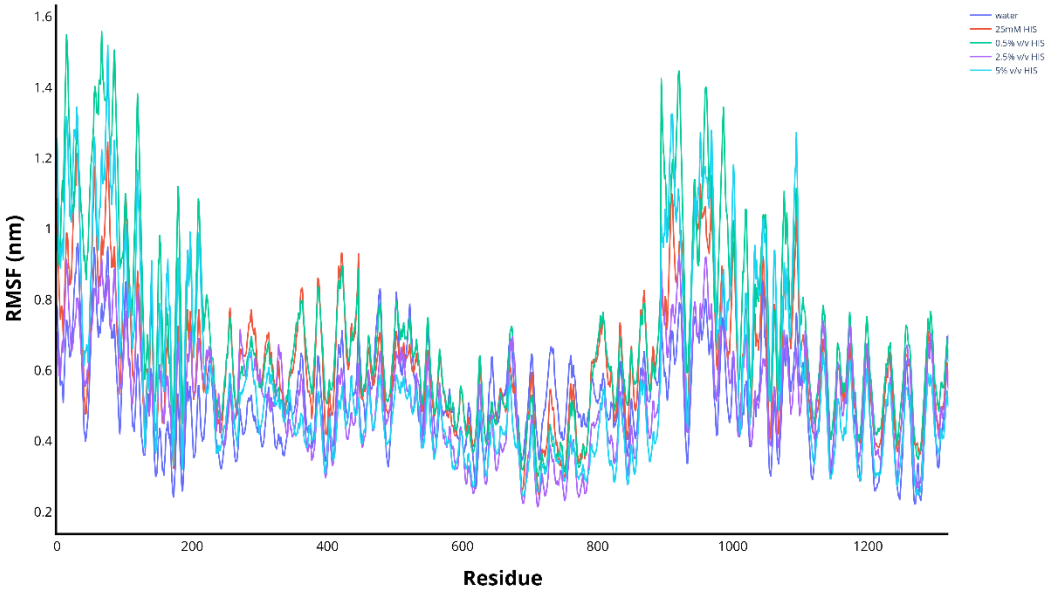


Figure 52: RMSF plots for NISTmAb in water and different concentrations of histidine.

The RMSF graphs shows the general trend of residue fluctuations is similar in water and histidine. When the mAb is in 0.5% v/v histidine, the fluctuations were higher compared to water and other concentrations.

3.4.10 Coarse grained simulation - MARTINI and SIRAH

Due to the size of the whole mAb, it is too time consuming to run all the atomic simulations in different cosolvents at a longer timescale. MARTINI CG was employed at the beginning as it is a well-known force field and have been used a lot in research. The simulation was initially carried out in water, as the 1 μs all-atomic simulation suggested the mAb should be stable inside water. When I run the first simulation, I didn't put any elastic forces and one of the Fab regions collapsed towards the rest of the mAb and remained for the rest of the simulation. Then I tried to apply elastic forces from 50 to 500 kcalmol-1 in replicas, sometimes the Fab region collapsed and sometimes it didn't. The collapsing is completely random, for example at 250 and 300, the Fab region was stable for 2 replicas and then collapsed at the last replica. I couldn't decide on a suitable force to run the simulation, so I decided to try to employ a different CG method. Examples of the mAb in CG representation of the first and last frames are shown below in Figure 53.

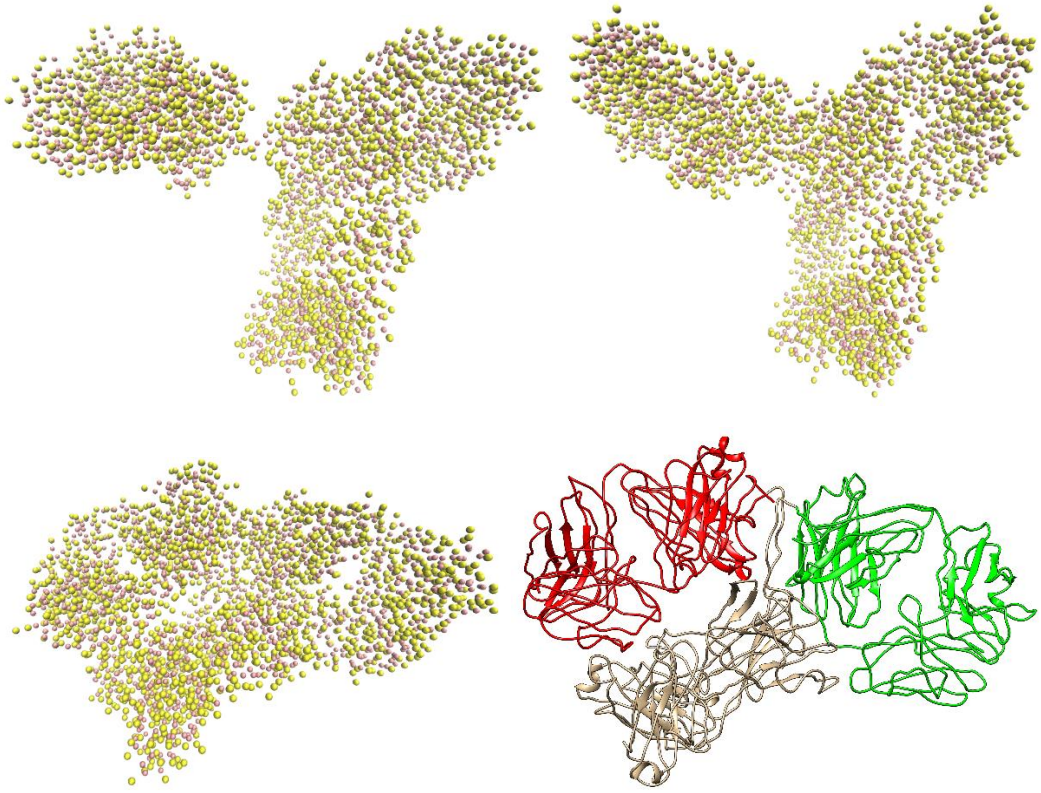


Figure 53: On the top left is the first frame of the CG model converted from atomic model and on the top right is the last frame of stabilised situations. On the bottom left is the last frame of

collapsed situation in CG model and on the left is the corresponding back mapped atomic model.

SIRAH force field was then used to convert the atomic structure to CG model and run the simulations. The mAb was placed in water first and then different amino acids were used as excipients. The first and last frames of CG model are shown below in Figure 54.

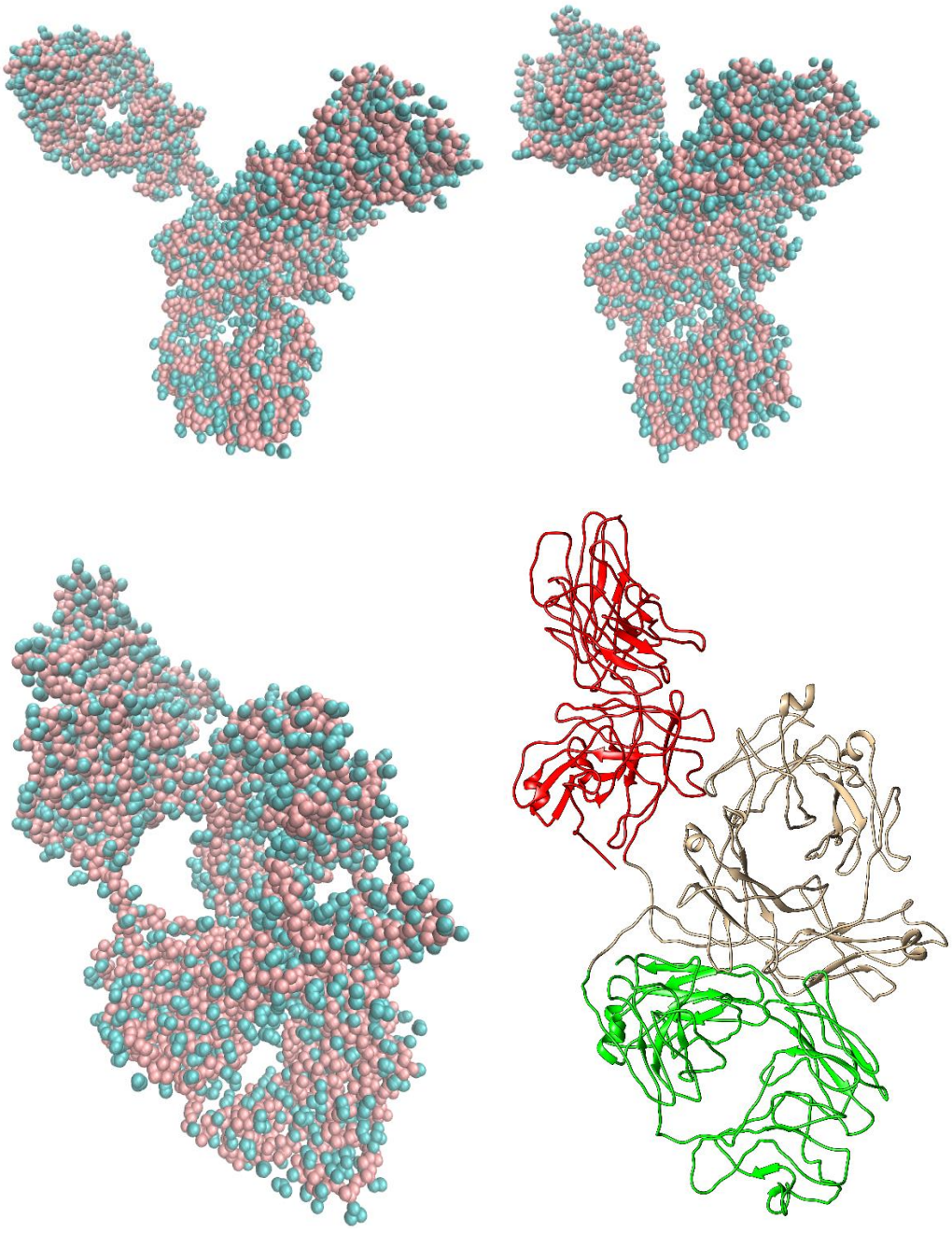


Figure 54: On the top left is the first frame of the CG model converted from atomic model and on the top right is the last frame when there is tiny movement of the hinge region. On the bottom left

is the last frame of larger movement in CG model and on the left is the corresponding back mapped atomic model.

When looking at the back mapped structure from SIRAH force field, the movement of the hinge region is more reasonable than the MARTINI force field. Then the mAb was placed in histidine, arginine, glycine, and proline. Alanine was not used because when converted to CG model, it is the same as glycine. The average RMSD graphs for four excipients are shown below in Figure 55.

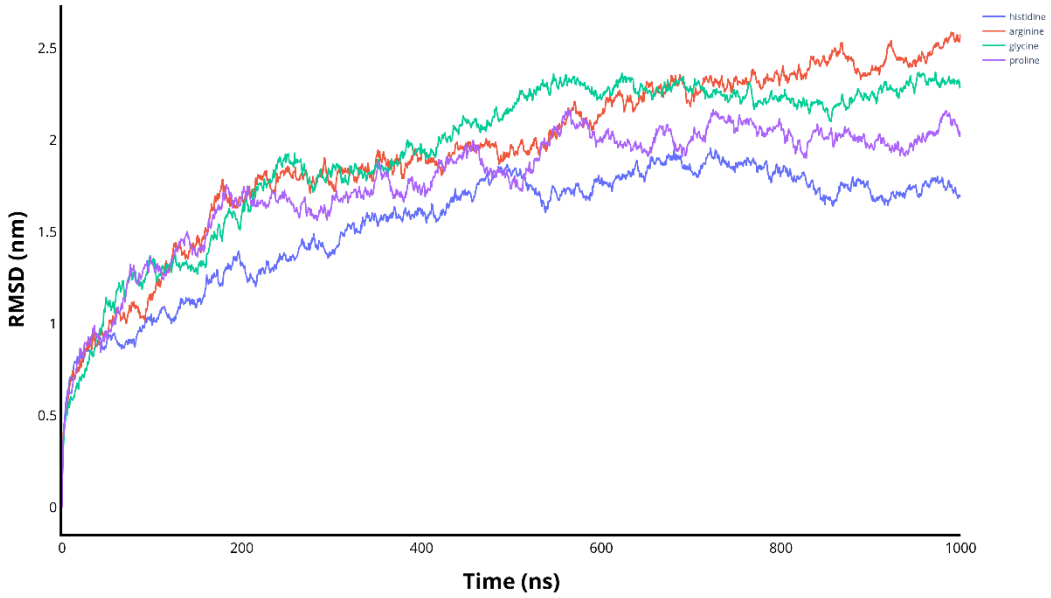


Figure 55: RMSD plot for CG simulation of mAb in 25mM of histidine, arginine, glycine and proline.

Histidine has a range of 1.555 to 2.032, arginine has a range of 1.943 to 2.869, glycine has a range of 1.940 to 2.517 and proline has a range of 1.777 to 2.480. Because there is no backbone selection of the CG model, the RMSD was calculated for all the protein atoms. The values are a little higher than calculating the RMSD for backbone only for the atomic structure. The huge jump at the first 200ns occurs because of the movement of the hinge region, then it increased in a

small degree and then plateaued. The mAb seems to be more stabilised in histidine and proline. The average RMSF graphs per bead for all the cosolvents are shown below in Figure 56.

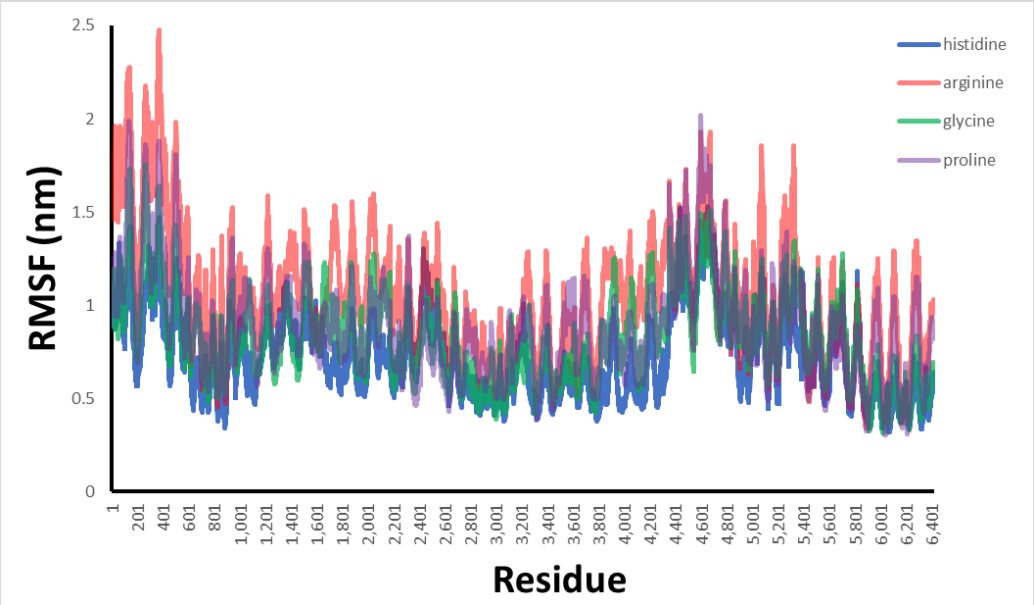


Figure 56: The RMSF plot for CG simulation of mAb in 25mM of histidine, arginine, glycine, and proline.

For CG simulation, it is mainly used to investigate the stability of the whole mAb in a longer timescale. Like all atomic simulation, the CG simulation shows the general trend of fluctuation of mAb is similar in different excipients. There are no obvious conformational changes, and the radius of gyration lies between 4.77 to 5.45 nm.

Beginning with how the model of NISTmAb was built, cosolvent simulations in amino acids: histidine, alanine, arginine, proline and glycine, sugars: sucrose, glucose, trehalose and mannitol were carried out at different concentrations. The chapter was mainly divided into two sections, one is results on Fab region, the other is results on whole mAb. Within the whole mAb, it is further divided into three sections, cosolvent simulations results, results for MD simulations in water and CG simulations. The analysis of both Fab and whole mAb shows that Fab can

be a reasonable representation for the whole mAb, therefore by running Fab region only, less time is required, and all atomic simulation can be carried out in a slightly longer timescale around 200 ns to 500 ns is possible.

Another hypothesis for this part of the project is RDF can be used as a possible indicator for aggregation propensity. The correlation between RDF and B_{22} is the key to this hypothesis. However, the results don't fully support this hypothesis. A weak correlation can be seen when amino acids were used as cosolvent as shown in Figures 23 and 46. As mentioned in the chapter, there are too little points to fully conclude the correlation. Also, to improve the comparison, a correlation between RDF and k_D can be investigated. Viscosity can be another property to be considered.

Chapter 4: Estrogen Receptor, NDP52 and L1Orf1

In this chapter, three proteins named ER, NDP52 and L1Orf1 were introduced. Simulations were running for each protein; experimental findings of these proteins were also mentioned briefly.

4.1 Estrogen receptor

The first part of this chapter is focusing on estrogen receptor (ER) which is an intracellular protein located in the cytosol. It belongs to the transcription regulator superfamily, stimulates gene transcription when it is activated by estrogen hormone. Steroid receptor ER can also be activated by steroid receptor coactivators (SRCs), these coactivators can be recruited both in the presence or absence of ligand, for example cyclin D1 can recruit SRC to ER without ligand binding.[115]

As a steroid hormone, estrogen affects growth and differentiation in mammals. It exerts its effect by several possible mechanisms and the classical one is binding to the nuclear receptor ERα and ERβ. In this mechanism, estrogens diffuse into the cell, binding to the receptors in the nucleus. Then dimerization occurs and the complex will bind to estrogen response elements (EREs) in the targeted genes and start transcription.[116] A conformational change within the ligand binding domain (LBD) of the ER is induced upon binding and coactivators can be recruited.[117] ERs can also regulate gene expression through protein-protein interactions, for example with activator protein 1 (AP1) and then recruits the coactivators as well. Nongenomic effects of estrogen can also occur using ERs located in or close to the plasma membrane, their activation can simulate kinase activity.[118]

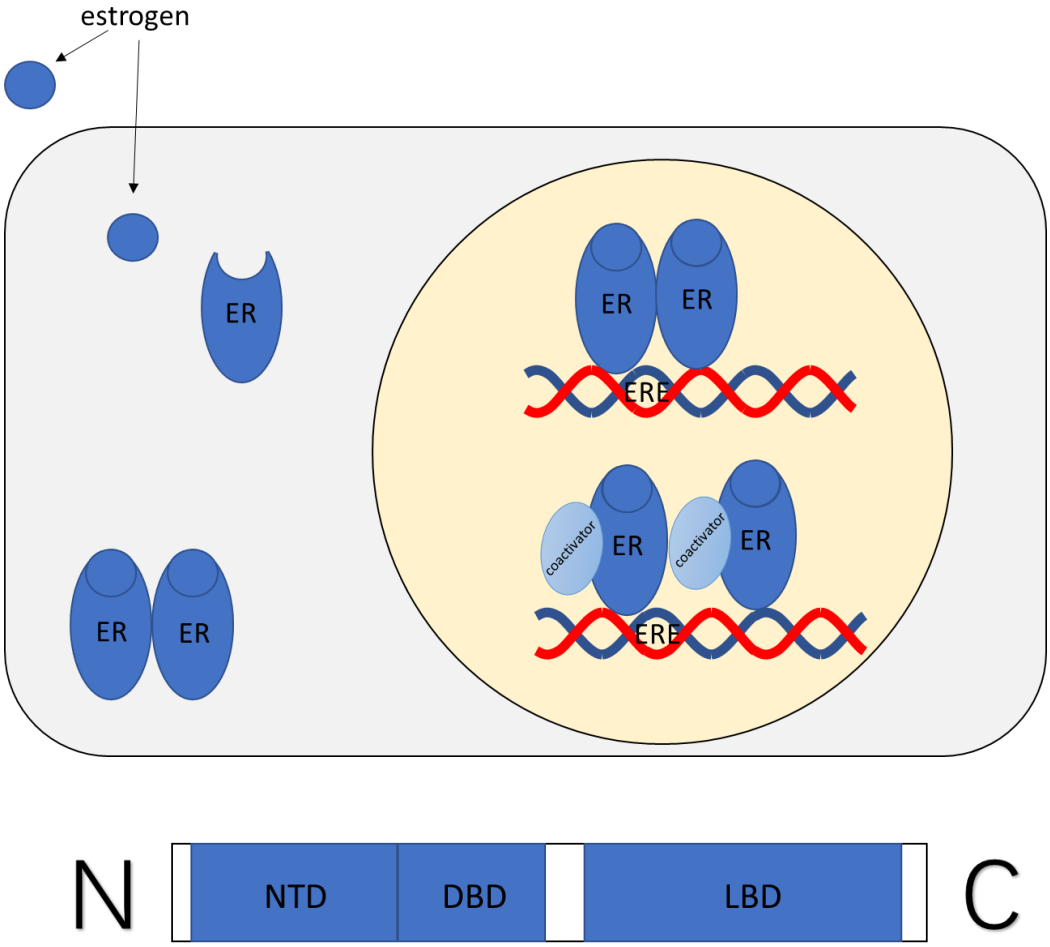


Figure 57 :The top panel is a schematic diagram of the mechanism of ER activation, cooactivators can be recruited. And the bottom panel shows the domains of ER; N-terminal domain (NTD), DNA-binding domain (DBD) and C-terminal ligand-binding domain (LBD).

In collaboration with Dr Wright's group, the effect of methylimidazolium ionic liquids (MILs) on the human estrogen receptor alpha (hERα) was investigated. MILs are a class of ionic liquids with a positively charged N-alkyl chain substituted methylimidazolium moiety, the cations always combine with anions such as chloride to give the liquid salt.[119] The ionic liquid 1-octyl-3-methylimidazolium (M8OI+) was detected in soil in landfill site.[120] M8OI has been shown as an activator of the hERα using a transactivation reporter gene assay[121] but the capability of other related MILs in activating the hERα remains unclear.

Chloride salts of 1-ethyl-3-methylimidazolium (EMI), 1-butyl-3-methylimidazolium (BMI), 1-hexyl-3-methylimidazolium (HMI), M8OI and 1-decyl-3-methylimidazolium (DMI) were examined for their stability in river water and metabolic human liver. Metabolism couldn't take place in short chain MILs (EMI, BMI, and HMI), however long chain MILs (M8OI and DMI) can undergo oxidative metabolism.

To model full-length ER, AlphaFold2[122] and other experimental data[123] were used, a model of hERα LBD-DBD dimer was built. There are two possible models fulfilled the site-directed mutagenesis and small-angle X-ray scattering restraints. As shown below in Figure 58, one is the DBD-LBD interface as one monomer and the other where domains are swapped.

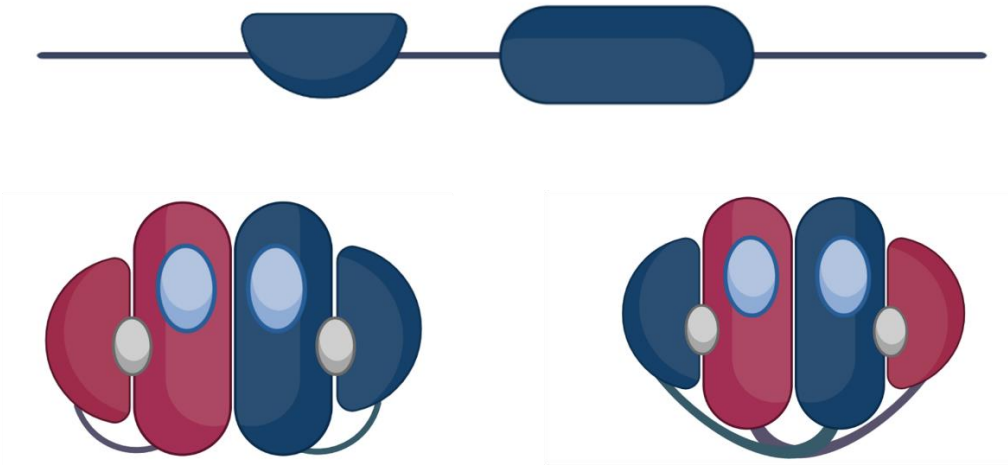


Figure 58: A schematic illustration of models drawn using PowerPoint where the DBD-LBD interface comprises each monomer separately (left) and one where the domains are swapped (right).

For molecular docking study, crystal structure of the ligand binding domain was used. (PDB: 1QKU). Solvent mapping of the models identified two druggable binding sites: an orthosteric (E2) binding site in each LBD domain and a smaller

allosteric binding site at each LBD-DBD interface as shown below in Figure 59. EMI is predicted to bind to the allosteric site.

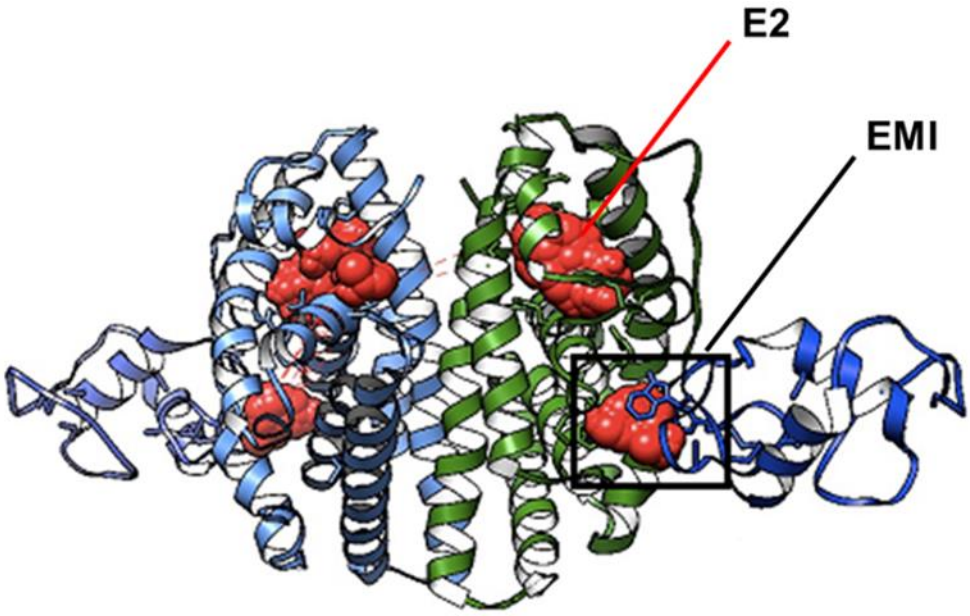


Figure 59: LBD and DBD of ER shown as flat ribbons, binding sites are shown as red regions. The allosteric site is inside the black rectangle. Picture is generated using Chimera, with ligands in the binding pockets, solid surfaces of the ligands are shown to give the red regions.

M8OI was predicted to bind to hERα at orthosteric site and the binding pose is shown below.

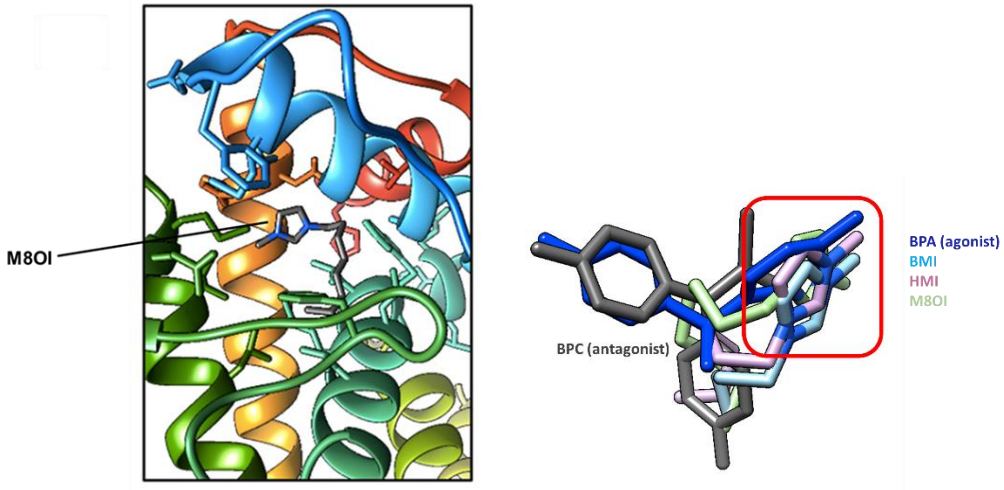


Figure 60: On the left-hand side showing M8OI bound to the orthosteric binding site in the hERα. Non hydrogen atoms of M8OI are coloured by element. Side chains of ER involved in interactions are shown. On the right-hand side are predicted binding modes BMI (cyan), HMI (pink) and M8OI (light green) overlayed with the binding mode of BPA (PDB code: 3UU7, blue – a hERα agonist)

and BPC (PDB code: 3UUC, grey - a hERα antagonist). The methylmidazolium rings of BMI, HMI and M8OI overlay with the ring that is unique for BPA but not BPC (the area selected by the red rectangle).

The toxicity and ER activation were assessed using ERα-HeLa-9903 cells. The MTT activity was shown below in Figure 61. There is a dose-dependent decrease in MTT activity when alkyl chain length increased, causing significant reductions in MTT activity at lower concentrations.

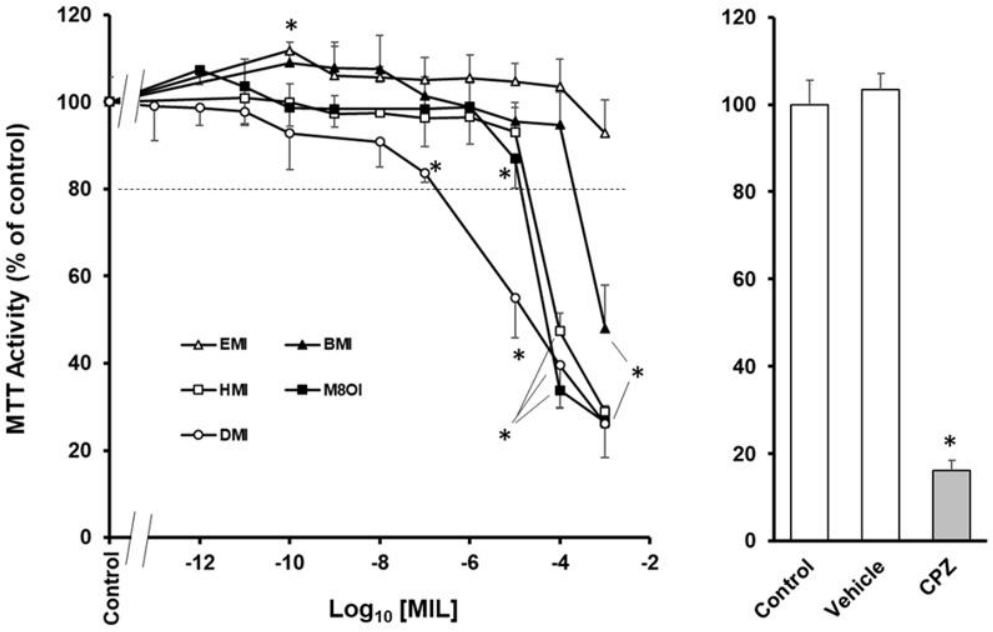


Figure 61: Effects of MILs and other ER-relevant compounds on MTT activities in ERα-HeLa-9903. ERα-HeLa-9903 were treated for 48 hours. During the last 2 hours, cells were also incubated with MTT prior to determination of MTT activity as outlined in methods section. Effect of the indicated MIL on MTT activity (left panel). Right panel, example of vehicle and positive controls routinely implemented within batch screens. Cells were incubated in control medium or additionally with vehicle control (0.1% (v/v) DMSO) or 200 µM chlorpromazine.

Treatment with EMI up to 1 mM did not lead to a reduction in MTT activity. With 1 mM BMI showed significant reduction. Similar reductions in MTT activity were seen with HMI, M8OI and DMI between 10-100 µM.

The hERα activation is shown in Figure 62, EMI, BMI and HMI were activators of hERα. M8OI was not positive for HeLa 9903 cells as the lower concentration is toxic to HeLa 9903 cells.

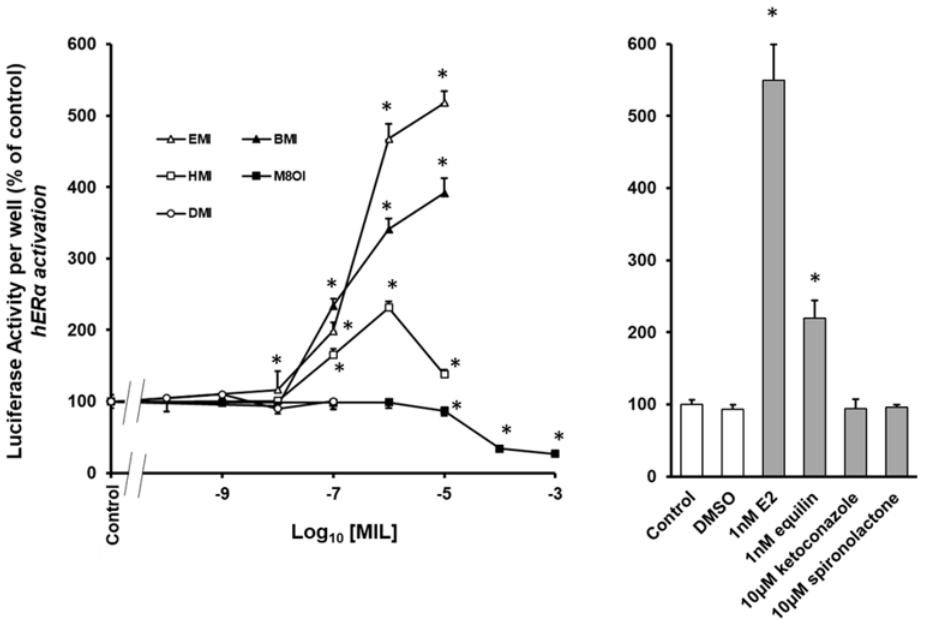


Figure 62: On the left-hand side is effect of the indicated MIL on luciferase activity; On the right-hand site is typical example of vehicle, OECD TG 4557 positive (E2 and equilin) and negative controls (ketoconazole, spironolactone) routinely implemented within batch screens. Cells were incubated in a control medium or additionally with vehicle control (0.1% (v/v) DMSO or positive/negative controls (added from 1000-fold molar concentrated DSMO stocks).

M8OI is a known activator of hERα from experimental data, but whether the short chain MILs will bind to hERα or not remains unclear. Our docking study agree with the experimental findings that M8OI can bind to hERα and suggests short chain MIL can bind to hERα as well. In vitro study by Dr Wright's group shown that short chain MILs are also activators of hERα, agrees with our docking results.

4.2 NDP52

This part of Chapter 4 is about NDP52. Domains like PAS or LBD can be regarded as “universal” intracellular sensors, yet this function is not limited to small molecule binding domains such as PAS or LBD. A lot of domains that used to be classified as scaffolding domains, can also exert the regulatory function, and be modulated by small molecules and ions. A good example of such domain is coiled-coil motif.

NDP52 belongs to the nuclear dot family. It contains a coiled-coil domain with redox-sensitive cysteines, upon oxidation leads to the formation of oligomers linked by disulphide bonds which is crucial for mitophagy. In the resting state, the NDP52 is in the cytosol, and it is proposed once reactive oxygen species (ROS) generated from damaged mitochondria, this triggers the oligomerization of NDP52 on the surface of mitochondria.

Figure 63: The domains of NDP52; including SKICH domain, the LC3-interation region (LIR) domain and coiled-coil domain.

The cargo receptor NDP52 is an autophagy receptor that can recognise ubiquitinated mitochondrial surface proteins and form an autophagic vesicle surround the mitochondria. Oxidation of NDP52 is required for PINK1/Parkin-dependent mitophagy. It contains cysteine residues that are redox sensitive and essential for disulphide bond formation and oligomerisation of the protein on damaged mitochondria.[124]

Monomer of NDP52 was obtained from AlphaFold 2.0[122], SKICH domains have a good agreement with the published crystal structures (PDB codes: 3VVV, 3VVW, 5Z7A, 5Z7L, 7EAA). The N-terminal SKICH domain and coiled coil domain were the focus of this project.

The coiled coil domain of NDP52 starts from residue Thr137, ends with Met349. To predict the multimer state of the coiled coil structure, CCFold[125] and AlphaFold-Multimer[126] were used, both showed a parallel dimer is preferred being a native complex rather than trimers. This is also supported by experimental data in Dr Korolchuk's lab. The other probable state is tetramer (dimer of dimer), the question is whether antiparallel or parallel model is favoured. AlphaFold Multimer predicted antiparallel model as the best-scoring model whereas CCBuilder2.0[127] predicted parallel tetramer as the best-scoring model.

The atomic modelling of dimer shows cysteines 163 and 321 can form C163-C163 and C321-C321 disulphide bonds, likely stabilising the dimer. The position of C153 is predicted to be away from the interface and can be involved in further crosslinking. The C153-C153 disulphide bond can stabilise the tetramer in an antiparallel conformation rather than the parallel conformation. Due to the orientation of those two outward-facing C153 per tetramer, higher order NDP52 oligomers can be formed through disulphide bridges. There is a C18 residue within the SKICH domain that can also contribute to the oligomerisation of NDP52.

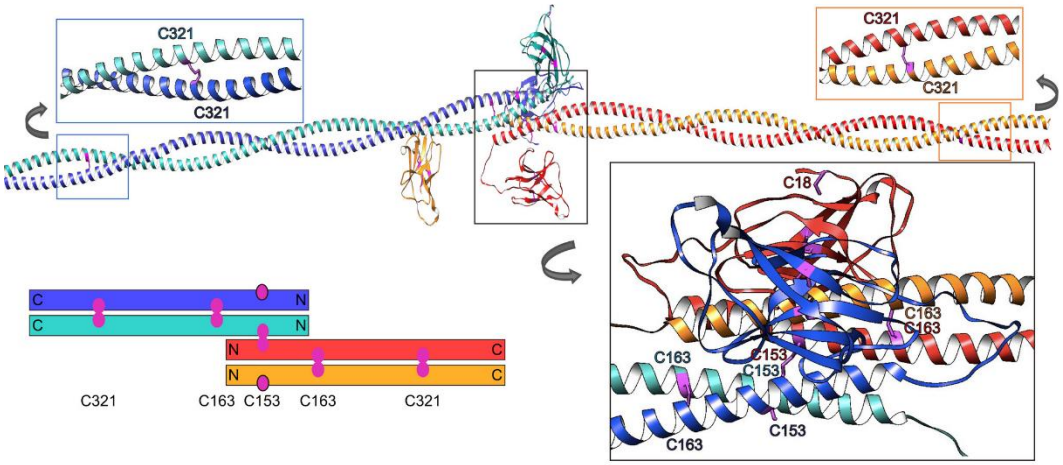


Figure 64: The conformation of antiparallel NDP52 tetramer and residue interactions. The schematic representation showing predicted orientation of cysteine residues in the coiled coil domains of the tetramer.

Models of part of the coiled coil domain of NDP52 (137 - 177) were built in both parallel and antiparallel orientation to start two sets of all-atom simulations in water.

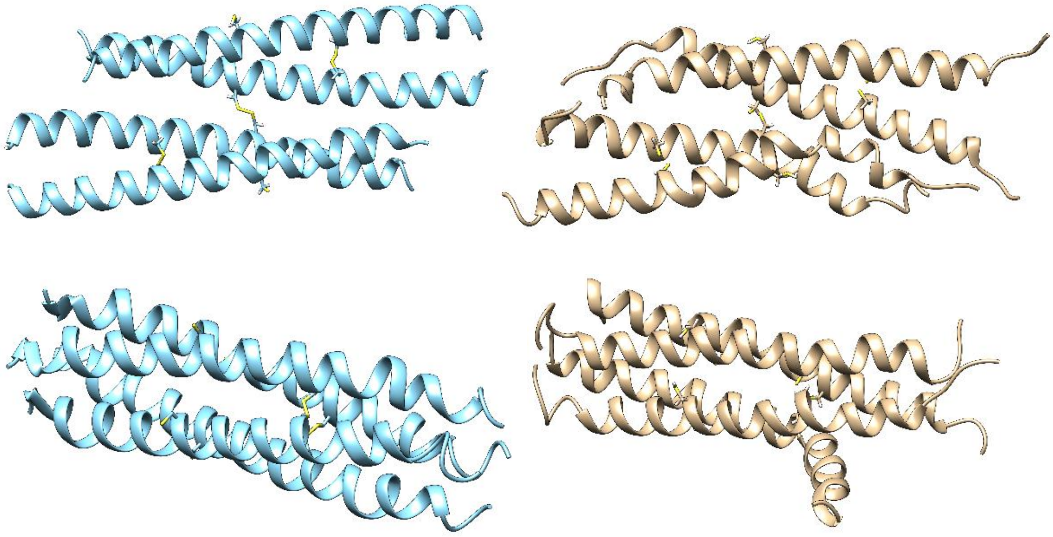


Figure 65: Models of NDP52 coiled-coil domain. On the top is the antiparallel model and the last frame after 100 ns atomic simulation. On the bottom is the parallel model and the last frame after 100 ns atomic simulation.

From the atomic simulation, we can tell that antiparallel orientation is more stable compared to the parallel one.

The NDP52 including the SKICH domain and coiled coil domain were built using protein-protein docking of SKICH domain on coiled coil domain, connecting loops were built by MODELLER. The model was then energy minimised. Full atomic model was converted to MARTINI coarse grained model using MARTINI version 2.2. A 500 ns simulation was carried out.

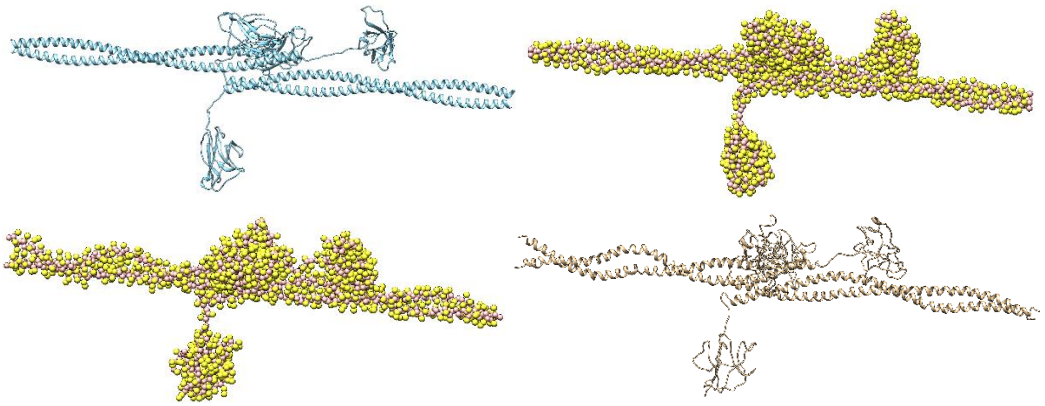


Figure 66: NDP52 tetramer models including SKICH domain and coiled coil domain. On the top left is the atomic model and on the right is the corresponding MARTINI model at the beginning of the simulation. The bottom left is the MARTINI model at the end of the simulation and on the right is the reverse mapped atomic structure.

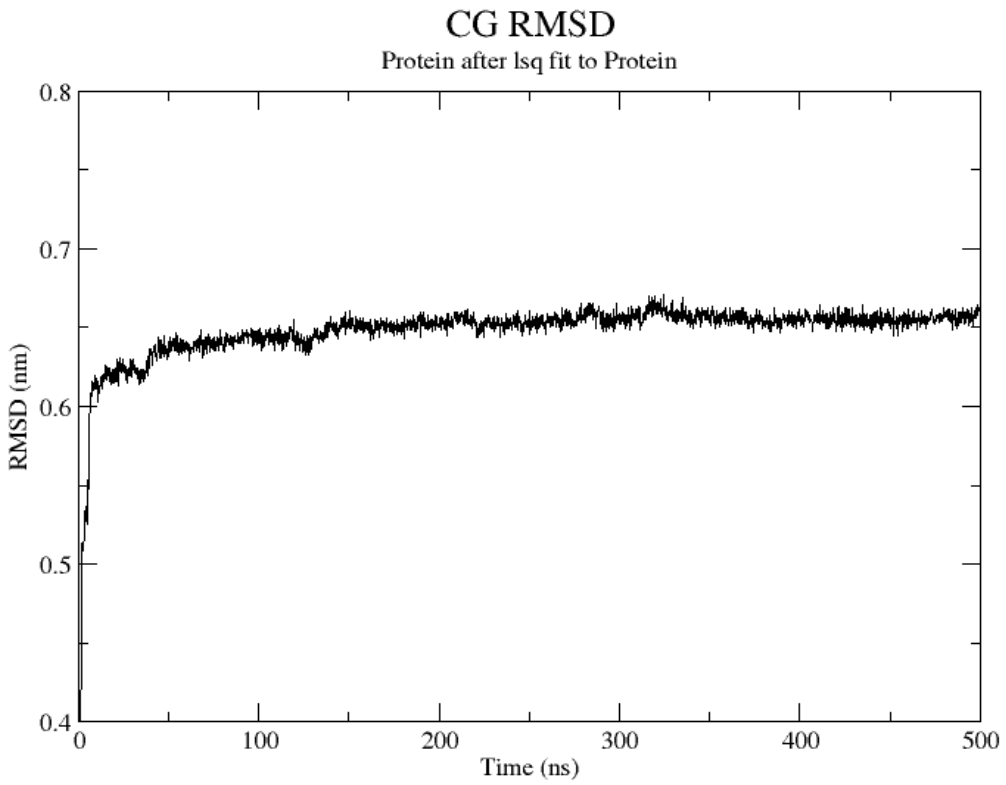


Figure 67: RMSD plot showing trajectories for 500 ns MARTINI coarse grain simulation of the antiparallel tetramer with SKICH domains.

From the RMSD plot, we can see the NDP52 was highly stable, there was a sharp rise in the first 20 ns and then the model was fluctuated within a 0.05 nm range.

From the experiments carried out by our collaborator, NDP52 forms DLC when exposed to oxidative stress or mitochondrial damage. Cysteine residues 18, 153, 163 and 321 are involved in oligomerisation of NDP52. It is confirmed when these four cysteines were mutated to serine, DLC cannot be formed when expose to hydrogen peroxide. The number of higher order oligomers were reduced in the mutants, which supports the finding that disulphide bridges can lead to oligomerisation of NDP52. Oxidation and oligomerisation of NDP52 lead to mitophagy is also supported by experimental data when C18, 153, 163, 321S mutant and WT were introduced to Hela cell line.

4.3 L1-ORF1p

The last part of Chapter 4 studies L1Orf1. The long interspersed nuclear element 1 (LINE-1) is an autonomous non-LTR retrotransposon evolving in mammalian genomes. ORF1p and ORF2p are two L1 encoded proteins required for retrotransposition.[128] ORF1p is responsible for RNA binding and ORF2p has endonuclease and reverse-transcriptase activities.[129] The process starts from binding of RNA polymerase II with the 5′-UTR promoter region of LINE-1. This will allow the transcription of full-length mRNA of LINE-1.[130] ORF1p and ORF2p are translated in the cytosol, so the mRNA is exported to the cytosol and combined to form a ribonucleoprotein (RNP) complex, which is the retrotransposition intermediate.[131, 132] The RNP is then translocated in the nucleus and mediate retrotransposition.

Figure 68: The domains of L1ORF1p; including coiled-coil domain, RNA recognition motif (RRM) domain and C-terminal domain (CTD).

Research by Cook et al[133] has shown there are highly conserved proline-directed protein kinase (PDPK) target sites and docking motifs which are crucial for L1 retrotransposition. Serine PDPK sites in the ORF1p can interact with proline isomerase Pin1 which is a key component of PDPK-mediated regulatory pathway.

To study the interaction between ORF1p and Pin1 and how it affects the retrotransposition, computational simulation and experiments were carried out.

For computational simulation, a model of Orf1p was built based on a hexamer of coiled coil domain (PDB: 6FIA) and RRM and CTD (PDB: 2LDY), the N-terminal disordered region is built according to AlphaFold. Pin1 (PDB:7EFJ) was also added.

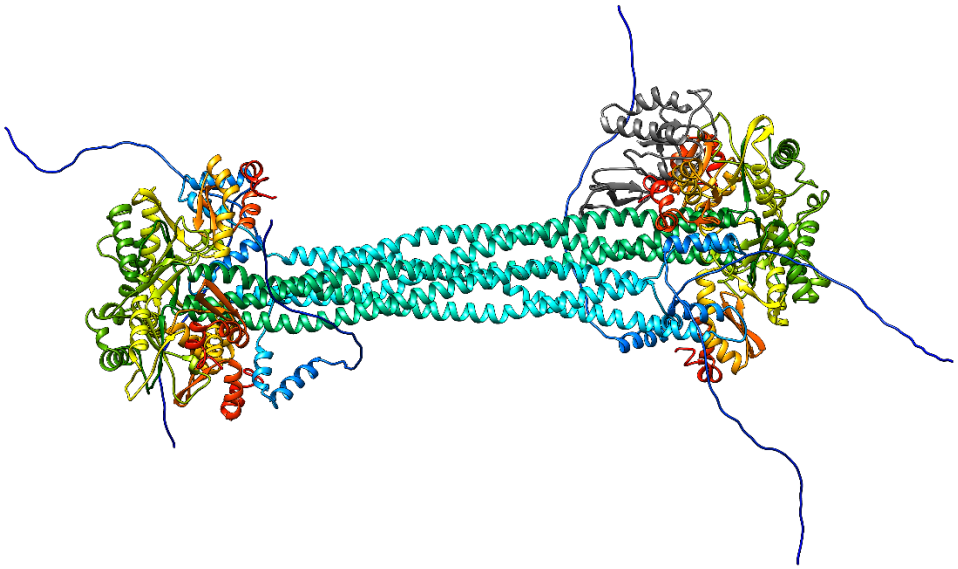


Figure 69: Computational model of ORF1p and Pin1. The ORF1p was displayed using rainbow description from N terminal to C terminal. Pin1 was coloured dark grey.

Mutation was introduced to the protein, Ser18 and Ser27 were mutated to phosphorylated serine according to the paper these two are potential phosphorylation sites[133], Arg141 was then mutated to alanine based on our hypothesis as it might interact with Pin1, the R141A mutant were also tested in cells by Dr Shukla. Two sets of simulations were carried out, one set is 50 ns all-atomic simulations for part of the ORF1p and full Pin1. The other set is 1 μs martini CG simulation for the full structure.

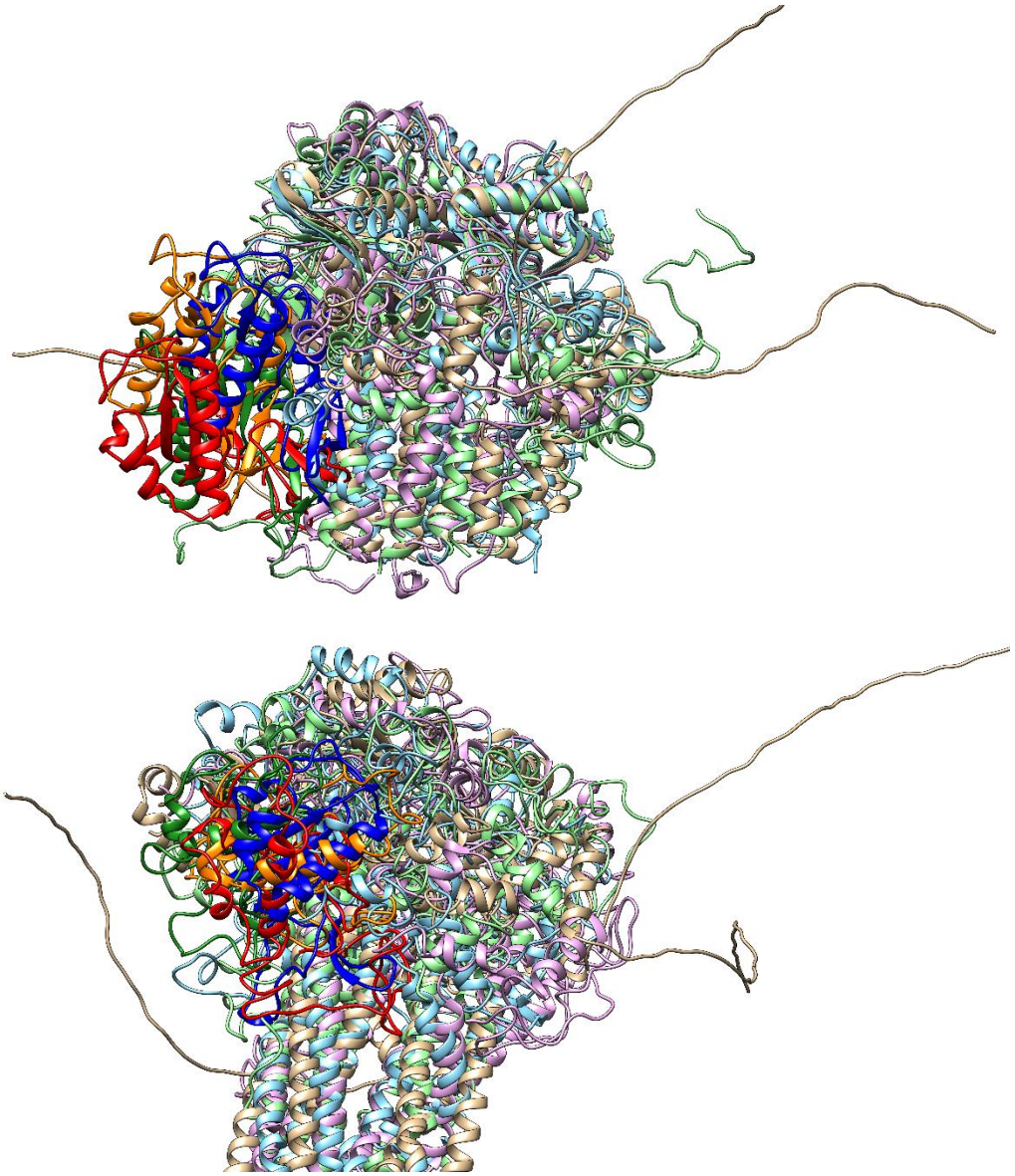


Figure 70: Top one is the all-atomic structure for overlapping of starting point and the last frame from WT, S1827A, S1827R141A. Bottom one is the CG structure for overlapping the starting structure and the last structure from WT, S1827A, S1827R141A. The ORF1p was coloured tan, cyan, light green and pink respectively. Pin1 was coloured blue, forest green, orange and red.

We can tell from Figure 70, for all simulations, the N terminal disordered loop moved towards the coiled coil region and the coiled coil region is quite stable. Pin1 moved slightly relative to ORF1p position. The last frame of all-atomic structure for S1827R141A mutant shows that Pin1 moved quite far away from ORF1p but in the CG simulation, the movement is not obvious.

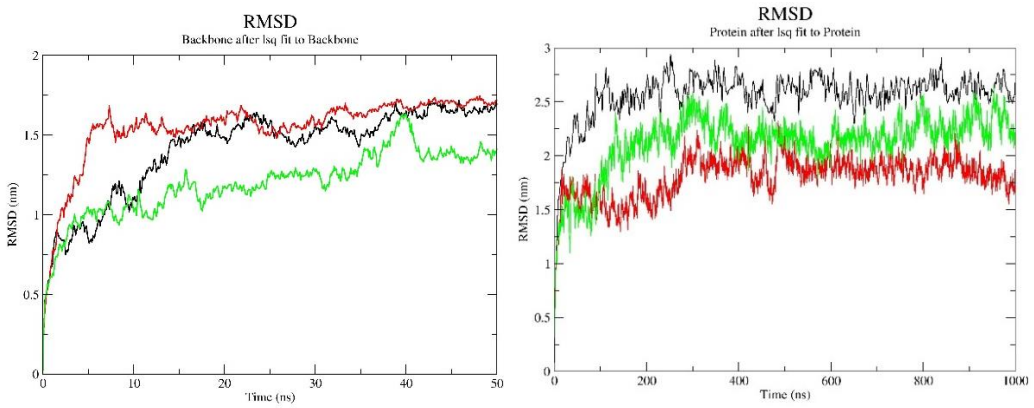


Figure 71: RMSD plots for trajectories of WT in black, S1827A in red and S1827R141A in green. On the left is 50 ns all-atomic simulation and on the right in 1 μs CG simulation.

For most of the simulation, there was a huge jump at the beginning and then fluctuated within a smaller range. The huge jump could because of movement of disordered loop. For all-atomic simulations, WT and S1827A reaches a similar value of 1.5 nm at 15 ns and 5 ns respectively. For S1827R141A, the RMSD was still increasing at a slower rate. For CG simulations, all of them fluctuated at a smaller extend up and down after they reached the stable conformation.

The purpose of this study is investigating how mutations affect the interaction between Orf1p and Pin1. For successful L1 retrotransposition, phosphorylation of Orf1p need to take place at proline-directed protein kinase (PDPK) sites. Pin1 binds to phosphorylated S/T-P motifs in Orf1p and plays a mechanistic role in L1 retrotransposition. By placing Pin1 close to S18 and S27 which are S-P motifs, running MD simulations for different mutants, observations were focused how Pin1 affect Orf1p. When R141 was mutated into alanine, Pin1 seems went apart to a small extent. Experiments done by Dr Shukla suggested retrotransposition was not affected by the mutation. It's just the start of a project, the relationship between Pin1 and Orf1p and how the interaction affect retrotransposition remains unclear.

For this chapter, we successfully showed the docking of short and long MILs to hERα in the binding sites and the experimental results agree that MIL are activators of hERα. From the computational docking study, the activation of hERα by MIL can be studied at an atomic level, the binding sites can be visualized which is an advantage over experimental study.

For the NDP52 study, simulations showing the importance of cystine residues, and the disulphide bridges are required for the stabilisation of the tetramer. Mutations from cysteine to serine reduces oligomerisation can be explained by the disruption of disulphide bridges, this reducing of oligomerisation affects mitophagy supported by cell studies.

As for Orf1p, there is no useful information learned from both computational and experimental studies. It is still early stage of research, the hypothesis of mutants of Orf1p have less interaction with Pin1 and retrotransposition would be affected needs more simulations and experiments to prove.

Chapter 5: Conclusions and Future Work

Multiscale molecular modelling and solvent mapping techniques can enable *in silico* predictions stability of proteins at different environments, once there is a reliable starting structure of the protein of interest available. Traditionally, it used to be experimental structures (obtained by X-ray crystallography or NMR), but models obtained by either homology modelling techniques or the cutting-edge tools such as AlphaFold may be successfully used.

Assessment of stability of protein therapeutics (mAbs) in different cosolvent solutions requires combining coarse-grain and all-atom techniques to investigate all relevant conformations of a large protein and putative binding pockets that may accommodate interacting cosolvents. Performing the all-atom MD simulations proved to be a necessary post-processing step for the coarse grain mAb models. Based on the results, the workflow proposed in this thesis could contribute to aid in risk assessment of early development of new formulations for therapeutic mAbs.

The workflow combining molecular modelling, multiscale simulations, and cosolvent mapping, developed for the formulation studies, can be readily transferred to study of biological phenomena. Examples provided in this dissertation involve large, highly flexible yet structurally and functionally diverse proteins. One of them, NDP52, is the redox-regulated key mitophagy receptor, which ability to sense ROS generated by damaged mitochondria is required for the efficiency of mitophagy. Multiscale modelling workflow assisted in gaining the understanding of the role of specific cysteine residues crucial for NDP52 ROS sensing, and also supported a particular arrangement of NDP52 oligomers, developing a structure-based model of redox-triggered NDP52 oligomerisation. The model has been supported by experimental data such as immunoblotting and

site-directed mutagenesis. The future work needs to validate NDP52 as a potential new therapeutic target, wherein identification of “druggable” sites and states, assisted by multiscale molecular modelling techniques can facilitate the drug discovery process, aiming to combat diseases associated with the loss of cellular homeostasis and enhance healthy ageing.

Another example covered in this work, namely L1 Orf1p, and its putative interactions with peptidyl-prolyl isomerase Pin1. In the absence of experimental structure of the complexes, the full-length L1Orf1p protein has been modelled in functionally relevant oligomeric states, its dynamics have been evaluated using coarse-grain and all-atom MD simulations, and its interactions with Pin1 have been investigated using protein-protein docking followed-up by MD simulations. This workflow allowed for identification of specific mutants which are likely to have an effect on Orf1p complexes with Pin1. Future works will require experimental validation of those mutants.

Finally, the workflow tested herein contributed to the explanation of the molecular mechanism underlying endocrine disruption by small molecular methylimidazolium liquids (MILs) confirmed *in vivo* in animal studies. These cosolvents are used in industrial processes such as biofuel production and are considered to be environmentally safe. The workflow predicted MILs binding to both the estrogen binding site (orthosteric) and a newly identified, allosteric site on the human estrogen receptor alpha. Based on structure-activity considerations involving both computational and experimental approaches, some MILs and their metabolites may remain a hazard to the population. The results indicate that MILs have the potential to exert adverse effects to humans, other animals and the environment in general. Future works will involve wider search on other putative molecular targets for those MILs and need to address the safety concerns highlighted by the work presented in this thesis.

The first billion atom simulation was done for a coarse-grained model of GATA4 gene locus, using 130,000 processor cores in 2019.[134] Recently, a whole cell of JCVI-syn3A which is a minimal cell was built at CG level, but simulations cannot be done yet. The whole system contains 561 million beads representing more than six billion atoms including water and ions.[135] The development of software and computing resources will allow the simulation of much more bigger systems in the future.

For this project, only one Fab region/mAb was simulated in the simulation box due to the time limit, if there is sufficient time and computing power, ideally more proteins can be put in the simulation box so the interaction between proteins can be clearly observed, and aggregation might happen at certain parts of the protein, the results could be more helpful towards understanding the cause of aggregation at an atomic level.

Reference

(1) Vickery, H. B. The origin of the word protein. *Yale J Biol Med* **1950**, *22* (5), 387-393. From NLM.

(2) Maehle, A.-H.; Prüll, C.-R.; Halliwell, R. F. The emergence of the drug receptor theory. *Nature Reviews Drug Discovery* **2002**, *1* (8), 637-641. DOI: 10.1038/nrd875.

(3) Rang, H. P. The receptor concept: pharmacology's big idea. *Br J Pharmacol* **2006**, *147 Suppl 1* (Suppl 1), S9-16. DOI: 10.1038/sj.bjp.0706457 From NLM.

(4) Kobilka, B. K. G protein coupled receptor structure and activation. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2007**, *1768* (4), 794-807. DOI: https://doi.org/10.1016/j.bbamem.2006.10.021.

(5) Rosenbaum, D. M.; Rasmussen, S. G.; Kobilka, B. K. The structure and function of G-protein-coupled receptors. *Nature* **2009**, *459* (7245), 356-363. DOI: 10.1038/nature08144 From NLM.

(6) Bockaert, J.; Philippe Pin, J. Molecular tinkering of G protein-coupled receptors: an evolutionary success. *The EMBO Journal* **1999**, *18* (7), 1723-1729. DOI: https://doi.org/10.1093/emboj/18.7.1723.

(7) Li, S.; Wong, A. H. C.; Liu, F. Ligand-gated ion channel interacting proteins and their role in neuroprotection. *Frontiers in Cellular Neuroscience* **2014**, *8*, Mini Review.

(8) Vogel, C.; Bashton, M.; Kerrison, N. D.; Chothia, C.; Teichmann, S. A. Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol* **2004**, *14* (2), 208-216. DOI: 10.1016/j.sbi.2004.03.011 From NLM.

(9) Behring, E. v. Ueber das Zustandekommen der Diphtherie-Immunität und der Tetanus-Immunität bei Thieren. *Deutsche Medizinische Wochenschrift* **1890**, *16*, 1113-1114.

(10) Takács, L.; Vazquez-Abad, M.-D.; Elliott, E. A. Chapter 23. Therapeutic monoclonal antibodies : history, facts and trends. In *Annual Reports in Medicinal Chemistry*, Vol. 36; Academic Press, 2001; pp 237-246.

(11) Davies, D. R.; Chacko, S. Antibody structure. *Accounts of Chemical Research* **1993**, *26* (8), 421-427. DOI: 10.1021/ar00032a005.

(12) Inbar, D.; Hochman, J.; Givol, D. Localization of antibody-combining sites within the variable portions of heavy and light chains. *Proc Natl Acad Sci U S A* **1972**, *69* (9), 2659-2662. DOI: 10.1073/pnas.69.9.2659 From NLM.

(13) Köhler, G.; Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **1975**, *256* (5517), 495-497. DOI: 10.1038/256495a0 From NLM.

(14) Morrison, S. L.; Johnson, M. J.; Herzenberg, L. A.; Oi, V. T. Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. *Proceedings of the National Academy of Sciences* **1984**, *81* (21), 6851-6855. DOI: 10.1073/pnas.81.21.6851.

(15) Jones, P. T.; Dear, P. H.; Foote, J.; Neuberger, M. S.; Winter, G. Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature* **1986**, *321* (6069), 522-525. DOI: 10.1038/321522a0 From NLM.

(16) Veitshans, T.; Klimov, D.; Thirumalai, D. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding and Design* **1997**, *2* (1), 1-22. DOI: https://doi.org/10.1016/S1359-0278(97)00002-3.

(17) Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc Natl Acad Sci U S A* **1992**, *89* (18), 8721-8725. DOI: 10.1073/pnas.89.18.8721 From NLM.

(18) Mahler, H. C.; Friess, W.; Grauschopf, U.; Kiese, S. Protein aggregation: pathways, induction factors and analysis. *J Pharm Sci* **2009**, *98* (9), 2909-2934. DOI: 10.1002/jps.21566 From NLM.

(19) Andrews, J. M.; Roberts, C. J. A Lumry−Eyring Nucleated Polymerization Model of Protein Aggregation Kinetics: 1. Aggregation with Pre-Equilibrated Unfolding. *The Journal of Physical Chemistry B* **2007**, *111* (27), 7897-7913. DOI: 10.1021/jp070212j.

(20) Rambaran, R. N.; Serpell, L. C. Amyloid fibrils: abnormal protein assembly. *Prion* **2008**, *2* (3), 112-117. DOI: 10.4161/pri.2.3.7488 From NLM.

(21) Neal, B. L.; Asthagiri, D.; Lenhoff, A. M. Molecular origins of osmotic second virial coefficients of proteins. *Biophys J* **1998**, *75* (5), 2469-2477. DOI: 10.1016/s0006-3495(98)77691-x From NLM.

(22) Roberts, C. J.; Das, T. K.; Sahin, E. Predicting solution aggregation rates for therapeutic proteins: approaches and challenges. *Int J Pharm* **2011**, *418* (2), 318-333. DOI: 10.1016/j.ijpharm.2011.03.064 From NLM.

(23) Herhut, M.; Brandenbusch, C.; Sadowski, G. Inclusion of mPRISM potential for polymer-induced protein interactions enables modeling of second osmotic virial coefficients in aqueous polymer-salt solutions. *Biotechnol J* **2016**, *11* (1), 146-154. DOI: 10.1002/biot.201500086 From NLM.

(24) Ruppert, S.; Sandler, S. I.; Lenhoff, A. M. Correlation between the osmotic second virial coefficient and the solubility of proteins. *Biotechnol Prog* **2001**, *17* (1), 182-187. DOI: 10.1021/bp0001314 From NLM.

(25) Valente, J. J.; Verma, K. S.; Manning, M. C.; Wilson, W. W.; Henry, C. S. Second virial coefficient studies of cosolvent-induced protein self-interaction. *Biophys J* **2005**, *89* (6),

4211-4218. DOI: 10.1529/biophysj.105.068551 From NLM.

(26) Strickley, R. G.; Lambert, W. J. A review of Formulations of Commercially Available Antibodies. *Journal of Pharmaceutical Sciences* **2021**, *110* (7), 2590-2608.e2556. DOI: https://doi.org/10.1016/j.xphs.2021.03.017.

(27) Narayanan, H.; Dingfelder, F.; Condado Morales, I.; Patel, B.; Heding, K. E.; Bjelke, J. R.; Egebjerg, T.; Butté, A.; Sokolov, M.; Lorenzen, N.; Arosio, P. Design of Biopharmaceutical Formulations Accelerated by Machine Learning. *Molecular Pharmaceutics* **2021**, *18* (10), 3843-3853. DOI: 10.1021/acs.molpharmaceut.1c00469.

(28) Tian, F.; Middaugh, C. R.; Offerdahl, T.; Munson, E.; Sane, S.; Rytting, J. H. Spectroscopic evaluation of the stabilization of humanized monoclonal antibodies in amino acid formulations. *International Journal of Pharmaceutics* **2007**, *335* (1), 20-31. DOI: https://doi.org/10.1016/j.ijpharm.2006.10.037.

(29) Falconer, R. J.; Chan, C.; Hughes, K.; Munro, T. P. Stabilization of a monoclonal antibody during purification and formulation by addition of basic amino acid excipients. *Journal of Chemical Technology & Biotechnology* **2011**, *86* (7), 942-948. DOI: https://doi.org/10.1002/jctb.2657.

(30) Chen, B.; Bautista, R.; Yu, K.; Zapata, G. A.; Mulkerrin, M. G.; Chamow, S. M. Influence of histidine on the stability and physical properties of a fully human antibody in aqueous and solid forms. *Pharm Res* **2003**, *20* (12), 1952-1960. DOI: 10.1023/b:pham.0000008042.15988.c0 From NLM.

(31) Baynes, B. M.; Trout, B. L. Rational Design of Solution Additives for the Prevention of Protein Aggregation. *Biophysical Journal* **2004**, *87* (3), 1631-1639. DOI: https://doi.org/10.1529/biophysj.104.042473.

(32) Baynes, B. M.; Wang, D. I.; Trout, B. L. Role of arginine in the stabilization of proteins against aggregation. *Biochemistry* **2005**, *44* (12), 4919-4925. DOI: 10.1021/bi047528r From NLM.

(33) Arakawa, T.; Ejima, D.; Tsumoto, K.; Obeyama, N.; Tanaka, Y.; Kita, Y.; Timasheff, S. N. Suppression of protein interactions by arginine: a proposed mechanism of the arginine effects. *Biophys Chem* **2007**, *127* (1-2), 1-8. DOI: 10.1016/j.bpc.2006.12.007 From NLM.

(34) Arakawa, T.; Timasheff, S. N. Stabilization of protein structure by sugars. *Biochemistry* **1982**, *21* (25), 6536-6544. DOI: 10.1021/bi00268a033.

(35) Lerbret, A.; Bordat, P.; Affouard, F.; Hédoux, A.; Guinet, Y.; Descamps, M. How Do Trehalose, Maltose, and Sucrose Influence Some Structural and Dynamical Properties of Lysozyme? Insight from Molecular Dynamics Simulations. *The Journal of Physical Chemistry B* **2007**, *111* (31), 9410-9420. DOI: 10.1021/jp071946z.

(36) Yadav, S.; Laue, T. M.; Kalonia, D. S.; Singh, S. N.; Shire, S. J. The Influence of

Charge Distribution on Self-Association and Viscosity Behavior of Monoclonal Antibody Solutions. *Molecular Pharmaceutics* **2012**, *9* (4), 791-802. DOI: 10.1021/mp200566k.

(37) Wang, L.; Yang, X.; Wang, Q.; Zeng, Y.; Ding, L.; Jiang, W. Effects of ionic strength and temperature on the aggregation and deposition of multi-walled carbon nanotubes. *Journal of Environmental Sciences* **2017**, *51*, 248-255. DOI: https://doi.org/10.1016/j.jes.2016.07.003.

(38) Weaver, K. D.; Kim, H. J.; Sun, J.; MacFarlane, D. R.; Elliott, G. D. Cyto-toxicity and biocompatibility of a family of choline phosphate ionic liquids designed for pharmaceutical applications. *Green Chemistry* **2010**, *12* (3), 507-513. DOI: 10.1039/B918726J.

(39) Gontrani, L. Choline-amino acid ionic liquids: past and recent achievements about the structure and properties of these really "green" chemicals. *Biophys Rev* **2018**, *10* (3), 873-880. DOI: 10.1007/s12551-018-0420-9 From NLM.

(40) Amoroso, R.; Hollmann, F.; Maccallini, C. Choline Chloride-Based DES as Solvents/Catalysts/Chemical Donors in Pharmaceutical Synthesis. *Molecules* **2021**, *26* (20). DOI: 10.3390/molecules26206286 From NLM.

(41) Schröder, C. Proteins in Ionic Liquids: Current Status of Experiments and Simulations. *Topics in Current Chemistry* **2017**, *375* (2), 25. DOI: 10.1007/s41061-017-0110-2.

(42) Introduction: Ionic Liquids. *Chemical Reviews* **2017**, *117* (10), 6633-6635. DOI: 10.1021/acs.chemrev.7b00246.

(43) Veríssimo, N. V.; Vicente, F. A.; de Oliveira, R. C.; Likozar, B.; Oliveira, R. P. d. S.; Pereira, J. F. B. Ionic liquids as protein stabilizers for biological and biomedical applications: A review. *Biotechnology Advances* **2022**, *61*, 108055. DOI: https://doi.org/10.1016/j.biotechadv.2022.108055.

(44) Almeida, J. S.; Capela, E. V.; Loureiro, A. M.; Tavares, A. P. M.; Freire, M. G. An Overview on the Recent Advances in Alternative Solvents as Stabilizers of Proteins and Enzymes. In *ChemEngineering*, 2022; Vol. 6.

(45) Doshi, N.; Demeule, B.; Yadav, S. Understanding Particle Formation: Solubility of Free Fatty Acids as Polysorbate 20 Degradation Byproducts in Therapeutic Monoclonal Antibody Formulations. *Molecular Pharmaceutics* **2015**, *12* (11), 3792-3804. DOI: 10.1021/acs.molpharmaceut.5b00310.

(46) Bennion, B. J.; Daggett, V. The molecular basis for the chemical denaturation of proteins by urea. *Proc Natl Acad Sci U S A* **2003**, *100* (9), 5142-5147. DOI: 10.1073/pnas.0930122100 From NLM.

(47) Macdonald, R. D.; Khajehpour, M. Effects of the protein denaturant guanidinium chloride on aqueous hydrophobic contact-pair interactions. *Biophys Chem* **2015**, *196*, 25-32. DOI: 10.1016/j.bpc.2014.08.006 From NLM.

(48) He, F.; Woods, C. E.; Litowski, J. R.; Roschen, L. A.; Gadgil, H. S.; Razinkov, V. I.; Kerwin, B. A. Effect of Sugar Molecules on the Viscosity of High Concentration Monoclonal Antibody Solutions. *Pharmaceutical Research* **2011**, *28* (7), 1552-1560. DOI: 10.1007/s11095-011-0388-7.

(49) Murphy, K. P.; Freire, E.; Paterson, Y. Configurational effects in antibody-antigen interactions studied by microcalorimetry. *Proteins* **1995**, *21* (2), 83-90. DOI: 10.1002/prot.340210202 From NLM.

(50) Murphy, K. P.; Xie, D.; Thompson, K. S.; Amzel, L. M.; Freire, E. Entropy in biological binding processes: estimation of translational entropy loss. *Proteins* **1994**, *18* (1), 63-67. DOI: 10.1002/prot.340180108 From NLM.

(51) Fisher, H. F.; Singh, N. Calorimetric methods for interpreting protein-ligand interactions. *Methods Enzymol* **1995**, *259*, 194-221. DOI: 10.1016/0076-6879(95)59045-5 From NLM.

(52) Smyth, M. S.; Martin, J. H. x ray crystallography. *Mol Pathol* **2000**, *53* (1), 8-14. DOI: 10.1136/mp.53.1.8 From NLM.

(53) Wang, L.; Sigworth, F. J. Cryo-EM and Single Particles. *Physiology* **2006**, *21* (1), 13-18. DOI: 10.1152/physiol.00045.2005 (acccessed 2025/05/09).

(54) Renaud, J.-P.; Chari, A.; Ciferri, C.; Liu, W.-t.; Rémigy, H.-W.; Stark, H.; Wiesmann, C. Cryo-EM in drug discovery: achievements, limitations and prospects. *Nature Reviews Drug Discovery* **2018**, *17* (7), 471-492. DOI: 10.1038/nrd.2018.77.

(55) Chua, E. Y. D.; Mendez, J. H.; Rapp, M.; Ilca, S. L.; Tan, Y. Z.; Maruthi, K.; Kuang, H.; Zimanyi, C. M.; Cheng, A.; Eng, E. T.; et al. Better, Faster, Cheaper: Recent Advances in Cryo-Electron Microscopy. *Annu Rev Biochem* **2022**, *91*, 1-32. DOI: 10.1146/annurev-biochem-032620-110705 From NLM.

(56) Jalily Hasani, H.; Barakat, K. Homology Modeling: an Overview of Fundamentals and Tools. *2017* **2017**, *10* (2), Homology Modeling; Comparative Modeling; Drug Discovery; Structural Prediction; Template Identification; Homology Modeling Tools. DOI: 10.15866/iremos.v10i2.11412

129-145.

(57) Muhammed, M. T.; Aki-Yalcin, E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chemical Biology & Drug Design* **2019**, *93* (1), 12-20. DOI: https://doi.org/10.1111/cbdd.13388.

(58) Hameduh, T.; Haddad, Y.; Adam, V.; Heger, Z. Homology modeling in the time of collective and artificial intelligence. *Computational and Structural Biotechnology Journal* **2020**, *18*, 3494-3506. DOI: https://doi.org/10.1016/j.csbj.2020.11.007.

(59) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer,

F. T.; de Beer, T. A P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **2018**, *46* (W1), W296-W303. DOI: 10.1093/nar/gky427 (acccessed 5/10/2025).

(60) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583-589. DOI: 10.1038/s41586-021-03819-2.

(61) Clackson, T.; Wells, J. A. A hot spot of binding energy in a hormone-receptor interface. *Science* **1995**, *267* (5196), 383-386. DOI: 10.1126/science.7529940 From NLM.

(62) Ciulli, A.; Williams, G.; Smith, A. G.; Blundell, T. L.; Abell, C. Probing Hot Spots at Protein−Ligand Binding Sites: A Fragment-Based Approach Using Biophysical Methods. *Journal of Medicinal Chemistry* **2006**, *49* (16), 4992-5000. DOI: 10.1021/jm060490r.

(63) Yin, H.; Hamilton, A. D. Strategies for Targeting Protein–Protein Interactions With Synthetic Agents. *Angewandte Chemie International Edition* **2005**, *44* (27), 4130-4163. DOI: https://doi.org/10.1002/anie.200461786.

(64) Alvarez-Garcia, D.; Barril, X. Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. *Journal of Medicinal Chemistry* **2014**, *57* (20), 8530-8539. DOI: 10.1021/jm5010418.

(65) Guvench, O.; MacKerell, A. D., Jr. Computational fragment-based binding site identification by ligand competitive saturation. *PLoS Comput Biol* **2009**, *5* (7), e1000435. DOI: 10.1371/journal.pcbi.1000435 From NLM.

(66) Lexa, K. W.; Carlson, H. A. Full Protein Flexibility Is Essential for Proper Hot-Spot Mapping. *Journal of the American Chemical Society* **2011**, *133* (2), 200-202. DOI: 10.1021/ja1079332.

(67) Ghanakota, P.; Carlson, H. A. Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. *The Journal of Physical Chemistry B* **2016**, *120* (33), 8685-8695. DOI: 10.1021/acs.jpcb.6b03515.

(68) Lal Gupta, P.; Carlson, H. A. Cosolvent Simulations with Fragment-Bound Proteins Identify Hot Spots to Direct Lead Growth. *J Chem Theory Comput* **2022**, *18* (6), 3829-3844. DOI: 10.1021/acs.jctc.1c01054 From NLM.

(69) Ghanakota, P.; Carlson, H. A. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *J Med Chem* **2016**, *59* (23), 10383-10399. DOI: 10.1021/acs.jmedchem.6b00399 From NLM.

(70) Schmidt, D.; Boehm, M.; McClendon, C. L.; Torella, R.; Gohlke, H. Cosolvent-Enhanced Sampling and Unbiased Identification of Cryptic Pockets Suitable for Structure-Based Drug Design. *Journal of Chemical Theory and Computation* **2019**, *15* (5), 3331-

3343. DOI: 10.1021/acs.jctc.8b01295.

(71) Szabó, P. B.; Sabanés Zariquiey, F.; Nogueira, J. J. Cosolvent and Dynamic Effects in Binding Pocket Search by Docking Simulations. *Journal of Chemical Information and Modeling* **2021**, *61* (11), 5508-5523. DOI: 10.1021/acs.jcim.1c00924.

(72) Matubayasi, N.; Masutani, K. Energetics of cosolvent effect on peptide aggregation. *Biophys Physicobiol* **2019**, *16*, 185-195. DOI: 10.2142/biophysico.16.0_185 From NLM.

(73) Sabanés Zariquiey, F.; de Souza, J. V.; Bronowska, A. K. Cosolvent Analysis Toolkit (CAT): a robust hotspot identification platform for cosolvent simulations of proteins to expand the druggable proteome. *Scientific Reports* **2019**, *9* (1), 19118. DOI: 10.1038/s41598-019-55394-2.

(74) Bruciaferri, N.; Eberhardt, J.; Llanos, M. A.; Loeffler, J. R.; Holcomb, M.; Fernandez-Quintero, M. L.; Santos-Martins, D.; Ward, A. B.; Forli, S. CosolvKit: a Versatile Tool for Cosolvent MD Preparation and Analysis. *Journal of Chemical Information and Modeling* **2024**, *64* (21), 8227-8235. DOI: 10.1021/acs.jcim.4c01398.

(75) Kulkarni, P. U.; Shah, H.; Vyas, V. K. Hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) Simulation: A Tool for Structure-Based Drug Design and Discovery. *Mini Rev Med Chem* **2022**, *22* (8), 1096-1107. DOI: 10.2174/1389557521666211007115250 From NLM.

(76) Vanommeslaeghe, K.; Guvench, O.; MacKerell, A. D., Jr. Molecular mechanics. *Curr Pharm Des* **2014**, *20* (20), 3281-3292. DOI: 10.2174/13816128113199990600 From NLM.

(77) Ekimoto, T.; Ikeguchi, M. Multiscale molecular dynamics simulations of rotary motor proteins. *Biophysical Reviews* **2018**, *10* (2), 605-615. DOI: 10.1007/s12551-017-0373-4.

(78) Engkvist, O.; Norrby, P.-O.; Selmi, N.; Lam, Y.-h.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational prediction of chemical reactions: current status and outlook. *Drug Discovery Today* **2018**, *23* (6), 1203-1218. DOI: https://doi.org/10.1016/j.drudis.2018.02.014.

(79) Singh, N.; Li, W. Recent Advances in Coarse-Grained Models for Biomolecules and Their Applications. In *International Journal of Molecular Sciences*, 2019; Vol. 20.

(80) Scheraga, H. A.; Khalili, M.; Liwo, A. Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. *Annual Review of Physical Chemistry* **2007**, *58* (1), 57-83. DOI: 10.1146/annurev.physchem.58.032806.104614 (acccessed 2022/12/14).

(81) Tozzini, V. Coarse-grained models for proteins. *Current Opinion in Structural Biology* **2005**, *15* (2), 144-150. DOI: https://doi.org/10.1016/j.sbi.2005.02.005.

(82) Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology* **1976**, *104* (1), 59-107. DOI:

https://doi.org/10.1016/0022-2836(76)90004-8.

(83) Ingólfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The power of coarse graining in biomolecular simulations. *Wiley Interdiscip Rev Comput Mol Sci* **2014**, *4* (3), 225-248. DOI: 10.1002/wcms.1169 From NLM.

(84) Pak, A. J.; Voth, G. A. Advances in coarse-grained modeling of macromolecular complexes. *Curr Opin Struct Biol* **2018**, *52*, 119-126. DOI: 10.1016/j.sbi.2018.11.005 From NLM.

(85) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chemical Reviews* **2016**, *116* (14), 7898-7936. DOI: 10.1021/acs.chemrev.6b00163.

(86) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *The Journal of Physical Chemistry B* **2004**, *108* (2), 750-760. DOI: 10.1021/jp036508g.

(87) Periole, X.; Marrink, S.-J. The Martini Coarse-Grained Force Field. In *Biomolecular Simulations: Methods and Protocols*, Monticelli, L., Salonen, E. Eds.; Humana Press, 2013; pp 533-565.

(88) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *The Journal of Physical Chemistry B* **2007**, *111* (27), 7812-7824. DOI: 10.1021/jp071097f.

(89) Wassenaar, T. A.; Pluhackova, K.; Böckmann, R. A.; Marrink, S. J.; Tieleman, D. P. Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. *Journal of Chemical Theory and Computation* **2014**, *10* (2), 676-690. DOI: 10.1021/ct400617g.

(90) Darré, L.; Machado, M. R.; Brandner, A. F.; González, H. C.; Ferreira, S.; Pantano, S. SIRAH: A Structurally Unbiased Coarse-Grained Force Field for Proteins with Aqueous Solvation and Long-Range Electrostatics. *Journal of Chemical Theory and Computation* **2015**, *11* (2), 723-739. DOI: 10.1021/ct5007746.

(91) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* **2005**, *26* (16), 1701-1718. DOI: https://doi.org/10.1002/jcc.20291.

(92) Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **2009**, *25* (5), 621-627. DOI: 10.1093/bioinformatics/btp036.

(93) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia,

B.; Beglov, D.; Vajda, S. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature Protocols* **2015**, *10* (5), 733-755. DOI: 10.1038/nprot.2015.043.

(94) Wells, J. A.; McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* **2007**, *450* (7172), 1001-1009. DOI: 10.1038/nature06526.

(95) Ma, B.; Shatsky, M.; Wolfson, H. J.; Nussinov, R. Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations. *Protein Science* **2002**, *11* (2), 184-197. DOI: https://doi.org/10.1110/ps.21302.

(96) Saphire, E. O.; Parren, P. W. H. I.; Pantophlet, R.; Zwick, M. B.; Morris, G. M.; Rudd, P. M.; Dwek, R. A.; Stanfield, R. L.; Burton, D. R.; Wilson, I. A. Crystal Structure of a Neutralizing Human IgG Against HIV-1: A Template for Vaccine Design. *Science* **2001**, *293* (5532), 1155-1159. DOI: 10.1126/science.1061692.

(97) Lin, X.; Zhang, X. Identification of hot regions in hub protein–protein interactions by clustering and PPRA optimization. *BMC Medical Informatics and Decision Making* **2021**, *21* (1), 143. DOI: 10.1186/s12911-020-01350-4.

(98) Keskin, O.; Ma, B.; Nussinov, R. Hot Regions in Protein–Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues. *Journal of Molecular Biology* **2005**, *345* (5), 1281-1294. DOI: https://doi.org/10.1016/j.jmb.2004.10.077.

(99) Reulecke, I.; Lange, G.; Albrecht, J.; Klein, R.; Rarey, M. Towards an Integrated Description of Hydrogen Bonding and Dehydration: Decreasing False Positives in Virtual Screening with the HYDE Scoring Function. *ChemMedChem* **2008**, *3* (6), 885-897. DOI: https://doi.org/10.1002/cmdc.200700319 (acccessed 2025/05/10).

(100) Schneider, N.; Lange, G.; Hindle, S.; Klein, R.; Rarey, M. A consistent description of HYdrogen bond and DEhydration energies in protein–ligand complexes: methods behind the HYDE scoring function. *Journal of Computer-Aided Molecular Design* **2013**, *27* (1), 15-29. DOI: 10.1007/s10822-012-9626-2.

(101) Morea, V.; Lesk, A. M.; Tramontano, A. Antibody modeling: implications for engineering and design. *Methods* **2000**, *20* (3), 267-279. DOI: 10.1006/meth.1999.0921 From NLM.

(102) Schiel, J. E.; Turner, A.; Mouchahoir, T.; Yandrofski, K.; Telikepalli, S.; King, J.; DeRose, P.; Ripple, D.; Phinney, K. The NISTmAb Reference Material 8671 value assignment, homogeneity, and stability. *Anal Bioanal Chem* **2018**, *410* (8), 2127-2139. DOI: 10.1007/s00216-017-0800-1 From NLM.

(103) Xu, A. Y.; Castellanos, M. M.; Mattison, K.; Krueger, S.; Curtis, J. E. Studying Excipient Modulated Physical Stability and Viscosity of Monoclonal Antibody

Formulations Using Small-Angle Scattering. *Molecular Pharmaceutics* **2019**, *16* (10), 4319-4338. DOI: 10.1021/acs.molpharmaceut.9b00687.

(104) Qin, S.; Zhou, H.-X. Calculation of Second Virial Coefficients of Atomistic Proteins Using Fast Fourier Transform. *The Journal of Physical Chemistry B* **2019**, *123* (39), 8203-8215. DOI: 10.1021/acs.jpcb.9b06808.

(105) Schleinitz, M.; Teschner, D.; Sadowski, G.; Brandenbusch, C. Second osmotic virial coefficients of therapeutic proteins in the presence of excipient-mixtures can be predicted to aid an efficient formulation design. *Journal of Molecular Liquids* **2019**, *283*, 575-583. DOI: https://doi.org/10.1016/j.molliq.2019.03.064.

(106) Formolo, T.; Ly, M.; Levy, M.; Kilpatrick, L.; Lute, S.; Phinney, K.; Marzilli, L.; Brorson, K.; Boyne, M.; Davis, D.; Schiel, J. Determination of the NISTmAb Primary Structure. In *State-of-the-Art and Emerging Technologies for Therapeutic Monoclonal Antibody Characterization Volume 2. Biopharmaceutical Characterization: The NISTmAb Case Study*, ACS Symposium Series, Vol. 1201; American Chemical Society, 2015; pp 1-62.

(107) Saurabh, S.; Kalonia, C.; Li, Z.; Hollowell, P.; Waigh, T.; Li, P.; Webster, J.; Seddon, J. M.; Lu, J. R.; Bresme, F. Understanding the Stabilizing Effect of Histidine on mAb Aggregation: A Molecular Dynamics Study. *Molecular Pharmaceutics* **2022**, *19* (9), 3288-3303. DOI: 10.1021/acs.molpharmaceut.2c00453.

(108) Maity, H.; O'Dell, C.; Srivastava, A.; Goldstein, J. Effects of Arginine on Photostability and Thermal Stability of IgG1 Monoclonal Antibodies. *Current Pharmaceutical Biotechnology* **2009**, *10* (8), 761-766. DOI: http://dx.doi.org/10.2174/138920109789978711.

(109) Platts, L.; Falconer, R. J. Controlling protein stability: Mechanisms revealed using formulations of arginine, glycine and guanidinium HCl with three globular proteins. *International Journal of Pharmaceutics* **2015**, *486* (1), 131-135. DOI: https://doi.org/10.1016/j.ijpharm.2015.03.051.

(110) Hagan, J. B.; Wasserman, R. L.; Baggish, J. S.; Spycher, M. O.; Berger, M.; Shashi, V.; Lohrmann, E.; Sullivan, K. E. Safety of L-proline as a stabilizer for immunoglobulin products. *Expert Rev Clin Immunol* **2012**, *8* (2), 169-178. DOI: 10.1586/eci.11.97 From NLM.

(111) Hung, J. J.; Dear, B. J.; Dinin, A. K.; Borwankar, A. U.; Mehta, S. K.; Truskett, T. T.; Johnston, K. P. Improving Viscosity and Stability of a Highly Concentrated Monoclonal Antibody Solution with Concentrated Proline. *Pharm Res* **2018**, *35* (7), 133. DOI: 10.1007/s11095-018-2398-1 From NLM.

(112) Kendrick, B. S.; Chang, B. S.; Arakawa, T.; Peterson, B.; Randolph, T. W.; Manning, M. C.; Carpenter, J. F. Preferential exclusion of sucrose from recombinant interleukin-1 receptor antagonist: role in restricted conformational mobility and compaction of native

state. *Proc Natl Acad Sci U S A* **1997**, *94* (22), 11917-11922. DOI: 10.1073/pnas.94.22.11917 From NLM.

(113) Jain, N. K.; Roy, I. Effect of trehalose on protein structure. *Protein Sci* **2009**, *18* (1), 24-36. DOI: 10.1002/pro.3 From NLM.

(114) Wen, L.; Chen, Y.; Liao, J.; Zheng, X.; Yin, Z. Preferential interactions between protein and arginine: Effects of arginine on tertiary conformational and colloidal stability of protein solution. *International Journal of Pharmaceutics* **2015**, *478* (2), 753-761. DOI: https://doi.org/10.1016/j.ijpharm.2014.12.038.

(115) Zwijsen, R. M.; Buckle, R. S.; Hijmans, E. M.; Loomans, C. J.; Bernards, R. Ligand-independent recruitment of steroid receptor coactivators to estrogen receptor by cyclin D1. *Genes Dev* **1998**, *12* (22), 3488-3498. DOI: 10.1101/gad.12.22.3488 From NLM.

(116) Nilsson, S.; Mäkelä, S.; Treuter, E.; Tujague, M.; Thomsen, J.; Andersson, G.; Enmark, E.; Pettersson, K.; Warner, M.; Gustafsson, J. A. Mechanisms of estrogen action. *Physiol Rev* **2001**, *81* (4), 1535-1565. DOI: 10.1152/physrev.2001.81.4.1535 From NLM.

(117) Rosenfeld, M. G.; Glass, C. K. Coregulator codes of transcriptional regulation by nuclear receptors. *J Biol Chem* **2001**, *276* (40), 36865-36868. DOI: 10.1074/jbc.R100041200 From NLM.

(118) Deroo, B. J.; Korach, K. S. Estrogen receptors and human disease. *J Clin Invest* **2006**, *116* (3), 561-570. DOI: 10.1172/jci27987 From NLM.

(119) Leitch, A. C.; Abdelghany, T. M.; Probert, P. M.; Dunn, M. P.; Meyer, S. K.; Palmer, J. M.; Cooke, M. P.; Blake, L. I.; Morse, K.; Rosenmai, A. K.; et al. The toxicity of the methylimidazolium ionic liquids, with a focus on M8OI and hepatic effects. *Food Chem Toxicol* **2020**, *136*, 111069. DOI: 10.1016/j.fct.2019.111069 From NLM.

(120) Probert, P. M.; Leitch, A. C.; Dunn, M. P.; Meyer, S. K.; Palmer, J. M.; Abdelghany, T. M.; Lakey, A. F.; Cooke, M. P.; Talbot, H.; Wills, C.; et al. Identification of a xenobiotic as a potential environmental trigger in primary biliary cholangitis. *J Hepatol* **2018**, *69* (5), 1123-1135. DOI: 10.1016/j.jhep.2018.06.027 From NLM.

(121) Leitch, A. C.; Lakey, A. F.; Hotham, W. E.; Agius, L.; Kass, G. E. N.; Blain, P. G.; Wright, M. C. The ionic liquid 1-octyl-3-methylimidazolium (M8OI) is an activator of the human estrogen receptor alpha. *Biochem Biophys Res Commun* **2018**, *503* (3), 2167-2172. DOI: 10.1016/j.bbrc.2018.08.008 From NLM.

(122) Jumper, J.; Hassabis, D. Protein structure predictions to atomic accuracy with AlphaFold. *Nat Methods* **2022**, *19* (1), 11-12. DOI: 10.1038/s41592-021-01362-6 From NLM.

(123) Huang, W.; Peng, Y.; Kiselar, J.; Zhao, X.; Albaqami, A.; Mendez, D.; Chen, Y.; Chakravarthy, S.; Gupta, S.; Ralston, C.; et al. Multidomain architecture of estrogen

receptor reveals interfacial cross-talk between its DNA-binding and ligand-binding domains. *Nature Communications* **2018**, *9* (1), 3520. DOI: 10.1038/s41467-018-06034-2.

(124) Kataura, T.; Otten, E. G.; Rabanal-Ruiz, Y.; Adriaenssens, E.; Urselli, F.; Scialo, F.; Fan, L.; Smith, G. R.; Dawson, W. M.; Chen, X.; et al. NDP52 acts as a redox sensor in PINK1/Parkin-mediated mitophagy. *The EMBO Journal* **2023**, *42* (5), e111372. DOI: https://doi.org/10.15252/embj.2022111372.

(125) Guzenko, D.; Strelkov, S. V. CCFold: rapid and accurate prediction of coiled-coil structures and application to modelling intermediate filaments. *Bioinformatics* **2018**, *34* (2), 215-222. DOI: 10.1093/bioinformatics/btx551 From NLM.

(126) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* **2021**, 2021.2010.2004.463034. DOI: 10.1101/2021.10.04.463034.

(127) Wood, C. W.; Woolfson, D. N. CCBuilder 2.0: Powerful and accessible coiled-coil modeling. *Protein Sci* **2018**, *27* (1), 103-111. DOI: 10.1002/pro.3279 From NLM.

(128) Moran, J. V.; Holmes, S. E.; Naas, T. P.; DeBerardinis, R. J.; Boeke, J. D.; Kazazian, H. H., Jr. High frequency retrotransposition in cultured mammalian cells. *Cell* **1996**, *87* (5), 917-927. DOI: 10.1016/s0092-8674(00)81998-4 From NLM.

(129) Feng, Q.; Moran, J. V.; Kazazian, H. H., Jr.; Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **1996**, *87* (5), 905-916. DOI: 10.1016/s0092-8674(00)81997-2 From NLM.

(130) Lavie, L.; Maldener, E.; Brouha, B.; Meese, E. U.; Mayer, J. The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res* **2004**, *14* (11), 2253-2260. DOI: 10.1101/gr.2745804 From NLM.

(131) Hohjoh, H.; Singer, M. F. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *Embo j* **1996**, *15* (3), 630-639. From NLM.

(132) Kulpa, D. A.; Moran, J. V. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum Mol Genet* **2005**, *14* (21), 3237-3248. DOI: 10.1093/hmg/ddi354 From NLM.

(133) Cook, P. R.; Jones, C. E.; Furano, A. V. Phosphorylation of ORF1p is required for L1 retrotransposition. *Proc Natl Acad Sci U S A* **2015**, *112* (14), 4298-4303. DOI: 10.1073/pnas.1416869112 From NLM.

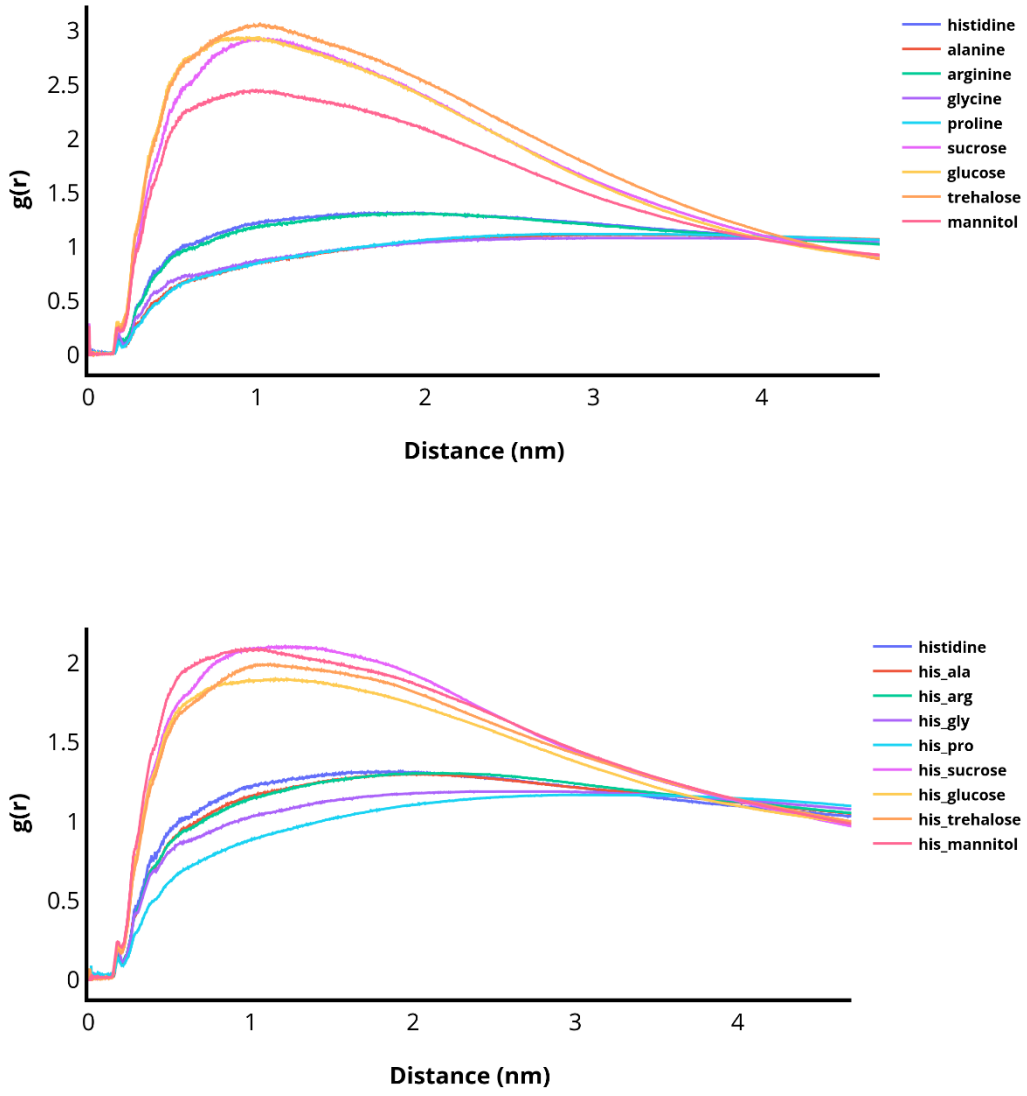
(134) Jung, J.; Nishima, W.; Daniels, M.; Bascom, G.; Kobayashi, C.; Adedoyin, A.; Wall, M.; Lappala, A.; Phillips, D.; Fischer, W.; et al. Scaling molecular dynamics beyond 100,000 processor cores for large-scale biophysical simulations. *Journal of Computational Chemistry* **2019**, *40* (21), 1919-1930. DOI: https://doi.org/10.1002/jcc.25840 (acccessed

2025/05/15).

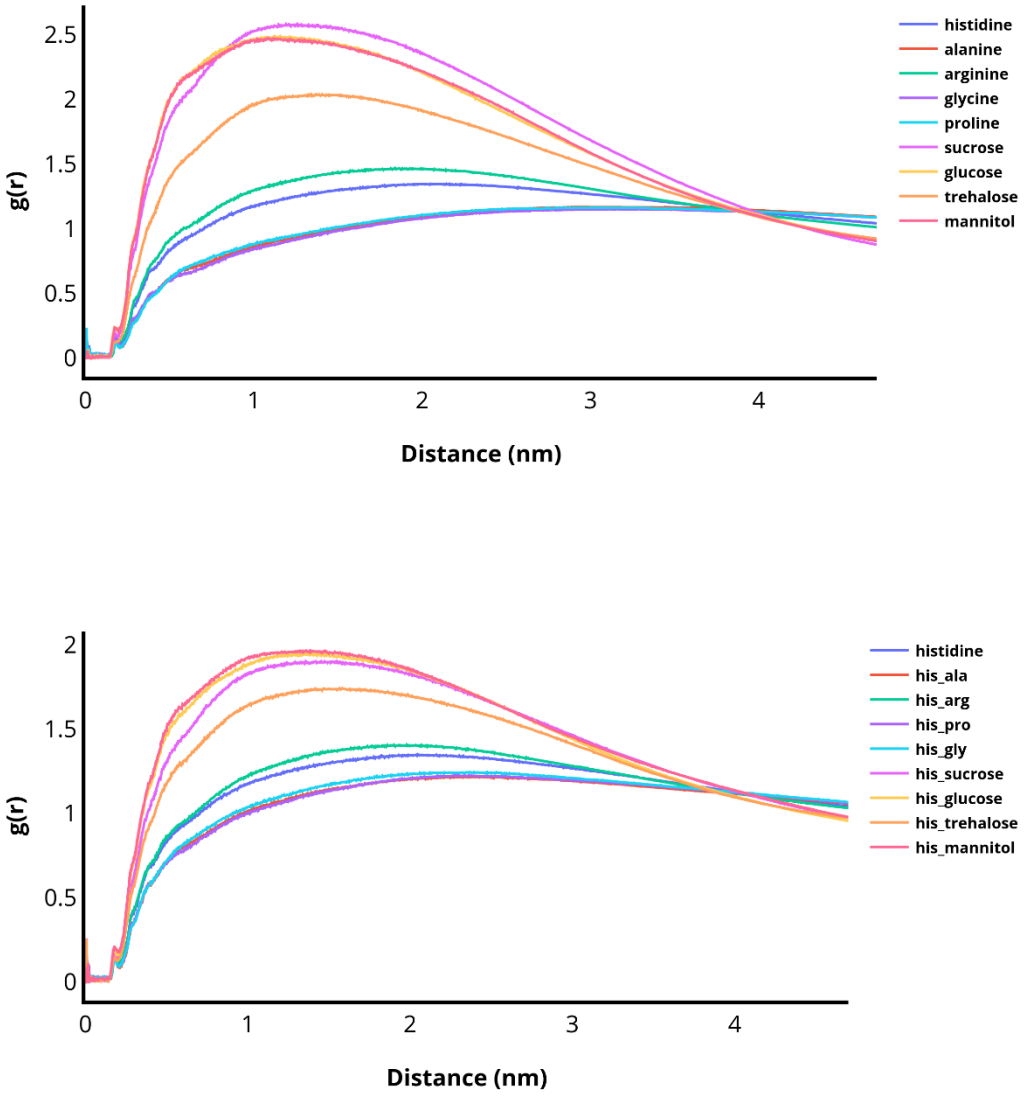
(135) Stevens, J. A.; Grünewald, F.; van Tilburg, P. A. M.; König, M.; Gilbert, B. R.; Brier, T. A.; Thornburg, Z. R.; Luthey-Schulten, Z.; Marrink, S. J. Molecular dynamics simulation of an entire cell. *Front Chem* **2023**, *11*, 1106495. DOI: 10.3389/fchem.2023.1106495 From NLM.

Appendix

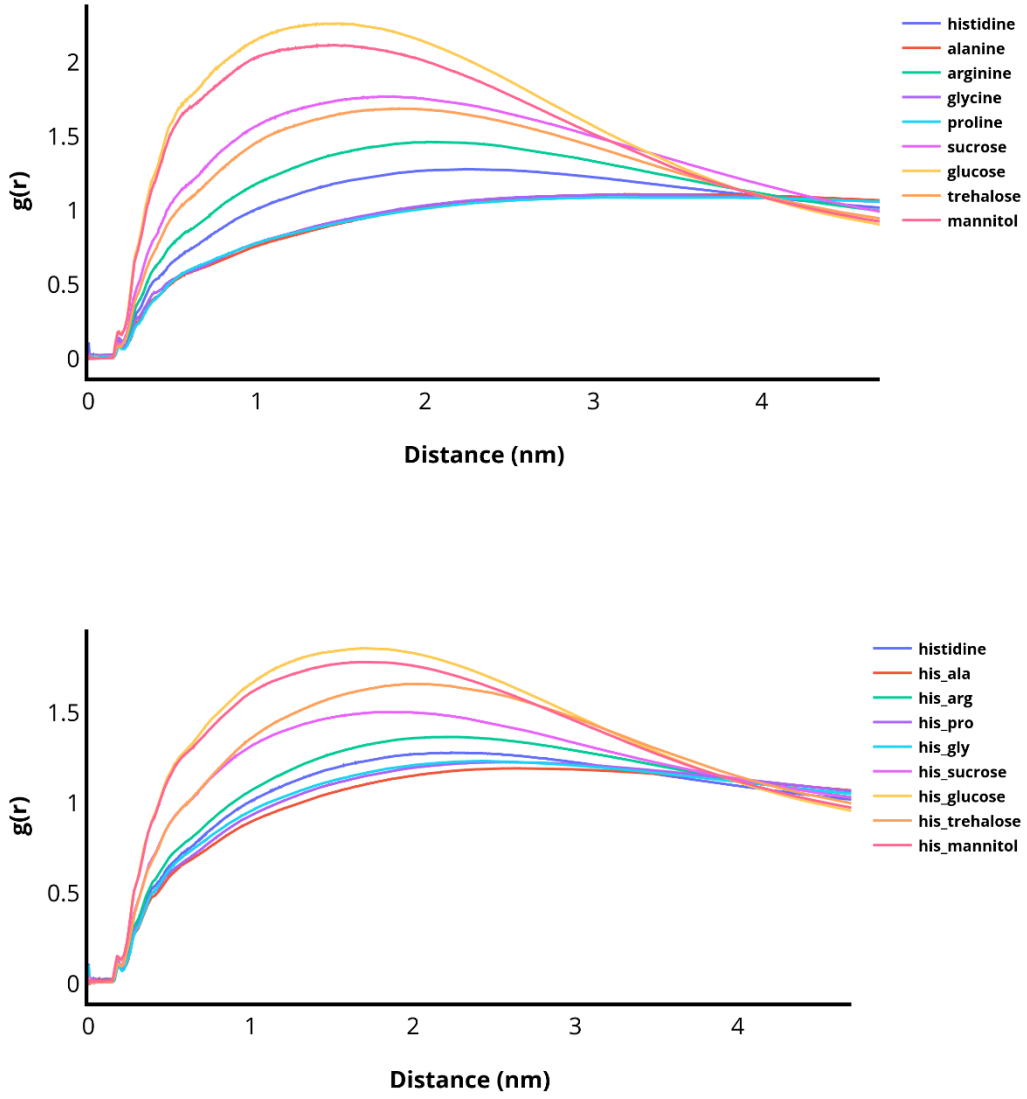
Fab - RDF



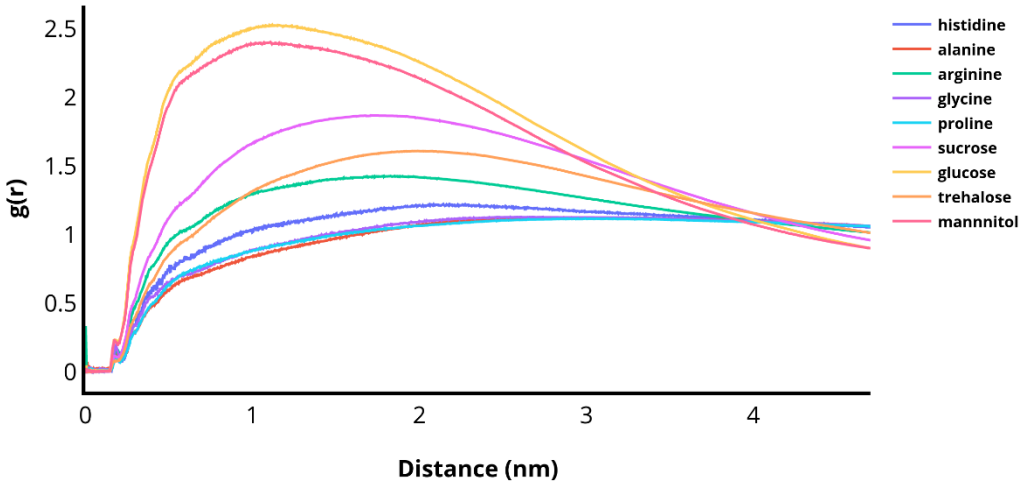
A 1: The average RDF plots of fab region of 3 replicas in 0.5% v/v solvents. The top graph is in single solvents - histidine, alanine, arginine, glycine, proline, sucrose, glucose, trehalose, and mannitol. The bottom graph is in a mixture of two solvents, 0.25% v/v histidine as the buffer with 0.25% v/v of the solvents mentioned above, pure histidine line was kept as a comparison.

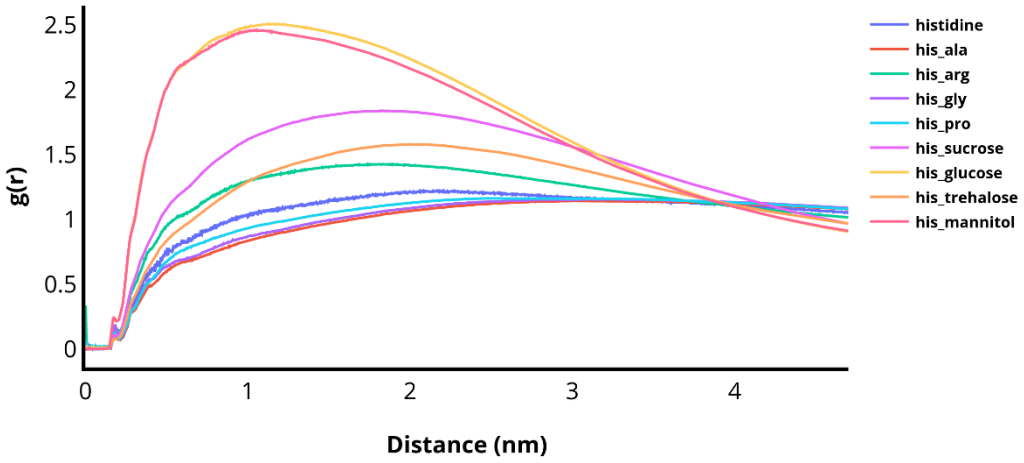


A 2: The average RDF plots of fab region of 3 replicas in 2.5% v/v solvents. The top graph is in single solvents - histidine, alanine, arginine, glycine, proline, sucrose, glucose, trehalose, and mannitol. The bottom graph is in a mixture of two solvents, 1.25% v/v histidine as the buffer with 1.25% v/v of the solvents mentioned above, pure histidine line was kept as a comparison.



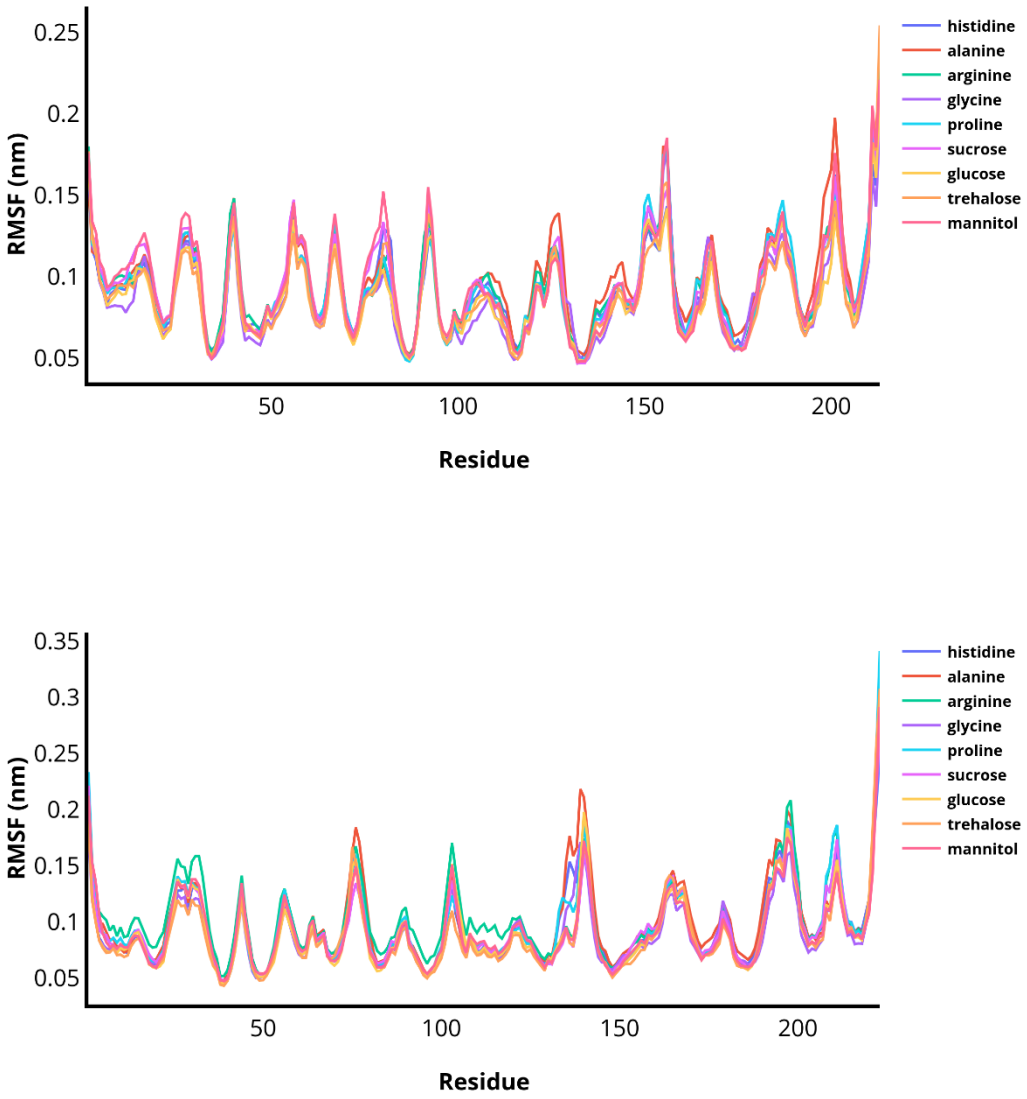
A 3: The average RDF plots of fab region of 3 replicas in 5% v/v solvents. The top graph is in single solvents - histidine, alanine, arginine, glycine, proline, sucrose, glucose, trehalose, and mannitol. The bottom graph is in a mixture of two solvents, 2.5% v/v histidine as the buffer with 2.5% v/v of the solvents mentioned above, pure histidine line was kept as a comparison.



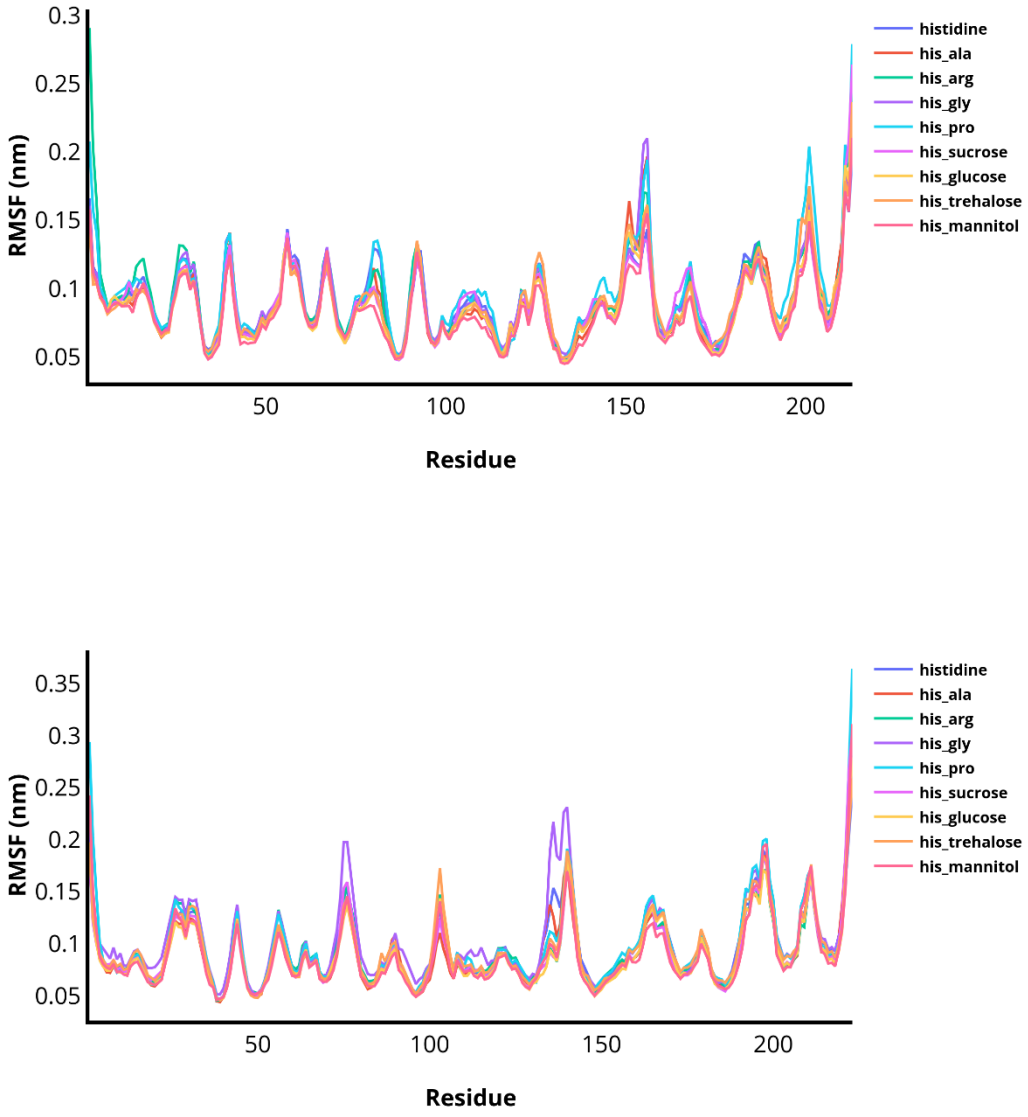


A 4: The average RDF plots of fab region of 3 replicas in different solvents. The top graph is in single solvents – 25 mM histidine, 200 mM alanine, 171 mM arginine, 200 mM glycine, 200 mM proline, 300 mM sucrose, 300 mM glucose, 300 mM trehalose, and 300 mM mannitol. The bottom graph is in a mixture of two solvents, 25mM histidine as the buffer with solvents mentioned above, pure histidine line was kept as a comparison.

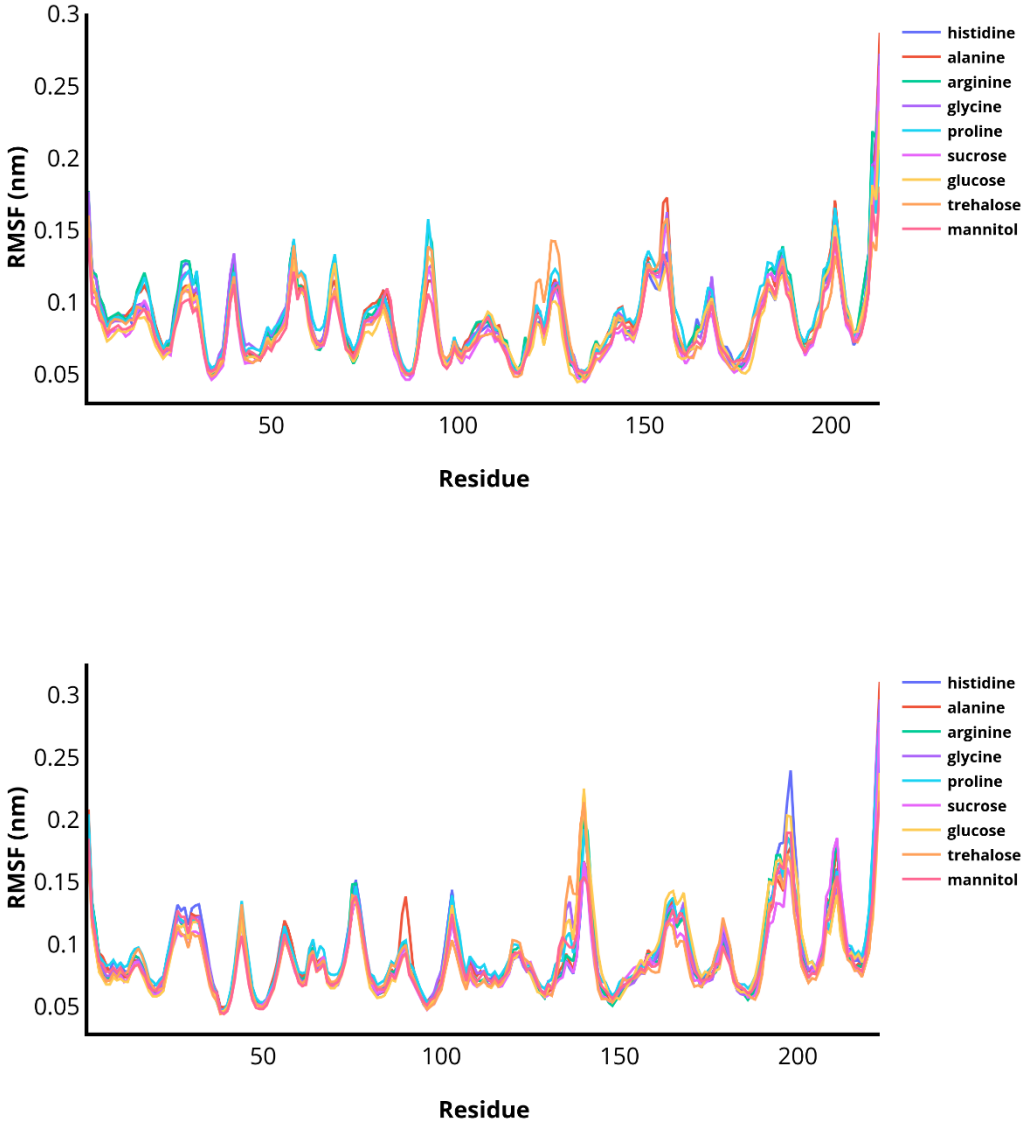
Fab - RMSF



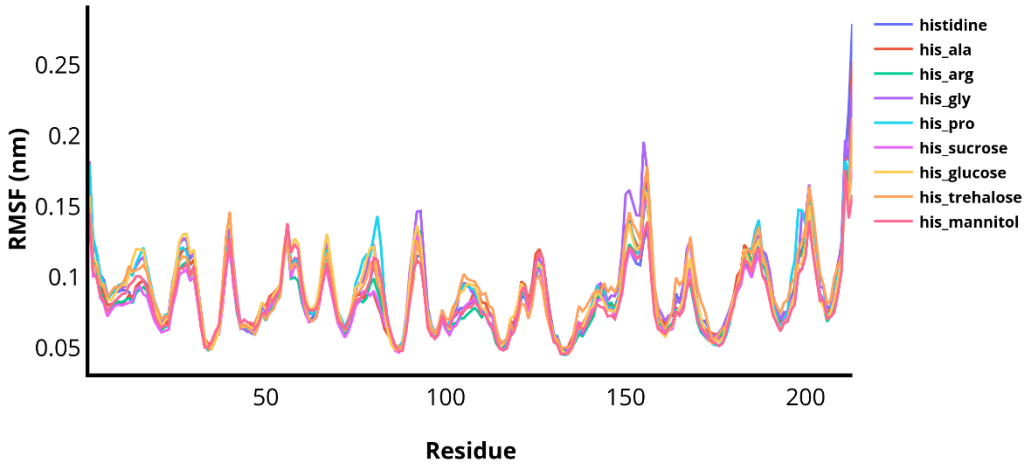
A 5: Average RMSF graphs showing the fluctuations of each residue of the light and heavy chains of the fab region in different 0.5% v/v single solvents. The top one is the light chain and the bottom one is the heavy chain.

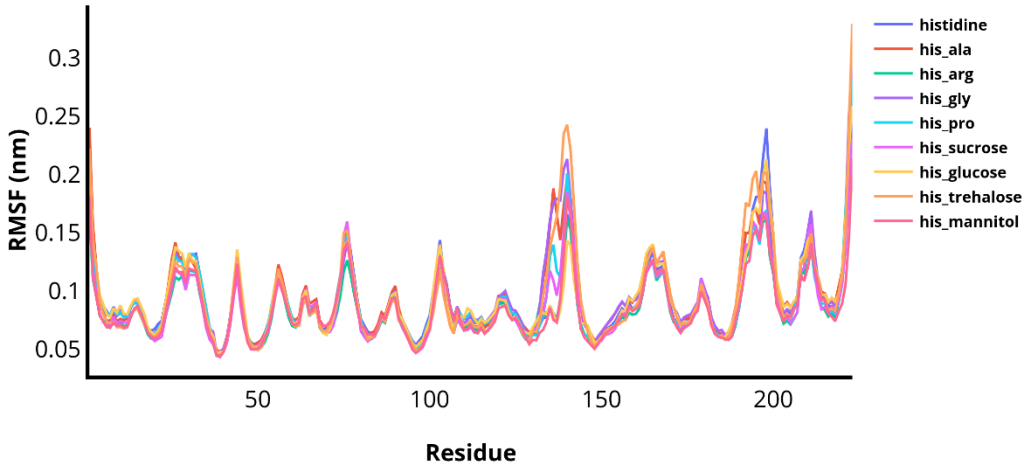


A 6: Average RMSF graphs showing the fluctuations of each residue of the light and heavy chains of the fab region in different 0.5% v/v cosolvents. The top one is the light chain and the bottom one is the heavy chain.

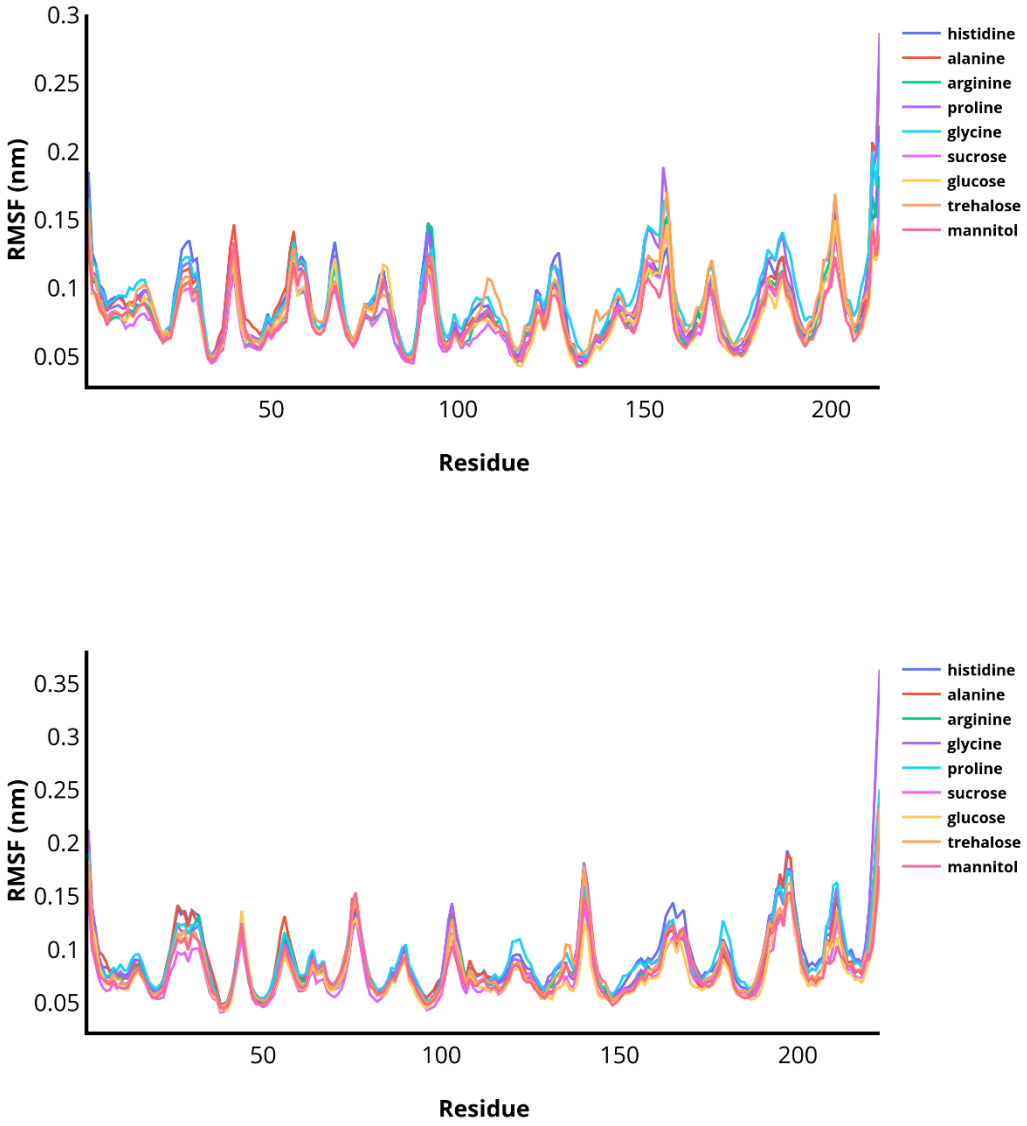


A 7: Average RMSF graphs showing the fluctuations of each residue of the light and heavy chains of the fab region in different 2.5% v/v single solvents. The top one is the light chain and the bottom one is the heavy chain.

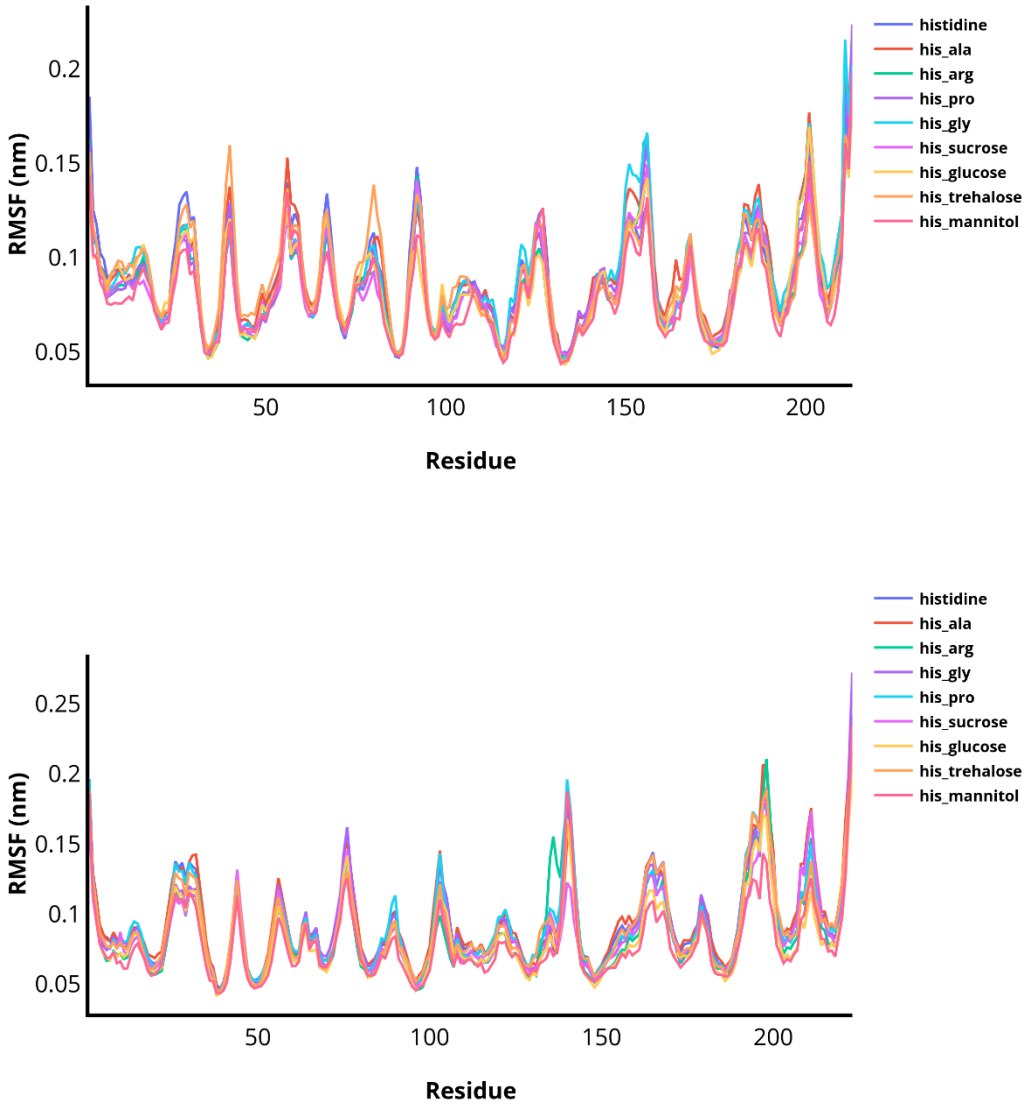




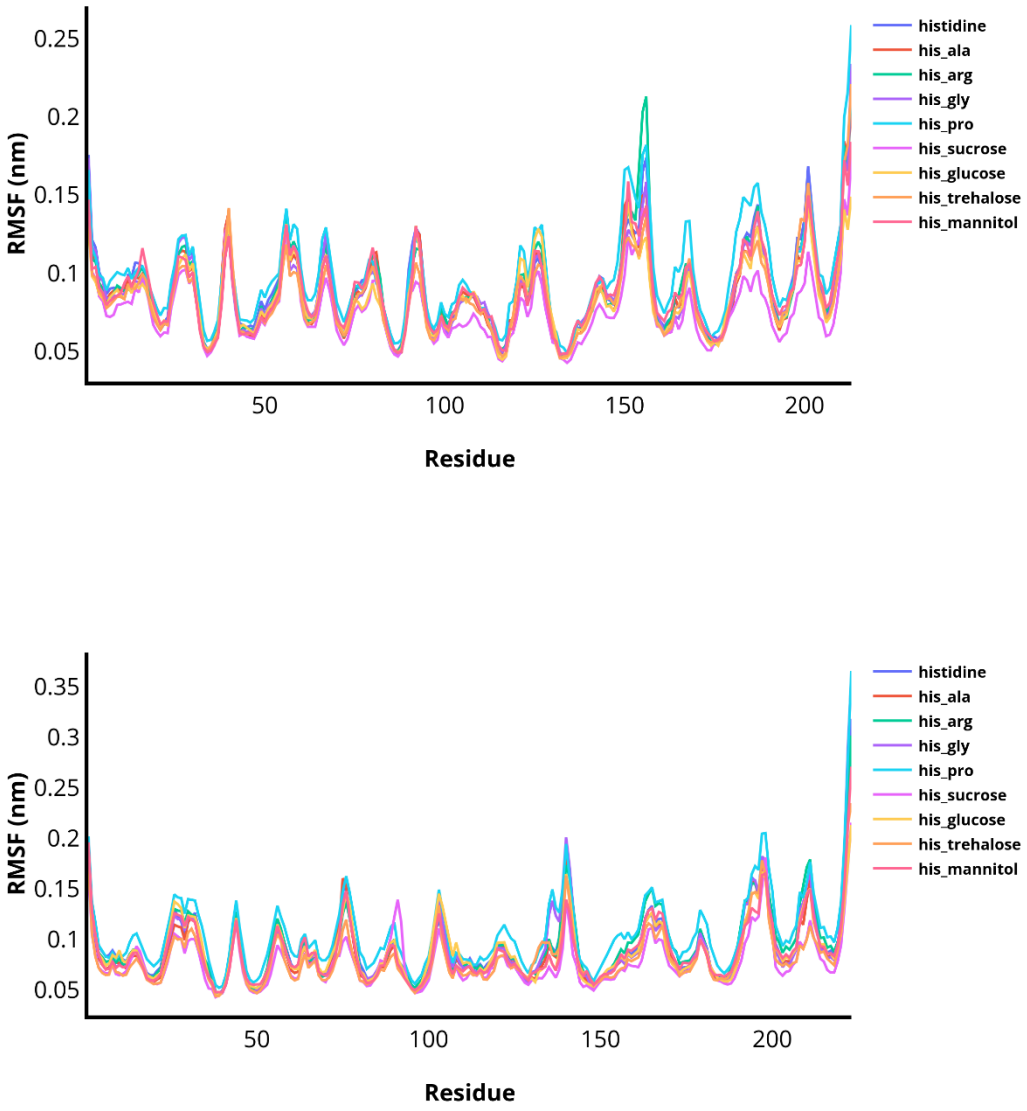
A 8: Average RMSF graphs showing the fluctuations of each residue of the light and heavy chains of the fab region in different 2.5% v/v cosolvents. The top one is the light chain and the bottom one is the heavy chain.



A 9: Average RMSF graphs showing the fluctuations of each residue of the light and heavy chains of the fab region in different 5% v/v single solvents. The top one is the light chain and the bottom one is the heavy chain.

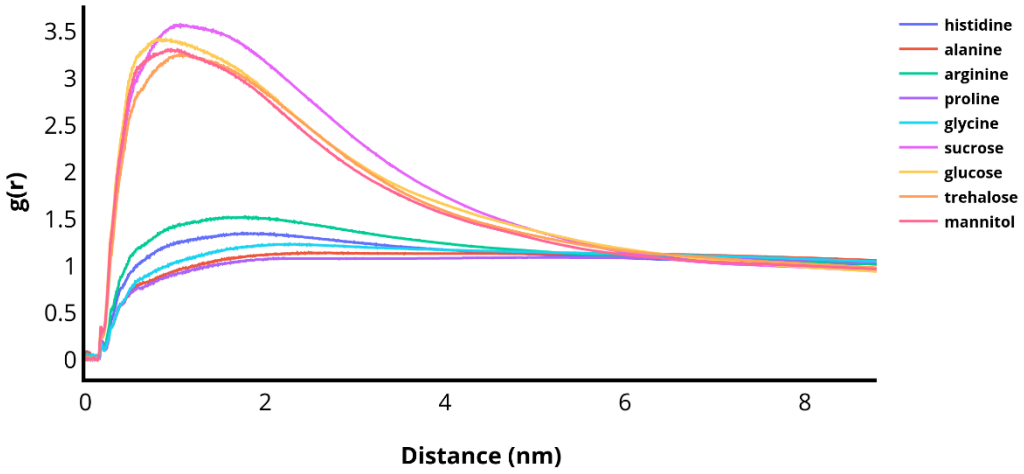


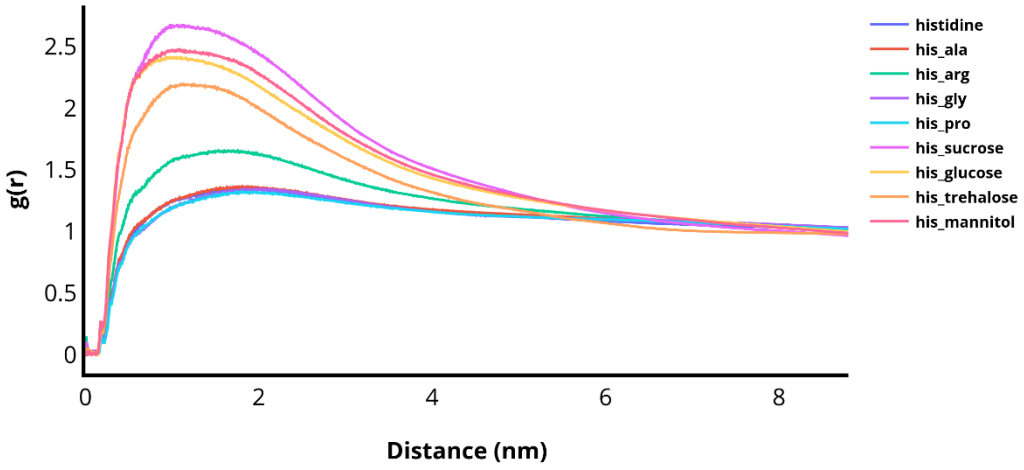
A 10: Average RMSF graphs showing the fluctuations of each residue of the light and heavy chains of the fab region in different 5% v/v cosolvents. The top one is the light chain and the bottom one is the heavy chain.



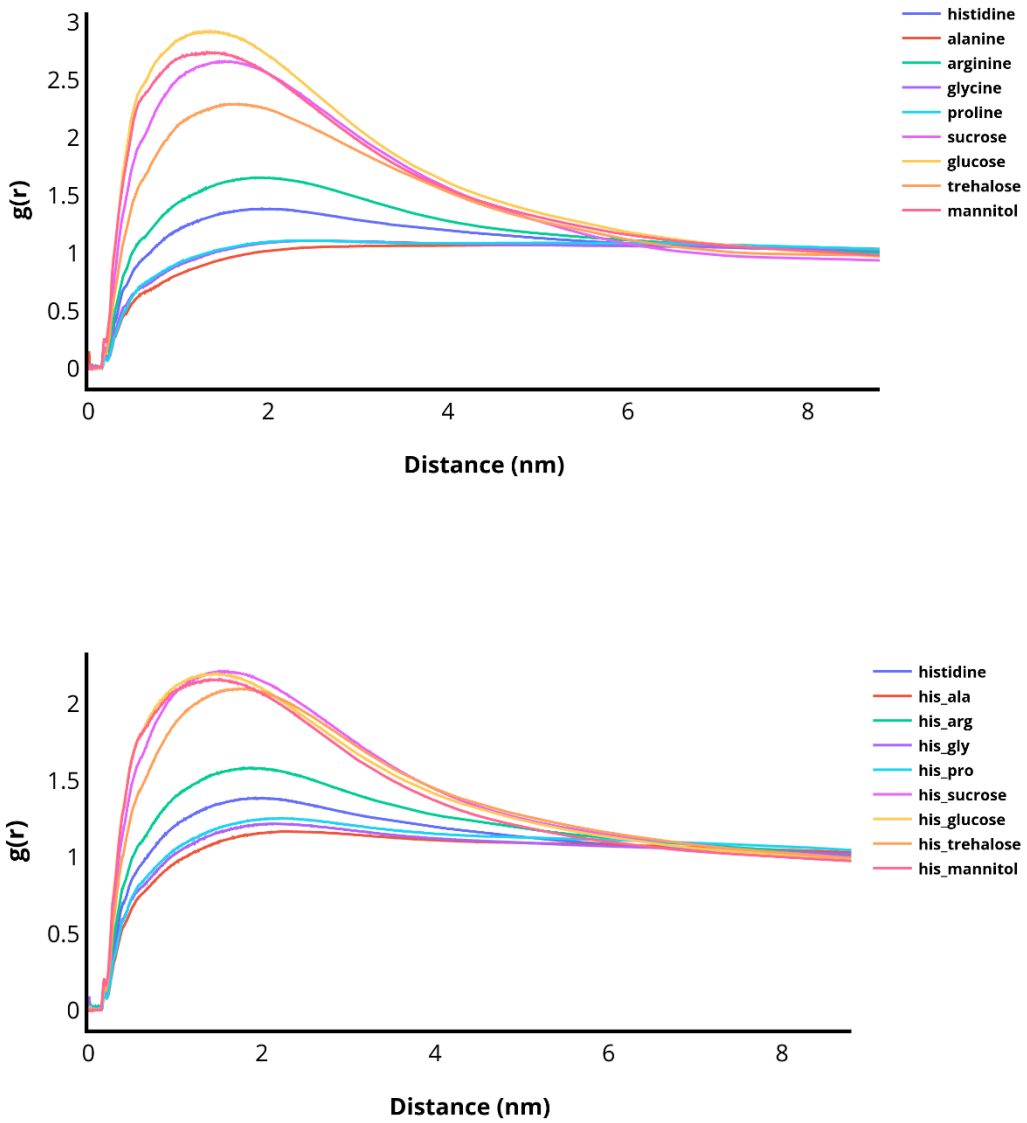
A 11: Average RMSF graphs showing the fluctuations of each residue of the light and heavy chains of the fab region in different cosolvents at paper concentrations. The top one is the light chain and the bottom one is the heavy chain.

mAb - RDF

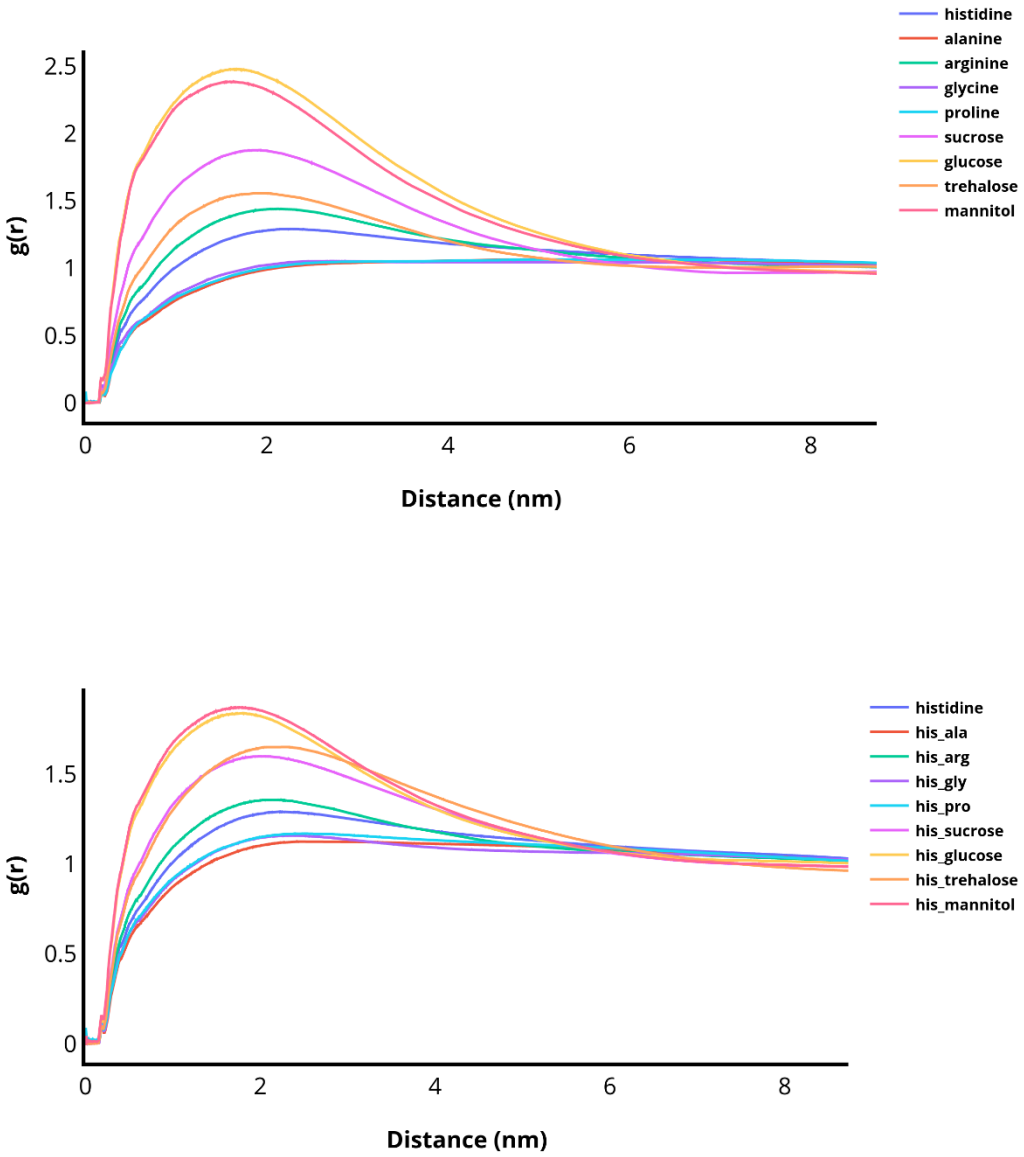




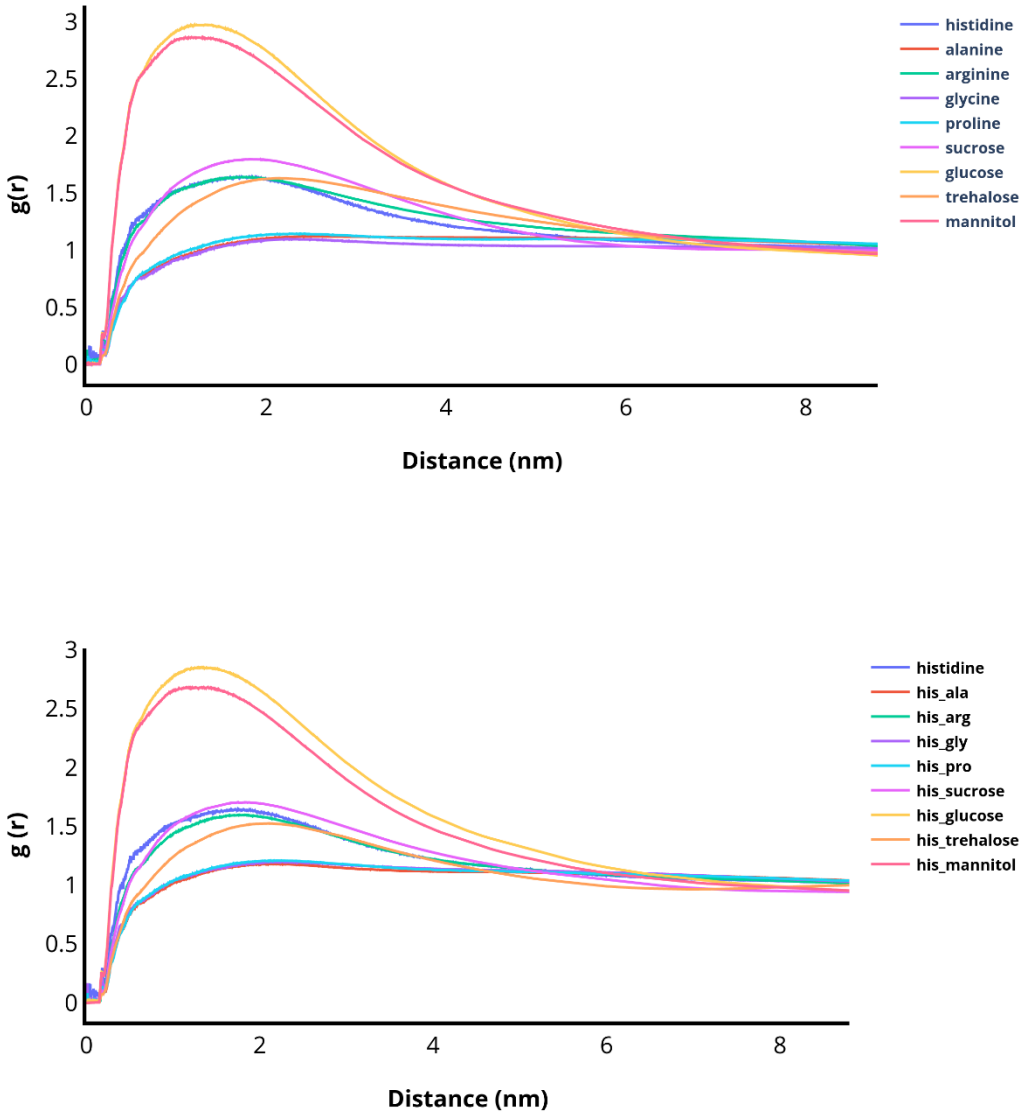
A 12: The average RDF plots of the whole mAb of 3 replicas in 0.5% v/v solvents. The top graph is in single solvents - histidine, alanine, arginine, glycine, proline, sucrose, glucose, trehalose, and mannitol. The bottom graph is in a mixture of two solvents, 0.25% v/v histidine as the buffer with 0.25% v/v of the solvents mentioned above, pure histidine line was kept as a comparison.



A 13: The average RDF plots of the whole mAb of 3 replicas in 2.5% v/v solvents. The top graph is in single solvents - histidine, alanine, arginine, glycine, proline, sucrose, glucose, trehalose, and mannitol. The bottom graph is in a mixture of two solvents, 1.25% v/v histidine as the buffer with 1.25% v/v of the solvents mentioned above, pure histidine line was kept as a comparison.

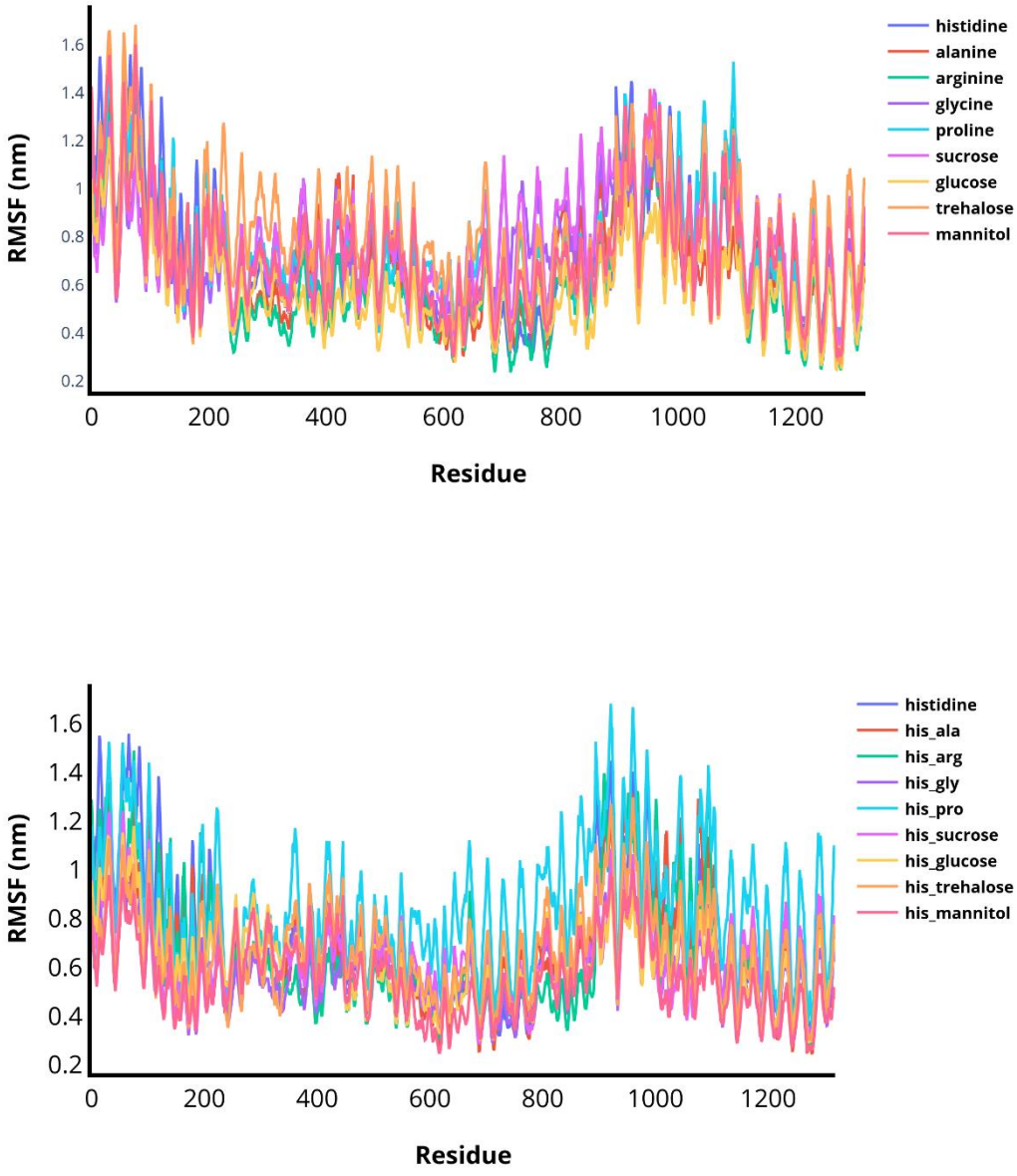


A 14: The average RDF plots of the whole mAb of 3 replicas in 5% v/v solvents. The top graph is in single solvents - histidine, alanine, arginine, glycine, proline, sucrose, glucose, trehalose, and mannitol. The bottom graph is in a mixture of two solvents, 2.5% v/v histidine as the buffer with 2,5% v/v of the solvents mentioned above, pure histidine line was kept as a comparison.

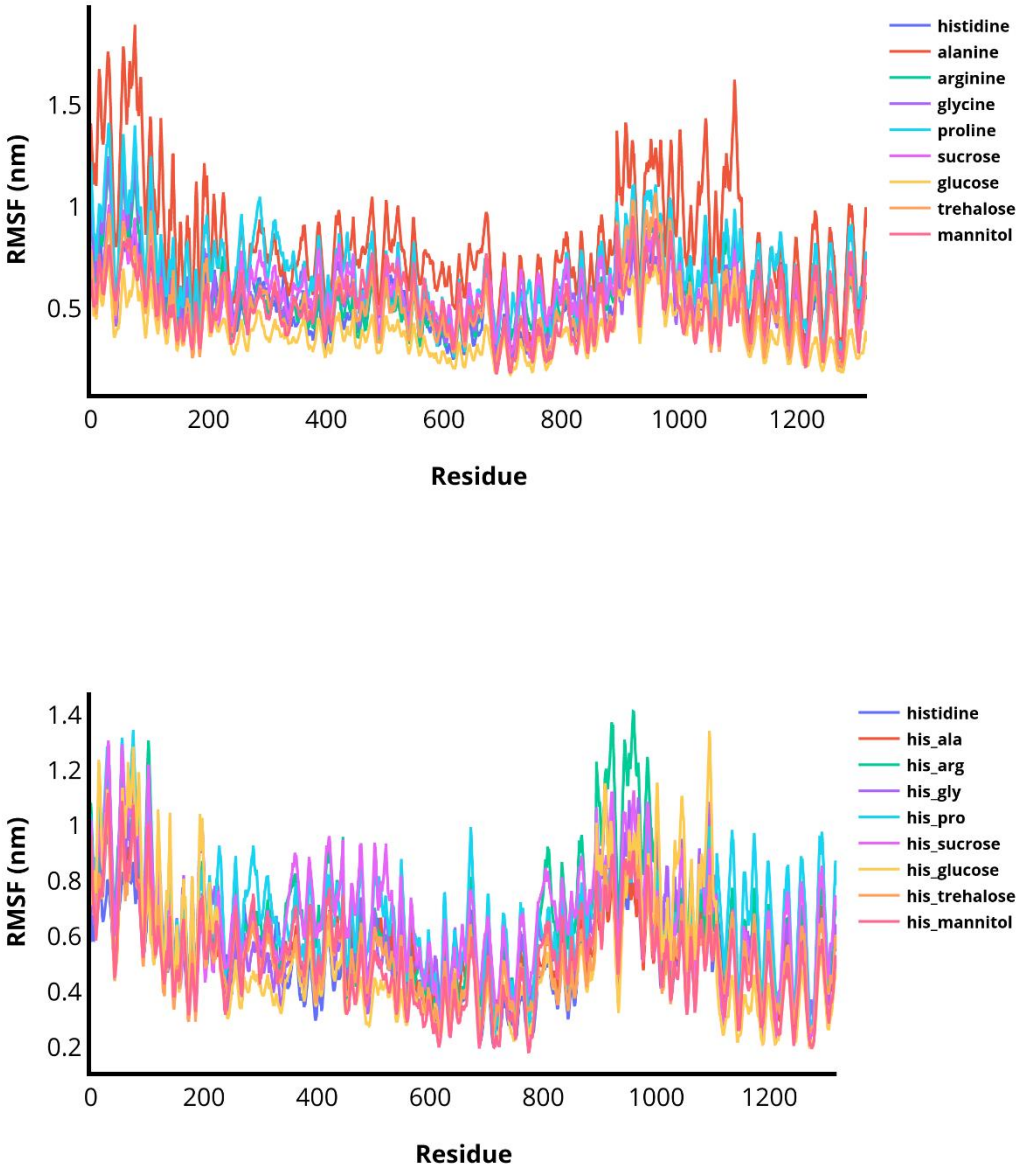


A 15: The average RDF plots of the whole mAb of 3 replicas in different solvents. The top graph is in single solvents – 25 mM histidine, 200 mM alanine, 171 mM arginine, 200 mM glycine, 200 mM proline, 300 mM sucrose, 300 mM glucose, 300 mM trehalose, and 300 mM mannitol. The bottom graph is in a mixture of two solvents, 25mM histidine as the buffer with solvents mentioned above, pure histidine line was kept as a comparison.

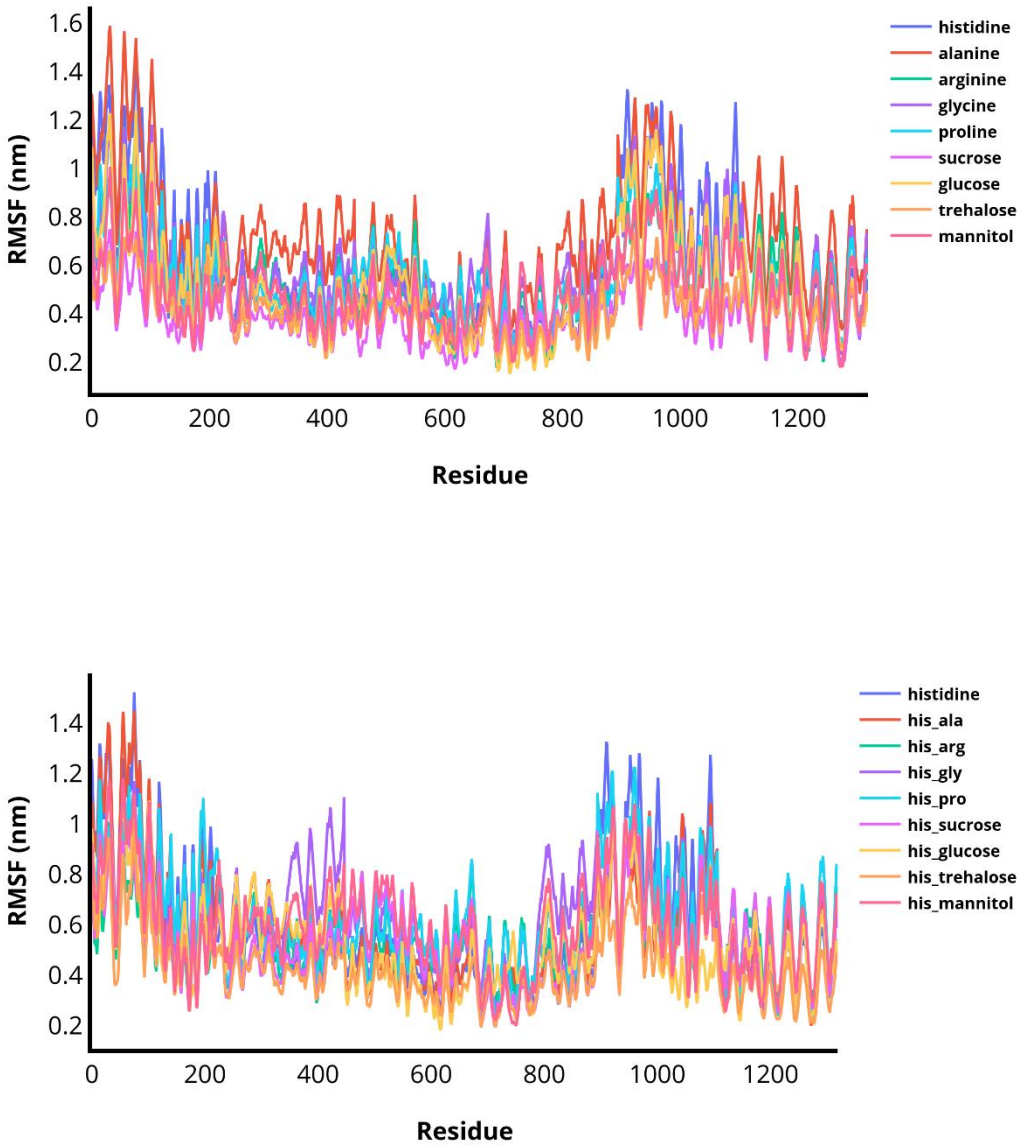
mAb - RMSF



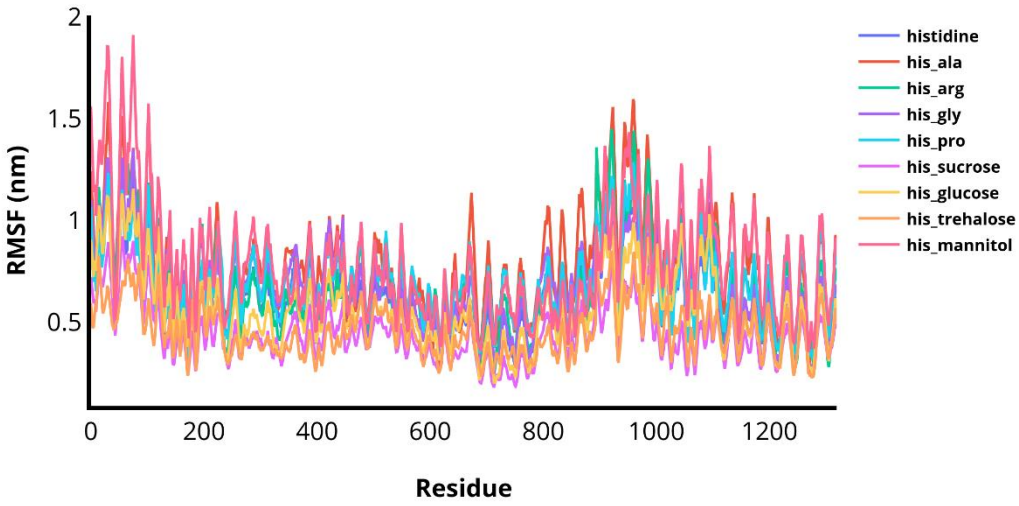
A 16: Average RMSF graphs showing the fluctuations of each residue of the of the whole mAb in different 0.5% v/v single and cosolvents. The top one is in single solvents and the bottom one is in cosolvents.



A 17: Average RMSF graphs showing the fluctuations of each residue of the of the whole mAb in different 2.5% v/v single and cosolvents. The top one is in single solvents and the bottom one is in cosolvents.



A 18: Average RMSF graphs showing the fluctuations of each residue of the of the whole mAb in different 5% v/v single and cosolvents. The top one is in single solvents and the bottom one is in cosolvents.



A 19: Average RMSF graphs showing the fluctuations of each residue of the of the whole mAb in different cosolvents at paper concentration.