Master's thesis
Master's Programme in Materials Research
Biophysics

Investigation of Nanodiscs as a Membrane Protein Environment

Veera Hägg

March 27, 2023

Supervisor(s): Prof. Ilpo Vattulainen, Dr. Shreyas Kaptan

Examiner(s): Prof. Ilpo Vattulainen

University of Helsinki
Faculty of Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Materials Research Biophysics

Tekijä — Författare — Author

Veera Hägg

Työn nimi — Arbetets titel — Title

Investigation of Nanodiscs as a Membrane Protein Environment

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	March 27, 2023	96

Tiivistelmä — Referat — Abstract

Nanodiscs are a synthetic model system for studying the behavior of cell membranes. They are used in experimental biological research to understand structural and functional properties of membrane proteins. Their utility is chiefly due to their water solubility and a relative native lipid environment for membrane proteins compared to other synthetic membrane systems. Though membrane proteins are frequently solubilized and stabilized in a nanodisc environment, the physical conditions that they are exposed to in a nanodisc have not been studied in detail. Additionally, the dynamic behavior of transmembrane proteins in a nanodisc environment has not been characterized with respect to a more typical planar bilayer environment.

The results presented in this thesis formulate an answer to these open questions through atomistic molecular dynamics simulations and machine learning methods. Nanodiscs and bilayer systems with identical lipid compositions are systematically studied, and separately, both types of systems with adenosine receptor $A_{2\alpha}R$ to understand the differences between the model systems. The membrane environment in the two systems is characterized by two well understood physical properties: the order parameter, and the diffusion of lipids in the membrane. The results not only affirm previous studies of nanodiscs but also provide novel insights into the membrane environment of the nanodisc systems. Finally, with the help of machine learning methods, the dynamical behaviour of the protein is shown to be significantly altered in the nanodisc system when compared to a planar bilayer environment. Specifically, it is shown that the activation behavior of $A_{2\alpha}R$ is dependent on model system used to reconstitute the protein.

Avainsanat — Nyckelord — Keywords

Nanodiscs, Lipids, G Protein-Coupled Receptors, MD

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1 Introduction **2**

2 Biological Background **5**

2.1 Lipids and Membranes 5

2.2 Proteins and Membrane Proteins 6

2.3 Nanodiscs 9

2.3.1 Structure 9

2.3.2 Scaffolding Protein 11

2.3.3 Lipid Properties 13

2.4 G Protein-Coupled Receptors 15

2.4.1 Structure 15

2.4.2 Activation and Signalling 17

2.4.3 Adenosine Receptor $A_{2\alpha}R$ 18

3 Methodology **21**

3.1 Molecular Dynamics Simulations 21

3.1.1 Initial Structure 21

3.1.2 Energy Minimization 22

3.1.3 Force Field 22

3.1.4 Newtonian Dynamics 24

3.1.5 Numerical Integration 25

3.1.6 Periodic Boundary Conditions 28

3.1.7 Statistical Ensembles 28

3.1.8 Thermostats and Barostats 29

3.1.9 Fine-Tuning Interactions 30

3.1.10 Constraints and Restraints 31

3.2 Biophysical Properties 31

3.2.1 Acyl Chain Order Parameter 32

3.2.2 Lateral Diffusion 32

3.3 Machine Learning Methods 34

3.3.1 Principal Component Analysis 34
3.3.2 Gaussian Mixture Models 36

4 Simulations **39**
4.1 Systems 39
4.2 Simulation Protocol 40

5 Results and Discussion **43**
5.1 Order Parameter 43
5.2 Diffusion 50
5.3 Protein Behaviour 60

6 Conclusions **76**

Appendix A Labels and Amino Acid Sequences of Membrane Scaffolding Proteins[1] **79**

Appendix B Diffusion Coefficient as a Function of Lagtime **80**

Bibliography **85**

1. Introduction

Life is organized in separate compartments surrounded by membranes called cells. Cells constitute the basic unit of structure and function that all life forms are built upon, whether mono- or multicellular. The primary cell structure contains a lipid membrane that separates the inner part of the cell from its outside environment. Inside the cell, genetic information is stored in chromosomes as deoxyribonucleic acid (DNA) chains containing the genetic information of the life form [2]. In addition, other biomolecules such as proteins, ribonucleic acid (RNA), and metabolites reside within the cell's cytoplasm.

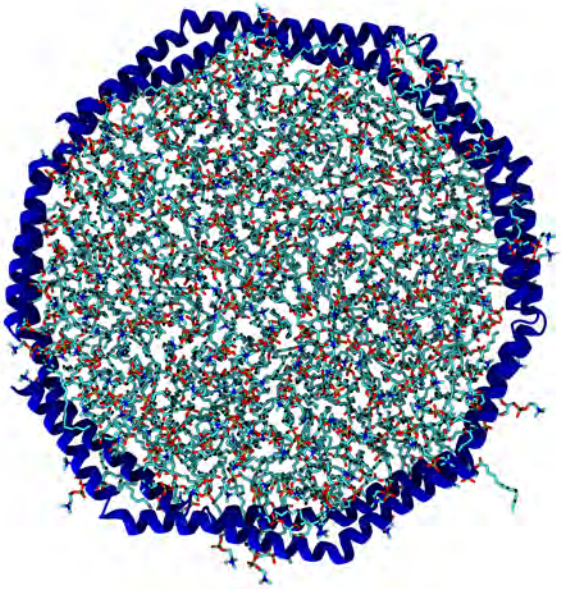
Cell membrane lipids are a diverse group of molecules with a wide range of reported physical and chemical properties. The number of typical lipid types found in cells is of the order of one thousand, and the lipid content of a cell varies significantly between tissues [3]. As of today, The Lipid Maps Structure Database (LMSD) (http://lipidmaps.org/data/structure/) reports over 47 000 known lipid structures in the database, of which a little over half are curated, and the other half predicted using computational methods. However, not all of them are central or abundant in the context of biological membranes [4]. Proteins, on the other hand, are a group of biomolecules that serve in multiple crucial functions within the membrane and inside a cell. For example, DNA replication, transportation of molecules, and catalysing metabolic reactions are all tended to by proteins [2]. The variety of proteins is even greater than lipids, as the Protein Data Bank (PDB) archive (https://www.rcsb.org/) [5] reports over 200 000 available protein structures and over a million computer-generated structure models.

As the numbers indicate, a plenty of work remains in the realm of structural research of proteins. An especially tough challenge has been posed by membrane proteins that are embedded into the plasma membrane with additional domains on the extra- and intracellular spaces as well. The hydrophobic surfaces of transmembrane proteins, in addition to their flexibility and relatively high instability, produce challenges at all levels of experimental work [6].

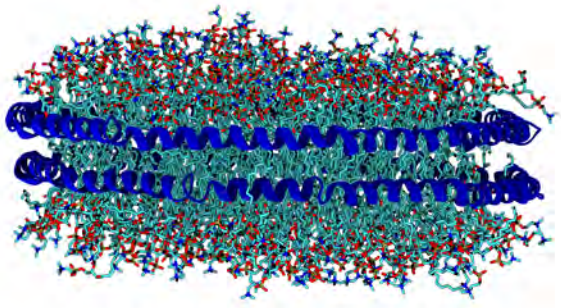
In 2002 Sligar and co-workers produced the first nanodiscs [7] (see structure in Figure 1.1), which since then have helped to tackle these issues. Nanodiscs are used in experimental biology research as a synthetic model of a membrane to assist in structural and functional studies of membrane proteins. Nanodiscs have gained a standing among

synthetic membrane models with the help of their water solubility and good representativeness of a native-like environment for membrane proteins. Membrane proteins can be solubilised and stabilized structurally in a nanodisc environment [8]. Still, the physical conditions that the membrane proteins are exposed to in a nanodisc have yet to be studied in detail, even though a native-like environment has a crucial role in the behaviour of a protein [9]. Especially if from a study conducted in a synthetic membrane, conclusions are hoped to be extended to understand the behaviour of proteins in their native environment. The quest to determine the similarities and differences in lipid environment that a membrane protein is exposed to in a nanodisc versus planar bilayer has become pressing due to recent findings, which indicate that there are differences in the behaviour of lipids within a nanodisc membrane based on their location [8] [9] [10].

Through lipid-protein interactions, the behaviour of the lipids is also tightly linked to the behavior of the membrane protein. Knowing this, an intriguing question can be raised: can the nanodisc in fact, though widely utilized, be depended on as a synthetic environment to reliably reproduce native-like protein behaviour? Since this question is hard to answer experimentally, the study detailed in this thesis provides an answer through computational means. To achieve this, molecular dynamics simulations were conducted to first study nanodisc and bilayer systems with similar lipid conditions and a well-behaved membrane protein to provide a comprehensive review of the lipid conditions within each system that proteins are exposed to. The lipid conditions were characterized by two properties: the order, and diffusion of lipids in the membrane. Secondly, with the help of machine learning methods, an answer has been formulated to the question posed above: Is the behaviour of a membrane protein altered depending on whether it is embedded in a nanodisc or a planar bilayer?



(a) Nanodisc, top view.



(b) Nanodisc, side view.

Figure 1.1: Structure of a nanodisc.

2. Biological Background

2.1 Lipids and Membranes

Lipids (Figure 2.1) are fundamental building blocks of all living organisms. Their nature is amphiphatic, meaning they have two distinguishable parts: hydrophilic and hydrophobic. The amphiphathic nature of lipids and the hydrophobic interactions that arise from it in contact with water allows lipids to aggregate and form structures through self-assembly [2]. This results in structures where several lipid molecules attach to each other by binding their hydro- and lipophilic parts together, such as bilayers, lipid droplets, vesicles, and liposomes. In bilayers, the outward facing, hydrophilic headgroups of lipids come into contact with the solvent on two sides, and in the middle, the two hydrophobic chains turn inward to face each other and shield themselves from the water. These lipid structures are crucial to living organisms since they are the foundation of the plasma membrane surrounding every living cell and the cell organelles.

The lipids of a cell membrane can be roughly grouped into three main types: phospholipids, glycolipids, and sterols, of which, in eukaryotic and prokaryotic cells, the majority are phospholipids. Phospholipids are composed of two fatty acids that attach from a glycerol molecule to phosphate and, finally a hydrophilic headgroup. The two fatty acid chains are monocarboxylic acids with usually an even number of carbon atoms and, in addition, are chain-like and non-cyclic. The order of a membrane is greatly influenced by the length and saturation of the lipid chains, and both saturated and unsaturated fatty acids are found in phospholipids. The unsaturated bonds produce kinks in the lipid chains, preventing them from packing as tightly as the straight saturated chains.

Sterols, such as cholesterol, are lipid molecules with a steroid ring. They play an essential role in mediating the fluidity of the membrane [2], as well as increasing their permeability and increasing mechanical strength [4]. Glycolipids form a tiny fraction of the total number of lipids in a cell membrane but also serve a crucial purpose. They covalently bond carbohydrates that serve a purpose in maintaining the stability of the membrane and cellular recognition [2].

The primary purpose of lipid bilayers is to serve as a biological barrier, separating the cell's insides from its environment. The cell membrane controls substance change

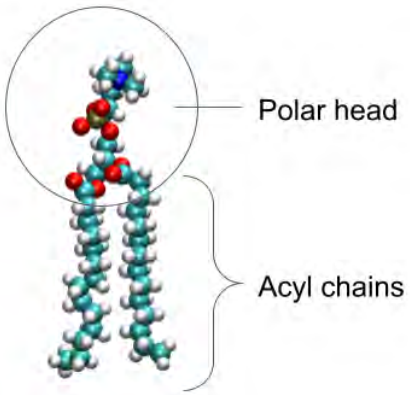


Figure 2.1: The amphiphatic nature of a phospholipid with a polar headgroup and two hydrophobic acyl chains.

to and from the cell, making healthy functioning of the membrane crucial for a cell. The membrane also plays a role in other cellular processes, such as cell adhesion and movement, in addition to regulating membrane proteins and cell signalling through lipid-protein interactions [3]. As a consequence of protecting the cytoplasm inside a cell, the cell membrane also creates a barrier that needs to be crossed every time that information about the environment in- or outside of the cell needs to be delivered to the opposite side of the membrane [2]. This is tended to by a specialized group of signalling proteins that reside in the membrane.

2.2 Proteins and Membrane Proteins

Proteins are a family of specialized organic macromolecules that perform some of the most complex molecular functions crucial for life. Proteins consist of amino acids that all share the same structure in their backbone, with which they form peptide bonds that attach two amino acids together linearly. Peptide bonds are covalent bonds that are inherent to proteins and they are formed by linking the CO-NH atoms in the protein backbone together [11]. All amino acids share the same backbone structure, but differ in the structure of their sidechains. Hence, the sidechain also determines the properties of the amino acid residue. Based on the sidechain properties, the amino acids can be classified into three groups: charged, polar, and hydrophobic. The classification leaves out four amino acids: glycine, proline, histidine, and cysteine. They are considered to be special cases since they cannot be strictly assigned to any of the previous categories due

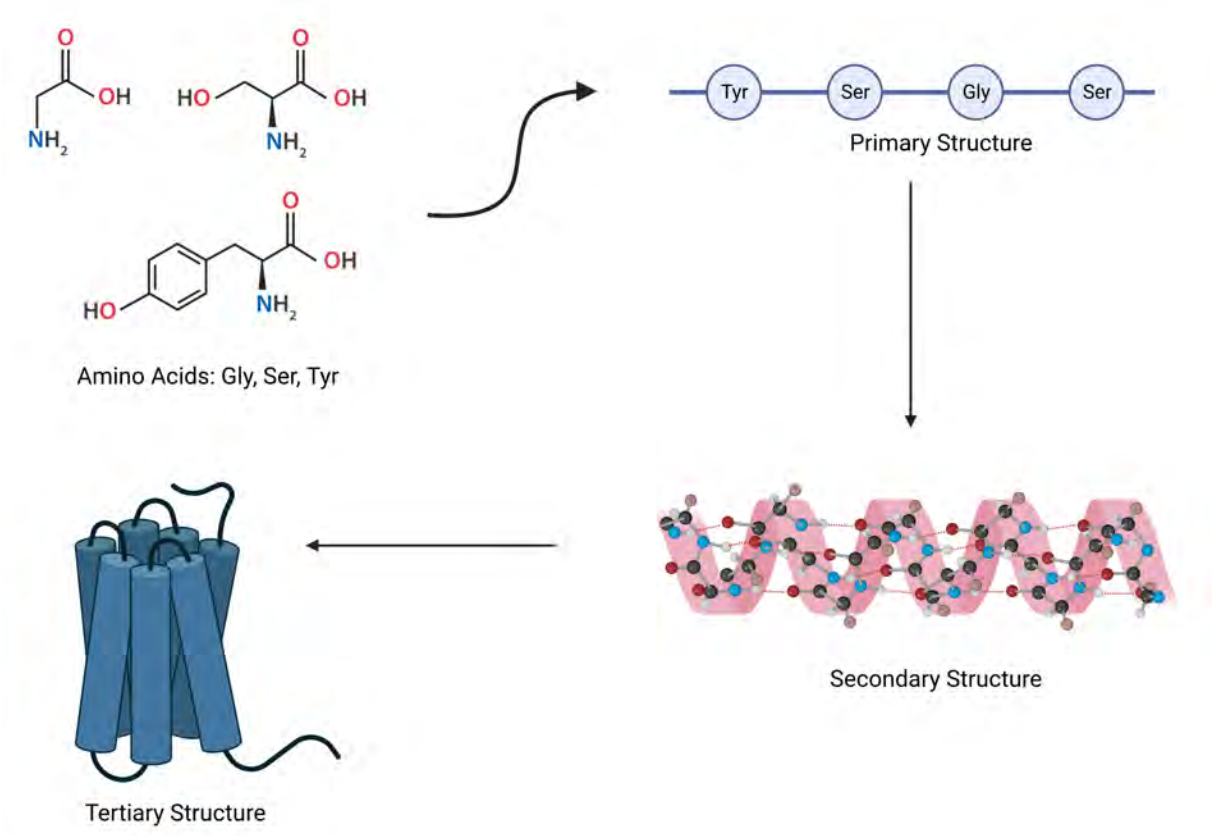


Figure 2.2: The process of protein folding from amino acids through primary and secondary structure into a functional protein. Created with BioRender.com.

to their sidechain structure [12].

The linear, ordered sequence of amino acids is called the primary structure of a protein. However, the primary structure itself is not enough for a protein to be functional. In order to achieve this goal, a protein needs to fold itself, see Figure 2.2. Secondary structure is the first level of protein folding that showcases localized organization and structural motifs. Several distinct secondary folding patterns are recognized in proteins, the main classes being α-helix, β-sheet and coil. Secondary structures are defined according to torsion angles between the backbone atoms of the peptide, where θ is the torsion angle around CA and N atoms and ϕ is the torsion angle around CA and C atoms. Of the classes of secondary structures, α-helices are the most conserved and stable, which is a result of the tight packing of amino acids within the helical structure that arise from hydrogen bonds between the carbonyl oxygen and amide hydrogen between subsequent residues. β-sheets are less stable compared to α-helices due to their lack of local interactions. But stacked together, hydrogen bonds are formed between the individual sheets, which stabilizes the secondary structure of β-sheets [11].

Tertiary structure can also be referred to as a super secondary folding. It arises from the interactions, such as hydrophobic interactions and disulfide bridges, between the

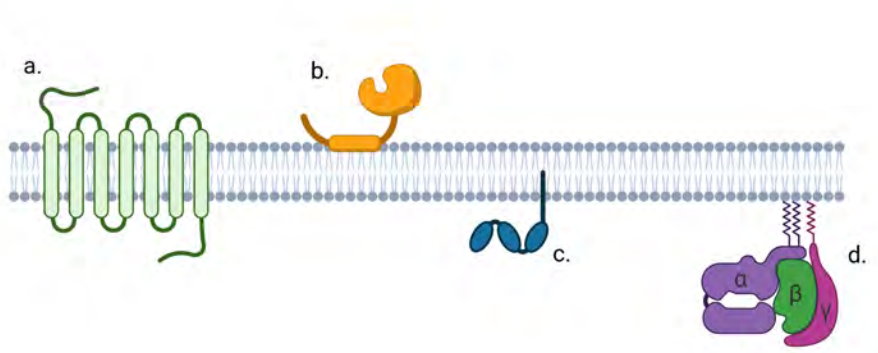


Figure 2.3: Membrane protein types: a. Integral membrane protein, b. Peripheral membrane protein interacting with membrane through an amphiphatic helix, c. Peripheral membrane protein anchored into the membrane through a hydrophobic chain, d. Lipid anchored protein: G protein with subunits α, β, and γ. Created with BioRender.com.

secondary motifs with each other and are also modified by the environmental effect. In other words, the protein folds and bends itself in a way that its interactions with the environment become as favourable as possible. The tertiary structure can also give rise to additional properties, both functional and non-functional [11]. Proteins can be fully functional after tertiary folding, but quaternary structures are also possible. Quaternary structure is the final functional form of a protein. Instead of folding of a single peptide chain, quaternary folding considers multiple folded chains merging together to form larger functional complexes.

The plasma membrane hosts membrane proteins that are responsible for many functions of the membrane, such as signalling and substance transportation across the membrane barrier. Hence many membrane proteins react to stimuli, such as the presence of chemical compounds or physical conditions like stress. The stimuli cause changes in the conformation of the protein receptor which in turn act as starting points for signalling cascades to make the protein perform specific functions or to stimulate molecules to do so [2]. Essential membrane proteins include, for example, integral membrane proteins that

pass all the way through the membrane one or multiple times, and peripheral membrane proteins that are connected to the membrane on one of its sides, see figure 2.3. Since the acyl chains of the membrane lipids are hydrophobic, it is also pivotal for the part of the integral membrane protein that stays inside the membrane to have neutral sidechains as well. Often these membrane embedded parts of the integral membrane proteins acquire an α-helical conformation. For peripheral proteins, it is enough to have a hydrophobic chain that will anchor it into the membrane [2]. A special group of membrane proteins are the, so called, lipid anchored proteins, of which G proteins serve as an example. Instead of being in direct contact with the membrane, the protein attaches to a single or multiple lipids through covalent bonds that anchor the protein onto the membrane. In addition to attaching the protein onto the membrane, the lipid also plays a role in mediating protein-protein interactions between the anchored protein and possible integral membrane proteins in its vicinity [13].

2.3 Nanodiscs

Nanodiscs are synthetic models of membranes used especially in research of membrane proteins. Even though they are not a native biological structure, nanodiscs have claimed a standing as one of the most relevant synthetic environments for membrane protein research, because compared to other synthetic environments, like micelles and bicelles, they represent a more native-like environment for proteins [14]. A native-like environment allows for the proteins under study to solubilise and stabilize as in a native environment, which consequently can yield realistic results of the membrane protein behaviour [8]. This, in addition to the fact that nanodiscs themselves are soluble in aqueous solutions making them easier to handle, makes nanodiscs extremely relevant in the area of membrane protein research [14]. Combined with cryo-electron microscopy, nanodiscs have recently played a part in unravelling novel structures for numerous membrane proteins [15] and notably, were used in the reconstruction of the active state of the full-length human insulin receptor [16].

2.3.1 Structure

Nanodiscs are macromolecular membrane assemblies, where the lipids are confined together with two long proteins that wrap around the membrane, referred to as the membrane scaffolding proteins (MSPs). As with other colloidal systems, nanodiscs are emergent and self-assemble when solvent is removed from the mix of the constituents [17].

Nanodiscs are composed of lipids of natural or synthetic origin that are surrounded and bound together by a scaffolding protein. In the case of protein research, a membrane

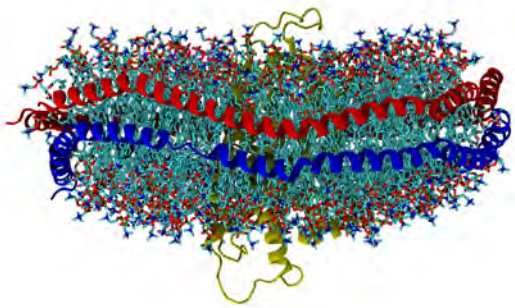


Figure 2.4: Nanodisc with an embedded transmembrane protein.

protein is also often embedded into the bilayer in the middle of the disc. Nanodiscs exhibit a multitude of different characteristics that make them desirable candidates for experimental biology research. Mainly, they are stable and structurally well defined. The stability of the nanodiscs comes down to the interactions between the lipid and the scaffolding protein that surrounds them: firstly, lipids have low solubility in water which drives them to form a bilayer in polar solvent and to minimize the unfavourable interactions between the two [14]. However, even in a bilayer formation some of the lipids are exposed to the solvent unless the bilayer transforms into a spherical aggregate, such as a vesicle.

Secondly, the scaffolding protein also plays a major role in the stability of nanodiscs. The scaffolding protein that surrounds the solvent exposed lipid acyl chains, is an amphiphatic helix who's already stable conformation will only be enhanced by the interactions between the lipids and itself. The MSP stabilizes the nanodisc structure by turning its neutral side chains towards the lipids and polar or charged side chains towards solvent, thus shielding the lipids chains completely from solvent and eradicating the remaining unfavourable interactions between the two. In this manner, the lipid-protein interactions between the lipids and the lipid-bound scaffolding protein stabilize the nanodisc structure. The amphiphatic nature of the scaffolding protein and its shielding effect to the

lipid-solvent interactions also gives rise to another important feature of nanodiscs that is their solubility in polar solvents, such as water. To conclude, the tendency of lipids to form membranes and the enhanced lipid-protein interactions of the MSP helix combined, give rise to the self-assembly of nanodiscs and their stability in dynamic equilibrium with the environment [14].

2.3.2 Scaffolding Protein

The first nanodiscs were produced using a membrane scaffolding protein derived from the human apolipoprotein Apo-A1, which is the main protein component of high density lipoprotein (HDL) [17]. HDLs partake in metabolism and trafficking of cholesterol in the human body and also play a role in the steroid hormone biosynthesis as a supplier, for which it can be merited as one of the key players in human health. In solvent, discoidal HDL is stabilized by two belts of Apo-A1 wrapping around it [1]. HDL was intensively studied during the 1950's and 60's and it was then noted for the first time that the helical configuration of the protein plays an important role in the protein-lipid interactions of HDL. The flexibility of lipoproteins was also noticed, in addition to the high dependency of the lipoproteins' properties from the lipid mixture in the HDL lipid-protein complex [18]. So, apolipoprotein Apo-A1 became the golden standard for nanodisc scaffolding proteins: the MSPs used in nanodiscs today are still either the original Apo-A1 or one of its derivatives [17].

Several different scaffolding proteins have been engineered after the original Apo-A1 that can be used to produce stable nanodiscs with different numbers of lipids and varying circumferences. The original human Apo-A1 sequence as a scaffolding protein contains 10 helices and an additional N-terminal helix (hexahistidine tag + Factor X recognition site) and is referred to as MSP1 [19]. All its derivatives are composed of the same helices with some of the ten helices either repeated or deleted. The N-terminal helix also has a longer modified version (hexahistidine tag and a linker containing a TEV protease site enabling removal of the tag) and the first helix (H1) has a few truncated versions of it, see Appendix A for more information. The rational engineering of the MSP's was done after it was uncovered that not all of the N-terminal amino acids were taking part in the formation of the belt around the nanodisc membrane that stabilizes the structure of the nanodisc [20]. Since all the residues in the engineered MSPs take part in lipid binding, they also allow for more precise control over the final size of the nanodisc and, consequently, to more easily accommodate membrane proteins of various shapes and sizes into the nanodiscs [21]. Generally, the scaffolding proteins with name starting with MSP1 consist of one repeat of the eleven helices with some possible repeated helices in the middle and those with name MSP2 are double-sized; meaning that all the helices are repeated in

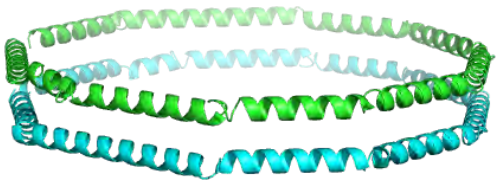


Figure 2.5: The double belt of membrane scaffolding protein MSP1E1.

the same order twice [19].

The scaffolding protein governs certain features, such as shape and size, of the nanodisc and they can be adjusted by truncating the MSP or fusing multiple of them together [1]. For a time, it was unclear how the scaffolding protein sits around the lipids. However, the finding of a correlation between the diameter of the nanodisc and the length of the MSP provided the evidence that the scaffolding protein in practice forms belts [22] around the bilayer of the nanodisc rather than a picket fence formation [23] that had been suggested previously [24]. This was later also confirmed by the means of solid state NMR [25], electron microscopy and other techniques [26] including simulations [23] [27], leading to general acceptance of the belt configuration. But most importantly, the scaffolding protein plays a pivotal role in the stability of the nanodisc as already discussed above. Considering the stability, especially the ratio between the number of lipids in the bilayer and the length of the scaffolding protein is crucial. The optimal length of MSP with respect to the number of lipids in a bilayer of the nanodisc are related in the following manner:

$$M = \frac{2(\pi r + \sqrt{\pi N S})}{L}, \tag{2.1}$$

where M is the length of the MSP (number of amino acid residues), r is mean radius of the MSP and L is the helical pitch per MSP residue. Finally, N is the number of lipids in bilayer and S the mean surface area of the lipid type [20].

The longest scaffolding proteins can be used to construct nanodiscs with a diameter up to ∼17 nm. Even though larger membrane sizes would also be possible, due to the ratio of the MSP length and number of lipids, larger nanodiscs are often unstable and collapse to spherical aggregates [14]. Multiple ideas have been proposed to maximize the stability

of nanodiscs: the modification of MSPs [28] as stated also above, and the optimization of lipid composition, since in nanodiscs it can be precisely controlled [29].

2.3.3 Lipid Properties

Even though nanodiscs are often used in structural and functional studies of membrane proteins [14] due to their stability and water solubility, growing evidence suggests that the structure and behaviour of the lipids inside nanodiscs are not strictly comparable to those in a normal planar bilayer [30]. If membrane protein function and structure are what we want to study with nanodiscs, this raises an obvious question about the conditions that a protein of interest faces when embedded in the nanodisc: Are the functional circumstances native-like enough to stabilize the protein structure and foster its usual activity [8]? The relevance of this question is in addition backed up by studies that have shown the sensitivity of mechanosensitive membrane protein channels to membrane pressure [31] and how lipid conformations have been observed to affect the structures of membrane proteins that are sensitive to allosteric binding of the lipids [32]. In conclusion, understanding the internal structure and lipid-protein dynamics in nanodiscs is extremely important.

The conformation of lipids' chains and headgroups and their dynamics have been shown to be modified under the influence of the MSPs that surround the nanodisc structure [10]. In parallel with these findings, it has also been reported that the lipids in nanodiscs indeed can be classified into three distinct categories that revel an internal structure within nanodiscs: central lipids (1) in the center of the disc, boundary lipids (2) that are in direct contact with the MSPs and intermediate lipids (3) in between [8]. The three groups of lipids have been shown to have drastically different properties:

1. **The central lipids** in the middle part of the nanodisc are characterized by significant ordering and consequently, slow rate of diffusion. Notably, the order in the central lipids is higher than that of normal planar bilayer systems and the features of the central lipids resemble the characteristics of a cholesterol-rich membrane [8]. Because of the tight packing of lipids, the area per lipid in the central area is smaller than the average area per lipid in the boundary regions. Due to the high order and subsequent tight packing, the central lipids are also more hydrophobic than the boundary lipids [33] and form the thickest part of the nanodisc [8].

2. **The boundary lipids** that are in direct contact with the scaffolding protein are, as opposed to the central lipids, characterized by disorientation and fast diffusion [8]. These lipids at the boundary of the nanodisc are prone to remain in fluid state even at temperatures below their main transition temperature [34]. Evidence also suggests that at the boundary area near the MSPs, the nanodisc bilayer is the thinnest and due to the

less tight packing of the lipids, they are more hydrated than in the middle of the nanodisc [8]. Around 60% of lipids in nanodiscs have been reported to be under the influence of the MSPs [35] and about 30% of all lipids have been approximated to be in direct contact with the scaffolding protein [8].

3. **The intermediate lipids** are the lipids that reside in between the two other groups. In the intermediate area, the lipids can translocate between the populations of central and boundary areas. The enthalpy of this transition has been observed to resemble the lipids going through a main phase transition. In addition, the number of fast lipids increases with increasing temperature which in practice means that the central lipids melt and transition into boundary lipids [8]. Due to this phenomenon and the fact that boundary lipids have a larger area per lipid than central lipids, the radii of nanodiscs increase with increasing temperature allowing the number of tightly packed central lipids to decrease [35].

Overall, on average the lipids of nanodiscs exhibit higher order than usual lipid bilayers, which has been explained by the influence of the scaffolding protein both experimentally [10] and in simulations [9]. Consequently, this is also expected to indicate lower entropy levels in nanodisc lipids that has been studied through simulations [9]. Also, the effect of the scaffolding protein divides the lipids into distinct groups of central and boundary lipids [8] that in addition to difference in lipid order, also differ in the hydration due to backfolding of lipids at the boundary regions. This allows water to penetrate at the edges of the nanodisc, confirmed both by simulations [9] and fluorescence studies [30]. The main conclusions of experimental data concerning the internal structure of nanodiscs discussed above have been confirmed by molecular dynamics simulations, with both all atom and coarse grained models [9]. However, all modelling studies of nanodiscs report perturbations in the circular structure of the nanodiscs that are described, for example, as elongated ellipsoids and distorted polygons [14]. These findings have been also employed to explain recent results from scattering studies, and these slightly distorted models of nanodiscs have been found to fit better than the ideal circular shape [36] [37] [38]. Hence, it has been concluded that it is probable for real nanodiscs to exhibit slightly irregular shapes rather than ideal circles. The sharp angles in the scaffolding protein have been so far explained by kinks at the proline residues that separate the helices of the scaffolding proteins from each other [33].

In the light of the interesting lipid characteristics in nanodiscs and knowing that lipid-protein interactions are crucial for the behaviour or membrane proteins, a comparison of biophysical properties of lipids in planar bilayers and nanodiscs is of essence. Since this could answer important questions about the effects of these distinct membrane environments on lipid behavior and protein-lipid interactions. Planar bilayers would serve as the conventional model for membrane systems and provide a comprehensive under-

standing of lipid organization and dynamics in a native state. Nanodiscs, on the other hand, would offer a unique, synthetic environment to study that mimics the membrane topology found in cells. By comparing the biophysical properties of lipids in these two systems, valuable insights could be gained into the role of confinement on lipid dynamics, organization, and finally, their impact on membrane protein function.

2.4 G Protein-Coupled Receptors

G Protein-Coupled Receptors (GPCRs) are a diverse group of membrane receptors. They are the largest superfamily of membrane receptors in eukaryotes with almost 1000 unique members only in humans [39]. GPCRs are signalling proteins that respond to a wide array of stimuli, through which the receptors modulate and control processes in the cell. For example, GPCRs are known to take part in senses of smell, vision, and taste, as well as regulation of behaviour, mood, immune system, and nervous system among other vital physiological events [40]. As of 2018, it has been estimated that about 35% of the FDA-approved drugs target 108 members of the GPCR family [41]. Over 35 atomic level structures of GPCRs are nowadays available [42] and in 2012 the Nobel Prize in Chemistry was awarded to Brian Kobilka and Robert Lefkowitz on the subject. The modulation of GPCR signalling still holds its place as a hot topic in pharmaceutical research.

2.4.1 Structure

According to a recent new classification system, the human GPCRs are classified into five main classes through phylogenetic analysis that are rhodopsin, secretin, glutamate, adhesion and frizzled/taste2 receptors [40]. Though each receptor is unique and highly specific to a particular signal, all the GPCRs share the same major structure: sometimes also referred to as the seven-pass-transmembrane domain (7TM) receptors, the GPCRs pass through the cell's plasma membrane seven times. The transmembrane domains are seven helices (H1-7) that connect together through six loops in- and outside of the cell in addition to the N- (outside the cell) and C-terminal (in cytoplasm) areas [43]. Some sources also differentiate a small cytoplasmic extra helix, called helix eight (H8) that follows straight after H7 and lies parallel to the membrane plane [40].

Although the transmembrane domains of GPCRs are similar to each other, the sequences of the receptors are diverse [44]. Still few features are considered to be crucial for GPCR activation and inactivation, and some corresponding SM/FM (conserved Structural Motifs that have roles as Functional Microdomains) elements have been identified in the known GPCR sequences [45]. These elements include groups of conserved residues that have been identified to be important as 'switches' [46] for the GPCR activation over

Figure 2.6: G protein (light purple) with GDP bound to adenosine receptor $A_{2\alpha}R$ (pink).

the family. The four most important motifs include:
1. The so-called 'ionic lock' that includes highly conserved residues in H3 and H6 [47]. These residues form a network of hydrogen bonds that bridges together the two transmembrane helices and are hypothesized to stabilize the inactive state of the receptor [48].
2. The hydrophobic cage, where a group of amino acids constrain a conserved arginine residue in H3 [49].
3. NPxxYxF motif in H7 or 'Tyrosine Toggle Switch' [50] that allows for a direct interaction of H7 and H8 [49], and a water-mediated interaction between H7 and H2 [51].
4. The Rotamer Toggle Switch which is an interaction of aromatic amino acids surrounding the tryptophan in the CWxP motif in H6 [52] that leads to a vertical rigid motion of the helix. Binding of a ligand triggers a chain of rearrangements in the extracellular part of the receptor that regulates the ionic lock described above. In the inactive state, this cluster of aromatic residues points towards H7 and in the active state undergoes a conformational change to pointing towards H5 [53].
Compared to the transmembrane domain structures, the termini and the extracellular loops of the GPCRs are extremely diverse both in lengths and sequences [44]. There is also evidence that they can strongly influence ligand binding and dynamics of the receptors [54].

2.4.2 Activation and Signalling

As implied in the name, GPCRs interact with so-called G proteins in the plasma membrane. There are three main signal transduction pathways that the GPCRs are known to be involved in: the phosphatidylinositol signal pathway and inhibition and stimulation of cAMP production [43]. In principle, a GPCR undergoes a conformational change when an extracellular signalling molecule binds to the receptor or through triggering of another signalling mediator, like light. On the cytoplasmic side of the bilayer, this conformational change mediated through the seven transmembrane helices allows for the GPCR to engage in an interaction with a nearby G protein.

G proteins exist as heterotrimers with three distinct subunits: α, β, and γ, with a nucleotide binding pocket in the α-subunit. To keep its inactive conformation, the $G_{\alpha\beta\gamma}$-trimer attaches itself to guanosine diphosphate (GDP), where the $\beta\gamma$-dimer prohibits the dissociation of the nuclide, hence stabilizing the conformation of the complex [55], see Figure 2.6. As the GPCR changes its conformation to active through a binding of an agonist, the receptor associates with the inactive $G_{\alpha\beta\gamma}$-trimer which may lead to the conformational change to activate the α-subunit of the G protein. The activated α-subunit exchanges GDP into guanosine triphosphate (GTP) rapidly, which in consequence initiates the dissociation of the α-unit from the $\beta\gamma$-dimer complex [55]. In this nucleotide free from, the GPCR and G protein complex is highly stable and has a high affinity for the agonist and a higher affinity for GTP than GDP [56]. This drives the binding of the GTP and the subsequent dissosiation of the GPCR and G protein and the eventual dissociation of the G_{α}-GTP and $G_{\beta\gamma}$ from each other [57]. For activation of the G protein, the GDP release has been determined to be the rate-limiting step [58] and despite available experimental data, the molecular level mechanism of the dissosiation step is still unclear. However, the freed GPCR is then ready to bind another $G_{\alpha\beta\gamma}$-trimer and the G_{α}-GTP and $G_{\beta\gamma}$ are available to continue a signal transduction cascade with other proteins within the cell [59]. The resulting signal is directly dependent on the type of the α-subunit, of which there are four distinct families: $G_{\alpha i}$ (inhibits cAMP pathway), $G_{\alpha s}$ (stimulates cAMP pathway), $G_{\alpha 12/13}$ (remodelling of actin cytoskeletal in cells during movement), and $G_{\alpha q/11}$ (stimulates the phosphatidylinositol signal pathway) [60]. Finally, the G protein cycle is completed with hydrolysis of GTP to GDP within the G_{α}-GTP subunit that allows for the $G_{\beta\gamma}$-dimer and the emerging G_{α}-GDP to rebind [61].

GPCRs have also been recently shown to form homo- and hetero-oligomers, and the binding of the oligomers versus single receptors to G proteins has been observed to trigger different signalling pathways [62]. However, the role of oligomerization in GPCR signalling still remains mostly uncertain. The GPCRs are believed to exist in a dynamic equilibrium between active and inactive states, and the binding of the ligands to shift the

equilibrium according to the type of the ligand. If the ligand is an agonist, the equilibrium shifts in favour of active states and if an inverse agonist, the equilibrium shifts toward inactive states. Ligands that are neutral antagonists, do not affect the equilibrium [63]. Also, it should be mentioned that in addition to the native ligand binding pocket on the extracellular side of the receptor (orthosteric site), where majority of the GPCR ligands bind, so-called allosteric ligands also exist that bind at other sites within the receptors [39]. Often the role of these allosteric modulators is to increase or decrease the binding affinity of the orthosteric ligands binding at the main site, thus influencing the dynamics, function and structure of the GPCRs [64], even to the point where the allosteric ligands may activate the receptor as strongly as the main ligand [61].

2.4.3 Adenosine Receptor $A_{2\alpha}R$

The adenosine receptor $A_{2\alpha}R$ (Figure 2.7) is one of the four adenosine receptor subtypes (A_1R, $A_{2\alpha}R$, $A_{2\beta}R$, A_3R) in the rhodopsin-like family of GPCRs that modulate adenosine [65]. Adenosine is a neuromodulator that plays a role in governing the activity of both neurons and glial cells and is formed by the degradation of adenosine triphosphate (ATP) [66] both in intra- and extracellular space. Adenosine also acts as a homeostatic modulator [67]. $A_{2\alpha}R$ is distributed widely in the limbic system and neocortex, but is most abundant in the ventral and dorsal striatum mainly in the synaptic and extrasynaptic space of GABA neurons [67]. The presense of $A_{2\alpha}R$ in synapses indicates its role in controlling synaptic transmission, though the exact physiological mechanism remains unclear [68]. However, some hypotheses suggest the controlling of synaptic transmission through the regulation of NMDA receptors or presynaptic mechanisms [69]. In neurons, $A_{2\alpha}R$ acts as a modulator of glutamate and dopamine release, making the receptor a potential drug target for, e.g., Parkinson's [70] and Alzheimer's disease [71], and other conditions such as drug addiction, insomnia, and pain [72]. $A_{2\alpha}R$ has also been observed to protect tissue from inflammation and damage by suppressing immune cells through regulation of immunosuppressant levels [73].

What distinguishes the structure of $A_{2\alpha}R$ from other structurally determined GPCRs is an allosteric binding pocket below the orthosteric pocket on the extracellular side [74] of the receptor. The allosteric pocket is also known as the sodium-ion binding pocket after the discovery that sodium can bind there [75]. $A_{2\alpha}R$ couples to $G_{\alpha s}$ that in turn activates adenylyl cyclase [67], finally leading to the stimulation of the intracellular cAMP signal transduction pathway. The activation of all rhodopsin-like GPCRs have been shown to involve large movements in the transmembrane helices H5-7 that lead to an opening of a hydrophobic binding site for the G protein on the intracellular side of the membrane [76]. In the case of $A_{2\alpha}R$, the crystal structures show the transmembrane

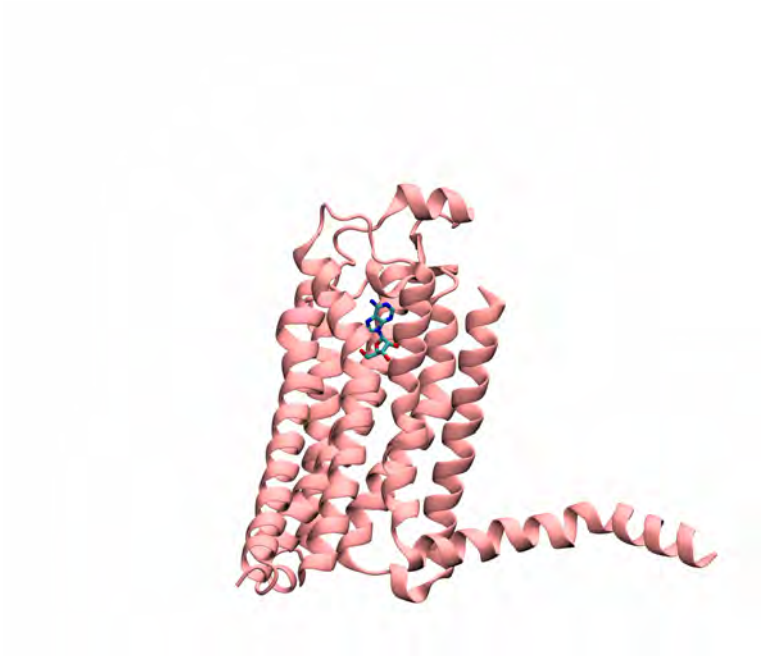


Figure 2.7: Structure of the adenosine receptor $A_{2\alpha}R$ with an adenosine bound to the ligand binding pocket.

events to be similar to other rhodopsin-like GPCR structures [77], but the magnitude of the changes appear to be smaller [78]. In simulations, $A_{2\alpha}R$ has been observed to rely on at least the 'ionic lock' and rotamer toggle switches when rearranging its transmembrane conformation from inactive to active state [79]. As for other GPCRs, the $A_{2\alpha}R$ has also been observed to engage in several hetero-oligomer conformations, for example with dopamine D_2 receptors, in addition to receptor tyrosine kinases, and glucocorticoid receptors [80].

3. Methodology

3.1 Molecular Dynamics Simulations

Molecular dynamics (MD) simulations can be used to study systems of molecular scale that would be unreachable by other means. The simulation systems are composed of particles that represent either atoms or groups of atoms that move according to the laws of classical mechanics. By tracking the movement of the particles and recording the system dynamics, models of the previously inaccessible systems can suddenly be within reach.

At the heart of the classical MD simulations lies the Born-Opperheimer approximation. It assumes for the electrons to stay in the ground state and have negligible effect on the movement of an atom due to the weight difference between electrons and the nucleus. Hence, the motion of atoms can be approximated only from the movements of the nucleus and potential energy functions are used to treat the eletron distributions [81].

3.1.1 Initial Structure

To begin an MD simulation, an initial starting structure is needed. Experimentally resolved structures of proteins are available through databases such as the PDB archive [5], additionally the initial structure can also be obtained, for example, from a previous simulation. The PDB structures often contain only the coordinates of the protein, and requires the user to add other molecules such as lipids or carbohydrates into the simulation system by themselves, as well as water and ions.

It is possible that the available protein structures are also missing some parts of the molecule. Single atoms can be simply added into the protein topology file, but bigger missing regions require help from modelling software such as MODELLER [82]. If the missing region in the molecule is too large to be reliably reconstructed through modelling, it might cause artefacts or the system might become unstable during simulation. Hence, without a good structure, a reliable simulation cannot be performed.

The initial structures yield coordinates for the system, but to start a simulation, velocities are also required. If they are not known from previous simulations, initial ve-

locities v_i are assigned randomly for each atom i from the Maxwell-Boltzmann distribution at a given temperature:

$$p(v_i) = \sqrt{\frac{m_i}{2\pi k_B T}} e^{\left(-\frac{m_i v_i^2}{2k_B T}\right)}, \tag{3.1}$$

where T is temperature, m_i is the mass of particle i, and k_B is the Boltzmann constant.

3.1.2 Energy Minimization

Ofter, the initial starting structure for a simulation does not respond to a physically relevant state or may not lead to a stable simulation. Hence, it is desirable to first optimize the interactions in the system by leading the system into a minimum energy state. Biological systems tend to be large and complex, meaning that they also have many local energy minima, all of which are impossible to sample. Because of that, the search is set out to find the nearest local energy minimum to the initial structure and start simulating from there. The local energy can be found by minimizing the potential energy V of the system as follows:

$$\frac{\partial V}{\partial \boldsymbol{r_i}} = 0 \tag{3.2}$$

$$\frac{\partial^2 V}{\partial \boldsymbol{r_i}^2} > 0, \tag{3.3}$$

where V is potential energy and $\boldsymbol{r_i}$ are the coordinates of particle i in the system.

Because of the complexity of biological systems, using analytical methods to find such minima is practically impossible. This is why numerical methods are used that can recognise and move down the potential energy landscape of the system towards lower energy. Such numerical methods that yield a sufficiently stable state of the system to start a biological simulation from, are for example the conjugate gradient and steepest descent. Also methods that use random sampling such as Monte Carlo methods and simulated annealing, can be used to find a local energy minimum [81].

3.1.3 Force Field

Force field is a construct that provides a model for how the different kinds of atoms in a simulation interact with each other through their potential energy. Consequently, the force field can be seen as the single most crucial component of the simulation protocol which affects the outcome of the simulated system. Five interactions construct the basis of

many force fields, though the potential energy function of some force fields might contain additional terms as well.

In a general from, the potential energy function for a system can be written as:

$$E_{\text{pot}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihedrals}} + E_{\text{L-J}} + E_{\text{Coulomb}}. \tag{3.4}$$

The first three terms of the equation are called the bonded interactions, which characterize the length, stretching, vibrations and shape of the bonds. During the simulation run, these interactions are calculated based on a list set at the beginning. The last two interactions are the so-called non-bonded interactions. They describe the interaction between neighbouring atoms and hence, are sensitive to the immediate neighbourhood of an atom at any moment in time. So, the non-bonded interactions are calculated based on a list of interactions that get updated periodically during the simulation run.

All terms written out, equation 3.4 gets the following form:

$$\begin{aligned} E_{\text{pot}} = &\sum_i {k_{\text{i}}}^{\text{bond}}(r_{\text{i}} - r_0)^2 + \sum_i {k_\theta}^{\text{angle}}(\theta_{\text{i}} - \theta_0)^2 + \sum_i {k_{\text{i}}}^{\text{dihed}}[1 + \cos(n_i\phi_{\text{i}} + \phi_0)] \\ &+ \sum_i \sum_{j \neq i} 4\epsilon_{\text{ij}} \left[\left(\frac{\sigma_{\text{ij}}}{r_{\text{ij}}}\right)^{12} - \left(\frac{\sigma_{\text{ij}}}{r_{\text{ij}}}\right)^{6} \right] + \sum_i \sum_{j \neq i} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{\epsilon_r r_{\text{ij}}}. \end{aligned} \tag{3.5}$$

The first term of the equation is related to the length and stretching of a bond between two atoms of the system with harmonic spring potential.

$$E_{\text{bonds}} = \sum_i {k_{\text{i}}}^{\text{bond}}(r_{\text{i}} - r_0)^2, \tag{3.6}$$

where ${k_i}^{bond}$ is the force constant related to the bond, r_i is the length of the bond and r_0 is the reference bond length.

The second term is related to bond angles and their vibrations. This is again described with harmonic spring potential like bond stretching previously.

$$E_{\text{angles}} = \sum_i {k_\theta}^{\text{angle}}(\theta_{\text{i}} - \theta_0)^2, \tag{3.7}$$

where ${k_\theta}^{angle}$ is the force constant, θ_i is bond angle and θ_0 is the reference angle.

The third term characterizes the torsion angles of the bonds. These dihedral angles are formed by four atoms and their values characterize the secondary structure of molecule as described in the previous chapters.

$$E_{\text{dihedrals}} = \sum_i {k_{\text{i}}}^{\text{dihed}}[1 + \cos(n_i\phi_{\text{i}}\phi_0)], \tag{3.8}$$

where ${k_i}^{dihead}$ is the force constant, n is multiplicity of the angle, ϕ_i is the phase factor and ϕ_0 its reference.

The fourth and the first non-bonded term describes the Lennard-Jones potential.

$$E_{\text{L-J}} = \sum_i \sum_{j \neq i} 4\epsilon_{\text{ij}} \left[\left(\frac{\sigma_{\text{ij}}}{r_{\text{ij}}} \right)^{12} - \left(\frac{\sigma_{\text{ij}}}{r_{\text{ij}}} \right)^{6} \right], \tag{3.9}$$

where σ_{ij} and ϵ_{ij} are parameters defined in the force field. The Lennard-Jones potential is split in two parts. The first term r^{-6} characterizes van der Waals interactions that arise from fluctions in the electronic distribution of neighbouring atoms. The second term r^{-12} characterizes repulsive interactions of neighbouring atoms. It has been shown to approximate the repulsive interactions well, even though the exponent does not have proper theoretical justification and has been chosen due to computational convenience.

The last term of the equation characterizes charged interactions which are described by the Coulomb law:

$$E_{\text{Coulomb}} = \sum_i \sum_{j \neq i} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{\epsilon_r r_{\text{ij}}}, \tag{3.10}$$

where q is charge, r_{ij} is distance between the two interacting atoms, ϵ_0 is dielectric constant of vacuum, and ϵ_r the relative dielectric constant.

3.1.4 Newtonian Dynamics

Classical molecular dynamics simulations are governed by Newton's second law of motion, which states that knowing the force acting on an atom with a known mass, it is possible to calculate its acceleration.

The Newton's equation of motion for a system:

$$m \frac{\partial^2 \boldsymbol{r}}{\partial t^2} = \boldsymbol{F}, \tag{3.11}$$

where m is the mass of an atom in the system, the second order derivative of the atom's position $\boldsymbol{r}$ is the acceleration of the atom due to force $\boldsymbol{F}$ acting on it.

At every time step of an atomistic simulation, the forces to each atom and the atoms' consequent accelerations are calculated. The forces can be obtained from a force field. Force fields provide a force model of potential functions that determine how different atoms interact with each other in the simulation system.

Potentials are partial derivatives of the forces:

$$\boldsymbol{F} = -\frac{\partial V}{\partial \boldsymbol{r}}, \tag{3.12}$$

where $\boldsymbol{F}$ is the force acting on an atom, expressed in the form of a negative partial derivative of potential V with respect to the position of the atom concerned.

3.1.5 Numerical Integration

To create a trajectory of a system with N atoms, the system needs to be propagated in time, as generally presented by Figure 3.1. Since the systems are generally too large for solving their development analytically, numerical integrators are utilized for this purpose. An integrator algorithm is used to integrate over the Newton's equation of motion for a small enough timestep, so that the forces in the system can be considered to be constants during it [83]. Because of this, the scale of the time step needs to very small, in atomistic simulations usually in the scale of femtoseconds. The integration is done for every atom in every time step and it gives out the forces acting on each of them. Of the forces, the accelerations of the atoms can be solved and when combined with coordinates and velocities from previous time steps, the coordinates and velocities of the new time step can be obtained. With the new velocities and positions for all atoms of the systems, the parameters for the next time step can now be calculated. Keeping the circle of integration going, a simulation trajectory can be obtained [83].

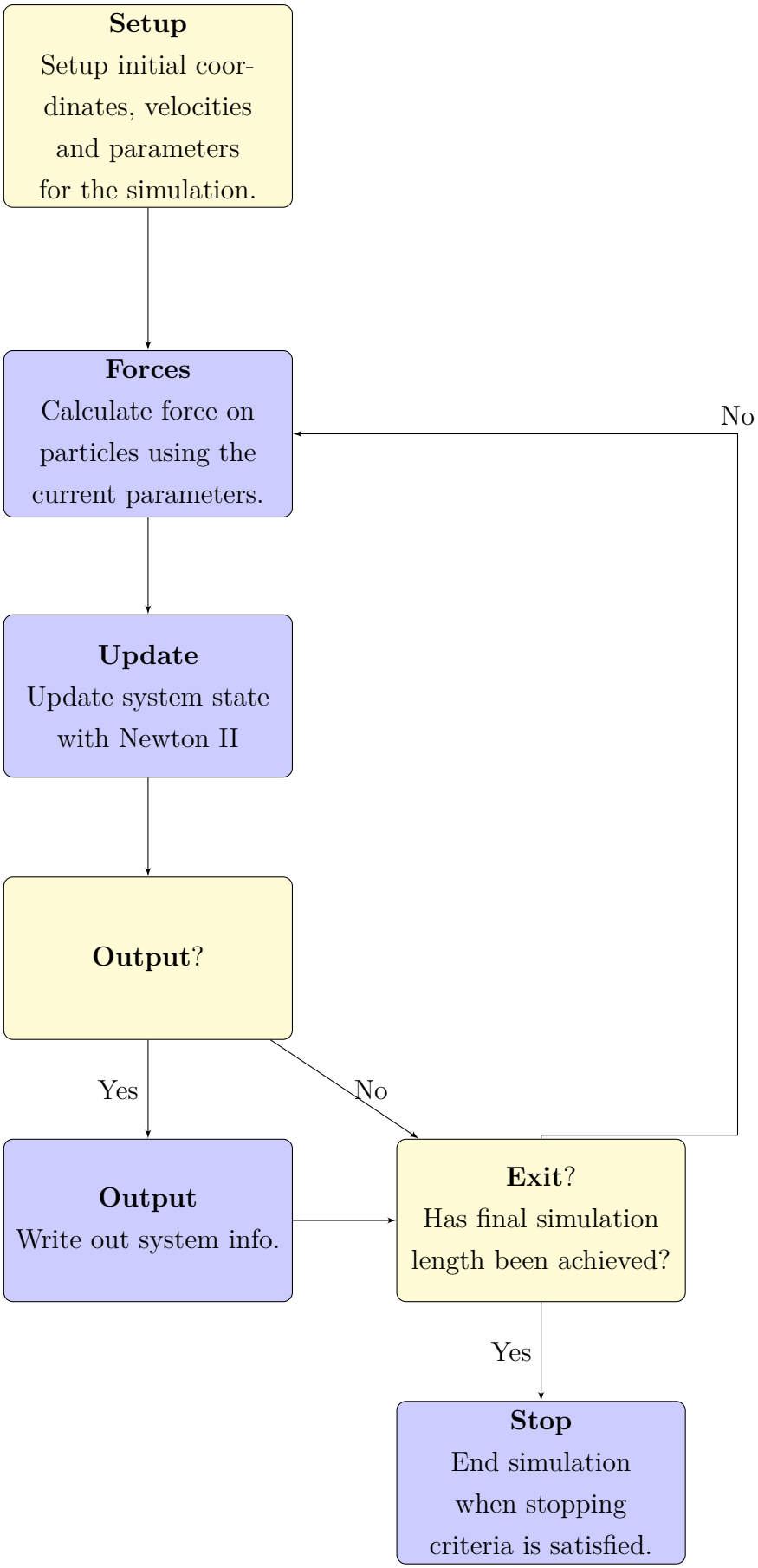


Figure 3.1: Schematic representation of the MD algorithm.

The validity of the integrator is crucial for the simulation to succeed. Most importantly the integrator has to be time reversible to avoid the system from developing into unphysical states. Also, it is desirable for the integrator to allow for as long a time step as possible to speed up the simulation process [81].

A simple algorithm for numerical integration in MD is the Verlet algorithm [84], where the positions of atoms are derived simply by expanding the expression for their positions with respect to time:

$$\boldsymbol{r}(t+\delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t-\delta t) + \delta t^2 \boldsymbol{a}(t). \tag{3.13}$$

However, a few issues are inherent to the Verlet algorithm [84]. The obvious main issue is that from the simple expression of the positions, the velocities do not automatically emerge and have to be computed separately. To obtain the velocities, for example the following expression can be used:

$$\boldsymbol{v}(t) = \frac{\boldsymbol{r}(t+\delta t) - \boldsymbol{r}(t-\delta t)}{2\delta t}, \tag{3.14}$$

where $\boldsymbol{v}$ and $\boldsymbol{r}$ are the velocity and position of an atom, respectively, and t is time.

To overcome the issue with velocities emerging in the integration, more sophisticated algorithms have been created. One of the most used numerical integrators, the leap frog [85], is a modification of the Verlet algorithm. The leap frog algorithm solves for particle positions $\boldsymbol{r}$ and velocities $\boldsymbol{v}$ as follows:

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t+\frac{1}{2}\delta t)\delta t \tag{3.15}$$

$$\boldsymbol{v}(t+\frac{1}{2}\delta t) = \boldsymbol{v}(t-\frac{1}{2}\delta t) + \boldsymbol{a}(t)\delta t. \tag{3.16}$$

Now both positions and velocities are obtained from the algorithm. Though as stated in equations 3.15 and 3.16, the velocities are always calculated a half step prior to the positions, so that the velocities can be used to obtain the new position of the atom at the full step. So, even though the leap frog is an improvement compared to the Verlet algorithm, it would still be desirable to modify the integrator to provide positions and velocities simultaneously.

Such conditions can be met with the Velocity Verlet algorithm [86], which also provides acceleration of each atom at every time step in addition to the positions and velocities. The Velocity Verlet algorithm works in four stages: first the algorithm calculates the positions of atoms from equation 3.17, then velocities at half step with 3.18. Next, using the coordinates obtained from 3.17 new forces are obtained at time step $t+\delta t$ to calculate acceleration $\boldsymbol{a}$ at time step $t+\delta t$. Lastly equation 3.19 is used to calculate velocities at the new time step $t+\delta t$.

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\delta t + \frac{1}{2}\boldsymbol{a}(t)\delta t^2 \tag{3.17}$$

$$\boldsymbol{v}(t+\frac{1}{2}\delta t) = \boldsymbol{v}(t) + \frac{1}{2}\boldsymbol{a}(t)\delta t \tag{3.18}$$

$$\boldsymbol{v}(t+\delta t) = \boldsymbol{v}(t+\frac{1}{2}\delta t) + \frac{1}{2}\boldsymbol{a}(t+\delta t)\delta t \tag{3.19}$$

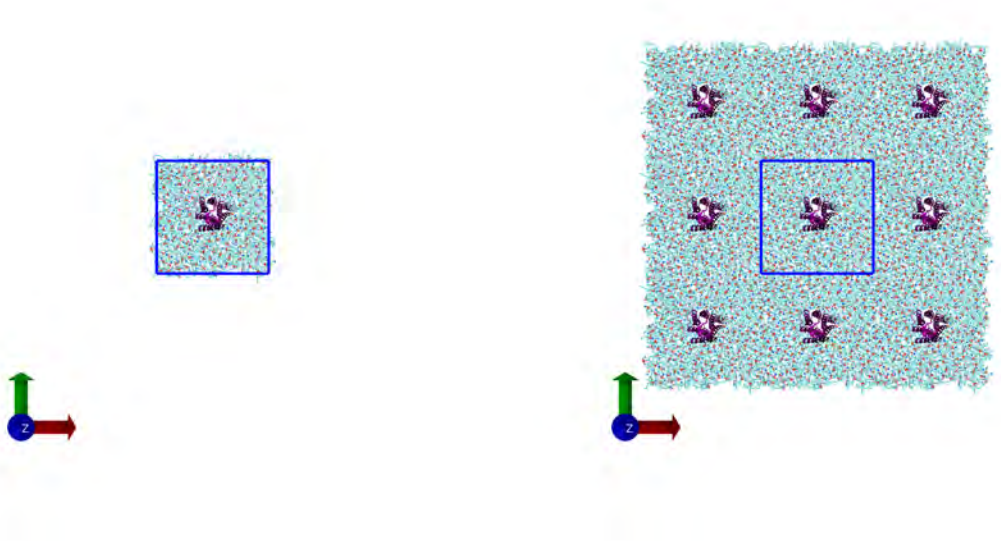
3.1.6 Periodic Boundary Conditions

The goal of MD simulations is usually to get a closer look at some part of the microscopic world that is not reachable through other methods. So it is safe to assume that the model system inside the simulation box is a part of a larger scheme, and would in reality be surrounded by a sea of lipids, proteins and other biological entities that together make up a membrane surrounding a living cell. On that note, it seems justified to conclude that the simulation box cannot be simulated as a separate entity, since in that scenario at all sides of the simulation box, the atoms in the box would have nothing to interact with. This problem with the vacuum surrounding the simulation box, is overcome with a simple construct called the periodic boundary conditions (PBCs) [81]. So, on every side of the simulation box, or the unit cell, resides an exact copy of the system itself which in fact is not a copy, but the system itself as visualized by image 3.2. This is accomplished by looping the atoms and interactions from one side of the unit cell to the other when they cross over the dimensions of the box, so that when something leaves the box, it will immediately re-emerge from the opposite side.

The same treatment also applies for all interactions and hence, it is possible for a particle to feel the attraction or repulsion of another particle on the far other side of the simulation box. In principle, it is possible for any two particles to interact within the box and also across all sides of the unit cell. This is why the so-called minimum image convention is used, in order to make sure that the same two atoms do not interact with each other in more than one way at the same time. Due to the minimum image convention, atoms of the system are only allowed to interact with the closest projections of each other, which consequently sets a limit for the maximum interaction distance between two atoms to be half of the minimum diameter of the unit cell [81].

3.1.7 Statistical Ensembles

Atomic details such as motion and structure are not usually relevant for studying the macroscopic properties of a simulation system, so the simulation is lightened by averaging over the details by methods of statistical mechanics. Macroscopic properties of the system



(a) Unit cell with a protein embedded in membrane. **(b)** Unit cell with its periodic images in plane.

Figure 3.2: Cubic unit cell and its nearest periodic images in xy-plane, snapshots of the second system in Table 4.1.

are always obtained as an ensemble average over a representative statistical ensemble. The most commonly used ensembles in MD methods are canonical (NVT), often used to equilibrate MD systems, where number of particles, volume, and temperature are kept constant. Also microcanonical ensemble (NVE) can be used, which conserves the number of particles, volume, and total energy. Finally, isothermal-isobaric (NpT) which is an often used ensemble in MD simulations, employed in order to mimic the conditions of real-life experiments. In NpT conditions the number of particles, pressure and temperature are constant. The chosen thermodynamic ensemble is set and maintained with the choice of barostat and thermostat.

3.1.8 Thermostats and Barostats

Statistical ensembles are obtained and maintained by employing thermostats and barostats to control the temperature and pressure of the system, in respective order.

The Berendsen thermostat [87] uses a weak coupling scheme to keep the temperature of the simulation system constant. The set reference temperature is reached and maintained by frictional constants that are used to scale the velocities of the atoms in the system. Even though the Berendsen temperature coupling is extremely efficient in stabilizing the temperature of the system, it also suppresses the kinetic energy fluctuations of

the particles and hence, the Berendsen thermostat does not produce true statistical ensembles [88]. Because the Berendsen thermostat reaches the desired temperature fast, it is often used in the equilibration stage of an MD simulation to set the reference temperature and more refined thermostats are used to run the production simulation.

A popular example of such a thermostat is the Nosé-Hoover thermostat [89]. In the Nosé-Hoover thermostat the frictional term that characterizes the strength of coupling between the system and its heat bath is more refined compared to the Berendsen coupling. This causes the temperature of the system to fluctuate around the reference value more than with the Berendsen thermostat. The Nosé-Hoover thermostat is known to produce correct canonical (NVT) ensembles. Hence it is often used in simulations after initially equilibrating the temperature with the Berendsen thermostat.

Pressure coupling in MD simulations can be implemented in three manners:
1. Isotropic, which is the most used way. In isotropic implementation, all the dimensions of the simulation box are scaled by the same measure.
2. Semi-isotropic coupling scales one dimension of the box independently and the other two by the same measure.
3. Anisotropic scales all box dimensions independently.

Alike with the thermostat, Berendsen barostat is often used in the equilibration phase of the simulation to equilibrate the system pressure. The Berendsen barostat works simply by applying a scaling factor to the dimensions of a simulation box. This scales the coordinates of all atoms, and the pressure of the system starts decaying exponentially toward the desired reference pressure. However, the Berendsen barostat, as its thermostat counterpart, does not produce the correct isothermal-isobaric (NpT) ensemble [90].

The Parrinello-Rahman barostat is in principle implemented on top of the simple Andersen barostat. The Andersen barostat sets the pressure of the system by controlling the dimensions of the simulation box by acting as a piston [91]. The Parrinello-Rahman barostat is simply an extension of the Andersen barostat which allows this pressure coupling scheme to be implemented for different shapes of boxes [90]. The Parrinello-Rahman barostat is also known to produce the correct NpT ensemble, which is why it is often used as a barostat in production simulations after equilibration.

3.1.9 Fine-Tuning Interactions

It would simply be computationally too expensive to calculate the non-bonded interactions of each atom with all other atoms in the system which is why cut-off methods are used to limit the calculation of non-bonded interactions to the nearest neighbors of a given

atom. This allows for faster and more efficient simulations. Due to the difference in the magnitude of the exponent of the non-bonded Lennard-Jones interactions (equation 3.9) the attractive interactions fade faster than repulsive interactions. This is why the Lennard-Jones interactions are usually handled with a cut-off list that ensures that the interactions are only calculated between atoms that are reasonably close to each other [81].

Electrostatic interactions have a longer range and using cut-off lists to reduce them would create artefacts to the simulation. To treat them in a realistic yet computationally effective manner, the more prominent short-range Coulombic interactions are calculated explicitly and long-range interactions are approximated with lattice-sum methods [81]. An example of such is the Particle Mesh Ewald (PME) method which uses reciprocal space to sum over the long-range electrostatic interactions. PME is currently one of the best and fastest ways to calculate long-range interactions in MD simulations [92].

3.1.10 Constraints and Restraints

If some part of the simulation system is desirable to be kept invariant during the simulation run, constraints and restraints can be used for this purpose according to the desired level of invariability. Constraints are boundary conditions set before the initialization of MD run that the model must satisfy. Constraints can be utilized to set, for example, positions, angles, distances or orientations of bonds invariable. Since constraints are set before starting the simulation to preserve the initial state of some part of the system, they are based on fixed lists that are not updated during the simulation run. Constraints are mostly used to increase the integration timestep of the simulation, but can also be useful when an experimentally determined structure is desired to be included in the MD simulation system. Restraints, on the other hand, are additional energy functions that aim to keep the system in a desired state. They do not, however, subject absolute invariability to the structure they restrain [81].

3.2 Biophysical Properties

The analysis of the order and lateral diffusion of lipids are crucial to gain a comprehensive understanding of the functional differences between planar bilayer and nanodisc systems. Acyl chain order parameter provides a quantitative measure of lipid chain conformational order, which directly impacts membrane fluidity, stability, and interactions with embedded transmembrane proteins. Evaluating lateral diffusion of lipids offers insights into the dynamic behavior of the lipids, shedding light on the mobility and exchange processes within the bilayer both in planar bilayer and nanodisc environment. Consequently, a

thorough analysis of these parameters is critical for elucidating the biophysical properties of lipids in both systems, and the consequent implications in membrane protein research.

3.2.1 Acyl Chain Order Parameter

Order parameter provides a quantitative measure to study how the lipid tails are oriented in a membrane. The calculation is conducted by measuring the orientation of a C-H bond in a lipid chain with respect to the normal of the bilayer, which often is the z-axis in membrane simulations. The calculation is performed for all C-H bonds in the lipid acyl chain and separately for both *sn-1* and *sn-2* chains of the lipid. The end result is usually averaged over all the lipids in the membrane and over the sampling time. Order parameter is calculated as:

$$S = \frac{3\langle \cos^2\theta - 1 \rangle}{2}, \tag{3.20}$$

where θ is the angle between z-axis or bilayer normal and the vector of a C-H bond. The angular brackets represent average over the number of lipids and time [93].

Often the order parameter is calculated for all C-H bonds and averaged over all lipids. However, since in this work the goal is to examine the lipid population around a membrane protein, the order parameter is calculated for each lipid separately and averaged over time. This way, the distribution of lipids in the systems can be studied as a function of distance from the protein.

In effect, the lower the value of the order parameter is, the more ordered the membrane is and the tighter the lipids are packed in it. Vice versa, for higher values of order parameter, the membrane can be expected to be less ordered and lipid movement to be faster and less restricted. Due to this inversionally proportional relationship of the membrane packing and order parameter value, the order parameter is also often expressed as its negative counterpart: $-S$ [93].

3.2.2 Lateral Diffusion

Typically, diffusion of lipids is quantified by its rigorous definition in terms of mean-squared displacements (MSD) of lipids at long time scales [94]. The MSD's can be calculated as:

$$\mathrm{MSD}(t) = \langle [\boldsymbol{r_i}(t) - \boldsymbol{r_i}(0)]^2 \rangle, \tag{3.21}$$

where $\boldsymbol{r_i}(t)$ is the position of the center-of-mass (COM) of lipid i at time t and the angular brackets denote average over all the lipids in the examined system and time.

From MSD, lateral diffusion coefficient in the membrane plane can be derived as [95]:

$$D_L = \lim_{t\to\infty} \frac{\mathrm{MSD}(t)}{4t}. \tag{3.22}$$

The long-time limit in equation 3.22 assumes that the lateral diffusion coefficient D_L is determined from the region, where MSD scales as a power-law in time as:

$$\mathrm{MSD}(t) \sim t^{\alpha}, \tag{3.23}$$

and $\alpha = 1$ [96].

However, since diffusion in the nanodisc systems studied in this case happens in a confined environment, the rigorous MSD method is not the technique to study the lateral diffusion of lipids. Instead, we can assume for the COM's of the lipids to follow random walk, meaning that the lengths of lipid displacements as well as the directions, should be random within a fixed time period. Consequently, studying the displacement distribution of the lipids over that time period, the distribution should show Gaussian behaviour that allows to determine the approximate lateral diffusion coefficient through fitting of a Gaussian over the distribution [97].

Considering a random walker in one-dimensional scenario, starting at time $t = 0$ at a starting position of $x = 0$ and over time t is found to have moved to distance $x + \Delta x$. The probability for this move to happen can be expressed as [98]:

$$P_{1d}(x,t)\Delta x = \frac{1}{\sqrt{4\pi D_{1d} t}} e^{\left(-\frac{x^2}{4D_{1d}t}\right)} \Delta x, \tag{3.24}$$

where D_{1d} is the diffusion coefficient in one dimension.

Similarly in two dimensions, where D_{2d} is the diffusion coefficient for two-dimensional diffusion, we have:

$$P_{2d}(xy,t)\Delta x\Delta y = \frac{1}{4\pi D_{2d} t} e^{\left(-\frac{(x^2+y^2)}{4D_{2d}t}\right)} \Delta x\Delta y. \tag{3.25}$$

Considering circular symmetry, it is feasible to change to spherical coordinates and study the displacement distributions of lipids radially in terms of $\Delta r, \Delta\Theta, \mathrm{and}, \Delta\Phi$. Integrating equation 3.25 over Θ and Φ, we acquire the probability distribution for the distance that a COM of a lipid has travelled within a fixed time interval as a function of r [99]:

$$P_{2d}(r,t)\Delta r = \frac{r}{2D_{2d} t} e^{\left(-\frac{r^2}{4D_{2d}t}\right)} \Delta r. \tag{3.26}$$

The fitting of equation 3.26 over the distribution of distances travelled by lipids, will now yield the two dimensional diffusion coefficient D_{2d} over the corresponding time interval.

3.3 Machine Learning Methods

The application of machine learning (ML) in the data analysis of MD simulations is increasingly becoming indispensable for driving scientific breakthroughs. ML algorithms can efficiently process the vast amounts of data generated in MD simulations, revealing key insights into the underlying molecular mechanisms with reduced human intervention. The ML methods used to analyse the results presented in this thesis highlight some of the key advantages of incorporating ML into MD simulations projects: uncovering hidden relationships within data with feature extraction, enhanced interpretability and a more comprehensive understanding of the complex phenomena happening in the systems.

3.3.1 Principal Component Analysis

Due to the nature of the method, working on biological systems with molecular dynamics simulations by default means working with inherently noisy and high dimensional data. Considering a typical simulation system, it often comprehends a number of atoms from anywhere between tens to millions. Hence, it might be desirable to reduce the dimensionality of the data before analysing it further. One of the established dimensionality reduction methods is the Principal Component Analysis (PCA) that allows to reduce the dimensionality of data while striving to preserve as much of its variance and, in effect, interpretability as possible [100]. In the realm of biological systems, this means, for example, finding and preserving the largest motions and mobile areas of the system.

In PCA, the objective is to summarize the original data in a small number of representative variables that collectively explain as much of the variability of the original dataset as possible. In other words, PCA seeks to find out a new, reduced feature space in which the original data is highly variable. In most cases, it is safe to presume that in a large dataset, not all the features of the data are as informative and interesting as others; interesting in this case meaning the highest variance along that dimension [101]. So, with the means of PCA, a set of orthogonal principal component (PC) vectors in the directions of the highest variance in the data can be obtained and used as a basis for a lower dimensional subspace to project the data into [100].

Considering a dataset $\mathbf{X}$, with n observations x and p features, and the dataset is assumed to be normalized to have mean zero [101].

The first principal component Z_i is a linear combination of the features

$X_1, X_2, ..., X_p$.

$$Z_1 = \Phi_{11}X_1 + \Phi_{21}X_2 + ... + \Phi_{p1}X_p, \tag{3.27}$$

where $\Phi_{11}, \Phi_{21}, ..., \Phi_{p1}$ are the loadings of the first principal component and together make up the PC loading vector:

$$\Phi_1 = (\Phi_{11}, \Phi_{21}, ..., \Phi_{p1})^T \tag{3.28}$$

The loadings $\Phi_{11}, \Phi_{21}, ..., \Phi_{p1}$ are constrained so that the sum of their squares equals one:

$$\sum_{j=1}^{p} \Phi_{j1}^2 = 1, \tag{3.29}$$

in order to normalize the PC vector and prohibit the variance from arbitrarily large absolute values.

To find the first principal component of dataset $\mathbf{X}$, with n observations x in p dimensional feature space, we look for a linear combination of the sample features that has the largest sample variance within the normalization constraint 3.29:

$$z_{i1} = \Phi_{11}x_{i1} + \Phi_{21}x_{i2} + ... + \Phi_{p1}x_{ip}, \tag{3.30}$$

where $\Phi_{11}, \Phi_{21}, ..., \Phi_{p1}$ are the elements of the first PC loading vector.

Using equation 3.30, an objective for the first principal loading vector can be formulated as the following optimization problem:

$$\text{maximize}_{\Phi_{11},...,\Phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \Phi_{j1}x_{ij} \right)^2 \right\} \text{ subject to} \sum_{j=1}^{p} \Phi_{j1}^2 = 1, \tag{3.31}$$

The problem above can be solved with the means of eigen decomposition by considering the PC loading vectors Φ_1, Φ_2, Φ_3... (3.28) as the eigenvectors of the matrix $\mathbf{X}^T\mathbf{X}$ and the variances of the components as corresponding eigenvalues.

After the first principal component Z_1 has been determined, the second principal component Z_2 can be determined again as a linear combination of the features $X_1, X_2, ..., X_p$ that has the largest variance out of all the linear combinations that are uncorrelated with the first principal component Z_1. As for the first principal component, the scores of the second PC acquire a corresponding form to 3.30:

$$z_{i2} = \Phi_{12}x_{i2} + \Phi_{22}x_{i2} + ... + \Phi_{p2}x_{ip} \tag{3.32}$$

By constraining Z_1 and Z_2 to have be uncorrelated, or in other words to have zero covariance with each other, it is also secured that the two principal components are

orthogonal to each other. To find the second principal component loading vector Φ_2, we again optimize the objective defined in 3.31, but with the added restraint that vector Φ_2 is orthogonal to Φ_1 as stated above.

If the dataset is large with multiple features, subsequent PC vectors can be solved in identical manner. Once the principal components are solved, they can be selected as the new orthonormal basis for the dataset and project the original data for low-dimensional views in the subspaces spanned by the PC vectors Φ_1, Φ_2, Φ_3...

To investigate how much information has been lost by projecting the observations onto the principal components, we can look at Proportion of Variance Explained (PVE). If the data are normalized to have a mean zero, the total variance in the dataset is:

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2, \tag{3.33}$$

and the PVE of the mth principal component can be acquired as:

$$PVE = \frac{\sum_{i=1}^{n} (\sum_{j=1}^{p} \Phi_{jm} x_{ij})^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2} \tag{3.34}$$

3.3.2 Gaussian Mixture Models

Mixture models are a tool to find subpopulations within data without any prior knowledge about possible substructures of the data. To find such underlying features in the data, mixture models fit distributions into the data in order to estimate the best combination of distributions to describe the dataset. Each distribution in the final combined model can be interpreted as a cluster, and by estimating the probabilities of each datapoint belonging in them, it is possible to identify subpopulations within the whole dataset [102].

In accordance with the name, in Gaussian mixture models, the fitted distributions $\aleph$ have the Gaussian form:

$$\aleph(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}} \tag{3.35}$$

where $\mathbf{x} = (x_1, x_2..., x_m)$ is the m-dimensional random variable, $\boldsymbol{\mu} = (\mu_1, \mu_2..., \mu_m)$ are the means, and $\boldsymbol{\sigma^2} = (\sigma_1^2, \sigma_2^2..., \sigma_m^2)$ variances of the distribution. When provided with the number of distributions K, the model will find the means, variances and coefficients π to maximize the likelihood of the data being from the distribution:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \aleph(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\sigma_k}^2)), \tag{3.36}$$

where $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$ [102].

The fitting of linear combinations of distributions to the data is managed by estimating the values of means $\boldsymbol{\mu}_k$, variances $\boldsymbol{\sigma}_k^2$, and coefficients π_k for each cluster. This is carried out by the Expectation Maximization (EM) algorithm, a general algorithm for learning about variables [103]. After setting the initial values for the parameters, the algorithm alternates between two steps - expectation and maximization - until the parameters converge. In the expectation step, the responsibilities r_{nk} are evaluated. In principle, the responsibilities determine the probabilities of data points to belong to component k in the combination of distributions. In the maximization step, the parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\sigma}_k^2$, and π_k are re-estimated according to the new values of corresponding responsibilities.

In the setting of parameter estimation for GMM, the EM-algorithm is implemented as illustrated by the following steps [102]:

1. **Initialise**

Set initial values $\boldsymbol{\mu}_k$, $\boldsymbol{\sigma}_k^2$, and π_k for the K components.

2. $\mathbf{E-step}$

Evaluate responsibilities r_{nk} for all components k and each datapoint $\mathbf{x}_n$:

$$r_{nk} = \frac{\pi_k \aleph(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)}{\sum j = 1^K \pi_j \aleph(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)} \tag{3.37}$$

3. $\mathbf{M-step}$

Recalculate all the corresponding parameters:

$$N_k = \sum_{n=1}^{N} r_{nk} \tag{3.38}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n \tag{3.39}$$

$$\boldsymbol{\sigma}_k^2 = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} |\mathbf{x}_n - \boldsymbol{\mu}_k|^2 \tag{3.40}$$

$$\pi_k = \frac{N_k}{N} \tag{3.41}$$

4. **Repeat**

Repeat E- and M-steps (2 and 3) until values of the parameters converge.

If there is not enough information about the dataset to determine how many components K are enough to satisfactorily describe the data, a set of criteria is needed to determine, which fit provides the best model. The same criteria can also be utilized, when the number of components is known, but the quality of the fit needs to be assessed to avoid overfitting which can result from adding too many free parameters to the model to increase its likelihood estimate. Two such criteria are often used for these purposes:

the Akaike Information Criterion (AIC) [104] and Bayesian Information Criterion (BIC) [105].

Considering a model with N data points, and p free parameters, and the maximum likelihood estimate of $\hat{L}$, the two criteria get the following forms:

$$\mathrm{AIC} = -2\log(\hat{L}) + 2p \tag{3.42}$$

$$\mathrm{BIC} = -2\log(\hat{L}) + \log(N)p \tag{3.43}$$

The shared form reveals the close relation of the two criteria, the difference appearing only in the second term. The penalty term of AIC remains constant, whereas the penalty term of BIC-criteria scales logarithmically with the accumulation of data points N. Due to its stronger penalty on the free parameters, the BIC-criterion is more likely to point towards a model of lower dimensionality than AIC [105].

4. Simulations

4.1 Systems

To examine differences between planar bilayer and nanodisc systems, eight simulation systems were built that varied in the type of the system and their lipid content. Two different types of lipids, POPC and DMPC, were used to build two planar and two nanodisc systems of each lipid. One system of each of the two system types was either a pure lipid planar bilayer or a pure lipid nanodisc system, and the other also had the transmembrane protein $A_{2\alpha}R$ in the membrane. Hence at the end, eight distinct systems were studied. Table 4.1 shows the detailed compositions of all systems used in this thesis.

The two phospholipids used in this study were POPC and DMPC. POPC (1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine) is a phospholipid which is naturally present and one of the most abundant lipid in eykaryotic cell membranes [106]. Hence, it is often used in studies of the membrane and to mimic the plasma membrane conditions, for example in nanodiscs [14]. DMPC (1,2-dimyristoyl-*sn*-glycero-3-phosphocholine), on the other hand, is a synthetic phospholipid. It is commonly used to mimic a mammalian plasma membrane in studies of plasma membranes and liposomes, and also to study the interactions of lipids with membrane proteins. DMPC's shortcoming is its shorter than average chain length compared to human membranes which sets limitations for its use [107].

The transmembrane protein in the systems, $A_{2\alpha}R$ is one of the four adenosine receptor subtypes in the rhodopsin-like family of GPCRs. PDB: 5G53 was used as a starting structure for $A_{2\alpha}R$ in the simulations and the missing residues were added into the structure using MODELLER [82]. The scaffolding protein used in all the nanodiscs was MSP1E1 [1].

Both planar and nanodisc systems in the study were originally constructed using CHARMM-GUI [108]. However, since CHARMM-GUI does not yet support the scaffolding protein MSP1E1 used in the nanodiscs, the nanodisc systems were initially built in CHARMM-GUI using another membrane scaffolding protein called MSP1D1. This MSP was chosen as the initial replacement for MSP1E1 due to its length that was the closest in number of residues to MSP1E1. The starting structures for MSP1E1 were constructed

Type	Lipid	No of Lipids	MSP	Membrane Protein	No of Replicas
Planar	POPC	300	-	-	10
Planar	POPC	321	-	A2AR	10
Planar	DMPC	414	-	-	10
Planar	DMPC	277	-	A2AR	10
Nanodisc	POPC	275	MSP1E1	-	10
Nanodisc	POPC	240	MSP1E1	A2AR	10
Nanodisc	DMPC	273	MSP1E1	-	10
Nanodisc	DMPC	260	MSP1E1	A2AR	10

Table 4.1: The simulation systems and their constituents.

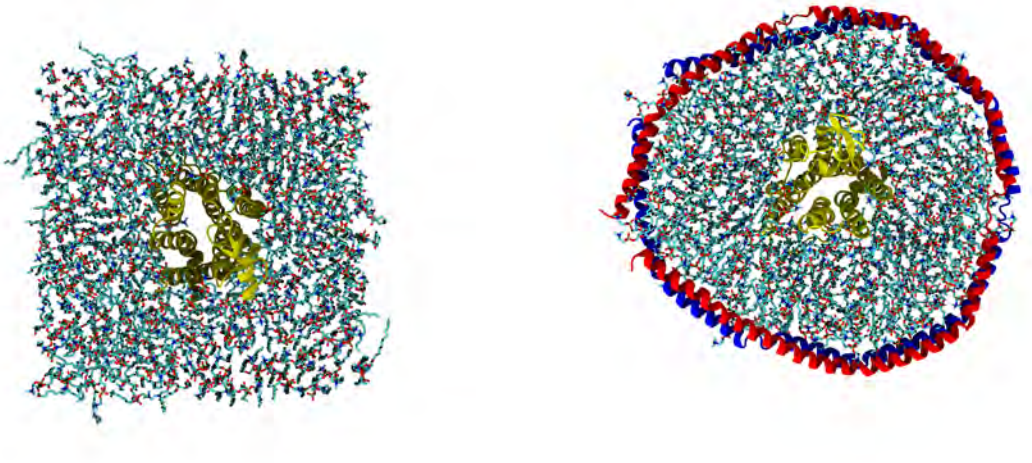
from FASTA sequences with the help of a recent nanodisc generation protocol [109], before carefully substituting the CHARMM-GUI generated MSP1D1 scaffolding proteins with the desired MSP1E1s in the nanodiscs. After the change of the scaffolding protein, the nanodisc systems were resolvated. All the equilibration steps and simulation runs were performed with the correct MSP1E1 scaffolding proteins.

All the systems were solvated in water, and ions were added to the system to neutralize charges, and to establish ion concentration of 0.15M KCl. The CHARMM36m force field was used for all proteins [110] and lipids [111] and the TIP3P model was used for water [112] with compatible parameters for ions [113]. For each of the eight systems, 10 replicas were run independently.

All simulation systems went through the same simulation protocol. First the systems were minimized and equilibrated in the NVT ensemble with a temperature of 310 K. Heavy atoms were restrained during the equilibration with decreasing constants on each repetition of the short equilibration periods. Also, side chains were set to have less rigorous restraints than backbone to allow for more efficient equilibration. After the equilibration, the systems were simulated for 500 ns in the NpT ensemble for a longer equilibration to give the nanodiscs sufficient time to equilibrate. After the long equilibration, a production run of 100 ns was conducted again in the NpT ensemble for analysis, Figure 4.2.

4.2 Simulation Protocol

All simulations were carried out using GROMACS 2021 [114]. The equations of motion were integrated using the leap frog algorithm with a 2 fs time step. All covalent bonds involving hydrogens were constrained using the LINCS algorithm [115]. Long-range electrostatic interactions were treated by the fast smooth particle mesh Ewald scheme [116] with a cutoff of 1.2 nm, Fourier spacing of 0.12 nm, and fourth-order interpolation was



(a) Planar bilayer system with an embedded transmembrane receptor $A_{2\alpha}R$.

(b) Nanodisc system with $A_{2\alpha}R$ and scaffolding protein MSP1E1.

Figure 4.1: Example of a planar bilayer system with an embedded transmembrane receptor $A_{2\alpha}R$ and its nanodisc counterpart.

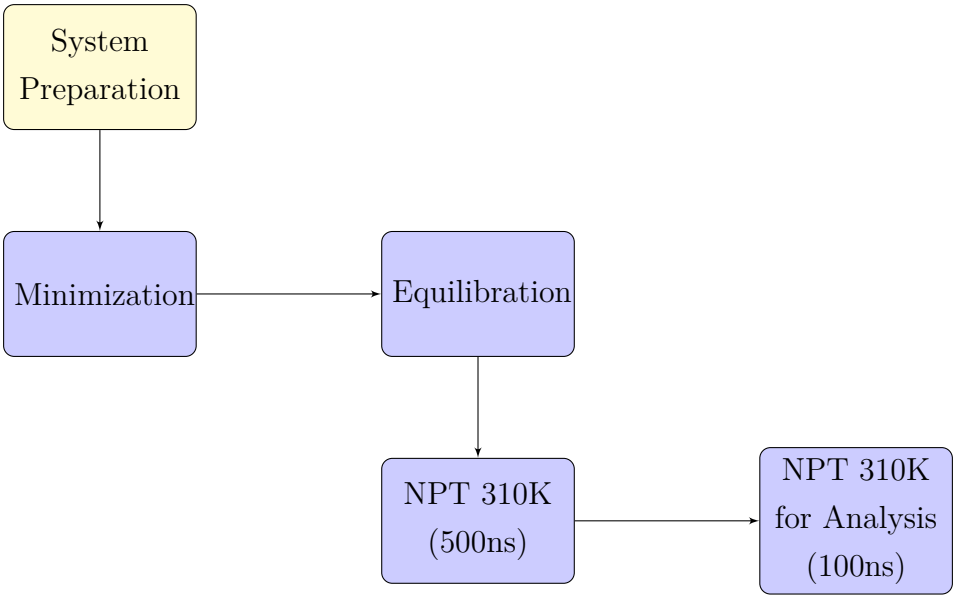


Figure 4.2: Schematics of the simulation protocol.

used. For van der Waals interactions, a Lennard-Jones potential with a force-switch between 1.0 and 1.2 nm was used.

Before the long equilibration and production runs, all systems were first minimized using the steepest descent algorithm and later equilibrated in stages with the heavy atoms of proteins kept restrained at 310 K and 1 atm pressure using the Berendsen thermostat and barostat [87].

All production simulations were performed in the NpT ensemble. In the long equilibration runs, the temperature was maintained at 310 K using the Berendsen thermostat [87]. The protein, membrane, and solvent (water and ions) were each coupled to separate heat baths at the reference temperature with a time constant of 1.0 ps. The production simulations, on the other hand, employed the same leap frog integrator and the Nosé-Hoover thermostat [89], [117] maintaining the temperature at 310 K. The constituents of the system were each coupled to separate heat baths as described before and the time constant was again 1.0 ps. Pressure was controlled semi-isotropically using the Berendsen barostat [87] with a reference pressure of 1 atm, a time constant of 5 ps, and compressibility of 4.5×10^{-5} bar^{-1} on the membrane plane.

The final simulation length for each system considering the long equilibration and the sampled production run was over 600 ns, of which the last 100 ns was used in analysis of the lipids. In addition, the first 500 ns, with 100 ns cut off from the beginning, was used in the machine learning analysis of membrane protein behaviour.

For the purposes of this thesis, all simulations were run on the supercomputer Mahti of CSC (IT Center for Science, Finland). PyMOL [118] and VMD [119] with Tachyon [120] renderer were used to construct the 3D images in this document. All the graphs in the results section were made with the Matplotlib Python package [121].

5. Results and Discussion

In order to investigate the differences between planar bilayers and nanodiscs, MD simulations with 8 different systems and two different lipid types were performed.

Each simulation system had 10 independent trajectories from which the results were derived by averaging over all trajectories. The analysis was conducted using the last 100 ns of the total 600 ns simulation per trajectory. At this stage in the simulations, the system and the analysed properties were assumed to have equilibrated to the point where they could reliably be quantified. Before analysis, in all the systems that had the transmembrane protein, the protein was centered and fixed to the center of the simulation box. In addition, in all the nanodisc systems, the nanodiscs were reconstructed so that they did not cross the periodic boundaries and centered into the simulation box from the scaffolding proteins or the transmembrane protein.

First, results concerning the lipids within the simulated planar and nanodisc systems will be discussed. These are composed of the analysis of the acyl chain order parameter and diffusion of lipids. Afterwards, the results of the machine learning approaches, principal component analysis and Gaussian mixture models, will be presented to decipher the behaviour of the transmembrane protein in these two systems. In the following discussion, it is argued how the lipid environment in each system affects the membrane protein in it.

5.1 Order Parameter

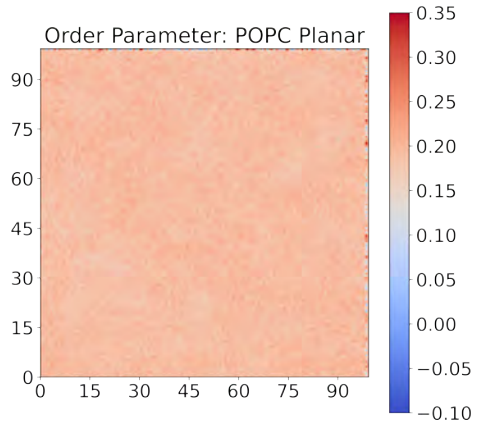
The acyl chain order parameter was calculated by using the carbon atoms of the *sn-1* chain of all the lipids in each system. The analysis of was done in two parts: the first step was to calculate the average order parameter of each individual lipid over the analysed length of the trajectory and projecting the resulting values on a 2D plane (xy-plane of the simulation system) in order to visualize the distribution of order parameter within the system. In the second step, the data were presented as a function of radial distance from the center of the system. The analysis was conducted in a similar manner for the planar and nanodisc systems.

As expected, the 2D order parameter distributions of planar bilayers without membrane protein 5.1 show a constant distribution values for the acyl chain order parameter.

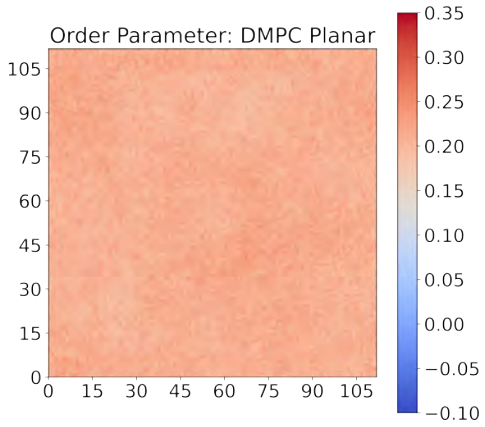
The system with POPC lipids exhibits a value of just below 0.2 and the system of DMPC lipids a value of a little over 0.2, meaning that the DMPC lipids are in a slightly more ordered state than the POPC lipids. This can be explained by the difference in the main transition temperatures of the two lipids at which they change from ordered gel phase to a disordered liquid crystalline phase: for POPC the main transition temperature is 271 K and for DMPC 297. As the simulations were conducted in the temperature of 310 K, the temperature is far further from the main transition temperature of POPC than DMPC, making POPC more likely to be disordered. The reason can also be found in the length of the lipid chains: as DMPC possesses shorter chains, it is more likely to be able to pack them efficiently and exhibit higher order than POPC.

In the 2D order parameter distributions of planar bilayers with membrane protein $A_{2\alpha}R$ at the center of the system, Figure 5.2, the effect of the membrane protein addition can be visualized. At the immediate vicinity of the protein, the lipids are relatively disordered compared to the rest of the system. This is an effect imposed by contact with the surface of the protein. The outer surface of the membrane protein is not flat and smooth, but has in- and out curved areas to its body. If there is no lipid residing in an inward curve, a tiny local vacuum is created in this space if no lipid is presently occupying it. As this is highly unfavourable, it creates an immense pressure with which the surrounding lipids press themselves towards the protein to fill the space. When a lipid comes in contact with the protein, the pressure imposed on it by the surrounding lipids makes it ordered by pressing it onto the protein surface. A lipid approaching an outward curve will also be pressed towards the surface of the protein by the surrounding lipids. However, since the surface protrudes outwards, the lipid cannot preserve an ordered straight acyl chain, but becomes disordered. This is the effect that can be seen in Figure 5.2 around the protein.

However, moving further away from the protein to the area, where the lipids are not in immediate contact with the protein anymore, but instead surround the protein completely. In this area, an increase in the order of the lipids is visible. Moving further away from the protein and towards the ends of the systems, the increased ordering caused by the protein can be discovered to weaken and finally descend towards the value of order parameter in the corresponding planar system without a membrane protein. This can be visualized better in the plot of radial distribution of order parameter in Figure 5.3.

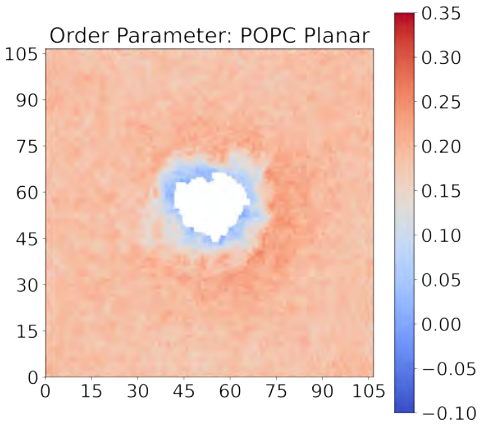


(a) POPC lipids in planar system.

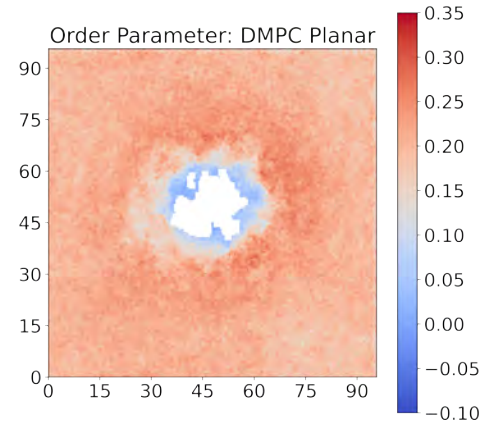


(b) DMPC lipids in planar system.

Figure 5.1: 2D distributions of order parameter in planar systems without $A_{2\alpha}R$.

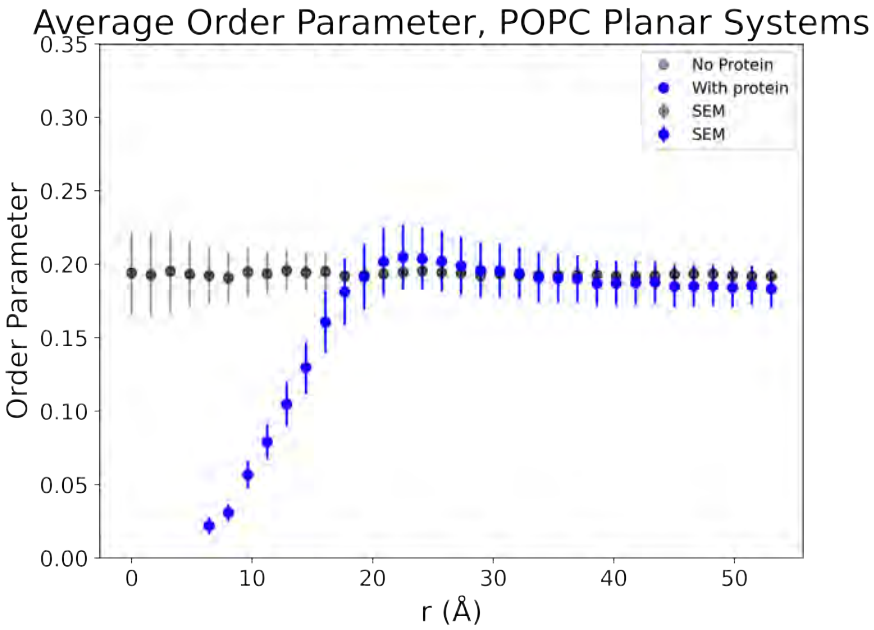


(a) POPC lipids and $A_{2\alpha}R$ in planar system.

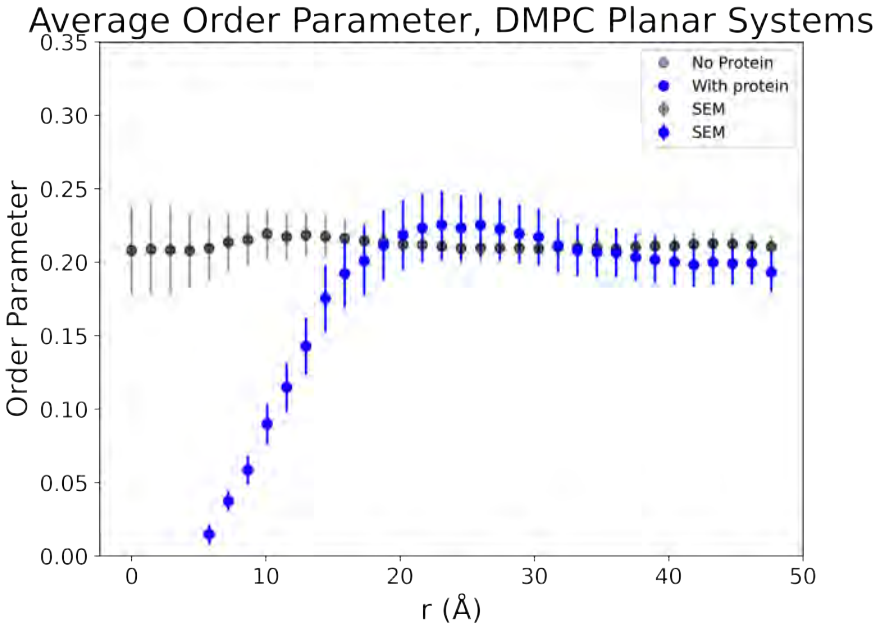


(b) DMPC lipids and $A_{2\alpha}R$ in planar system.

Figure 5.2: 2D distributions of order parameter in planar systems with $A_{2\alpha}R$.



(a) Planar systems with POPC.



(b) Planar systems with DMPC.

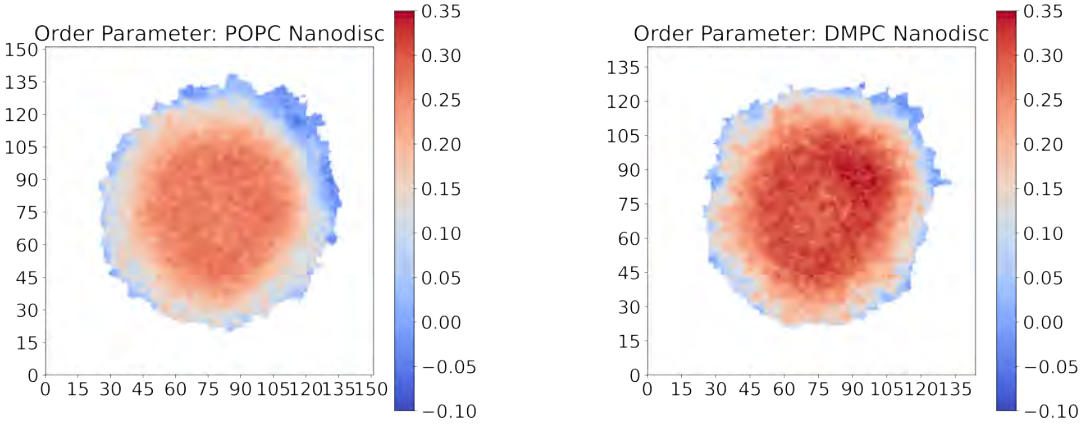
Figure 5.3: Order parameter as a function of radius in planar systems, uncertainty of data points expressed in form of Standard Error of the Mean (SEM).

As previously published in multiple studies, the 2D visualization of order parameter in nanodiscs without a membrane protein, reveals a distribution of three distinct regions, Figure 5.4. In the center of the nanodisc, the order is the highest and in the edges of the nanodisc and in the immediate vicinity of the membrane scaffolding protein, the lipids are extremely disordered. In between the two regions, a third intermediate region can be detected, in which the lipids are changing between the two regions and the value of order parameter is somewhere between the two extremes. Similarly with respect to the planar systems, the overall order is higher in the DMPC nanodiscs (around 0.3 in the middle of the nanodisc) than in the POPC nanodisc (a bit over 0.25 in the middle) as read from Figure 5.6.

In Figure 5.5, which shows the 2D visualization of order parameter in nanodisc systems with $A_{2\alpha}R$, the same three regions can be seen within the lipids as in the nanodisc systems without membrane proteins in Figure 5.4. Additionally, the local disordering effect of the added membrane protein is visible, as described previously in the case of the planar systems with $A_{2\alpha}R$.

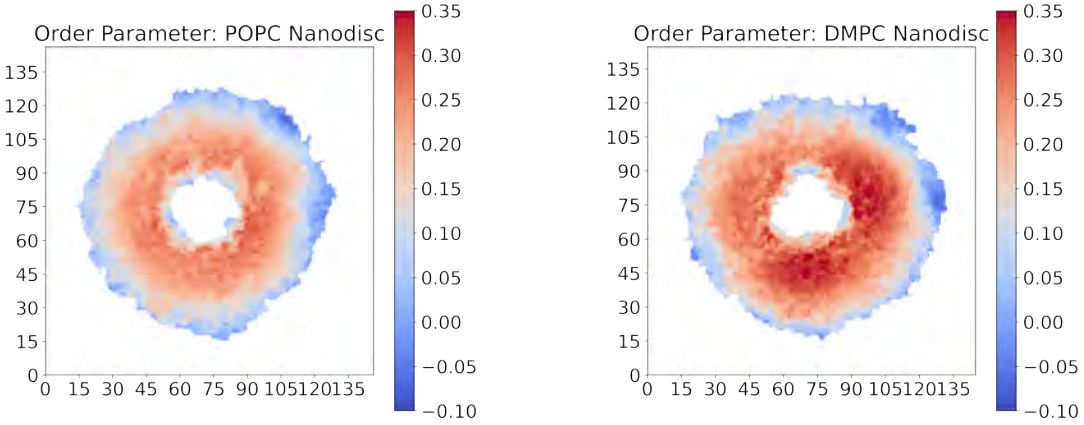
However, considering the radial distribution of order parameters in nanodiscs in Figure 5.6, unlike in the planar systems, the addition of the $A_{2\alpha}R$ does not raise the value of the order parameter higher than in the corresponding nanodisc with the same lipid type, but without the membrane protein. This is likely explained by the small size of the nanodisc.

To conclude, the results confirm the previous findings of the three region distribution of lipid order within the nanodisc systems [8] that is not present in planar bilayers. In addition, the radial visualization of order parameters in the systems provides a more systematic way of quantifying the change of order parameter values in nanodiscs. If no transmembrane protein is not present in the middle of the nanodisc, in the middle of the nanodisc, the order of the membrane is significantly higher than in a planar membrane. In both nanodiscs and planar bilayers the order of the membrane protein disintegrates in the vicinity of $A_{2\alpha}R$ due to the rough interface of the protein. The results also indicate that in a nanodisc environment, the membrane protein never experiences a lipid order similar to that in a planar bilayer.



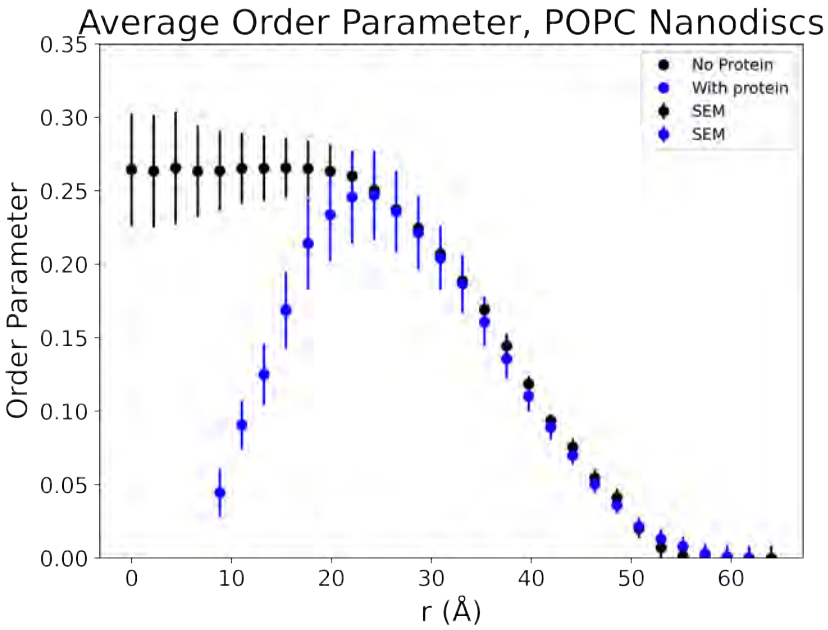
(a) POPC lipids in nanodisc. **(b)** DMPC lipids in nanodisc.

Figure 5.4: 2D distributions of order parameter in nanodisc systems without $A_{2\alpha}R$.

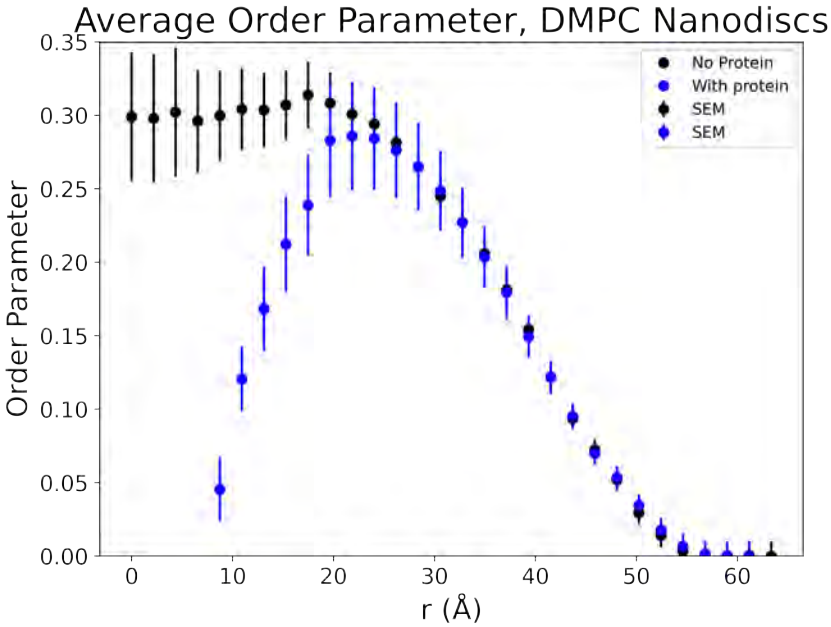


(a) POPC lipids and $A_{2\alpha}R$ in nanodisc. **(b)** DMPC lipids and $A_{2\alpha}R$ in nanodisc.

Figure 5.5: 2D distributions of order parameter in nanodisc systems with $A_{2\alpha}R$.



(a) Nanodisc systems with POPC.



(b) Nanodisc systems with DMPC.

Figure 5.6: Order parameter as a function of radius in nanodisc systems, uncertainty of data points expressed in form of Standard Error of the Mean (SEM).

5.2 Diffusion

Due to the enclosed nature of nanodiscs, the diffusion analysis of lipids in the simulation systems was conducted by calculating the 2D jump lengths Δr of the COM of each lipid in a system over a fixed lagtime of 10 ns. Then a skewed Gaussian of the form presented by equation 3.26 was fitted over the distribution in order to acquire the diffusion coefficient.

The results of the diffusion analysis will be presented in radial plots showing the diffusion coefficient as a function of distance from the center of the system, corresponding to the similar plots in the previous acyl chain order parameter analysis. In addition, the fitting of the skewed Gaussian over the distribution of lipid jump lengths will be showcased for all systems.

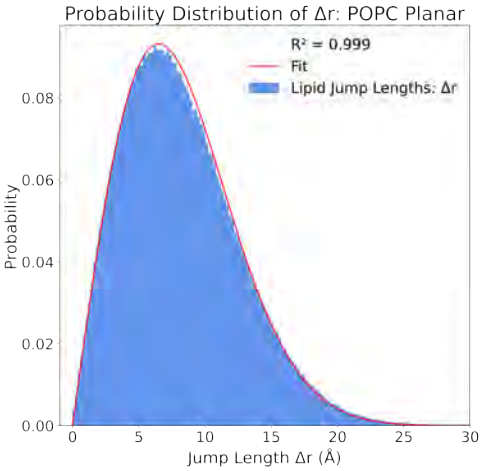
In appendix B, the dependence of the diffusion coefficient from lagtime in the different simulation systems, planars and nanodiscs, is shown in detail. In principle, lagtime means the timeframe within which diffusion is quantified, here in the range of 0.1 to 50 ns. In all systems, excluding the very beginning and end of the graphs, the relationship between the two is linear. The lagtime used in the following analysis is 10 ns.

In the following Figures 5.7, 5.8, 5.9, and 5.10, the fitting of the skewed Gaussian function over the lipid jump length distribution is shown. The fits can be seen to be the best for the planar systems without $A_{2\alpha}R$. The function fit is inferior for the planar system with a transmembrane protein and both nanodisc systems, as also characterized by the R^2 values shown in the Figures that characterize the quality of the fitting. However, as the R^2's indicate, the fits are still good enough to provide reasonable results. The poor fits rise from the fact that the fitting in these figures has been done for all lipids in the systems in which multiple regions of lipids with distinct diffusion coefficients exist: lipids by the protein, and lipids further and far away from the protein in the system.

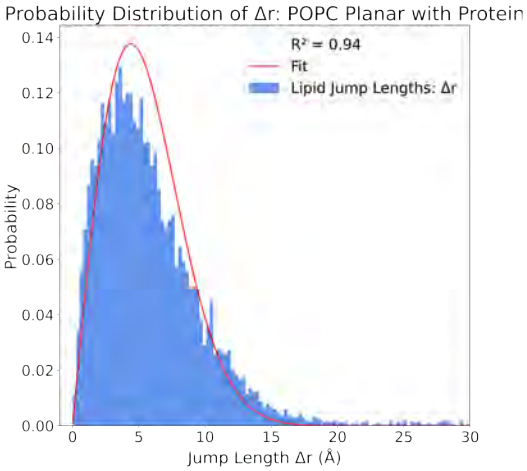
R^2 is the so-called coefficient of determination, and is calculated using equation 5.1:

$$R^2 = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y})^2}, \tag{5.1}$$

where $\sum(y_i - f_i)^2$ is the residual sum of squares and $\sum(y_i - \bar{y})^2$ is the total sum of squares. In principle, the measure gives an interpretation of how well the model explains the observed data by comparing the fitted function to the observed data.

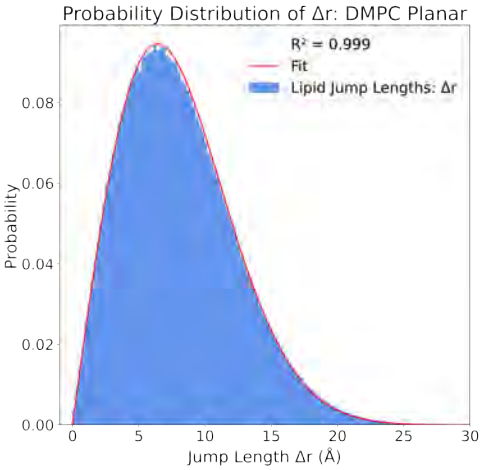


(a) Planar system with POPC and without $A_{2\alpha}R$.

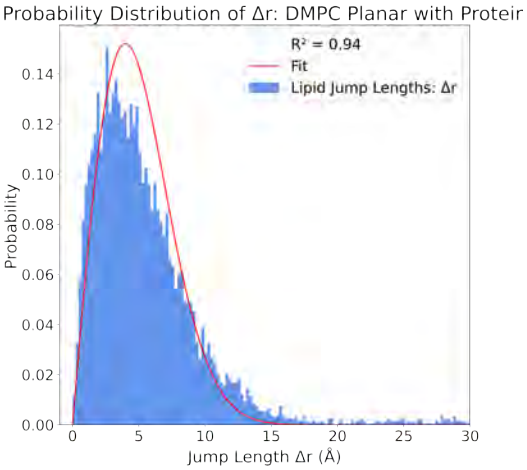


(b) Planar system with POPC and $A_{2\alpha}R$.

Figure 5.7: Fitting of skewed Gaussian distribution over jump lengths of lipids in planar POPC systems.

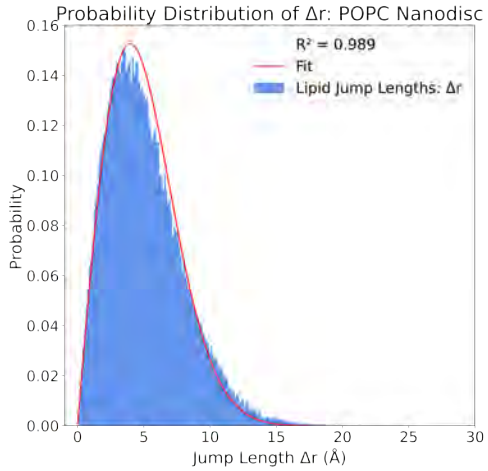


(a) Planar system with DMPC and without $A_{2\alpha}R$.

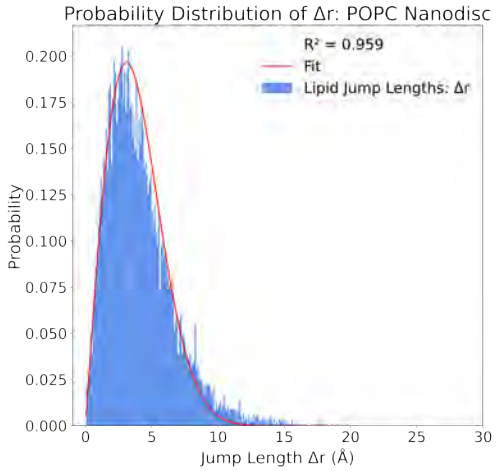


(b) Planar system with DMPC and $A_{2\alpha}R$.

Figure 5.8: Fitting of skewed Gaussian distribution over jump lengths of lipids in planar DMPC systems.

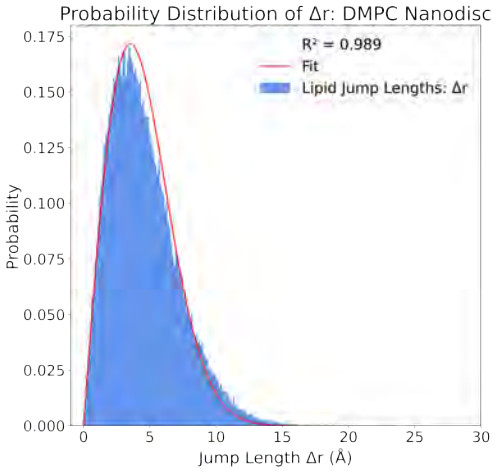


(a) Nanodisc system with POPC and without $A_{2\alpha}R$.

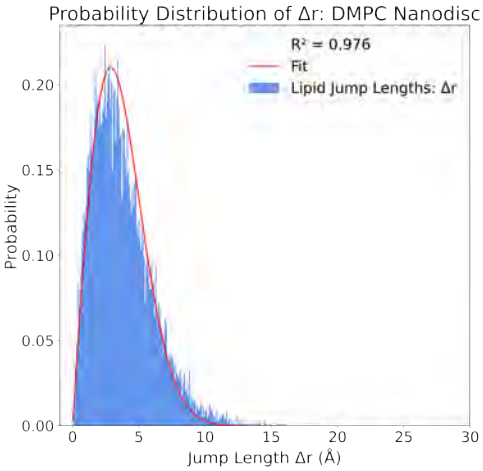


(b) Nanodisc system with POPC and $A_{2\alpha}R$.

Figure 5.9: Fitting of skewed Gaussian distribution over jump lengths of lipids in nanodisc POPC systems.



(a) Nanodisc system with DMPC and without $A_{2\alpha}R$.



(b) Nanodisc system with DMPC and $A_{2\alpha}R$.

Figure 5.10: Fitting of skewed Gaussian distribution over jump lengths of lipids in nanodisc DMPC systems.

From the fits above, an average diffusion coefficient can be acquired for each of the eight systems. These values can be found in Figure 5.11 and Table 5.1. The diffusion coefficient for the POPC and DMPC planar bilayers have been studied before, and the results shown here are well aligned with both previous computational [99] and experimental [122] work.

As indicated by visualizing Figure 5.11, the diffusion coefficient is always slightly higher in systems with POPC lipids compared to systems with DMPC lipids. This is expected due to the earlier results concerning the acyl chain order parameter. The order of the membrane and the diffusivity of the lipids are known to be connected, such that with higher order the diffusion in the system is slowed down. As DMPC lipids were concluded to have a higher order than POPC in the simulated systems, it is not surprising that the diffusion is consequently slightly slower in the DMPC systems. Considering the planar systems with proteins, the addition of the transmembrane protein seems to halve the average diffusion coefficient compared to the systems with lipids only.

On the other hand, the diffusion coefficients in nanodiscs without transmembrane protein are already in the same order of magnitude, as the planar systems that have the added membrane protein. And when moving on to the nanodisc systems with $A_{2\alpha}R$, the diffusion coefficients are reduced even further, but not quite halved as was the case with the planar bilayers. Hence, according to these results, the nanodisc environment can be observed to have a significant effect on the diffusivity of the lipids. The trend of DMPC systems having a lower diffusion coefficient than POPC, is continued in the nanodiscs in similar fashion as in the planar systems.

Type	Lipid	Membrane Protein	Diffusion coefficient (cm^2/s)	Uncertainty (cm^2/s)
Planar	POPC	-	$2.1 \cdot 10^{-7}$	$3 \cdot 10^{-8}$
Planar	POPC	A2AR	$0.98 \cdot 10^{-7}$	$3 \cdot 10^{-8}$
Planar	DMPC	-	$2.0 \cdot 10^{-7}$	$3 \cdot 10^{-8}$
Planar	DMPC	A2AR	$0.81 \cdot 10^{-7}$	$2 \cdot 10^{-8}$
Nanodisc	POPC	-	$0.82 \cdot 10^{-7}$	$4 \cdot 10^{-9}$
Nanodisc	POPC	A2AR	$0.50 \cdot 10^{-7}$	$2 \cdot 10^{-9}$
Nanodisc	DMPC	-	$0.63 \cdot 10^{-7}$	$3 \cdot 10^{-9}$
Nanodisc	DMPC	A2AR	$0.44 \cdot 10^{-7}$	$2 \cdot 10^{-9}$

Table 5.1: Average diffusion coefficients in the simulated systems.

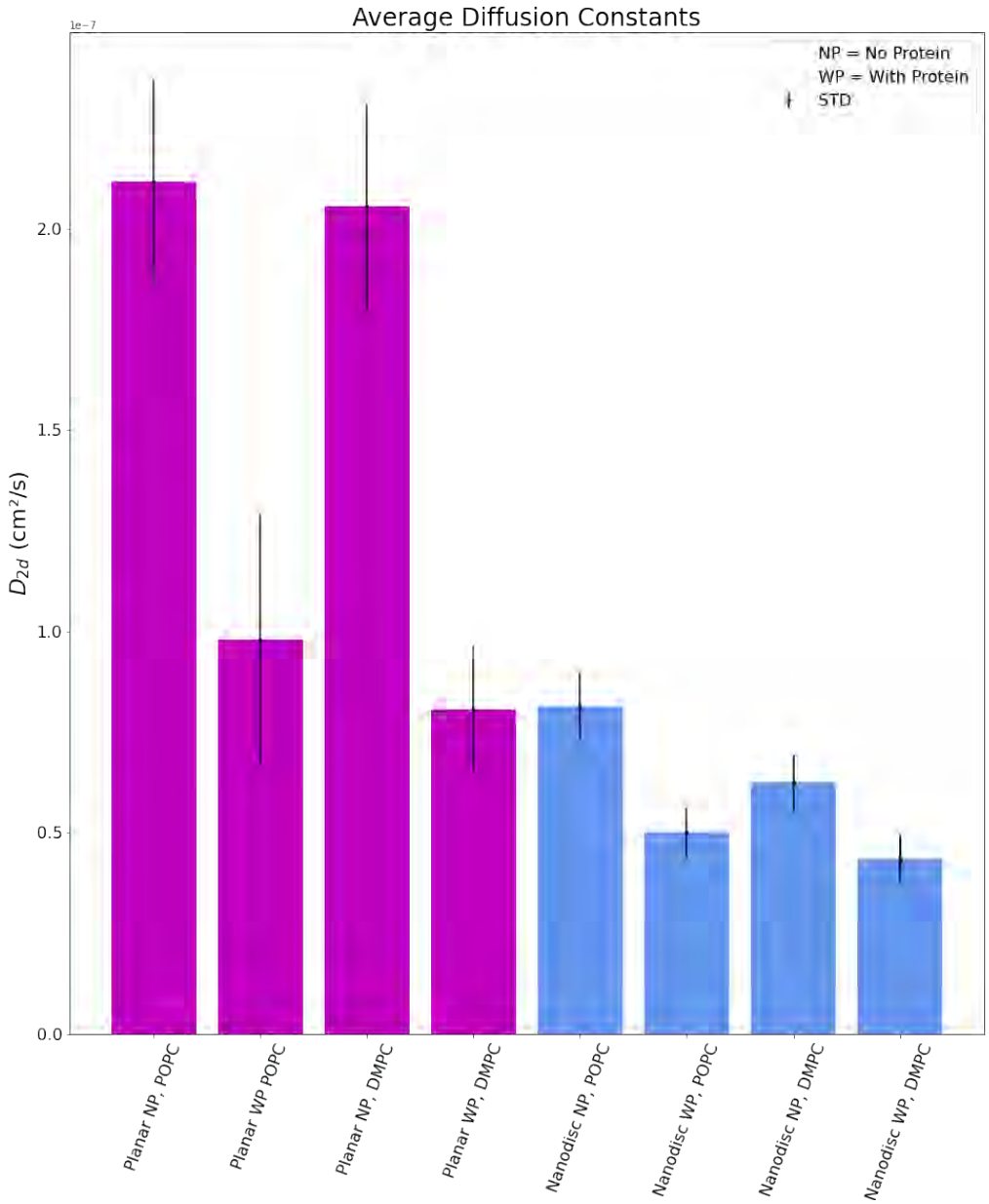
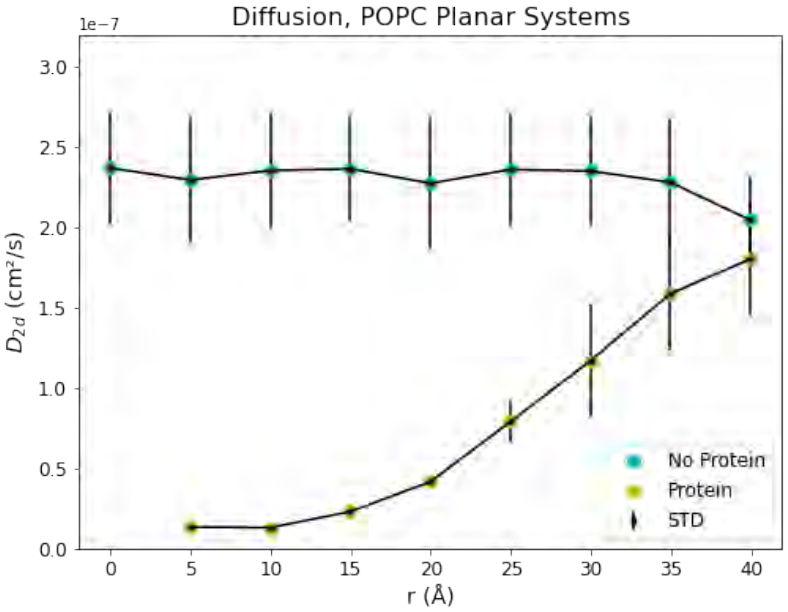


Figure 5.11: Visual representation of Table 5.1, average diffusion coefficients in the simulated systems, uncertainty expressed in form of Standard Deviation (STD).

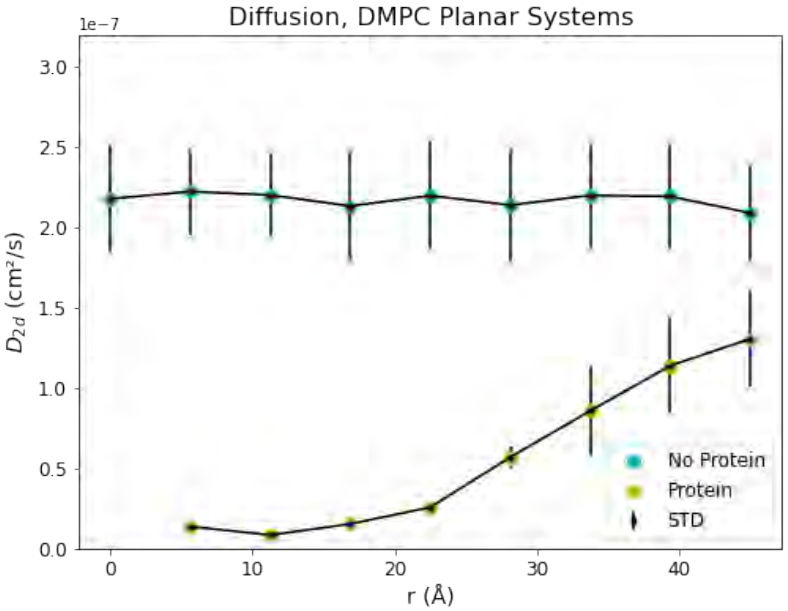
Figure 5.12 displays diffusion coefficients as a function of radial distance from the center of the system towards the edges. The diffusion coefficient of the plain lipid bilayers is not majorly altered throughout the length of the system. In the planar bilayers that have the transmembrane protein, the diffusion starts very low at the center of the system where $A_{2\alpha}R$ is located, and rises towards the edges of the system towards the value of diffusion coefficient in the simple lipid bilayer systems. As discussed earlier, the diffusion is overall slightly slower in the DMPC systems than POPC, and consequently, the diffusion coefficients are lower as well.

In the planar POPC lipid systems, the diffusion coefficients of the two systems (with and without $A_{2\alpha}R$) rise to meet the same order of value in the end, but for the DMPC this does not quite hold true. In the end, the effect of the transmembrane protein on the diffusion of lipids, should be fairly local and dissipate quickly as moving away from the protein. There could be multiple explanations to why this does not show in the DMPC graph. The chosen lagtime could offer an explanation to the behaviour, possibly in a larger timescale, the diffusion at the edges of the two systems might be more comparable.

Another possible explanation could be protein crowding, resulting from the periodic boundary conditions. In studies of the effect before, it was disclosed that the effect of protein crowding provides a significant change in the diffusion of lipids when the lipids-to-protein ratio in simulation systems is 200:1 compared to the situation where the ratio is comparable to infinity:1 [96]. Since the ratio in the planar simulation systems with lipids and transmembrane protein here is in the order of 300:1, there might be a chance that the effect of protein crowding could offer at least a part of the explanation to the effect seen here. Also, it should be noted that the number of lipids in the POPC planar bilayer system with the membrane protein is a bit higher than in its DMPC counterpart. The number of lipids in the planar bilayers was originally chosen to be comparable to that in the nanodisc systems.



(a) Planar systems with POPC.



(b) Planar systems with DMPC.

Figure 5.12: Diffusion coefficient as a function of radius in planar systems, uncertainty expressed in form of Standard Deviation (STD).

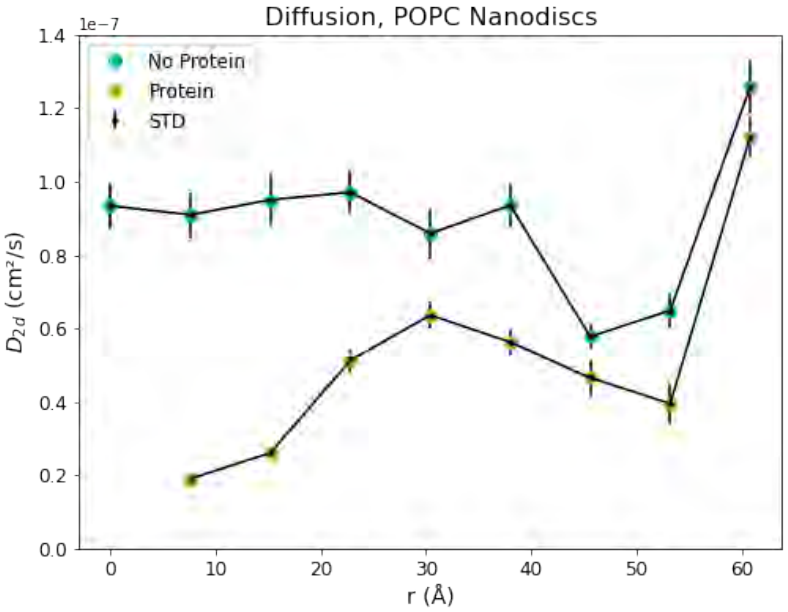
Diffusion coefficients in nanodiscs are more than half lower than those in the planar bilayers. The radial distributions of diffusion coefficients in Figure 5.13, also display a significant drop in the scale of magnitude. Considering the nanodisc systems with no $A_{2\alpha}R$, the graph shows the diffusion to hold a stable value at the center of the system. Moving away from the center of the disc, the value of diffusion coefficient stays stable for almost until the edges of the disc, until the lipids counter the ordering effect imposed by the membrane scaffolding protein shortly before coming in contact with it. The effect is the same as was discussed previously in the case of changes in diffusion coefficient imposed by the transmembrane protein in planar bilayers. This is also the area where the diffusion is at its lowest in nanodiscs. As soon as the lipids come into contact with the scaffolding protein itself, the diffusion speeds up quickly. Knowing that the order parameter and rate of diffusion are related, this is in line with the earlier result of drastic drop in order parameter at the very edges of the nanodisc systems.

The radial treatment of the diffusion coefficient has provided new insights into the diffusion patterns inside a nanodisc. Previously it has been concluded that the slowest diffusion in nanodiscs happens at the very center of the discs [8], but the results presented here indicate that even though the diffusion is highest at the very edges of the nanodiscs, the lowest diffusion rate in fact is not in the center, but closer towards the edges of disc. In the very center of the nanodisc, the diffusion rate is somewhere between the two extreme regions.

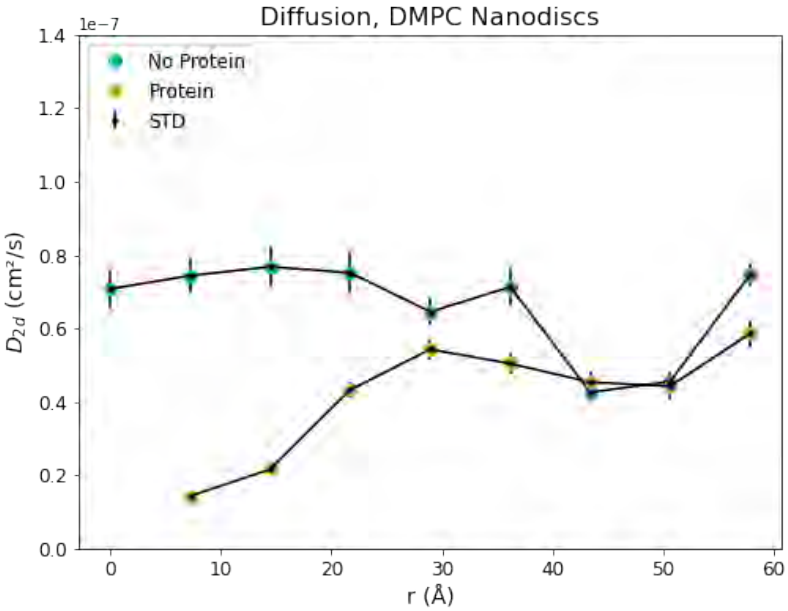
The behaviour of diffusion coefficients as a function of radius in the nanodisc systems with $A_{2\alpha}R$ as seen in Figure 5.13, in the middle of the system follows a similar pattern as in the corresponding planar systems. Close to the transmembrane receptor, diffusion is extremely low and starts slowly rising when moving towards the edges of the system. However, when the vicinity of the membrane scaffolding protein at the edges of the nanodisc is reached, the rate of diffusion starts decreasing again. This happens until the lipids again come in contact with the scaffolding protein and at the edges of the system, the diffusion reaches its highest value in these systems as well.

As in the planar systems, in nanodiscs also the diffusion of POPC lipids is slightly faster than DMPC lipids. Illustrated by Figure 5.13, in nanodiscs with $A_{2\alpha}R$, the diffusion in the center of the systems never reach the diffusion rate that they do in the nanodiscs with the same lipid, but no protein inserted. However, in the DMPC nanodiscs, the lowest rate of diffusion in both types of discs is the same, but in POPC nanodiscs, the lowest value of diffusion in the nanodisc without a transmembrane protein never reaches that of the system with the protein. Also, in the POPC nanodiscs, the diffusion coefficients at the edges of the two types of nanodisc systems are significantly higher than that of the central part of the nanodisc. On the other hand, this does not hold true for the DMPC nanodisc systems in which the rate of diffusion at the edges of the two systems is comparable to

the diffusion in the center of the nanodisc.



(a) Nanodisc systems with POPC.



(b) Nanodisc systems with DMPC.

Figure 5.13: Diffusion coefficient as a function of radius in nanodisc systems, uncertainty expressed in form of Standard Deviation (STD).

5.3 Protein Behaviour

In order to find out whether the behaviour of the transmembrane protein $A_{2\alpha}R$ is different in planar bilayer and nanodisc environments, two machine learning techniques where employed: dimensionality reduction with principal component analysis and clustering with Gaussian mixture models. For the analysis, 500 ns of data from all simulation systems with $A_{2\alpha}R$ was used of which the first 100 ns was considered not equilibrated enough and cut out. From the remaining 400 ns for all the repeats of each system, the CA atom coordinates of $A_{2\alpha}R$ were taken for the analysis.

To begin with, PCA was performed for all the data without labels indicating which dataset it was originally from: planar or nanodisc systems. The result was plotted in the space of the first two principal components for easy interpretability. Figure 5.14 shows the result of this simple analysis after the origin of each data point had been traced back and marked which dataset they originated from. The first two principal components were decided on as a starting point, since they explain the largest amounts of variance in the data in regards to all principal components and due to the possibility to visualize the results in a simple two dimensional graph.

Already Figure 5.14 showcases that the two datasets, one coming from the transmembrane protein coordinates of planar bilayers and the other in nanodiscs, do not have the same distribution. The data points coming from the planar systems cover a larger area compared to the data coming from the nanodisc systems that are concentrated in an inverted C-like shape in the middle of the graph.

This is further highlighted in Figure 5.15 that shows the distributions of the two datasets projected along the first seven principal components of the dataset. In all dimensions, the distributions are only partially overlapping, signifying that along the given dimension, the datasets do not share the same distribution. Only after reaching principal components five and six, the shape of the distribution starts to converge towards a shared form.

The initial clustering using Gaussian mixture models was conducted on the dataset obtained from the simple PCA dimensionality reduction shown in Figure 5.14. Since there was no initial knowledge about underlying clusters in the data, the first step was to choose a clustering model and Figure out how many clusters to form. This was determined by emplying the BIC-criteria as displayed in Figure 5.16. In the Figure, the colored bars stand for four independent types of covariance parameters: 1: "full": where each cluster has its own general covariance matrix, 2: "tied": in which all clusters share the same general covariance matrix, 3: "diag": in which each cluster has its own diagonal covariance matrix, and 4: "spherical": where each cluster has its own single variance.

The BIC-score should be minimized and the number of clusters should be chosen

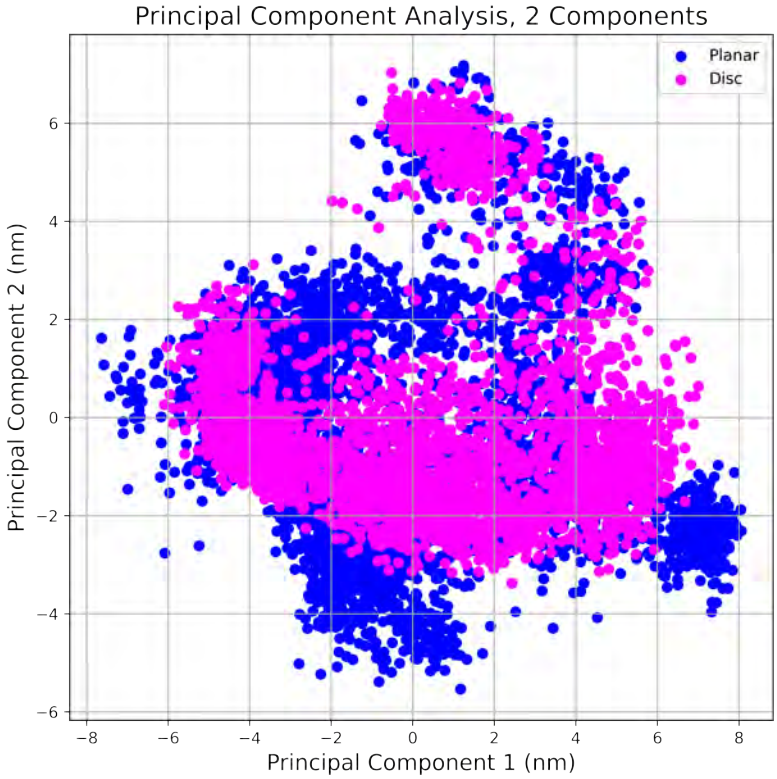


Figure 5.14: Protein coordinates from all simulations projected onto the space of the first two principal components.

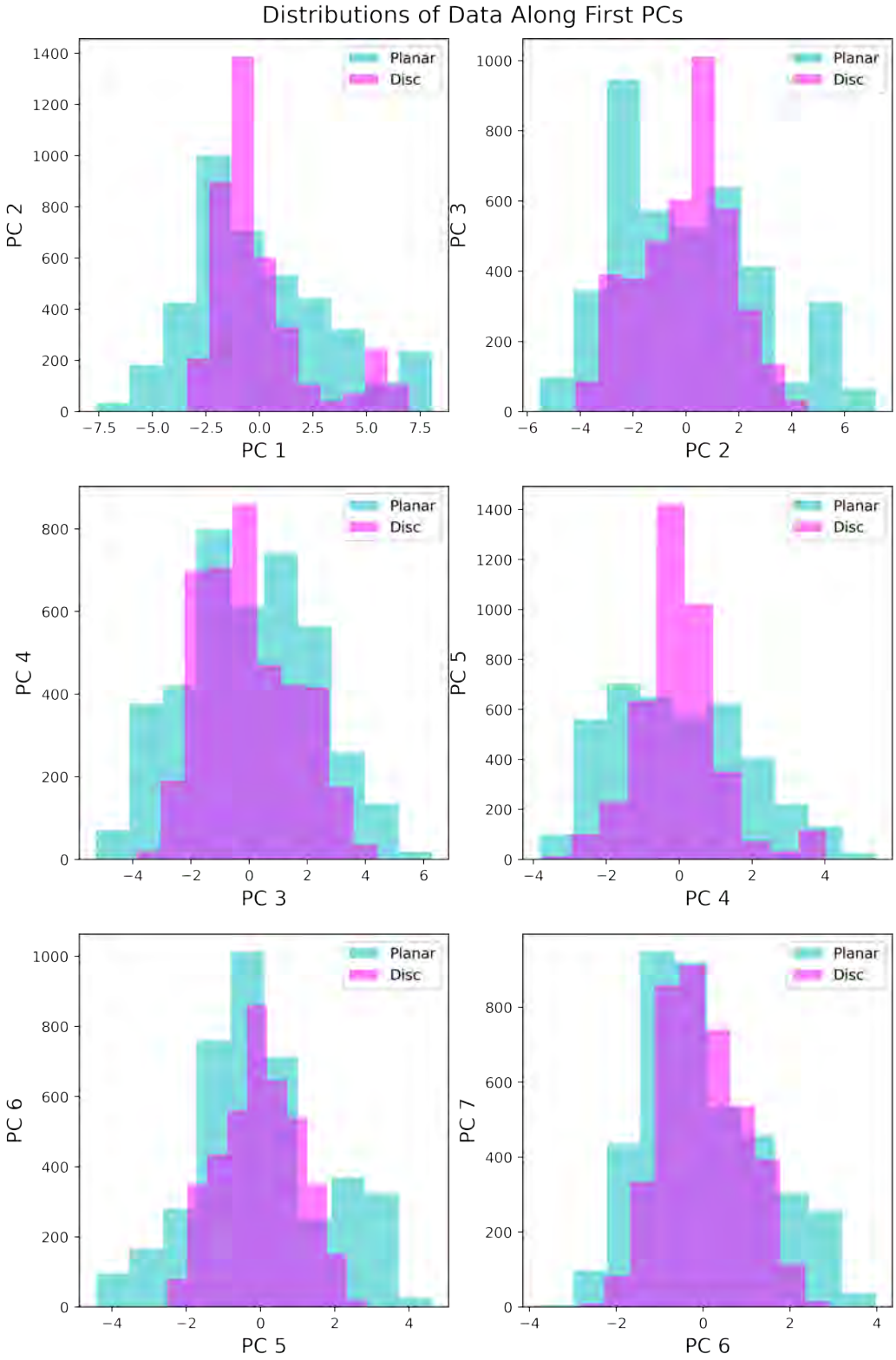


Figure 5.15: Distributions of data coming from planar and nanodisc systems along the first principal components.

to be as low as possible. Based on Figure 5.16, a clustering with a full model and 7 components or clusters was chosen. Even though the BIC-score slightly improves also after this point, the change after the seventh component does not provide enough improvement to compromise the robustness of the clustering by possibly overfitting with a larger number of components. A bigger and smaller number of clusters was also empirically tried, but the seven components clustering provided the best results. In this case over all the components, the full model for variance had always the lowest BIC-score.

With the chosen model, full model with seven components, the clustering was carried out by firstly training the model. This was done by using the unlabeled data from the PCA run presented above. Afterwards, the newly trained model was used to predict clusters on the PCA reduced coordinate dataset. The prediction was done on both planar and nanodisc data separately. Plotting all the data together, the clustered data are seen in Figure 5.17 with the centroids of the clusters marked with black crosses. Of course, this graph does not yet tell much about differences between the two types of systems.

However, when the two types of systems are separated as in Figure 5.18, intriguing differences start to emerge. Cluster five is only visible in the data coming from the planar dataset and cluster 1 has significantly more data points in the nanodisc dataset than planar. Figure 5.19 quantifies the phenomenon and indeed, clusters one and five are only or mostly populated by data from one of the systems. This can be interpreted as a sign that the two datasets are different. Transferring back into the context of MD simulations, this would mean structures that the other type of system does not explore during the simulation trajectory.

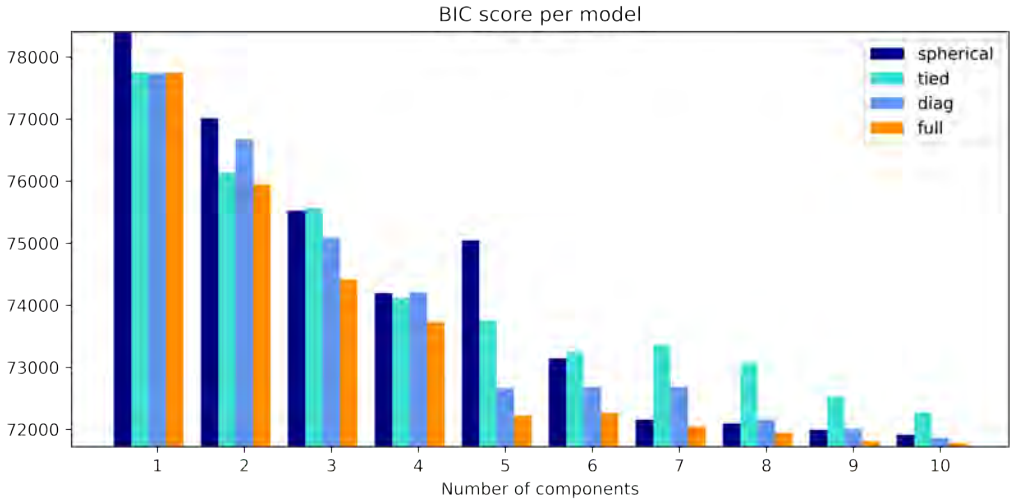


Figure 5.16: Values of BIC-criterion for different Gaussian mixture models.

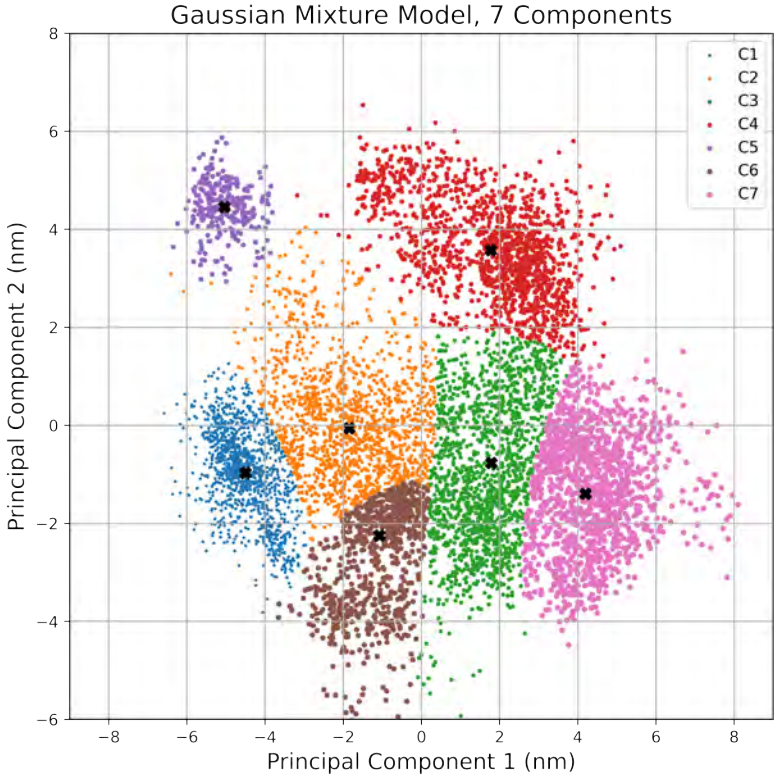
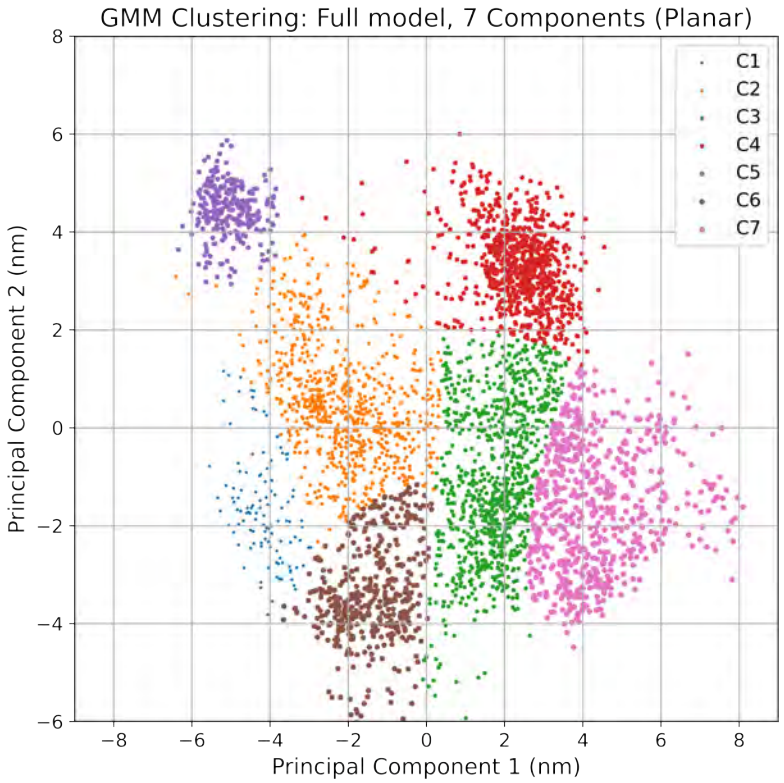
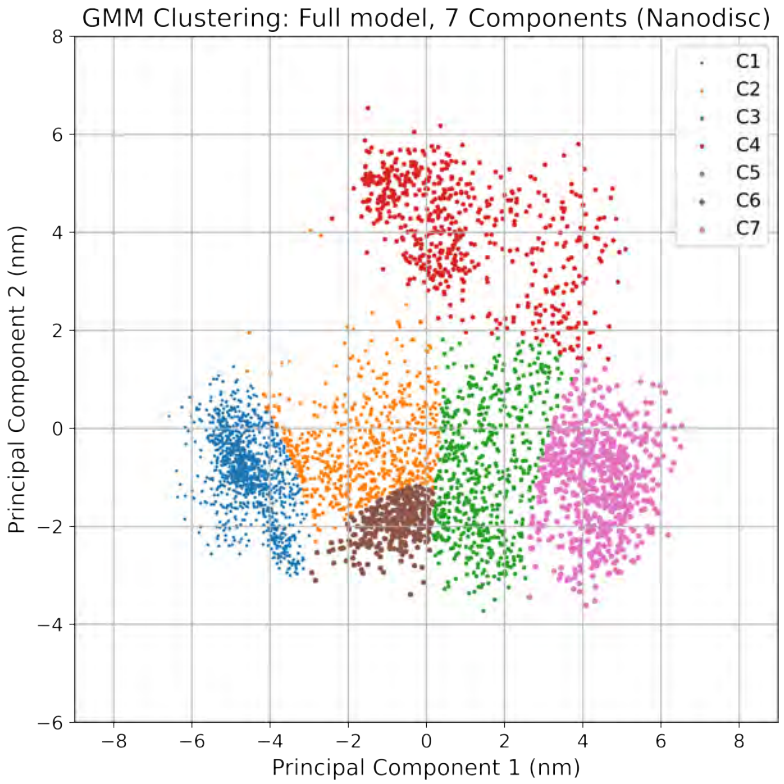


Figure 5.17: Clustering in the dimension of the first two principal components.



(a) Planar systems



(b) Nanodisc

Figure 5.18: Visualization of the origin datasets of data points in the clusters of Figure 5.17.

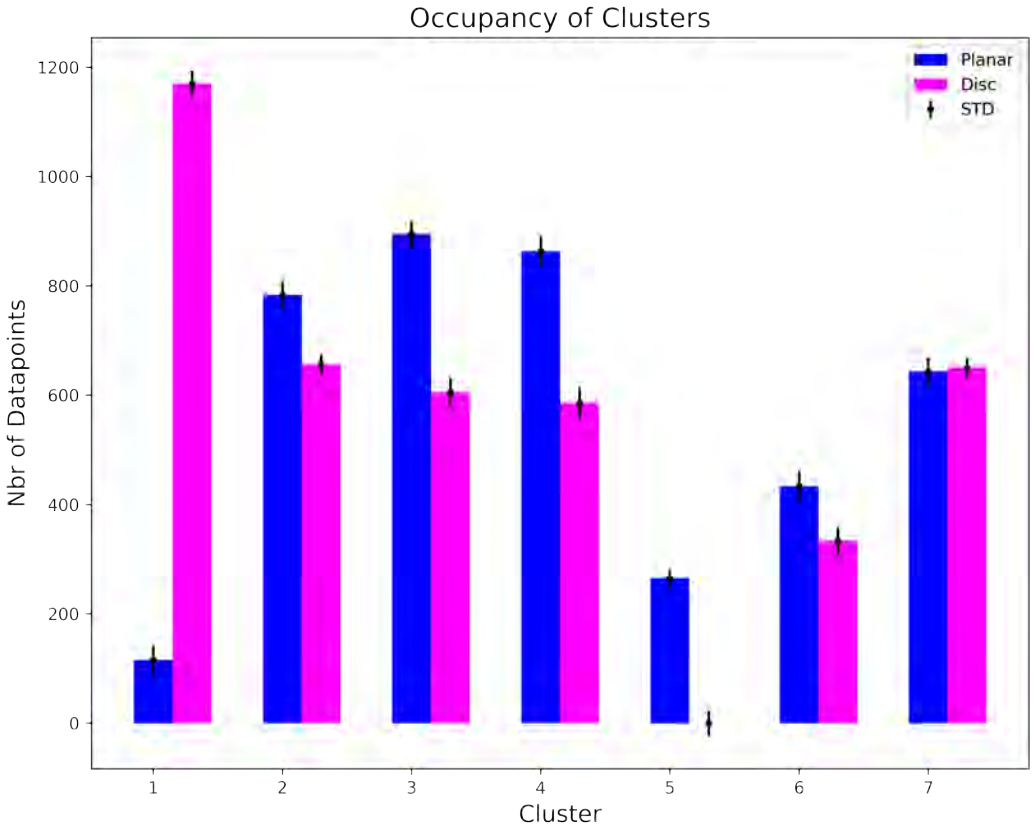


Figure 5.19: Amount of data points in each cluster of Figure 5.18.

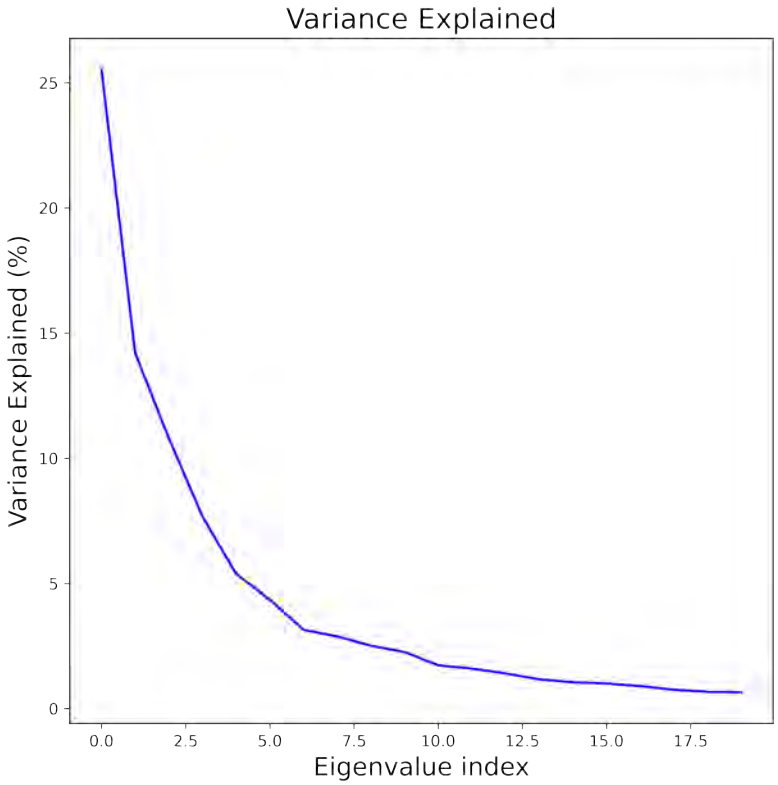


Figure 5.20: Variance explained by the first principal components of the full dataset.

Since already using only the two first principal components of the dimensionality reduction offered a good separation between the two datasets, the next logical step was to go back and look at the next principal components and decide whether they had more to offer. As discussed already above in the case of Figure 5.15, the distributions of the planar and nanodiscs datasets do not share a common form until the principal components exceed four dimensions. From that, it is safe to deduct that the first two principal components are not enough to capture all the major differences in the two datasets.

Thus, the clustering was conducted again, but this time with higher dimensional data. To decide how many dimensions to take into account, the variance explained by each principal component of the PCA as seen in Figure 5.20 was determined. As the percentage of variance explained by each principal component in the plot starts majorly decreasing after the fourth principal component, and because in Figure 5.15, the largest differences in the two distributions are before the fifth principal component, the clustering was determined to be conducted again with the data from the first four principal components.

This decision was also supported by Figure 5.21, which shows a pairplot of the

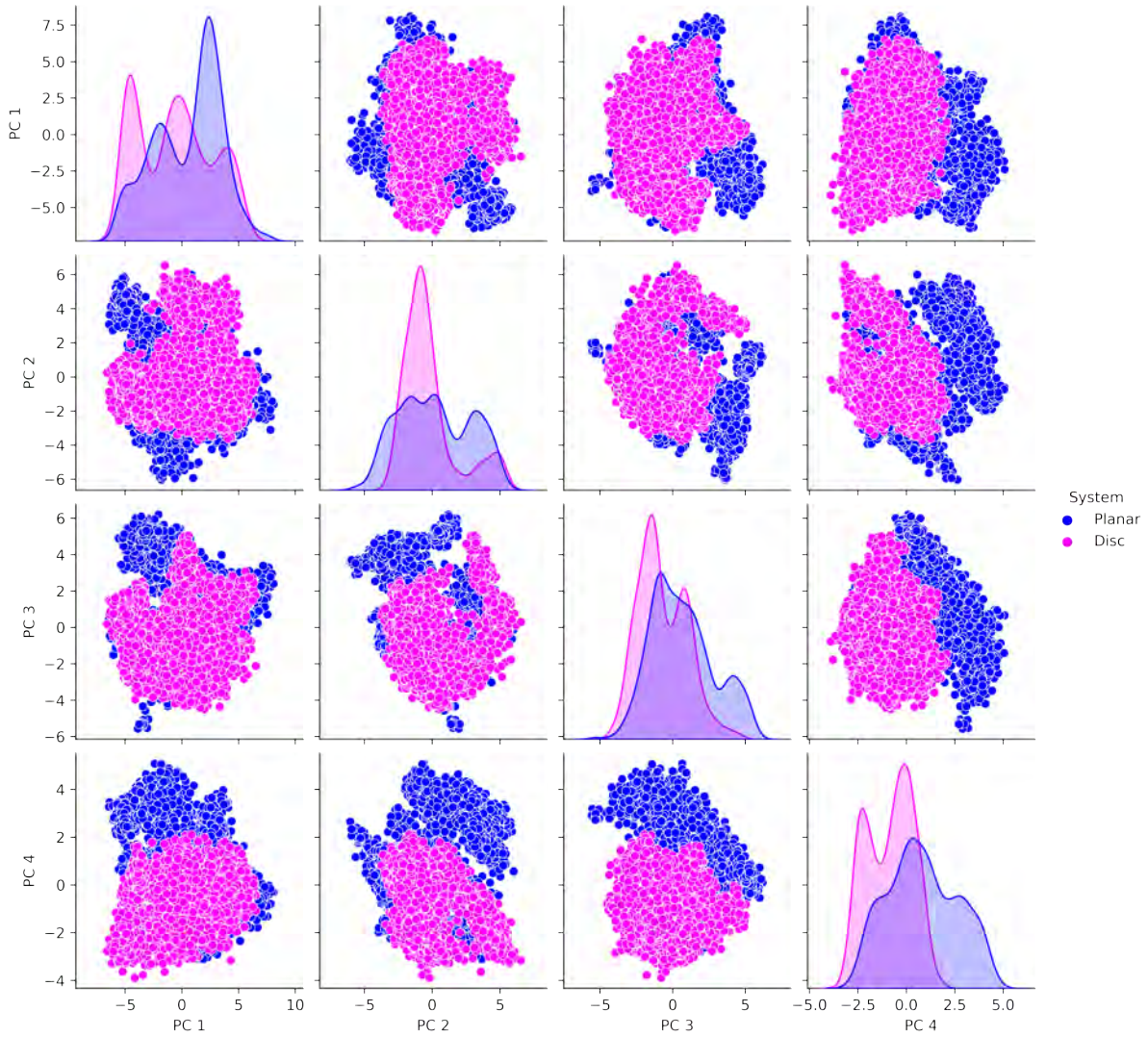


Figure 5.21: Pairplot of principal component analysis in the dimensions of the first four PCs, showing the separation and distributions of the two datasets (planar and nanodisc) in each dimension.

labeled data and their distributions from both planar and nanodisc datasets along the first four dimensions. This is due to the fact that all the frames of the plot with the fourth principal component show the best separation between the two datasets.

The clustering model for the second round of clustering was the same, full model with seven components. The training of the model was again done with all the data and the clusters were predicted separately for the data points coming from planar and nanodisc systems. A pairplot of the predicted clusters along all the four dimensions is shown in Figure 5.22 and Figure 5.23 shows the number of data points in each cluster. In the latter graph, the separation of the clusters can be observed to have been improved compared to the previous clustering model with just the first two principal components. Additionally, from the histograms of the graph, it can be noted that the third and fourth

PCs bring a major number of data points into the clusters that are not as populated in the first and second dimensions.

In the improved clustering model with four PCs, two clusters with mainly data points from planar systems (clusters 1 and 5) can be found and a third cluster with mostly data points from the nanodisc systems' coordinates (cluster 3). This serves as a nice improvement in the robustness of the clustering model compared to the previous one, which only provided one cluster which was populated by data from the planar systems and another cluster which was mainly populated by data from nanodisc systems. This is a strong indication that the two systems do not share the same underlying distribution. Translating back to the original simulations from which the data points originated from, this would correspond to the situation that a transmembrane protein in a nanodisc environment would behave differently from a protein embedded in a simple planar bilayer.

The next incentive was to find the structures of $A_{2\alpha}R$ that correspond to the centroids of the clusters that have the biggest differences in populations (clusters 1, 3, and 5 in Figure 5.23), visualize them and see whether they differ from each other and whether the differences can be linked to the activation of $A_{2\alpha}R$. The representative structures are visualized in Figures 5.24 (cluster 1), 5.25 (cluster 3), and 5.26 (cluster 5). Clusters 1 and 5 are both representative structures from clusters that are mainly populated by data points from planar bilayers and cluster 3 represents a structure from a cluster with mainly nanodisc data.

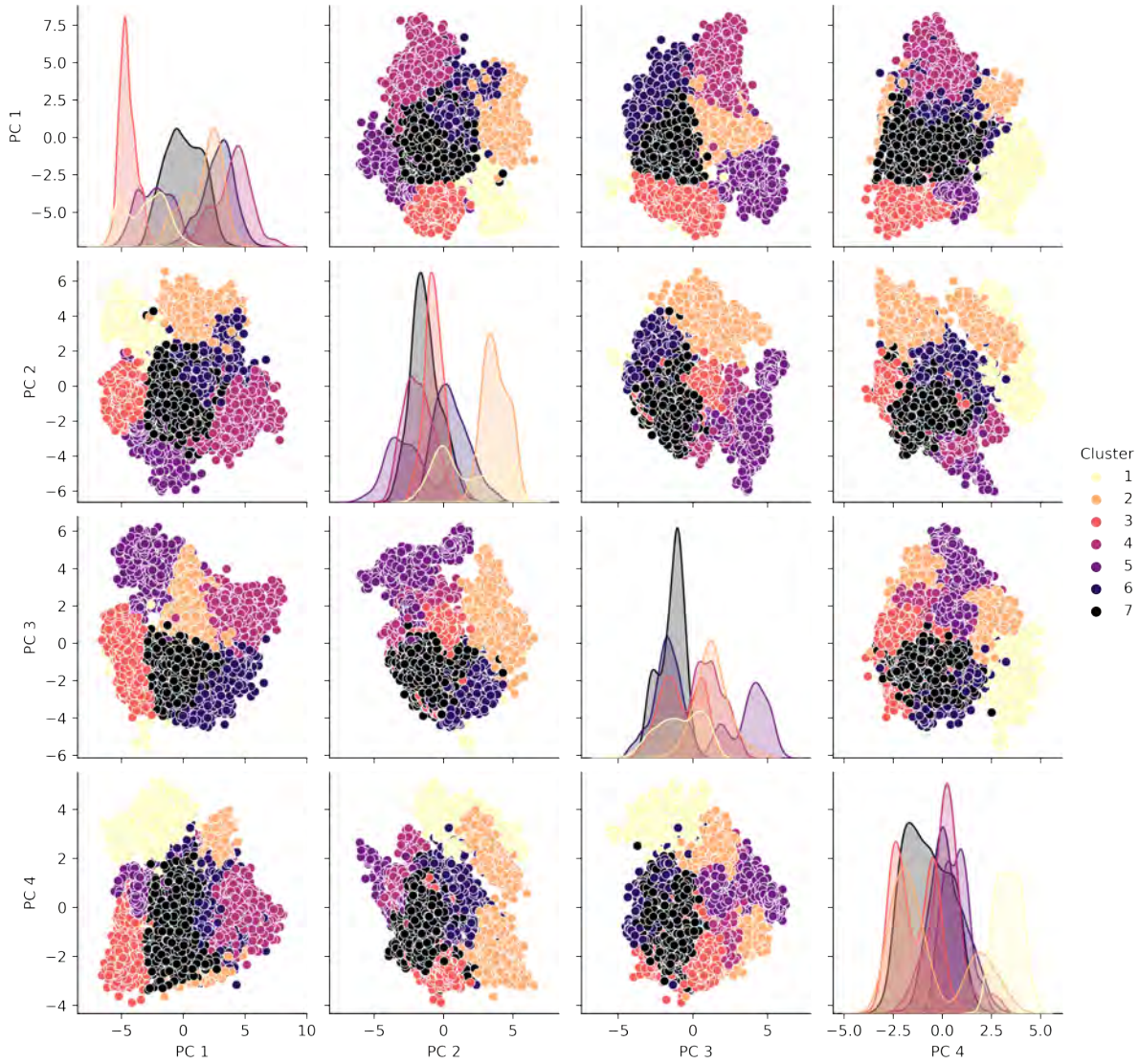


Figure 5.22: Pairplots of clusters in the dimension of the first four PCs, showing the amount of data points each PC dimension contributes to each cluster.

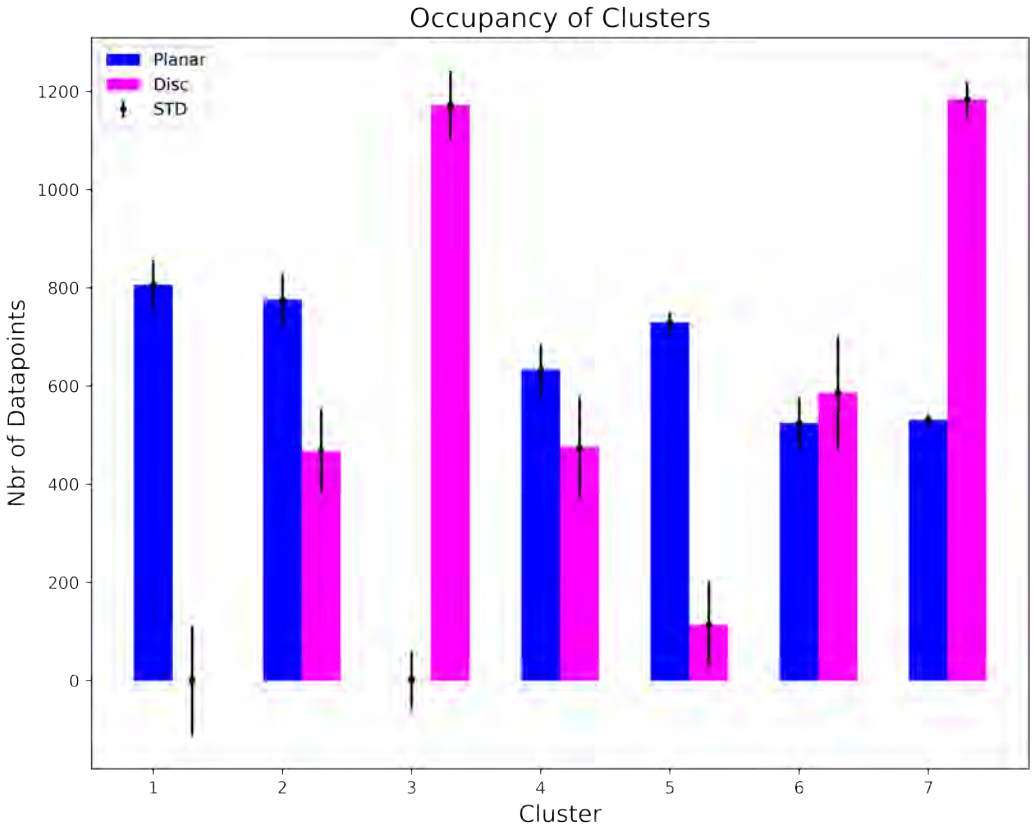
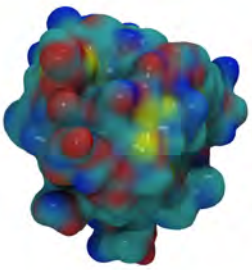
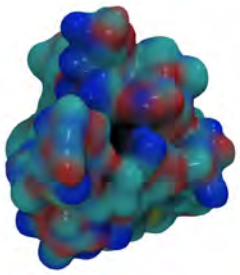


Figure 5.23: Amount of data points in each cluster of Figure 5.22.



(a) Cluster 1, ligand binding site: closed.



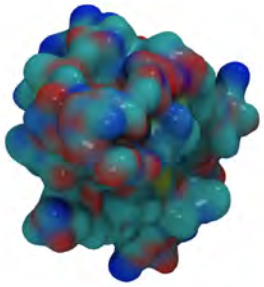
(b) Cluster 1, G protein binding site: partially open.

Figure 5.24: Representative structure of cluster 1 (planar bilayer).

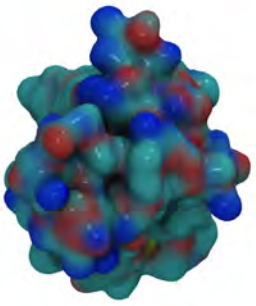
In cluster one, the ligand binding site is seen to be in a closed state and the G protein binding site in a relatively wide, partially open state. In cluster five, the other planar cluster, the ligand binding site is partially closed and the G protein binding site is completely open and a view to the channel inside the receptor is revealed. In cluster three, the nanodisc cluster, both the ligand binding and the G protein binding site are in a tight, almost closed position.

As a simple model of activation, with an open ligand binding site, $A_{2\alpha}R$ can be considered potentially ready to bind a ligand to the open site. On the other hand, with an open G protein binding site, the $A_{2\alpha}R$ can be assumed to be ready to bind a G protein. As in both of the planar data populated clusters, one and five, the G protein binding site is seen to be in an open state, the structures in these clusters could be interpreted to possibly be available to activate through binding of a G protein. However, in the nanodisc cluster number three, neither of the binding sites are open, but are neither tightly closed. In this state, the binding of a ligand or a G protein should not be possible, and the receptor seems to exhibit a kind of intermediate state where it would not be susceptible for activation through either binding site.

In conclusion, the representative structures of planar clusters exhibit a state in which the receptor is potentially able to activate through binding of a G proteins, and that of the nanodisc cluster a state in which the receptor would be incapable of activating neither through ligand binding nor G protein binding. Since all the three clusters possess very few data points from the opposing dataset, it can be concluded that the $A_{2\alpha}R$ in a

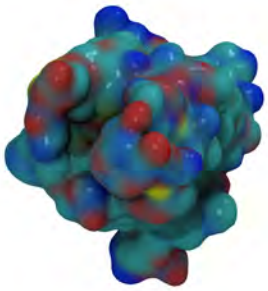


(a) Cluster 3, ligand binding site: closed.

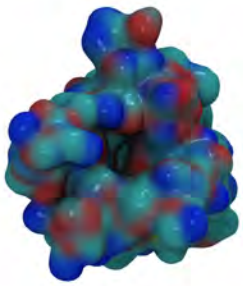


(b) Cluster 3, G protein binding site: partially closed.

Figure 5.25: Representative structure of cluster 3 (nanodisc).



(a) Cluster 5, ligand binding site: partially closed.



(b) Cluster 5, G protein binding site: open.

Figure 5.26: Representative structure of cluster 5 (planar bilayer).

planar bilayer might be more likely to be activated than $A_{2\alpha}R$ in a nanodisc environment. In a nanodisc, the lipid environment seems to push the receptor to exhibit a possible intermediate state in which neither of its binding sites might not available. With the evidence presented, it can be concluded that the nanodisc environment in fact affects the behaviour of adenosine receptor $A_{2\alpha}R$ by not letting the protein explore all its possible open conformations.

6. Conclusions

Nanodiscs are synthetic membranes that are widely used in structural studies of transmembrane proteins. In addition to the lipid dynamics in nanodiscs, the protein's reaction to its environment is poorly understood. In this thesis, a comprehensive study of two lipid characteristics: order parameter and diffusion, has been conducted. Differences in lipid behaviour between lipid only systems and systems with transmembrane proteins were investigated for planar bilayers and nanodiscs.

The results presented in the thesis shed new light on the behaviour of both membrane lipids and the embdedded protein, adenosine receptor $A_{2\alpha}R$. Three main conclusions can be drawn from the results:

First, the nanodisc environment affects the ordering of lipids. In nanodiscs, the results concerning order parameter confirm results reported in the literature before [8]. Three regions of order can be found in nanodiscs: a central ordered region, a disordered region on the edges of the disc, and an intermediate region between the two. In both the planar and nanodisk bilayers it is shown that the lipid order substantially decreases close to the protein. The radial distribution function of the order parameter shows that this effect dissipates rapidly as distance from the protein surface increases.

In addition, the results detail the effect of a transmembrane protein on the ordering. The minor disruptive effect of the transmembrane protein is also visible in nanodiscs. However, presumably due to the nanodisc system's limited space, the transmembrane protein's ordering effect is not clearly visible in nanodiscs. This can be seen best from the graphs showing the radial behaviour of the order parameter.

Second, the nanodisc environment also affects the diffusion of lipids. Although evidence of the affected lipid diffusion in nanodiscs was already found in literature before this study, the new results presented in this thesis are based on a more robust method of calculating diffusion coefficients. According to the results, the average diffusion in planar bilayers is roughly halved with the addition of the protein. The average diffusion coefficient in nanodiscs without transmembrane protein is already reduced to about half of that in a plain planar bilayer and reduces even further with the addition of the membrane protein. Investigation of the radial behaviour of lipid diffusion in nanodiscs yields three distinct regions of diffusion: a region of fast diffusion in the middle of the nanodisc, a

region of fast diffusion in the vicinity of the concave surface of the membrane scaffolding protein, and a region of slow diffusion between the two. This latter regime of slow diffusion has not been reported in literature earlier. In both planar and the nanodisk systems the addition of the membrane protein produces a region of slow diffusion in the vicinity of the protein. The effect dissipates as the distance from the protein surface increases. These results provide new insights into the distributions of diffusion behaviours in nanodiscs, as the radial behaviour of diffusion in nanodiscs has not been reported before. Previously in literature, the diffusion regions in nanodiscs have only been characterized in concert with the order parameter, slow in the ordered area in the middle of the disc and fast by the scaffolding protein [8]. However, the results presented in this thesis also reveal a third previously uncharacterized area between the two, where the slowest diffusion in a nanodisc happens.

Third, the behaviour of the adenosine receptor $A_{2\alpha}R$ is different in a nanodisc environment than in a planar bilayer. The protocol of dimensionality reduction with principal component analysis and clustering with Gaussian mixture models has provided the striking result that the two datasets, one originating from the structures of $A_{2\alpha}R$ in planar systems and the other in nanodiscs differ substantially from each other. Moreover, the cluster analysis and visualization of the representative structures have revealed that the difference in $A_{2\alpha}R$ behaviour can be linked to the activation mechanism of the receptor. This manifests itself in a manner that the receptor in a nanodisc environment is much more likely to reside in a configuration intermediate to the G protein receptive or the ligand receptive state, in which neither binding site is not available. In comparison, the receptor in a planar bilayer is more likely to be found in a state that exhibits potential for activation, where its G protein binding site is open.

As the results concerning the activation of the $A_{2\alpha}R$ are qualitative primarily at this point, much work remains to be done to investigate the matter further. By incorporating a potential quantitative measure of $A_{2\alpha}R$ activation into the protocol, a more conclusive answer could be derived for the question how the nanodisc environment affects the transmembrane receptor. Also, the respective times that $A_{2\alpha}R$ spends in each of the identified conformations require further investigations. In addition to that, the analysis done for $A_{2\alpha}R$ in this thesis could also in future be repeated with other GPCRs, as well as other membrane proteins to generalize the result.

The investigation of lipid characteristics could also be expanded further in order to understand further how the lipid environment in nanodiscs changes the behaviour of the transmembrane protein. For example, an analysis on the pressure profile in nanodiscs could be useful in order to see whether the protein is exposed to a higher compression in a nanodisc which could force it towards a closed conformation. In addition, a study of the nanodisc radius on the lipid characteristics could provide interesting insights. Since

the results presented here and previously indicate an area of stable order and diffusion in the middle of the nanodisc, how large should a nanodisc diameter be to appear to a transmembrane protein as a planar bilayer? Furthermore, taking into consideration the ultimately limited size of nanodiscs before they become too big and collapse into spheres, it might not even be possible to reach the lipid conditions where nanodisks can provide a local environment similar to that of a planar bilayer. Some studies incorporating multiple sizes of nanodiscs already exist, such as reports of lower ordering of lipids with a larger disc diameter [8]. This could prove to be an interesting path to investigate further.

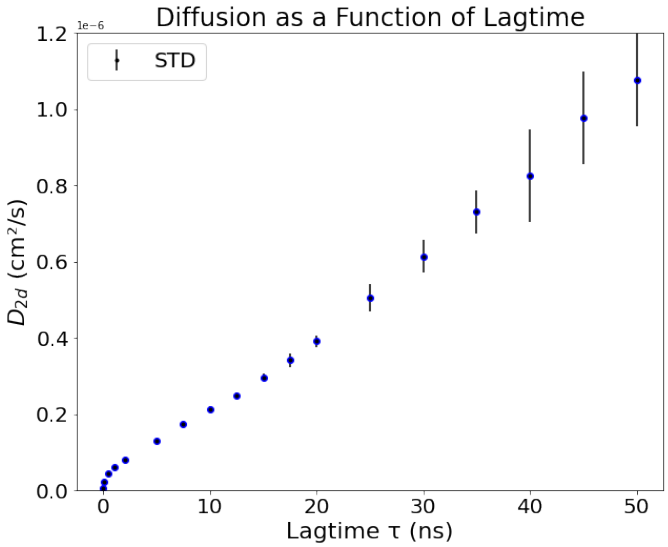
In conclusion, the results presented in this thesis provide novel insights into the lipid and transmembrane protein behaviour in a nanodisc environment compared to a planar bilayer. The question posed in the introduction is also answered: Does a nanodisk provide an environment identical to that of a planar bilayer? The results demonstrate that neither the lipid environment nor the protein behavior is comparable in the two systems.

Appendix A. Labels and Amino Acid Sequences of Membrane Scaffolding Proteins[1]

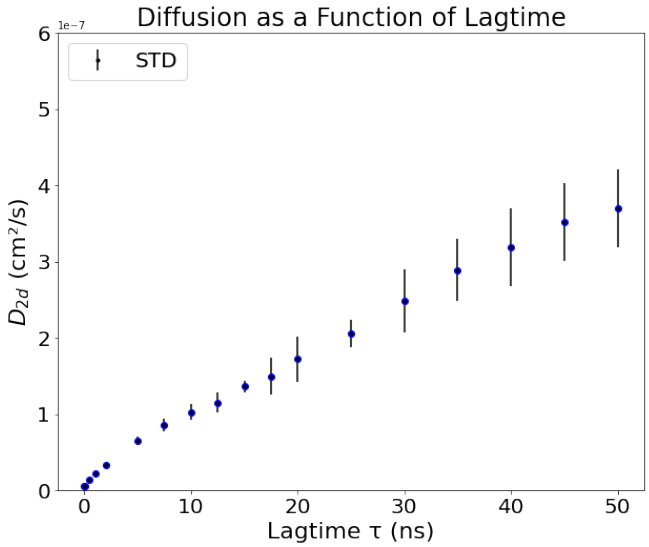
Abbreviation	Description	Amino Acid Sequence
H1	Helix 1	LKLLDNWDSVTSTFSKLREQLG
H1Δ(1-11)	Truncated Helix 1	STFSKLREQLG
H1Δ(1-17)	Truncated Helix 1	REQLG
H2	Helix 2	PVTQEFWDNLEKETEGLRQEMS
H3	Helix 3	KDLEEVKAKVQ
H4	Helix 4	PYLDDFQKKWQEEMELYRQKVE
H5	Helix 5	PLRAELQEGARQKLHELQEKLS
H6	Helix 6	PLGEEMRDRARAHVDALRTHLA
H7	Helix 7	PYSDELRQRLAARLEALKENGG
H8	Helix 8	ARLAEYHAKATEHLSTLSEKAK
H9	Helix 9	PALEDLRQGLL
H10	Helix 10	PVLESFKVSFLSALEEYTKKLNTQ
FX	Original N-terminus	MGHHHHHHIEGR
TEV	Modified N-terminus	MGHHHHHHHDYDIPTTENLYFQG

Abbreviated Name	Composition
MSP1	FX-H1-H2-H3-H4-H5-H6-H7-H8-H9-H10
MSP2	FX-H1-H2-H3-H4-H5-H6-H7-H8-H9-H10- GT-H1-H2-H3-H4-H5-H6-H7-H8-H9-H10
MSP1E1	FX-H1-H2-H3-H4-H4-H5-H6-H7-H8-H9-H10
MSP1E2	FX-H1-H2-H3-H4-H5-H4-H5-H6-H7-H8-H9-H10
MSP1E3	FX-H1-H2-H3-H4-H5-H6-H4-H5-H6-H7-H8-H9-H10
MSP1D1	TEV-H1Δ(1-11)-H2-H3-H4-H5-H6-H7-H8-H9-H10
MSP1D2	TEV-H2-H3-H4-H5-H6-H7-H8-H9-H10
MSP2N1	TEV-H1Δ(1-11)-H2-H3-H4-H5-H6-H7-H8-H9-H10- GT-H1Δ(1-11)-H2-H3-H4-H5-H6-H7-H8-H9-H10
MSP2N2	TEV-H1Δ(1-11)-H2-H3-H4-H5-H6-H7-H8-H9-H10- GT-H2-H3-H4-H5-H6-H7-H8-H9-H10
MSP2N3	TEV-H1Δ(1-11)-H2-H3-H4-H5-H6-H7-H8-H9-H10- GT-H1Δ(1-17)-H2-H3-H4-H5-H6-H7-H8-H9-H10

Appendix B. Diffusion Coefficient as a Function of Lagtime

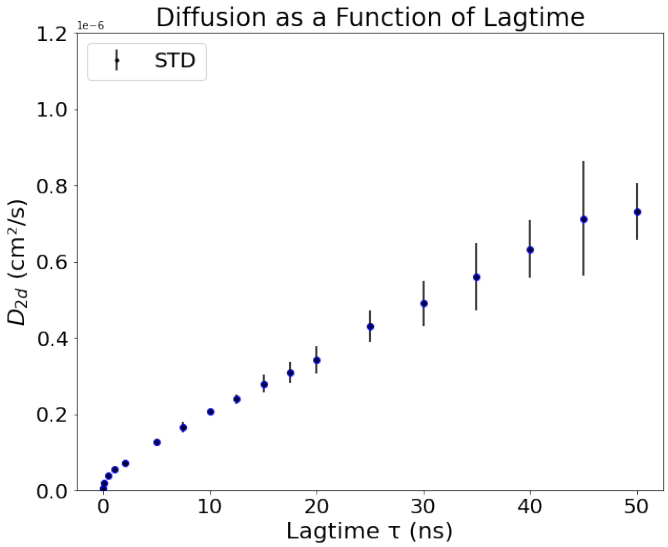


(a) POPC lipids in planar system without $A_{2\alpha}R$

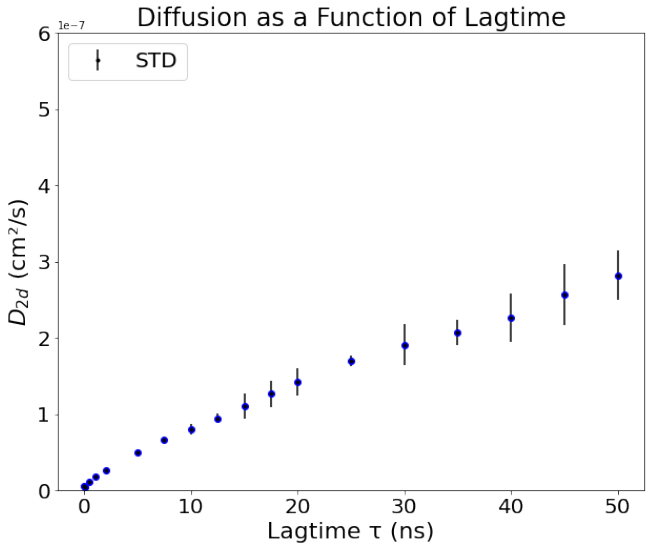


(b) POPC lipids in planar system with $A_{2\alpha}R$

Figure B.1: Diffusion coefficient as a function of lagtime, POPC planar systems.

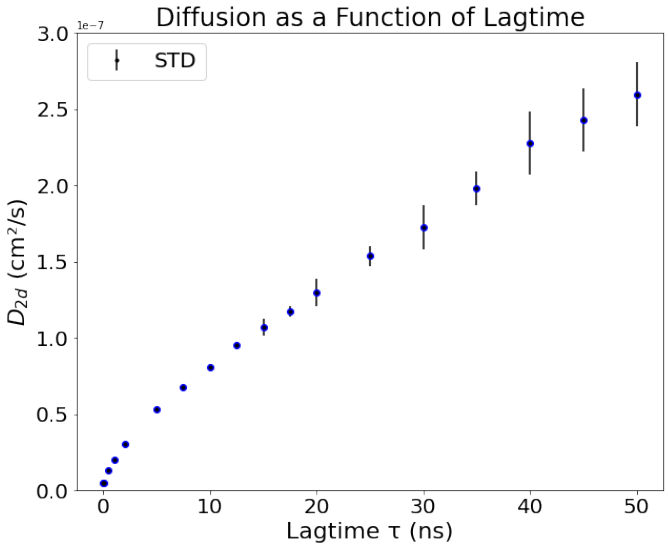


(a) DMPC lipids in planar system without $A_{2\alpha}R$

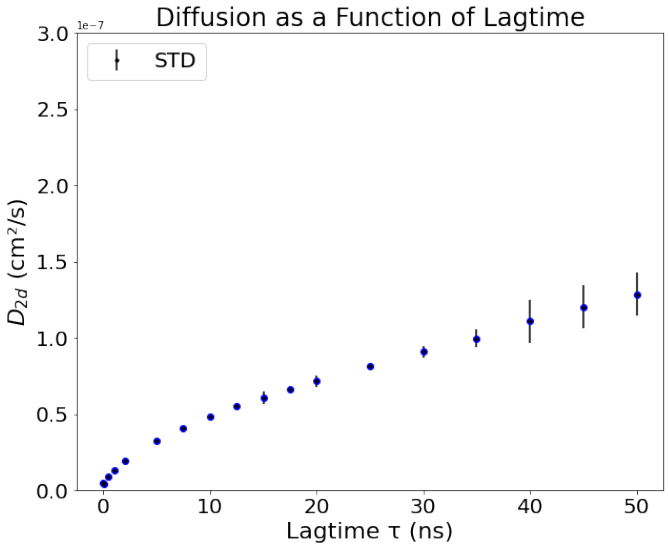


(b) DMPC lipids in planar system with $A_{2\alpha}R$

Figure B.2: Diffusion coefficient as a function of lagtime, DMPC planar systems.

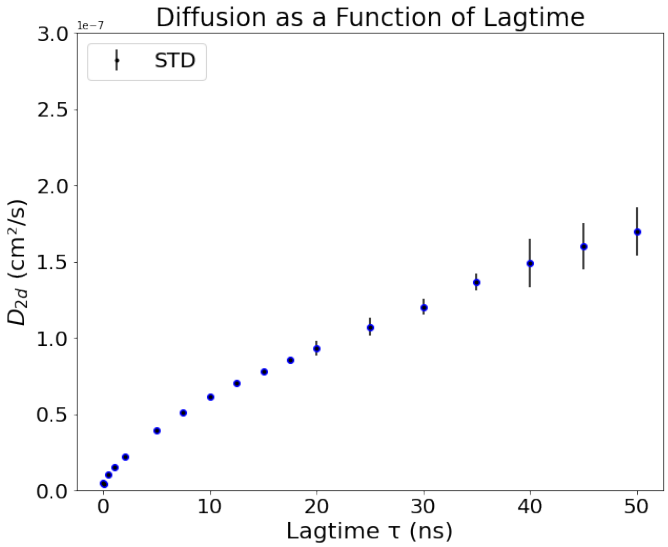


(a) POPC lipids in nanodisc system without $A_{2\alpha}R$

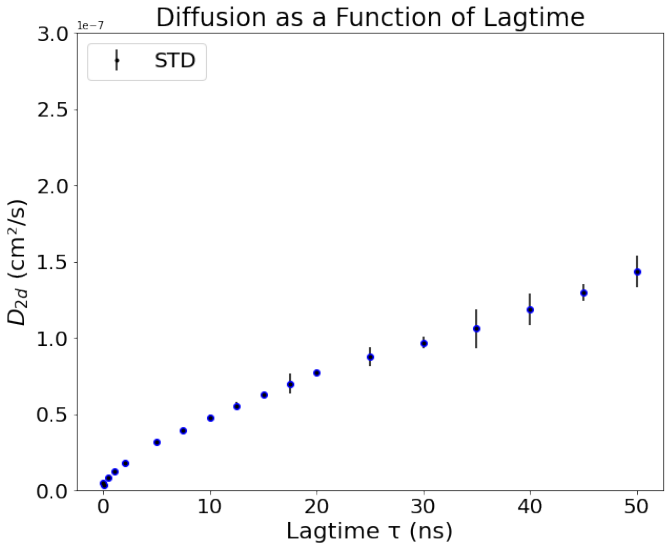


(b) POPC lipids in nanodisc system with $A_{2\alpha}R$

Figure B.3: Diffusion coefficient as a function of lagtime, POPC nanodisc systems.



(a) DMPC lipids in nanodisc system without $A_{2\alpha}R$



(b) DMPC lipids in nanodisc system with $A_{2\alpha}R$

Figure B.4: Diffusion coefficient as a function of lagtime, DMPC nanodisc systems.

Bibliography

[1] Yelena V Grinkova, Ilia G Denisov, and Stephen G Sligar. Engineering extended membrane scaffold proteins for self-assembly of soluble nanoscale lipid bilayers. *Protein Engineering, Design and Selection*, 23:843–848, 2010.

[2] Thomas D Pollard, William C Earnshaw, Jennifer Lippincott-Schwartz, and Graham Johnson. *Cell biology E-book*. Elsevier Health Sciences, 2016.

[3] Andrej Shevchenko and Kai Simons. Lipidomics: coming to grips with lipid diversity. *Nature Reviews Molecular Cell Biology*, 11:593–598, 2010.

[4] Giray Enkavi, Matti Javanainen, Waldemar Kulig, Tomasz Róg, and Ilpo Vattulainen. Multiscale simulations of biological membranes: the challenge to understand biological phenomena in a living substance. *Chemical Reviews*, 119:5607–5774, 2019.

[5] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[6] Elisabeth P Carpenter, Konstantinos Beis, Alexander D Cameron, and So Iwata. Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, 18:581–586, 2008.

[7] Timothy H Bayburt, Yelena V Grinkova, and Stephen G Sligar. Self-assembly of discoidal phospholipid bilayer nanoparticles with membrane scaffold proteins. *Nano Letters*, 2:853–856, 2002.

[8] Piotr Stepien, Bozena Augustyn, Chetan Poojari, Wojciech Galan, Agnieszka Polit, Ilpo Vattulainen, Anna Wisnieska-Becker, and Tomasz Rog. Complexity of seemingly simple lipid nanodiscs. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1862:183420, 2020.

[9] Mohsen Pourmousa and Richard W Pastor. Molecular dynamics simulations of lipid nanodiscs. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1860:2094–2107, 2018.

[10] Karsten Mörs, Christian Roos, Frank Scholz, Josef Wachtveitl, Volker Dötsch, Frank Bernhard, and Clemens Glaubitz. Modified lipid and protein dynamics in nanodiscs. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1828:1222–1229, 2013.

[11] Engelbert Buxbaum. Protein structure. *Fundamentals of protein structure and function*, pages 15–64, 2015.

[12] Donald Voet and Judith G Voet. *Biochemistry*. John Wiley & Sons, 2010.

[13] Patrick J Casey and Miguel C Seabra. Protein prenyltransferases. *Journal of Biological Chemistry*, 271:5289–5292, 1996.

[14] Ilia G Denisov and Stephen G Sligar. Nanodiscs in membrane biochemistry and biophysics. *Chemical Reviews*, 117:4669–4713, 2017.

[15] Rafael Fernandez-Leiro and Sjors HW Scheres. Unravelling biological macromolecules with cryo-electron microscopy. *Nature*, 537:339–346, 2016.

[16] Theresia Gutmann, Kelly H Kim, Michal Grzybek, Thomas Walz, and Ünal Coskun. Visualization of ligand-induced transmembrane signaling in the full-length human insulin receptor. *Journal of Cell Biology*, 217:1643–1649, 2018.

[17] Timothy H Bayburt and Stephen G Sligar. Membrane protein assembly into nanodiscs. *FEBS Letters*, 584:1721–1727, 2010.

[18] Christie G Brouillette, GM Anantharamaiah, Jeffrey A Engler, and David W Borhani. Structural models of human apolipoprotein ai: a critical analysis and review. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1531:4–46, 2001.

[19] TK Ritchie, YV Grinkova, TH Bayburt, IG Denisov, JK Zolnerciks, WM Atkins, and SG Sligar. Reconstitution of membrane proteins in phospholipid bilayer nanodiscs. *Methods in Enzymology*, 464:211–231, 2009.

[20] Ilia G Denisov, Yelena V Grinkova, Anne A Lazarides, and Stephen G Sligar. Directed self-assembly of monodisperse phospholipid bilayer nanodiscs with controlled size. *Journal of the American Chemical Society*, 126:3477–3487, 2004.

[21] Ilia G Denisov and Stephen G Sligar. Nanodiscs for structural and functional studies of membrane proteins. *Nature Structural & Molecular Biology*, 23:481–486, 2016.

[22] Christie G Brouillette, GM Anantharamaiah, Jeffrey A Engler, and David W Borhani. Structural models of human apolipoprotein ai: a critical analysis and review. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1531:4–46, 2001.

[23] James C Phillips, Willy Wriggers, Zhigang Li, Ana Jonas, and Klaus Schulten. Predicting the structure of apolipoprotein ai in reconstituted high-density lipoprotein disks. *Biophysical Journal*, 73:2337–2346, 1997.

[24] Jere P Segrest, Martin K Jones, Anthony E Klon, Christopher J Sheldahl, Matthew Hellinger, Hans De Loof, and Stephen C Harvey. A detailed molecular belt model for apolipoprotein AI in discoidal high density lipoprotein. *Journal of Biological Chemistry*, 274:31755–31758, 1999.

[25] Ying Li, Aleksandra Z Kijac, Stephen G Sligar, and Chad M Rienstra. Structural analysis of nanoscale self-assembled discoidal lipid bilayers by solid-state nmr spectroscopy. *Biophysical journal*, 91:3819–3828, 2006.

[26] Vishwanath Koppaka, Loraine Silvestro, Jeffrey A Engler, Christie G Brouillette, and Paul H Axelsen. The structure of human lipoprotein AI: evidence for the "belt" model. *Journal of Biological Chemistry*, 274:14541–14544, 1999.

[27] Anthony E Klon, Jere P Segrest, and Stephen C Harvey. Molecular dynamics simulations on discoidal HDL particles suggest a mechanism for rotation in the apo AI belt model. *Journal of molecular biology*, 324:703–721, 2002.

[28] Mahmoud L Nasr, Diego Baptista, Mike Strauss, Zhen-Yu J Sun, Simina Grigoriu, Sonja Huser, Andreas Plückthun, Franz Hagn, Thomas Walz, James M Hogle, et al. Covalently circularized nanodiscs for studying membrane proteins and viral entry. *Nature Methods*, 14:49–52, 2017.

[29] Bozena Augustyn, Piotr Stepien, Chetan Poojari, Edouard Mobarak, Agnieszka Polit, Anna Wisniewska-Becker, and Tomasz Rog. Cholesteryl hemisuccinate is not a good replacement for cholesterol in lipid nanodiscs. *The Journal of Physical Chemistry B*, 123:9839–9845, 2019.

[30] Minoru Nakano, Masakazu Fukuda, Takayuki Kudo, Masakazu Miyazaki, Yusuke Wada, Naoya Matsuzaki, Hitoshi Endo, and Tetsurou Handa. Static and dynamic properties of phospholipid bilayer nanodiscs. *Journal of the American Chemical Society*, 131:8308–8312, 2009.

[31] OH Samuli Ollila, Martti Louhivuori, Siewert J Marrink, and Ilpo Vattulainen. Protein shape change has a major effect on the gating energy of a mechanosensitive channel. *Biophysical Journal*, 100:1651–1659, 2011.

[32] Moutusi Manna, Miia Niemelä, Joona Tynkkynen, Matti Javanainen, Waldemar Kulig, Daniel J Müller, Tomasz Rog, and Ilpo Vattulainen. Mechanism of allosteric regulation of B2-adrenergic receptor by cholesterol. *eLife*, 5:e18432, 2016.

[33] Stefan Bibow, Yevhen Polyhach, Cédric Eichmann, Celestine N Chi, Julia Kowal, Stefan Albiez, Robert A McLeod, Henning Stahlberg, Gunnar Jeschke, Peter Güntert, et al. Solution structure of discoidal high-density lipoprotein particles with a shortened apolipoprotein ai. *Nature Structural & Molecular Biology*, 24:187–193, 2017.

[34] Denis Martinez, Marion Decossas, Julia Kowal, Lukas Frey, Henning Stahlberg, Erick J Dufourc, Roland Riek, Birgit Habenstein, Stefan Bibow, and Antoine Loquet. Lipid internal dynamics probed in nanodiscs. *ChemPhysChem*, 18:2651–2657, 2017.

[35] Ilia G Denisov, Mark A McLean, Andrew W Shaw, Yelena V Grinkova, and Stephen G Sligar. Thermotropic phase transition in soluble nanoscale lipid bilayers. *The Journal of Physical Chemistry B*, 109:15580–15588, 2005.

[36] Søren Roi Midtgaard, Martin Cramer Pedersen, Jacob Judas Kain Kirkensgaard, Kasper Kildegaard Sørensen, Kell Mortensen, Knud J Jensen, and Lise Arleth. Self-assembling peptides form nanodiscs that stabilize membrane proteins. *Soft Matter*, 10:738–752, 2014.

[37] Nicholas Skar-Gislinge, Jens Bæk Simonsen, Kell Mortensen, Robert Feidenhans'l, Stephen G Sligar, Birger Lindberg Møller, Thomas Bjørnholm, and Lise Arleth. Elliptical structure of phospholipid bilayer nanodiscs encapsulated by scaffold proteins: casting the roles of the lipids and the protein. *Journal of the American Chemical Society*, 132:13713–13722, 2010.

[38] Nicholas Skar-Gislinge, Søren AR Kynde, Ilia G Denisov, Xin Ye, Ivan Lenov, Stephen G Sligar, and Lise Arleth. Small-angle scattering determination of the shape and localization of human cytochrome P450 embedded in a phospholipid nanodisc environment. *Acta Crystallographica Section D: Biological Crystallography*, 71:2412–2421, 2015.

[39] Naomi R Latorraca, AJ Venkatakrishnan, and Ron O Dror. GPCR dynamics: structures in motion. *Chemical Reviews*, 117:139–155, 2017.

[40] Irina S Moreira. Structural features of the G protein/GPCR interactions. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1840:16–33, 2014.

[41] Alexander S Hauser, Misty M Attwood, Mathias Rask-Andersen, Helgi B Schiöth, and David E Gloriam. Trends in GPCR drug discovery: new agents, targets and indications. *Nature Reviews Drug Discovery*, 16:829–842, 2017.

[42] Vignir Isberg, Stefan Mordalski, Christian Munk, Krzysztof Rataj, Kasper Harpsøe, Alexander S Hauser, Bas Vroling, Andrzej J Bojarski, Gert Vriend, and David E Gloriam. GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Research*, 44:D356–D364, 2016.

[43] Dorothea Haasen, Andreas Schnapp, Martin J. Valler, and Ralf Heilker. G protein-coupled receptor internalization assays in the high-content screening format. In *Measuring Biological Responses with Automated Microscopy*, volume 414 of *Methods in Enzymology*, pages 121–139. Academic Press, 2006.

[44] AJ Venkatakrishnan, Xavier Deupi, Guillaume Lebon, Christopher G Tate, Gebhard F Schertler, and M Madan Babu. Molecular signatures of G protein-coupled receptors. *Nature*, 494:185–194, 2013.

[45] Dorota Latek, Pawel Pasznik, Teresa Carlomagno, and Slawomir Filipek. Towards improved quality of GPCR models by usage of multiple templates and profile-profile comparison. *PloS One*, 8:e56742, 2013.

[46] Bartosz Trzaskowski, Dorota Latek, Shuguang Yuan, Umesh Ghoshdastider, A Debinski, and Slawomir Filipek. Action of molecular switches in GPCRs-theoretical and experimental studies. *Current Medicinal Chemistry*, 19:1090–1109, 2012.

[47] Reiner Vogel, Mohana Mahalingam, Steffen Lüdeke, Thomas Huber, Friedrich Siebert, and Thomas P Sakmar. Functional role of the "ionic lock" -an interhelical hydrogen-bond network in family A heptahelical receptors. *Journal of Molecular Biology*, 380:648–655, 2008.

[48] Daniel M Rosenbaum, Søren GF Rasmussen, and Brian K Kobilka. The structure and function of G protein-coupled receptors. *Nature*, 459:356–363, 2009.

[49] Cassandra Prioleau, Irache Visiers, Barbara J Ebersole, Harel Weinstein, and Stuart C Sealfon. Conserved helix 7 tyrosine acts as a multistate conformational switch in the 5HT2C receptor: identification of a novel "locked-on" phenotype and double revertant mutations. *Journal of Biological Chemistry*, 277:36577–36584, 2002.

[50] Karel Konvicka, Frank Guarnieri, Juan A Ballesteros, and Harel Weinstein. A proposed structure for transmembrane segment 7 of G protein-coupled receptors incorporating an asn-pro/asp-pro motif. *Biophysical Journal*, 75:601–611, 1998.

[51] Thomas E Angel, Sayan Gupta, Beata Jastrzebska, Krzysztof Palczewski, and Mark R Chance. Structural waters define a functional channel mediating activation of the GPCR, rhodopsin. *Proceedings of the National Academy of Sciences*, 106:14367–14372, 2009.

[52] Lei Shi, George Liapakis, Rui Xu, Frank Guarnieri, Juan A Ballesteros, and Jonathan A Javitch. β2 adrenergic receptor activation: Modulation of the proline kink in transmembrane 6 by a rotamer toggle switch. *Journal of Biological Chemistry*, 277:40989–40996, 2002.

[53] Irache Visiers, Juan A Ballesteros, and Harel Weinstein. Three-dimensional representations of G protein-coupled receptor structures and mechanisms. In *Methods in Enzymology*, volume 343, pages 329–371. Elsevier, 2002.

[54] Gregory V Nikiforovich, Christina M Taylor, Garland R Marshall, and Thomas J Baranski. Modeling the possible conformations of the extracellular loops in G protein-coupled receptors. *Proteins: Structure, Function, and Bioinformatics*, 78:271–285, 2010.

[55] Beata Jastrzebska. GPCR: G protein complexes - the fundamental signaling assembly. *Amino Acids*, 45:1303–1314, 2013.

[56] Søren GF Rasmussen, Brian T DeVree, Yaozhong Zou, Andrew C Kruse, Ka Young Chung, Tong Sun Kobilka, Foon Sun Thian, Pil Seok Chae, Els Pardon, Diane Calinski, et al. Crystal structure of the β2 adrenergic receptor-gs protein complex. *Nature*, 477:549–555, 2011.

[57] Klaus Peter Hofmann, Patrick Scheerer, Peter W Hildebrand, Hui-Woog Choe, Jung Hee Park, Martin Heck, and Oliver P Ernst. A G protein-coupled receptor at work: the rhodopsin model. *Trends in Biochemical Sciences*, 34:540–552, 2009.

[58] Maxime Louet, David Perahia, Jean Martinez, and Nicolas Floquet. A concerted mechanism for opening the GDP binding pocket and release of the nucleotide in hetero-trimeric g-proteins. *Journal of Molecular Biology*, 411:298–312, 2011.

[59] Gregory J Digby, Robert M Lober, Pooja R Sethi, and Nevin A Lambert. Some G protein heterotrimers physically dissociate in living cells. *Proceedings of the National Academy of Sciences*, 103:17789–17794, 2006.

[60] Nina Wettschureck and Stefan Offermanns. Mammalian G proteins and their cell type specific functions. *Physiological Reviews*, 85:1159–1204, 2005.

[61] Emma T van der Westhuizen, Celine Valant, Patrick M Sexton, and Arthur Christopoulos. Endogenous allosteric modulators of G protein-coupled receptors. *Journal of Pharmacology and Experimental Therapeutics*, 353:246–260, 2015.

[62] Wanling Song, Anna L Duncan, and Mark SP Sansom. Modulation of adenosine A2A receptor oligomerization by receptor activation and PIP2 interactions. *Structure*, 29:1312–1325, 2021.

[63] Lester A Rubenstein and Richard G Lanzara. Activation of G protein-coupled receptors entails cysteine modulation of agonist binding. *Journal of Molecular Structure: THEOCHEM*, 430:57–71, 1998.

[64] Patrick R Gentry, Patrick M Sexton, and Arthur Christopoulos. Novel allosteric modulators of G protein-coupled receptors. *Journal of Biological Chemistry*, 290:19478–19488, 2015.

[65] Bertil B Fredholm, Adriaan P IJzerman, Kenneth A Jacobson, Joel Linden, and Christa E Müller. International union of basic and clinical pharmacology. LXXXI. nomenclature and classification of adenosine receptors - an update. *Pharmacological Reviews*, 63:1–34, 2011.

[66] K Fuxe, D Marcellino, DO Borroto-Escuela, Michele Guescini, V Fernandez-Duenas, S Tanganelli, A Rivera, F Ciruela, and LF Agnati. Adenosine-dopamine interactions in the pathophysiology and treatment of CNS disorders. *CNS Neuroscience & Therapeutics*, 16:e18–e42, 2010.

[67] Catarina V Gomes, Manuella P Kaster, Angelo R Tomé, Paula M Agostinho, and Rodrigo A Cunha. Adenosine receptors and brain diseases: neuroprotection and neurodegeneration. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1808:1380–1399, 2011.

[68] Nelson Rebola, Rafael Lujan, Rodrigo A Cunha, and Christophe Mulle. Adenosine A2A receptors are essential for long-term potentiation of NMDA-EPSCs at hippocampal mossy fiber synapses. *Neuron*, 57:121–134, 2008.

[69] Nelson Rebola, Paula M Canas, Catarina R Oliveira, and Rodrigo A Cunha. Different synaptic and subsynaptic localization of adenosine A2A receptors in the hippocampus and striatum of the rat. *Neuroscience*, 132:893–903, 2005.

[70] Diana G Ferreira, Vânia L Batalha, Hugo Vicente Miranda, Joana E Coelho, Rui Gomes, Francisco Q Gonçalves, Joana I Real, José Rino, António Albino-Teixeira, Rodrigo A Cunha, et al. Adenosine A2A receptors modulate α-synuclein aggregation and toxicity. *Cerebral Cortex*, 27:718–730, 2017.

[71] Cyril Laurent, S Burnouf, B Ferry, VL Batalha, Joana E Coelho, Y Baqi, E Malik, E Mariciniak, S Parrot, Anneke Van der Jeugd, et al. A2A adenosine receptor deletion is protective in a mouse model of tauopathy. *Molecular Psychiatry*, 21:97–107, 2016.

[72] Sergi Ferré, Ivan Diamond, Steven R Goldberg, Lina Yao, Susanna MO Hourani, Zhili L Huang, Yoshihiro Urade, and Ian Kitchen. Adenosine A2A receptors in ventral striatum, hypothalamus and nociceptive circuitry: implications for drug addiction, sleep and pain. *Progress in Neurobiology*, 83:332–347, 2007.

[73] Akio Ohta and Michail Sitkovsky. Role of g protein-coupled adenosine receptors in downregulation of inflammation and protection from tissue damage. *Nature*, 414:916–920, 2001.

[74] Veli-Pekka Jaakola, Mark T Griffith, Michael A Hanson, Vadim Cherezov, Ellen YT Chien, J Robert Lane, Adriaan P Ijzerman, and Raymond C Stevens. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science*, 322:1211–1217, 2008.

[75] Wei Liu, Eugene Chun, Aaron A Thompson, Pavel Chubukov, Fei Xu, Vsevolod Katritch, Gye Won Han, Christopher B Roth, Laura H Heitman, Adriaan P IJzerman, et al. Structural basis for allosteric regulation of GPCRs by sodium ions. *Science*, 337:232–236, 2012.

[76] Thue W Schwartz, Thomas M Frimurer, Birgitte Holst, Mette M Rosenkilde, and Christian E Elling. Molecular mechanism of 7TM receptor activation - a global toggle switch model. *Annual Review of Pharmacology and Toxicology*, 46:481–519, 2006.

[77] Xavier Deupi and Jörg Standfuss. Structural insights into agonist-induced activation of g protein-coupled receptors. *Current Opinion in Structural Biology*, 21:541–551, 2011.

[78] Guillaume Lebon, Tony Warne, Patricia C Edwards, Kirstie Bennett, Christopher J Langmead, Andrew GW Leslie, and Christopher G Tate. Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature*, 474:521–525, 2011.

[79] Gleb V Novikov, Victor S Sivozhelezov, and Konstantin V Shaitan. Investigation of the conformational dynamics of the A2A adenosine receptor by molecular dynamics simulation. *Biophysics*, 58:482–492, 2013.

[80] Dasiel O Borroto-Escuela, Sonja Hinz, Gemma Navarro, Rafael Franco, Christa E Müller, and Kjell Fuxe. Understanding the role of adenosine A2AR heteroreceptor complexes in neurodegeneration and neuroinflammation. *Frontiers in Neuroscience*, 12:43, 2018.

[81] Lindahl, Abraham, Hess, and van der Spoel. Gromacs 2021.5 manual, January 2022.

[82] Andrej Šali and Tom L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234:779–815, 1993.

[83] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[84] Loup Verlet. Computer" experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical review*, 159:98, 1967.

[85] Roger W Hockney. The potential calculation and some applications. *Methods in Computational Physics*, 9:136, 1970.

[86] William C Swope, Hans C Andersen, Peter H Berens, and Kent R Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76:637–649, 1982.

[87] Herman JC Berendsen, JPM van Postma, Wilfred F Van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81:3684–3690, 1984.

[88] Tetsuya Morishita. Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath. *The Journal of Chemical Physics*, 113:2976–2982, 2000.

[89] Shuichi Nose. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 52:255–268, 1984.

[90] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52:7182–7190, 1981.

[91] Hans C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72:2384–2393, 1980.

[92] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An nlog(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98:10089–10092, 1993.

[93] Thomas J Piggot, Jane R Allison, Richard B Sessions, and Jonathan W Essex. On the calculation of acyl chain order parameters from lipid simulations. *Journal of Chemical Theory and Computation*, 13:5683–5696, 2017.

[94] Paul Heitjans and Jörg Kärger. *Diffusion in condensed matter: methods, materials, models*. Springer Science & Business Media, 2006.

[95] Edward J Maginn, Richard A Messerly, Daniel J Carlson, Daniel R Roe, and J Richard Elliot. Best practices for computing transport properties 1. self-diffusivity and viscosity from equilibrium molecular dynamics. *Living Journal of Computational Molecular Science*, 1:6324–6324, 2019.

[96] Matti Javanainen, Henrik Hammaren, Luca Monticelli, Jae-Hyung Jeon, Markus S Miettinen, Hector Martinez-Seara, Ralf Metzler, and Ilpo Vattulainen. Anomalous and normal diffusion of proteins and lipids in crowded lipid membranes. *Faraday Discussions*, 161:397–417, 2013.

[97] Emma Falck, Tomasz Róg, Mikko Karttunen, and Ilpo Vattulainen. Lateral diffusion in lipid membranes through collective flows. *Journal of the American Chemical Society*, 130:44–45, 2008.

[98] George B Benedek and Felix MH Villars. *Physics with illustrative examples from medicine and biology: statistical physics*. Springer Science & Business Media, 2000.

[99] Timo Vuorela, Andrea Catte, Perttu S. Niemelä, Anette Hall, Marja T. Hyvönen, Siewert-Jan Marrink, Mikko Karttunen, and Ilpo Vattulainen. Role of lipids in spheroidal high density lipoproteins. *PLoS Computational Biology*, 6:1–14, 10 2010.

[100] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[101] James Gareth, Witten Daniela, Hastie Trevor, and Tibshirani Robert. *An introduction to statistical learning: with applications in R*. Spinger, 2013.

[102] Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.

[103] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:1–22, 1977.

[104] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19:716–723, 1974.

[105] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978.

[106] Edward Egelman. *Comprehensive biophysics*. Elsevier, 2012.

[107] Przemyslaw Jurczak, Kosma Szutkowski, Slawomir Lach, Stefan Jurga, Paulina Czaplewska, Aneta Szymanska, and Igor Zhukov. Dmpc phospholipid bilayer as a potential interface for human cystatin c oligomerization: analysis of protein-liposome interactions using nmr spectroscopy. *Membranes*, 11:13, 2020.

[108] Yifei Qi, Jumin Lee, Jeffery B Klauda, and Wonpil Im. Charmm-gui nanodisc builder for modeling and simulation of various nanodisc systems. *Journal of Computational Chemistry*, 40:893–899, 2019.

[109] Lisbeth R Kjølbye, Leonardo De Maria, Tsjerk A Wassenaar, Haleh Abdizadeh, Siewert J Marrink, Jesper Ferkinghoff-Borg, and Birgit Schiøtt. General protocol for constructing molecular models of nanodiscs. *Journal of Chemical Information and Modeling*, 61:2869–2883, 2021.

[110] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L De Groot, Helmut Grubmüller, and Alexander D MacKerell Jr. Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14:71–73, 2017.

[111] Jeffery B Klauda, Richard M Venable, J Alfredo Freites, Joseph W O'Connor, Douglas J Tobias, Carlos Mondragon-Ramirez, Igor Vorobyov, Alexander D MacKerell Jr, and Richard W Pastor. Update of the charmm all-atom additive force field for lipids: validation on six lipid types. *The Journal of Physical Chemistry B*, 114:7830–7843, 2010.

[112] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79:926–935, 1983.

[113] Dmitrii Beglov and Benoit Roux. Finite representation of an infinite bulk system: solvent boundary potential for computer simulations. *The Journal of Chemical Physics*, 100:9050–9063, 1994.

[114] Lindahl, Abraham, Hess, and van der Spoel. Gromacs 2021 source code, January 2021.

[115] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. Lincs: a linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18:1463–1472, 1997.

[116] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *The Journal of Chemical Physics*, 103:8577–8593, 1995.

[117] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31:1695, 1985.

[118] LLC Schrödinger and Warren DeLano. Pymol.

[119] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[120] John Edward Stone et al. An efficient library for parallel ray tracing and animation. *University of Missouri, Rolla*, Master's Thesis, 1998.

[121] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9:90–95, 2007.

[122] Andrey Filippov, Greger Orädd, and Göran Lindblom. Influence of cholesterol and water content on phospholipid lateral diffusion in bilayers. *Langmuir*, 19:6397–6400, 2003.