Keble College, University of Oxford

Using Molecular Simulations to Parameterize Discrete Models of Protein Movement in the Membrane

Author:

Thomas A. Hirst-Dunton

Supervisors:

Dr. James M. Osborne

Prof. David J. Gavaghan

Prof. Mark S. P. Sansom

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Oxford.

April, 2015

The work presented in this thesis centres on the development of a work-flow in which coarse-grained molecular dynamics (MD) simulations of a planar phospholipid bilayer, containing membrane proteins, is used to parameterize a larger-scale simplified bilayer model. Using this work-flow, repeat simulations and simulations of larger systems are possible, better enabling the calculation of bulk statistics for the system. The larger-scale simulations can be run on commercial hardware, once the initial parameterization has been performed.

In the simplified representation, each protein was initially only represented by the position of its centre of mass and later with the inclusion of its orientation. The membrane protein used throughout most of this work was the bacterial outer membrane protein NanC, a member of the KdgM family of proteins. To parameterize the motion and interaction of proteins using MD, the potential of mean force (PMF) for the pairwise association of two proteins in a bilayer was calculated for a variety of orientational combinations, using a modified umbrella sampling procedure. The relative orientations chosen represented extreme examples of the contact regimes between the two proteins: they approximately corresponded to maxima and minima of the solvent inaccessible surface area, calculated when the proteins were in contact. These PMFs showed that there was a correlation between the buried surface area and the depth of the potential well in the PMF; this is something that, to date, has only been observed in these relatively-'featureless' membrane proteins (but is seen in globular proteins), where the effect of the interactions with lipids in the bilayer plays a larger role. Features in the PMF were observed that resulted from the preferential organization of lipids in the region between the two proteins. These features were small wells in the PMF, which occurred at protein separations that corresponded to the intervening lipids being optimally packed between the proteins. This result further highlighted the role that the lipids in the bilayer played in the interaction between the NanC proteins.

The simplified bilayer model was parameterized using the PMFs and the relationship between buried surface area and potential well depth. The initial model included only the proteins' positions. A series of Monte Carlo simulations were performed in order to compare the system behaviour to that of an equivalent MD simulation. Initially, the MD simulation and our parameterized model did not show a good agreement, so a Monte Carlo scheme that incorporated cluster-based movements was implemented. The agreement between the MD simulation and the simulations of our model using the cluster-based scheme, when comparing diffusive and clustering behaviour, was good. Including the orientation-dependent features of the parameterization resulted in the emergence of behaviour that was not clearly detectable in the MD simulation.

Finally, attempts were made to parameterize the model using PMFs for the association of rhodopsin from the literature. Rhodopsin was a much more complicated protein to represent: there was not a clear correlation between surface area and the features of the PMF, and the geometry of the interaction between two rhodopsins was more complicated. Simulations of the 'rows-of-dimers' system of rhodopsin, observed in disc membranes, was not entirely well represented by the model; for such a closely packed system, where the number of lipids is much closer to the number of proteins, the use of an implicit-lipid model meant that the effect of the reduced lipid mobility was not adequately captured. However, the model accurately captures the orientational composition of the system. Future work should be focussed on incorporating explicit representations of the lipid in the system so that the behaviour of close-packed systems are better represented.

Contents

Acknowledgements

My time in Oxford has been long: much longer than I'd ever anticipated. When I first arrived at Keble College in 2005 to begin my undergraduate degree in physics, I had never properly considered the fact that I could still be here in 10 years' time, finally completing my DPhil. 18-year-old me would never have realized this, but I think Oxford is such a wonderful city in which to live and I'm sure I wouldn't be enjoying myself half as much if I were anywhere else. I am very grateful that I've been able to spend such a large portion of my life so far in this great place.

My DPhil journey began in 2009 when I started the Life Sciences Interface programme at the Doctoral Training Centre. It was the opportunities made available to me during this year-long programme, along with the experiences I had and the choices I made, that enabled me to put down the pipettes and pick up the keyboard; before I started, the majority of my research experience had been in laboratories: putting small amounts of clear liquid into small amounts of other clear liquid, working with large pieces of electronic equipment, and dealing with regular and unexplainable disappointments. I'm really glad that I made the decision to start on the DTC programme, as it gave me the opportunity to change the direction in which my academic life was heading and has led to me being able to embark on a career in which I'm really able to both push and enjoy myself. My time at the DTC was also great fun; it was brilliant to be able to work alongside such a great group of fellow students. We may have been playing cards more often than doing actual work, but were enjoying the work nonetheless.

My project supervisors—Ozzy, Dave and Mark—have been incredibly helpful throughout the entire project. Ozzy's willingness to take any piece of junk that I had cobbled together for him and still provide thorough and insightful comments and make invaluable suggestions was something that no other student I knew had experienced with their supervisors. Dave and Mark brought decades of collective experience with them that allowed me to see what I was doing with a much wider perspective and understand where I should be taking the work; such insight was always invaluable. Without all of my supervisors' hard work, I would definitely not have a thesis in which to write these acknowledgements.

My day-to-day research experience has been made both enjoyable and fruitful as a result of the people alongside whom I spent most of my time. All of the students, post-docs and academics in both the Computational Biology group and the Structural Bioinformatics and Computational Biochemistry group with whom I've both worked and socialized have made the daily grind of completing this thesis that much more enjoyable; having a good long moan about things, dragging someone aside to scribble ideas on a whiteboard, playing a quick game of Doom, or escaping early for an ale or two really makes the work that much easier. Particular thanks go to Joe Goose, without whose help and collaboration I would have struggled to get my paper published.

But to to talk of research choices and to thank research groups is to tell only part of the story of my DPhil. It is at Keble, as part of the MCR, where I've spent so much of my time outside of work. Whether playing board games or drinking games, strange sports like croquet or stranger ones like kubb, whiling away the summer on punts or with a miserable attempt at competitive rowing, bursting into an impromptu and uninvited rendition of "The Twelve Days of Christmas" in hall or an impromptu and uninvited rendition of "Everybody Dance Now" in a club trying to close, the memories of my time in Keble MCR will stay with me forever. I was initially unsure how I felt about my hand having been somewhat forced into staying at Keble for my DPhil, but time and time

again I've said how fortunate I am that it was; the people I've met, whether just passing through for a year or in it for the long haul, have made my time in Oxford so much fun and left me with the fondest of memories.

As far as being in it for the long haul is concerned, there is clearly one winner. Someone with whom I've embarked on a rather longer journey and who's been a massive part of my life while working towards my DPhil. Hannah's influence on my work is undeniable (I've used 'whom' three times so far in these acknowledgements and there are loads of semi-colons!). Whether she was offering a comforting cuddle when I was in the doldrums of the second year or a kick up the arse when the fourth year rolled into the fifth, I am so very grateful for all of her input, both direct and indirect. All manner of troubles that may have occurred during the day have a habit of melting away when you are returning to such a wonderful life at home. I may have gained a DPhil during my time here but, much more importantly, I've also managed to find someone who'll have me for a husband and wants to embark on much more rewarding and challenging projects to come.

Abbreviations

AFM atomic force microscopy

ATP adenosine triphosphate

CG coarse-grained

CGMD coarse-grained molecular dynamics

EGF epidermal growth factor

FCS Fourier correlation spectroscopy

FRAP fluorescence recovery after photobleaching

FRET Förster resonance energy transfer

GPCR G protein-coupled receptor

GROMACS Groningen machine for chemical simulations

MC Monte Carlo

MCS Monte Carlo step(s)

MD molecular dynamics

MDR mean squared rotation

MSD mean squared deviation

N-BAR N-terminal bin–amphiphysin–rvs

OMP outer membrane protein

OMPLA outer membrane phospholipase A1

PC phosphatidylcholine

PIP phosphatidylinositol phosphate

PMF potential of mean force

POPE 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine

POPG 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoglycerol

SASA solvent-accessible surface area

VMMC Virtual-move Monte Carlo

WHAM weighted histogram analysis method

Chapter 1

Introduction

The cellular membrane is a complex and essential part of all living organisms: many of the most important metabolic processes are affected by the behaviour linked to the membrane environment. The work presented in this thesis is focussed on the cellular membrane and its processes. Modelling and simulation is used to develop a work-flow aimed at improving our understanding of the movement of and interaction between proteins in the membrane.

1.1 The importance of the cellular membrane

At the most basic level, the cellular membrane separates the cell from its surroundings; it provides a compartment within which interactions beneficial to the cell's survival are more likely to occur (Deamer et al. 2002). The origins of cell membranes are a topic of debate (Ruiz-Mirazo et al. 2014), but much credence is given to theories based around the notion of early membranes having a similar structure but formed from simpler molecules (Rendón et al. 2012). The membranes surrounding prebiotic life may have only passively provided a more productive environment for self-replication to occur than in the surrounding sea, but these simple structures have evolved into a varied and multi-functional part of all organisms.

The cellular membrane, most often observed as a bilayer structure wrapped around to form a vesicle, is host to a multitude of molecules whose interactions and operations form part of an organisms metabolic process. Accounting for approximately a quarter of the coding regions of an organism's genome (Nilsson et al. 2005), membrane-associated proteins control the transport of solutes between a cell and its surroundings, facilitate cellular movement, and regulate many aspects of cellular behaviour.

Throughout much of this thesis we are concerned with membrane proteins found in the outer membranes of Gram-negative bacteria; Gram-negative bacteria are surrounded by two concentric membranes, which are separated by a periplasmic region. The outer membrane is composed of phospholipids in the inner (i.e. periplasmic) leaflet of the bilayer, and of lipopolysaccharides in the outer leaflet. Within this outer membrane many species of outer membrane proteins (OMPs) are found; OMPs are a class of integral membrane proteins whose secondary structures are almost exclusively β-barrels (Koebnik et al. 2000). Many of these β-barrel OMPs are porins (OmpC, OmpF, LamB, NanC, for example), through which small (approximately 500 g mol^{-1}) molecules can diffuse across the membrane. Porins provide a route for many antibiotics into bacterial cells and are potential targets for vaccines (Nikaido 2001).

1.2 The role of computer simulation in biological research

There are many challenges associated with experiments on cell membranes *in vivo* and *in vitro*, so it is important that we strive to improve the methods by which we can simulate membrane environments. Computer simulations, often described as *in silico* experiments, provide a complementary approach to both *in vitro* and *in vivo* studies (Stansfeld et al. 2011), enabling us to probe the microscopic interactions underlying membrane processes. It is through *in silico* experiments that we can hope to reduce the time taken to perform and the cost of undertaking research in important areas, such as drug discovery, and attempt to understand more fully the processes governing protein dynamics in the bilayer (Efremov et al. 2004; Burrage et al. 2011; Lyubartsev et al. 2011). Simulations can be performed on many scales, from the motion of the individual atoms in a system to the representation of an entire organism. The multi-scale nature of biological processes present significant simulation challenges. How can we effectively unite processes that occur on different time and length scales?

Molecular dynamics (MD) simulations have been used to explore a range of membrane proteins (Stansfeld et al. 2011), in addition to related approaches such as Monte Carlo (Janosi et al. 2010) and Brownian dynamics (Cui et al. 2008) simulations. In particular, simulations using a coarse-grained approximation (Monticelli et al. 2008) have been used to increase the time and length scales that are accessible to MD simulations. By using coarse-grained MD simulations, it has been possible to perform simulations containing hundreds of proteins and thousands of lipids. Such simulations may be used to investigate protein clustering on a large scale and the effect of the motion of lipids.

The behaviour of large systems may exhibit significant differences from what would be expected if we were to extrapolate the behaviour observed in simulations of smaller systems. It is by performing large simulations that emergent behaviour, which is inherent to systems above a certain size, could be investigated. It is also important to try to investigate membrane systems that have a diverse array of constituents, something

which will better approximate membranes *in vivo*. However, using MD to investigate such systems requires a lot of computational power. The use of supercomputers, with thousands of compute cores dedicated to each simulation, is required to complete the simulations in a reasonable length of time (usually days). With the increasing availability of supercomputing time, coupled with the decrease in cost, it has become easier to carry out such simulations, but it would be advantageous to have more computationally efficient simulations, especially for running repeat simulations to generate statistics describing the system behaviour.

1.3 The aim of this thesis

In this thesis we use modelling and simulation to bridge the gap between two different scales. By using coarse-grained MD to characterize the interaction between a pair of OMPs, we are able to understand more about the rules that govern protein-protein interaction in the bacterial outer membrane; in this work we use a sialic acid porin, NanC, which is found in *E. coli*. Using this characterization we develop and parameterize a discrete model of the membrane, in which the NanC proteins are represented by their centres of mass in a planar patch of membrane with implicit lipids. We create this simplified representation to enable repeat simulation of a membrane system so that we can calculate statistics describing the behaviour of the proteins. We are able to do so because simulating it has greatly reduced requirement for computational resources.

Throughout this thesis we wish to compare the performance of our simplified model to data so that we can assess its performance. The data we use for most of the comparisons are obtained from a large membrane simulation performed by Dr. Joseph Goose. Large coarse-grained MD simulations, the largest containing hundreds of OMPs, were performed. The simulations have so far primarily been used to study the diffusion of lipids and proteins (Goose et al. 2013; Chavent et al. 2014), but a publication on the clustering of proteins using these large membrane simulations is in preparation. The OMPs simulated

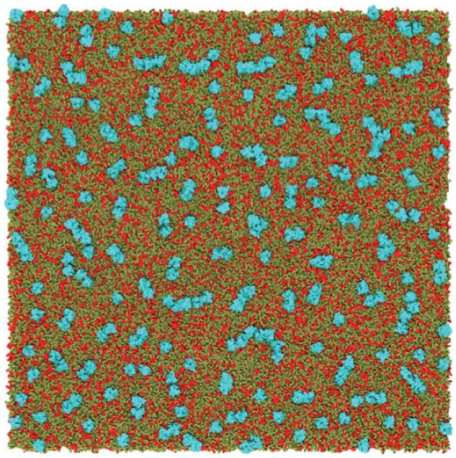

Figure 1.1: A coarse-grained representation of a bilayer system containing 256 OmpA proteins (cyan), 28260 POPE molecules (brown) and 9420 POPG molecules (red). Adapted with permission from Chavent et al. (2014).

were the proteins FhuA, LamB, NanC, OmpA and OmpF. An example of the scale of the simulations can be seen in Figure 1.1 for the large membrane simulation containing 256 OmpA proteins. We use data from the one of the simulations of NanC extensively throughout this thesis as a means to evaluate the performance of our simplified model. The MD simulations were performed using a bilayer constructed from approximately 38000 lipids. The lipid species used to create the bilayer were POPE and POPG, at a ratio of 3:1. The bilayer contained 256 NanC proteins and was built by successively tiling smaller patches of membrane to create a larger patch, equilibrating at each stage. The protein density corresponds to an initial protein separation of approximately 7.5 nm, assuming proteins were initially placed on a uniform square grid. The NanC simulations were 1 μs long and took a week to run on thousands of computer cores. Comparisons between simulations of our model and the MD simulations were made using the positions and orientations of the NanC proteins in the MD simulation: the same degrees of freedom that are represented in our reduced model.

1.4 The structure of this thesis

The work presented in this thesis is arranged into the following chapters. Literature on *in vivo* and *in vitro* methods used to study membrane biology is reviewed in Chapter 2. This discusses research into protein movement and clustering and the interactions that play a role in such behaviours. The literature on studying membranes and membrane proteins using MD and using other larger-scale discrete simulation methods is reviewed at the beginnings of Chapters 3 and 4, respectively; these are the chapters containing our results from employing such methods. Chapter 3 presents the results of our investigations into the interaction between a pair of NanC proteins, characterized by the potential of mean force (PMF); the work in this chapter resulted in the publication: "The free energy landscape of dimerization of a membrane protein, NanC." (Dunton et al. 2014). In Chapter 4 we develop a simple model of the membrane that contains proteins whose interactions are based on the PMFs presented in Chapter 3. This model is extended in Chapter 5 to include orientational degrees of freedom, whose behaviour are also parameterized by the PMFs of Chapter 3. The conclusions of the research presented in this thesis are presented in Chapter 6.

Chapter 2

Membrane Biology Literature

This chapter reviews research from the literature that focuses on cell membranes and the behaviour of membrane proteins, both *in vivo* and *in vitro*. Starting with a discussion of the historical representation of the cell membrane, the fluid-mosaic model of the membrane is discussed. Challenges to the fluid-mosaic model are presented in view of recent experimental evidence. Research relating to the diffusion and clustering or proteins in the membrane, including hydrophobic mismatch, lipid rafts and anomalous diffusion is reviewed. Finally, studies on the role and character of protein-protein interactions in the membrane discussed, alongside an overview of the techniques used.

2.1 Introduction

Cells, first discovered by Hooke (1665), are the building blocks of all organisms. Early work on the theory of cells began in the first half of the 19th century, but it was with the work of Overton, in which he hypothesized that the membrane was made of lipid molecules, where some of the first steps in the understanding of membrane biology were taken (Overton 1895). Gorter et al. (1925) demonstrated that the amount of "lipoids" present in a cell were exactly sufficient to cover its surface twice, leading them to suggest that these lipids formed a membrane that was two molecules thick. Electron microscope images of the membranes of axons led to the development of the unit membrane theory, in which it was suggested that the lipid bilayer was surrounded on either side by a layer of proteins (Robertson 1972). This view was later shown to be incorrect and was replaces by the fluid-mosaic model of Singer et al. as the *de facto* model of cellular membranes.

Firstly, we look at the fluid-mosaic model of the membrane and how its accuracy as a description of cell membranes has been questioned and updated since its inception in the 1970s. Next we discus research into the diffusion and clustering of membrane proteins *in vivo* and *in vitro*. Finally, we review work that has focussed on understanding and characterizing the interactions between proteins in the membrane.

2.2 The fluid-mosaic model of cell membranes

The fluid-mosaic model of membranes was initially proposed by Singer et al. (1972); it was based on fluorescence experiments in which two, initially segregated, antigens, tagged with different fluorophores, were seen to mix upon membrane fusion (Frye et al. 1970). The fluid-mosaic model describes the membrane as a phospholipid bilayer structure. Phospholipids are amphipathic, the result of which is that they preferentially organize themselves so that their hydrophilic head groups are in contact with water and their hydrophobic tails are shielded from water. The bilayer structure is formed of two layers of

lipids, in both of which the lipids have their head groups in contact with the surrounding water-based environment and their tails in contact with the tails of the other layer of lipids. The fluid-mosaic model describes the bilayer as an environment populated with integral membrane proteins. Some of these proteins span the entire bilayer, in contact with the environment on either side of the bilayer; these proteins can act as transporters of ions across the membrane or as part of signalling processes. Other proteins may be embedded in only one leaflet of the bilayer, acting as markers to other metabolic processes by identifying the membrane or as part of an enzymatic process. The fluid-mosaic model describes the bilayer as mostly consisting of phospholipids, with the integral membrane proteins freely diffusing in the plane of the membrane.

At over 40 years old, the fluid-mosaic model has persisted due to its success as a general model that describes many of the behaviours observed in the complicated environment of cell membranes, but it has not done so without attracting criticism (Engelman 2005) and revision (Nicolson 2014). Most of its initial assertions have been challenged during the course of its existence; for instance, its assertions that the membrane is a random two-dimensional fluid, that the membrane has a uniform thickness to which proteins' hydrophobic regions are tuned, and that integral membrane proteins occupy a very small fraction of the membrane surface area have all been challenged. Non-random lateral organization of the membrane has been observed as a result of lipid behaviour (Lingwood et al. 2009) and as a result of the interaction of membrane proteins with cytoskeletal structures and the extracellular matrix (Kusumi et al. 2005). For instance, optical tweezers were used to drag the immobilized protein Qa-2 across a cell membrane and discrete positions were identified at which the protein experienced a high resistance; these positions corresponded to cytoskeletal attachment sites (K. Suzuki et al. 2001). Lipid bilayer phase and thickness, as well as varying as a result of non-uniform bilayer compositions, can be modulated by the presence of proteins: Killian et al. (1996) and Mitra et al. (2004) used nuclear magnetic resonance and electron microscopy techniques, respectively, to show

that the presence of proteins and peptides could lead to changes in the thickness of lipid bilayers. In many situations cell membranes are crowded environments, with an area fraction of 25% or greater taken up by membrane proteins: Dupuy et al. (2008) shaved the surface of a cell using NaOH to leave only the transmembrane α-helices in the membrane; using this shaved membrane they made measurements of the density to estimate the protein area-occupancy. The protein area-occupancy is also effectively increased when the membrane is populated by proteins that have large extra-membrane regions, which create steric contacts outside of the membrane: the Kv channel protein Kv1.2 (Long et al. 2005) and F_1F_0-ATP synthase (Giraud et al. 2012) both have large extra-membrane regions that will affect the minimum separation of proteins in the membrane.

2.3 Membrane protein diffusion and clustering

The diffusion and clustering of proteins in the membrane is vital to many metabolic processes. Voltage-gated ion channels are precisely distributed in the membranes of neurons so that signals are transmitted correctly (Choquet et al. 2003). Also, any breakdown in the organization of proteins in the membrane, possibly resulting from mutations to the proteins themselves, can lead to a number of diseases (Cobbold et al. 2003). Because of the important role played by the motion and clustering of membrane proteins, it is a very active area of research.

Movement in the membrane is noisy, but there are a host of experimental techniques that enable us to gather information (Owen et al. 2009). Fluorescent techniques can be used to measure bulk diffusion characteristics of proteins in the membrane across a range of time and length scales. Ramadurai et al. (2009) used fluorescence correlation spectroscopy (FCS) to investigate lateral diffusion of proteins in vesicle membranes and were able to verify that the diffusion coefficient varies with the radius of the protein; in FCS the autocorrelation of fluorescence intensity measurements taken from a small volume, is used to discern the diffusive behaviour of tagged proteins. Another factor

that can affect the diffusion of proteins is hydrophobic mismatch: the difference between the length of the hydrophobic region of the protein and the thickness of the membrane. Fluorescence recovery after pattern photobleaching (FRAP) was employed by Gambin, Reffay, et al. (2010) to investigate how protein diffusion coefficients were affected by variations in hydrophobic mismatch; FRAP involves photobleaching a region of membrane that contains fluorescently-tagged proteins and then observing the recovery of fluorescence, driven by the diffusion of still-fluorescing proteins into the photobleached region, to study the diffusive behaviour. Brown (2006) was able to use FRAP to investigate the diffusion properties of proteins with varying affinities for rafts. Diffusion of membrane proteins is often described by the Saffman-Delbrück model (Saffman et al. 1975), which was verified to hold in both the small and large protein regimes by Weiß et al. (2013) using FCS; the Saffman-Delbrück model represents membrane proteins as cylinders, which are embedded in a three-dimensional planar fluid (representing the bilayer) surrounded by a less-viscous fluid (representing the cytosol), and uses fluid-dynamics to derive the equations of motion for the protein cylinders.

There has been much work on studying the anomalous diffusion of membrane proteins, where the diffusion has a nonlinear relationship with time. Single molecule tracking experiments (in which the individual trajectories of fluorescently-tagged proteins are measured) have demonstrated that under certain conditions membrane proteins undergo anomalous diffusion (Schütz et al. 1997; Weiss et al. 2003; Lenne et al. 2006; Spillane et al. 2014). Recruitment of proteins to lipid microdomains (Lasserre et al. 2008), interaction with the cytoskeletal network (Lenne et al. 2006) and coupling between the two bilayer leaflets (Spillane et al. 2014) all contribute to the anomalous diffusion of certain proteins. However, not all membrane proteins undergo anomalous diffusion: Crane et al. (2008) demonstrated that the diffusion of aquaporin-1 was not affected by the disruption of the cytoskeletal network and underwent long-range normal diffusion. By studying the movement of membrane proteins that cooperate as part of some process, it is possible to

understand more about the mechanics of the operation. Spector et al. (2010) measured the diffusion coefficients of OmpF and BtuB, using single molecule tracking *in vivo*, and showed that the formation of the colicin translocon complex depends on a collision between the relatively mobile BtuB and the relatively-immobile OmpF.

As well as playing a role in modulating protein function, as observed with the lipid PIP_2 and various ion channels (Suh et al. 2008) and with EGF receptor (Coskun et al. 2011), lipids and other non-protein membrane constituents play an important role in controlling protein movement in the membrane. For instance, the function and cortical-actin-mediated stabilization of ATP-binding cassette transporters were found to be highly dependent on their being located in a lipid raft (Kok et al. 2014). Lipid microdomains, or lipid rafts, are regions within the lipid bilayer that are enriched with sphingolipids and sterols such as cholesterol (known to increase membrane thickness and rigidity (Roduit et al. 2008)) and are thought to play an active role in many processes requiring lateral organization in the membrane (Simons et al. 1997; Lingwood et al. 2009).

Whilst the majority of discussion as to the nature of membrane protein cluster formation has focussed on lipid rafts, it should be noted that lateral organization of membranes is a more general property; lateral organization of lipids (Mileykovskaya et al. 2000) and proteins (Spector et al. 2010) has been observed in *E. coli* with studies on the patterns formed by fluorescently-tagged lipids during growth and with single molecule tracking, respectively, and is also observed in many more bacterial species (Mileykovskaya et al. 2005). The functional importance of protein clustering is demonstrated by the above example with the colicin translocon. Clustering of proteins may also be the result of the preferential association of combinations of proteins, for instance, as observed in syntaxins (Sieber et al. 2007), where it is the protein-protein interaction that is key to understanding behaviour.

2.4 Protein-protein interactions in the membrane

It is through protein-protein interaction that many metabolic processes are controlled and more still involve some form of protein-protein interaction. Membrane protein complexes are involved in the assembly of other membrane proteins, for instance, the β-barrel proteins of various bacterial outer membranes (Höhr et al. 2015), or as part of protein transport, for instance in the bacterial holo-translocon studied by Schulze et al. (2014) who assessed the stoichoimetry and activity of the complex by looking at relative expression levels of the constituent proteins. Only by understanding the way in which the proteins interact with each other in the membrane can we get a full understanding of how these kinds of processes are facilitated by the membrane environment and protein features. Useful information that we hope to gain about protein-protein interactions includes: knowledge of binding modes, association-dissociation equilibria, and biophysical mechanisms controlling the interactions.

The existence of interactions between proteins can be identified using proteomics (Ngounou Wetie et al. 2013) and fluorescence techniques (Lowder et al. 2011), leading to the creation of protein interaction networks. However, it is only through more-detailed qualitative and quantitative studies of protein-protein interactions that we can understand how the dynamics of the membrane play a role in these interactions. Fluorescence techniques like Förster resonance energy transfer (FRET), which was first used to study membranes by Fernandez et al. (1976), can measure binding affinities using tagged proteins by detecting the interaction between different types of fluorophore (detectable in the emission of certain wavelengths of light) that have been used to tag various different components of the system. For instance, the membrane associated dimerization of N-BAR domains (Capraro et al. 2013), and the oligomerization state of transporters (Sergeev et al. 2012) and G protein-coupled receptors (GPCRs) (Maurel et al. 2008; Albizu et al. 2010) have all been investigated using FRET. Single particle tracking has also been used to characterize the monomer-dimer equilibrium rate constants of a GPCR (Kasai et al. 2011) and to

study the dimerization state of glycosylphosphatidylinositol-anchored proteins, which are associated with lipid rafts (K. G. N. Suzuki et al. 2012). Using quantum-dots as fluorescent labels, the dimerization of erbB1 was studied in order to understand the role that ligand binding had on the lifetime of the dimers; it was found that the ligand binding resulted in a longer-lived dimer (Low-Nam et al. 2011).

Other techniques for understanding the energetics of protein-protein interaction involve biochemical, rather than physical, techniques. For instance, the dimerization of transmembrane α-helices in glycophorin was investigated by Fleming et al. (2001) using sedimentation equilibrium analytical ultracentrifugation; they found that the energetics of helix-helix interaction could be explained using simple protein-protein interaction principles and that the contributions to the energy of association were not uniformly distributed across the interaction site. Sedimentation equilibrium using analytical centrifugation involves spinning a container of the proteins of interest in solution at a speed that it high enough to force the proteins toward the side of the container, but not high enough to form a pellet. This action creates a gradient of proteins across the container, as the action of the centrifuge is balanced by the diffusion of the proteins. Once equilibrium is reached, the form of this gradient can be used to calculate the proportion of proteins that are oligomerized because its form is a function of the mass of the proteins, not their shape. Ebie et al. (2007) also used this method to discover that the dimerization energy of an erythropoietin receptor was different for the murine and human variants of the protein, suggesting that they may perhaps play different roles in the two organisms. This work on α-helical proteins suggested a linear correlation between the strength of interaction and the buried surface area of the protein complex. There are many published examples of experimentally determined dimerization energies for α-helical membrane proteins and peptides (Russ et al. 1999; Cristian et al. 2003; Doura et al. 2004; Ebie et al. 2007; MacKenzie et al. 2008), but relatively few for β-barrel proteins: one important example being the dimerization free energy of the phospholipase OMPLA, which was

found to be in the region -25 to -35 kJ mol^{-1} (Stanley et al. 2006) and was not observed to correlate with the buried surface area (Ebbie Tan et al. 2008).

Recently there have been advances in the use of atomic force microscopy (AFM), first developed by Binnig et al. (1986), to study membrane proteins in a bilayer environment. Work by Ando (2012) and Casuso et al. (2012), who used high-speed AFM to study the dynamics of membrane proteins in supported bilayer membranes, has shown that AFM promises to be an invaluable experimental technique as the field of high-speed AFM matures.

2.5 Summary

The cell membrane continues to be a very active focus of research, with a multitude of techniques employed, both *in vivo* and *in vitro*, and from a range of disciplines, in order to understand the operation of this important system. With membrane proteins playing a key role in many diseases and as the targets for many pharmaceuticals, their importance is unlikely to change. It is for this reason that we must also use modelling and simulation to complement the work done *in vivo* and *in vitro*, and improve our ability to reason about the processes that govern behaviour of the cellular membrane. Literature related to *in silico* experimentation using molecular dynamics and using discrete-protein simulation techniques is reviewed in Chapters 3 and 4, respectively.

Chapter 3

Characterizing the Free Energy Landscape of NanC dimerization with the Potential of Mean Force

This chapter describes the use of molecular simulations to calculate the free energy of association for a pair of bacterial outer membrane proteins. This is done by calculating orientationally-restrained potentials of mean force for various combinations of the relative orientations of two NanC proteins, as a functions of their separation. The aim of this characterization is to gain a deeper understanding of this three-dimensional free energy landscape (the proteins' separation and their two orientations) by looking at specific hypersurfaces through the higher-dimensional free energy landscape. In later chapters this characterization of the interaction is used to parameterize a simplified model of the bacterial outer membrane.

Contents

3.1 Introduction

The complexity of the bilayer environment makes the detailed study of membrane proteins difficult. The advent of new techniques constantly brings new insight into membrane proteins, but the difficulty in doing so makes computer simulations of model systems a useful approach. Using simulations we are able to provide a deeper understanding of the interactions that determine membrane organization at the level of proteins and small molecules.

3.1.1 Molecular dynamics

Molecular dynamics (MD) simulations of a system traditionally represent molecules in near-complete atomic detail. Originally developed by Alder et al. (1959), simulations proceed by integrating the equations of motion for the system.

3.1.1.1 The equations of motion

Newton's second law states that the rate of change of an object's momentum is equal to the net force acting on that object. This is captured in the following equation for a system of N objects,

$$\mathbf{F}_i = \frac{d\mathbf{p}_i}{dt}, \tag{3.1}$$

where $\mathbf{F}_i$ is the net force on the ith object and $\mathbf{p}_i$ is its momentum. Since Newton's law is only valid for constant mass systems, it is usually represented as

$$\mathbf{F}_i = m_i \frac{d\mathbf{v}_i}{dt}, \tag{3.2}$$

where m_i is the ith objects mass and $\mathbf{v}_i$ its velocity.

MD simulations apply Newton's second law to a system of interacting atoms, evolving the system state by integrating Equation (3.2) for all of the atoms in the system. The equations of motion are stated as a set of pairs of coupled first-order ordinary differential equations

$$\begin{aligned} \mathbf{F}_i &= m_i \frac{d\mathbf{v}_i}{dt}, \\ \mathbf{v}_i &= \frac{d\mathbf{x}_i}{dt}, \end{aligned} \tag{3.3}$$

where $\mathbf{x}_i$ is the position of the ith particle. The force on the ith particle is given by the negative gradient of the potential energy

$$\mathbf{F}_i = -\frac{\partial V(\{\mathbf{x}_i\}_N)}{\partial \mathbf{x}_i}. \tag{3.4}$$

where the potential energy, V, in MD simulations is a function of the positions of the N particles in the system, $\{\mathbf{x}_i\}_N$.

3.1.1.2 Numerical integration of the equations of motion

For a system of more than two objects, there is no analytical solution to the equations of motion, so the evolution of the system must be evaluated using a numerical integration scheme. The numerical schemes used in MD simulations are usually selected based on their ability to satisfy time-reversibility (integrating forward n steps and then backward n steps returns the system to its initial state), their symplectic nature (they conserve a

slightly perturbed energy of the underlying system) and their high-order numerical error with low computational complexity. One such scheme, the one that we employ in our MD simulations in this chapter, is the leapfrog method.

The leapfrog method can be derived as follows. Firstly, we make a Taylor approximation for the position $\mathbf{x}_i$ of a particle at time $t + \Delta t$, giving

$$\mathbf{x}_i(t + \Delta t) = \mathbf{x}_i(t) + \frac{d\mathbf{x}_i(t)}{dt}\Delta t + \frac{1}{2}\frac{d^2\mathbf{x}_i(t)}{dt^2}\Delta t^2 + \mathcal{O}(\Delta t^3). \tag{3.5}$$

Next, we make a similar approximation for the velocity $\mathbf{v}_i$ of a particle at time $t + \frac{1}{2}\Delta t$, giving

$$\mathbf{v}_i(t + \frac{1}{2}\Delta t) = \mathbf{v}_i(t) + \frac{1}{2}\frac{d\mathbf{v}_i(t)}{dt}\Delta t + \frac{1}{8}\frac{d^2\mathbf{v}_i(t)}{dt^2}\Delta t^2 + \mathcal{O}(\Delta t^3). \tag{3.6}$$

Since $\mathbf{v}_i(t) = \frac{d\mathbf{x}_i(t)}{dt}$, we can substitute Equation (3.6) into Equation (3.5), giving

$$\mathbf{x}_i(t + \Delta t) = \mathbf{x}_i(t) + \mathbf{v}_i(t + \frac{1}{2}\Delta t)\Delta t + \mathcal{O}(\Delta t^3). \tag{3.7}$$

Then subtracting the reverse approximation of Equation (3.6), given by

$$\mathbf{v}_i(t - \frac{1}{2}\Delta t) = \mathbf{v}_i(t) - \frac{1}{2}\frac{d\mathbf{v}_i(t)}{dt}\Delta t + \frac{1}{8}\frac{d^2\mathbf{v}_i(t)}{dt^2}\Delta t^2 - \mathcal{O}(\Delta t^3), \tag{3.8}$$

from Equation (3.6), we get

$$\mathbf{v}_i(t + \frac{1}{2}\Delta t) = \mathbf{v}_i(t - \frac{1}{2}\Delta t) + \frac{d\mathbf{v}_i(t)}{dt}\Delta t + \mathcal{O}(\Delta t^3). \tag{3.9}$$

By combining Equations (3.2), (3.4) and (3.9), we get

$$\mathbf{v}_i(t + \frac{1}{2}\Delta t) = \mathbf{v}_i(t - \frac{1}{2}\Delta t) - \frac{1}{m}\frac{\partial V(\{\mathbf{x}_i(t)\}_N)}{\partial \mathbf{x}_i}\Delta t + \mathcal{O}(\Delta t^3). \tag{3.10}$$

Equations (3.7) and (3.10) are the coupled equations that are solved in the leapfrog method. From these equations it is clear why it is called the leapfrog method because we

are evaluating the positions and velocities at interleaved time-steps.

3.1.2 Understanding protein-protein interactions using molecular dynamics simulations

Atomistic MD simulations typically use a time-step length in the range of femtoseconds, which make them very computationally demanding, often infeasible, when studying large systems and/or long time scales. The membrane processes in which we are interested—protein association-dissociation and protein diffusion—occur on the time scale of microseconds, and as such are currently not easily accessible to atomistic simulations. To remedy this, enabling simulation of processes over timescales of 10–100 μs, which is appropriate for characterising protein association-dissociation events, we can try to reduce the dimensionality of the system. Using a system with a reduced number of degrees of freedom reduces the computational requirements of the numerical integration scheme. Various methods of dimensionality-reduction exist. For instance, by representing the bonded interactions between atoms using functions of the bond angles and dihedral angles, as in torsion-angle dynamics (Stein et al. 1997), the complexity of calculating the potential functions is reduced. It is also possible to reduce the complexity of a simulated system by using translations and rotations of clusters of atoms as rigid bodies, although such cluster moves can only be implemented in Monte Carlo simulations of molecular systems because the size and frequency of the moves made will have complicated dependencies on the size and shape of the cluster.

To reduce the dimensionality of our system we will use a coarse-graining procedure, in which the atomistic system is mapped to a coarse representation where, on average, four heavy atoms are mapped to one coarse-grained particle. The procedure that we use is an extension of the Martini coarse-grained model (Marrink, Risselada, et al. 2007; Monticelli et al. 2008). As well as reducing the number of particles in the system, the Martini model also uses simplified and truncated forms of the potential. The non-bonded

interactions are represented by a Lennard-Jones 12-6 potential of the form

$$V_{ij} = 4\epsilon_{ij}\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right], \tag{3.11}$$

where V_{ij} is the potential between particles i and j separated by a distance r_{ij}; ϵ_{ij} ranges between 5.6 kJ mol^{-1}, for strongly polar particles, and 2.0 kJ mol^{-1}, for polar and apolar particles; and σ is 0.47 nm. When the particles i and j are charged, the non-bonded interaction also includes a Coulomb-type interaction energy. Both the Lennard-Jones and Coulomb potentials have infinite range, so they are implemented in a shifted form in the Martini model, with a cut-off distance of 1.2 nm. The bonded interactions between the atoms, for instance the atoms in a proteins, are represented by harmonic interactions for bond length and angle and multimodal dihedral potentials. An elastic network of harmonic restraints is also applied to the proteins in order to preserve their secondary structure; the secondary structure would be stabilized by directional hydrogen bonds in an atomistic representation, but these are not represented in Martini's coarse-grained model (Marrink and Tieleman 2013).

These coarse-grained models have been used to study protein-protein interactions by looking for protein-protein interfaces in self-assembly simulations (Chng et al. 2011; Prakash et al. 2011), in which many repeats of long coarse-grained simulations are combined to give an insight into the possible binding configurations of the proteins. The approach of using self-assembly simulations to study protein-protein interactions is not always sufficient: Arnarez et al. (2013) found that simulations of 100 μs were not long enough to sample thoroughly some of the association-dissociation events.

3.1.3 Using the potential of mean force to describe protein-protein interactions

Another approach for investigating the nature of protein-protein interactions is to calculate the potential of mean force (PMF), which is a measure of the free energy of the interaction as a function of some reaction coordinate, or coordinates (Torrie et al. 1977; Roux 1995). This is a useful quantity to calculate because by understanding how the free energy changes for a particular reaction coordinate relevant to an aspect of the system that we are trying to understand, we are able to surmise how the dynamics of the system along that reaction coordinate progress. The simulations used to calculate the PMF are subject to the same sampling limitations as the self-assembly approaches discussed above. Specifically relevant to the investigation of membrane proteins, there is the issue of the slow lipidation and delipidation of the protein-protein interface. Total sampling times required to achieve convergence of the PMF are often in the range 0.1–1 ms (for instance D. Sengupta et al. (2010) performed 21 simulations of up to 8 μs to calculate the PMF for association of glycophorin A), which again necessitate the use of coarse grained models. Much work investigating protein-protein interactions has focussed on the calculation of PMFs for α-helical peptides and proteins (Hénin, Pohorille, et al. 2005; Janosi et al. 2010; D. Sengupta et al. 2010; Johnston et al. 2011; Benjamini et al. 2013). Periole, Knepp, et al. (2012) undertook one of the most advanced studies on protein-protein interactions for an α-helical protein, rhodopsin, where the PMF was calculated for various relative protein orientations, which were identified as interesting candidates from a set of large self-assembly simulations. They found that there was a strong interaction in certain orientations, and almost no interaction in others. This suggested specific binding modes of rhodopsin, which enable the formation of chain-like structures as observed in rod membranes. However, to date, there has not been much work on calculating PMFs of β-barrel proteins. Casuso et al. (2012) calculated the PMF for the association of OmpF trimers; the PMF was used to explain the observed clustering behaviour of the proteins

using high-speed atomic force microscopy. Various intermediate states were identified corresponding to relative orientations with differing amounts of protein contact, with the orientations having greater contact occurring with a lower free energy. The validity of coarse-grained models of protein systems to study protein-protein interactions, compared to using an atomistic representation, was investigated by May et al. (2013) for two soluble proteins. Looking at the interaction energy of the TCR–pMHC complex and the MP1–p14 scaffolding complex, they found that the free energy of association was at least as good using the coarse-grained Martini model as when using a fully-atomistic representation in comparison to experimentally determined values. There have not, however, been any validations of coarse-grained protein-protein interaction calculations using membrane proteins, where it is likely that different interactions will play an important role.

Through studies like these, where the PMF for a given protein or pair of proteins is calculated, we are able to characterize a specific protein-protein interaction in a manner that is not possible using only self-assembly simulations. We can investigate the interaction in detail, enabling a full characterization of the proteins' behaviour along some predetermined reaction coordinate. Such an assessment is useful not only for investigating the kinetics of the system, but also for informing us about the nature of its dynamics.

3.1.4 The structure of this chapter

This chapter describes the calculation of PMFs for the bacterial outer membrane protein (OMP) NanC, whose crystal structure is shown in Figure 3.1 and was calculated with a resolution of 2 Å (Wirth et al. 2009). NanC is a β-barrel protein, acting as a porin for the transport of sialic acid into *E. coli*. The pore is lined by two strings of basic residues which face each other, which is assumed to facilitate the diffusion of oligosaccharides (Wirth et al. 2009) and has an overall charge of -1. It is a relatively-featureless protein, which does not have any specific binding modes, which makes it an ideal candidate to investigate more general properties of protein-protein interactions in the bilayer. Firstly, we introduce

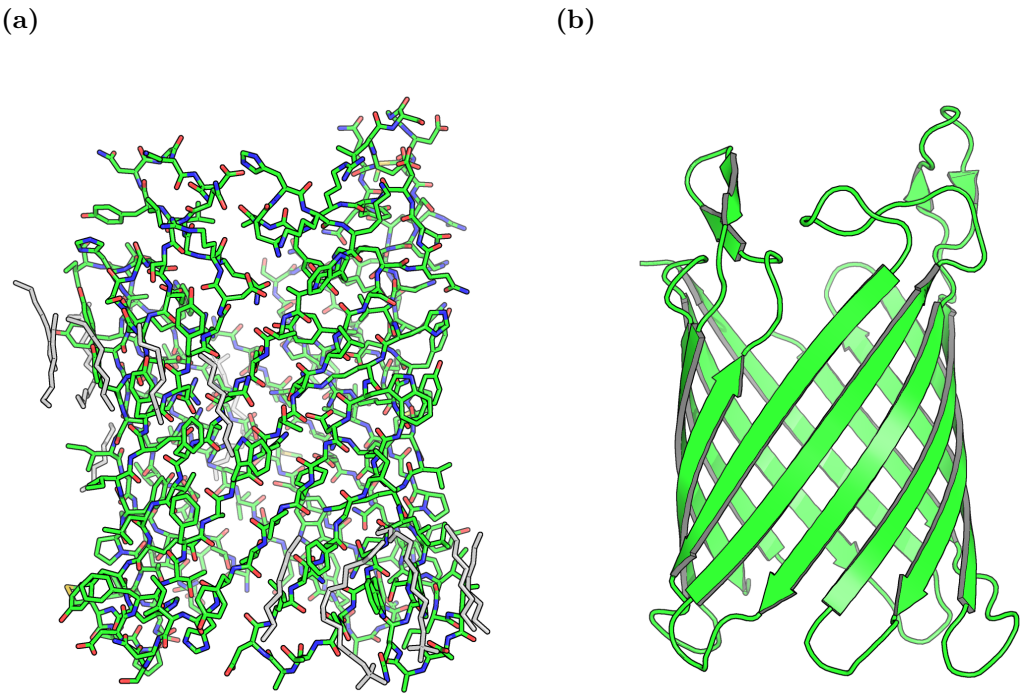

Figure 3.1: **(a)** Crystal structure of NanC obtained from from the Protein Data Bank (Wirth et al. 2009). Carbon atoms are shown in green, oxygen in red, nitrogen in blue, and sulphur in yellow. Atoms in the structure file that do not belong to the protein are shown in grey and consist of detergent molecules from the purification process. **(b)** The secondary structure of NanC. It is a β-barrel constructed of 12 anti-parallel β-sheets.

the theory behind free energy calculations and the methods that we use throughout the chapter. Next, we illustrate, by example, the use of umbrella sampling to calculate the PMF for the association of NanC in the bilayer. Then we introduce a novel method for investigating the PMF for specific protein-protein interfaces using a restrained system. Finally, we look at some interesting features of these restrained PMFs and the relation to the biophysical processes that may cause them.

3.2 Theoretical background to potential of mean force calculations

The PMF is used to characterize the change in a system's free energy (the total energy available to do work) as one or more reaction coordinates change. With knowledge of the changes in free energy that result from a particular process we are able to say whether or not that process will occur spontaneously or would require the input of energy. The reaction coordinates could be any combination of the internal coordinates of the system, for instance, the angle of a peptide bond in a protein, the distance between a ligand and its binding site on a protein, or, as in our case, the distance between two proteins in a lipid bilayer. To characterize the free energy landscape of dimerization for two proteins, we will calculate the PMF as a function of the inter-protein separation.

3.2.1 Thermodynamic ensembles

The state of a thermodynamic system (a system defined as a volume in space separated from its surroundings by a boundary) is completely described by the state of all its constituent particles: the microstate of the system. For a given system, we can define macroscopic observables—such as the volume, pressure or temperature—that describe the macroscopic behaviour of the system. For any particular set of values of these macroscopic observables, the system could be in one of many microstates. The macrostate of a system is described by a distributions of microstates, called a statistical ensemble.

The macroscopic variables that describe the state of the system are known as state variables. A system is said to be in thermodynamic equilibrium when its macroscopic state, described by its state variables, no longer changes with time. A system that is in thermodynamic equilibrium can be described by a thermodynamic ensemble. Various thermodynamic ensembles exist, the properties of each are determined by the nature of the boundary between the system its surroundings. For instance, the microstates

of an isolated system (a system with fixed size and which does not exchange energy or particles with its surroundings) is described by the microcanonical ensemble. If an isolated system is put in contact with an external heat bath, so that it can exchange energy with its surroundings in order to keep constant temperature, then the distribution of its microstates will be described by the canonical ensemble. The canonical ensemble has constant particle number, N, volume, V, and temperature, T, and so is sometimes referred to as the NVT ensemble.

For a system described by the canonical ensemble, with discrete microstates, the probability, P_s, of being in a given microstate, s, is given by

$$P_s = \frac{e^{-\beta E_s}}{\sum_s e^{-\beta E_s}}, \tag{3.12}$$

where E_s is the energy of the microstate and the inverse-temperature, β, is defined as

$$\beta = k_B T, \tag{3.13}$$

where T is the temperature and k_B is Boltzmann's constant. The normalization factor for the probability in Equation (3.12) is called the partition function and is the sum of Boltzmann factors for the microstates of the system. For a discrete system described by the canonical ensemble, the partition function is given by

$$Z = \sum_s e^{-\beta E_s}, \tag{3.14}$$

where the sum is taken over all microstates. When the microstates of a system are continuous, for instance when the energy of a given state is described by a Hamiltonian, $\mathcal{H}$, that is a function of the positions, $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, and momenta, $\mathbf{p} = \{\mathbf{p}_1, \dots, \mathbf{p}_N\}$,

of the N particles in the system, the partition function is given by

$$Z = \frac{1}{h^{3N}N!} \iint e^{-\beta\mathcal{H}(\mathbf{x},\mathbf{p})} d\mathbf{x} d\mathbf{p}, \tag{3.15}$$

where h is a constant with dimensions of action, usually taken to be Plank's constant (since this is the value that is obtained by taking the classical limit of the quantum description of the canonical ensemble) (Huang 1987).

3.2.2 Defining the potential of mean force in the isothermal-isobaric ensemble

Throughout this chapter we work with the isothermal-isobaric thermodynamic ensemble. This ensemble is similar to certain experimental conditions, those that involve systems that are kept at constant temperature and pressure without exchange of particles. An isothermal-isobaric system differs from a canonical system in that its size changes in order to maintain a constant pressure, p, which it does by performing work on the system's surroundings: isothermal-isobaric systems are also referred to as NpT systems.

The partition function for the isothermal-isobaric ensemble can be obtained from the partition function for the canonical ensemble, given in Equation (3.15). We wish to change the canonical partition function's dependency on volume to a dependency on pressure to arrive at the partition function for the isothermal-isobaric ensemble. Such a change between conjugate variables, from extrinsic to intrinsic, for thermodynamic equations of state can be achieved using the Legendre transform (Huang 1987). This means that our isothermal-isobaric partition function has the form

$$\Xi(N, p, T) = C \int Z(N, V, T) e^{-\beta p V} dV, \tag{3.16}$$

where C is a normalization factor, which we don't need to determine for the following discussion.

The free energy of a system is the total energy available to the system that is able to be used to do work (Blundell et al. 2010). The free energy is a function of the state variables of the system, and as such there are multiple expressions for the free energy, each of which is a function of different state variables. For a system described by a given ensemble, it will be a particular free energy function that is minimized when the system is at equilibrium. For instance, for NpT systems it is the Gibbs free energy, which is a function of pressure, temperature and particle number, that is minimized when the system is in equilibrium (Gibbs 2014). The Gibbs free energy, $G(N, p, T)$, can be expressed in terms of the partition function of the isothermal-isobaric ensemble, $\Xi(N, p, T)$ using

$$G(N, p, T) = -\frac{1}{\beta} \ln \Xi(N, p, T). \tag{3.17}$$

The PMF is a measure based on the free energy of the system. Central to the notion of the PMF is the definition of a reaction coordinate. In a system described by the positions, $\mathbf{x}$, and momenta $\mathbf{p}$, of the N constituent atoms, the reaction coordinate, ξ, can then be described by some function, $\hat{\xi}$ of the positions of the atoms

$$\xi = \hat{\xi}(\mathbf{x}), \tag{3.18}$$

which means that the system will be at a given value of the reaction coordinate when its positions satisfy the above equation.

When determining the PMF we are interested in the free energy of the system for a given value of the reaction coordinate, $\hat{G}(N, p, T, \xi)$, so we need to calculate the partition function similarly. To do so we need only to include contributions to the partition function from states that have a specific value of the reaction coordinate, which we can do by introducing a delta function to the integral of the partition function, resulting in

$$\hat{\Xi}(N, p, T, \xi) = \frac{C}{h^{3N} N!} \int e^{-\beta p V} dV \iint \delta(\xi - \hat{\xi}(\mathbf{x})) e^{-\beta \mathcal{H}(\mathbf{x}, \mathbf{p})} d\mathbf{x} d\mathbf{p}. \tag{3.19}$$

The delta function will have the effect of removing contributions to the partition function from states where the value of the reaction coordinate is not ξ.

The probability density for a continuous system in a state with positions $\mathbf{x}$ and momenta $\mathbf{p}$ (the analogue of Equation (3.12)) is given by

$$\pi(\mathbf{x},\mathbf{p}) = \frac{e^{-\beta\mathcal{H}(\mathbf{x},\mathbf{p})}}{\iint\ e^{-\beta\mathcal{H}(\mathbf{x}',\mathbf{p}')}d\mathbf{x}'d\mathbf{p}'}, \tag{3.20}$$

and is called the distribution function. We can include the reaction coordinate explicitly in the expression for the distribution function using $\hat{\pi}(\mathbf{x},\mathbf{p},\xi) = \delta(\xi - \hat{\xi}(\mathbf{x}))\pi(\mathbf{x},\mathbf{p})$, which requires no normalization factor because $\int\ \delta(y)dy = 1$. We can thus obtain an expression for the distribution function as a function of the reaction coordinate, $\rho(\xi)$, by integrating over the positions and momenta, giving

$$\begin{aligned}\rho(\xi) &= \iint\ \hat{\pi}(\mathbf{x},\mathbf{p},\xi)d\mathbf{x}d\mathbf{p}, \\ &= \frac{\iint\ \delta(\xi - \hat{\xi}(\mathbf{x}))e^{-\beta\mathcal{H}(\mathbf{x},\mathbf{p})}d\mathbf{x}d\mathbf{p}}{\iint\ e^{-\beta\mathcal{H}(\mathbf{x},\mathbf{p})}d\mathbf{x}d\mathbf{p}}.\end{aligned} \tag{3.21}$$

Combining Equations (3.16), (3.19) and (3.21) we get

$$\rho(\xi) = \frac{\hat{\Xi}(N,p,T,\xi)}{\Xi(N,p,T)}. \tag{3.22}$$

Using this expression in the reaction-coordinate dependent expression for the free energy, similar to Equation (3.17), gives

$$\begin{aligned}\hat{G}(N,p,T,\xi) &= -\frac{1}{\beta}\ln\hat{\Xi}(N,p,T,\xi), \\ &= -\frac{1}{\beta}\ln\rho(\xi) + \frac{1}{\beta}\ln\Xi(N,p,T).\end{aligned} \tag{3.23}$$

Using Equation (3.23) we are able to calculate the changes in the free energy at different points along the reaction coordinate by calculating the changes in the distribution function,

$\pi(\xi)$. The free energy as a function of the reaction coordinate is the PMF, which we will represent using $\hat{G}(N, p, T, \xi) \rightarrow \mathcal{W}(\xi)$. By evaluating the differences between different points along the reaction coordinate we arrive at the familiar expression for the PMF of

$$\mathcal{W}(\xi) = \mathcal{W}(\xi^*) - \frac{1}{\beta} \ln \left[\frac{\rho(\xi)}{\rho(\xi^*)} \right], \tag{3.24}$$

where ξ^* is an arbitrary value of the reaction coordinate.

3.2.3 Calculating potentials of mean force

To calculate the PMF through simulation, we can approximate the reaction coordinate-dependent probability density, $\rho(\xi)$, by sampling configurations along the reaction coordinate. Using an unbiased simulation we would theoretically be able to calculate the PMF using either of the above methods, but for any sufficiently complicated system the calculation would not converge because the sampling would be poor. In order for the PMF calculation to converge, we are required to obtain adequate sampling for configurations corresponding to all the regions of the reaction coordinate in which we are interested. There are many approaches for biasing a simulation to encourage sampling of energetically unfavourable states of the reaction coordinate (states that would not be sampled sufficiently in a free simulation). For example: metadynamics, where a series of Gaussian potentials are gradually added to bias an unrestrained simulation (Laio et al. 2002; Barducci et al. 2008); steered molecular dynamics, where a non-equilibrium force is used to steer a simulation between two states and Jarzynski's equality is used to obtain the PMF (Jarzynski 1997); and applying an adaptive biasing force, where a continually-changing force is applied to converge on the PMF along some reaction coordinate (Darve et al. 2001; Hénin and Chipot 2004). For this work we have chosen to employ umbrella sampling to encourage a thorough exploration of the entirety of the reaction coordinate, and the weighted histogram analysis method to obtain the PMF from

the umbrella sampling simulations. This method is explained in the following section.

3.2.4 Umbrella sampling and the weighted histogram analysis method

Umbrella sampling employs a series of simulation windows, in each of which the system is restrained to some small region of the reaction coordinate using a harmonic (umbrella) potential (Torrie et al. 1977), defined by

$$w(\mathbf{x}) = \frac{k_t}{2}\left(\hat{\xi}(\mathbf{x}) - \xi_i\right)^2, \tag{3.25}$$

where ξ_i is the value of the reaction coordinate for simulation window i, and k_t is the force constant of the harmonic restraint. For each position of the harmonic restraint, independent simulations are run, in which the system will explore the region close to the centre of the restraining potential. The positions of the sampling windows (ξ_i) and the strength of the potential (k_t) need to be judiciously selected to ensure that there is smooth sampling of the entire reaction coordinate and that each point is sampled by multiple simulation windows. This approach ensures that we are sampling the entire region of the reaction coordinate, but makes no adjustments for the amount of sampling required in a given region to obtain a converged PMF.

In a membrane system containing two proteins, where we wish to calculate the PMF as a function of the separation of the proteins, we would place the umbrella potentials at various positions along the line connecting the proteins' centres of mass. The umbrella potentials would be distributed along the entire range of the reaction coordinate in which we are interested, starting from a small inter-protein separation, where the proteins would be in contact, up to a large separation, where the proteins no longer affect each other.

From each of the multiple independent simulations run for the series of positions and strengths of the umbrella potential we obtain a distribution describing the sampling of the reaction coordinate which is an estimator of the probability density of the biased

system in the region of the restraining potential. In the case of a membrane protein system, these will be distributions of the distance between the proteins' centers of mass, biased by the effect of the umbrella potential of the particular simulation window.

The most common approach for obtaining a PMF from these sampling distributions is to use the weighted histogram analysis method (WHAM). The WHAM procedure divides the entire reaction coordinate into a series of B bins, into which we count the number of samples from each of the S simulation windows. Following the maximum likelihood derivation of Bartels et al. (1997), we explain the WHAM procedure below. The number of samples in the jth bin from the ith simulation window is given by n_{ij}. We choose B to be sufficiently large, such that the set of probabilities that an unbiased sample of the reaction coordinate is in each of the bins, $\{P_j\}$, which describes the shape of the histogram, is a good estimator for the probability density of the reaction coordinate, $\rho(\xi)$. In the ith simulation, the probability of a reaction coordinate sample being in bin j is given by

$$P_{ij} = f_i c_{ij} P_j, \tag{3.26}$$

where the factor $c_{ij} = e^{-\beta w_i(\xi)}$ accounts for the effect of the biasing potential, $w_i(\xi)$, and can be justified by relating the biased probability density, $\rho_i(\xi)$, to the unbiased probability density, $\rho(\xi)$,

$$\rho_i(\xi) = \frac{\iint \ \delta(\xi - \hat{\xi}(\mathbf{x}))e^{-\beta(\mathcal{H}(\mathbf{x},\mathbf{p})-w_i(\xi))}d\mathbf{x}d\mathbf{p}}{\iint \ e^{-\beta(\mathcal{H}(\mathbf{x},\mathbf{p})-w_i(\xi))}d\mathbf{x}d\mathbf{p}} = \frac{e^{-\beta w_i(\xi)}}{\left\langle e^{-\beta w_i(\xi)} \right\rangle}\rho(\xi); \tag{3.27}$$

and f_i in Equation (3.26) is a normalization factor given by

$$f_i = \frac{1}{\sum_j c_{ij} P_j}. \tag{3.28}$$

The histogram counts are assumed to follow a multinomial distribution and, given the probabilities $\{P_{ij}\}$ for the ith simulation, the likelihood of the histogram counts $\{n_{ij}\}$ is

given by

$$P\left(n_{i1},\ldots,n_{iB}|P_{i1},\ldots,P_{iB}\right)=\frac{\left(\sum_j n_{ij}\right)!}{\sum_j n_{ij}!}\prod_j (P_{ij})^{n_{ij}}. \tag{3.29}$$

The total likelihood for observing the counts in all of the S simulations is given by the product of the likelihoods for the individual simulations

$$P\left(\{n_{ij}\}|\{P_j\}\right)\propto\prod_i\prod_j (P_{ij})^{n_{ij}}. \tag{3.30}$$

Taking the logarithm of this gives

$$\ln\left[P\left(\{n_{ij}\}|\{P_j\}\right)\right]=\sum_i\sum_j n_{ij}\ln P_{ij}+C, \tag{3.31}$$

where C is a constant. Substituting for P_{ij} using Equation (3.26) gives

$$\ln\left[P\left(\{n_{ij}\}|\{P_j\}\right)\right]=\sum_i N_i\ln f_i+\sum_j M_j\ln P_j+C', \tag{3.32}$$

where N_i is the total number of samples from the ith simulation, M_j is the number of samples in the jth bin from all simulations, and further constants have been combined with C in C'. The maximum likelihood estimates of $\{P_j\}$ can be found by firstly differentiating the above to give

$$\frac{\partial}{\partial P_j}\ln\left[P\left(\{n_{ij}\}|\{P_j\}\right)\right]=-\sum_i N_i c_{ij} f_i+\frac{M_j}{P_j}, \tag{3.33}$$

and then equating this with zero to get an expression for each of the $\{P_j\}$, which is given by

$$P_j=\frac{M_j}{\sum_i N_i c_{ij} f_i}. \tag{3.34}$$

Equations (3.28) and (3.34) are a set of coupled equations, which are solved iteratively until converged estimates for $\{P_j\}$ are obtained. Using the values obtained for $\{P_j\}$ as

an approximation to $\rho(\xi)$, we construct an estimate for the PMF using Equation (3.24) (on page 31).

3.3 Calculating the PMF for association of NanC

3.3.1 Using a coarse-grained model of a bilayer

3.3.1.1 The coarse-grained model

To be able to calculate the PMF for protein association in a bilayer system, we need to perform molecular dynamics simulations of a system with hundreds of lipids and multiple proteins for $\sim$100μs. Using a fully-atomistic model of the system would not be computationally tractable and so we must turn to a coarse-grained representation of the system. The reduction in the number of constituent particles and a simplification of the interaction potentials, where a cut-off is implemented for long range interactions, results in a speed up of multiple orders of magnitude.

The coarse-grained model used throughout this chapter is a modification of the Martini model (Marrink, Risselada, et al. 2007; Monticelli et al. 2008), in which approximately four non-hydrogen atoms are represented by a singe coarse-grained particle (Bond and Sansom 2006; Bond, Holyoake, et al. 2007; Bond, Wee, et al. 2008). The Martini model was initially developed to simulate systems containing lipids and surfactants. This model was modified by Bond and Sansom (2006) to include a coarse-grained representation for proteins. A protein is represented by a single coarse-grained particle for each amino acid backbone atom group and the amino acid side chains are included by the addition of up to two extra coarse-grained particles.

3.3.1.2 From the crystal structure to a coarse-grained model of NanC

The coarse-grained NanC model used in this work was created by Goose et al. (2013). To create the model, the crystal structure of NanC (2WJQ) was first obtained from the

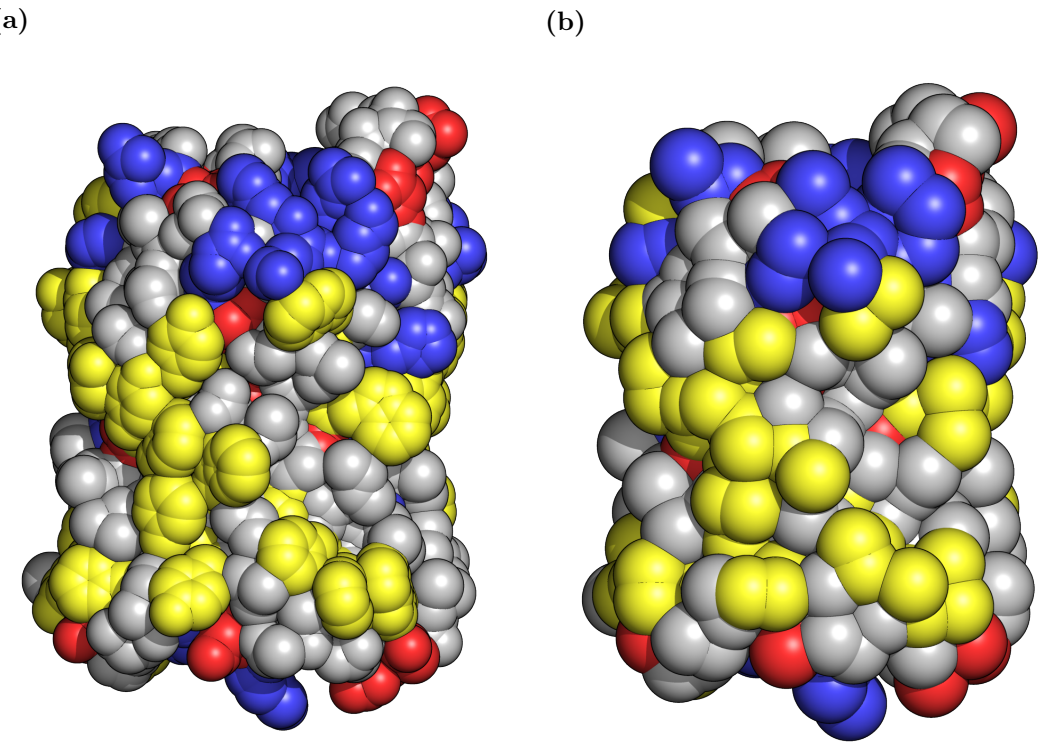

Figure 3.2: (a) Atomistic model of NanC in the plane of the membrane. Acidic residues are shown in red, basic residues in blue, aromatic residues in yellow, and neutral residues in grey. **(b)** Coarse-grained model of NanC, in the plane of the membrane. Acidic residues are shown in red, basic residues in blue, aromatic residues in yellow, and neutral residues in grey.

Protein Data Bank (`http://www.rcsb.org`). The crystallographic water molecules and other non-protein molecules were removed from the structure and then the incomplete loop region, from residue 43 to residue 52, was completed using *Modeller* (9v8) (Fiser et al. 2003). The complete atomistic model of NanC is shown in Figure 3.2(a). The coarse-grained model of NanC, following the coarse-graining process outlined in Bond and Sansom (2006), is shown in Figure 3.2(b).

3.3.2 Molecular dynamics simulations

The MD simulations performed in this chapter use the *GROMACS* simulation software package using simulation parameters outlined in Table 3.1.

Parameter	**Value**
Integrator	Leapfrog
Time-step	40 fs
Thermostat	Berendsen
Temperature groups	1. protein, 2. lipid, 3. solvent and ions
Thermostat time-constant	1 ps
Reference temperature	310 K
Barostat	Berendsen
Pressure coupling	semi-isotropic (separate barostats perpendicular to and in the plane of the membrane)
Compressiblity	5×10^{-6} bar^{-1}
Barostat time-constant	1.0 ps
Reference pressure	1.0 bar
Umbrella samping	
Coarse-grained force field	modified Martini (Bond, Wee, et al. 2008; Monticelli et al. 2008)
Translational protein restraint, k_t	1000 kJ mol^{-1} nm^{-2}
Strong translational protein restraint (detailed simulation windows), k_t	10 MJ mol^{-1} nm^{-2}
Rotational protein restraint, k_r	1000 kJ mol^{-1} nm^{-2}

Table 3.1: Simulation parameters used for MD simulations throughout this chapter.

3.3.3 Calculating a PMF for the pairwise association of NanC

To demonstrate the umbrella sampling procedure, we will calculate the PMF for the pairwise association of two NanC proteins in a lipid bilayer constructed from 424 coarse-grained POPE molecules. The bilayer contains the two coarse-grained NanC proteins. The bilayer is solvated and the entire system is made neutral with the addition of counter-ions. Details of the methods used to construct the system used for our molecular dynamics simulations are given in Appendix A (on page 193). To ensure we obtain a converged PMF from the umbrella sampling process we must calculate well-sampled distributions from each of the simulation windows (the distributions must be smoothly varying), and for the WHAM procedure to converge the distributions from multiple adjacent simulation windows must overlap.

3.3.3.1 Sampling window distribution

To ensure that the distributions calculated from adjacent simulation windows overlap sufficiently we can vary both the inter-window spacing and the umbrella force constant. Increasing the force constant will encourage the system to explore a smaller region of the reaction coordinate, which is essential to be able to resolve more detailed features in the PMF. However, if we use an increased force constant to restrict sampling to a smaller region, we must then decrease the inter-window separation in order to maintain sufficient overlap between the distribution from adjacent simulation windows.

The initial window separation used when calculating this PMF was 0.1 nm, with a restraining force constant of 1000 kJ mol^{-1} nm^{-2}, as used in many other studies, for instance that of D. Sengupta et al. (2010). After running production simulations using these parameters, we are able to assess any regions of the reaction coordinate that require more targeted sampling with a smaller window separation and increased force constant. These are usually regions with detailed features, which we wish to resolve more clearly, or regions with steep changes in potential, which are not sampled thoroughly.

We can investigate this by analysing the amount of overlap between adjacent window distributions. In Section 3.5.2 (on page 50) we apply just such a refinement to our more advanced umbrella sampling method. The starting configuration for each window was obtained following the methods described in Appendix A.iii (on page 195).

3.3.3.2 Umbrella sampling simulations

Production simulations of 2 μs were run for four different starting configurations in each umbrella sampling window, resulting in 8 μs of production simulation for each window. The simulation windows were distributed evenly from an inter-protein separation of 2.8 nm to 8 nm, resulting in a total of 424 μs of production simulation.

From the distributions obtained from the production simulations the WHAM process constructs the PMF by optimally joining the sections of the unbiased distribution function. This process is be performed using the `g_wham` tool in *GROMACS*. We used `g_wham` with a tolerance of 10^{-6}, which means the procedure is considered to have converged once the changes in the probability of the profile are smaller than 10^{-6}. The PMF created by the WHAM process is shown in Figure 3.3. The PMF has a depth of 70 kJ mol^{-1}, which is of the same order of magnitude to calculations for similar membrane protein systems (Casuso et al. 2012). The WHAM process does not give us an absolute value of the free energy at a given point, but the relative change in free energy in relation to neighbouring points. If we wish to have some absolute measure we must choose an appropriate point on the profile to set to zero. From Figure 3.3 we see that the change in the PMF with increasing separation is negligible at the edge of the simulation region, at an inter-protein separation of 8 nm; this separation is used to define the separation at which the proteins do not experience a change in free energy as a result of the presence of the other protein: the point at which the PMF is set to zero.

The interaction between the NanC proteins is strong. A key factor in determining the association of proteins in the membrane is the amount of hydrophobic mismatch - the

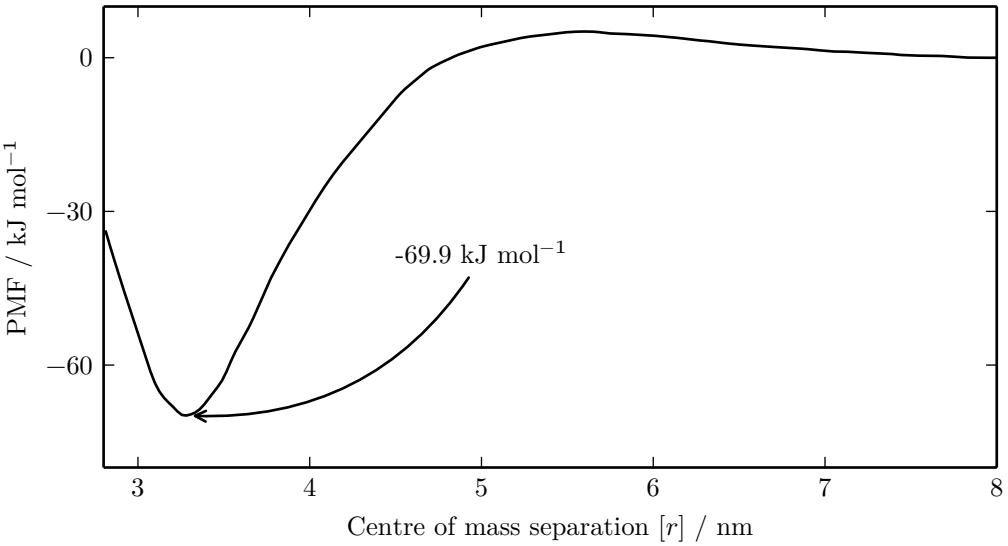

Figure 3.3: A PMF of the association of two NanC proteins in a pure POPE bilayer. As indicated, the depth of the PMF is 69.9 kJ mol^{-1}. At a separation of 8 nm there is negligible change in the PMF with increasing separation.

difference in the size of the hydrophobic region of the proteins and the hydrophobic region of the bilayer. When there is a difference in the sizes of these two hydrophobic regions, the bilayer deforms in order to compensate, which has an associated energetic cost. It known that hydrophobic mismatch can be a driver of proteins aggregation (Parton et al. 2011). It has also been shown that changing the hydrophobic mismatch in a system containing simple WALP peptides can lead to a 50% increase in the energy of their association when calculated sing molecular simulations (Castillo et al. 2013). For proteins that have a larger surface area exposed to the bilayer, the energetic cost of mismatch increases. For the protein OmpF, Casuso et al. (2012) calculated an energy of association of ∼150 kJ mol^{-1}. In our system, these smaller NanC proteins still exhibit a very strong interaction, thought to be dominated by the effects of minimizing hydrophobic mismatch.

3.3.3.3 Assessing the coverage and convergence of the PMF

Distributions were generated from the production simulations for each umbrella sampling window. From the distributions in Figure 3.4 we can see that there is good overlap between the distributions from adjacent windows. This is encouraging as it means that the WHAM process will lead to a converged PMF. It is also useful to note that the spacing of the distributions remains relatively constant; there are not any large gaps, which would indicate a high energy region that was not being adequately sampled. For the sampling windows at larger separations, the distribution for any one window overlaps with the four surrounding distributions; for the sampling windows at smaller separations the amount of overlap increases such that a single distribution overlaps with the surrounding six distributions. Each of the distributions appears to be approximately smoothly varying, with the exception of the few points near the peak, where the width of the peak approaches the same size as the histogram bin separations, thus giving a slightly angular profile.

Another approach to assessing whether there has been adequate sampling of the reaction coordinate is to repeat the calculation of the PMFs using subsets of the production simulation data. We did this by dividing the 2 μs of production simulation data at each simulation window for each configuration into 0.5 μs sections and calculating four separate PMFs. These PMFs are shown overlaid in Figure 3.5, where we see that there is very good agreement when the proteins are far apart, but some variation at small separations. The close agreement when far apart is something we would expect, as with a combined total of 2 μs (including all four starting configurations), we sample a lot of motion of the lipids and the four configurations represent a broad variety of the rotational states of the two proteins. There is reasonably good agreement between the depth of the potential wells in each of the individual PMFs, however, at very small separations, once the PMF begins to increase again, this discrepancy becomes quite noticeable. This is a region of the PMF where the proteins are, in effect, being squashed together by the umbrella potential. The

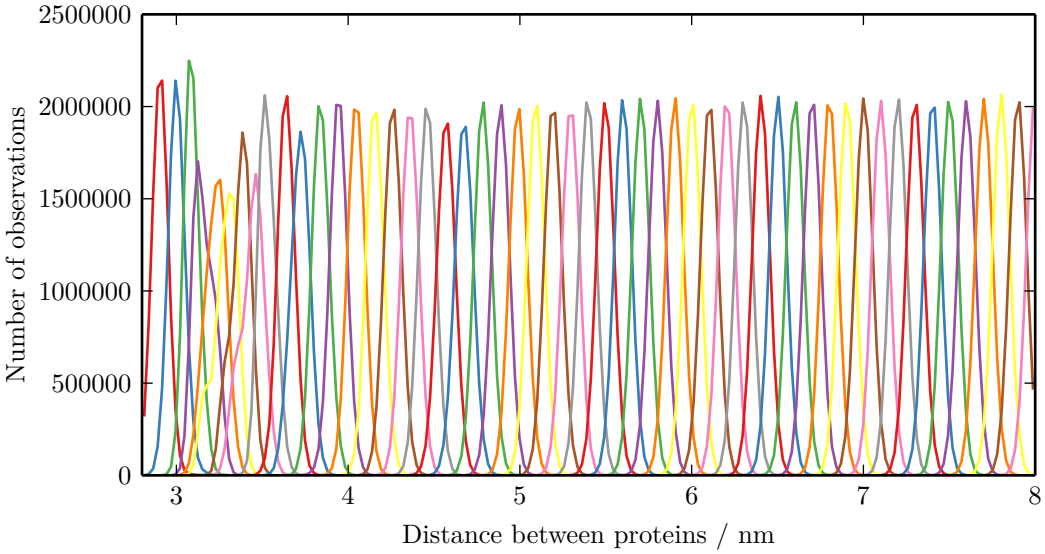

Figure 3.4: Distributions of positions along the reaction coordinate sampled during initial umbrella sampling simulations. The reaction coordinate has been divided into 200 bins to show the sampling in each bin from all the simulation windows. The result of this binning is a slightly angular appearance at the peaks of the histogram traces, where the peak width is approximately the same size as the bin separation.

lipids around the protein in such a scenario will be less mobile and the sampling of the various relative orientations of the proteins will also be reduced, because they are unable to rotate easily against one another. To improve the agreement in this section of the PMF would require a much greater simulation time to ensure that we sampled enough configurations, however, the dynamics of the protein-protein interaction are dominated by the shape of the potential from the minimum in the direction of increasing separation, so this would not be a useful effort.

3.4 Applying rotational restraints to the proteins

The energy landscape of a system is a map of the conformations of the system to the energy of that conformation. We are interested in the free energy landscape of our bilayer system, as this will describe the behaviour of the system by indicating how it transitions between stable or metastable minima and the energy required to do so. In particular we

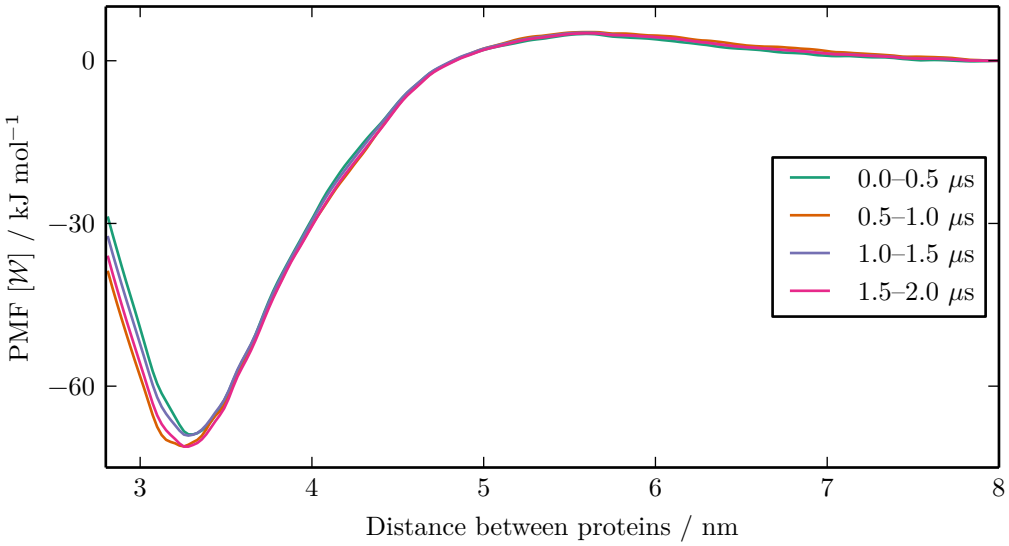

Figure 3.5: A comparison of PMFs calculated from non-intersecting subsets of the production simulation data used to calculate the PMF shown in Figure 3.3.

aim to obtain a projection of the free energy landscape in a particular coordinate system. For the PMF presented in Figure 3.3, we have projected the free energy landscape onto a coordinate system that is represented solely by the separation of the two proteins. This is a useful measure if we are interested only in the kinetics of dimerization for the proteins. However, we also wish to gain more information about the likely routes for dimerization in a more complex coordinate representation; specifically, we wish to understand the role that the relative-orientations of the two proteins have on the free energy of their interaction. For a simple system containing membrane proteins that are rotationally asymmetric about their pore axis, there are complexities to the free energy surface that depend on the relative orientations of the two proteins.

We wish to extend the standard approach to characterizing the dimerization by introducing extra reaction coordinates; in this case we introduce two angular coordinates, which correspond to the relative orientations of the proteins, measured around their pore axis, approximately perpendicular to the membrane. A full characterization of this extended projection of the free energy landscape is currently computationally intractable;

a crude estimate of the amount of simulation time required for a three-dimensional PMF compared to a one dimensional PMF can be obtained by taking the cube of the total length of production simulations used to calculate the PMF presented above, resulting in a total production simulation requirement on the order of $(100\ \mu\text{s})^3 = 1\ s$, which is vastly beyond the times that are feasible with current technology. As a result of this, we chose to look at slices through this three-dimensional projection of the free energy surface by calculating PMFs based on a restrained system of two NanC proteins, where the relative orientations have been chosen to represent some interesting extremes of the geometry of their interaction.

3.4.1 Selecting relative orientational configurations

Perpendicular to the plane of the membrane, in a direction parallel to the pore axis, NanC has an approximately elliptical cross-section. This can been seen in Figure 3.6, where the view is taken from the extracellular side of the protein. When selecting relative orientations of the proteins to consider when calculating restrained PMFs, we wished to exploit this shape to investigate the role of NanC's geometry on the PMF.

In a system with two NanCs, the orientation of one protein will determine how much of it could be in contact with the other protein when they are in close proximity. If the protein is oriented so that a less-curved surface is directed towards the other protein, they will be in a state with increased contact. If the protein is orientated so that a more-curved face is directed towards the other protein, it will be in a state with decreased contact. We use this as a geometrical basis to inform our choice of configurations for which we wish to calculate PMFs. We use four relative orientations of two NanC proteins to calculate four different PMFs. These orientations are shown in Figure 3.7 for our coarse-grained model of NanC, where the black arrows are drawn from the centre of the protein, through the C_α coarse-grained particle of the isoleucine residue 209. This vector, when projected onto the plane of the membrane, is used to define the orientation of the

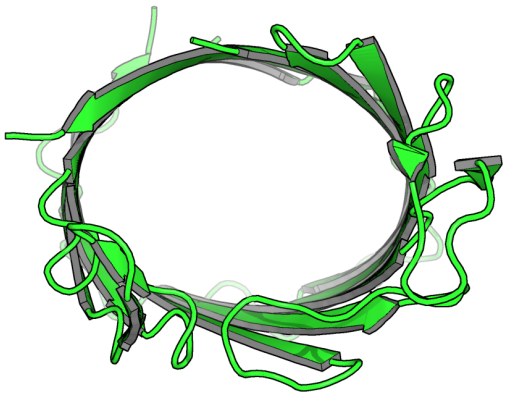

Figure 3.6: NanC secondary structure viewed along the pore axis from the extracellular side. The approximately elliptical nature of its structure is clearly visible. To explore the effect that this shape has on protein-protein interactions, we use the two less-curved faces of the protein (the top and bottom sections of the barrel, as shown here) and the two more-curved faces (the left and right sections of the barrel) as directions in which we wish to calculate restrained PMFs.

protein.

The four orientational configurations chosen represent three different contact regimes of the two proteins. Firstly, in Figures 3.7(a) and 3.7(b) the two NanC proteins are oriented with two less-curved surfaces of the proteins in contact. The other extreme of the contact between the two proteins is shown by the configuration in Figure 3.7(d), where two of the more-curved surfaces of the protein are in contact. Finally, in Figure 3.7(c) we have an intermediate configuration in which a less-curved surface of one protein is in contact with a more-curved surface of the other.

3.4.2 Enforced rotation of proteins in *GROMACS*

To restrain the orientation of the proteins, we have chosen to use the enforced rotation potentials implemented in *GROMACS*. We used these because they enabled rotational control of the protein as a whole, rather than between individual particles (e.g. as done by Periole, Knepp, et al. (2012)). It was this consideration that lead to us using harmonic restraints, rather than constraints. Other simulation software packages allow for the

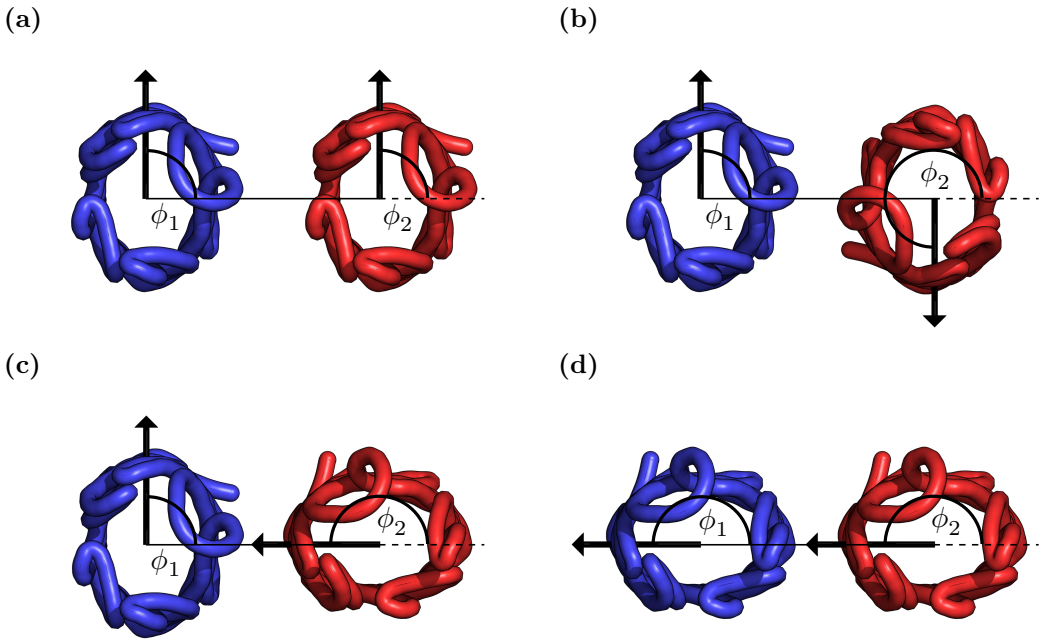

Figure 3.7: The four orientational combinations of the pair of NanC proteins. The red and blue tubes are traces through the C_α particles of each residue, with the black arrows used to define the orientational angle of the proteins; the line is drawn from the proteins' centre of mass through the C_α particle of isoleucine residue 209. The orientations are described by the planar membrane angles ϕ_1 and ϕ_2. **(a)** $(\phi_1, \phi_2) = (90°, 90°)$ **(b)** $(\phi_1, \phi_2) = (90°, 270°)$ **(c)** $(\phi_1, \phi_2) = (90°, 180°)$ **(d)** $(\phi_1, \phi_2) = (180°, 180°)$

use of collective variables which would have facilitated our use of constrained molecular dynamics, but this is not possible using *GROMACS*. Using the enforced rotation enables the orientation of the entire protein to be controlled with minimal perturbation of the internal dynamics of the system (Kutzner et al. 2011). There are various choices of rotational potentials that can be applied in *GROMACS*, but they all are based on the principle of comparing the current state of the restrained group to some reference state of the group. The reference states of our proteins are given by the positions shown in Figure 3.7.

The enforced rotation procedure compares the positions of the particles that are being rotated with reference positions, and applies a restoring force based on some potential function. There are various choices for this potential function, with different properties. The most simple of these is an isotropic potential, where the restoring force is based solely on the distance between the particle's current position, $\mathbf{x}_i$, and its reference position, $\mathbf{y}_i$, and is defined by the potential

$$\Phi_{\text{iso}} = \frac{k_r}{2} \sum_{i=1}^{N} \left[\Omega(t)(\mathbf{y}_i - \mathbf{u}) - (\mathbf{x}_i - \mathbf{u})\right]^2, \tag{3.35}$$

where k_r is the force constant of the potential, the sum is over the N particles being restrained, $\Omega(t)$ is a rotation matrix could be used to provide a steadily rotating reference state, and $\mathbf{u}$ is the position of the rotation axis. The effect of applying another external potential is to further bias the sampling of the reaction coordinate of interest for the umbrella sampling: the inter-protein separation. This effect is accounted for by including an extra term in the biased histogram of probabilities in Equation (3.26) (on page 33) as an extra factor $e^{\Phi_{\text{iso}}/k_B T}$, a factor that depends only on the value of the biasing potential for a given sample.

This simple enforced rotation potential in Equation (3.35) can restrain the positions to some reference state, but there are components of the force that act in all directions. We

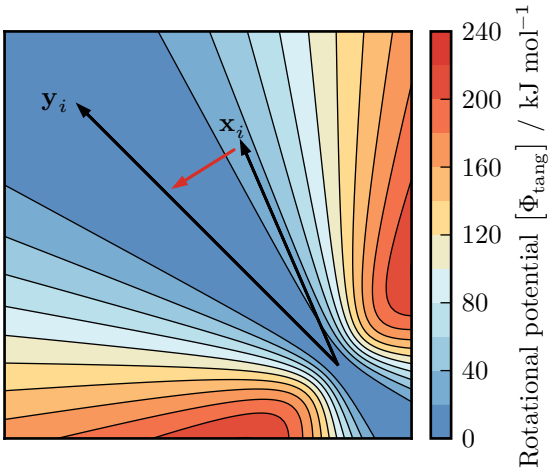

Figure 3.8: The rotational potential, Φ_{tang}, shown in the plane of the membrane. The force exerted on a particle at $\mathbf{x}_i$ is tangential to the axis of rotation, and acts to rotate the particle towards the same angle as its reference position, $\mathbf{y}_i$. The direction of the force is illustrated by the red arrow. Values of $k = 200$ kJ mol^{-1} nm^{-2} and $\epsilon = 0.01$ nm^2 were used in this potential.

wish to restrain the protein in a purely rotational manner; no components of force parallel to the axis of rotation, or in a radial direction perpendicular to the axis of rotation, but only in a tangential direction. In this way we can be sure that there is no effect on the internal structure of the protein. We also want the rotation axis to pass through the protein's centre of mass, parallel to the z-axis of the system, which is approximately perpendicular to the membrane. Kutzner et al. (2011) show that the desired force can be generated if based on a potential of the form

$$\Phi_{\text{tang}} = \frac{k_r}{2} \sum_{i=1}^{N} \frac{\left[(\hat{\mathbf{v}} \times (\mathbf{x}_i - \mathbf{x}_c)) \cdot \Omega(t)(\mathbf{y}_i - \mathbf{y}_c)\right]^2}{\left\|\hat{\mathbf{v}} \times (\mathbf{x}_i - \mathbf{x}_c)\right\|^2 + \epsilon}, \tag{3.36}$$

where $\hat{\mathbf{v}}$ is a unit vector in the direction of the rotation axis, $\mathbf{x}_c$ and $\mathbf{y}_c$ are the current and initial positions of the centre of mass, respectively, and ϵ is a small parameter used to avoid a singularity in the potential at the axis of rotation. The form of this potential, in the plane of the membrane, is shown in Figure 3.8.

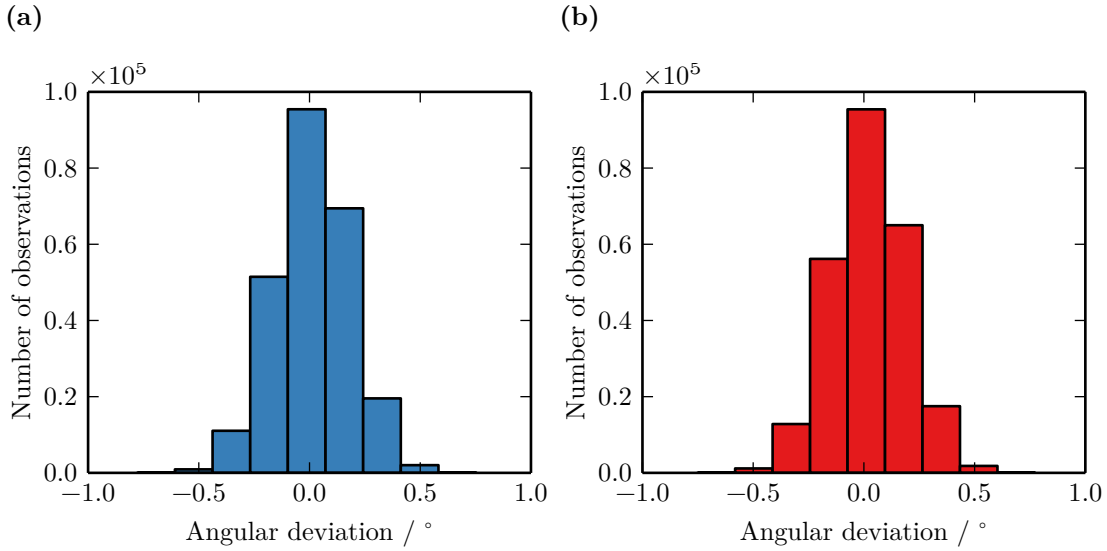

Figure 3.9: Angular deviation during orientationally restrained umbrella sampling. The histograms were calculated using 1 μs of production simulation from a simulation window with the proteins restrained at positions corresponding to an inter-protein separation of 6 nm. **(a)** Angular deviation of protein 1. **(b)** Angular deviation of protein 2.

3.4.3 Analysis of orientational deviation

To see how the rotational restraints perform, we can analyse the variation on the rotation angle. We implement the restraints with a force constant $k_r = 1000$ kJ mol^{-1} nm^{-2} and $\epsilon = 0.01$ nm^2. A force constant greater that 200 kJ mol^{-1} nm^{-2} was shown to produce consistent rotation of the γ subunit of F_1-ATP synthase (Kutzner et al. 2011). We wanted to keep the orientational deviation below 1°, and using this value of the force constant achieves that. A histogram of the deviation in angle for the two proteins during 1 μs of production simulation is shown in Figure 3.9. Here we can clearly see that the deviation is less than 1° for all observed cases.

3.5 Umbrella sampling of rotationally-restrained proteins

3.5.1 Sampling window preparation

The umbrella sampling procedure follows a similar approach as that used in the unrestrained simulations discussed in Section 3.3.3.1 (on page 38), again using the pulling procedure outlined in Appendix A.iii (on page 195). The proteins are restrained orientationally during the pulling procedure using the potential introduced in Equation (3.36). For each of the simulation windows, an equilibration is run with the proteins restrained in the x-y directions for 1 μs. We now have a series of equilibrated systems in each of which the proteins are at a separation required for our umbrella sampling.

3.5.2 Stronger force restraints at close range

The nature of these orientationally-restrained PMFs is such that they are likely to pick up more detail in the free energy surface. If there are fewer degrees of freedom characterizing the free energy surface, we are averaging over more of configuration space. As the number of degrees of freedom is increased, we are looking at a smaller volume in configuration space, and will resolve more detailed features. The region where such details will be most prevalent is at small protein-protein separations, as there will be more complex interactions involved when the intervening lipids do not act in such a fluid manner, caused by their reduced mobility. It is therefore of interest to increase our sampling resolution at these short distances so any intricacies in the PMF profiles are more clearly resolved.

In Section 3.3.3.1 (on page 38) we used simulation windows separated by 0.1 nm. Here we increased the sampling at these short length scales by introducing more simulation windows, distributed at half the separation used in the unrestrained PMFs, corresponding to a separation of 0.05 nm. We also wish to explore with more precision, which we can do by increasing the force constant, k_t, of the umbrella potential for these windows to 10 MJ mol^{-1} nm^{-2}.

3.5.3 Sampling lipidated and delipidated states

For the simulation windows where the proteins are separated by only a few lipids, the mobility of these intervening lipids will be dramatically reduced. It will take a very long production simulation to be able to adequately sample the lipidation-delipidation events that occur at these close distances. In order for us to attempt to improve the sampling of these events, we manually remove lipids from between the two proteins in a series of the simulation windows, moving them to the bulk of the bilayer, and run another 1 μs of equilibration. Without intervening in this way, we would not be able to ensure we sample the slow lipidation-delipidation process. The dynamics of the lipidation-delipidation process itself will not be sampled, but we approximate the behaviour of the process by including adequate sampling of the two end states.

3.6 Obtaining PMFs for the restrained system

From the equilibrated simulation windows we ran production simulations for 4 μs for each simulation window with a force constant of 1000 kJ mol^{-1} nm^{-2}. For the simulation windows with a stronger force constant of 10 MJ mol^{-1} nm^{-2}, which are separated by 0.05 nm, we ran 2 μs of production simulation for the lipidated states and 4 μs of production simulation for the delipidated states. Longer simulations were done for the delipidated states as the lower force constant states were not run with explicitly delipidated configurations.

3.6.1 PMFs

To combine the production simulations for the orientationally-restrained simulations we again used WHAM. The four PMFs are shown in Figure 3.10. The PMFs were set to zero at an inter-protein separation of 8 nm, where the potentials become approximately constant.

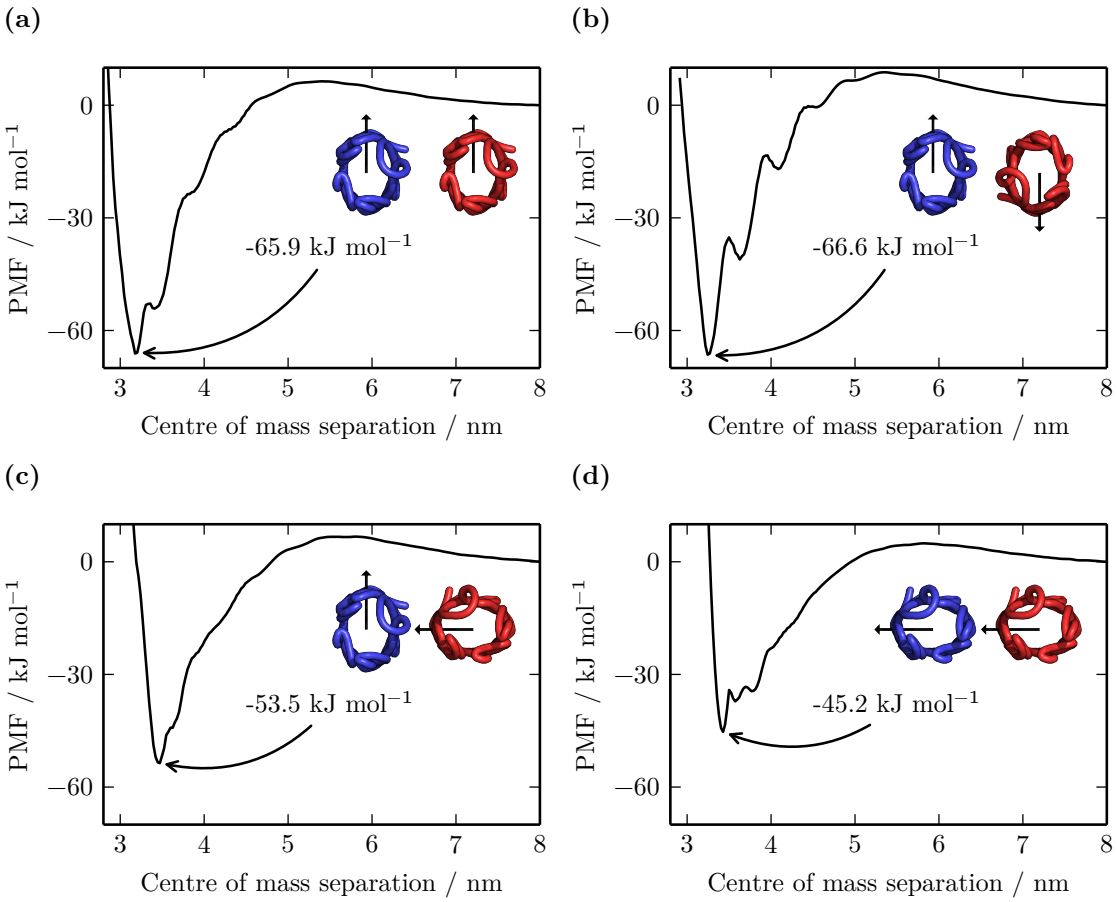

Figure 3.10: PMFs of the orientationally restrained NanC proteins. The PMFs are for orientational configuration of: **(a)** $(\phi_1, \phi_2) = (90^\circ, 90^\circ)$ **(b)** $(\phi_1, \phi_2) = (90^\circ, 270^\circ)$ **(c)** $(\phi_1, \phi_2) = (90^\circ, 180^\circ)$ **(d)** $(\phi_1, \phi_2) = (180^\circ, 180^\circ)$.

We categorized the PMFs in Figure 3.10 by the depths of their potential well, which resulted in three categories of well depth. The first category contains the PMFs in Figures 3.10(a) and 3.10(b), which both have depths of approximately -66 kJ mol^{-1} occurring at inter-protein separations of approximately 3.2 nm. This first category corresponds to the orientational configurations of maximal contact, $(\phi_1, \phi_2) = (90^\circ, 90^\circ)$ and $(\phi_1, \phi_2) = (90^\circ, 270^\circ)$, where less-curved surfaces of both proteins are brought into contact (shown in Figures 3.7(a) and 3.7(b), respectively). This depth is of a similar magnitude to that calculated for interacting pairs of OmpF trimers by Casuso et al. (2012). It is interesting to note that the depths of the PMFs for these two parallel and anti-parallel orientational configurations are approximately the same. The next category contains the PMF in Figure 3.10(c), with a potential well depth of approximately -52 kJ mol^{-1} occurring at a separation of approximately 3.5 nm. This corresponds to the intermediate orientational configuration in Figure 3.7(c), where a less-curved surface of one protein is brought into contact with a more-curved surface of the other. The decrease in the depth of the PMF indicates that the configurations with two less-curved surfaces in contact are more stable than this intermediate contact configuration, where $(\phi_1, \phi_2) = (90^\circ, 180^\circ)$. The third category contains the PMF shown in Figure 3.10(d), which is the shallowest of the four PMFs with a potential well depth of approximately -45 kJ mol^{-1}, occurring at an inter-protein separation of 3.5 nm. This PMF corresponds to the orientational configuration with minimal protein contact, where more-curved surfaces of both proteins are brought into contact (shown in Figure 3.7(d)). This configuration is the least stable of the four configurations considered here. The correlation between the depth of the PMFs and the orientational configuration of the proteins suggests that the strength of the interaction may correlate with the overall extent of the resultant protein-protein interface.

3.6.2 Convergence of calculated PMFs

As performed for the case of the orientationally-unrestrained proteins in Section 3.3.3.3 (on page 41), we have assessed the convergence of the PMFs by calculating them using non-intersecting subsets of the production simulation data. These comparison plots are shown in Figure 3.11. We can see that there is some discrepancy between the two PMFs calculated for each orientational configuration, but the features are generally the same and the differences are small in comparison to the size of the features observed in each of the profiles.

3.7 Buried surface area determines the depth of the PMF potential well

In order to formally characterize the dependence of the PMF depth on the orientation of the proteins, which we have suggested is related to the extent of the protein-protein interface, we calculated the solvent-accessible surface area (SASA) of the two proteins as a function of the separation of their centres of mass. The SASA is calculated using a spherical probe whose size determines the level of detail captured in the surface for a specific set of atoms or particles (Eisenhaber et al. 1995).

3.7.1 Calculation of the buried surface area

We used a probe with a radius of 0.47 nm, which is twice the radius of the coarse grained particles (0.235 nm) and should be a reasonable measure for the size of a lipid. We chose this size of probe because it is the lipids that are the solvent of interest when two proteins come together in a membrane. To obtain the buried surface area of the proteins at various positions along the reaction coordinate, we analysed the surface area of the simulation windows that used the higher translational restraining potential of 10 MJ mol^{-1} nm^{-2}. This was to enable the analysis of the surface area on a finer scale, using the window

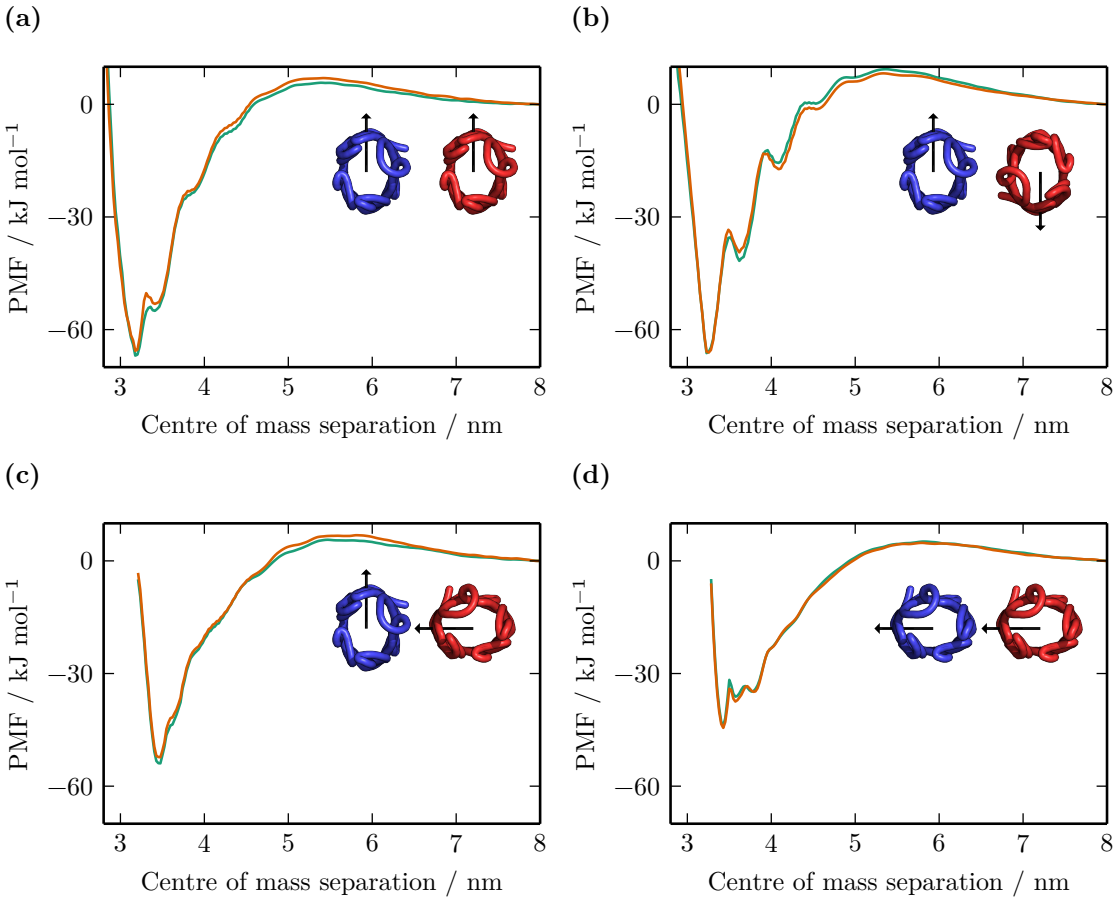

Figure 3.11: Convergence of the rotationally restrained PMFs. The two profiles were calculated using equally sized and distributed, non-overlapping subsets of the production simulations used to calculate the PMFs in Figure 3.10. The plots are for orientational configurations of: **(a)** $(\phi_1, \phi_2) = (90^\circ, 90^\circ)$ **(b)** $(\phi_1, \phi_2) = (90^\circ, 270^\circ)$ **(c)** $(\phi_1, \phi_2) = (90^\circ, 180^\circ)$ **(d)** $(\phi_1, \phi_2) = (180^\circ, 180^\circ)$.

separations of 0.05 nm. Using the higher force constant also ensured that the surface area was measured for a conformation that was sampled closer to the centre of the window. If we used the windows with the weaker force constant, we would be measuring the surface area for conformations with separations that could differ significantly from the position of the window centre: in regions where the gradient of the PMF was steep, the force on the protein would act move it away from the window centre. The SASA was calculated for each of the proteins individually and then for the pair of proteins together. The buried surface area is given by the difference between the surface area of the proteins calculated individually and the surface area of the combined, two-protein system. All of the surface area calculations were carried out using the `g_sas` tool in *GROMACS* on 1 μs of production simulation.

3.7.2 Dependence of the buried surface area on the potential well depth

For each of the four orientational configurations, Figure 3.12(a) shows the buried surface area as a function of distance from the minimum of their respective PMFs. We chose this measure since we wanted to remove the effect that the orientational variation in protein radius has on the location of the minimum. In Figure 3.12(a) we see that there is a stratification of the buried surface areas in the region around the location of the minima of the PMFs. As with the PMF depths in Figure 3.10 (on page 52), the buried surface areas can also be divided into three categories. The buried surface area is largest for the orientations $(\phi_1, \phi_2) = (90^\circ, 90^\circ)$ and $(\phi_1, \phi_2) = (90^\circ, 270^\circ)$, where the two less-curved surfaces are brought together. The next largest buried surface area, around the minimum of the PMF, is for the orientation $(\phi_1, \phi_2) = (90^\circ, 180^\circ)$, where one more-curved surface is brought into contact with one less-curved surface. The smallest buried surface area, around the minimum of the PMF, is for the orientation $(\phi_1, \phi_2) = (180^\circ, 180^\circ)$, where two more-curved surfaces are brought into contact. The correlation between the depth

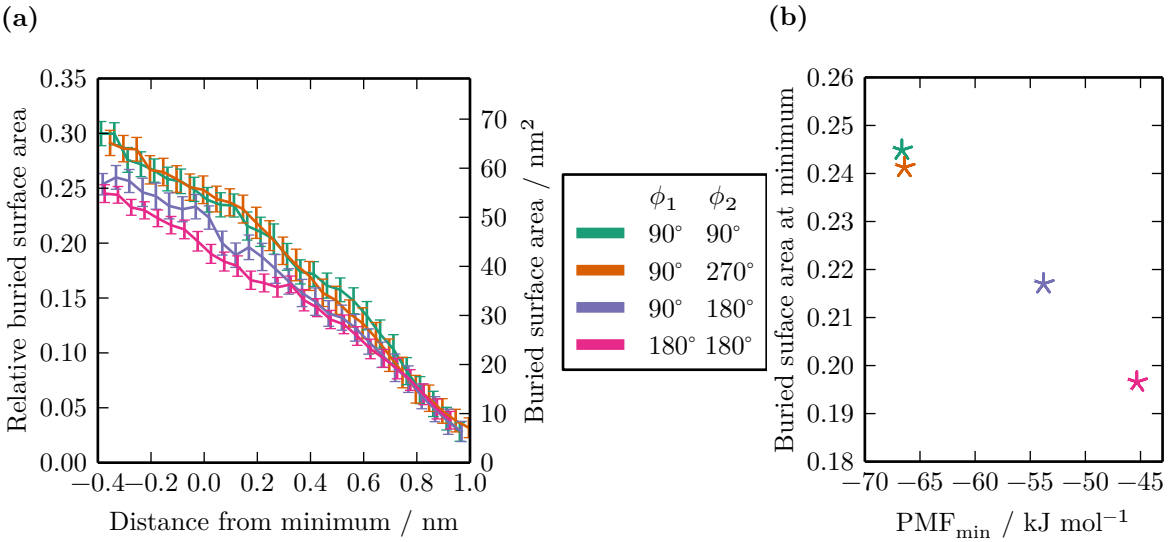

Figure 3.12: **(a)** Buried surface area for each of the orientational configurations as a function of distance from the minimum of the PMF. **(b)** Buried surface area at the minimum of the PMF plotted against the depth of the minimum of the PMF.

of the PMF and the buried surface area can be seen in Figure 3.12(b), in which these two quantities are plotted. We see that there is a negative correlation between the two quantities: for a protein orientation with a larger buried surface area, the minimum of the PMF is deeper.

3.8 Protein-lipid-protein effects lead to metastable states in the potentials of mean force

As well as the global minima of the potential wells in the PMFs of Figure 3.10 (on page 52), there are also multiple local minima, which occur at a variety of centre of mass separations. When using the terms global and local here, we mean a global minima in the context of a single PMF, and local minima as other minima within the same PMF. For example, the PMF for the orientational configuration $(\phi_1, \phi_2) = (90°, 270°)$, shown in Figure 3.13, has a global minimum (labelled α) and two higher-energy local minima

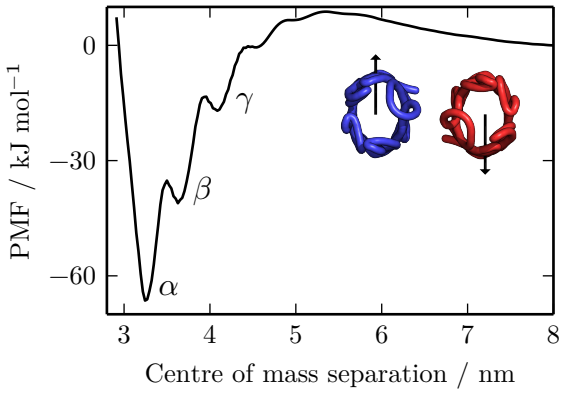

Figure 3.13: PMF for the orientational configuration $(\phi_1, \phi_2) = (90°, 270°)$, where the global minimum, α, and two local minima, β and γ, have been labelled.

(labelled β and γ). By fitting quadratic curves to the minima in Figure 3.13 we calculated their locations as 3.26 nm, 3.62 nm and 4.07 nm for α, β and γ, respectively.

The nature of the global (α) and local (β and γ) minima is illustrated by the simulation snapshots shown in Figure 3.14. These snapshots were taken from the simulation windows used to calculate the PMF in Figure 3.13, for an orientational configuration of $(\phi_1, \phi_2) = (90°, 270°)$. The snapshot shown in Figure 3.14(a) is from the umbrella sampling window in which the proteins were restrained at a separation of 3.3 nm, which is closest to the minimum at 3.26 nm in Figure 3.13, labeled α. We see that there is one lipid molecule between the two proteins at this global minimum. It should be noted that this is the only lipid in between the two proteins; there is no equivalent lipid on the extracellular side of the membrane (the view from the other side of the membrane is shown in Figure 3.15), so the global minimum configuration for this orientation has space for one lipid on the periplasmic side of the membrane. A snapshot from the umbrella sampling window with the proteins restrained at a separation of 3.6 nm is shown in Figure 3.14(b), which is the window closest to the minimum at 3.62 nm, labelled β in Figure 3.13. We can see that there are two lipid molecules between the two proteins in this snapshot. The snapshot in Figure 3.14(c) is taken from the umbrella sampling window in which the proteins are restrained at a separation of 4.1 nm, which is the

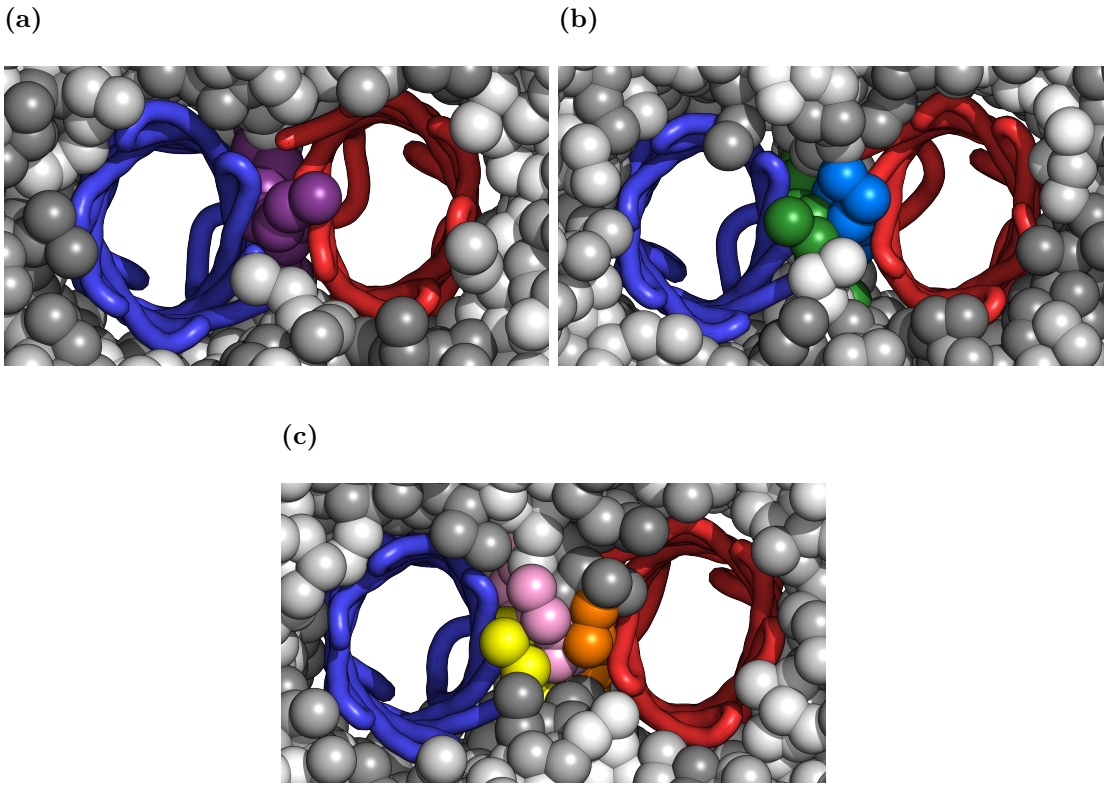

Figure 3.14: Snapshots from simulation windows of the orientational configuration $(\phi_1, \phi_2) = (90°, 270°)$. **(a)** At a separation of 3.1 nm one lipid can fit between the proteins. **(b)** At a separation of 3.6 nm two lipids can fit between the proteins. **(c)** At a separation of 4.1 nm three lipids can fit between the proteins.

window closest to the minimum at 4.07 nm, labelled γ in Figure 3.13, in which we see that three lipid molecules can occupy the space between the two proteins. These observations suggest that the existence of these metastable states is a result of protein-lipid-protein interactions in this orientationally-restrained system.

3.8.1 Analysis of lipid density around a single protein

To investigate the hypothesis that these metastable states were the result of the lipid ordering between the proteins, we calculated the lipid distribution around a freely diffusing NanC in a POPE bilayer. We used a 5 μs simulation of a single NanC protein in a POPE

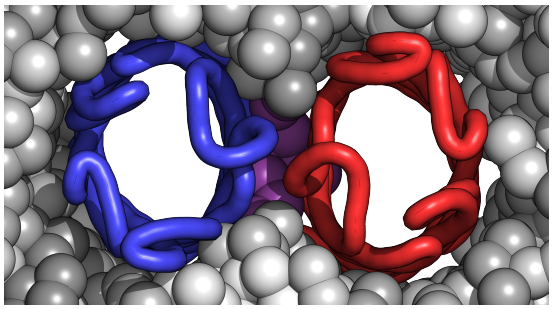

Figure 3.15: Extracellular side of the membrane showing the singular intervening lipid. This is taken from the same simulation frame as Figure 3.14(a).

bilayer. This simulation consisted of a single coarse-grained NanC protein model (the same model used for the PMF calculations) embedded in a 25 nm square membrane constructed from coarse-grained POPE molecules. To analyse the lipid distribution around the single protein, we rotated each frame of the trajectory so that the NanC protein was aligned with its position at the start of the simulation. From this aligned trajectory we were able to calculate the position of the lipid particles in relation to the protein for the entire simulation. The particle density was calculated for a 6 nm square region around the NanC protein for each of the particles in the coarse-grained lipid molecules. The distribution for a specific coarse-grained particle in the lipid tail is shown in Figure 3.16, where distinct annuli are visible, indicating regions of preferred occupation.

3.8.2 Peaks in lipid density around a single protein

We calculated the mean lipid distribution in a direction that corresponds to the direction of the other protein for the orientational configuration $(\phi_1, \phi_2) = (90°, 270°)$, which is the same relative direction for both of the proteins, as they have the same less-curved surface oriented toward the other protein. This direction is marked in Figure 3.16 by the dashed lines. To obtain this mean lipid density, we averaged over the radial projection of the two-dimensional density. To do this we projected the two-dimensional density onto a

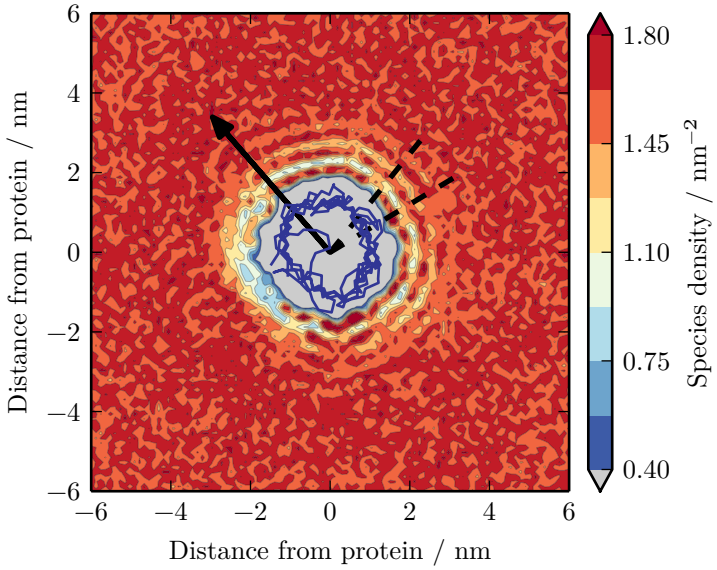

Figure 3.16: Lipid particle density around a freely diffusing NanC. The density is shown for the third coarse-grained particle in the palmitate chain of each lipid molecule. The black arrow is the same as that used in Figure 3.7 (on page 46) to show the proteins' orientations and passes through the C_α particle of isoleucine residue 209. The dashed lines indicate the direction in which the other protein is located when in the orientational configuration $(\phi_1, \phi_2) = (90°, 270°)$.

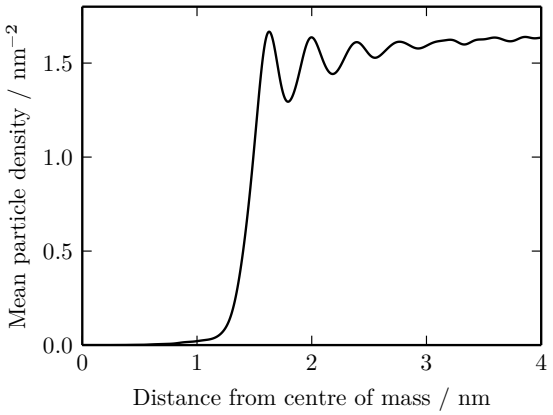

Figure 3.17: Mean distribution of all lipid particles in both leaflets taken in the angular range corresponding to the orientational configuration $(\phi_1, \phi_2) = (90°, 270°)$. This data was collected from a 5 μs coarse-grained simulation of a freely diffusing NanC protein in a POPE bilayer.

series of 4 nm radial lines, emanating from the protein's centre of mass at regular angular intervals. The lines were at 0.5° intervals within the region marked by the dashed radial lines in Figure 3.16, which represent an angular window of ±10° from the direction of the other proteins in the PMF simulations. The average lipid distribution across both leaflets and all coarse-grained lipid particles in this direction is shown in Figure 3.17, where again we can see there are preferred distances from the protein at which the lipids are observed.

3.8.3 Aligning peaks in the lipid density distributions to predict metastable state locations

We can use the mean lipid distribution calculated in the direction of the other protein, shown in Figure 3.17, to predict the separations at which the intervening space between the two proteins will be optimally packed by specific numbers of lipids. Assuming that the lipid behaviour in the presence of two proteins is not too dissimilar from that shown in Figure 3.16, we hypothesize that the optimum separation of proteins will occur when peaks in the lipid density plot for each protein overlap. This corresponds to a situation

where we optimally pack the intervening region with a certain number of lipids; that number is the number of aligned pairs of peaks in the distribution that occur between the two proteins.

To perform this calculation we use the mean lipid density calculated in the direction of the other protein, shown in Figure 3.17. By reversing the x-axis of that plot, and overlaying it on the original plot we have an approximate map for the optimal distributions of lipids around two proteins in close proximity. The relative separation of the two proteins can be changed by moving the reversed plot along the x-axis. If we use this composite plot to align a specific number of peaks in the two distributions, we can use the location of the centre of the protein in the reversed plot to estimate the optimum separation for that number of intervening lipids. The result of this procedure for one, two and three intervening lipids is illustrated in Figure 3.18. For the case of a single intervening lipid we aligned the first peaks of both plots (shown in Figure 3.18(a)). For the case of two intervening lipids we aligned the first peak with the second peak of the other plot (shown in Figure 3.18(b)). Finally for the case of three intervening lipids, we aligned the first peak with the third peak of the other plot as well as aligning the second peaks of both plots (shown in Figure 3.18(c)).

Following the overlaying and alignment procedure, for the minimum labelled α in the PMF in Figure 3.13 (on page 58) that occurs at a separation of 3.26 nm, we predict an optimal separation of 3.24 nm with one intervening lipid. For the first metastable state labelled β in Figure 3.13 (on page 58) that occurs at a separation of 3.62 nm, we predict a separation of 3.63 nm with two intervening lipids. For the second metastable state labelled γ in Figure 3.13 that occurs at a separation of 4.07 nm, we predict a separation of 4.02 nm with three intervening lipids. Our predictions for the metastable state locations are in close agreement with their location in the PMF. This supports our suggestion that the metastable states observed in the PMFs are due to the protein-lipid-protein effects caused by the distribution of lipids between the two NanC proteins.

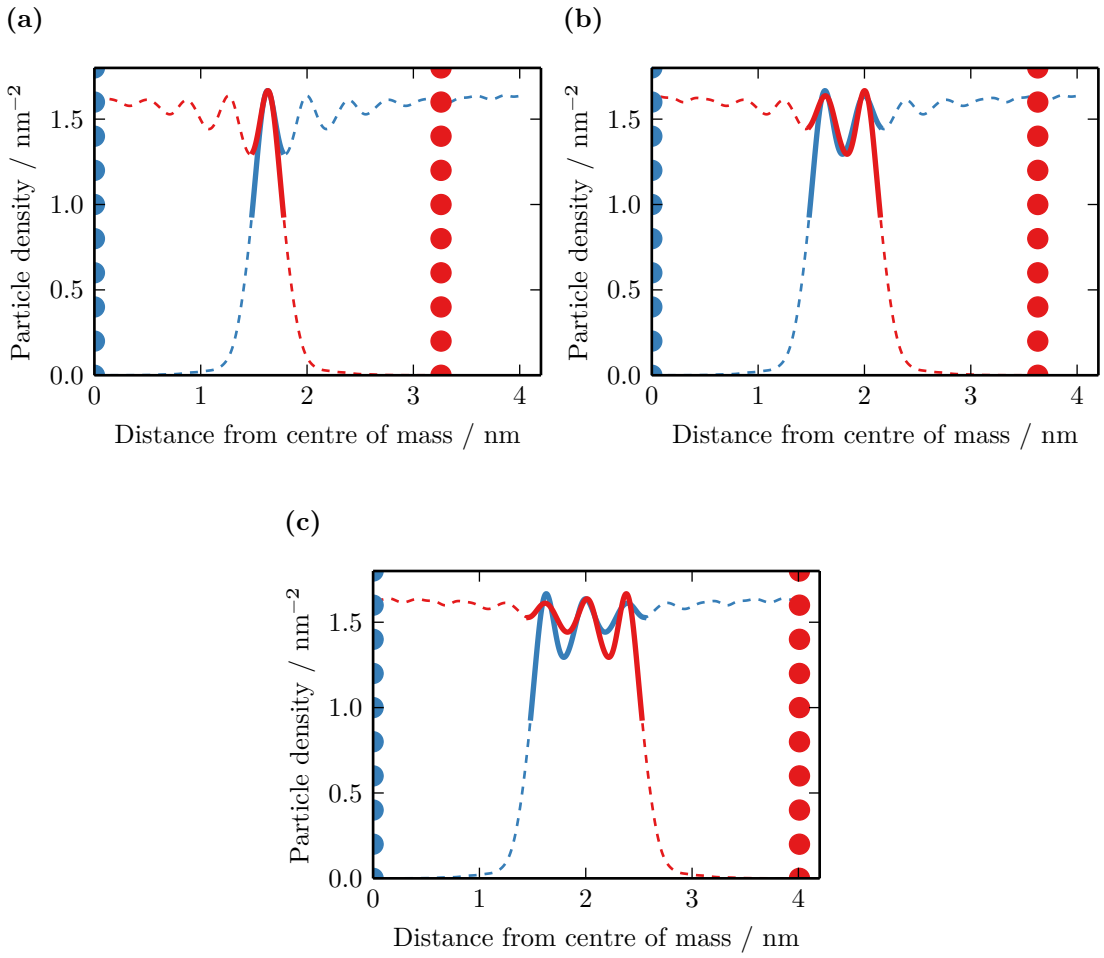

Figure 3.18: Predicting the location of metastable states in the PMF calculated for the orientational configuration $(\phi_1, \phi_2) = (90°, 270°)$ (shown in Figure 3.13 (on page 58)), by aligning the peaks in the mean lipid particle density with itself, reversed in the x-direction. In each plot the location of the centres of the two proteins are shown by the vertical row of dots, and the corresponding optimal lipid distributions are shown by the dashed lines. The section of each distribution corresponding to the intervening region between the two proteins is shown by the solid lines. These sections are defined as the regions in which both lipid density profiles are greater than half their maximal value. The alignments for one **(a)**, two **(b)**, and three **(c)** intervening lipids are shown. The vertical rows of dots, indicating the prediction of the optimal separation, are located at 3.24 nm **(a)**, 3.63 nm **(b)**, and 4.02 nm **(c)**.

The procedure followed above was very successful at predicting the locations of the restrained metastable states for the orientational configuration $(\phi_1, \phi_2) = (90°, 270°)$, where it is the same face of each protein that is facing the other protein. This is as a result of the fact that the overlaying procedure involves using the same lipid density plot for both proteins in this case. For the other orientational configurations the protein faces that are brought into contact are different and as a result the lipid density profiles are different. Without such a symmetric operation it is not obvious how best to align the profiles, and this is most likely the reason why these observed features are most pronounced for the orientations $(\phi_1, \phi_2) = (90°, 270°)$. For the other configurations the pattern of the restrained metastable states is less regular and their shapes are less pronounced.

Such features are not usually observed in PMFs calculated with proteins that are free to rotate (for example the unrestrained PMF calculated above, shown in Figure 3.3 (on page 40)). By restraining the proteins' orientations, we are sampling from a multi-dimensional PMF along a specific reaction pathway, which is in contrast to the case of a rotationally-unrestrained PMF where, by averaging over the rotational degrees of freedom in the free energy landscape, these features are not observed. In the orientationally-unrestrained case the proteins would be able to rotate in order to alter the distance between their surfaces (if they are not perfectly rotationally symmetric) so that the intervening region could be optimally packed with lipids without leaving any voids. However, for a system with rotationally restrained proteins, there is an optimal separation at which multiple lipid molecules can occupy the intervening space between the proteins.

3.9 Summary

In this chapter we introduced the concept of using MD simulations to investigate protein-protein interactions. We demonstrated this process by calculating the PMF for the association of a pair of bacterial outer membrane proteins in a pure POPE lipid bilayer,

using a coarse-grained model of the system. From this initial calculation we demonstrated the procedure that we extended in the following section. The PMF calculated shows many features common to such interactions: a deep potential well at small inter-protein separations and a small barrier at a slightly larger separation. The depth of this PMF was approximately -70 kJ mol^{-1}, which although not directly comparable to an experimentally determined value, is of a similar magnitude to the depth of -150 kJ mol^{-1} calculated for OmpF trimers by Casuso et al. (2012). The trimer system is much larger—each monomer is larger and they form a trimer—so if the relationship demonstrated in this chapter between buried surface area and PMF depth holds for OmpF, we would expect it to be deeper.

We developed a method for applying rotational restraints to the proteins so that their orientation in the plane of the membrane was fixed. Using such restraints we were able to calculate restrained PMFs for specific relative orientations of the proteins, for which we chose four different orientational combinations with different levels of contact. The PMFs calculated for these systems showed several interesting features. The depth of the PMFs varied between some of the orientational configurations. We showed that the depth of the PMF was negatively correlated with the amount of buried surface area: with a larger buried surface area between the two proteins, the depth of the PMF was greater. This result is contrary to the result published by Periole, Knepp, et al. (2012), but in their rhodopsin system a complex series of binding modes were identified, along with some relative orientations that showed approximately no binding affinity. With the relatively-featureless outer membrane protein, NanC, we demonstrated that in a more simple system, a dependence of the PMF depth on the buried surface area was still observed, as is often found for soluble proteins.

We also identified features in the PMFs that were attributed to protein-lipid-protein interactions. In such a highly restrained system we were able to access behaviour of the lipids over which we would average in a system where the protein were free to rotate.

Specifically, for a given direction from the protein, there appears to be an annular lipid distribution, which, when combined with the nature of a two protein interaction, leads to the existence of states of preferred separation. In these states the intervening space between the proteins was optimally populated by a specific number of lipid molecules. Such behaviour is previously unreported and marks a significant step into understanding the way in which lipid behaviour plays a dominant role in determining the strength and form of protein-protein interactions. In an unrestrained system the proteins would rotate such that these states relax into an energetically more favourable state, and we would not be able to observe the complexities of the higher dimensional free energy surface for protein-protein interaction in our NanC system.

In the following chapters we will use these PMFs to parameterize a yet coarser-grained model of a bilayer. By representing the proteins in the system as points embedded in a two-dimensional bilayer which have some intrinsic orientation, we will be able to use the PMFs calculated in this chapter to parameterize the protein-protein interaction strength (extending the approach introduced by Yiannourakou et al. (2010)). This simplified membrane model is introduced in the following chapter and extended to include an anisotropic potential, based on the orientationaly-restrained PMFs, in Chapter 5.

Chapter 4

Using an Isotropic Model of NanC to Simulate Diffusion and Interaction in a Bilayer

Using a one-dimensional potential of mean force, we demonstrate the parameterization of an individual-based discrete model of NanC in a bilayer. We identify acceptable values of the simulation and model parameters for various bilayer environments. We then establish a framework for analysing simulations of the model, enabling us to assess the performance of individual-based movement schemes for our bilayer system. This framework is used to investigate the effect of varying the parameters describing the protein-protein interaction, from which we elucidate their role in the dynamics of the system. Comparing the simulation results with a coarse-grained molecular dynamics simulation of a similar system shows that protein clustering occurs too quickly. A more advanced cluster-based Monte Carlo method is employed, which improves the agreement between our model and the MD simulation.

Contents

4.1 Introduction

In this chapter we develop a discrete protein model of NanC, parameterized using the PMFs calculated in Chapter 3. A discrete protein model is one in which the proteins are represented as a single entity, rather than as a collection of atoms or particles, as was the case in the coarse-grained MD simulations. This is a simplification of the system in which we are interested, but a reduction in the degrees of freedom is required if we want to run many repeat simulations using commercial hardware. This reduced representation of a membrane system, and of other similar systems, is often performed in order to simulate larger and more complicated systems. This level of coarse-graining, where we are still able to resolve individual proteins (as opposed to continuum approximations, where proteins are represented by some local protein concentration) is usually divided into two distinct schemes: on-lattice simulations and off-lattice simulations.

4.1.1 On-lattice simulations

On-lattice models divide the spatial domain into a set of lattice sites, to which the objects in a simulation are confined. The occupancy of each site is usually restricted to a maximum of one object. The use of a lattice ensures that only integer arithmetic is

required when calculating distances and separations, greatly facilitating the simulation of the model.

The cellular automaton is one of the simplest incarnations of a lattice-based model. The lattice sites evolve according to a specific set of rules in a cellular automata simulation. When simulating a system where the objects are proteins, the set of rules are designed to result in the random translation of proteins from occupied sites to unoccupied, neighbouring (or nearby) sites. This type of simulation has been used to investigate the effects of lipid rafts and cytoskeletal corralling on membrane protein diffusion and interaction (Nicolau, Burrage, et al. 2006; Nicolau, Hancock, et al. 2007). The effect of lipid rafts was included by defining regions of the membrane as raft regions, within which different diffusion properties were applied; the rafts were characterised by a shorter protein hopping distance. Simulations showed that the clustering of proteins within rafts increased when the proteins had a specific affinity for that raft; in the absence of differing affinities, varying the diffusive properties of different protein types in specific rafts would also lead to their segregation. A cellular automaton, described by simple rules, is useful for simulating systems populated by objects that are not strongly interacting, however, they fail to accurately capture the motion of systems where the clustering of objects is favoured.

Lattice gas simulations introduce an extra level of complexity to the cellular automaton approach. A stochastic translation of an object on the lattice is proposed and the resulting change in energy is calculated. The probability of accepting a move is dependent on this energy change; lattice gas simulations are lattice-based Monte Carlo simulations. Goldman et al. (2004) used lattice gas simulations to investigate the clustering characteristics of membrane proteins, both point-like and more complex proteins, whilst modifying the strength of protein-protein interactions. Lattice gas simulations have also been used to study the periodicity of protein clusters in growing bacteria (Wang et al. 2008) and the cluster phases of eukaryotic coat proteins, used in intracellular trafficking (G.

Huber et al. 2011). However, lattice gas simulations are poorly suited to capturing the concerted motion of clusters (something which is abundant in membrane systems), since the probability that each object in a large and strongly-interacting cluster will move in the same direction is vanishingly small.

Lattice-Boltzmann methods are commonly used in computational fluid dynamics simulations, but have also been used with some success in the simulation of lipid bilayers. Wagner et al. (2007) used lattice-Boltzmann methods to model the phase separation of lipids in a membrane consisting of two coupled leaflets. However, given that such simulations include explicit treatment of the lipids in the membrane, they can be computationally costly; however, by continuing to represent the dynamics of the lipid as an explicit fluid, hydrodynamic interactions are incorporated.

Instead of relying on a domain modelled by a fixed lattice, Collins et al. (2010) used an adaptive lattice to investigate the dynamics and dimerization of epidermal growth factor receptor. They used densely-distributed lattice sites to represent a lipid-raft-type region, which was contained within a larger domain of sparsely distributed sites, representing the rest of the membrane section. Techniques like this are very successful at investigating the biochemical nature of a system, but advanced knowledge is needed of the regions in which clustering is likely or intended.

In an attempt to benefit from the reduced computational cost of lattice models, but improve the spatial resolution of simulations, dynamic lattice models are a hybrid of on-lattice and off-lattice simulations. In a continuous spatial domain the proteins are connected by a series of links; the interaction between two objects is only considered if they are connected by a link, and the total number of links in the system remains constant. The benefit of using a dynamic lattice model is that the number of interactions remains constant throughout the simulation, so although you have to perform non-integer arithmetic, the number of calculations is fixed. This makes a dynamic lattice model well suited to systems that are densely packed, like the lipid-sterol membranes simulated by

Polson et al. (2001), from which they were able to obtain phase diagrams for the system that were in good agreement with experimental results.

4.1.2 Off-lattice models

Off-lattice models have a continuous spatial domain, so calculating the position and movement of objects involves floating-point arithmetic. The effect of this is to increase the computational cost of such simulations, but it enables more detailed analysis of the process of clustering. A major contribution to the increase in computational operations required with off-lattice models is the need to verify constantly the separation of objects. This becomes an important factor when working with crowded systems, where the implementation of cell-lists, and other techniques to reduce this cost, are increasingly employed.

Langevin dynamics simulates the motion of particles that are in a fluid coupled to an external heat bath. By integrating the laws of motion for a subset of the constituents of the complete system, and incorporating the effect of the rest of the system by thermal noise terms, it is possible to study complicated systems of interacting particles that are immersed in an implicit solvent. In the non-inertial limit, where the timescale of the thermal fluctuations of the solvent is much larger than the corresponding timescale for momentum correlation, the motion of system may be represented by Brownian dynamics (Huang 1987). Brownian dynamics simulations are widely used, and have been employed to measure biochemical reaction rates in both the cytosol and the membrane. Andrews et al. (2010) used a Brownian dynamics simulation package called *SmolDyn*[1] to investigate a pheromone sensing system in yeast by looking at the roles that different proteins played. They were able to suggest a protein that would be a good candidate for identifying the fittest partner cell with which to mate. The *SmolDyn* software package can simulate large systems of proteins in one to three dimensions. In this work, their system was an entire

[1] http://www.smoldyn.org

cell; in this cell, the proteins were able to adsorb onto the cellular membrane and continue to diffuse in two dimensions. However, *SmolDyn*, like many Brownian dynamics packages, works by simulating the motion of freely diffusing particles, which can be a sensible choice for diffusion in the cytoplasm, but is less suitable when considering diffusion in the membrane. Any inter-protein interactions are stochastic in nature, and occur when the proteins are below some threshold separation. Another shortcoming of Brownian dynamics packages, such as *ChemCell*[2], is that the proteins are often represented as dimensionless particles, so any interaction are represented as reactions from substrates to products, which are unable to represent details of clustering. There is also no facility to include longer range interactions, which can be present in the membrane as a result of hydrophobic mismatch. *ReaDDy*, a recently developed simulation package created by Schöneberg et al. (2013), overcomes many of these shortcomings and enables the explicit inclusion of inter-protein interactions that are a function of separation, but without at least a simple attempt at multi-body interactions, interactions like those found in Chapter 3 are not easily represented.

Off-lattice Monte Carlo methods date back to the landmark paper by Metropolis et al. (1953), in which the authors proposed a method for evolving a system stochastically that is dependent on the energy change resulting from transitions. In membrane simulations, the general method is to propose small movements of objects in the system, which are accepted with a probability that is dependent on the energetic cost of the move. Simulations like these benefit from the fact that the transitions are based on evaluation of potentials, which is a simpler calculation than the evaluation of the forces acting on objects in the system. However, it an be complicated to construct appropriately-smart moves that have a high probability of acceptance. P. Sengupta et al. (2004) used a Monte Carlo model to simulate cholesterol organization in the membrane. They included both a short-range, hard sphere interaction, and a long-range repulsive interaction between

[2]http://www.sandia.gov/~sjplimp/cell.html

the cholesterol molecules in the lipid membrane. They used this system to look at the phase transitions of the membrane as a function of the system parameters (density and temperature), observing that the system developed a lateral order for certain values of these parameters. Yet, simulations like this, which do not technically represent real dynamics of the system, are limited in their scope to represent the actual motion of proteins in the membrane. With a long enough simulation, you can ensure that there is an adequate sampling of configuration space (with annealing to avoid staying in local minima), but the motion is not dynamically-realistic. Also, these simulations neglect the important role that clustering plays in systems of strongly interacting particles, particularly the role of the diffusion and reshaping of clusters of objects in the system.

To help overcome these problems with Metropolis Monte Carlo simulations that use movements of individual proteins, various linking schemes have been developed to facilitate the movement of clusters of interacting particles. Statically-linked Monte Carlo simulations allow for movement of the system on various length scales (Wolff 1989; Liu et al. 2004; Troisi et al. 2005b). In these methods, a cluster of connected objects is selected and moved at each time-step, instead of just moving one object. The static linking approaches construct the cluster to be moved based only on the initial state of the system. The effect of such an approach is to create unphysical movements of the clusters of particles. The movement of interacting particles would not be based only on the initial configuration of the system, but would instead move down gradients in the system energy. This is something that is addressed in dynamically-linked Monte Carlo methods, such as the virtual move Monte Carlo scheme of Whitelam and Geissler (2007). The introduction of virtual moves to the process of selecting the moving clusters means that the clusters are not only selected based on their current configuration, but also the configuration that would result from the proposed move (Whitelam 2011). The effect of this is to add dynamically-realistic movement to an off-lattice Monte Carlo simulation. This model was used by Villar et al. (2009) to simulate the clustering of homomeric protein complexes,

but has not (at the time of writing) been implemented in the simulation of membrane protein systems.

Other more complex schemes of protein movement, such as agent-based simulations, where objects or clusters of objects evolve according to some pre-defined set of rules, bear some similarity to the on-lattice cellular automata, as they can encode more complex protein-protein interactions. These rules do not have to be based on physically realistic processes, but aim to improve the sampling of configuration space. Troisi et al. (2005a) developed an agent-based simulation scheme for the self assembly of two-dimensional protein clusters. Their approach involved a system-wide process for cluster evolution, in which the clusters are ranked based on their interaction energy. When clusters are formed that are not energetically-optimal, compared to other clusters of the same size, then there is an increased probability of splitting. This model was used to simulate an atomic system, resulting in a lower energy final state than the same system simulated using traditional Monte Carlo methods (Fortuna et al. 2010). However, as discussed by Jankowski et al. (2009), this approach is limited by the fact it only maintains clusters that are constructed from optimal smaller clusters. Agent-based models like this are also restricted in the sense that they may find energetically optimal clusters of proteins, but the route taken to reach them may not be dynamically realistic.

In this chapter we develop a protein model based on the interactions, characterized by PMFs, between pairs of NanC proteins that were calculated in Chapter 3. Next we introduce the theoretical basis for the Monte Carlo scheme implemented. We then compare the performance of our model to MD simulations of a similar system. Finally, the VMMC scheme of Whitelam and Geissler (2007) is implemented to improve the clustering behaviour of the simulations.

4.2 Monte Carlo simulation theory

Monte Carlo simulation is a technique to sample random microstates of a system from a probability distribution corresponding to a specific thermodynamic ensemble (Frenkel et al. 2002). The microstates in a system of N particles are defined by the set of centre of mass positions and momenta, and angular orientations and momenta for each of the particles. From classical statistical mechanics we know that the density of microstates should be

$$\rho\left(\mathbf{r}^N, \mathbf{p}^N, \mathbf{\Omega}^N, \mathbf{L}^N\right) \propto e^{-\beta\mathcal{H}(\mathbf{r}^N, \mathbf{P}^N, \mathbf{\Omega}^N, \mathbf{L}^N)}, \tag{4.1}$$

where $\mathbf{r}^N$ and $\mathbf{p}^N$ are the centre of mass positions and momenta of the particles; $\mathbf{\Omega}^N$ and $\mathbf{L}^N$ are the angular orientations and momenta of the particles; β is $1/k_BT$, and $\mathcal{H}\left(\mathbf{r}^N, \mathbf{p}^N, \mathbf{\Omega}^N, \mathbf{L}^N\right)$ is the Hamiltonian. The kinetic terms of the Hamiltonian are trivially integrated over, as they are quadratic in the momenta, resulting in a constant that is independent of position (Frenkel et al. 2002). This leaves the density of states as

$$\rho\left(\mathbf{r}^N, \mathbf{\Omega}^N\right) \propto e^{-\beta V(\mathbf{r}^N, \mathbf{\Omega}^N)}, \tag{4.2}$$

where $V\left(\mathbf{r}^N, \mathbf{\Omega}^N\right)$ is the potential function; it is from this distribution that we sample when using Monte Carlo methods. The Monte Carlo scheme developed by Metropolis et al. (1953) samples states from this distribution by firstly randomly generating a trial state; this trial state is conditionally accepted with a specified probability, whose form results from the need to satisfy detailed balance. Detailed balance requires that for any transition between two states, the ratio of the forward and reverse transition rates is equal to the ratio of the probability of being in those states, this statement is captured by the equation

$$\pi(\mu)P(\mu \to \nu) = \pi(\nu)P(\nu \to \mu), \tag{4.3}$$

where, for states μ and ν, π is the probability of being in the given state, and P is the probability of transitioning between the states in the given direction. The probability of the transition is the product of the probability of generating the move and the probability of accepting the move. Assuming that the probability of generating the move is the same in each state, then

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{P_{\text{gen}}(\mu \to \nu)P_{\text{acc}}(\mu \to \nu)}{P_{\text{gen}}(\nu \to \mu)P_{\text{acc}}(\nu \to \mu)} = \frac{P_{\text{acc}}(\mu \to \nu)}{P_{\text{acc}}(\nu \to \mu)} = \frac{\pi(\mu)}{\pi(\nu)}, \tag{4.4}$$

and the ratio of probabilities of being in the two states is equal to the ratio of their Boltzmann factors. This is given by

$$\frac{P_{\text{acc}}(\mu \to \nu)}{P_{\text{acc}}(\nu \to \mu)} = e^{-\beta(E(\nu)-E(\mu)}, \tag{4.5}$$

where E is the energy of being in the given state. Metropolis et al. (1953) proposed that the acceptance probability for the trial state, which satisfies the above equation, should be of the form

$$P_{\text{acc}} = \min\left\{1, e^{-\beta(E(\nu)-E(\mu))}\right\}. \tag{4.6}$$

One of the benefits of Monte Carlo simulations is that it does not involve integrating forces, as is required in the MD simulations in Chapter 3. This means we are able to make larger transitions between trial steps without having to worry about the stability of the simulation and each iteration only involves the calculation of the energy of a particular state, but we are limited to making moves that have a reasonable probability of acceptance. However, as a result of there being no equations of motion, the trajectory generated between states does not necessarily represent the realistic dynamics of the system. If we only make local transitions between states, the system trajectory will approximate realistic diffusive dynamics (Whitelam 2011). Yet for this to be applicable, the energy between states needs to be slowly varying, which is something that we cannot

guarantee, and do not expect for a system of proteins in a membrane. As a result of this, in systems with strong interactions, Monte Carlo simulations under-sample the motion of clusters. This occurs because for a cluster of tightly bound particles to diffuse, each particle in the cluster needs to individually be moved in the same direction. It is likely that such a series of moves would result in high energy intermediate states, which would make such movements unlikely. To encourage the movement of clusters of particles, which are strongly interacting, various cluster based Monte Carlo schemes have been developed. These schemes either propose some static linking scheme each for each interaction (Wolff 1989; Troisi et al. 2005b) or link clusters based on the gradients of the interactions between particles (Whitelam and Geissler 2007).

4.3 Parameterizing a discrete model of NanC diffusion

Here we describe the model representing our lipid bilayer system, which is used throughout this chapter and built upon in Chapter 5. The features described herein are ubiquitous in this thesis, but the implementation of transitions between states and other features of the simulation method will vary as we aim to capture more of the complexity of protein movement and interaction within a bilayer environment.

4.3.1 The discrete model

As in Chapter 3, we consider the case of proteins embedded in a bilayer environment. In the MD simulations we represented the proteins by a reduced set of atomic coordinates, where approximately four non-hydrogen atoms were represented by one coarse-grain particle. Here we reduce the number of degrees of freedom even further, to the limit at which we can still resolve individual proteins. This is done by representing the proteins by a single point at their centre of mass. So for our system of N proteins, their translational

state is fully described by the set of positions of their centres of mass

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, \tag{4.7}$$

where $\mathbf{x}_i$ is the position of the ith protein's centre of mass.

In Chapter 3, our proteins were embedded in a bilayer that consisted of lipid molecules, which were themselves coarse-grained using the same ratio of atoms to coarse-grained particles as for the proteins. The interactions between the particles of the lipids and those of the proteins were explicitly defined for the system. In our discrete model we implement an implicit lipid environment, where the effect of the lipids on the motion of and interaction between proteins is defined through appropriate parameterization of the model for protein motion and protein-protein interaction, rather than by explicitly accounting for the positions, motions and properties of individual lipid molecules in the bilayer and the water and ions in the surrounding solvent. So the translational state of the entire system, not just the proteins, is in fact described by X.

Previously, our MD system was represented in three spatial dimensions. Yet, the motion of and interaction between the proteins is mostly captured by their positions within the plane of the lipid bilayer itself; there is no protein movement outside of the bilayer. The limiting of many features of protein movement and clustering to their position in the plane of the bilayer can also be highlighted by the distribution of the protein's pore axis angle. In order to calculate this we used a coarse-grained MD simulation of a single NanC to measure the deviation of the protein's pore axis from the membrane normal. The distribution of this angle is shown in Figure 4.1. The simulation data used to calculate this angle distribution is the same simulation data that was presented in Chapter 3, which was used to calculate the lipid distribution around an individual protein. In Figure 4.1 we can see that there is only a relatively small angular deviation of $19^\circ \pm 6^\circ$. It is for these reasons that we chose to represent our membrane in two dimensions, with the proteins'

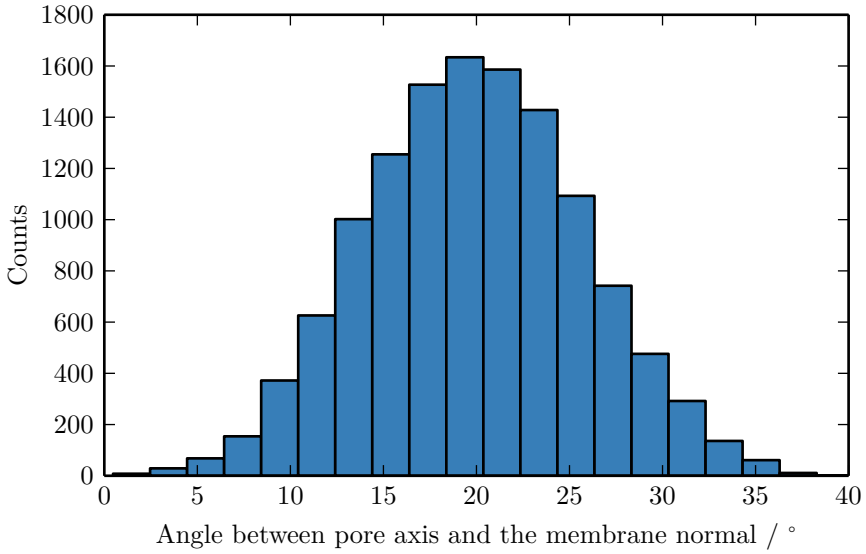

Figure 4.1: Distribution of pore axis angle measured relative to the membrane normal. The distribution has a mean of 19° and a standard deviation of 6°.

positions, X, being described by two-component vectors specifying the proteins' locations in the plane of the membrane. Our membrane system has periodic, toroidal boundaries, much like those that are employed in the MD simulations in Chapter 3.

The interactions between the proteins in our model are described by pairwise interactions. The interaction energy, ϵ_{ij}, for two isolated proteins, i and j, in our model is a function of the positions of those two proteins,

$$\epsilon_{ij} \equiv \epsilon_{ij}\left(\mathbf{x}_i, \mathbf{x}_j\right). \tag{4.8}$$

The final feature of our model that we need to introduce is our implementation of a simple three-body interaction scheme, similar to that employed by Yiannourakou et al. (2010). So although the interaction energy between proteins is pairwise, we implement a three-body scheme that introduces a simple occlusion of proteins. If there are any proteins occupying the region between the proteins i and j, then $\epsilon_{ij} = 0$. This has the effect of restricting any interactions between proteins to those that have a direct

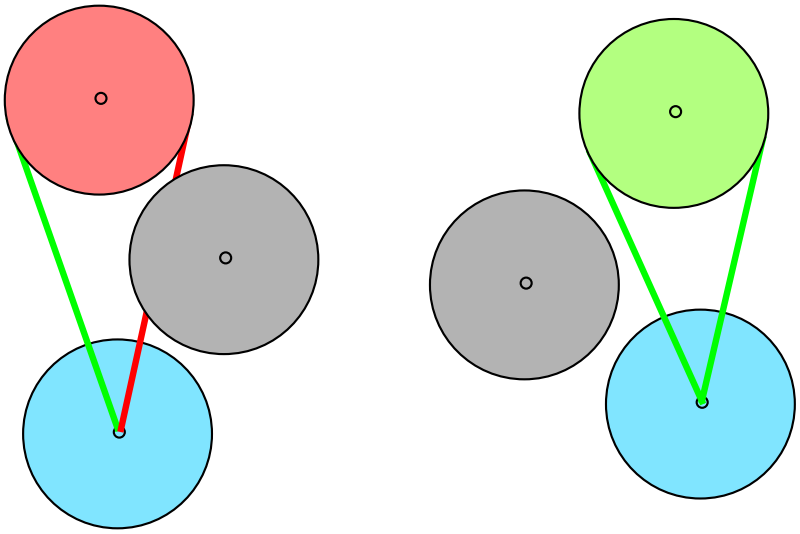

Figure 4.2: An illustration of protein occlusion is given using two different cases, both involving three proteins: one blue, one grey and one red/green. The red protein is occluded from its corresponding blue protein, whereas the green protein is not occluded from its corresponding blue protein. In both cases, the grey protein is a nearby protein which has already been considered when calculating the interaction energy. For the trio of proteins on the left, the grey protein acts to occlude the red protein from the blue, as its edge crosses the red line, drawn from the centre of our protein of interest. For the trio of proteins on the right, the grey protein does not cross either of the two green lines, and therefore the green protein is not occluded by the grey.

line-of-sight to each other. The three-body occlusion is illustrated in Figure 4.2.

We have chosen to implement this three-body effect as the interactions observed in Chapter 3 were dominated by lipid effects, with any electrostatic or van der Waals interaction between two proteins having no direct effect on their interaction. The exception to this behaviour occurs at very short inter-protein separations, but in such circumstances there is no possibility for another protein to separate the two proteins being considered, so no occlusion can occur.

4.3.1.1 Sampling configuration space using physically-realistic trial moves

To simulate the movement and interaction of the proteins embedded in our two-dimensional implicit membrane, we need to define a set of rules that govern how they move. We evolve the system by defining some discrete time scale, such that at time t, when the proteins are at positions X^t, we apply our chosen rule for updating positions and advance time by δt, which leaves the proteins at $X^{t+\delta t}$.

We implement a Monte Carlo simulation scheme by randomly sampling the system's configuration space to study its behaviour. Above we stated that in the case of small transitions between proposed states, where the individual particles in the system are moved using small local movements, the application of a Monte Carlo sampling scheme reproduces the diffusive dynamics observed when integrating the laws of motion.

The transitions involve small translations of individual proteins in a random direction by an amount sampled uniformly from the range $[0, \delta l]$. By sampling in this manner we ensure that there is no preferentially chosen directions in which the proteins will move.

To ensure physically realistic moves, we restrict the maximum translation distance, δl, to some fraction, $0 < \delta \leq 1$, of the protein diameter, σ, where $\delta = \delta l/\sigma$. In Section 4.4 we calculate the bounds on δ that are required for our scheme to result in simulation trajectories that are independent of our choice of δ.

Our bilayer model is evolved by continuously iterating over the proteins in the system and proposing individual trial moves for them. We consider one Monte Carlo step to be the proposal of a trial translation for each protein in the membrane, on average. This is on average because we select the protein for which we will propose a transition in a random manner, therefore, in any given Monte Carlo step, one protein may move more than once, or not at all. One Monte Carlo step is the fundamental time-step in our discrete simulations.

4.3.1.2 Criteria for accepting trial states

As explained above, the Monte Carlo technique involves two stages. Firstly, the proposal of a transition from the current state to some trial state, which we outlined above. Secondly, the trial state is assessed using some criterion, which results in either the acceptance of rejection of the trial state. The Metropolis criterion for assessing a trial move, resulting from a transition from state μ to a state ν, is defined by the probability, P_{acc}, which was given in Equation (4.6). The model includes only pairwise interaction energy terms between proteins, so the total energy, E, for a certain state is given by

$$E = \frac{1}{2} \sum_{\{i,j\}} \epsilon_{ij}, \tag{4.9}$$

where the sum is over the all pairs, $\{i, j\}$, of proteins. Using this criteria demands that we have a suitable characterization of the energy for proteins at a given centre of mass separation, which includes the effect of the lipids, water and ions. We discuss our parameterization of the pairwise interaction energy in Section 4.3.2.

4.3.1.3 Interpreting the time scale

Our membrane model can be evolved using the procedures outlined above, but our intrinsic time scale, the Monte Carlo step, is highly dependent on our choice of the maximum allowable translation, δl. To apply a physically realistic time scale to the simulation of our model we follow the approaches of both Romano et al. (2011) and Jabbari-Farouji et al. (2012) by using a scaling approach to regain a realistic time scale for our simulations.

The first of the scaling approaches discussed, developed by Jabbari-Farouji et al. (2012), relates the length of time for a single Monte Carlo step to the mean squared deviation

(MSD) of the particles. After n Monte Carlo steps the MSD, $\langle \Delta r^2(n) \rangle$, which is given by

$$\langle \Delta r^2(n) \rangle = 1/N \sum_{i=1}^{N} |\mathbf{x}_i(n) - \mathbf{x}_i(0)|^2, \tag{4.10}$$

should vary linearly with n, although the slope of the linear relationship may vary depending on the whether the diffusive regime is the short-time or long-time regime (whether the particles have interacted with each other or not). For sufficiently small n the MSD should be governed by the infinite-dilution diffusion coefficient, which for the two-dimensional case should result in $\langle \Delta r^2(n) \rangle \simeq 4D_0(n\delta t)$, where D_0 is the infinite-dilution diffusion coefficient and δt is the length of time for a single Monte Carlo step. Following this reasoning, Jabbari-Farouji et al. (2012) obtain a Monte Carlo step size of

$$\frac{\delta t}{\tau_B} = \lim_{n \to 0} \frac{\langle \Delta r^2 (n) \rangle}{n\sigma^2}, \tag{4.11}$$

where τ_B is the Brownian time scale, defined as the time required for the protein with an infinite-dilution diffusion coefficient, D_0, to diffuse over its diameter, σ, and is $\tau_B \equiv \sigma^2/(4D_0)$.

The second time-scaling method used is that of Romano et al. (2011), where the length of time for a Monte Carlo step is scaled based on the maximum translation distance, δl; the protein diameter, σ; and the rate of trial move acceptance, A, and is given by

$$\frac{\delta t}{\tau_B} = A\frac{\delta l^2}{\sigma^2}. \tag{4.12}$$

This scaling is based on the relationship found by Sanz et al. (2010), which related the MSD to the current acceptance rate by $\langle \Delta r^2(n) \rangle \simeq A\delta l^2$ for small δl. In this scaling it is assumed that rejected moves have the effect of slowing down the physical time for each Monte Carlo step, because the proteins will not have moved as much.

Using both of these approaches to apply a physical time scale to Monte Carlo simulations

of our model requires that we have a suitable characterization of the diffusion coefficient in the infinite-dilution limit. We discuss our application of the infinite-dilution diffusion coefficient to scale the time steps in Section 4.3.3.

4.3.2 Characterizing the pairwise protein-protein interaction energy using the potential of mean force

Having defined our model, we now need to define the functions that we will use to describe the protein-protein interaction, along with the parameters of those functions. In this section we will use the results of Chapter 3 to parameterize the pairwise protein-protein interaction energy.

4.3.2.1 Why the PMF is a suitable characterization of the protein-protein interaction energy

We stated above that in order to effectively apply the Metropolis criterion for judging whether trial moves should be accepted or rejected, we needed to have some measure for the pairwise energy of interaction for two proteins at a given separation in the membrane. This energy needs to account for the effect of the lipids in the bilayer, and the water and ions that surround it. In Chapter 3 we calculated the potential of mean force (PMF) for two NanC proteins in a bilayer as a function of separation. The PMF gives the change in free energy as a function of some reaction coordinate, or set of coordinates. It was defined in Chapter 3 as the potential on a sub-set of the system that results from averaging over the force on that sub-set due to the rest of the system, where the force here comes from the coarse-grained MD forcefield used to calculate the PMF. So the PMF calculated in Chapter 3, which was calculated as a function of inter-protein separation (defined by the separation of the proteins' centres of mass), gives us the change in energy that results from changing the protein separation and takes into account the force contributions from all of the lipids in the bilayer and the water and ions that surround the bilayer. This means

we are able to use the PMF as our measure for the pairwise protein-protein interaction energy because it was originally defined to estimate an energy that is equivalent to the energy between proteins in our discrete model.

4.3.2.2 Representing the PMF by a function of inter-protein separation

In order that we can use the PMF for rotationally unrestrained proteins, we need to obtain an appropriately fitted function of the inter-protein separation that adequately represents the data. If we are to use the PMF in our model we need to be able to obtain its value for any inter-protein separation.

In order to get a good fit of the PMF data, we have used the sum of two functions, through which we have tried to capture the main features of the profile. Firstly, we used a Morse potential to fit the potential well of our PMF, which is generally used to model bound states of atoms, but enables us to match the position, depth and width of the well. The Morse potential has the form

$$V_{\text{Morse}} = \epsilon_{\min} \left(1 - \exp\left[-\sigma_{\min}(r - r_{\min})\right]\right)^2 , \tag{4.13}$$

where $\epsilon_{\min}$ is the depth of the potential well, $r_{\min}$ is the location of the minimum, and $\sigma_{\min}$ controls the width of the potential well. However, it is more commonly used in its offset form

$$\begin{aligned} \hat{V}_{\text{Morse}} =& V_{\text{Morse}} - \epsilon_{\min} \\ =& \epsilon_{\min} \left(\exp\left[-2\sigma_{\min}(r - r_{\min})\right] - 2\exp\left[-\sigma_{\min}(r - r_{\min})\right]\right) , \end{aligned} \tag{4.14}$$

where $\lim_{r\to\infty} \hat{V}_{\text{Morse}} = 0$. Throughout this work, we exclusively use the offset form of Equation (4.14), so we set $V_{\text{Morse}} \equiv \hat{V}_{\text{Morse}}$ for ease of representation. Secondly, we used a Gaussian potential to fit the small energy barrier that occurs at an inter-protein

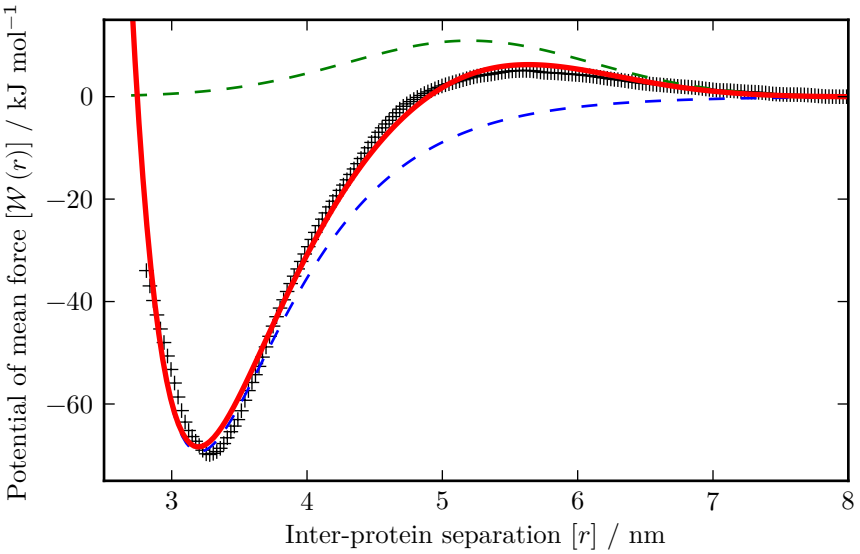

Figure 4.3: Fitting the PMF data (black crosses) using a Morse potential (dashed blue line) and a Gaussian (dashed green line). The combined function is shown by the solid red line.

separation of approximately 5.5 nm. The Gaussian potential has the form

$$V_{\text{Gauss}} = \epsilon_{\text{barrier}} \exp\left[-\frac{(r - r_{\text{barrier}})^2}{2\sigma_{\text{barrier}}^2}\right], \tag{4.15}$$

where $\epsilon_{\text{barrier}}$ is the height of the Gaussian, r_{barrier} is the location of the Gaussian maximum, and σ_{barrier} controls its width. Both these functions are shown by dashed lines in Figure 4.3.

The combined interaction potential between the proteins in our model is thus given by

$$\begin{aligned} V =& V_{\text{Morse}} + V_{\text{Gauss}} \\ =& \epsilon_{\text{min}} \left(\exp\left[-2\sigma_{\text{min}}(r - r_{\text{min}})\right] - 2\exp\left[-\sigma_{\text{min}}(r - r_{\text{min}})\right]\right) + \\ & \epsilon_{\text{barrier}} \exp\left[-\frac{(r - r_{\text{max}})^2}{2\sigma_{\text{barrier}}^2}\right], \end{aligned} \tag{4.16}$$

which is a function of six parameters: the height, location and width of both the Morse and Gaussian potentials. The potential function was fitted to the PMF data using a least

Parameter	Value
ϵ_{min}	69.3 kJ mol^{-1}
$\epsilon_{\text{barrier}}$	10.9 kJ mol^{-1}
r_{min}	3.2 nm
r_{barrier}	5.2 nm
σ_{min}	1.5 nm^{-1}
σ_{barrier}	0.9 nm

Table 4.1: Parameter values obtained from a least squares fitting of the interaction potential given by Equation (4.16) to the PMF calculated using MD simulation in Chapter 3. The PMF and the interaction potential are shown in Figure 4.3.

squares method implemented in the optimization package of the Python *SciPy* library[3]. We wished to ensure that both V_{Morse} and V_{Gauss} had physically meaningful forms, so that they were both realistic potentials in isolation. This can be seen by the fact that both the blue dashed line and the green dashed line follow the shape of the respective features reasonably closely. To do so we were required to place some constraints on the fit, which were: 2.5 nm $\leq r_{\text{min}} \leq$ 4nm and 4 nm $\leq r_{\text{barrier}} \leq$ 6nm. Without any constraints, the forms of V_{Morse} and V_{Gauss} that result in the best fit to the data were unphysically large and in unrealistic locations. By constraining the parameters we are thus able to alter the potential in a more manageable manner by modifying the parameters of the individual component. For example, a doubling in the well depth requires an approximate doubling of ϵ_{min} and a halving of the barrier height requires an approximate halving of the value of $\epsilon_{\text{barrier}}$. With a more precise fit to the data, but using unrealistic components, we would not be able to have as much control over the inter-protein interaction. The results of the parameter fitting are given in Table 4.1.

[3] `http://www.scipy.org`

4.3.3 Estimating the infinite-dilution diffusion coefficient required for time-scaling

Having specified the form of the protein-protein interaction energy above, the other parameter that we need to specify before we can perform simulations of our model is the diffusion coefficient in the infinite-dilution limit.

Since our model does not include any hydrodynamic effects, we can use MD simulation of a single protein to calculate a value for the diffusion coefficient. We attempt to approximate the infinite-dilution limit by considering the motion of a single protein in a large simulation box. There will be some finite size effects in such a system, but by choosing a large simulation box, we hope to minimize these. Using a 5 μs coarse-grained MD simulation of a single NanC in a periodic POPE bilayer with a simulation box side length of 28.3 nm, corresponding to a protein area fraction of 1.1%, we calculated the mean squared displacement (MSD) of the protein. The MSD gives a measure of the diffusion of the protein. In a given time step a protein may be displaced by some amount δr. The mean squared displacement at time t_1 is thus given by

$$\langle \Delta r^2 \rangle = \langle \sum_{t=0}^{t_1} \delta r \left(t\right)^2 \rangle, \tag{4.17}$$

where the mean, $\langle ... \rangle$, is taken over all of the proteins in the system. The MSD calculated using MD is shown as a function of simulation time in Figure 4.4, where we can see that it is linear. For free diffusion in a two-dimensional region, the long-time relationship between the MSD, $\langle \Delta r^2 \rangle$, and the time, t, is given by

$$\langle \Delta r^2 \rangle = 4 D_0 t, \tag{4.18}$$

where D_0 is the translational diffusion coefficient. By fitting a straight line to the data in Figure 4.4, we are able to calculate an estimate of the infinite-dilution diffusion coefficient,

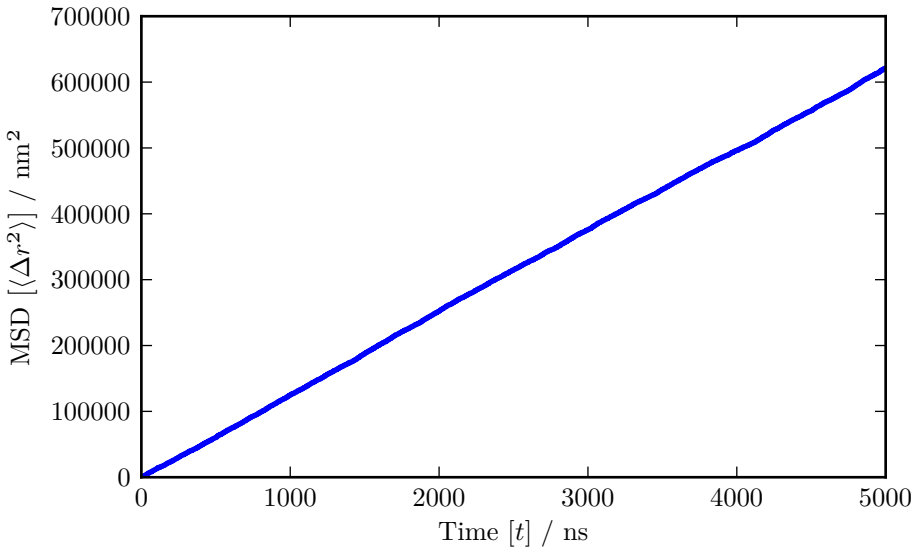

Figure 4.4: Mean square deviation (MSD) of a single NanC as a function of time, t, obtained from a 5 μs coarse-grained MD simulation of a 28.3 nm-square periodic pure POPE bilayer.

D_0, of 0.031 nm^2 ns^{-1}.

The limitation in our approach to characterizing the diffusive behaviour of the proteins using coarse-grained MD simulations is that the movement is sped up by the coarse-graining process. This is as a result of the smoothing of the inter-atomic interaction potentials and accounts for an approximately four-fold increase in the speed of coarse-grained MD simulations (Monticelli et al. 2008). However, since we will compare simulations of our model to coarse-grained MD simulations in Section 4.5, which suffer from the same time-scaling issues, this should not be a problem.

4.4 The convergence of system statistics and the regions of parameter validity

In this section we use Monte Carlo simulations to calculate the MSD of the proteins in our discrete individual-based model and assess the convergence of its error with increasing

number of repeat simulations. We then use the MSD to investigate the choice of the step size, δ, used in our simulations and what effect this choice has on our ability to implement the diffusion-based time-scaling scheme in Equation (4.11). Finally, we compare the performance of the diffusion-based time-scaling scheme with that of the acceptance-based time-scaling scheme introduced in Equation (4.12). To do this we analyse the overlap of the MSD trajectories for various values of δ, which, in the ideal case, should overlap perfectly.

4.4.1 Simulating the system

The initial state of our bilayer system consists of a number of proteins arranged in a square lattice formation. To vary the system density, we can change the initial inter-protein separation. In this section we use initial inter-protein separations ranging from 6 nm to 8 nm, which correspond to a protein area fraction approximately between 12 and 22%.

As outlined in Section 4.3, our system is evolved using the Monte Carlo method, where we implement the Metropolis acceptance criterion given in Equation (4.6) and generate trial states of the system by applying small transitions to individual proteins. Given the set of parameters describing our model via the interaction energy, ϵ_{ij}, and the protein self-diffusion, D_0, we are able to control the simulation by varying either δ or the system temperature, T. However, our values of ϵ_{ij} and D_0 were both calculated using a system temperature $T = 310$ K, so we shall use this value for all of our simulations.

The simulations of the model were performed using a simulation program written in Scala. Scala runs on the Java Virtual Machine, making it easy to run the software on a variety of systems. Snapshots of the simulation of 256 proteins after 100 and 1000 Monte Carlo steps are shown in Figure 4.5.

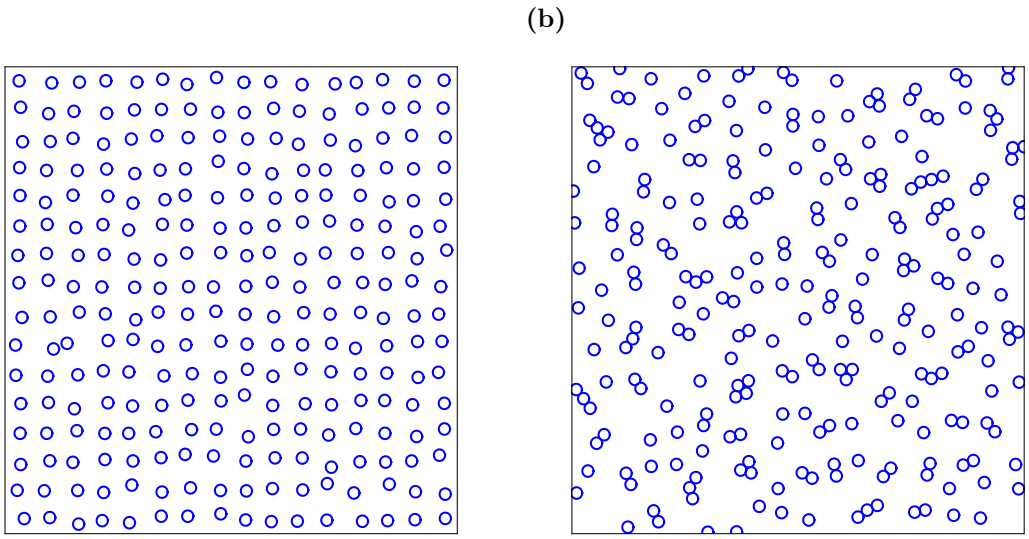

Figure 4.5: Positions of 256 proteins distributed in a planar and periodic patch of membrane. The proteins were initially placed on a square grid with an inter-protein spacing of 8 nm. **(a)** Positions of the proteins after 100 Monte Carlo steps. **(b)** Positions of the proteins after 1000 Monte Carlo steps.

4.4.2 Characterizing the simulation data

The output from our simulation is a time-series of spatial coordinates for each protein in the system, along with a record of the acceptance rate for our proposed trial moves. For any given protein, the trajectory will differ significantly from its neighbours and, given that the random numbers generated for repeat simulations are independent, the trajectories will differ significantly for the same protein in different simulations. Given that the individual protein trajectories are so varied, in order to analyse the system we must characterize it by its bulk properties. To use these bulk properties to describe our system we should have some idea about how they converge with an increasing number of independent repeat simulations. If a given bulk property and its error do not converge, then it will not be a particularly useful descriptor of our system.

To demonstrate the calculation of bulk properties for the system and the analysis of their convergence, we calculate the proteins' MSD, $\langle \Delta r^2 \rangle$, defined in Section 4.3.3.

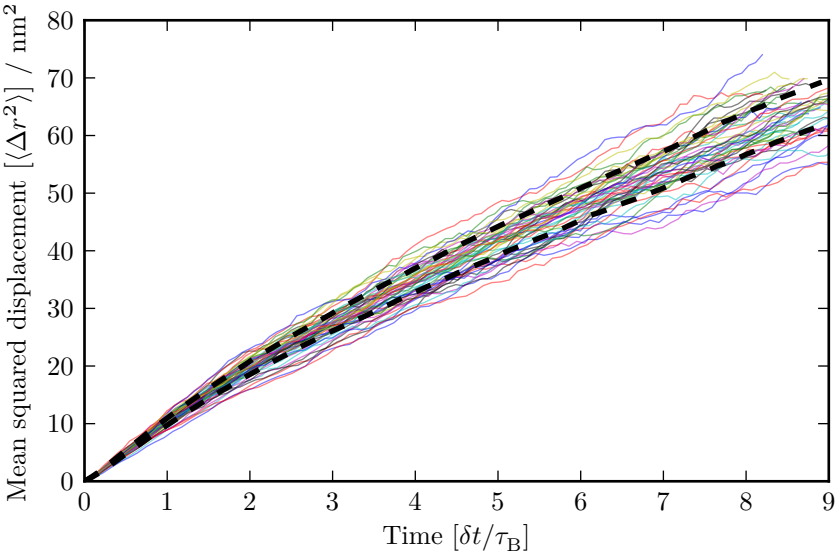

Figure 4.6: Mean squared deviations for repeat runs of a simulation of 256 proteins with $\delta = 0.03$ and a protein density of approximately 14%. Only 48 of the 192 simulations are shown here, for clarity. The region between the black dashed lines is one standard deviation either side of the mean of the data.

We must take care to track the position irrespective of any boundaries, such that the value of r is the sum of the individual transitions that were accepted for a given protein, rather than the closest separation between the current position of a protein and its initial position.

In Figure 4.6 we show the MSDs, taken over a system of 256 proteins for a series of repeat simulations. Here we can see that the MSD calculated for different simulated systems are not identical; they follow a similar profile, but do not overlap significantly. This highlights the need to run repeats for stochastic simulation, if we are to obtain reliable estimates for the bulk statistics of the system. There are 48 example trajectories shown in Figure 4.6, taken from a set of 192 repeats. The dashed lines show the region that is one standard deviation away from the mean MSD across the data set.

In Figure 4.6 and throughout this chapter, we measure time in units of the Brownian time scale, which was introduced in Equation (4.11) as the time it takes for a protein to diffuse a distance equal to its diameter. This is a useful time scale to use as it enables

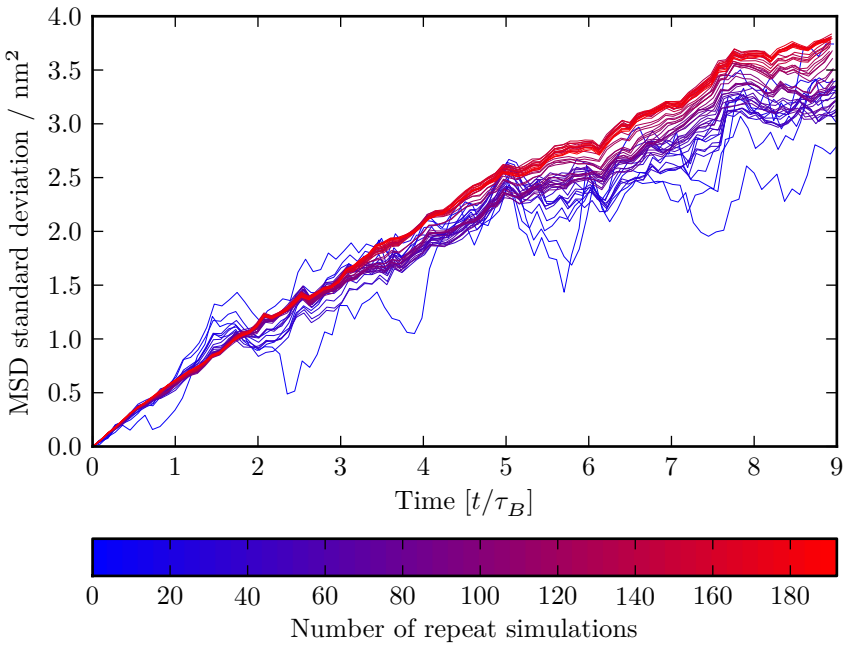

Figure 4.7: Standard deviation of the mean squared displacement shown for an increasing number of repeat simulations. The data shown are for the addition of 4 repeat simulations for each trace plotted, up to a total of 192 repeat simulations.

quick comparison between results from simulations that use different parameters. It is, however, simple to transform to time in seconds using $\tau_{\mathrm{B}} \equiv \sigma^2/4D_0$ where σ is the protein's diameter and D_0 is its infinite-dilution diffusion coefficient.

Characterization of the error for a given statistic is an essential step in being able to interpret the behaviour of the system. Figure 4.7 shows how the value of the standard deviation of the MSDs in Figure 4.6 converges with an increasing number of repeat simulations. This result demonstrates that by repeating the same simulation many times, using independent random number sequences, we are able to evaluate bulk statistics for the system with a converged error.

4.4.3 Changing the step size and its effect on the validity of diffusion-based time-scaling

The step-size, δ, is one of the parameters we can change, which will have a direct effect on the evolution of our system. Yet our choice of δ, at least in the diffusion-based time-scaling of Equation (4.11), is restricted by the requirement that we are able to resolve the initial diffusive behaviour of the proteins, when they are not influenced by the effect of the surrounding proteins.

In Figure 4.8, we can see this short-time diffusive regime in the plot of the MSD for a system with a protein density of approximately 14% (corresponding to an initial protein separation of 7.5 nm). At short time scales the diffusion obeys the relationship shown in Equation (4.18), where we have added the black line to the plot to show how the MSD should evolve were it not for the presence of the other proteins. The diffusion matching scheme of Equation (4.11) requires that we are able to identify this short-time diffusive regime and calculate the mean squared deviation of the proteins as the number of Monte Carlo steps tends to zero. This places a requirement on our choice of δ, because if our maximum transition size is too large, we will not be able to resolve this short-time diffusive behaviour. Figure 4.8 illustrates the criteria applied to assess the suitability of a given step size for using the diffusion matching scheme. We stipulate that in order to be able to extrapolate a straight line fit to the initial section of the MSD, the MSD must not have deviated outside the region marked by the red dashed lines by the third Monte Carlo step[4]. It is clear from the zoomed-in inset axes that the MSD in this instance falls well within these bounds.

We have used this criterion to assess the choice of δ for a range of protein densities by performing a parameter sweep, recording an acceptability score in the range $[0, 1]$, the results of which are shown in Figure 4.9. The acceptability score is a measure of how

[4]This choice is somewhat arbitrary, but since we are scaling based only on the first Monte Carlo step, having a good straight-line fit for the first three seems a reasonable criterion.

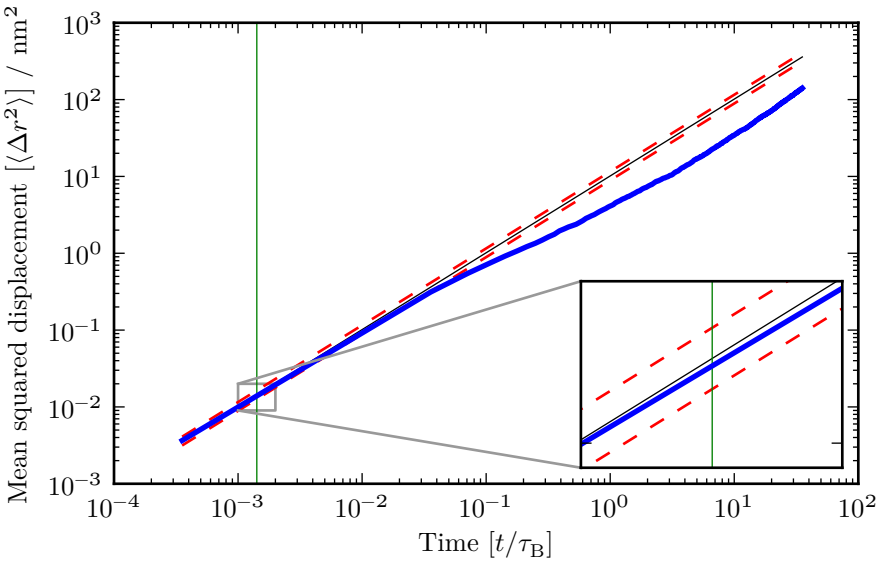

Figure 4.8: Mean squared displacement is plotted as the thick blue line on a logarithmic scale. We are more easily able to identify the region during which the proteins have not been affected by the presence of the others in the system. The thin black line shows the profile for normal diffusion of an isolated protein. We can see that the MSD for this system diverges from the line between $t/\tau_B = 10^{-2}$ and $t/\tau_B = 10^{-1}$. The dashed red lines indicate the limits of acceptability for the MSD to deviate from the ideal line, which is evaluated at the threshold indicated by the vertical green line. The acceptability score is linearly mapped to the range $[0, 1]$ based on how close the MSD is to either of the limits (red lines) at the location of the threshold (green line). A score of one means that the MSD is on the central line, whereas a score of zero means the MSD is on, or beyond, either of the two red boundary lines. From the zoomed inset axes we can see that this MSD is acceptable; it has a score of approximately 0.8.

much the MSD deviates from a linear relationship with the simulation time. When there is no deviation, the MSD is a linear function of the simulation time, the acceptability score is 1. When the deviation is equal to or greater than the region defined by the dashed red lines in Figure 4.9 after the third Monte Carlo step, the acceptability score is 0. This is a somewhat arbitrary measure, but the pattern in acceptability score for various simulation parameters, rather than the absolute values, is what is most indicative of the simulation behaviour. It is clear from Figure 4.9 that there is a significant portion of our parameter sweep that does not enable us to use the diffusion-based time-scaling method. The region with high protein density and high δ scores poorly. It is to be expected that it is in this region that we find the worst matching to the ideal diffusion scenario. For large values of δ the proteins will move more in each Monte Carlo step, and will, therefore, interact with the other proteins in the system sooner. Similarly, for the higher density systems, the proteins are initially placed closer to the other proteins and will also interact with other proteins sooner. Therefore, it is for the lower protein densities or for smaller transition step sizes that we will be able to match the short term diffusive motion to the expected diffusive behaviour in the infinite-dilution limit. These parameter combinations are those with the higher acceptability scores, which are shown in pale yellow and white in Figure 4.9.

4.4.4 Data collapse

Having identified the regions of parameter space allowed for both δ and the protein density, which should enable us to apply the diffusion-based time-scaling technique using the short-time infinite-dilution diffusion coefficient, we are able to see how both of the time-scaling methods perform by reducing some simulation data onto the same time scale using various values of δ. If our scaling has worked correctly, the MSDs for different values of δ should coincide.

From Figure 4.10 we can see that for the case of a system with the proteins initially

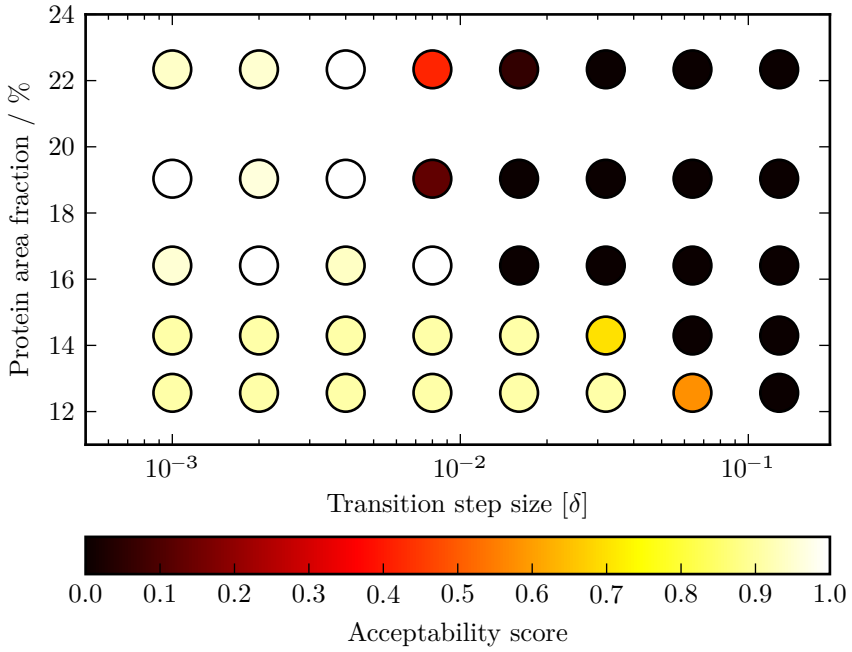

Figure 4.9: Acceptability scores for a range of values of the transition step size δ, and the protein density. A score of one means the MSD for the given parameters is very well suited to matching with the diffusion coefficient as the number of Monte Carlo steps tends to zero. The region with the worst scores is the region of high δ and high protein density.

distributed on a square grid at 8 nm separations we get good data collapse, across almost all values of δ, for both the diffusion and acceptance scaling regimes. It is only for the largest δ value shown that there is any noticeable deviation, and this was predicted by its poor acceptability score shown in Figure 4.9. However, in Figure 4.11 we can see an example of where the diffusion-based time-scaling breaks down, as the MSDs for different values of δ do not collapse well onto our real time scale. In the parameter sweep shown in Figure 4.8 the most dense system studied corresponded to a protein surface area accounting for approximately 22% of the membrane, which corresponds to the MSDs in Figure 4.11 with an initial protein separation of 6 nm. From our parameter sweep we find that only the first two or three values of δ should give acceptable scaling under the diffusion-based scheme. We see in Figure 4.11 that the smallest values of δ investigated do show some degree of appropriate collapse, but it is not as good as the data collapse observed with the larger initial protein separation shown in Figure 4.10. The MSDs for larger values of δ shown in Figure 4.11 do not result in satisfactory data collapse, which is as we expect given their poor acceptability scores. The acceptance-based time-scaling method produces better data collapse at larger values of δ as shown in Figure 4.11B, but there is still some discrepancy around the change in diffusive regimes in the case of larger values of δ.

4.5 Analysing the performance of Monte Carlo simulation of our model

In the previous section we looked at analysing simulations of our model using the MSD and how we are required to constrain our simulation parameters (δ and the protein density) based on the form of the MSD. In this section we extend our analysis of the model by investigating the clustering behaviour of proteins in the system. The clustering behaviour is of interest because many protein functions are affected by the presence of

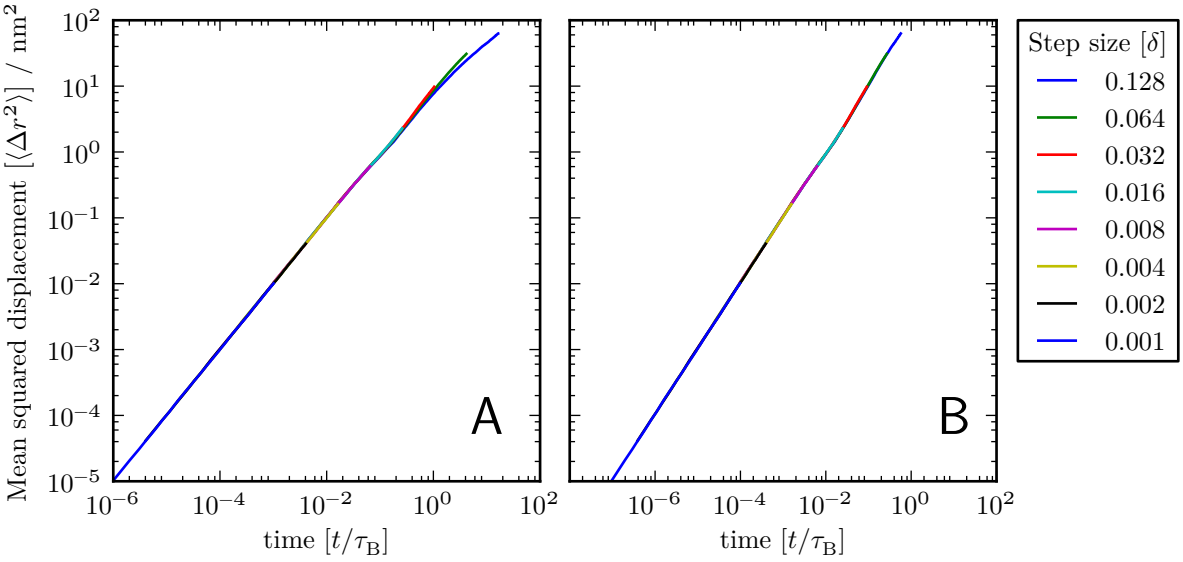

Figure 4.10: Overlaid MSDs for an initial protein separation of 8 nm. The MSD profiles overlap for both the diffusion (A) and acceptance (B) scaling. Note that the lines all start at the same point and are overlaid with the shortest step size on top so that the lines below are visible; using a shorter step size results in a shorter simulation time for a given number of steps.

other proteins in their immediate neighbourhood. It is also of relevance to the model in general, because if we were to extend it to try to capture the motion of protein subunits, rather than just individual proteins, it would require that we understand the clustering in order that we are able to make predictions about the formation of proteins and protein complexes from these subunits.

4.5.1 Protein clustering metrics

Here we describe the cluster analysis metric that we will use to compare the effect of changing features of our model and to compare the simulation results to those of other simulation paradigms, notably large-scale coarse-grained MD simulations.

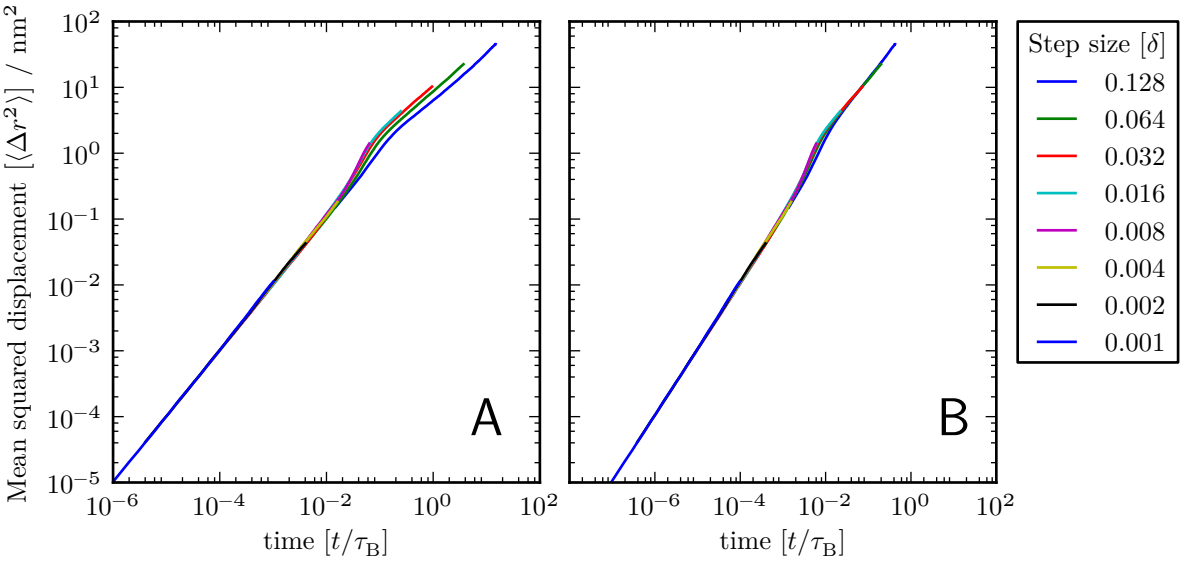

Figure 4.11: Overlaid MSDs for an initial protein separation of 6 nm. The MSD profiles do not overlap for both the diffusion (A) and acceptance (B) scaling. There is some agreement between the smaller values of δ, but for the larger values there is significant divergence. There is better agreement between data sets for the acceptance scaling (B), but this still breaks down for the largest values of δ. Note that the lines all start at the same point and are overlaid with the shortest step size on top so that the lines below are visible; using a shorter step size results in a shorter simulation time for a given number of steps.

4.5.1.1 Defining protein clusters

In order that we are able to characterize our bilayer system using some measure of protein clustering, we first need to have a concrete definition of what it means for proteins to be clustered. The most simple, and generally applicable, method for defining protein clustering is to consider the inter-protein separation. We calculate the clustering linkage hierarchy for a given time-step by joining the proteins together, from shortest inter-protein separation to longest, to form a tree structure. Inter-protein distances are considered between both individual proteins and proteins that have already been clustered, so that we have an ordering of cluster links in our linkage tree. We get from this tree structure to a desired flat clustering by defining some threshold separation beyond which the proteins are no longer considered to be in the same cluster. All the linkages in the tree that are formed from links that are shorter than this threshold are collapsed into our clustering. We have chosen to use a relatively tight threshold, with a maximum inter-protein separation of 3.3 nm, to define clustered proteins (the minimum of the potential well in Figure 4.3 is at an inter-protein separation of approximately 3.2 nm). This was done to ensure that our clustered proteins represent proteins in states of contact, where there would not be any room for lipids between them.

4.5.1.2 Protein clustering rate

Using our definition of clustered proteins, we are able to characterize the clustering state of the system by calculating the distribution of proteins among clusters of various sizes for any given simulation time-step. Figure 4.12 shows the results of analysing the distribution of cluster sizes for a simulation of 256 proteins initially arranged in a grid with spacing 7.5 nm (corresponding to a protein area fraction of approximately 14%) for 10^3 Monte Carlo steps with a transition step-size of $\delta = 0.03$, which is one of the acceptable parameter choices. We can see from Figure 4.12 that the fraction of proteins in a cluster of a given size appears to be converging for clusters of size one (individual

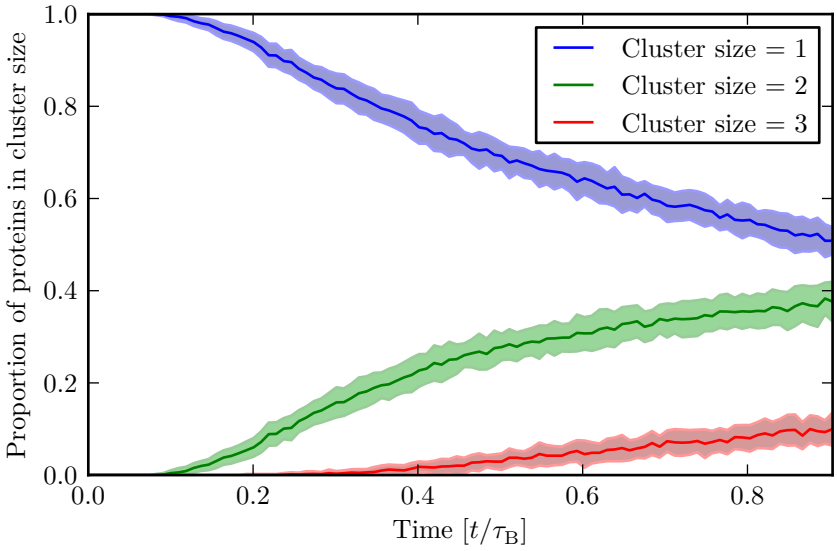

Figure 4.12: The distribution of proteins among clusters of size one, two and three, show as they vary with simulation time. Here we note that at the start, all the proteins are in clusters of size one, as the initial condition for the simulation is proteins in isolation. We can see that for all three cluster sizes the number of proteins in clusters of that size appears to be converging. The data was obtained from a 10^3 Monte Carlo step simulation of 256 proteins, with $\delta = 0.03$ and a protein are fraction of ∼14%. The simulation was repeated 192 times.

proteins), two and three proteins. Here we have not displayed the simulation data until convergence (this is shown later in the chapter) because we are interested particularly in validating our model parameters during the transient phase of its evolution.

4.5.2 Analysis of protein behaviour

4.5.2.1 Verification of step size independence for accepted step sizes

In Section 4.4.3 we calculated the regions of parameter space in which we would get good agreement between simulations of varying step size. To verify that our acceptability score for the step sizes was suitable, we ran simulations of the model with different step sizes so that we could compare them using a clustering metric. If the score was a suitable measure of acceptability for the step size, then when we use two differing values of δ, both with high scores, the trajectory of the metrics should agree. This is

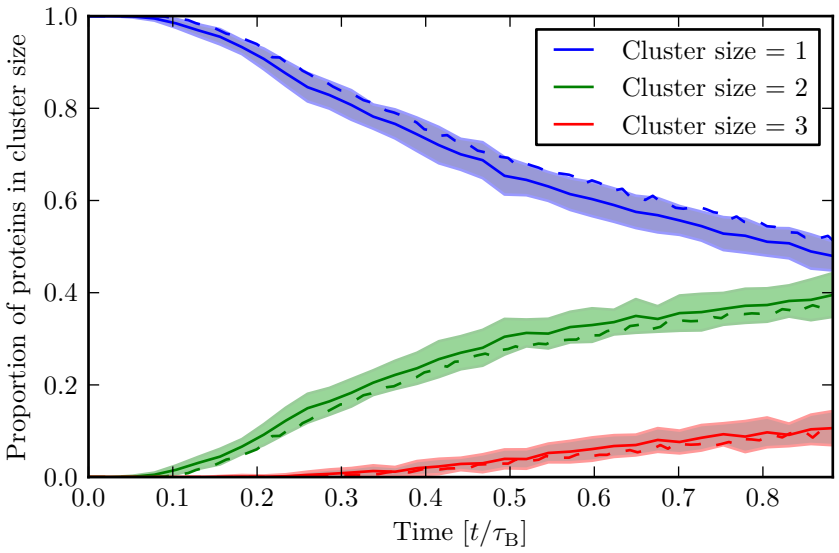

Figure 4.13: A comparison of the cluster size distribution for two simulations with a different value of the step size, δ. The dashed line is for $\delta = 0.03$ and the solid line is for $\delta = 0.016$, both of which had an acceptability score close to one. There is a good agreement between the two sets of simulation results. The shaded region represents one standard deviation for the data used to calculate the cluster sizes for $\delta = 0.016$.

shown in Figure 4.13, where we can see that there is good overlap between the cluster distributions for values of $\delta = 0.016$ and $\delta = 0.03$. We also tested the acceptability scores by comparing a simulation using a value of δ with a high score ($\delta = 0.03$), with a simulation using a value of δ with a low score ($\delta = 0.128$). Figure 4.14 shows the results of this comparison, in which we can see that there is poor agreement between the results from the two simulations. This comparison serves as a good test that our scoring method for our choice of δ is a good measure for identifying acceptable parameter regimes.

4.5.2.2 Effect of variations in the inter-protein interaction energy

We can investigate the properties of our model by comparing the cluster metrics after a modification of one of the model parameters. We modified two features of the interaction potential between proteins in order to understand the role they played in protein clustering. The first is the depth of the potential well in the interaction potential and the second

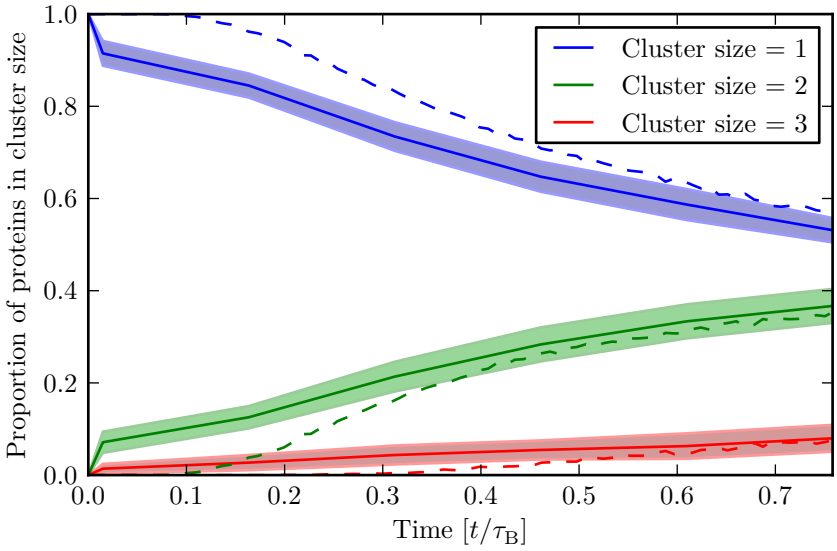

Figure 4.14: A comparison of the cluster size distribution for two simulations with a different value of the step size, δ. The dashed line is for $\delta = 0.03$ and the solid line is for $\delta = 0.128$, which had a low acceptability score. There is poor agreement between the two sets of simulation results. The shaded region represents one standard deviation for the data used to calculate the cluster sizes for $\delta = 0.128$.

is the height of the barrier (which occurs at an inter-protein separation slightly larger than that of the potential well). These modifications are shown in Figure 4.15 and the modified parameters are given in Table 4.2.

The effects of the various modifications on the cluster size distributions are shown in Figure 4.16. We can see from the distribution trajectories that each of the modifications has a different effect. Firstly, the modification of the barrier, shown in Figures 4.16(a) and 4.16(b), show that the rate at which the clusters form is strongly affected by the height of the barrier potential. A high energy barrier dramatically reduces the rate at which clusters form, whereas removing the barrier has the effect of increasing the cluster formation rate. This is perhaps expected, as the probability that a given protein gets close enough to another protein to form a cluster will be dependent on the energy barrier between them. In the case of no barrier we can see that the cluster distributions converge to the same value, and assume that the same would be the case for the double height

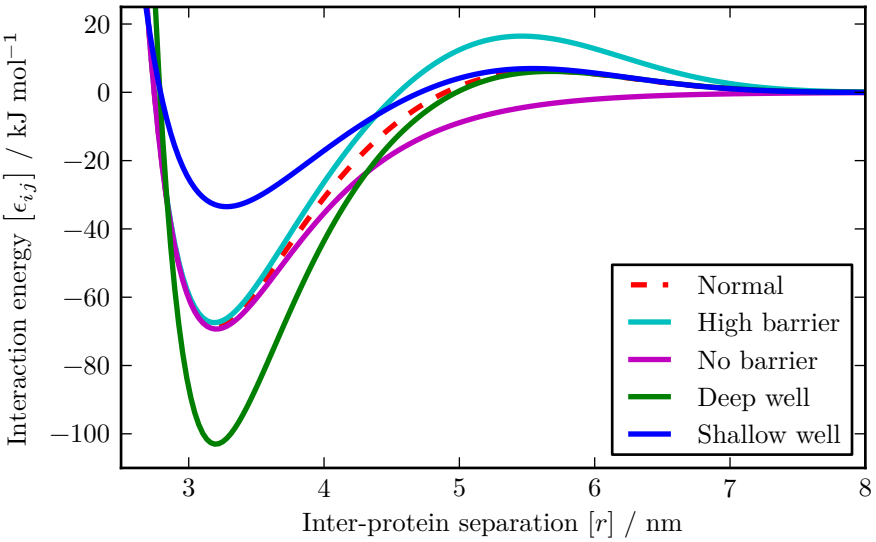

Figure 4.15: The modified inter-protein interaction energies used to test the effect of the various features on protein clustering are shown here. The red dashed line is the original parameterized form of the pairwise interaction energy. We modified it by: doubling the barrier height; removing the barrier; increasing the well depth by 50%; and reducing the well depth by 50%.

Parameter	High barrier	No barrier	Deep well	Shallow well
ϵ_{min}	69.3 kJ mol^{-1}	69.3 kJ mol^{-1}	138.6 kJ mol^{-1}	34.7 kJ mol^{-1}
$\epsilon_{\text{barrier}}$	21.8 kJ mol^{-1}	0.0 kJ mol^{-1}	10.9 kJ mol^{-1}	10.9 kJ mol^{-1}
r_{min}	3.2 nm	3.2 nm	3.2 nm	3.2 nm
r_{barrier}	5.2 nm	5.2 nm	5.2 nm	5.2 nm
a	1.5 nm^{-1}	1.5 nm^{-1}	1.5 nm^{-1}	1.5 nm^{-1}
b	0.9 nm	0.9 nm	0.9 nm	0.9 nm

Table 4.2: Parameter values for the interaction energy function in Equation (4.16) used to create the modified energy functions in Figure 4.15.

barrier if we ran a longer simulation. So we can surmise that the barrier height appears to control the rate of cluster formation, but not the equilibrium state of the system. Secondly, the effect of modifications to the potential well depth in our inter-protein interaction energy are shown in Figures 4.16(c) and 4.16(d). We can see that the modifications made to the potential well have a similar effect on the protein clustering rate. The effect of a deeper well leads to a faster formation rate, although the increase is much smaller for a 50% increase in well depth, than it is for removing the barrier. The effect of a shallower well has a more pronounced effect on the rate than a deeper well, leading to a decrease in the clustering rate. Where the effects of a modification of the well depth differs from that for the barrier height is in the long-term cluster distribution. We can see from Figures 4.16(c) and 4.16(d) that the equilibrium cluster distributions are different. From this we can infer that the equilibrium state of the system is affected by modifications in the absolute energy change between an isolated protein and a clustered protein, but only the dynamics of reaching that equilibrium are affected by modifying the barrier height.

4.5.2.3 A comparison with a large-scale coarse-grained MD simulation

We have looked at the behaviour of the model under modification of its parameters, but we have not yet analysed the performance of the simulation method to test how well it reproduces the actual behaviour of the system. We parameterized our model using coarse-grained MD data, therefore it is reasonable to expect our simulation technique to reproduce the behaviour of an equivalent system that was simulated in its entirety using coarse-grained MD. The data for comparison comes from a simulation performed by Dr. Joseph Goose, introduced in Chapter 1. The MD simulation represents 1000 ns of simulation time. The simulated system is the same as the one we have been using throughout this section: 256 NanC proteins with an initial separation of approximately 7.5 nm. However, the lipid bilayer in the large MD simulation is constructed from two different lipid species, POPE and POPG at a ratio of 3:1, whereas our PMF and diffusion

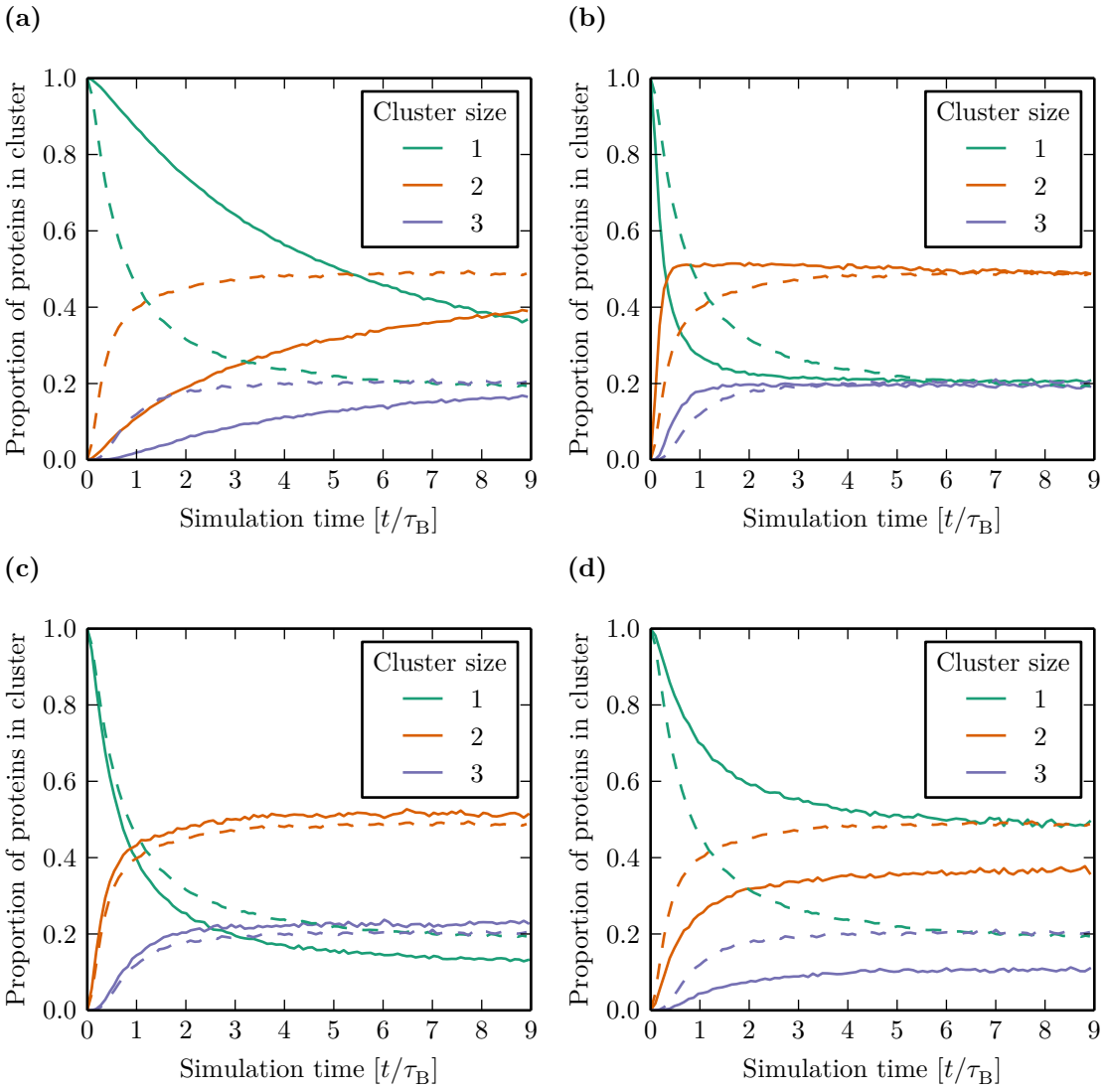

Figure 4.16: Comparing the effect on cluster size distributions during a 10^4 MCS simulation of 256 proteins, with $\delta = 0.03$, when we introduce modifications to the inter-protein interaction energy. The modified distributions are shown by the solid lines and the unmodified distribution by the dashed lines. The modifications are a high barrier **(a)**; no barrier **(b)**; a deep well **(c)**; and a shallow well **(d)**.

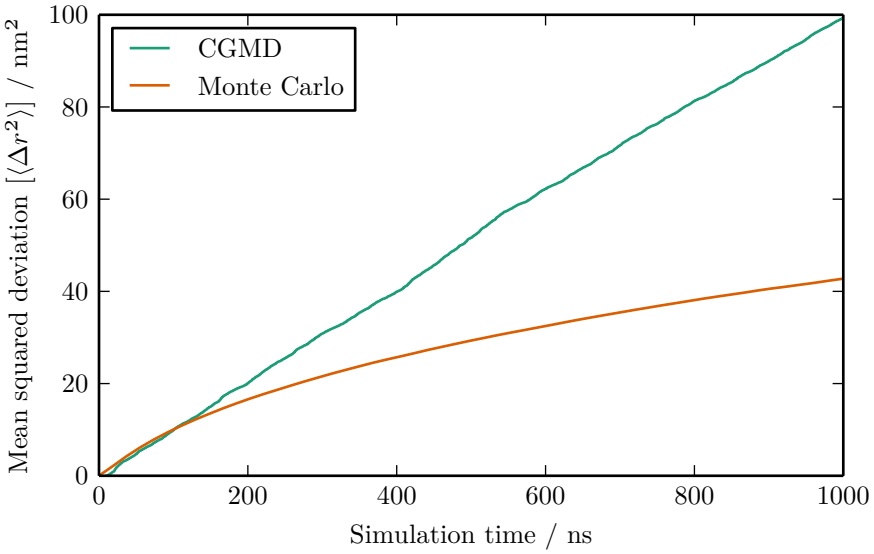

Figure 4.17: The MSD for the MD simulations is shown compared to the MSD for a Monte Carlo simulation of our model, using similar simulation parameters. The MSD for the MD simulation was calculated using a polling interval of 2 ns.

coefficient were calculated using MD simulations of NanC proteins in a pure POPE bilayer. The coarse-grained MD simulation took over a week to run on thousands of computer cores, whereas our simulations were run on single processors and took approximately 6 minutes to complete 10^4 MCS. If our model were to accurately represent the motion of NanC proteins in a bilayer, we assume that the difference in the lipid bilayer composition should only lead to minor discrepancies.

Figure 4.17 shows a comparison between the MSD for both the coarse-grained MD simulation and our Monte Carlo simulation. It is obvious that there is some discrepancy between the MSD for the two simulation methods. The reduced mobility could be due to the difference in lipid bilayer constitution, but this is only likely to cause a discrepancy in the gradient, rather than producing a different form of the MSD. The likely cause of this discrepancy is our use of an individual-based Monte Carlo scheme. In such schemes, especially for systems of strongly interacting particles, the movement of clusters is repressed.

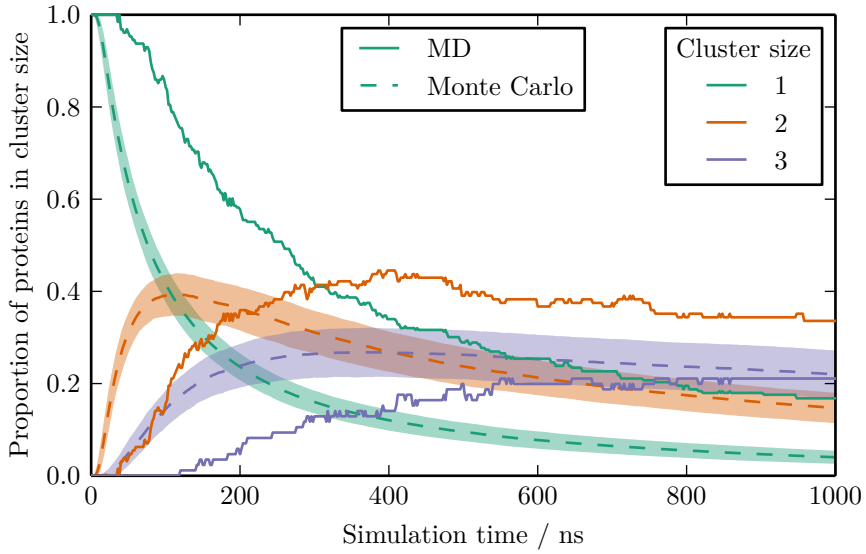

Figure 4.18: The cluster size distributions for the MD simulation (solid lines) are compared to the cluster sizes for a Monte Carlo simulation of our model (dashed lines). The cluster threshold for this comparison was set at 4 nm. It is clear that there is not very good agreement between the two sets of data. The solid coloured regions indicate one standard deviation for the Monte Carlo simulations.

We also compare the cluster size distribution between the simulations of our model and the coarse-grained MD simulations, where we have used a cluster threshold separation of 4 nm. This is shown in Figure 4.18, where it is again apparent that there is a significant difference between the results from our simulations and those from the coarse-grained MD. In this instance, the proteins in our simulation form clusters at a much faster rate than in the MD simulation. One other likely cause of this discrepancy is that we have chosen to ignore hydrodynamic interactions. The proteins in our system will move at random, even when there are proteins nearby. In reality, the proximity of other proteins will have the effect of reducing the mobility of the surrounding lipids, which in turn will reduce the protein mobility in that region.

4.6 Introducing rigid body moves of protein clusters

The individual-move Monte Carlo simulations performed above did not agree well with the large scale MD simulations with which we compared them. We argued above that one of the issues with the simulation of our model was that in only considering individual proteins for movement each time, we were unlikely to achieve any movement of clusters of proteins once they have formed. To attempt to fix this shortcoming, we have implemented a cluster-based Monte Carlo scheme, which aims to improve the dynamical realism to the exploration of configuration space. The scheme we have implemented is called virtual move Monte Carlo, developed by Whitelam and Geissler (2007), and it creates moving clusters of proteins based on the gradient of the pairwise interactions between proteins, rather than only based on the energy of the initial state as done in other schemes (Wolff 1989; Troisi et al. 2005b).

4.6.1 Virtual move Monte Carlo

Dynamic cluster-linking schemes in Monte Carlo simulations are designed to iteratively build up a cluster, for which a trial move will be proposed, by assessing the consequences of the move in such a way that the linking is not strongly coupled to the particle configuration in the initial state (Wolff 1989). In the virtual move Monte Carlo scheme particles that interact with the current cluster are added in an iterative manner; their addition is based on the difference in the energy change resulting from the movement of the cluster if the new particle were and were not added. The algorithm proceeds as follows and is based on the description presented in Whitelam (2011).

Starting with a randomly chosen seed particle, and a randomly chosen transition (translation or rotation) for that particle, which moves it from a state μ to a state ν, we recursively consider links between particles, i, that are in the current cluster realization, $\mathcal{C}$, and those particles that are not, j. We attempt to form a partial link between the

particles i and j, which occurs with a probability

$$p_{ij}\left(\mu \to \nu\right) = \Theta\left(n_c - n_{\mathcal{C}}\right) \mathcal{I}_{ij}\left(\mu\right) \max\left\{0, 1 - e^{\beta(E_{ij}(\mu) - E_{i'j}(\mu))}\right\}, \tag{4.19}$$

which depends on making a virtual move of i from its position in state μ to its position in state ν. In Equation (4.19) $E_{ij}(\mu)$ is the pairwise energy between the particles in state μ and $E_{i'j}(\mu)$ is the pairwise energy between the two particles following the virtual move of i. The factor $\mathcal{I}_{ij}(\mu)$ is one if i and j interact in state μ and zero otherwise. The factor $\Theta(n_c - n_{\mathcal{C}})$ is used to terminate the linking procedure prematurely in order to stop clusters from moving more than is realistic; if the number of particles in the current cluster, $n_{\mathcal{C}}$, exceeds n_c (where $n_c = \lceil \eta^{-1} \rceil$ and η is a uniform random variable in the range $[0, 1]$) then the move is rejected straight away.

If the partial link is rejected then the link remains unformed and is not considered again during this move. If the partial link does form, then it is converted to a full link with probability $f_{ij}(\mu \to \nu) = \min\left\{1, p_{i'j'}(\nu \to \mu)/p_{ij}(\mu \to \nu)\right\}$, where $p_{i'j'}(\nu \to \mu)$ is equivalent to the partial link probability, except that both proteins start in the state ν and the virtual move is reversed.

We continue considering all links between the current cluster, $\mathcal{C}$, and those particles not yet in the cluster. Once all possible links have been considered, resulting in a set, $\mathcal{R}$, of full links, then the acceptance rate for the move is given by

$$P_{\text{acc}}(\mu \to \nu | \mathcal{R}) = \mathcal{D}(\mathcal{C}) \min\left\{1, \prod_{\{ij\}_{n \leftrightarrow o}} e^{-\beta(E_{ij}(\nu) - E_{ij}(\mu))}\right\}. \tag{4.20}$$

If there are partial links between particles in the cluster and particles not in the cluster then the move is rejected. The product is taken over pairs of particles that do not interact in state μ and have positive energy in state ν or have positive energy in state μ and do not interact in state ν. All other contributions to maintaining detailed balance are accounted for during the linking procedure, so these are the only terms that need to be

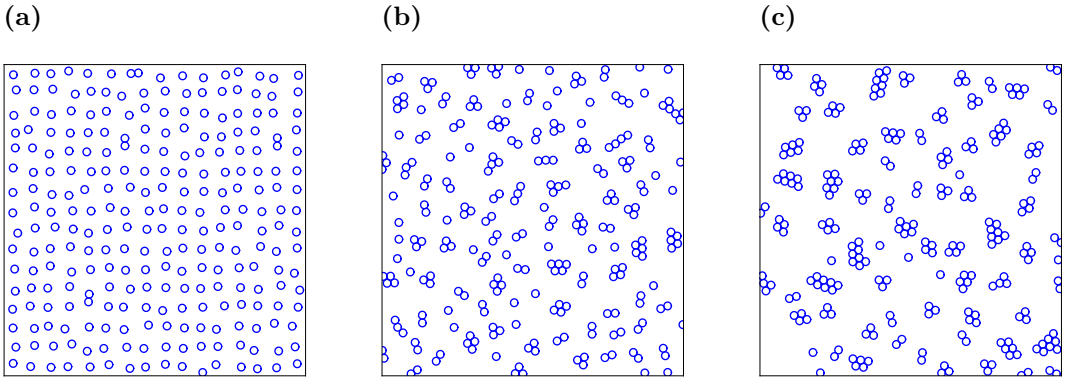

Figure 4.19: Positions of the proteins after 100 **(a)**, 10^4 **(b)**, and $2{\times}10^4$ **(c)** Monte Carlo steps using the VMMC method to simulate their movement.

calculated for the acceptance rate. The factor $\mathcal{D}(\mathcal{C})$ is used to modulate the diffusivity of clusters. In this work we use a factor of $1/\sqrt{n_{\mathcal{C}}}$, which will approximately correspond to the reciprocal of the cluster radius. Although for large inclusions, such as large-scale lipid domains, diffusion in a membrane will have a diffusion constant that follows the scaling rules outlined by Saffman et al. (1975), it was found that for small inclusions, like proteins, a diffusion constant that scales with $1/r$ for inclusions of radius r is more appropriate (Gambin, Lopez-Esparza, et al. 2006).

Simulations of the protein model using VMMC were performed using the same software that was used for the Metropolis Monte Carlo simulations, introduced in Section 4.4.1. Examples of the system after 100, 10^4 and $2{\times}10^4$ Monte Carlo steps are shown in Figure 4.19.

4.6.2 Comparing cluster-based Monte Carlo scheme with CGMD data

Using a cluster-based Monte Carlo scheme with a reasonable approach to including realistic dynamics, as described above, enables us to address some of the discrepancies seen in Figures 4.17 and 4.18. For simulations of the same number of Monte Carlo steps, with exactly the same parameters, we have repeated the comparison for both the MSD of the proteins and the distribution of proteins among clusters of different sizes. We

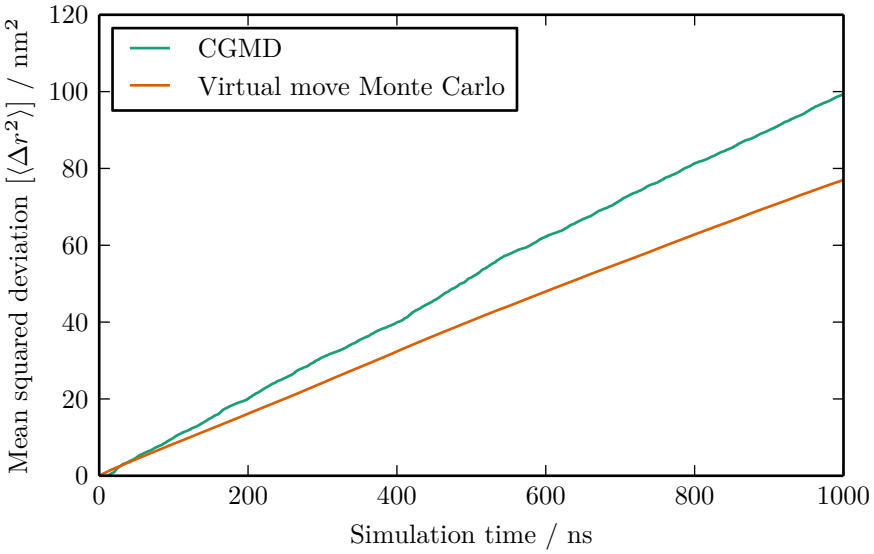

Figure 4.20: MSD of the isotropic NanC proteins simulated using VMMC compared to the MSD of coarse-grained NanC proteins simulated using molecular dynamics. There is much better agreement between these two lines that there is for the Monte Carlo simulations shown in Figure 4.17 (on page 110).

ran 512 repeat VMMC simulations, each with 256 proteins, and at the same protein density as the CGMD simulation with which we are making the comparison. For the MSD the comparison between the VMMC scheme and the CGMD data is shown in Figure 4.20. We can see that there is much better agreement between the two MSDs than in Figure 4.17. This implies that we are capturing the movement of the proteins better with the cluster-based Monte Carlo scheme than before. The lines do not agree perfectly, but there are many approximations in our model, so we would never expect a perfect agreement. However, it is clear that by including the explicit cluster-based motion in the simulation scheme, we are improving the ability of the model to capture the behaviour observed in the CGMD simulations.

The comparison for the proportion of proteins in clusters of various sizes between the CGMD and VMMC simulations is shown in Figure 4.21. Here there is another drastic improvement, with the CGMD cluster proportions within, or close to within, one standard deviation of the cluster proportions observed in the VMMC simulations. This

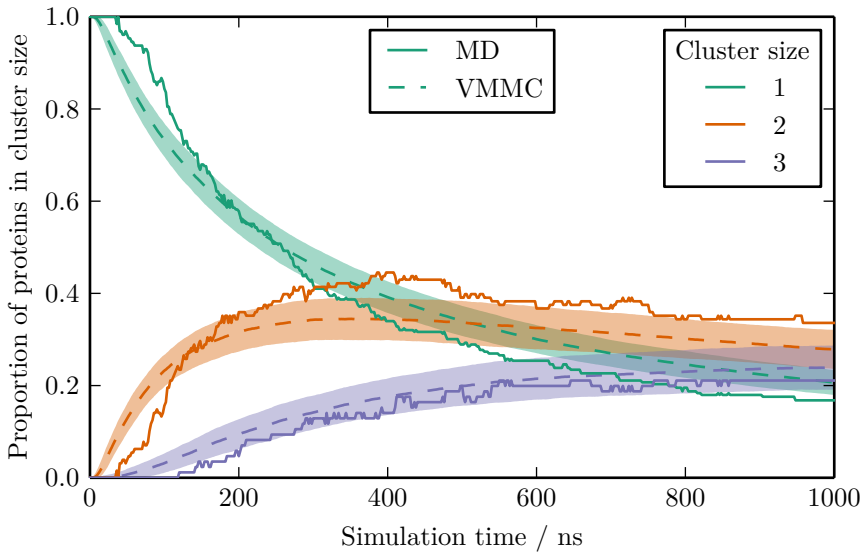

Figure 4.21: Distribution of the proteins in the system among clusters of size one, two and three. Here we compare the results from one CGMD simulation with the results of 512 VMMC simulations. The VMMC simulations are plotted with one standard deviation marked by the block of colour surrounding the corresponding line. There is much better agreement between these two sets of data than there is for the Monte Carlo simulation shown in Figure 4.18 (on page 111).

lends further support to our argument for including cluster-based moves into the Monte Carlo scheme in order to better capture the dynamics of the system. Here we have found that the motion of proteins into clusters and the movement of those clusters is reasonably well represented in the VMMC simulations of our isotropic NanC model.

4.7 Summary

In this chapter we have presented our model, which represents a lipid bilayer system populated with membrane proteins. We have outlined a simple individual-based Monte Carlo simulation scheme and tested the simulation parameters to ensure that they are suitable for our approach to scaling the simulation time. Simulation of our bilayer model can be analysed using various metrics, such as the mean squared displacement for individual proteins and the size of the protein clusters that form. We have introduced

these metrics here so that they may also be used in Chapter 5 as a means for comparison between the various incarnations of our model.

The parameterization of our model was investigated through modifications of the shape of features of the interaction energy profile. We were able to characterize how the depth of the potential well in the interaction energy and the height of the barrier affected the clustering rate of proteins in the system.

Using this simple simulation scheme to investigate the behaviour of our model, we have been able to compare its performance with large-scale MD simulations performed by Dr. Joseph Goose. This simple simulation technique produces qualitatively similar behavior, but the dynamics of the simulations were not in agreement. We hypothesized that the introduction of cluster-based moves and hydrodynamic interactions to the method we used to simulate our model might improve its replication of coarse-grained MD simulations. In the final section we implemented using the virtual move Monte Carlo scheme of Whitelam and Geissler (2007). Using this scheme the quantitative agreement between the movement of the proteins and the distribution of proteins among clusters of different sizes was greatly improved.

In the next chapter we will be extending our model to use a more complicated interaction energy parameterization, derived from the orientationally-restrained PMFs calculated in Chapter 3. Given our investigation of the effect of changes to the interaction potential in this chapter, it is likely that the variations in potential calculated as a function of orientation in Chapter 3 will lead to measurable differences in the behaviour of the system.

Chapter 5

Including Anisotropic Interactions in the Discrete NanC Model

In this chapter we extend the discrete model of NanC, introduced in Chapter 4, by incorporating the anisotropy found in the protein-protein interactions in Chapter 3. With the development of this model we aim to be able to create a work-flow in which we can use any parameterization of protein-protein interactions based on their separation and relative orientations. We develop a protocol for investigating cluster features of an anisotropic system: the eccentricity of clusters, the pairwise orientational correlations of adjacent proteins and the alignment of protein orientations within a cluster are all used to assess the system. Finally, we investigate the application of this scheme to orientation-dependent PMFs from the literature.

Contents

5.1 Modifying the Monte Carlo model to include anisotropic proteins

As a first attempt at parameterizing a Monte Carlo model of a two-dimensional bilayer in Chapter 4, we characterized the pairwise protein-protein interactions using a function of only the inter-protein separation. However, in Chapter 3 we observed a significant difference in the strength of the protein-protein interaction as the relative orientations of the proteins were changed. Our representation of a bilayer populated with isotropically-interacting proteins was a simplified model. In this section we will extend our initial model to include anisotropic interactions between the proteins. Additionally we will use the orientation-dependent PMFs calculated in Chapter 3 to parameterize these interactions.

To include the effects of protein orientation in our model, we have to make two modifications to the procedure outlined in Chapter 4. Firstly, we have to extend the specification of the model, such that the protein orientations are explicitly included in both our description of the state of the system and in the update rules (which modify the system's state) by including rules for the modification of these orientations in a manner analogous to those for the protein positions. Secondly we need to parameterize these new interactions and processes, over the whole range of the system's configuration space, using information obtained through the MD simulations of Chapter 3.

5.1.1 Modifying the protein interaction model

To modify the model to incorporate the anisotropy in the protein-protein interaction, which we measured in Chapter 3, we firstly need to extend the description of the proteins and their interactions to account for their current orientational state. In Chapter 4 our system state was fully represented by the set of positions of the proteins' centres of mass

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, \tag{5.1}$$

where our system contains N proteins. We now need to include a variable for each protein that characterizes its orientational state. The proteins are relatively stable in the membrane, as their pore axis does not deviate very far from the normal to the membrane (a mean angular deviation of $19^\circ \pm 6^\circ$ was calculated in Section 4.3.1 (on page 79)). This means that a sensible measure of their orientational state, as employed in Chapter 3, is the angle about the pore axis, which will be approximately in the plane of the membrane, because the angle between the pore axis and the membrane normal is small. Using this variable to characterize the orientational state of each protein means that the total state of the system is now specified by the set $\Psi = X \cup \Phi$, where Φ is the set of the individual

protein orientations, ϕ_i, and is given by

$$\Phi = \{\phi_1, ..., \phi_N\}. \tag{5.2}$$

The ϕ_i are the same angles that were used to described the orientational configurations of the proteins (shown in Section 3.4.1 (on page 44)) when calculating the PMFs in Chapter 3.

In our system of anisotropic proteins, where their state is specified by both position and orientation, the interaction energy for the two proteins, ϵ_{ij}, is now a function of their positions and orientations

$$\epsilon_{ij} \equiv \epsilon_{ij}\left(\mathbf{x}_i, \phi_i, \mathbf{x}_j, \phi_j\right). \tag{5.3}$$

5.1.2 Monte Carlo update rules

In the isotropic model of Chapter 4, we applied transitions to the system by proposing translations to individual proteins. With the proteins described by the total state X^t at time t, we applied our translational update rule such that following some time interval δt, the system was in state $X^{t+\delta t}$. In this anisotropic model we follow the same approach, except that the system state is described by Ψ^t at time t and following a time interval of δt it will then be in state $\Psi^{t+\delta t}$. We need to extend our update rules to cover transitions between states that incorporate both translational and rotational moves, in order that we fully specify the possible transitions between the states.

The translational transition rules can be carried across directly from the isotropic case, where we propose a translation of a single protein in a random direction by some distance uniformly selected from the interval $[0, \delta l]$. We can incorporate rotational moves in an analogous manner, by proposing a rotation, $\delta\phi$, where $\delta\phi$ is selected uniformly from the interval $[-\delta\alpha, \delta\alpha]$ and $\delta\alpha$ is the maximum absolute rotation for a single trial transition. Given that the mean squared rotation (MSR) about a single axis is given

by $\langle \phi^2 \rangle = 2D_R t$, by equating the time, t, with the time in the expression for the MSD, $\langle r^2 \rangle = 4D_T t$, to give a scaling ratio for rotational moves to translational moves of

$$\delta\alpha = \delta l \sqrt{\frac{D_R}{2D_T}}, \tag{5.4}$$

where D_R is the rotational diffusion coefficient, and D_T is the translational diffusion coefficient (Berg 1993).

5.1.3 Required anisotropic parameters

In Chapter 4 we modelled the pairwise protein-protein interaction potential using a combined Morse and Gaussian potential given by Equation (4.16) (on page 88), which is completely defined by the four parameters $r_{\min}$, $\sigma_{\min}$, r_{barrier}, and σ_{barrier}, which define the position and width of the Morse potential, and the position and width of the Gaussian potential, respectively. For our anisotropic model, the simplest approach to extending the isotropic potential outlined above, so that it reflects the anisotropic nature of the protein-protein interactions, is to replace the four constant parameters with parameters that are functions of the orientations of the two proteins. Although, since we want our system to be rotationally invariant, strictly they need to be functions of their relative orientations, $\hat{\phi}_i$ and $\hat{\phi}_j$, given by

$$\begin{aligned} \hat{\phi}_i =& \phi_i - \arctan\left(\frac{\mathbf{e}_2 \cdot (\mathbf{x}_j - \mathbf{x}_i)}{\mathbf{e}_1 \cdot (\mathbf{x}_j - \mathbf{x}_i)}\right), \\ \hat{\phi}_j =& \phi_j - \arctan\left(\frac{\mathbf{e}_2 \cdot (\mathbf{x}_i - \mathbf{x}_j)}{\mathbf{e}_1 \cdot (\mathbf{x}_i - \mathbf{x}_j)}\right), \end{aligned} \tag{5.5}$$

where $\mathbf{e}_1$ and $\mathbf{e}_2$ are unit vectors in the x and y directions, respectively.

These relative angles are illustrated in Figure 5.1. The parameters are replaced by

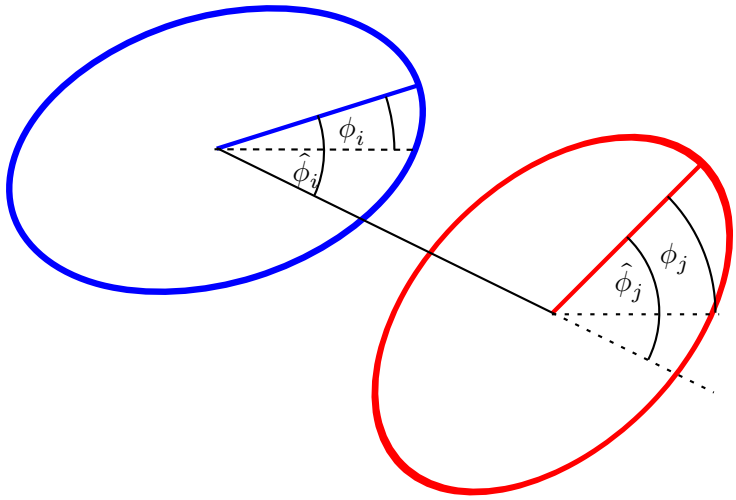

Figure 5.1: Relative angles of the two ellipses. The angles $\hat{\phi}_i$ and $\hat{\phi}_j$ are measured relative to the line between the centres of the ellipses i and j shown in blue and red, respectively. ϕ_i and ϕ_j are both measured relative to the x axis.

functions:

$$\begin{aligned} r_{\text{min}} &\to r_{\text{min}}(\hat{\phi}_i, \hat{\phi}_j), & \sigma_{\text{min}} &\to \sigma_{\text{min}}(\hat{\phi}_i, \hat{\phi}_j), & \epsilon_{\text{min}} &\to \epsilon_{\text{min}}(\hat{\phi}_i, \hat{\phi}_j), \\ r_{\text{max}} &\to r_{\text{max}}(\hat{\phi}_i, \hat{\phi}_j), & \sigma_{\text{barrier}} &\to \sigma_{\text{barrier}}(\hat{\phi}_i, \hat{\phi}_j), & \epsilon_{\text{barrier}} &\to \epsilon_{\text{barrier}}(\hat{\phi}_i, \hat{\phi}_j). \end{aligned} \tag{5.6}$$

In order to fully parameterize our model, we have to determine the form of these functions using the free energy profiles obtained in Chapter 3. The determination of the functional forms of these parameters will be performed in Section 5.2.

5.1.4 Calculating the rotational diffusion coefficient

As well as modifying the interaction potential, we must also obtain a value for the rotational diffusion coefficient of the proteins. This can be achieved using a process similar to that used to measure the translational diffusion coefficient in Chapter 4. The relationship between the mean squared rotation (MSR) and time is given by

$$\langle \Delta\phi^2 \rangle = 2D_R t, \tag{5.7}$$

where $\langle \Delta\phi^2 \rangle$ is the MSR, D_R is the rotational diffusion coefficient and t is time.

We can calculate the rotational diffusion coefficient using the same MD simulation of a single NanC protein diffusing in a large lipid bilayer that we used in Section 4.3.3 (on page 90) to calculate the translational diffusion coefficient. The mean squared rotation (MSR) of the NanC protein in the MD simulation is shown as a function of simulation time in Figure 5.2. There is clearly a linear relationship, as we expect, and by fitting a straight line to the data we are able to calculate a rotational diffusion coefficient, D_R, of 5.62×10^{-3} rad^2 ns^{-1}.

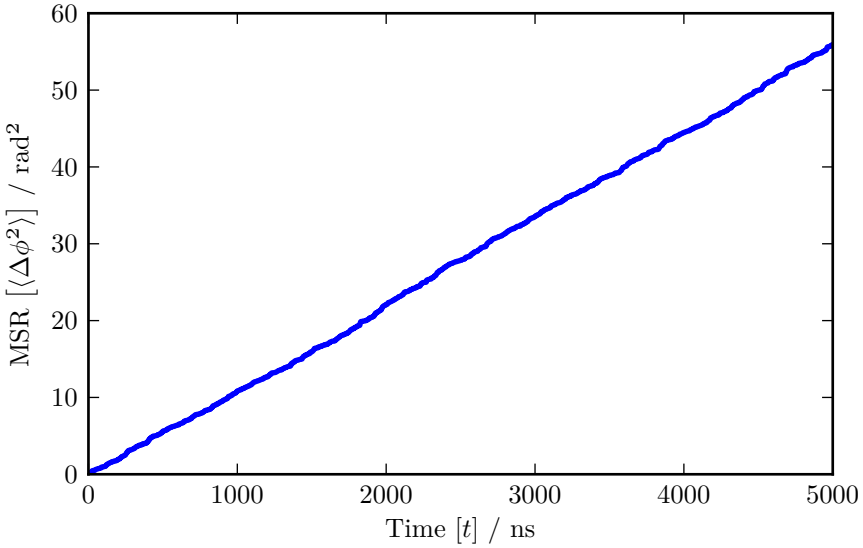

Figure 5.2: Mean square rotation (MSR) of a single NanC as a function of time, t, obtained from a 5 μs coarse-grained MD simulation of a 28.3 nm-square periodic pure POPE bilayer.

5.2 Fitting the anisotropic PMFs

To parameterize the interactions between the proteins of our model, we have to interpolate the potential for orientation values that are between those for which we calculated the PMFs. We require our interpolated potential to be a function of three variables: the inter-protein separation, r_{ij}, and the relative orientations of the two proteins, $\hat{\phi}_i$ and

$\hat{\phi}_j$. To obtain this interpolation we firstly fit a function to the PMFs, for which we use a combination of a Morse and a Gaussian potential, as we did in Chapter 4. Starting from these fitted potentials we enforce some simplifying assumptions about the nature of the interactions in order to interpolate the potential for the entirety of our model's configuration space.

5.2.1 Justification of the fitting function

For the case of anisotropic PMFs, we are fitting to a curve that bears many similarities to the isotropic PMF, which was fitted in Chapter 4. At very short-ranges the potential is repulsive. At slightly larger distances, the potential decreases into a deep well. With increasing distance, outside of the potential well, there is a small energetic barrier, which eventually decreases to zero with increasing separation. As these general features are the same in both the isotropic and anisotropic case, we will use the same function to fit to the PMF data: the combined Morse and Gaussian potential defined in Equation (4.16) (on page 88).

There is, however, a difference between the anisotropic PMFs and the isotropic PMF. As identified and discussed in Chapter 3, the anisotropic PMFs exhibit various metastable local minima. In Chapter 3 we discussed a possible cause of these metastable states being the organization of lipids between the proteins. Such features are not usually observed in PMFs calculated between pairs of bilayer constituents and are only observed in the results of our calculations because we employed a method requiring a highly restrained system to obtain them. We observe these metastable states because the PMFs we calculated were slices through a larger, multidimensional free energy surface, which represented a particular reaction pathway (which would normally be averaged over). As discussed, these metastable states are of great interest in trying to elucidate the intricacies of protein-protein interactions in the bilayer, particularly the role played by protein-lipid-protein interactions. However, the relevance of such states, insofar as being

actually observed in the dynamical behaviour of a bilayer system, is not as important; in an unrestrained system, the proteins are able to take a less direct path across the free energy surface, such that they would not be caught in any of these metastable states. For instance, if a protein had a position and orientation that corresponded to one of these states, it would not be likely to reside there long, but would be able to rotate, which would thus change the local potential caused by the location of neighbouring proteins. If the physical cause of such states is the organization of the lipids in the intervening space between the proteins, then any protein that was not at an optimal distance from another protein would, by a combination of both rotation and translation, be able to alter the size of this intervening region so that it was the correct size to maintain the optimal lipid packing. It is for this reason that we will not to try to represent these metastable states in the parameterization of our model. Not only are they likely to play a small role in the dynamics of proteins in the bilayer, but the dependence of these features on the relative protein orientations is likely to be complex and poorly captured by any interpolation we attempt to perform.

5.2.2 Numerical fitting procedure

We fitted the four anisotropic PMFs, which were functions of the inter-protein separation for four different relative orientations, using the same method employed in Chapter 4 for the isotropic PMF. Again, the Python *SciPy* library was used to perform a curve fitting using the combined Morse and Gaussian potentials, where we constrained the location of the potential well in the Morse potential and the location of the barrier in the Gaussian potential to be near the well and barrier of the PMFs, respectively. The fitted PMFs are shown in Figures 5.3(a) to 5.3(d) and the values of the parameters in Equation (4.16) (on page 88) are given in Table 5.1. We can see the relationships between the depths of the PMFs, which were identified in Chapter 3, both from these figures and from the fitting parameters in the table; the PMFs for $(90°, 90°)$ and $(90°, 270°)$ have the deepest

	Protein orientations, $(\hat{\phi}_i, \hat{\phi}_j)$			
Parameter	$(90°, 90°)$	$(90°, 270°)$	$(90°, 180°)$	$(180°, 180°)$
ϵ_{min}	65.5 kJ mol^{-1}	66.5 kJ mol^{-1}	51.4 kJ mol^{-1}	46.5 kJ mol^{-1}
$\epsilon_{\text{barrier}}$	7.2 kJ mol^{-1}	9.0 kJ mol^{-1}	7.2 kJ mol^{-1}	5.3 kJ mol^{-1}
r_{min}	3.2 nm	3.2 nm	3.5 nm	3.5 nm
r_{barrier}	5.3 nm	5.6 nm	5.6 nm	6.0 nm
σ_{min}	2.4 nm^{-1}	2.9 nm^{-1}	2.4 nm^{-1}	3.0 nm^{-1}
σ_{barrier}	0.8 nm	0.7 nm	0.8 nm	0.6 nm

Table 5.1: Parameter values for the interaction energy function in Equation (4.16) (on page 88) fitted to the four orientationally dependent PMFs calculated in Chapter 3.

potential wells, $(180°, 180°)$ has the shallowest, and the potential well of the $(90°, 180°)$ orientational configuration has a depth that is between that of the two other categories.

5.3 Interpolating the parameterization of the potentials of mean force

The four PMFs in Figure 5.3 describe the change in free energy with changing inter-protein separation for four different orientational combinations. However, because we wish to parameterize a discrete model of the bilayer using these PMFs, we have to interpolate the PMF for the rest of configuration space. To do so we will make several basic assumptions about the behaviour of the fitted parameters used above. These assumptions are based on our understanding of the nature of the interactions that are involved in controlling the behaviour described by a given parameter. We will use various methods to interpolate between the relevant parameters based on these dependencies. This will give us a consistent system, whose behaviour we can compare with the isotropic model presented in Chapter 4.

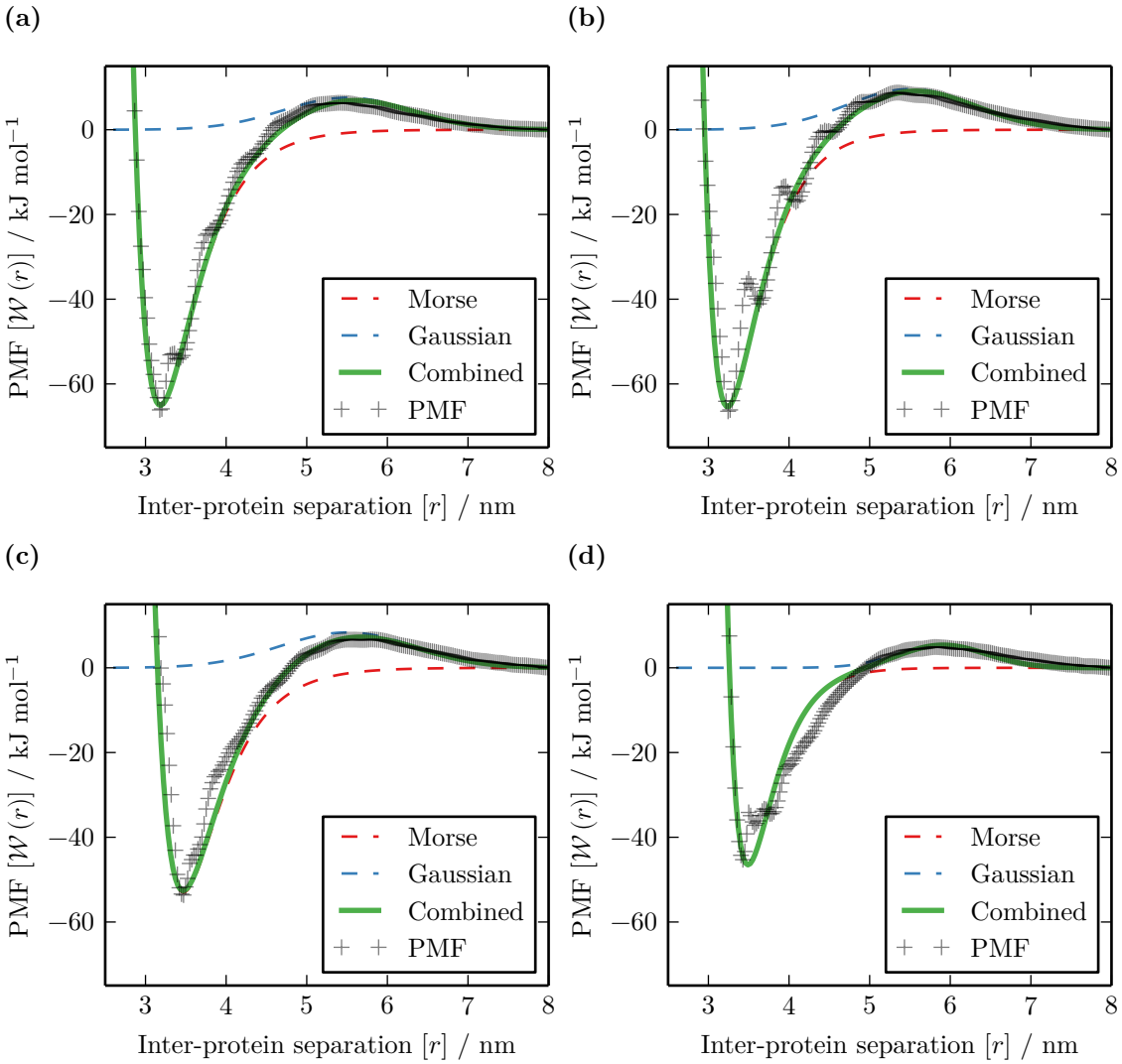

Figure 5.3: Fitting the PMF data (grey crosses) for the protein orientational configurations: **(a)** $(\phi_1, \phi_2) = (90°, 90°)$; **(b)** $(\phi_1, \phi_2) = (90°, 270°)$; **(c)** $(\phi_1, \phi_2) = (90°, 180°)$; and **(d)** $(\phi_1, \phi_2) = (180°, 180°)$. The fit is a combination of a Morse potential (dashed red line) and a Gaussian (dashed blue line). The combined function is shown by the solid green line.

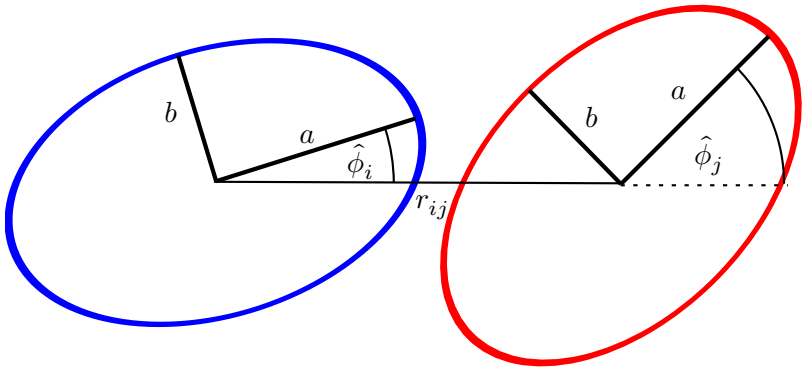

Figure 5.4: Two proteins, i and j, represented by ellipses with semi-major and semi-minor axis lengths of a and b, respectively. Their centres of mass are separated by a distance r_{ij} and their orientations are described by the angles $\hat{\phi}_i$ and $\hat{\phi}_j$, which are measured relative to the line that passes through both of their centres of mass.

The basis for the interpolation schemes used in this parameterization is that we are able to represent the two proteins as ellipses in a two-dimensional membrane. This is an appropriate representation because the cross-sectional shape of NanC is approximately elliptical, so when describing the orientational state of the system, we do so using the analogous representation of two ellipses in a plane. In this representation the shapes of the two proteins are described by the lengths of their semi-major and semi-minor axes, a and b, respectively. Their centres of mass are separated by a distance r_{ij} and their orientations are described by the angles $\hat{\phi}_i$ and $\hat{\phi}_j$, which are measured relative to the line connecting their centres of mass. These parameters are shown in Figure 5.4.

As discovered in Chapter 3, the depths of the PMFs show a correlation with the buried surface area of the two proteins and so we will use the behaviour of the buried surface of two ellipses, as a function of their orientation, to interpolate the parameter describing the depth of the PMF, ϵ_{min}. The position of the minimum, r_{min}, and the position of the barrier, r_{barrier}, are features that we expect to be dependent on the distance between the two proteins. We have made the assumption that surface specific effects do not play a large role in these parameters. This decision was informed by the observations of Niemelä et al. (2010) that simulations of lipids diffusing in concert with proteins appeared to do so

as a result of a "wall effect" due to the moving protein rather than the direct interaction, especially for the lipids that were greater than one layer removed from the protein, whose only interactions are with adjacent lipids and not with the protein itself. Also, our own observations in Chapter 3 that the effect of lipids in close proximity exhibited a strong effects on the interaction, as a result of the packing between the proteins, suggests that this is a reasonable assumption. These dependencies on protein separation upon contact are unlikely to be wholly accurate, but given the elliptical representation is also an approximation, any discrepancies in distance are likely to be small in comparison to the reduced representation of the proteins themselves. Finally, we assume that the width of the potential well, σ_{min}; the barrier height, $\epsilon_{\text{barrier}}$, and the barrier width, σ_{barrier}, are all constant, which we can see from Table 5.1 is a reasonable assumption, as these do not change by a large amount. For the parameters of the barrier, this is also a reasonable assumption because these properties are strongly dependent on the behaviour of the lipids, which, although complicated at small separations (see Chapter 3), will likely be determined by lipid-lipid interactions in a low mobility environment instead of the interactions of the specific proteins. These proposed dependencies are summarized in Table 5.2, and the ability to fit the PMF data using these interpolation schemes will be demonstrated in Section 5.3.2.

In the rest of this section we will derive a mapping for these parameters based on the orientation of two ellipses. For the three non-constant parameters we need to interpolate between sets of parameters. For a given parameter, p, we interpolate between two specific values using a function of the following form:

$$p\left(\hat{\phi}_i, \hat{\phi}_j\right) = p_0 + [p_1 - p_0]\, M\left(\hat{\phi}_i, \hat{\phi}_j\right), \tag{5.8}$$

where $M(\hat{\phi}_i, \hat{\phi}_j)$ is a continuous mapping function on $[0, 1]$; and p_0 and p_1 are the values of the parameter p when $M = 0$ and $M = 1$, respectively.

Parameter	Interpolation scheme
ϵ_{min}	Buried surface area of ellipses in contact
$\epsilon_{\text{barrier}}$	Constant
r_{min}	Separation of ellipses in contact
r_{barrier}	Separation of ellipses in contact
σ_{min}	Constant
σ_{barrier}	Constant

Table 5.2: Interpolation schemes used for the PMF fitting parameters. The schemes are all based on the representation of the proteins by two ellipses in a two-dimensional membrane domain.

5.3.1 Interpolation schemes based on an elliptical representation of the proteins

In Chapter 3 we showed that the depth of the potential well was correlated with the buried surface area of the protein-protein complex. In the section above we introduced a simple angle-dependent mapping scheme which interpolates between the extreme values of a parameter in the fitting function of the PMFs. Here we derive two mapping functions, $M_{\text{sep}}(\hat{\phi}_i, \hat{\phi}_j)$ and $M_{\text{buried}}(\hat{\phi}_i, \hat{\phi}_j)$, which are based on the separation of two ellipses when just in contact and on the buried surface area of those two ellipses, respectively. As we are representing our proteins as ellipses, the mapping functions that we choose should have a periodicity of 180°, should vary smoothly with angle, and should have values in the range $[0, 1]$.

In Figure 5.5 the two ellipses are shown at a separation, r_{ij}, when they are just in contact. The buried surface area, as defined in Chapter 3, is calculated using a probe of radius r_{p}, which tracks the edge of the proteins and where again we use a value of 0.47 nm (twice the size of the coarse-grained MD particles) to maintain consistency with the buried surface calculations performed in Section 3.7 (on page 54). In our simple

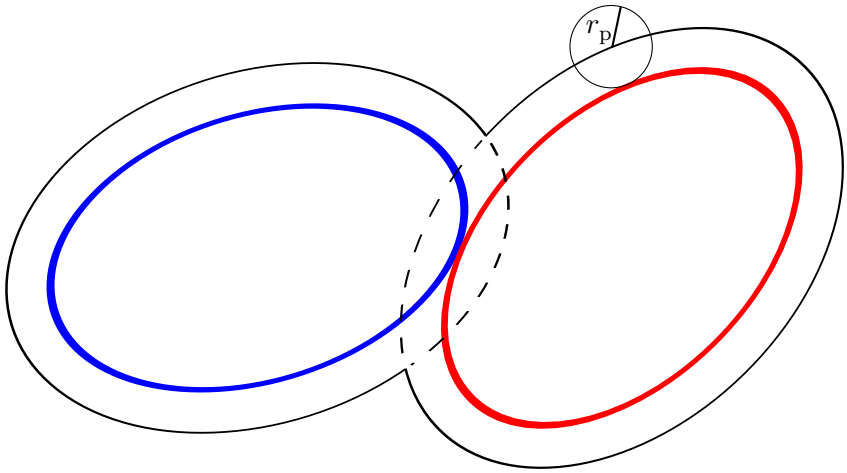

Figure 5.5: For the two ellipses, E_i and E_j, shown in blue and red respectively, the buried surface area is defined using a probe of radius r_{p} (as in Chapter 3). When the two ellipses are in contact, the solvent accessible surface area is defined by the trace of the probe around the ellipses, given that the probe must maintain contact with at least one ellipse and cannot overlap with either of them. This trace is shown by the solid section of the thin black line around the ellipses. The buried surface area is given by the difference in length between the solid section of the thin black line and the combined perimeter of both ellipses consisting of both the solid and dashed sections of the black lines; this corresponds to the difference between the surface area when in contact and when at a separation larger than the probe.

elliptical representation in two dimensions, the circular probe will trace out the larger black ellipses, centred around each of the proteins. The region in which the probe cannot fit is that defined by the sections of the ellipses which overlap. To calculate the buried surface area we have to find the arc length of this overlapping section. The first step of this is to calculate the intersections of two identical ellipses separated by a distance r_{ij} with their semi-major axes at angles $\hat{\phi}_i$ and $\hat{\phi}_j$ to the line connecting their centres of mass. We then need to evaluate the intersections both to calculate the separation, for given $\hat{\phi}_i$ and $\hat{\phi}_j$, at which the proteins are just in contact, and to calculate the angular bounds of the intersection of the two larger ellipses traced by the probe.

5.3.1.1 Calculating the intersection of two ellipses

The equation of an ellipse is described by a quadratic in two variables

$$E = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} \alpha_{00} & \alpha_{01} \\ \alpha_{10} & \alpha_{11} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \beta_0 & \beta_1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \gamma = 0, \tag{5.9}$$

where $E < 0$ is inside the ellipse and $E > 0$ is outside. If E_i is an ellipse centred on the origin and E_j is an ellipse at a distance r_{ij} along the x axis, then the two ellipses are described by

$$E_i = 0 = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} b^2 \cos^2 \hat{\phi}_i + a^2 \sin^2 \hat{\phi}_i & (b^2 - a^2) \cos \hat{\phi}_i \sin \hat{\phi}_i \\ (b^2 - a^2) \cos \hat{\phi}_i \sin \hat{\phi}_i & b^2 \sin^2 \hat{\phi}_i + a^2 \cos^2 \hat{\phi}_i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - a^2 b^2, \tag{5.10}$$

and

$$\begin{aligned} E_j = 0 = & \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} b^2 \cos^2 \hat{\phi}_j + a^2 \sin^2 \hat{\phi}_j & (b^2 - a^2) \cos \hat{\phi}_j \sin \hat{\phi}_j \\ (b^2 - a^2) \cos \hat{\phi}_j \sin \hat{\phi}_j & b^2 \sin^2 \hat{\phi}_j + a^2 \cos^2 \hat{\phi}_j \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\ & + \begin{pmatrix} -2r_{ij}(a^2 \sin^2 \hat{\phi}_j + b^2 \cos^2 \hat{\phi}_j) \\ 2r_{ij} \sin \hat{\phi}_j \cos \hat{\phi}_j (a^2 - b^2) \end{pmatrix}^T \begin{pmatrix} x \\ y \end{pmatrix} + r_{ij}^2 (b^2 \cos^2 \hat{\phi}_j + a^2 \sin^2 \hat{\phi}_j) - a^2 b^2. \end{aligned} \tag{5.11}$$

To find the intersection of the ellipses we have to solve $E_i = E_j = 0$. Two polynomials have a common root if and only if their Bézout determinant is zero (Atallah et al. 2009). For two quadratics, $f(x) = f_0 + f_1 x + f_2 x^2$ and $g(x) = g_0 + g_1 x + g_2 x^2$, the entries of the Bézout matrix, b_{ij}, can be calculated from

$$b_{ij} = \sum_{k=1}^{m_{ij}} f_{j+k-1} g_{i-k} - f_{i-k} g_{j+k-1}, \tag{5.12}$$

where $m_{ij} = \min\{i, 3 - j\}$. The resulting Bézout matrix, B_2, for our polynomials of

degree two is

$$B_2(f,g) = \begin{pmatrix} f_1g_0 - f_0g_1 & f_2g_0 - f_0g_2 \\ f_2g_0 - f_0g_2 & f_2g_1 - f_1g_2 \end{pmatrix}. \tag{5.13}$$

Setting its determinant, $\hat{B} = |B_2|$, to zero gives

$$(f_2g_1 - f_1g_2)(f_1g_0 - f_0g_1) - (f_2g_0 - f_0g_2)^2 = 0. \tag{5.14}$$

We can rewrite the ellipse equations, Equations (5.10) and (5.11), as quadratics in x, with coefficients that are functions of y, giving

$$\begin{aligned} Q_k(x,y) &= q_0^{(k)} + q_1^{(k)}x + q_2^{(k)}x^2 \\ &= \left(\alpha_{11}^{(k)}y^2 + \beta_1^{(k)}y + \gamma^{(k)}\right) + \left((\alpha_{01}^{(k)} + \alpha_{10}^{(k)})y + \beta_0^{(k)}\right)x + \left(\alpha_{00}^{(k)}\right)x^2, \end{aligned} \tag{5.15}$$

where the $\alpha_{mn}^{(k)}$, $\beta_m^{(k)}$ and $\gamma^{(k)}$ are the components of the matrix equation for ellipse E_k, with $k = i, j$. The Bézout determinant for Equation (5.15) is given by the quartic $\hat{B}(y) = u_0 + u_1y + u_2y^2 + u_3y^3 + u_4y^4$, where the coefficients are given by

$$\begin{aligned} u_0 &= v_2v_{10} - v_4^2, \qquad u_1 = v_0v_{10} + v_2(v_7 + v_9) - 2v_3v_4, \\ u_2 &= v_0(v_7 + v_9) + v_2(v_6 - v_8) - v_3^2 - 2v_1v_4, \\ u_3 &= v_0(v_6 - v_8) + v_2v_5 - 2v_1v_3, \qquad u_4 = v_0v_5 - v_1^2, \end{aligned} \tag{5.16}$$

where

$$
\begin{aligned}
v_0 &= 2\left(\alpha_{00}^{(i)}\alpha_{01}^{(j)} - \alpha_{00}^{(j)}\alpha_{01}^{(i)}\right), & v_1 &= \alpha_{00}^{(i)}\alpha_{11}^{(j)} - \alpha_{00}^{(j)}\alpha_{11}^{(i)}, \\
v_2 &= \alpha_{00}^{(i)}\beta_0^{(j)} - \alpha_{00}^{(j)}\beta_0^{(i)}, & v_3 &= \alpha_{00}^{(i)}\beta_1^{(j)} - \alpha_{00}^{(j)}\beta_1^{(i)}, \\
v_4 &= \alpha_{00}^{(i)}\gamma^{(j)} - \alpha_{00}^{(j)}\gamma^{(i)}, & v_5 &= 2\left(\alpha_{01}^{(i)}\alpha_{11}^{(j)} - \alpha_{01}^{(j)}\alpha_{11}^{(i)}\right), \\
v_6 &= 2\left(\alpha_{01}^{(i)}\beta_1^{(j)} - \alpha_{01}^{(j)}\beta_1^{(i)}\right), & v_7 &= 2\left(\alpha_{01}^{(i)}\gamma^{(j)} - \alpha_{01}^{(j)}\gamma^{(i)}\right), \\
v_8 &= \alpha_{11}^{(i)}\beta_0^{(j)} - \alpha_{11}^{(j)}\beta_0^{(i)}, & v_9 &= \beta_0^{(i)}\beta_1^{(j)} - \beta_0^{(j)}\beta_1^{(i)}, \\
v_{10} &= \beta_0^{(i)}\gamma^{(j)} - \beta_0^{(j)}\gamma^{(i)}.
\end{aligned}
\tag{5.17}
$$

The intersections of the ellipses can be found by solving $\hat{B}(y) = 0$ and then discarding the solutions with non-zero imaginary components. The corresponding values of x can be found by solving $Q_i(x, y) = 0$ and any solutions, (x, y), which do not also satisfy $Q_j(x, y) = 0$ are discarded.

5.3.1.2 Ellipse separation at minimal contact

To solve these equations a numerical scheme was implemented in python using the *NumPy* and *SciPy* packages. For a given orientational combination of the two proteins, specified by their angles $\hat{\phi}_i$ and $\hat{\phi}_j$, the separation at which there is only one intersection (i.e. the point of minimum contact) is found by performing a binary search on the value of r, where the initial bounds on r were set to $[1.9b, 2.1a]$, since $a > b$. In this range of separations there are either no intersections, one intersection or two intersections. For a given value of r, if there are no intersections found, then we reduce r for the next step, and if there are two intersections found, then we increase r. Using a numerical solution will not return the exact separation at which there is only one intersection, but we set a 10^{-6} nm tolerance distance between the two intersection points for the convergence of the binary search, and below which we consider the ellipses to be just in contact. This value was chosen because the procedure only took a short time to converge and there was

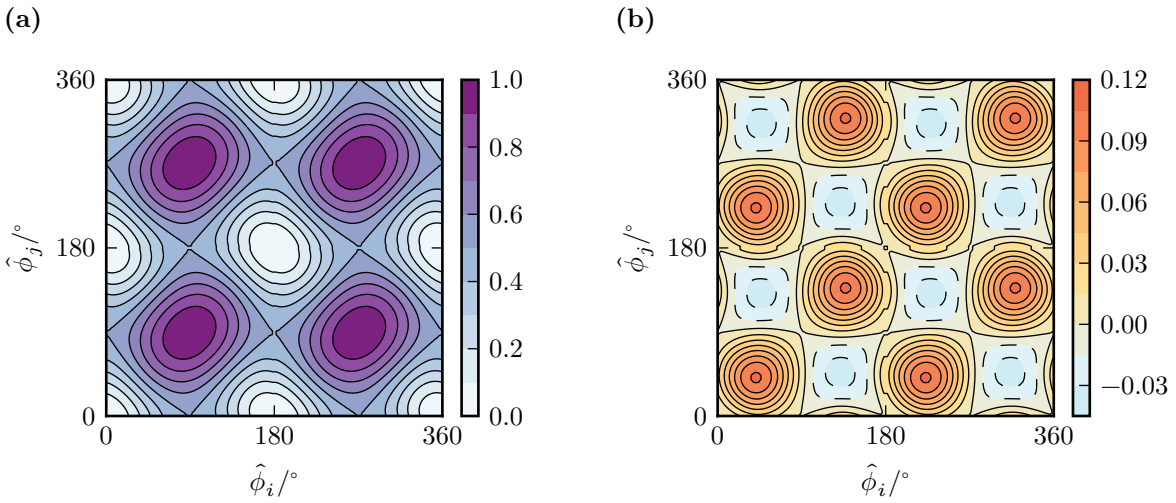

Figure 5.6: The mapping function $M\left(\hat{\phi}_i, \hat{\phi}_j\right)$, introduced in Equation (5.8), describes a general transition between extremal values of the parameters in our interpolation scheme when the relative angles of the proteins change. **(a)** The form of the mapping function when based on the separation of ellipses' centres of mass when their surfaces are in contact. **(b)** This difference between this interpolation scheme based on separation and the simple interpolation scheme (shown in Figure 5.7).

no noticeable change in the interpolation scheme at this level of accuracy. The mapping function, $M_{\text{sep}}(\hat{\phi}_i, \hat{\phi}_j)$, for the separation of the two ellipses is shown in Figure 5.6(a). Values for a and b were set to, 1.75 nm and 1.6 nm, respectively, which are half the corresponding values of r_{min} from Table 5.1 (on page 128).

As a means for comparison, we can calculate a simple interpolation scheme that does not rely on the geometrical properties of the elliptical representation of the proteins. In this simple scheme the mapping function is given by

$$M\left(\hat{\phi}_i, \hat{\phi}_j\right) = \frac{1}{4}(2 - \cos 2\hat{\phi}_i - \cos 2\hat{\phi}_j). \tag{5.18}$$

This function satisfies the need to be smoothly varying in the range $[0, 1]$ and to have the same periodicity as our elliptical protein representation, as can be seen in Figure 5.7. The difference between this simple scheme and the interpolation scheme based on the separation of the two ellipses is shown in Figure 5.6(b).

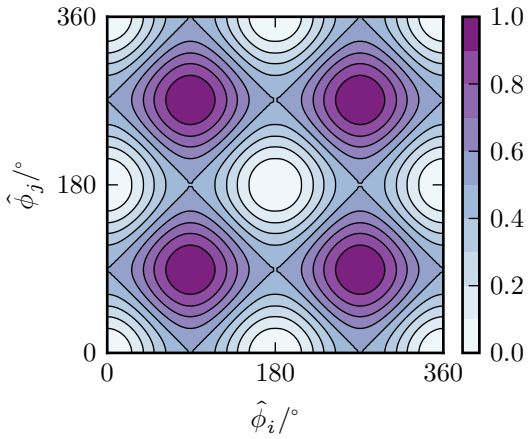

Figure 5.7: A simple mapping scheme that satisfies the need to be smoothly varying between 0 and 1 with periodicity 180° in both variables. The interpolation is given by Equation (5.18).

5.3.1.3 Buried surface area as a function of ellipse orientations

After obtaining the approximate value of r for minimum contact, to calculate the buried surface area we need to find the two positions of intersection for the larger ellipses, defined by the trace of the probe, in Figure 5.5. The angle that these intersections make with the origin can then be calculated and the arc length of the buried section for each ellipse may be found by integrating along the ellipse between these angles. The arc length of an ellipse between two angles, θ_1 and θ_2, is given by

$$L = a \int_{\theta_1}^{\theta_2} \sqrt{1 - \epsilon^2 \sin^2 \theta'} d\theta', \tag{5.19}$$

where $\epsilon = \sqrt{1 - (b/a)^2}$ is the eccentricity of the ellipse. This can be calculated numerically using the `ellipeinc` function in the *SciPy Python* package.

The procedure outlined above was employed to calculate the relative buried surface area over the entire range of rotation of the two ellipses. The relative buried surface area is shown in Figure 5.8(a).

We can again compare this mapping, based on the buried surface area, to the simple

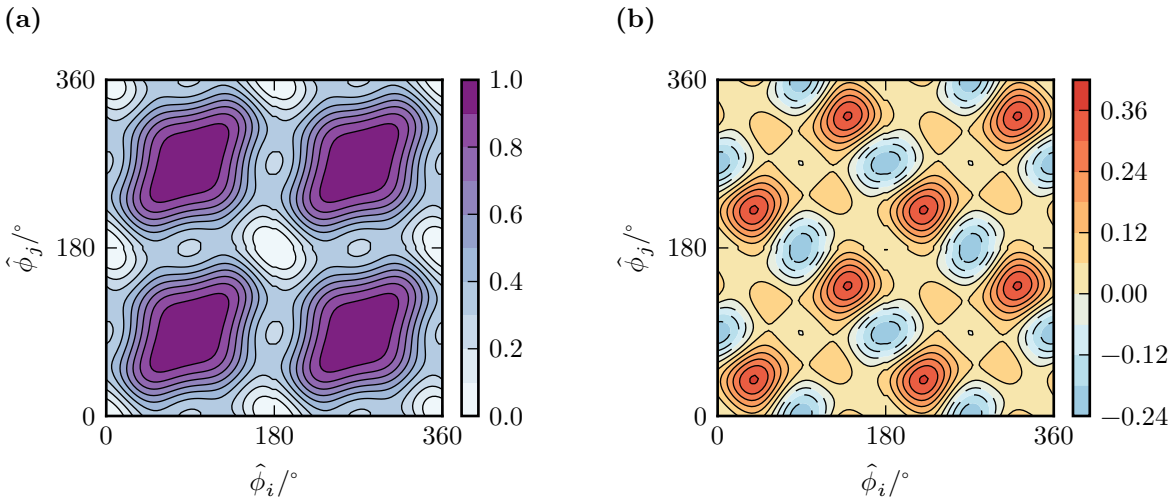

Figure 5.8: The mapping function $M\left(\hat{\phi}_i, \hat{\phi}_j\right)$, introduced in Equation (5.8), describes a general transition between extremal values of the parameters in our interpolation scheme when the relative angles of the proteins change. **(a)** The form of the mapping function when based on the buried surface area of two ellipses when in contact, as a function of the angles of the two ellipses. **(b)** Difference between the interpolation scheme based on buried surface area and the simple interpolation scheme (shown in Figure 5.7).

scheme given by Equation (5.18); the difference between the two schemes is shown in Figure 5.8(b). There is a more significant difference between these two schemes than for the difference between the separation-based scheme and the simple scheme (shown in Figure 5.6(b)). To justify our use of the buried surface area interpolation scheme, we can assess the performance of this more advanced interpolation scheme by computing the depth of the PMF with orientational configuration $(\hat{\phi}_i, \hat{\phi}_j) = (90^\circ, 180^\circ)$, which is shown in Table 5.3. We can see from these values that the interpolation scheme using the buried surface area in the ellipse representation is much better than the simple interpolation scheme.

5.3.2 The final mapping

Using the interpolation schemes summarized in Table 5.2 and calculated above, we demonstrate how our proposed potential function compares to the PMF data, from which

$\epsilon_{\min}$ for $(\hat{\phi}_i, \hat{\phi}_j) = (90°, 180°)$		
Simple interpolation	Buried surface interpolation	From PMF
56.3 kJ mol^{-1}	51.9 kJ mol^{-1}	51.4 kJ mol^{-1}

Table 5.3: Comparing the performance of the simple interpolation scheme and the interpolation scheme using the buried surface area for the potential well depth parameter, $\epsilon_{\min}$, in the orientational state $(\hat{\phi}_i, \hat{\phi}_j) = (90°, 180°)$.

it was originally derived, in Figure 5.9. We can see that the interpolated potentials still fit the data quite well, considering that many of the originally fitted parameters have been replaced by the averages across all fittings. The region in which this parameterization seems to perform worst is the position of the minimum, $r_{\min}$ for the orientation $(\hat{\phi}_i, \hat{\phi}_j) = (90°, 180°)$, shown in Figure 5.9(c), but this discrepancy is not large.

5.4 Characterizing the behaviour of an anisotropic protein model

In this anisotropic model, all of the proteins in the system now have an intrinsic orientation. Having introduced orientational dependence to the inter-protein potential, we wish to investigate the effect that this dependency has on the behaviour of our model system. The previous metric for characterizing the protein behaviour, cluster size, is still of use in assessing the suitability of our model, but here we also introduce metrics that enable us to determine how these orientational degrees of freedom behave and metrics that describe the shapes of the clusters.

We introduce three such metrics here, which will enable us to see the effect that an anisotropic potential has on the system. The first of these is the eccentricity of clusters, which describes the shape of the clusters as a whole. The second is a pairwise correlation metric, which will enable us to ascertain to what extent the anisotropic pairwise potential affects the orientational correlation between pairs of proteins. The third metric is used to

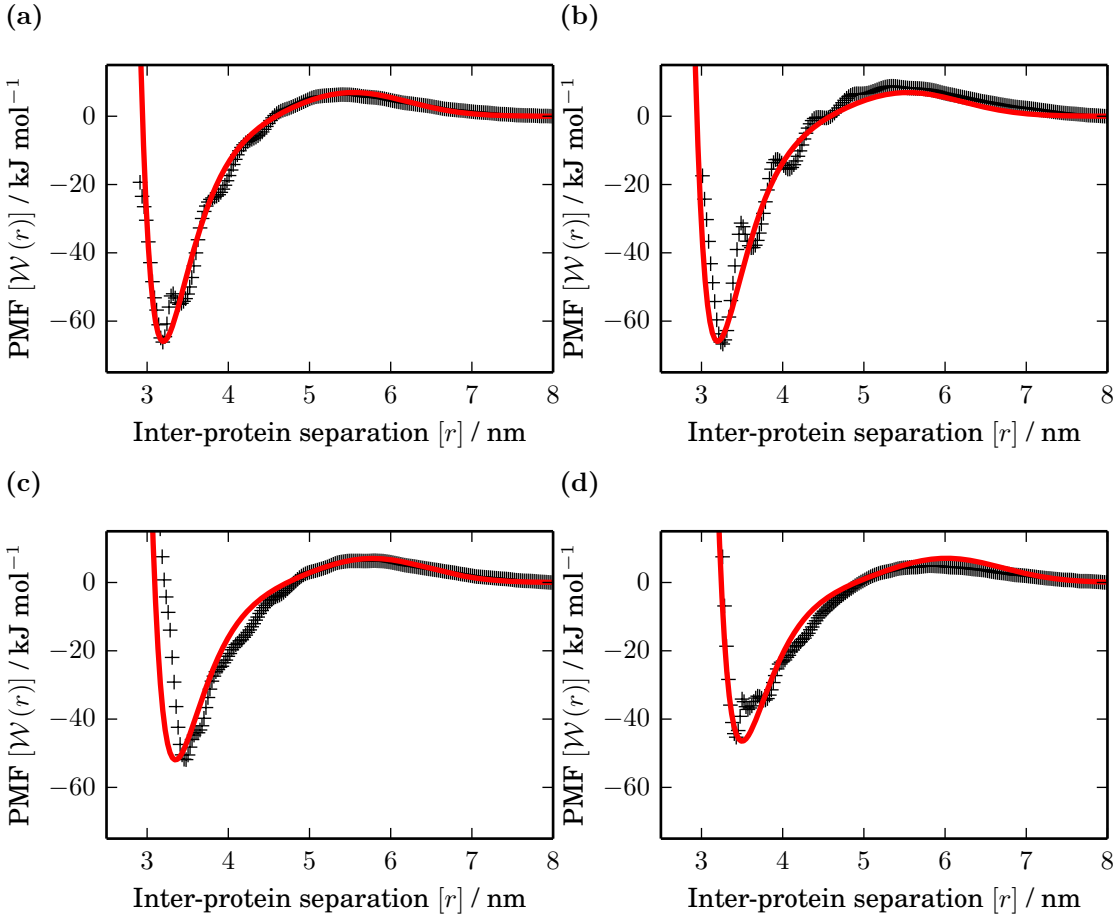

Figure 5.9: The final parameterization based on applying the interpolation rules stated in Table 5.2 compared to the original PMFs calculated in Chapter 3. The parameterizations are given for the orientational configurations $(\hat{\phi}_i, \hat{\phi}_j) = (90°, 90°)$; $(\hat{\phi}_i, \hat{\phi}_j) = (90°, 270°)$; $(\hat{\phi}_i, \hat{\phi}_j) = (90°, 180°)$; and $(\hat{\phi}_i, \hat{\phi}_j) = (180°, 180°)$ in **(a)**, **(b)**, **(c)**, and **(d)**, respectively. The potential functions using the proposed interpolation scheme are shown by the red lines and the PMFs by the black crosses.

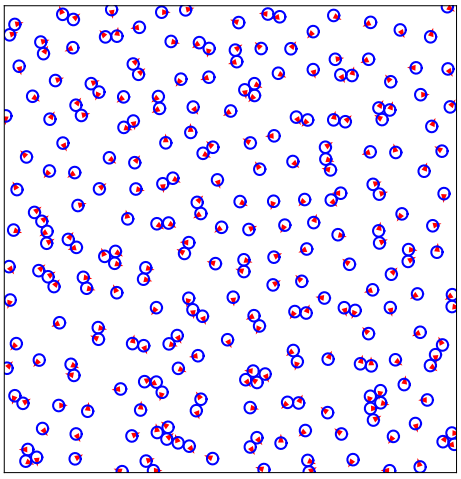

Figure 5.10: An example of the anisotropic system after $5{\times}10^3$ Monte Carlo steps. The red arrows on each protein correspond to the direction of that protein's 0°. The alignment of the protein orientations, in either a parallel or anti-parallel arrangement, can be seen in many of the elongated clusters.

characterize the alignment of the orientations of proteins in a given cluster, which shows the extent to which the pairwise potential leads to large-scale orientational alignment.

An example of the anisotropic protein simulations is shown in Figure 5.10. From this figure some examples of these metrics can be observed. For instance, in many of the elongated clusters (which we define below as clusters that are highly eccentric) the protein orientations are aligned in a parallel or anti-parallel manner. This correlation between cluster eccentricity will be analysed later in this chapter.

5.4.1 Cluster eccentricity

We define the eccentricity using the notion of a minimum enclosing circle for the protein centres in a given cluster. The eccentricity is measured in terms of the radius of the minimum enclosing circle, r_e, which is illustrated in Figure 5.11. The eccentricity is defined as zero when r_e is equal to the smallest possible enclosing circle for a cluster of a given size, and one when r_e is equal to the radius of the largest minimum enclosing circle for a given cluster size. The smallest minimum enclosing circle for a cluster is also

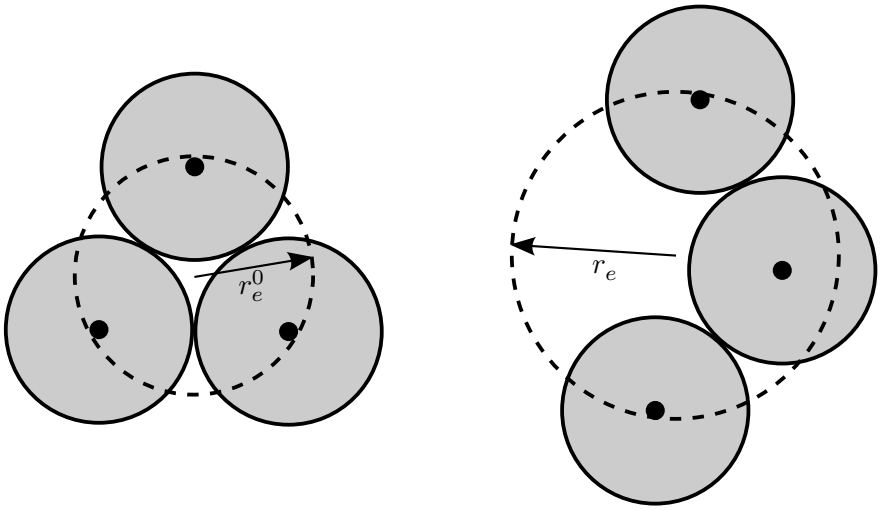

Figure 5.11: For any given cluster the minimum enclosing circle is the smallest circle that encloses the centres of all the protein in the cluster. It has radius r_e and is shown on the right. For a specific cluster size we also know (up to a cluster size of 13) the size of the smallest possible minimum enclosing circle, which has radius r_e^0 and is shown on the left.

illustrated in Figure 5.11. The size of the smallest enclosing circles have been proved up to a cluster size of 13 and there are estimates up to a size of 20 (Graham et al. 1998), which is large enough for our purposes. Given the proportion of proteins in clusters of size one, two and three, shown in Figure 4.12 (on page 104), it is unlikely that we will observe enough clusters with sizes larger than 20 to produce meaningful statistics, even if such clusters do form.

Figure 5.12 shows how the eccentricity of the clusters of a given size vary throughout a simulation of 10^4 Monte Carlo steps using the data for simulations of the isotropic model in Chapter 4. We can see that the eccentricity of the clusters start off high and settle down to some smaller value.

We can see that, excluding the initial section of the simulation, the mean eccentricity stays approximately constant throughout. There is, however, a large variation in the eccentricities, which is indicated by the relatively large error bars, which represent one

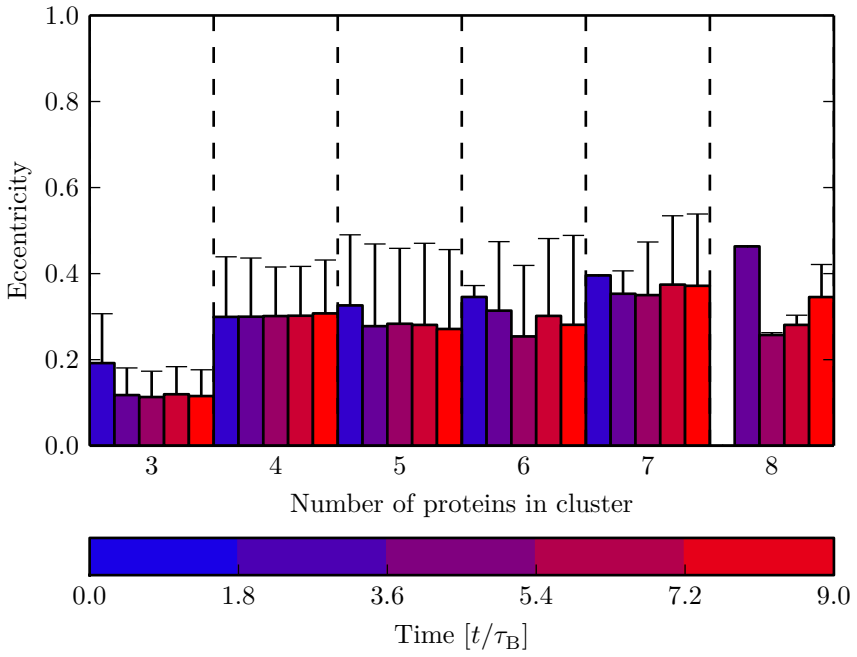

Figure 5.12: Cluster eccentricity for clusters of various sizes. The eccentricity appears to increase, with increasing cluster size, but there is significant variation in the eccentricities. The simulation contained 256 proteins with an area fraction of ∼14%, with $\delta = 0.03$ for 10^4 Monte Carlo steps. The coloured bars for each cluster size correspond to averages taken over subsequent simulation intervals, each lasting $2{\times}10^3$ Monte Carlo steps. The results were also averaged over 192 independent repeat simulations.

standard deviation. The fact that this measure has a large variability does not mean we can not use it to characterize the behaviour of the system under a modification of some parameter, or even a modification of the simulation technique in its entirety. Any deviations in either the mean of the eccentricity or its variance should still indicate that certain parameter modifications affect the behaviour of the model. However, since the measure is reasonably stable throughout the simulation, we will only be looking at the eccentricity of clusters of a given size averaged across the entire simulation.

5.4.2 Pairwise orientation correlation

The first measure used to interpret the orientational behaviour of the system is the pairwise correlation of angles for any two proteins in the system. For each pair of proteins that are below a threshold separation of 4 nm, we record their orientational angles, $\hat{\phi}_i$ and $\hat{\phi}_j$. In this analysis we have chosen to use a fairly generous threshold separation, so as to negate the effects of the variation in the position of the minimum of the potential well as a function of orientation. This can be seen in Figure 5.9 (on page 141), where at a separation of 4.0 nm the potential is approximately half that at the minimum. The pairwise orientation correlations are then separated into bins. From this metric we can see the effect that the potential has in the context of a complex system with a variety of protein cluster configurations. One would perhaps expect, were you to analyse a pair of proteins in isolation, that the distribution of pairwise angles would converge to the form of the interpolation scheme used for the minimum of the potential, since they would just be sampling the distribution of states given by that potential. However, in a larger system this will not be the case and the nature of the orientational correlation that occurs is not immediately obvious.

To measure the effect that the anisotropic potential has on the pairwise correlation of angles we ran simulations using two different interaction regimes. Simulations were run for a system of 256 proteins using both the Monte Carlo scheme introduced in Chapter 4

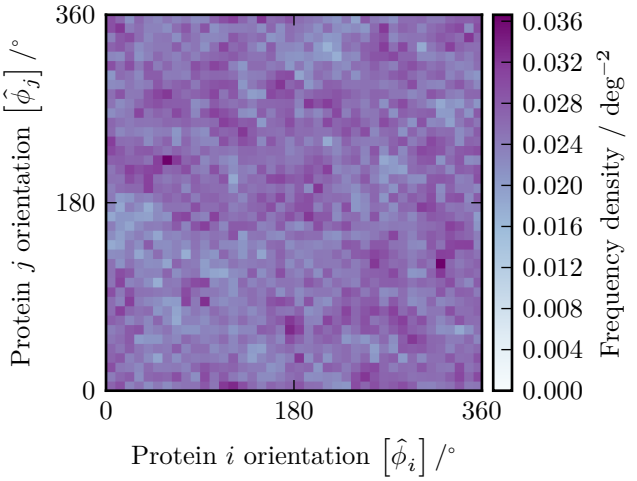

Figure 5.13: Histogram of pairwise orientation correlation for an isotropic inter-protein potential. For 1200 repeats of a simulation of 256 proteins the pairwise correlation is recorded for all protein pairs below a threshold separation of 4.0 nm.

and the extended version introduced in Section 5.1 with a spatial step size $\delta = 0.3$ for 10^4 steps. We ran 1200 repeat simulations, which resulted in a reasonably smooth distribution of the metrics. Firstly, we used the isotropic potential from Chapter 4, to act as a control. Next we used the anisotropic potential introduced in Section 5.3.

We can see in Figure 5.13 that there is no correlation between the orientation of two clustered proteins with an isotropic potential, as we would expect. The introduction of the anisotropic potential, where the depth of the potential well is interpolated using the scheme in Figure 5.8(a), results in a significant difference to the orientation correlation between the proteins. In Figure 5.14 we can see that the distribution of orientations between pairs of clustered proteins in no longer uniform. There are distinct peaks that occur in regions that are distributed in a similar pattern as the depth of the potential, although the specific location of peaks does not correspond exactly with the peaks in the interpolations scheme. The difference in peak location is likely a result of the complex interactions within the clusters: each protein is not always interacting with only one other protein.

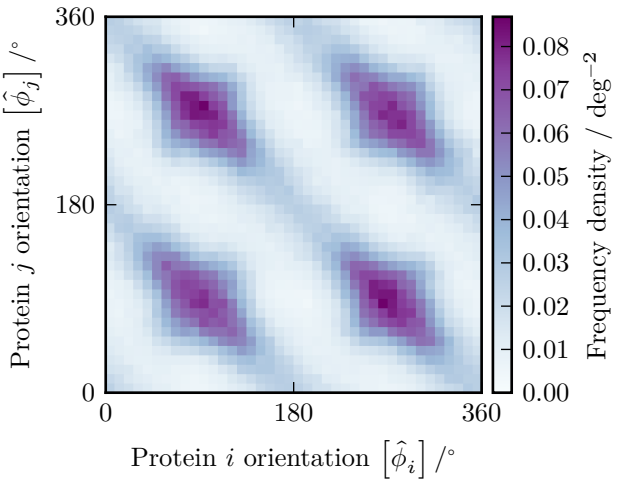

Figure 5.14: Histogram of pairwise orientation correlation for an anisotropic inter-protein potential. For 1200 repeats of a simulation of 256 proteins the pairwise correlation is recorded for all protein pairs below a threshold separation of 4.0 nm.

This metric is useful for seeing the effect that an anisotropic potential has on the distribution of angle correlations, but does not give us any insight into the way in which these effects manifest themselves in the context of the wider system.

5.4.3 Eccentricity of clusters and alignment of clustered proteins

The next metric used to analyse the orientational correlation looks at the cluster scale, rather than at the protein scale. This means we are able to see what effect, if any, an anisotropic potential has on the system as a whole. Analysing the same simulation data above, we threshold the protein connections to find clusters as before.

Taking each cluster individually, we firstly can calculate the eccentricity of the cluster as defined in Section 5.4.1. Given the size of the simulation systems we have used and the length of simulations performed, clusters of size three and four were present in large enough numbers to give smooth distributions of the eccentricity of clusters (clusters of size two have an undefined eccentricity). For clusters of size three we have plotted histograms of the eccentricity for both the anisotropic and isotropic potential and compared these

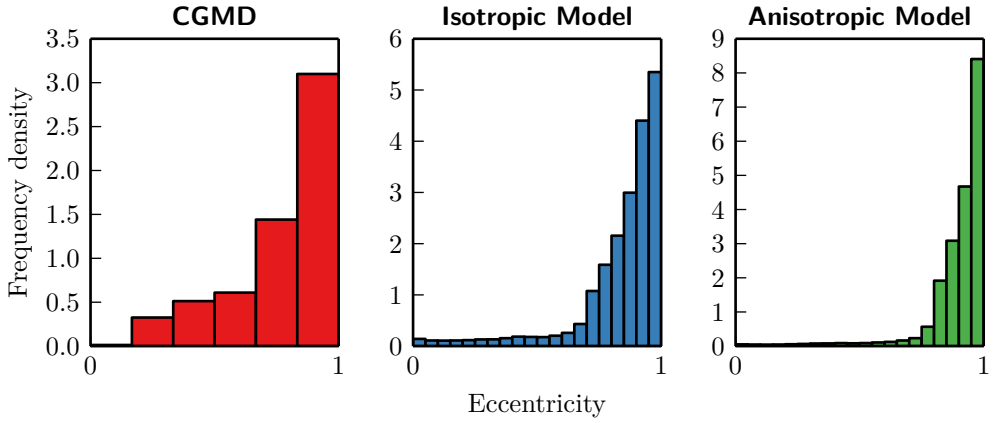

Figure 5.15: Histograms showing the eccentricity distribution of clusters containing three proteins from the CGMD simulation data, and MC simulations using an anisotropic and an isotropic potential. The number of samples for the three histograms are: $N_{\mathrm{CGMD}} = 107295$, $N_{\mathrm{Iso}} = 2033854$, and $N_{\mathrm{Aniso}} = 4751408$.

to the eccentricity of clusters measured from the CGMD simulation of a similar system, these histograms are shown in Figure 5.15 (the CGMD simulations were performed by Dr. Joseph Goose and were introduced in Chapter 1). From these histograms we can see that in both the anisotropic and isotropic cases the eccentricity distribution is very similar in shape to that obtained from the CGMD simulation. The equivalent histograms for clusters of size four are shown in Figure 5.16, where here we are limited by very small numbers of clusters in the CGMD simulations, restricting the amount we can infer from comparisons. The anisotropic distribution is skewed slightly towards higher values of eccentricity, which is in better agreement with the CGMD data. We can see that the peak in both the anisotropic and isotropic potential appears to be consistent with that in the CGMD data. Comparing the anisotropic and isotropic simulation data to each other we see that, as in the case of clusters of size three, there is a slight increase in the amount of eccentric clusters observed, although this is a much more muted effect with clusters of size four. From these comparisons of the eccentricity of the clusters, we have shown that the behaviour is matched reasonably well by both models, but that there is a slightly stronger agreement once we include the anisotropy of the protein-protein interaction.

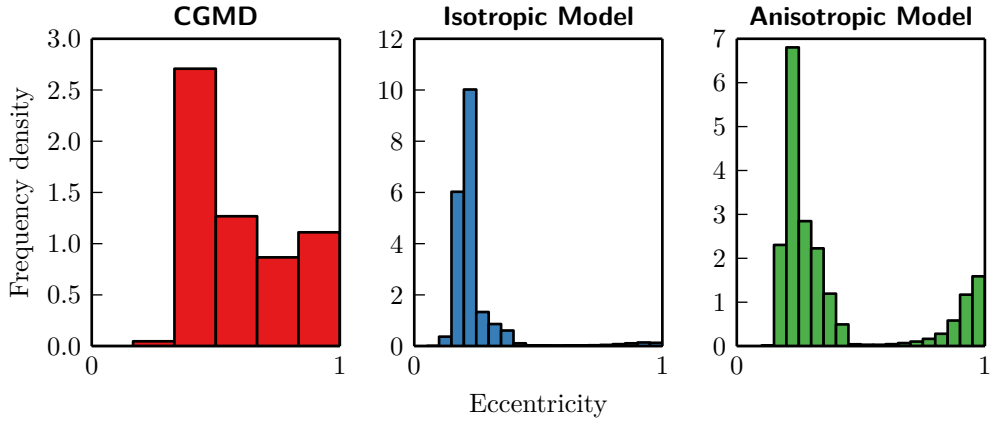

Figure 5.16: Histograms showing the eccentricity distribution of clusters containing four proteins from the CGMD simulation data, and MC simulations using an anisotropic and an isotropic potential. The number of samples for the three histograms are: $N_{\mathrm{CGMD}} = 79080$, $N_{\mathrm{Iso}} = 1579896$, and $N_{\mathrm{Aniso}} = 1845828$.

A further cluster-scale analysis we can perform is to compare the differences between the orientations of the proteins throughout the whole cluster, characterizing the alignment of proteins throughout the entire cluster, not just between pairs of proteins that are in contact. However, we need to account for the fact that our anisotropic potential has rotational symmetry of degree two. To account for this we can multiply the orientations of the proteins by two and take the modulo with respect to 360°, which results in the alignment of orientations that were originally pointing in opposite directions. With this correction applied to the orientations of the proteins, we can then calculate the distribution of differences between the orientations. An alignment score of one is defined as having no angular separation between the corrected orientations for a given pair of proteins. An alignment score of zero corresponds to having a corrected angular difference of 180°. For clusters of size two we have calculated the orientational alignment, which is shown by the histograms in Figure 5.17, where again we have performed the analysis for both the anisotropic and isotropic potentials and the CGMD data. The isotropic potential does not show a noticeable deviation from the uniform distribution, however, in the case of the anisotropic potential, there is a marked increase in the amount of protein

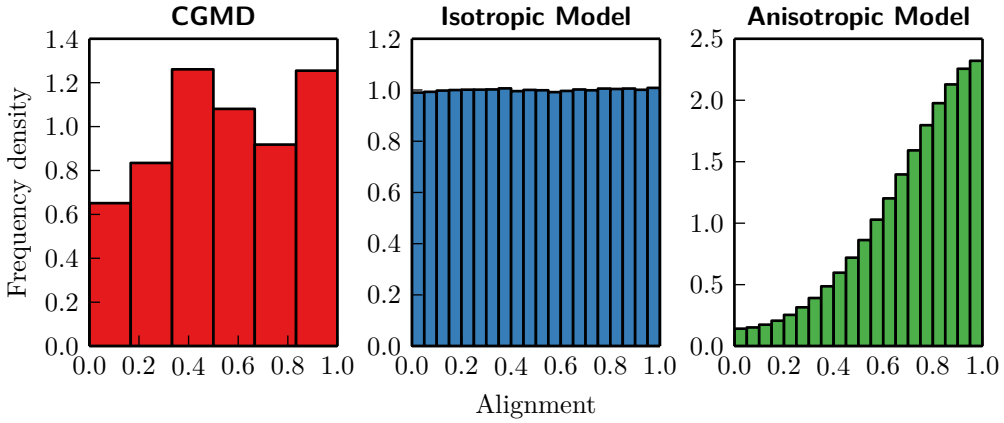

Figure 5.17: Histograms showing the alignment distribution of clusters containing two proteins from the CGMD simulation data, and MC simulations using an anisotropic and an isotropic potential. The number of samples for the three histograms are: $N_{\mathrm{CGMD}} = 216570$, $N_{\mathrm{Iso}} = 1769500$, and $N_{\mathrm{Aniso}} = 1494552$.

alignment. This is something that we would expect to find, given that the potential is anisotropic, but it is a larger effect than that observed in the CGMD data, although any conclusions drawn need to be made with care, given the small sample size of the CGMD data. Similar trends are observed for the alignment of proteins in clusters of size three in both the anisotropic and isotropic models, as shown in Figure 5.18. The alignment of the CGMD clusters appears to be more uniform, but the number of samples for the CGMD was approximately half that for the clusters of size two so any patterns are likely to be less significant.

Since we are calculating this metric for each of the proteins in a cluster, we can record it alongside the eccentricity of the cluster. By combining these two metrics we can see how the anisotropic potential affects both the protein clustering and the alignment of protein orientations within those clusters.

For clusters of size three, these histograms are shown in Figure 5.19; there is nothing to be learnt from an equivalent plot for clusters of size two since their eccentricities are all the same. The top row of the figure shows the data for all of the clusters in the simulations and the bottom row shows the same data after the application of a

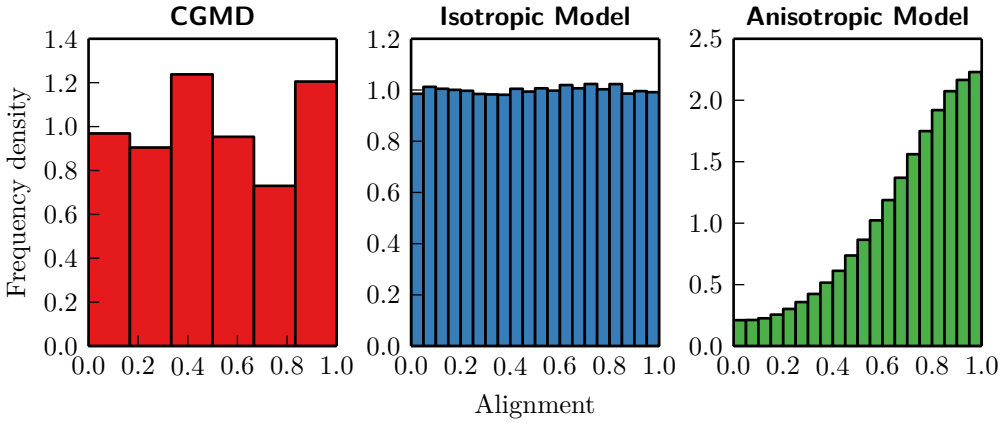

Figure 5.18: Histograms showing the alignment distribution of clusters containing three proteins from the CGMD simulation data, and MC simulations using an anisotropic and an isotropic potential. The number of samples for the three histograms are: $N_{\mathrm{CGMD}} = 107295$, $N_{\mathrm{Iso}} = 2033854$, and $N_{\mathrm{Aniso}} = 4751408$.

biasing procedure that corrects for the non-uniform sampling of cluster eccentricities. The biasing was performed in order to adjust for the extremely skewed distribution of eccentricities. The biasing proceure involved scaling the sampling in each row of the histrogram by dividing each number of samples in that row, giving the effect that the data for eccentricities was sampled uniformly. As before, it is apparent that there is only a small sample size for the CGMD data, but we can see the data is peaked toward clusters with high eccentricity, and demonstrates a slight bias towards more orientational alignment (with a peak close to one) for those eccentric clusters. The interesting result is visible in the comparison between the isotropic and anisotropic potential. For the isotropic case, we see that the clusters are not very eccentric and there is no observable dependence on alignment. This is more clear in the histogram biased to sample clusters of different eccentricities uniformly, where no pattern emerges. However, in the anisotropic case we see a shift in the eccentricity, as noted above, but more importantly we see a change in the alignment of the clusters based on their eccentricity. This is most clearly seen in the biased sampling histogram where we see that the most eccentric clusters are likely to be highly aligned, whereas the less eccentric clusters have more uniform

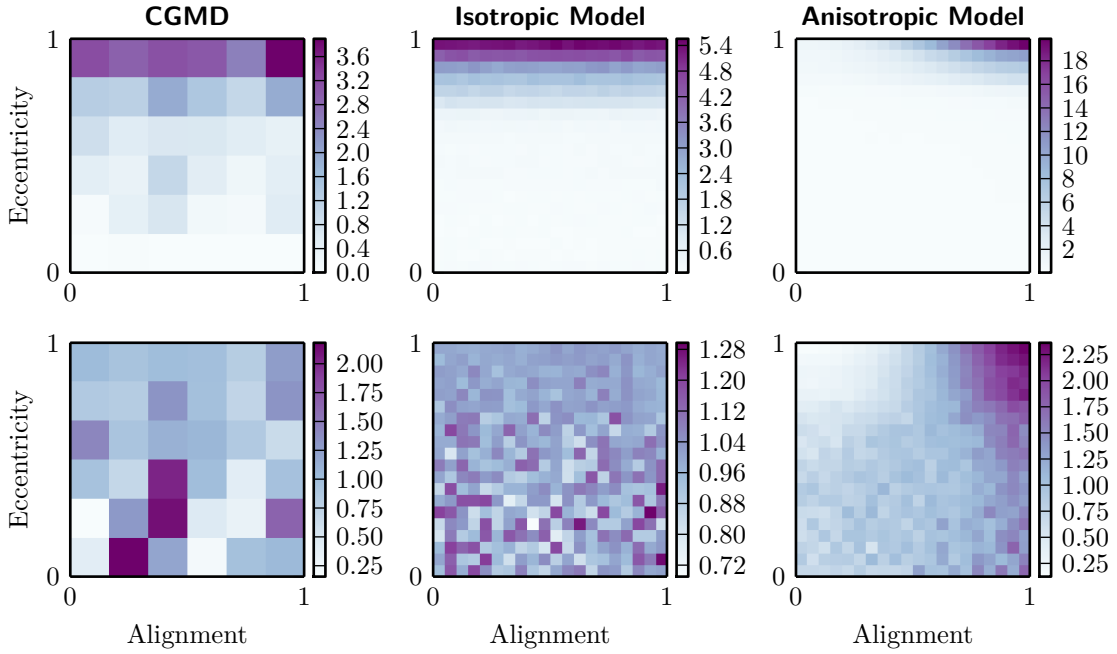

Figure 5.19: Frequency density plots of the alignment of proteins in clusters of size three with the eccentricity of those clusters. These are constructed for clusters of protein from the CGMD simulation and clusters from the isotropic and anisotropic potential simulations. Here we show the histograms sampled from the clusters in an unbiased manner **(top)** and also with a biasing correction applied to approximate a uniform sampling of cluster eccentricities **(bottom)**. The number of samples for the three histograms are: $N_{\mathrm{CGMD}} = 107295$, $N_{\mathrm{Iso}} = 2033854$, and $N_{\mathrm{Aniso}} = 4751408$.

alignment. This is evidence that the introduction of an anisotropic potential leads to behaviour that emerges at the scale of clusters of proteins. There is not such a strong bias in the CGMD data for eccentric clusters of size three, which is perhaps indicative of more complicated behaviour that is not captured by our model.

5.5 Using other published PMFs to parameterize our discrete model

All of the simulations so far have been based on the PMFs calculated for the interaction of NanC in Chapter 3. In this section we will instead parameterize our discrete protein model using information from PMFs published in the literature.

The system that we will attempt to parameterize is that of a membrane populated by rhodopsin proteins; PMFs for rhodopsin, using a range of orientational configurations, were published by Periole, Knepp, et al. (2012). It is from this paper that we will gather information about the pairwise interaction of rhodopsin and parameterize our model.

5.5.1 Obtaining the generalized PMF features from the paper

The PMFs for the G-protein coupled receptor, rhodopsin, were calculated in a bilayer consisting of $(C20:1)_2PC$ lipids, which have been shown to minimize the effect of hydrophobic mismatch on rhodopsin interaction (Botelho et al. 2006; Periole, T. Huber, et al. 2007). Rhodopsin's structure consists of seven transmembrane helices and one amphipathic helix, which is aligned with the surface of the cytoplasmic membrane leaflet. Through a series of extended self-assembly MD simulations, lasting up to 64 μs, the distribution of protein contacts as a function of orientation was analyzed. From this distribution, they identified the most probable orientational states and obtained the PMFs using umbrella sampling. The system used to calculate the PMFs was similar to that used for NanC in Chapter 3, but the method used to perform the orientational restraint was the virtual bond algorithm of Boresch et al. (2003). This algorithm restrains the proteins based around three dihedral restraints, two bond angle restraints and one distance restraint.

It was clear from their self-assembly simulations that the strongest attraction between pairs of rhodopsins was when they were aligned in a tail-to-tail configuration, with the amphipathic helices interlocking. The PMFs were thus calculated for the following orientations (as measured in the plane of the membrane in a similar manner to that used for NanC): $(10^\circ, 10^\circ)$, the tail-to-tail conformation; $(180^\circ, 180^\circ)$ and $(-152^\circ, -152^\circ)$, which are similarly aligned as the tail-to-tail configuration; and $(-90^\circ, 90^\circ)$ and $(-90^\circ, -90^\circ)$, which are more closely packed configurations with a larger buried surface area. This set of relative orientations is illustrated in Figure 5.20.

The PMFs calculated for these five orientational configurations of rhodopsin are similar

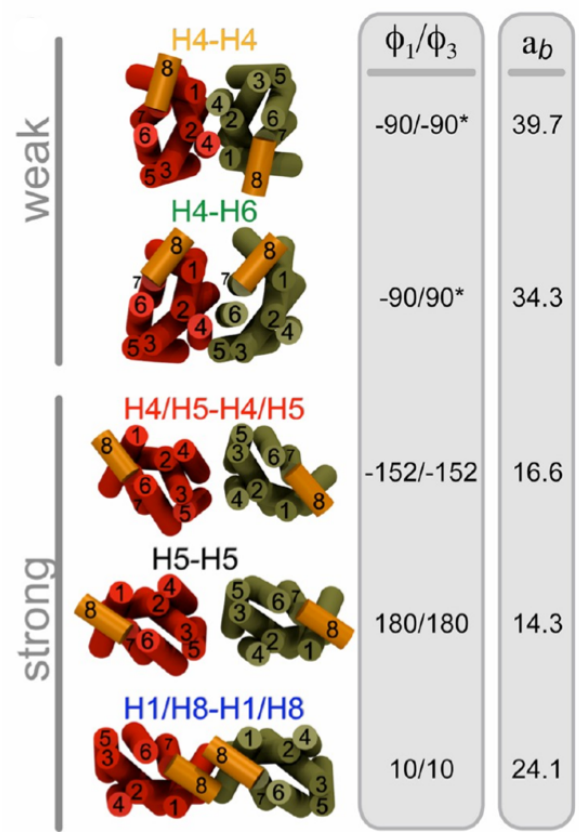

Figure 5.20: Relative orientations of rhodopsin used to calculate the PMFs shown in Figure 5.21. The relative angles of the two proteins are shown alongside each orientational configuration, along with the buried surface area in nm^2. The configurations have been grouped by the strength of their interactions. This figure was adapted with permission from Periole, Knepp, et al. (2012).

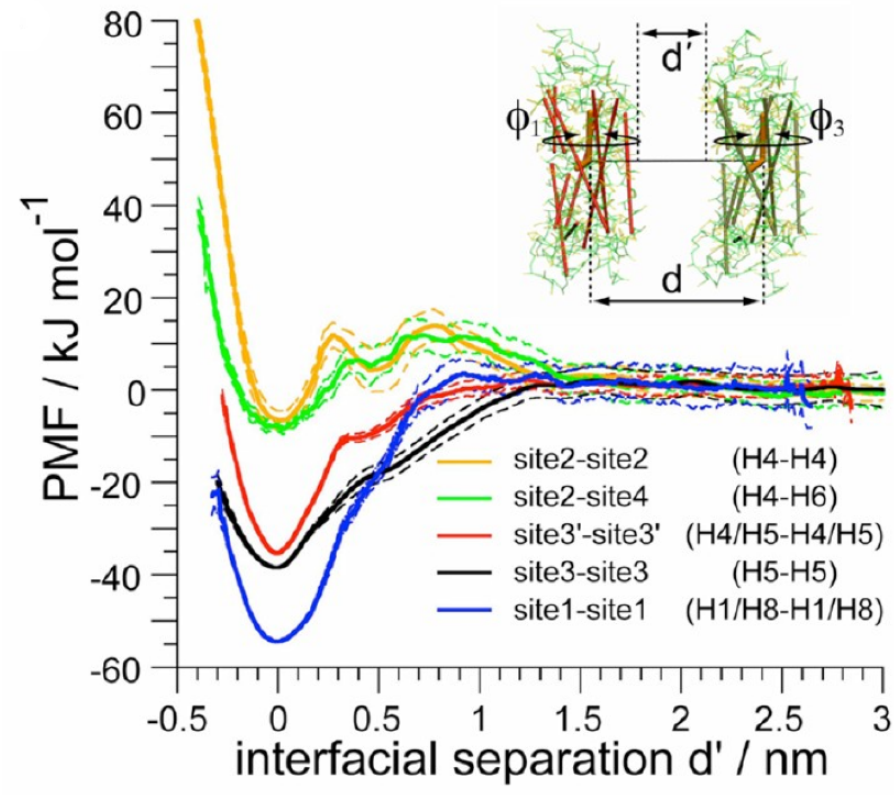

Figure 5.21: PMFs for multiple relative orientations of rhodopsin. The relative orientations correspond to those illustrated in Figure 5.20. It is important to note that, contrary to the convention throughout this work, the separations of the proteins are given as an interfacial separation instead of as a separation of the proteins' centres of mass. This figure was adapted with permission from Periole, Knepp, et al. (2012).

in character to those calculated for NanC, as can be seen in Figure 5.21. They have a potential well at small separation, and some of them have a small barrier at slightly larger separation. There are also signs of features in the PMF that are located in a similar position as those we have earlier demonstrated were a product of the protein-lipid-protein interactions.

Contrary to our findings for NanC, the strength of the rhodopsin interaction is not determined by the buried surface area between the proteins. This is something that we would perhaps expect, as the structure of rhodopsin is more complicated than that of the relatively-featureless NanC. The strongest interaction appears in a configuration in

	Rhodopsin orientations, $(\hat{\phi}_i, \hat{\phi}_j)$				
Parameter	$(10°, 10°)$	$(180°, 180°)$	$(-152°, -152°)$	$(-90°, 90°)$	$(-90°, -90°)$
ϵ_{min}	55 kJ mol^{-1}	38 kJ mol^{-1}	35 kJ mol^{-1}	8 kJ mol^{-1}	5 kJ mol^{-1}
$\epsilon_{\text{barrier}}$	0 kJ mol^{-1}	0 kJ mol^{-1}	0 kJ mol^{-1}	10 kJ mol^{-1}	10 kJ mol^{-1}
r_{min}	4 nm	4 nm	4 nm	3 nm	3 nm
r_{barrier}	5 nm	5 nm	5 nm	4 nm	4 nm
σ_{min}	1.5 nm^{-1}	1.5 nm^{-1}	1.5 nm^{-1}	1.5 nm^{-1}	1.5 nm^{-1}
σ_{barrier}	1.2 nm	1.2 nm	1.2 nm	1.2 nm	1.2 nm

Table 5.4: Parameter values obtained from the PMFs calculated by Periole, Knepp, et al. (2012) for the interaction of rhodopsin as various orientations. The orientations are illustrated in Figure 5.20.

which the amphipathic helices of the two rhodopsins are interlocked and with interactions between residues on two of the transmembrane helices.

From the PMFs presented, we are able to infer the approximate depths and widths of the potential wells, and the height, width and relative positions of any barrier. These parameters can be used in our combined Morse-Gaussian potential model for the interaction of two proteins. Obtaining the location of the minima from the PMFs presented is a little more complicated, as they are presented as a function of interfacial separation, and not as a function of the separation of their centres of mass. However, we obtain estimates for the locations of the minima using the approximate width of rhodopsin along a given axis; all of the PMFs are presented for rotationally equivalent angles, so we are able to approximate their separation by their width along this axis. Here we have taken the width along the long axis of rhodopsin to be 4 nm and along the short axis to be 3 nm; these are not precise figures, but given that the other parameters are estimated from the figures, it is unlikely that this estimation of the radius will have a large effect on the simulation. The parameters, obtained from the PMFs presented in the paper and the approximation of the width of the protein, are given in Table 5.4.

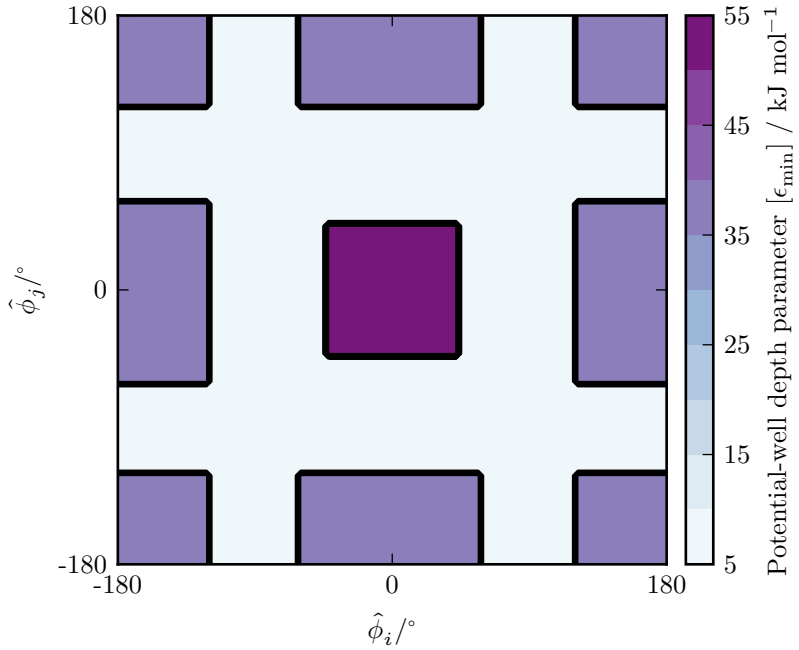

Figure 5.22: Interpolation scheme for the potential-well depth parameter, $\epsilon_{\min}$, in the Morse-Gauss potential in Equation (4.15) (on page 88). The three regions are clearly marked by the different values of $\epsilon_{\min}$: the region around the tail-to-tail orientation is centred on $(0^\circ, 0^\circ)$; the other strong interactions occur at all the combinations that are related to the tail-to-tail configuration by a 180° rotation of one or both of the proteins; and the weak interactions occupy the in the intervening space.

The dihedral angles used in the paper to classify the orientations of the two proteins are both measured outwardly from the line connecting the two proteins centres: an angle of 0° corresponds to the amphipathic helix being closest to the other protein, the so called "tail-to-tail" configuration.

To obtain our parameterization in the full rotation of both proteins, we must make certain simplifying assumptions about the form of the potential. Given the explanation of the behaviour of the proteins in the paper, these assumptions are not likely to be completely inaccurate, but they are a source of potential error in the model. Our parameterization for the potential-well depth parameter, $\epsilon_{\min}$, is shown in Figure 5.22.

The first region of interest in this parameterization if for the strongest interaction in the tail-to-tail configuration, and a small region of angles around it; in our system we will

take this to be the 90° region centred on 0°. The other strong interactions occur around the opposite alignment of the proteins, where the amphipathic proteins are oriented away from the other protein. Here we make the assumption that all interactions that are aligned similarly (both proteins long axes aligned with the line connecting their centres), excluding the tail-to-tail configuration, are of the same strength, which we base on the depths of the $(180°, 180°)$ and $(-152°, -152°)$ PMFs. From the self-assembly simulation of Periole, Knepp, et al. (2012), the regions of orientational-configuration space corresponding to these alignments in which there is significant interaction of the proteins occurs for a larger range of angles than it does for the tail-to-tail configuration, and as such, we choose to make these regions in our parameterization correspondingly larger, with widths of 135°. For the remaining interactions, we have used the weak PMFs as a guide. We use the same interpolation scheme for all of the parameters in Table 5.4, where the parameters for each of the three regions, tail-to-tail, strong and weak, are obtained from: the $(10°, 10°)$ PMF parameters; an average of the $(180°, 180°)$ and the $(-152°, -152°)$ PMF parameters; and an average of the $(-90°, 90°)$ and the $(-90°, -90°)$ PMF parameters, respectively.

The parameterization in Figure 5.22 has very sharp changes in potential between plateaus of constant values. As our justification for the parameterization scheme chosen is limited, a more reasonable scheme would be for the parameters to vary smoothly between their peak values; we have no evidence to support any shape features in the PMFs between the rhodopsin proteins, so we have chosen to use a more idealized, smoothly-varying potential in our model of the rhodopsin system. To smooth the features of the interpolation in Figure 5.22 we used a Gaussian kernel of width 36°, one tenth of the orientational domain size. The result of this smoothing process is shown by our final interpolation scheme for the potential-well depth parameter, ϵ_{min}, in Figure 5.23. The application of the smoothing kernel leaves the parameter values at the orientational configurations of the PMFs unchanged, but smooths the sharp transitions between the

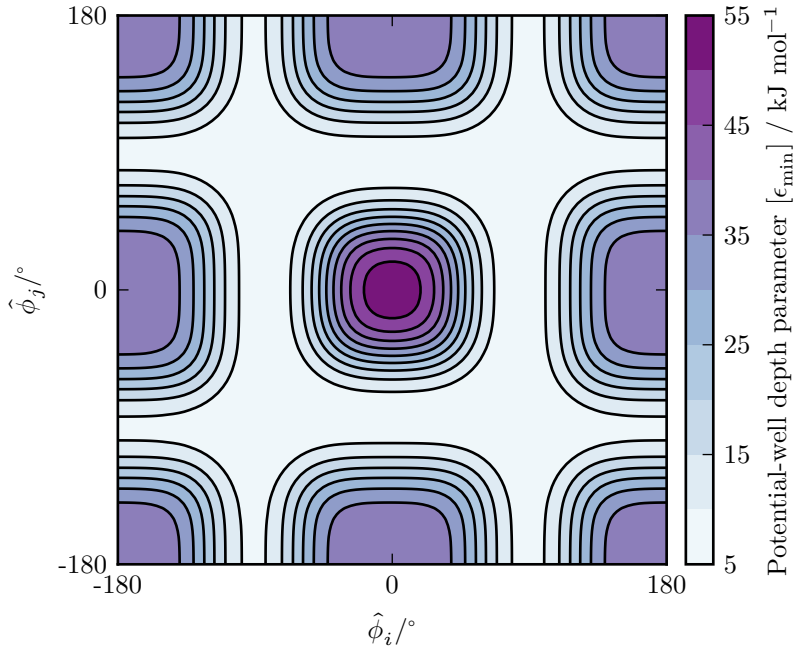

Figure 5.23: Interpolation scheme for the potential-well depth parameter, $\epsilon_{\min}$, in the Morse-Gauss potential in Equation (4.15) (on page 88) shown in Figure 5.22 following the application of a Gaussian smoothing kernel. The values of the parameters at the orientations on which we have based the initial parameterization are left unchanged by the process.

regions.

5.5.2 Investigating the stability of rhodopsin rows-of-dimers

The coarse-grained rhodopsin model that Periole, Knepp, et al. (2012) used to calculate their PMFs was used to investigate the stability of a system of rows-of-dimers, similar to those seen in atomic force microscope observations (Fotiadis et al. 2003; Liang et al. 2003). The rows-of-dimers system consisted of pairs of rhodopsin proteins arranged in the tail-to-tail configuration as dimers, which in turn were arranged in rows perpendicular to the axis of the dimer. Periole, Knepp, et al. (2012) used a system consisting of 16 rhodopsin proteins, arranged in two parallel rows of four dimers. This is shown in Figure 5.24, where the simulation's unit-cell, containing the 16 rhodopsins, is marked.

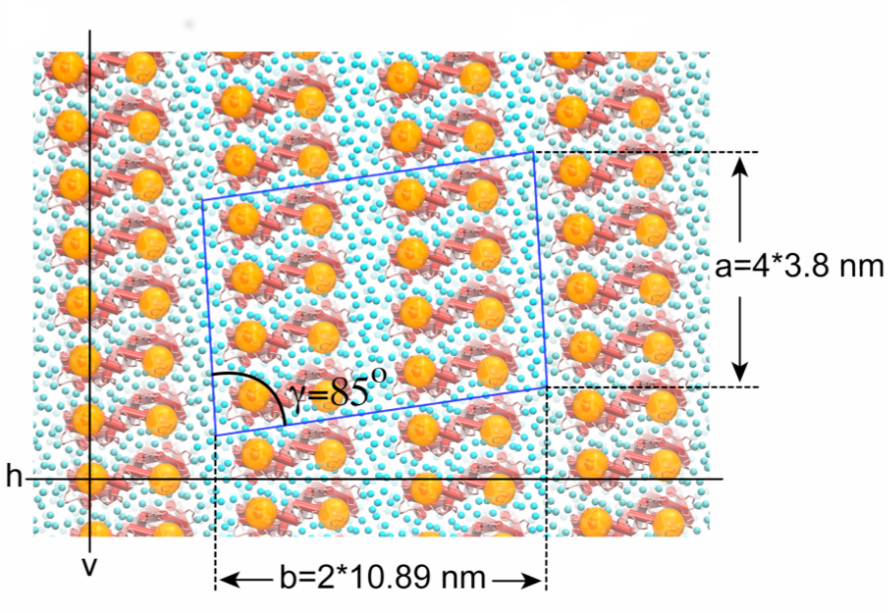

Figure 5.24: Initial positions for the simulation of 16 rhodopsins using MD. The rhodopsin pairs are arranged in the tail-to-tail configuration, and the unit-cell has an angle of 85°. Adapted with permission from Periole, Knepp, et al. (2012).

Our simulation system consists of 96 rhodopsin proteins, arranged in four rows of 16 dimers. We use the same unit-cell dimensions but with an angle of 90°; our representation of the rhodopsin system is symmetric around 0°, and this simplification follows from the symmetry of the idealized potential interpolation.

5.5.2.1 Time scaling methods in a close-packed system

The methods used to assign a timescale to the Monte Carlo simulation both rely on the initial system being dispersed enough that the constituents do not interact with each other. With the rhodopsins arranged in rows of dimers, we do not satisfy this condition. Consequently the scaling method will no longer result in a reasonable timescale for the Monte Carlo steps in our simulation; the diffusion coefficient will not necessarily be a representative parameter for the motion of the individual proteins in our system. The proteins will be moving in a system with a greatly reduced mobility. This means that

by applying the scaling scheme, we will obtain a timescale that is much shorter than the actual timescale over which our simulation will evolve the system. It will, in effect, provide us with some form of lower bounding estimate for the timescale. This may not be a quantitatively useful tool for investigating the dynamics, but it is an indicator as to whether we are observing motion that is characteristic for timescales over similar orders of magnitude as that used in the simulation of Periole, Knepp, et al. (2012). Throughout this section we use the units ns* to signify that these are now a lower bound on the simulation time and not a reasonable representation on the simulation time. The diffusion constant used in this calculation was 0.00014 nm*2 ns^{-1}, as measured by fluorescence microscopy experiments of Govardovskii et al. (2009) using gecko rod membranes.

The simulations of our system of 96 rhodopsins were 10^5 Monte Carlo steps in length. We performed 240 repeat simulations, which enabled us to investigate the behaviour of our system through the behaviour of the ensemble, rather than using a single instance, as was the case with the MD simulation. From the plot of the mean squared deviation in Figure 5.25 we can see that the proteins are in the same diffusive regime throughout; there are no changes in diffusive behaviour. The lower bound on the time scale is a simulation length of just of 2 μs*. This is shorter than than the 16 μs MD simulation performed by Periole, Knepp, et al. (2012), but as we are unable to quantify by how much our simulation time is an underestimate, the only reasonable comparison that we are able to make about the simulation lengths is that they are likely to be of a similar order of magnitude. We should, therefore, be able to observe similar behaviours in both simulations, if our model is accurately capturing the interactions responsible for the dynamics of the rhodopsin system.

5.5.2.2 Visualizing the simulations of rhodopsins arranged in rows of dimers

The 16 μs coarse-grained simulation of Periole, Knepp, et al. (2012) demonstrated that the rows-of-dimers system was stable. From the initial configuration of the system, as

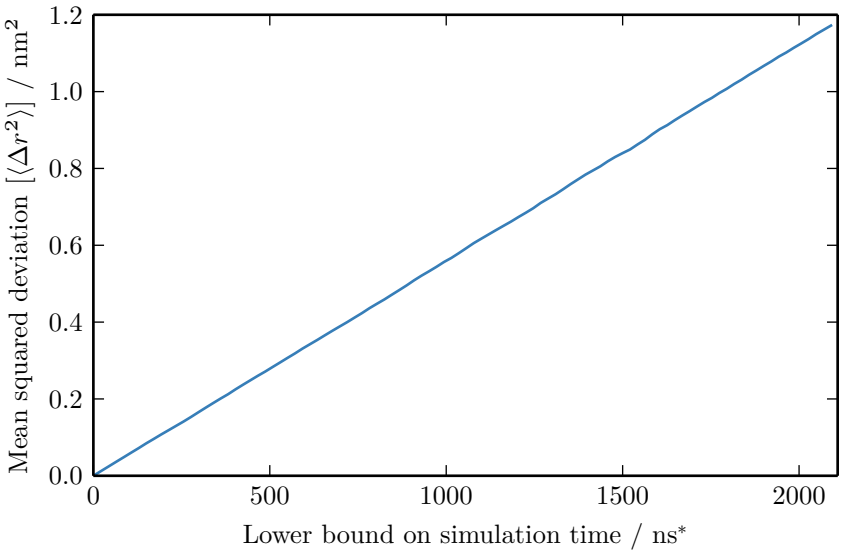

Figure 5.25: The mean squared deviation of the rhodopsin proteins. The lower bound on simulation time was calculated using the diffusion-scaling method introduced in Chapter 4.

shown in Figure 5.24, the system evolved to a final configuration shown in Figure 5.26. It is clear that the rows of proteins have maintained their structure, with only a single pair being out of alignment.

To visualize the rows-of-dimers structure in our simulation we have plotted the positions of all proteins at two points during the simulation. The first is after 1000 ns*, shown in Figure 5.27, and the second after 2000 ns*, shown in Figure 5.28. In these figures, the positions of the proteins are represented by a coloured circle at the position of their centre of mass. Proteins that form a dimer are shown using different shades of the same colour. From both these figures we can see that the overall structure of the rows-of-dimers is maintained, with the orientations of the proteins pairs being well constrained by the interaction between the proteins visible as a small spread in the angles. However, the positions of the proteins was less consistent throughout the simulation, with a clear spread in their locations visible in both Figures 5.27 and 5.28.

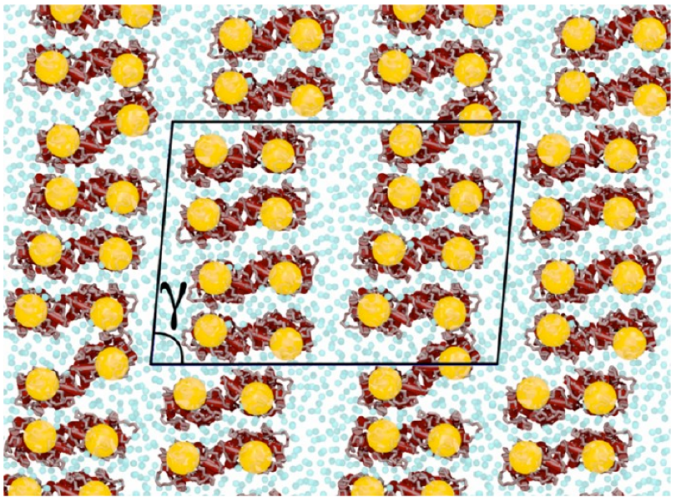

Figure 5.26: Rhodopsin positions after 16 μs of coarse-grained MD simulation of 16 rhodopsins. Adapted with permission from Periole, Knepp, et al. (2012).

5.5.2.3 Using dimer separation as a measure of stability

The assess the instability in the positions of the proteins in their dimers, we can look at the distributions of dimer separation for the proteins in the simulation. Figure 5.29 compares the distributions of dimer separation for the rows-of-dimers simulation at both 1000 and 2000 ns*, corresponding to the protein positions shown in Figures 5.27 and 5.28, respectively. Out assessment of the instability visible in the positions is corroborated by the distributions of the dimer separation; it is clear that there is an increase in the dimer separation between 1000 and 2000 ns*.

We can see this instability more clearly in Figure 5.30, where it is clear that the mean dimer separation is slowly increasing throughout the simulation, and the distribution is also getting wider.

From these data it is apparent that our model is not sufficiently representing the interactions that are integral to maintaining the stability of the rhodopsin rows-of-dimers system. From Figure 5.30 it is clear that the stabilization of the dimers themselves is dependent on more than just the interaction potential that is calculated for a pair of

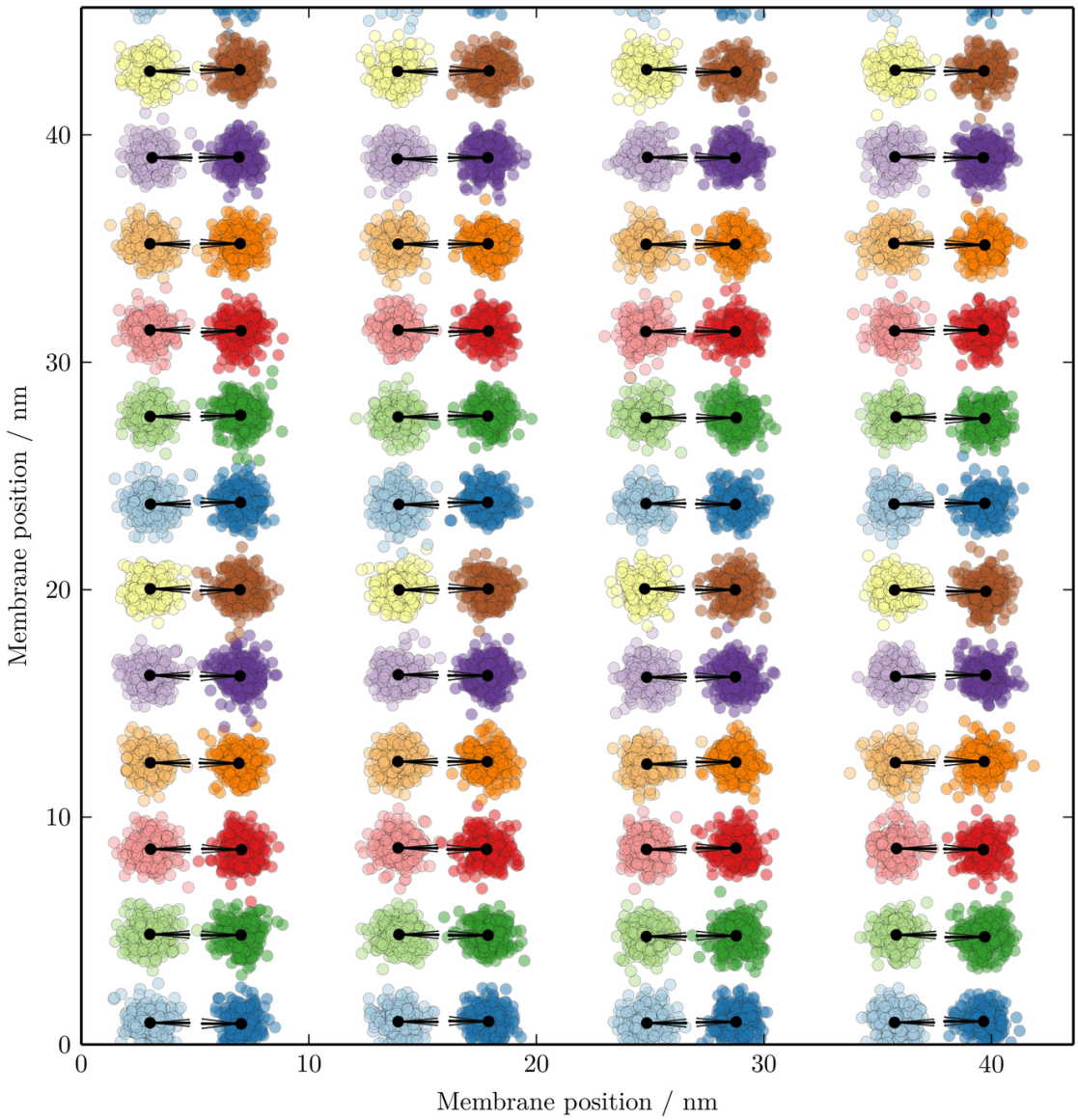

Figure 5.27: Positions of 96 rhodopsin proteins after 1000 ns* of simulation in a rows-of-dimers configuration. The positions of rhodopsins in 240 repeat simulations are shown, with each circle positioned at the centre of mass of a specific protein in one of the repeat simulations. The proteins are coloured so that each pair of proteins is identifiable (e.g. green paired with light green, red paired with light red and so on). The black dots represent the initial position of the corresponding proteins. The thick black lines are the mean orientation or the protein and the thin black lines represent one standard deviation of the distribution of orientations.

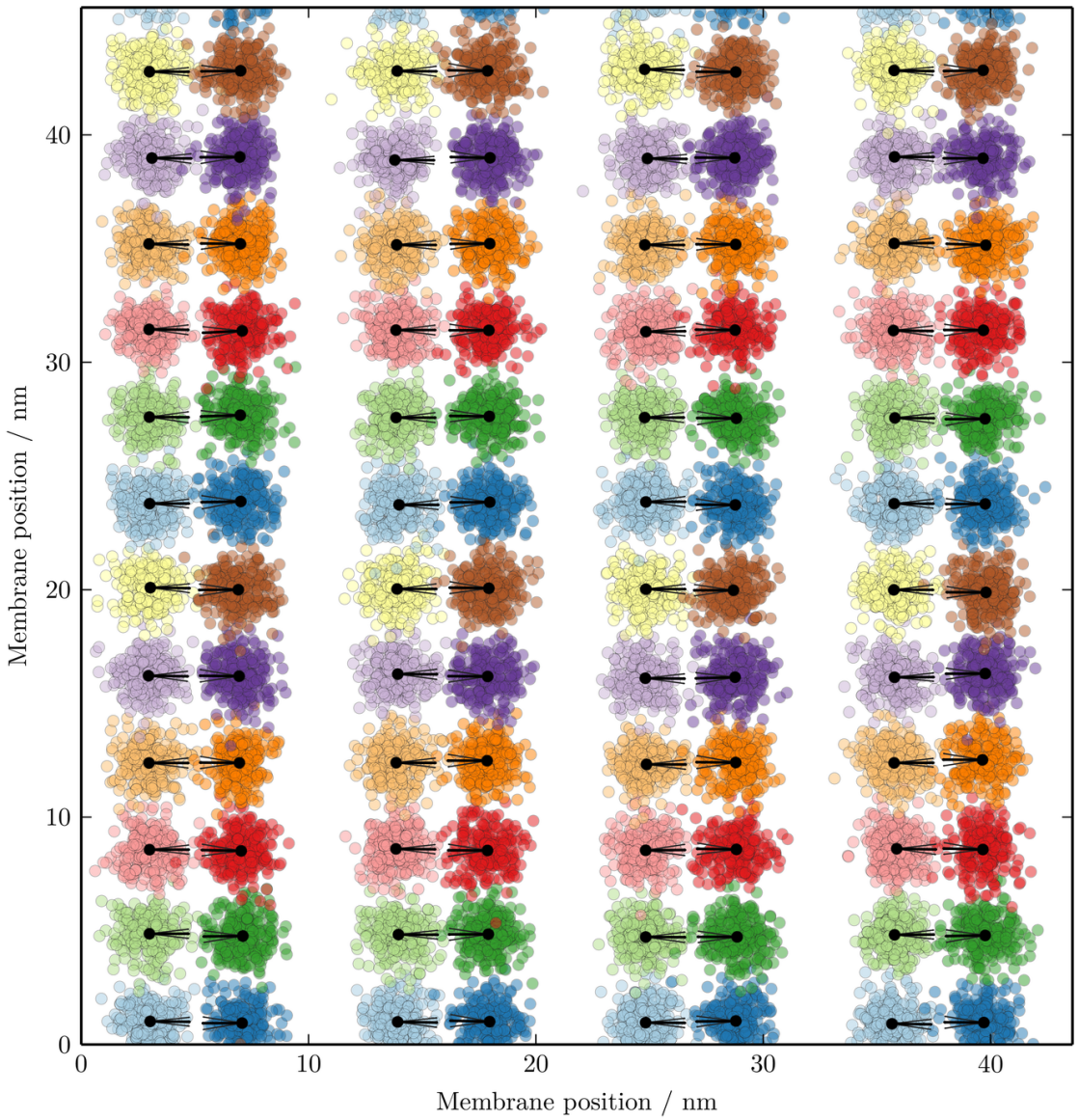

Figure 5.28: Positions of 96 rhodopsin proteins after 2000 ns* of simulation in a rows-of-dimers configuration. The positions of rhodopsins in 240 repeat simulations are shown, with each circle positioned at the centre of mass of a specific protein in one of the repeat simulations. The proteins are coloured so that each pair of proteins is identifiable (e.g. green paired with light green, red paired with light red and so on). The black dots represent the initial position of the corresponding proteins. The thick black lines are the mean orientation or the protein and the thin black lines represent one standard deviation of the distribution of orientations.

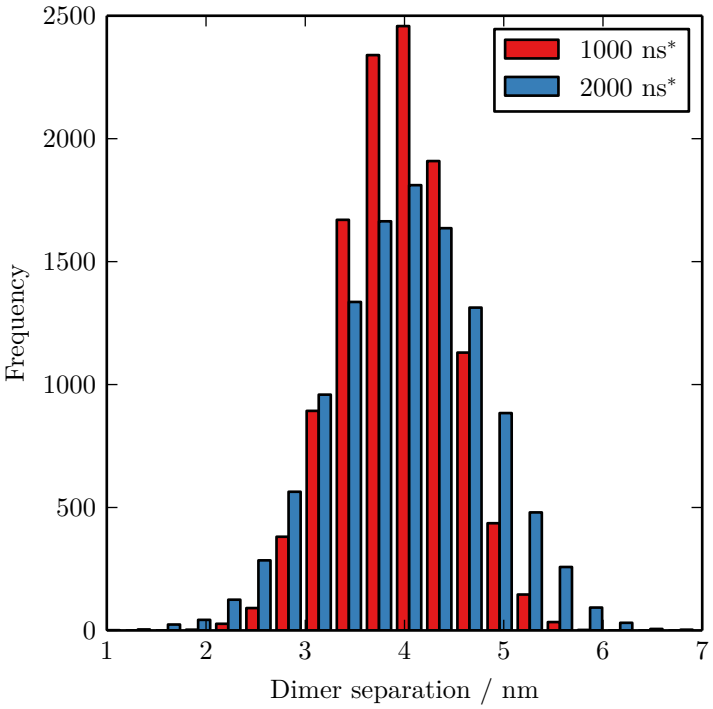

Figure 5.29: Distributions of dimer separation at 1000 ns* and 2000 ns*. The dimer separation is the distance between the centres of mass of the two proteins in each dimer.

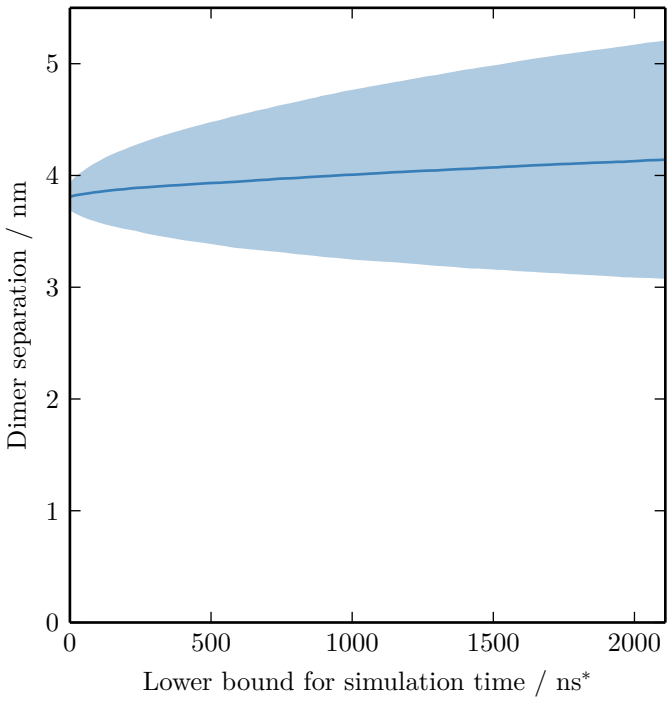

Figure 5.30: Mean dimer separation throughout the entire simulation of 96 rhodopsin proteins arranged in a rows-of-dimers configuration. The light blue region corresponds to one standard deviation.

isolated rhodopsins. The most likely reason that our model is not stable in this situation is that we are using implicit lipids; this has been a reasonable approximation in the more sparsely packed system of NanC, but does not adequately represent the effect of individual lipids in this more densely packed rhodopsin system. In Figure 5.26 we see that there is often only room for a single lipid between adjacent rows of dimers and there are fewer than 10 lipids separating adjacent rows. In a densely packed system like this, the lipid mobility will be much reduced, so our implicit lipid model that acts as a fluid-like bilayer will be particularly unsuitable in capturing the protein-lipid-protein interactions, or probably the dimer-lipid-dimer interactions, that are likely operating to stabilize the system as well as the interactions that were characterized by the PMF calculations, on which our parameterization is based. However, the parameterization of the protein-protein interactions appears to capture the orientational behaviour reasonably well, as indicated by the rotational stability of the proteins shown by the mean orientation and spread of the orientations illustrated by the standard deviations shown in Figures 5.27 and 5.28. So although the model doesn't capture all of the processes that are important in describing the system, it does appear to describe some of the system's behaviours relatively well.

5.6 Summary

In this chapter we extended the model developed in Chapter 4 with the introduction of rotational degrees of freedom for the proteins. The development of the extended model attempted to follow the pattern of development used for the isotropic model of protein-protein interaction. The anisotropic model was parameterized using MD simulations of the NanC system: rotational diffusion properties were obtained from the mean squared rotation of NanC and the interaction profile was obtained from the rotationally restrained PMFs (presented in Chapter 3). The fitting of the PMFs to a generalized model of orientationally-dependent pairwise interaction was performed for

the orientational configurations used in the original MD simulations. To complete the parameterization of the anisotropic interactions between pairs of proteins, the parameters for the rotational potential were interpolated.

The interpolation scheme used was based on an idealized representation of the NanC proteins as objects with an elliptical cross-section. Using this representation we were able to calculate how certain parameters in the interaction profile may change with the orientations of the two proteins, subject to several assumptions. The two assumptions were based on the results in Chapter 3 where we showed that local minima in the PMF correlated with the ability of lipids to fit in the intervening space, and the extension of this being that the position of features in our generalized interaction profile will also be dependent on the ability for lipids to fit between the two proteins, which is strongly dependent on the distance between the proteins surfaces, and, in the limit of small separation, the distance at which they are in contact. For the strength of the interaction between pairs of NanC, we found in Chapter 3 that there was a dependency on the buried surface area of the protein pair when they were in contact. Using our elliptical model we calculated an expression for the buried surface area when the ellipses were in contact. By numerically solving this expression, we were able to calculate how the buried surface area for two ellipses changed with their relative orientations. This buried-surface-based interpolation scheme was used for the parameter that determined the depth of the potential well in our anisotropic model of the pairwise protein-protein interaction; this interpolation scheme was shown to give a much improved parameterization when compared to a simple, smoothly varying, interpolation scheme.

We performed virtual-move Monte Carlo simulations of the anisotropic model, using the method employed in Chapter 4, to explore differences between the anisotropic and isotropic models, and compare them to data obtained from the CGMD simulations performed by Dr. Joseph Goose. The anisotropic model showed improved behaviour on some of the clustering analysis metrics, for instance the alignment of clusters of size

two and eccentricity of clusters of size four, however the CGMD simulation, although extensive for a simulation of its type, did not provide sufficient data points for meaningful comparisons of a lot of the metrics.

As a final implementation of our anisotropic model, we attempted to parameterize it using data from other published PMFs for the pairwise interaction of membrane proteins. We used PMFs for the association of rhodopsin proteins to parameterize our model (Periole, Knepp, et al. 2012). This parameterization used a much simpler approach to the interpolation, as there was no simple geometric basis to the shape of the PMFs that would be amenable to representation within our anisotropic model. We used this model to attempt to simulate a rows-of-dimers system of rhodopsins, similar to those seen in disk membranes (Fotiadis et al. 2003; Liang et al. 2003). This was a much more closely packed system than our NanC system. The result of this was firstly that our time-scaling methods were not applicable, meaning that we were only able to use what we argued would be some form of lower-bound on the simulation time. Secondly, we observed significant instability in the dimers when measured by the average dimer separation. This observation is something that would perhaps be expected when using our model for such a close-packed system because our intrinsic model of lipids is not a good representation when the number of proteins in the system approaches the number of lipids. This would have a particularly marked effect on the behaviour of the lipids themselves, as a result of their reduced mobility, were an important factor in the stability of the dimers within the membrane. The dimers that formed did have relatively stable orientational configurations, perhaps suggesting that the orientational stability of the individual proteins in their dimers had a greater dependence on the contributions captured in the PMFs calculated for the sparse system by Periole, Knepp, et al. (2012).

Chapter 6

Conclusions

This chapter summarizes the work presented in this thesis and the conclusions we have drawn. Firstly, the PMFs calculated in Chapter 3 are assessed and the conclusions we made regarding the correlation between buried surface area and potential well depth are presented. We also discuss our analysis of the protein-lipid-protein interactions that this work highlighted. Next we talk about the performance of our simplified bilayer model and how it reproduced some of the clustering dynamics observed in the more computationally-expensive MD simulations we were using as a comparison. Finally we conclude with an appraisal of the application of our model to simulating a close-packed 'rows-of-dimers' configuration of rhodopsin, parameterized using PMFs from the literature. This simulation, although not fully capturing the translational stability of rhodopsin dimers, did capture the rotational stability of the dimers reasonably well.

6.1 Free energy landscape of NanC

The first stage of our work-flow for parameterizing a discrete model of membrane proteins in a simplified bilayer was to characterize the interaction between a pair of NanC proteins, which was our protein of interest. Our initial characterization, presented in Chapter 3, used the one-dimensional potential of mean force; the form of the PMF was similar in form to other published PMFs for membrane proteins: it had a deep potential well at small inter-protein separations and a small barrier at a slightly larger separation. The depth of approximately -70 kJ mol^{-1}, which although not directly comparable to an experimentally determined value, was of a similar magnitude to the depth of the PMF calculated by Casuso et al. (2012) for OmpF trimers: -150 kJ mol^{-1}.

The next step of the characterization of the protein-protein interaction was to introduce orientational degrees of freedom. We did this with the addition of rotational restraints to the proteins when calculating the PMFs for four different orientational combinations of NanC; these orientations corresponded to various values of the buried surface area (the area that is inaccessible to lipids when the proteins are in contact). Using these PMFs we demonstrated that the depth of the PMF was negatively correlated with the amount of buried surface area: for a larger buried surface area between the two proteins, the depth of the PMF was greater. This result is contrary to the result published by Periole, Knepp, et al. (2012), but in their rhodopsin system a complex series of binding modes were identified, along with relative orientations that showed approximately no binding affinity. Our orientationally restrained PMFs also identified features that were attributed to protein-lipid-protein interactions. In such a highly restrained system we are able to observe behaviour of the lipids that would be averaged over in a system in which the protein were free to rotate. These features, manifested as local minima in the PMFs, corresponded to separations at which the lipid between the two proteins were optimally packed. This was corroborated by analysing the distribution of lipids around a single NanC protein, where rings in the average lipid density were observed. Using the radius

of these rings, we were able to predict the positions at which we would expect to observe optimal lipid-packing between two proteins and there was good agreement between these predictions and the features in the PMFs.

In the following chapters we used these PMFs to parameterize a coarser-grained model of a bilayer. By representing the proteins in the system as points embedded in a two-dimensional bilayer, which have some intrinsic orientation, we were able to use these PMFs to parameterize the protein-protein interaction strength (extending the approach introduced by Yiannourakou et al. (2010)).

6.2 Discrete simulation of an implicit-lipid bilayer

The remaining chapters of the thesis centred around the development of a simplified model representation of the bilayer, which we parameterized using the MD simulations in Chapter 3.

6.2.1 Isotropic protein-protein interaction

Our first representation of the protein-protein interaction was using an isotropic model. The proteins were represented by the position of their centre of mass and the interaction was parameterized using the one-dimensional PMF calculated for NanC in Chapter 3. Simulations were initially performed using a Metropolis Monte Carlo method, with individual movements of the proteins. Using this method we investigated the regions of validity of the simulation parameters, such that the time-scaling techniques of Romano et al. (2011) and Jabbari-Farouji et al. (2012) remained valid. Through comparison of the model with data from coarse-grained MD simulations of a similar system performed by Dr. Joseph Goose (introduced in Chapter 1), it was apparent that the individual move scheme, although qualitatively similar, did not result in the same diffusive clustering behaviour that was observed in the MD simulations. In our simulations, as the proteins interacted with each other, they formed clusters, and the larger the size of the cluster,

the smaller the probability that the Monte Carlo method would result in a concerted movement of the cluster as a whole. Such movements were something that we would expect to see in a membrane system and are typical of the movement of proteins in the MD simulations. To improve the simulations of the model, we implemented a Monte Carlo simulation method with cluster-based moves called virtual-move Monte Carlo, which is known to produce reasonable dynamics (Whitelam and Geissler 2007). Implementing the virtual-move Monte Carlo simulations improved the quantitative agreement between the diffusive clustering behaviour of our model and the coarse-grained model simulated with Molecular Dynamics. However, given the depth of the minimum of the PMFs used to parameterize the interactions in the system, we were unlikely to observer much interesting behaviour within and between the clusters. With such strong interaction energies, the proteins were very unlikely to dissociate once they had formed clusters. If the depth of the interaction potential was less pronounced, then it would be likely that we would have observed more interesting clustering behaviour included the exchange of proteins between clusters and more varied cluster shapes.

6.2.2 Anisotropic protein-protein interactions

The isotropic model developed in Chapter 4 was extended to include an angular dependence to the protein-protein interactions in Chapter 5. The interactions were parameterized using the orientationally-restrained PMFs calculated in Chapter 3, and using assumptions based on the relationship between buried surface area and PMF depth, to guide the interpolation of the interaction potentials. The interpolation was based on an idealized representation of the proteins as having an elliptical cross-section. We then used this representation to calculate the buried surface area as a function of the orientation of the two proteins. The interpolation scheme performed well when compared with a basic interpolation scheme. We performed virtual-move Monte Carlo simulations of the anisotropic model, using the same method employed in Chapter 4, to explore

differences between the anisotropic and isotropic models, and to compare them both to data obtained from the coarse-grained MD simulations performed by Dr. Joseph Goose. The anisotropic model showed better agreement with the coarse-grained MD simulation data for some of the clustering metrics considered, but for metrics applied to larger cluster sizes the comparison was difficult because there were not enough data points from the MD simulation to make a meaningful comparison.

The final test we applied to the model was to attempt to parameterize it using PMFs for the association of rhodopsin from the literature (Periole, Knepp, et al. 2012). As we did not have the raw data from the published PMFs, it was a much more crude parameterization. We also implemented a much simpler method for the interpolation. Using this model of rhodopsin we attempted to simulate a rows-of-dimers system, similar to those seen in disk membranes (Fotiadis et al. 2003; Liang et al. 2003). The rhodopsin system was more closely packed than our NanC system, which invalidated the time-scaling methods used throughout the rest of Chapters 4 and 5, since they relied on an initial period of the simulation during which the proteins were not interacting with each other. The dimers in our system were not stably bound, which was evident by an increasing bond length. Our model does not include explicit representations of the lipids and in a very close-packed system, such as the rows-of-dimers configuration of rhodopsins, the reduced mobility of the lipids will have a large effect on the mobility of the proteins and, therefore, the stability of the dimers. However, the parameterization did result in relatively stable orientational configurations, perhaps suggesting that the orientational stability of the individual proteins in their dimers had a greater dependence on the contributions captured in the PMFs calculated for the sparse system by Periole, Knepp, et al. (2012).

6.3 Summary

In this thesis we have presented results that have demonstrated a method for parameterizing a discrete model of membrane proteins using molecular dynamics. The calculation of the PMFs for NanC highlighted an important relationship between the strength of the interaction and the buried surface area of the proteins when in contact. We also presented results that demonstrated the important role that protein-lipid-protein interactions play in the movement of proteins in the membrane. In order to characterize the interactions of further proteins, it would perhaps be useful to investigate other methodologies that could capture more information in a similar amount of simulation time.

The use of our free energy calculations to parameterize a discrete model was successful in capturing some of the behaviour of the same system when simulated using MD: a much more computationally expensive process. The model was found lacking in some aspects, particularly relating to behaviours that involve extreme behaviours of lipids. These kinds of regimes would be much better represented using a model that included explicit lipids in its representation of the bilayer. Future investigation using this work-flow would be well served by including such lipids in the model that we have developed.

Bibliography

Albizu, L., M. Cottet, M. Kralikova, S. Stoev, R. Seyer, I. Brabet, T. Roux, H. Bazin, E. Bourrier, L. Lamarque, C. Breton, M.-L. Rives, A. Newman, J. Javitch, E. Trinquet, M. Manning, J.-P. Pin, B. Mouillac, and T. Durroux (2010). "Time-resolved FRET between GPCR ligands reveals oligomers in native tissues." In: Nature chemical biology 6.8, pp. 587–594.

Alder, B. J. and T. E. Wainwright (1959). "Studies in molecular dynamics. I. General method". In: Journal of Chemical Physics 31, p. 459.

Ando, T. (2012). "High-speed atomic force microscopy coming of age." In: Nanotechnology 23.6, p. 062001.

Andrews, S. S., N. J. Addy, R. Brent, and A. P. Arkin (2010). "Detailed simulations of cell biology with Smoldyn 2.1." In: PLOS Computational Biology 6.3, e1000705.

Arnarez, C., J.-P. Mazat, J. Elezgaray, S.-J. Marrink, and X. Periole (2013). "Evidence for cardiolipin binding sites on the membrane-exposed surface of the cytochrome bc1." In: Journal of the American Chemical Society 135.8, pp. 3112–3120.

Atallah, M. J. and M. Blanton (2009). "General Concepts and Techniques". Algorithms and Theory of Computation Handbook, Second Edition. CRC Press.

Barducci, A., G. Bussi, and M. Parrinello (2008). "Well-tempered metadynamics: A smoothly converging and tunable free-energy method". In: Physical Review Letters 100.2, p. 020603.

Bartels, C. and M. Karplus (1997). "Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations". In: Journal of Computational Chemistry 18.12, pp. 1450–1462.

Benjamini, A. and B. Smit (2013). "Lipid mediated packing of transmembrane helices-a dissipative particle dynamics study". In: Soft Matter 9.9, pp. 2673–2683.

Berg, H. C. (1993). "Random walks in biology". Princeton, NJ: Princeton University Press.

Binnig, G., C. F. Quate, and C. Gerber (1986). "Atomic force microscope." In: Physical Review Letters 56.9, pp. 930–933.

Blundell, S. J. and K. M. Blundell (2010). "Concepts in Thermal Physics". Oxford University Press.

Bond, P. J., J. Holyoake, A. Ivetac, S. Khalid, and M. S. P. Sansom (2007). "Coarse-grained molecular dynamics simulations of membrane proteins and peptides." In: Journal of Structural Biology 157.3, pp. 593–605.

Bond, P. J. and M. S. P. Sansom (2006). "Insertion and assembly of membrane proteins via simulation." In: Journal of the American Chemical Society 128.8, pp. 2697–2704.

Bond, P. J., C. L. Wee, and M. S. P. Sansom (2008). "Coarse-grained molecular dynamics simulations of the energetics of helix insertion into a lipid bilayer." In: Biochemistry 47.43, pp. 11321–11331.

Boresch, S., F. Tettinger, M. Leitgeb, and M. Karplus (2003). "Absolute Binding Free Energies: A Quantitative Approach for Their Calculation". In: Journal of Physical Chemistry B 107.35, pp. 9535–9551.

Botelho, A. V., T. Huber, T. P. Sakmar, and M. F. Brown (2006). "Curvature and hydrophobic forces drive oligomerization and modulate activity of rhodopsin in membranes." In: Biophysical Journal 91.12, pp. 4464–4477.

Brown, D. A. (2006). "Lipid rafts, detergent-resistant membranes, and raft targeting signals." In: Physiology (Bethesda, Md.) 21.6, pp. 430–439.

Burrage, K., P. M. Burrage, A. Leier, T. T. Marquez-Lago, and D. V. Nicolau Jr (2011). "Stochastic Simulation for Spatial Modelling of Dynamic Processes in a Living Cell". In: Design and Analysis of Biomolecular Circuits. New York: Springer-Verlag.

Capraro, B. R., Z. Shi, T. Wu, Z. Chen, J. M. Dunn, E. Rhoades, and T. Baumgart (2013). "Kinetics of endophilin N-BAR domain dimerization and membrane interactions." In: Journal of Biological Chemistry 288.18, pp. 12533–12543.

Castillo, N., L. Monticelli, J. Barnoud, and D. P. Tieleman (2013). "Free energy of WALP23 dimer association in DMPC, DPPC, and DOPC bilayers." In: Chemistry and Physics of Lipids 169, pp. 95–105.

Casuso, I., J. Khao, M. Chami, P. Paul-Gilloteaux, M. Husain, J.-P. Duneau, H. Stahlberg, J. N. Sturgis, and S. Scheuring (2012). "Characterization of the motion of membrane proteins using high-speed atomic force microscopy." In: Nature Nanotechnology 7.8, pp. 525–529.

Chavent, M., T. Reddy, J. Goose, A. C. E. Dahl, J. E. Stone, B. Jobard, and M. S. P. Sansom (2014). "Methodologies for the analysis of instantaneous lipid diffusion in MD simulations of large membrane systems." In: Faraday Discussions 169, pp. 455–475.

Chng, C.-P. and S.-M. Tan (2011). "Leukocyte integrin αLβ2 transmembrane association dynamics revealed by coarse-grained molecular dynamics simulations." In: Proteins 79.7, pp. 2203–2213.

Choquet, D. and A. Triller (2003). "The role of receptor diffusion in the organization of the postsynaptic membrane." In: Nature reviews. Neuroscience 4.4, pp. 251–265.

Cobbold, C., A. P. Monaco, A. Sivaprasadarao, and S. Ponnambalam (2003). "Aberrant trafficking of transmembrane proteins in human disease". In: Trends in Cell Biology 13.12, pp. 639–647.

Collins, S., M. Stamatakis, and D. G. Vlachos (2010). "Adaptive coarse-grained Monte Carlo simulation of reaction and diffusion dynamics in heterogeneous plasma membranes." In: BMC Bioinformatics 11, p. 218.

Coskun, Ü., M. Grzybek, D. Drechsel, and K. Simons (2011). "Regulation of human EGF receptor by lipids". In: Proceedings of the National Academy of Sciences of the United States of America, pp. 9044–9048.

Crane, J. M. and A. S. Verkman (2008). "Long-range nonanomalous diffusion of quantum dot-labeled aquaporin-1 water channels in the cell plasma membrane." In: Biophysical Journal 94.2, pp. 702–713.

Cristian, L., J. D. Lear, and W. F. DeGrado (2003). "Determination of membrane protein stability via thermodynamic coupling of folding to thiol-disulfide interchange." In: Protein Science 12.8, pp. 1732–1740.

Cui, M., M. Mezei, and R. Osman (2008). "Modeling dimerizations of transmembrane proteins using Brownian dynamics simulations." In: Journal of computer-aided molecular design 22.8, pp. 553–561.

Darve, E. and A. Pohorille (2001). "Calculating free energies using average force". In: Journal of Chemical Physics 115.20, pp. 9169–9183.

Deamer, D., J. P. Dworkin, S. A. Sandford, M. P. Bernstein, and L. J. Allamandola (2002). "The First Cell Membranes". In: Astrobiology 2.4, pp. 371–381.

Doura, A. K. and K. G. Fleming (2004). "Complex interactions at the helix-helix interface stabilize the glycophorin A transmembrane dimer." In: Journal of Molecular Biology 343.5, pp. 1487–1497.

Dunton, T. A., J. E. Goose, D. J. Gavaghan, M. S. P. Sansom, and J. M. Osborne (2014). "The free energy landscape of dimerization of a membrane protein, NanC." In: PLOS Computational Biology 10.1, e1003417–e1003417.

Dupuy, A. D. and D. M. Engelman (2008). "Protein area occupancy at the center of the red blood cell membrane." In: Proceedings of the National Academy of Sciences of the United States of America 105.8, pp. 2848–2852.

Ebbie Tan, A. Z. and K. G. Fleming (2008). "Outer membrane phospholipase A dimer stability does not correlate to occluded surface area". In: Biochemistry 47.46, pp. 12095–12103.

Ebie, A. Z. and K. G. Fleming (2007). "Dimerization of the erythropoietin receptor transmembrane domain in micelles". In: Journal of Molecular Biology 366.2, pp. 517–524.

Efremov, R. G., D. E. Nolde, A. G. Konshina, N. P. Syrtcev, and A. S. Arseniev (2004). "Peptides and proteins in membranes: What can we learn via computer simulations?" In: Current Medicinal Chemistry 11.18, pp. 2421–2442.

Eisenhaber, F., P. Lijnzaad, P. Argos, C. Sander, and M. Scharf (1995). "The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies". In: Journal of Computational Chemistry 16.3, pp. 273–284.

Engelman, D. M. (2005). "Membranes are more mosaic than fluid." In: Nature 438.7068, pp. 578–580.

Fernandez, S. M. and R. D. Berlin (1976). "Cell surface distribution of lectin receptors determined by resonance energy transfer." In: Nature 264.5585, pp. 411–415.

Fiser, A. and A. Sali (2003). "Modeller: generation and refinement of homology-based protein structure models." In: Methods in enzymology 374, pp. 461–491.

Fleming, K. G. and D. M. Engelman (2001). "Specificity in transmembrane helix-helix interactions can define a hierarchy of stability for sequence variants." In: Proceedings of the National Academy of Sciences of the United States of America 98.25, pp. 14340–14344.

Fortuna, S. and A. Troisi (2010). "Agent-based modeling for the 2D molecular self-organization of realistic molecules." In: Journal of Physical Chemistry B 114.31, pp. 10151–10159.

Fotiadis, D., Y. Liang, S. Filipek, D. A. Saperstein, A. Engel, and K. Palczewski (2003). "Atomic-force microscopy: Rhodopsin dimers in native disc membranes." In: Nature 421.6919, pp. 127–128.

Frenkel, D. and B. Smit (2002). "Understanding Molecular Simulation: From Algorithms to Applications". Taylor & Francis.

Frye, L. D. and M. Edidin (1970). "The rapid intermixing of cell surface antigens after formation of mouse-human heterokaryons." In: Journal of cell science 7.2, pp. 319–335.

Gambin, Y., R. Lopez-Esparza, M. Reffay, E. Sierecki, N. S. Gov, M. Genest, R. S. Hodges, and W. Urbach (2006). "Lateral mobility of proteins in liquid membranes revisited." In: Proceedings of the National Academy of Sciences of the United States of America 103.7, pp. 2098–2102.

Gambin, Y., M. Reffay, E. Sierecki, F. Homblé, R. S. Hodges, N. S. Gov, N. Taulier, and W. Urbach (2010). "Variation of the lateral mobility of transmembrane peptides with hydrophobic mismatch." In: Journal of Physical Chemistry B 114.10, pp. 3559–3566.

Gibbs, J. W. (2014). "Elementary Principles in Statistical Mechanics". Courier Dover Publications.

Giraud, M.-F., P. Paumard, C. Sanchez, D. Brèthes, J. Velours, and A. Dautant (2012). "Rotor architecture in the yeast and bovine F1-c-ring complexes of F-ATP synthase." In: Journal of Structural Biology 177.2, pp. 490–497.

Goldman, J., S. S. Andrews, and D. Bray (2004). "Size and composition of membrane protein clusters predicted by Monte Carlo analysis". In: European Biophysics Journal 33.6, pp. 506–512.

Goose, J. E. and M. S. P. Sansom (2013). "Reduced Lateral Mobility of Lipids and Proteins in Crowded Membranes". In: PLOS Computational Biology 9.4, e1003033.

Gorter, E. and F. Grendel (1925). "On Bimolecular Layers Of Lipoids On The Chromocytes Of The Blood." In: The Journal of experimental medicine 41.4, pp. 439–443.

Govardovskii, V. I., D. A. Korenyak, S. A. Shukolyukov, and L. V. Zueva (2009). "Lateral diffusion of rhodopsin in photoreceptor membrane: a reappraisal." In: Molecular vision 15, pp. 1717–1729.

Graham, R. L., B. D. Lubachevsky, K. J. Nurmela, and P. R. J. Östergård (1998). "Dense packings of congruent circles in a circle". In: Discrete Mathematics 181.1-3, pp. 139–154.

Hénin, J. and C. Chipot (2004). "Overcoming free energy barriers using unconstrained molecular dynamics simulations". In: Journal of Chemical Physics 121.7, pp. 2904–2914.

Hénin, J., A. Pohorille, and C. Chipot (2005). "Insights into the recognition and association of transmembrane alpha-helices. The free energy of alpha-helix dimerization in glycophorin A." In: Journal of the American Chemical Society 127.23, pp. 8478–8484.

Hess, B., C. Kutzner, D. van der Spoel, and E. Lindahl (2008). "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation". In: Journal of Chemical Theory and Computation 4.3, pp. 435–447.

Höhr, A. I. C., S. P. Straub, B. Warscheid, T. Becker, and N. Wiedemann (2015). "Assembly of β-barrel proteins in the mitochondrial outer membrane." In: Biochimica et biophysica acta 1853.1, pp. 74–88.

Hooke, R. 1.-1. (1665). "Micrographia: or, Some physiological descriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon." In:

Huang, K. (1987). "Statistical mechanics". New York: Wiley.

Huber, G., H. Wang, and R. Mukhopadhyay (2011). "Protein-coat dynamics and cluster phases in intracellular trafficking". In: Journal of Physics-Condensed Matter 23.37, p. 374105.

Jabbari-Farouji, S. and E. Trizac (2012). "Dynamic Monte Carlo simulations of anisotropic colloids". In: Journal of Chemical Physics 137.5, p. 054107.

Jankowski, E. and S. C. Glotzer (2009). "A comparison of new methods for generating energy-minimizing configurations of patchy particles". In: Journal of Chemical Physics 131.10.

Janosi, L., A. Prakash, and M. Doxastakis (2010). "Lipid-modulated sequence-specific association of glycophorin A in membranes." In: Biophysical Journal 99.1, pp. 284–292.

Jarzynski, C. (1997). "Nonequilibrium Equality for Free Energy Differences". In: Physical Review Letters 78.14, pp. 2690–2693.

Johnston, J. M., M. Aburi, D. Provasi, A. Bortolato, E. Urizar, N. A. Lambert, J. A. Javitch, and M. Filizola (2011). "Making structural sense of dimerization interfaces of delta opioid receptor homodimers." In: Biochemistry 50.10, pp. 1682–1690.

Kasai, R. S., K. G. N. Suzuki, E. R. Prossnitz, I. Koyama-Honda, C. Nakada, T. K. Fujiwara, and A. Kusumi (2011). "Full characterization of GPCR monomer-dimer dynamic equilibrium by single molecule imaging." In: Journal of Cell Biology 192.3, pp. 463–480.

Killian, J. A., I. Salemink, M. R. de Planque, G. Lindblom, R. E. Koeppe, and D. V. Greathouse (1996). "Induction of nonbilayer structures in diacylphosphatidylcholine model membranes by transmembrane alpha-helical peptides: importance of hydrophobic mismatch and proposed role of tryptophans." In: Biochemistry 35.3, pp. 1037–1045.

Koebnik, R., K. P. Locher, and P. van Gelder (2000). "Structure and function of bacterial outer membrane proteins: barrels in a nutshell." In: Molecular Microbiology 37.2, pp. 239–253.

Kok, J. W., K. Klappe, and I. Hummel (2014). "The Role of the Actin Cytoskeleton and Lipid Rafts in the Localization and Function of the ABCC1 Transporter". In: Advances in Biology 2014.2, pp. 1–11.

Kusumi, A., C. Nakada, K. Ritchie, K. Murase, K. Suzuki, H. Murakoshi, R. S. Kasai, J. Kondo, and T. Fujiwara (2005). "Paradigm shift of the plasma membrane concept from the two-dimensional continuum fluid to the partitioned fluid: high-speed single-molecule tracking of membrane molecules." In: Annual Review of Biophysics and Biomolecular Structure 34, pp. 351–378.

Kutzner, C., J. Czub, and H. Grubmüller (2011). "Keep it flexible: driving macromolecular rotary motions in atomistic simulations with GROMACS". In: Journal of Chemical Theory and Computation 7.5, pp. 1381–1393.

Laio, A. and M. Parrinello (2002). "Escaping free-energy minima." In: Proceedings of the National Academy of Sciences of the United States of America 99.20, pp. 12562–12566.

Lasserre, R., X.-J. Guo, F. Conchonaud, Y. Hamon, O. Hawchar, A.-M. Bernard, S. M. Soudja, P.-F. Lenne, H. Rigneault, D. Olive, G. Bismuth, J. A. Nunès, B. Payrastre, D. Marguet, and H.-T. He (2008). "Raft nanodomains contribute to

Akt/PKB plasma membrane recruitment and activation." In: Nature chemical biology 4.9, pp. 538–547.

Lenne, P. F., L. Wawrezinieck, F. Conchonaud, O. Wurtz, A. Boned, X. J. Guo, H. Rigneault, H. T. He, and D. Marguet (2006). "Dynamic molecular confinement in the plasma membrane by microdomains and the cytoskeleton meshwork". In: The EMBO journal 25.14, pp. 3245–3256.

Liang, Y., D. Fotiadis, S. Filipek, D. A. Saperstein, K. Palczewski, and A. Engel (2003). "Organization of the G protein-coupled receptors rhodopsin and opsin in native membranes." In: Journal of Biological Chemistry 278.24, pp. 21655–21662.

Lingwood, D. and K. Simons (2009). "Lipid Rafts As a Membrane-Organizing Principle". In: Science 327.5961, pp. 46–50.

Liu, J. and E. Luijten (2004). "Rejection-free geometric cluster algorithm for complex fluids." In: Physical Review Letters 92.3, p. 035504.

Long, S. B., E. B. Campbell, and R. Mackinnon (2005). "Crystal structure of a mammalian voltage-dependent Shaker family K+ channel." In: Science 309.5736, pp. 897–903.

Lowder, M. A., J. S. Appelbaum, E. M. Hobert, and A. Schepartz (2011). "Visualizing protein partnerships in living cells and organisms." In: Current Opinion in Chemical Biology 15.6, pp. 781–788.

Low-Nam, S. T., K. A. Lidke, P. J. Cutler, R. C. Roovers, P. M. P. van Bergen en Henegouwen, B. S. Wilson, and D. S. Lidke (2011). "ErbB1 dimerization is promoted by domain co-confinement and stabilized by ligand binding." In: Nature structural & molecular biology 18.11, pp. 1244–1249.

Lugtenberg, E. J. and R. Peters (1976). "Distribution of lipids in cytoplasmic and outer membranes of Escherichia coli K12." In: Biochimica et biophysica acta 441.1, pp. 38–47.

Lyubartsev, A. P. and A. L. Rabinovich (2011). "Recent development in computer simulations of lipid bilayers". In: Soft Matter 7.1, pp. 25–39.

MacKenzie, K. R. and K. G. Fleming (2008). "Association energetics of membrane spanning alpha-helices." In: Current Opinion in Structural Biology 18.4, pp. 412–419.

Marrink, S.-J., H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries (2007). "The MARTINI force field: coarse grained model for biomolecular simulations." In: Journal of Physical Chemistry B 111.27, pp. 7812–7824.

Marrink, S.-J. and D. P. Tieleman (2013). "Perspective on the Martini model." In: Chemical Society reviews 42.16, pp. 6801–6822.

Maurel, D., L. Comps-Agrar, C. Brock, M.-L. Rives, E. Bourrier, M. A. Ayoub, H. Bazin, N. Tinel, T. Durroux, L. Prézeau, E. Trinquet, and J.-P. Pin (2008).

"Cell-surface protein-protein interaction analysis with time-resolved FRET and snap-tag technologies: application to GPCR oligomerization." In: Nature Methods 5.6, pp. 561–567.

May, A., R. Pool, E. van Dijk, J. Bijlard, S. Abeln, J. Heringa, and K. A. Feenstra (2013). "Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins." In: Bioinformatics (Oxford, England).

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). "Equation of State Calculations by Fast Computing Machines". In: Journal of Chemical Physics 21.6, pp. 1087–1092.

Mileykovskaya, E. and W. Dowhan (2000). "Visualization of phospholipid domains in Escherichia coli by using the cardiolipin-specific fluorescent dye 10-N-nonyl acridine orange." In: Journal of bacteriology 182.4, pp. 1172–1175.

– (2005). "Role of membrane lipids in bacterial division-site selection". In: Current Opinion in Microbiology 8.2, pp. 135–142.

Mitra, K., I. Ubarretxena-Belandia, T. Taguchi, G. Warren, and D. M. Engelman (2004). "Modulation of the bilayer thickness of exocytic pathway membranes by membrane proteins rather than cholesterol". In: Proceedings of the National Academy of Sciences 101.12, pp. 4083–4088.

Monticelli, L., S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink (2008). "The MARTINI coarse-grained force field: extension to proteins". In: Journal of Chemical Theory and Computation 4.5, pp. 819–834.

Ngounou Wetie, A. G., I. Sokolowska, A. G. Woods, U. Roy, J. A. Loo, and C. C. Darie (2013). "Investigation of stable and transient protein-protein interactions: Past, present, and future." In: Proteomics 13.3-4, pp. 538–557.

Nicolau Jr, D. V., K. Burrage, R. G. Parton, and J. F. Hancock (2006). "Identifying optimal lipid raft characteristics required to promote nanoscale protein-protein interactions on the plasma membrane." In: Molecular and Cellular Biology 26.1, pp. 313–323.

Nicolau Jr, D. V., J. F. Hancock, and K. Burrage (2007). "Sources of anomalous diffusion on cell membranes: a Monte Carlo study." In: Biophysical Journal 92.6, pp. 1975–1987.

Nicolson, G. L. (2014). "The Fluid - Mosaic Model of Membrane Structure: Still relevant to understanding the structure, function and dynamics of biological membranes after more than 40 years". In: Biochimica et biophysica acta 1838.6, pp. 1451–1466.

Niemelä, P. S., M. S. Miettinen, L. Monticelli, H. Hammaren, P. Bjelkmar, T. Murtola, E. Lindahl, and I. Vattulainen (2010). "Membrane proteins diffuse as dynamic complexes with lipids." In: Journal of the American Chemical Society 132.22, pp. 7574–7575.

Nikaido, H. (2001). "Preventing drug access to targets: cell surface permeability barriers and active efflux in bacteria." In: Seminars in Cell & Developmental Biology 12.3, pp. 215–223.

Nilsson, J., B. Persson, and G. von Heijne (2005). "Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes." In: Proteins 60.4, pp. 606–616.

Overton, C. E. (1895). "Ueber die osmotischen eigenschaften der lebenden pflanzen- und tierzelle," Zurich: Fäsi & Beer.

Owen, D. M., D. Williamson, C. Rentero, and K. Gaus (2009). "Quantitative microscopy: protein dynamics and membrane organisation." In: Traffic 10.8, pp. 962–971.

Parton, D. L., J. W. Klingelhoefer, and M. S. P. Sansom (2011). "Aggregation of model membrane proteins, modulated by hydrophobic mismatch, membrane curvature, and protein class." In: Biophysical Journal 101.3, pp. 691–699.

Periole, X., T. Huber, S.-J. Marrink, and T. P. Sakmar (2007). "G protein-coupled receptors self-assemble in dynamics simulations of model bilayers." In: Journal of the American Chemical Society 129.33, pp. 10126–10132.

Periole, X., A. M. Knepp, T. P. Sakmar, S.-J. Marrink, and T. Huber (2012). "Structural determinants of the supramolecular organization of G protein-coupled receptors in bilayers." In: Journal of the American Chemical Society 134.26, pp. 10959–10965.

Polson, J. M., I. Vattulainen, H. Zhu, and M. J. Zuckermann (2001). "Simulation study of lateral diffusion in lipid-sterol bilayer mixtures". In: The European physical journal. E, Soft matter 5.4, pp. 485–497.

Prakash, A., L. Janosi, and M. Doxastakis (2011). "GxxxG motifs, phenylalanine, and cholesterol guide the self-association of transmembrane domains of ErbB2 receptors." In: Biophysical Journal 101.8, pp. 1949–1958.

Raetz, C. R. H. (1978). "Enzymology, genetics, and regulation of membrane phospholipid synthesis in Escherichia coli". In: Microbiological Reviews 42.3, pp. 614–659.

Ramadurai, S., A. Holt, V. V. Krasnikov, G. van den Bogaart, J. A. Killian, and B. Poolman (2009). "Lateral diffusion of membrane proteins." In: Journal of the American Chemical Society 131.35, pp. 12650–12656.

Rendón, A., D. G. Carton, J. Sot, M. García-Pacios, L.-R. Montes, M. Valle, J.-L. R. Arrondo, F. M. Goñi, and K. Ruiz-Mirazo (2012). "Model systems of precursor cellular membranes: long-chain alcohols stabilize spontaneously formed oleic acid vesicles." In: Biophysical Journal 102.2, pp. 278–286.

Robertson, J. D. (1972). "Discussion paper: experimental basis of the unit membrane theory." In: Annals of the New York Academy of Sciences 195, pp. 356–365.

Roduit, C., F. G. van der Goot, P. De Los Rios, A. Yersin, P. Steiner, G. Dietler, S. Catsicas, F. Lafont, and S. Kasas (2008). "Elastic membrane heterogeneity of living cells revealed by stiff nanoscale membrane domains." In: Biophysical Journal 94.4, pp. 1521–1532.

Romano, F., C. De Michele, D. Marenduzzo, and E. Sanz (2011). "Monte Carlo and event-driven dynamics of Brownian particles with orientational degrees of freedom." In: Journal of Chemical Physics 135.12, p. 124106.

Roux, B. (1995). "The calculation of the potential of mean force using computer simulations". In: Computer Physics Communications 91.1-3, pp. 275–282.

Ruiz-Mirazo, K., C. Briones, and A. De La Escosura (2014). "Prebiotic systems chemistry: New perspectives for the origins of life". In: Chemical Reviews 114.1, pp. 285–366.

Russ, W. P. and D. M. Engelman (1999). "TOXCAT: a measure of transmembrane helix association in a biological membrane." In: Proceedings of the National Academy of Sciences of the United States of America 96.3, pp. 863–868.

Saffman, P. G. and M. Delbrück (1975). "Brownian motion in biological membranes." In: Proceedings of the National Academy of Sciences of the United States of America 72.8, pp. 3111–3113.

Sanz, E., E. Sanz, D. Marenduzzo, and D. Marenduzzo (2010). "Dynamic Monte Carlo versus Brownian dynamics: A comparison for self-diffusion and crystallization in colloidal fluids." In: Journal of Chemical Physics 132.19, p. 194102.

Schöneberg, J. and F. Noé (2013). "ReaDDy–a software for particle-based reaction-diffusion dynamics in crowded cellular environments." In: PLOS One 8.9, e74261.

Schulze, R. J., J. Komar, M. Botte, W. J. Allen, S. Whitehouse, V. A. M. Gold, J. A. L. A. Nijeholtb, K. Huard, I. Berger, C. Schaffitzel, and I. Collinson (2014). "Membrane protein insertion and proton-motive-force-dependent secretion through the bacterial holo-translocon SecYEG-SecDF-YajC-YidC". In: Proceedings of the National Academy of Sciences of the United States of America 111.13, pp. 4844–4849.

Schütz, G. J., H. Schindler, and T. Schmidt (1997). "Single-molecule microscopy on model membranes reveals anomalous diffusion." In: Biophysical Journal 73.2, pp. 1073–1080.

Sengupta, D. and S.-J. Marrink (2010). "Lipid-mediated interactions tune the association of glycophorin A helix and its disruptive mutants in membranes." In: Physical Chemistry Chemical Physics 12.40, pp. 12987–12996.

Sengupta, P., R. R. P. Singh, D. L. Cox, and A. Slepoy (2004). "Lateral organization of cholesterol molecules in lipid-cholesterol assemblies". In: Physical Review E 70.2 Pt 1, p. 021902.

Sergeev, M., A. G. Godin, L. Kao, N. Abuladze, P. W. Wiseman, and I. Kurtz (2012). "Determination of membrane protein transporter oligomerization in native tissue using spatial fluorescence intensity fluctuation analysis." In: PLOS One 7.4, e36215.

Sieber, J. J., K. I. Willig, C. Kutzner, C. Gerding-Reimers, B. Harke, G. Donnert, B. Rammner, C. Eggeling, S. W. Hell, H. Grubmüller, and T. Lang (2007). "Anatomy and dynamics of a supramolecular membrane protein cluster." In: Science 317.5841, pp. 1072–1076.

Simons, K. and E. Ikonen (1997). "Functional rafts in cell membranes." In: Nature 387.6633, pp. 569–572.

Singer, S. J. and G. L. Nicolson (1972). "The Fluid Mosaic Model of the Structure of Cell Membranes". In: Science 175.4023, pp. 720–731.

Spector, J., S. Zakharov, Y. Lill, O. Sharma, W. A. Cramer, and K. Ritchie (2010). "Mobility of BtuB and OmpF in the Escherichia coli outer membrane: implications for dynamic formation of a translocon complex." In: Biophysical Journal 99.12, pp. 3880–3886.

Spillane, K. M., J. Ortega-Arroyo, G. de Wit, C. Eggeling, H. Ewers, M. I. Wallace, and P. Kukura (2014). "High-speed single-particle tracking of GM1 in model membranes reveals anomalous diffusion due to interleaflet coupling and molecular pinning." In: Nano letters 14.9, pp. 5390–5397.

Spoel, D. van der, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen (2005). "GROMACS: Fast, flexible, and free". In: Journal of Computational Chemistry 26.16, pp. 1701–1718.

Stanley, A. M., P. Chuawong, T. L. Hendrickson, and K. G. Fleming (2006). "Energetics of outer membrane phospholipase A (OMPLA) dimerization." In: Journal of Molecular Biology 358.1, pp. 120–131.

Stansfeld, P. J. and M. S. P. Sansom (2011). "From coarse grained to atomistic: a serial multiscale approach to membrane protein simulations". In: Journal of Chemical Theory and Computation 7.4, pp. 1157–1166.

Stein, E. G., L. M. Rice, and A. T. Brünger (1997). "Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation." In: Journal of magnetic resonance (San Diego, Calif. : 1997) 124.1, pp. 154–164.

Suh, B.-C. and B. Hille (2008). "PIP2 is a necessary cofactor for ion channel function: how and why?" In: Annual Review of Biophysics 37, pp. 175–195.

Suzuki, K. and M. P. Sheetz (2001). "Binding of cross-linked glycosylphosphatidylinositol-anchored proteins to discrete actin-associated sites and cholesterol-dependent domains." In: Biophysical Journal 81.4, pp. 2181–2189.

Suzuki, K. G. N., R. S. Kasai, K. M. Hirosawa, Y. L. Nemoto, M. Ishibashi, Y. Miwa, T. K. Fujiwara, and A. Kusumi (2012). "Transient GPI-anchored protein homodimers are units for raft organization and function." In: Nature chemical biology 8.9, pp. 774–783.

Torrie, G. M. and J. P. Valleau (1977). "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". In: Journal of Computational Physics 23.2, pp. 187–199.

Troisi, A., V. Wong, and M. A. Ratner (2005a). "An agent-based approach for modeling molecular self-organization." In: Proceedings of the National Academy of Sciences of the United States of America 102.2, pp. 255–260.

– (2005b). "Self-assembly on multiple length scales: a Monte Carlo algorithm with data augmentation." In: Journal of Chemical Physics 122.2, p. 024102.

Villar, G., A. W. Wilber, A. J. Williamson, P. Thiara, J. P. K. Doye, A. A. Louis, M. N. Jochum, A. C. F. Lewis, and E. D. Levy (2009). "Self-assembly and evolution of homomeric protein complexes." In: Physical Review Letters 102.11, p. 118106.

Wagner, A., S. Loew, and S. May (2007). "Influence of monolayer-monolayer coupling on the phase behavior of a fluid lipid bilayer". In: Biophysical Journal 93.12, pp. 4268–4277.

Wang, H., N. S. Wingreen, and R. Mukhopadhyay (2008). "Self-organized periodicity of protein clusters in growing bacteria." In: Physical Review Letters 101.21, p. 218101.

Weiss, K., A. Neef, Q. Van, S. Kramer, I. Gregor, and J. Enderlein (2013). "Quantifying the diffusion of membrane proteins and peptides in black lipid membranes with 2-focus fluorescence correlation spectroscopy." In: Biophysical Journal 105.2, pp. 455–462.

Weiss, M., H. Hashimoto, and T. Nilsson (2003). "Anomalous protein diffusion in living cells as seen by fluorescence correlation spectroscopy." In: Biophysical Journal 84.6, pp. 4043–4052.

Whitelam, S. (2011). "Approximating the dynamical evolution of systems of strongly interacting overdamped particles". In: Molecular Simulation 37.7, pp. 606–612.

Whitelam, S. and P. L. Geissler (2007). "Avoiding unphysical kinetic traps in Monte Carlo simulations of strongly attractive particles." In: Journal of Chemical Physics 127.15, p. 154101.

Wirth, C., G. Condemine, C. Boiteux, S. Bernèche, T. Schirmer, and C. M. Peneff (2009). "NanC crystal structure, a model for outer-membrane channels of the acidic sugar-specific KdgM porin family." In: Journal of Molecular Biology 394.4, pp. 718–731.

Wolff, U. (1989). "Collective Monte Carlo updating for spin systems." In: Physical Review Letters 62.4, pp. 361–364.

Yiannourakou, M., L. Marsella, F. J.-M. de Meyer, and B. Smit (2010). "Towards an understanding of membrane-mediated protein-protein interactions." In: Faraday Discussions 144, pp. 359–367.

Appendices

Appendix A

Simulating a Bilayer Using Molecular Dynamics

The sections in this appendix explain the three main stages in preparing the system for umbrella sampling simulations. These steps were performed using tools provided by the *GROMACS* molecular dynamics simulation software package.The steps taken to build the lipid bilayer, insert the proteins into this preformed bilayer, and then generate the configurations used for the umbrella sampling simulation windows are all described in this appendix.

A.i Building a bilayer

The model bilayer we are using is a pure POPE bilayer with periodic boundary conditions. The bilayer is solvated and Na^+ counter-ions were added to maintain system neutrality. We use POPE as it is the major lipid constituent in the bacterial outer membrane (Lugtenberg et al. 1976; Raetz 1978). Using a more realistic membrane composition, with a mixture of lipids and lipopolysaccharides in an asymmetric bilayer, would increase the volume of the state space of the bilayer system, and would thus require a corresponding increase in the simulation time required to achieve convergence. Using a more realistic membrane would be a natural extension to the work presented here.

Our coarse-grained bilayer model consists of a rectangular box, with periodic boundaries in the x, y, and z directions. We arrange the bilayer to lie in the x-y plane. Before the proteins are embedded in the bilayer, we must firstly create a section of bilayer that is equilibrated. In order to ensure that the tension in both leaflets remains approximately equal, we require that there are the same number of lipids in each. To build our bilayer system we construct and array of lipids using the *GROMACS* molecular dynamics simulation package and the various tools included with it (Spoel et al. 2005; Hess et al. 2008). By manually arranging two layers of lipids in a tightly packed grid, we can create an approximate bilayer structure using `editconf`. This bilayer structure is then solvated using `genbox` and neutralized, where appropriate, using `genion`. All of these tools are included in *GROMACS*. Our bilayer system can now be energy minimized, to relax any poor contacts created during the construction phase, and then allowed to correctly self assemble with a full molecular dynamics simulation. Once the membrane was equilibrated, as judged by the convergence of fluctuations in the simulation box size and in the bilayer itself, we were able to insert the proteins into the membrane.

A.ii Embedding the proteins

We chose to embed the proteins in a pre-formed membrane, rather than letting the whole system self-assemble from a mixed state, as this enabled us to ensure that we had equal sized membrane leaflets throughout the entirety of the construction phase. To embed proteins in a preformed bilayer we can use the *GROMACS* tool `g_membed`, which is specifically designed for the insertion of proteins into a bilayer. `g_membed` inserts the proteins by artificially changing their shape so that they are compressed in the plane of the membrane and extended perpendicular to it. This stretched protein is then inserted into the system containing the membrane, and any lipid, solvent or ion molecules that overlap with the protein are removed. The protein is then slowly returned to its correct shape over the course of a simulation; we chose to do this over a 1 μs simulation, which was slow enough that the membrane responded to the growing proteins in a smooth manner. Following the embedding, we further equilibrated the system, with harmonic positions restraints, in the x and y direction, applied to the C_α CG particles of the protein. This is necessary because the protein embedding process will leave the bilayer in a disrupted state.

A.iii Obtaining the umbrella sampling initial configurations

To obtain initial configurations for the umbrella sampling simulations, we used the pull code in *GROMACS*. This feature allows you to apply a force to a group of particles throughout a simulation. That force can either be a constant force in a specified direction, a harmonic force centred at a specific position, or one of several other forms, none of which we use in this thesis.

To generate the starting configurations, the proteins, which are embedded in the bilayer at a reasonable separation, are slowly pulled into contact using a harmonic

force whose centre moves such that the proteins are slowly pulled together. From an initial separation of 5nm, the NanC proteins are pulled into contact over the course of a 2 μs simulation; pulling the proteins slowly ensures that the system remains close to equilibrium throughout the simulation. Once the proteins are in contact, the system is equilibrated for 2 μs with a static potential restraining the proteins in their position of contact. Next the reverse of the initial pulling procedure is performed: the proteins are slowly pulled apart, during a 5 μs simulation, after which they reach 8 nm separation, the maximum separation at which we perform umbrella sampling.

This final pulling simulation is analysed, and the system configurations that correspond to an inter-protein separation with a value equal to one of our desired window positions (either 0.1 nm or 0.05 nm intervals along the reaction coordinate) are used as the starting configurations for our window simulations. However, to remove any residual effect that the pulling procedure had on the bilayer, each of these configurations is equilibrated for 2 μs.