Multiscale Simulations and Markov State Models of Peripheral Membrane Proteins



Sarah-Beth T. A. Amos

Department of Biochemistry

University of Oxford

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

Wadham College

Michaelmas 2019

The *viva voce* examination was held at the Department of Biochemistry, University of Oxford on 5th November 2019. The examiners were Prof. Elena Seiradake (University of Oxford) and Prof. Franca Fraternali (King's College, London). A copy of this thesis has been deposited in the Bodleian Library, Oxford.

Acknowledgements

"If I have seen further, it is by standing on the shoulders of Giants."

Isaac Newton

Firstly I must thank Prof Mark Sansom for supervising me with such enthusiasm and humour. It has been a great privilege to work for him. I have learned so much from him and I hope one day I can hold as much information in my brain as he does.

I would like to extend particular thanks to colleagues at UCB, notably Dr Jiye Shi for his supervision and guidance. I would particularly like to thank Dr Richard Taylor for his help and interest in my work. Thanks also to Dr Silvia Lovera, for going out of her way to offer her expertise when I needed help. I must thank the BBSRC and UCB for funding this project through an iCASE studentship, and Wadham College, Oxford for travel funding.

I am grateful to many colleagues in the lab and must offer my thanks firstly to Dr Antreas Kalli, who taught me a number of methods used in this thesis. I am tremendously grateful to: Dr Anna Duncan, who is not only the chicest academic I know, but also one of the kindest and most insightful; Dr Robin Corey, for always being willing to debate more nebulous aspects of my project and for many pep talks on bad days;

Dr Phillip Stansfeld, for always offering sage advice; Dr Keith Cassidy, for offering his thoughts and jokes; Dr Charlotte Lynch, for always being willing to talk about maths and music; Gianni Klesse, for never failing to complain and commiserate with me and for his generous friendship (thanks for all the chats about wine and Bayesian statistics); and Victor Everett, for unconditional belief and support, which means so much to me.

Oxford is a transient place but I hope that the friendships I have made here will last a lifetime. Words are not enough to fully describe how humbled I am to have met such wonderful people. I am truly grateful to Annabella Massey for her never-ending wisdom and delightful company, and to Gareth Molyneux for his generosity, warmth, and support. Dr Colleen Curran has always been willing to share a glass of Rhône with me, on bad days and good, and I hope we will continue this tradition until we are very old. Particular thanks go to Kate Sim (for coffee), Dan Harkin (for cake), and Cian O'Concubhair (for dessert toast) for their wonderful friendship and support. I have also been fortunate to meet so many wonderful people through MCR PresCom (shout out to Dr Matthew Naiman, Dr Joshua Bull, and Dr Fergus Cooper for all the encouragement from the other side of submission), Blind Tasting Society, and Wadham MCR; they are too numerous to mention here, but I hope that they will appreciate a more general yet heartfelt thanks. Friends further afield include Dr Justyna Kozłowska (you're the best), Dr Jamie Macpherson (for always believing in me and picking up the phone to call me when I needed it most), Angela and Phil (for being there through all highs and lows), Rebecca and Will (for their lovely company and hotel services), and Dr Jan Domański (for being #1 cheerleader). I would like to extend my thanks to my previous supervisors, Dr James Mason and Dr Chris Lorenz, for supporting me always and for putting me on this path. To all of you, I can never

express the full extent of my gratitude.

To Ben, for his love and for bringing light into my darkest hours. To my sister Elle, for her eternal wisdom, support, and unconditional love. Finally, to my parents, Rosemary and Jonathan, for their love and support without which this endeavour would not have been possible. It is the greatest privilege as a daughter to have you as my parents, and I hope I have made you proud.

This thesis is dedicated to my family. - SB Amos 2019

Publications

Publications originating from this thesis:

- Membrane Recognition and Binding by the Phosphatidylinositol Phosphate Kinase PIP5K1A: A Multiscale Simulation Study. **Amos, SBTA**; Kalli, AC; Shi, J; Sansom, MSP. *Structure*. 2019. DOI: 10.1016/j.str.2019.05.004.
- Computational and NMR studies of α-synuclein reveal membrane binding characteristics. **Amos, SBTA**; Schwarz, T; Cossins, BP; Shi, J; Baker, TS; Taylor, RC; Konrat, R; Sansom, MSP. *In preparation.*

Prizes received as a result of DPhil work:

- 2017 Peter Beaconsfield Prize in Medical Sciences
- 2017 SCI Young Lipid Scientist Award
- 2019 Women in Software Power List

Abstract

Peripheral membrane proteins (PMPs) interact with the cell membrane. They are important for many cell processes and are important therapeutic targets. This thesis seeks to investigate the membrane recognition and binding processes that are integral to the function of two PMP systems. Phosphatidylinositol phosphates (PIPs) are lipid signaling molecules that play key roles in many cellular processes. The PIP5K1A kinase catalyses phosphorylation of PI4P to form PIP_2, which in turn interacts with membrane and membrane-associated proteins. We present a multiscale simulation study of the PIP5K1A kinase and its interaction with PIPs in the membrane. The study showed that the PIP5K1A kinase interacts with PIPs via its activation loop, leading to reorientation and binding. It also showed that the dimeric PIP5K1A binds with PIPs via only one catalytic site at a time, which has important implications for the role of membrane curvature and composition in PMP binding. This is addressed by large-scale simulations of PIP5K1A signalling. Furthermore, Markov state models (MSMs) were applied to the activation loop. This work showed that the loop remains disordered in the membrane but adopts specific conformations that may help explain its specificity. These methods were extended to another PMP system of interest, α-synuclein. The results showed that interaction with the membrane is led by the inter-helical region, and offers a structural perspective of a break in the helical structure, which is supported by experimental studies. This may inform our understanding of α-synuclein aggregation, and more generally aid in our understanding of PMP signalling.

List of Abbreviations

AT	Atomistic
CD	Circular Dichroism
CG	Coarse-grained
COM	Centre Of Mass
DAG	Diacylglycerol
DHPC	Diheptanoylphosphocholine
DOPC	Dioleoylphosphocholine
DOPE	Dioleoylphosphoethanolamine
DOPS	Dioleoylphosphoserine
DSSP	Define Secondary Structure of Proteins (algorithm)
IP_3	Inositol Trisphosphate
MD	Molecular Dynamics
MSM	Markov State Model
NAC	Non-amyloid Component
NMR	Nuclear Magnetic Resonance
PBC	Periodic Boundary Conditions
PCA	Principal Components Analysis
PI4P	Phosphatidylinositol-4-phosphate
PIP	Phosphatidylinositol phosphate
PIP_2	Phosphatidylinositol 4,5-bisphosphate
PIP_3	Phosphatidylinositol 3,4,5-trisphosphate
PIPK	Phosphatidylinositol Phosphate Kinase

POPC	Palmitoyloleoyl Phosphatidylcholine
POPG	Palmitoyloleoyl Phosphatidylglycerol
POPS	Palmitoyloleoyl Phosphatidylserine
SSP	Secondary Structure Propensity
tICA	Time-lagged Independent Component Analysis

Table of contents

Abstract **vi**

List of Abbreviations **vii**

Table of contents **ix**

List of figures **xiii**

List of tables **xvi**

1 Introduction **1**

1.1 The cell membrane 1

1.1.1 Lipids 3

1.1.2 Peripheral Membrane Proteins 6

1.1.3 The Role of Disordered Regions 10

1.1.4 Protein-lipid Interactions 12

1.1.5 Previous Computational Studies of PMPs 15

1.2 Markov State Models 19

1.2.1 Estimation and Validation of Markov State Models 20

1.2.2 Software Advances & Application to Protein Simulation 22

1.3 Thesis Organisation 23

2 Theory and Methods **25**

2.1 Classical Molecular Dynamics Simulations 25

2.1.1 Integrating the Equations of Motion 27

2.1.2 Computing the Forces 30

2.1.3 Force Fields 33

2.1.4 Periodic Boundary Conditions 35

2.1.5 Treatment of Electrostatic Interactions 36

2.1.6 Temperature and Pressure Coupling 40

2.1.7 Energy Minimisation 42

2.1.8 Coarse-grained Molecular Dynamics Simulations 42

2.1.9 Tools for Analysing Molecular Dynamics Simulations 46

2.1.10 Limitations of Molecular Simulations 49

2.2 Markov State Models 50

2.2.1 Theory of Markov State Models 50

2.2.2 MSM workflow 53

2.3 Machine Learning Techniques 55

2.3.1 Principal Component Analysis 56

2.3.2 Clustering 59

3 Membrane Recognition & binding by PIP5K1A **63**

3.1 Introduction 63

3.2 Methods 67

3.3 Results & Discussion 71

3.3.1 Binding of PIP5K1A to a Substrate-Containing Lipid Bilayer . . 71

3.3.2 Dimer Binding 80

3.3.3 The bound dimer 86

3.4 Conclusions 93

4 Markov State Models of the PIP5K1A Activation Loop 98

4.1 Introduction . . . 98

4.2 Methods . . . 101

4.3 Results & Discussion . . . 103

4.3.1 The activation loop in the membrane . . . 103

4.3.2 Lipid contacts show the activation loop leads binding and makes diverse lipid contacts . . . 107

4.3.3 tICA shows the loop switches between two slowly equilibrating populations . . . 111

4.3.4 Markov state models of the activation loop . . . 117

4.4 Conclusion . . . 124

5 Computational Studies of α-synuclein 126

5.1 Introduction . . . 126

5.2 Methods . . . 130

5.3 Results & Discussion . . . 132

5.3.1 CG simulations show that the initial binding sequence is led by the inter-helical region . . . 132

5.3.2 CG simulations in mixed lipid bilayers indicates time-resolved lipid preference . . . 137

5.3.3 Atomistic simulations show great variation in the initial binding sequence and corroborate experimental data . . . 141

5.3.4 Markov state models provide a view for understanding the initial binding of a-synuclein . . . 147

5.4 Conclusion . . . 156

6 Large-scale Simulations of PIP5K1A in Mixed Lipid Membranes 157

6.1 Introduction 157
6.2 Methods 161
6.3 Results & Discussion 162
6.3.1 Contact analysis identifies dimerisation sites in line with experimental data 163
6.3.2 The dimerisation interface is independent of lipid composition . 167
6.3.3 Cluster analysis shows time-progression of protein dimerisation and oligomerisation 171
6.3.4 Cluster and diffusion analysis show lipids cluster in simulated signalling events 176
6.3.5 HDBSCAN shows lipid clusters forming over time 179
6.3.6 Simulations of dimeric PIP5K1A with a complex membrane indicate a binding preference for curved regions 180
6.4 Conclusions and Future Directions 184

7 Conclusions & Future Directions **186**

7.1 Membrane recognition by peripheral membrane proteins 187
7.2 Markov state models of the PIP5K1A activation loop 188
7.3 Insights into the structure and function of α-synuclein 189
7.4 Simulation of cell signalling events 191
7.5 The frontiers of biomolecular simulation 192

References **193**

List of figures

1.1 The lipid bilayer with integral and peripheral membrane proteins. . . . 3

1.2 Interconversion of PIPs. 5

1.3 PMP domains and their lipid substrates. 7

1.4 PIP kinase signalling. 8

1.5 The transition matrix and the Markov state model. 20

2.1 Flowchart summarising the MD algorithm. 30

2.2 Schematic representation of periodic boundary conditions in 2D. 35

2.3 Martini model CG representations of amino acid residues. 45

2.4 Visual outline of the construction of Markov state models from MD simulation trajectories. 52

2.5 Visual outline of the MSM workflow. 54

2.6 Vectors indicating the first 2 principal components of 2D dataset. . . . 57

2.7 Comparison of clustering algorithms. 60

3.1 PIP5K1A Protein, Substrate and Product. 66

3.2 PIP5K1A/Membrane Encounter Simulations. 72

3.3 PIP5K1A/Membrane encounter controls. 74

3.4 Membrane Contacts of PIP5K1A 76

3.5 Orientation density maps. 78

3.6 PIP5K1A dimer-PI4P contact profile. 81
3.7 PIP5K1A dimer at a membrane surface. 83
3.8 PIP5K1A Dimer Interactions with PI4P at a Membrane Surface. 85
3.9 Atomistic Simulations of the membrane bound-PIP5K1A Dimer. 87
3.10 PIPs around the bound dimer. 88
3.11 Bound dimer controls. 90
3.12 Hydrogen bonding and RMSF in the PIP5K1A dimer system. 91
3.13 Final Configuration of the Membrane-Bound Dimer. 94

4.1 Structure of the PIP5K1A kinase monomer. 99
4.2 PIP5K1A-PI4P contacts. 104
4.3 PIP5K1A-PI4P contacts. 105
4.4 Penetration of activation loop residues into the membrane. 106
4.5 Graphical representation of the three starting structures for simulation. 107
4.6 PI4P-residue contacts over time for PIP5K1A. 108
4.7 Initial binding RMSD and secondary structure calculations. 110
4.8 Feature selection. 112
4.9 Dimensionality reduction. 114
4.10 All trajectories in tICA space (grey landscape) over time. 116
4.11 MSM of the PIP5K1A activation loop. 118
4.12 2D projection of tICs 1 and 2 with representative structures of the loop. 120
4.13 Graphical representation of an example folding trajectory in PCA space viewed from front and back. 122

5.1 The sequence and structure of α-synuclein. 128
5.2 Coarse grained simulations of α-synuclein with a POPG membrane. . . 133
5.3 Coarse grained simulations of α-synuclein with a POPG membrane identify important residue contacts. 135

5.4 CG simulations of α-synuclein with a POPG membrane with varying degrees of helicity. . . . 136
5.5 CG simulations of the interaction of α-synuclein with a DOPC/DOPE/DOPS 2:5:3 membrane. . . . 138
5.6 CG simulations of α-synuclein with a mixed DOP membrane. . . . 140
5.7 Atomistic simulations of α-synuclein with a POPG membrane. . . . 142
5.8 Secondary structure analysis of the α-synuclein monomer. . . . 144
5.9 Secondary structure analysis of the -synuclein monomer binding to a POPG membrane. . . . 146
5.10 Residue-residue distances of the α-synuclein monomer binding to a POPG membrane. . . . 147
5.11 tICA analysis of all atomistic simulations based on dihedral featurisation. 148
5.12 Markov state model of residues 60-70. . . . 149
5.13 Model of the two structures derived from the MSM. . . . 151
5.14 Residues of the MSM mapped onto PDB structures 1XQ8 and 6FLT. . 153

6.1 Simulation setup and protein clustering. . . . 164
6.2 Protein-protein interactions for dimers in one of the PI4P only replicates. 166
6.3 Protein contacts made by each residue across all replicates. . . . 168
6.4 Comparison between simulation and structural data of dimeric interfaces. 170
6.5 Protein clustering over all replicates in 2D. . . . 173
6.6 Protein clustering over all replicates in 1D. . . . 175
6.7 2D density analysis of PIPs in the membrane. . . . 176
6.8 Lipid clustering and diffusion in mixed PIP simulations. . . . 177
6.9 Lipid clustering over the course of one mixed PIP simulation. . . . 179
6.10 Snapshots showing progression of a simulation of 9 PIP5K1A dimers with a complex asymmetric lipid membrane. . . . 182

List of tables

3.1 Table of large-scale simulations in simple symmetric composition bilayers. [a] Bilayer composition: 75% PC, 20% PS, and 5% PI4P (80%, 20% for PC/PS bilayers). [b] Of the five initial simulations of 1 μs duration, three were extended to 3 μs. 68

5.1 Table of α-synuclein simulations. 132

6.1 Table of large-scale PIP5K1A monomer simulations in simple symmetric composition bilayers. 163

6.2 Table of large-scale PIP5K1A dimer simulations in complex asymmetric composition bilayers. 181

Chapter 1

Introduction

"Everything that living things do can be understood in terms of the jigglings and wigglings of atoms."

Richard Feynman

1.1 The cell membrane

Every living cell is surrounded by a semi- and selectively-permeable barrier, the membrane. This statement is now widely accepted after the developments through the 17th, 18th, and 19th Centuries after the term 'cell' was first proposed by Robert Hooke in 1665 [1]. The cell membrane was thought to be an unessential secondary structure by the 1890s, until early electrophysiology and osmotic studies paved the way for the general acceptance of its existence in the early 20th Century. The idea that these membranes would be composed of lipids extends back to the late 19th Century [2] and after a series of experiments we arrived at the modern concept of the fluid mosaic model. Singer and Nicholson presented a fluid mosaic model for the cell membrane in 1972 [3], describing it as 'a two-dimensional oriented solution of integral proteins ...

in the viscous phospholipid bilayer'. Whilst this concept remains useful in our overall understanding of the cell membrane, this view has undergone some revision [4] and the true complexity of biological membranes is now beginning to be revealed. In the model, lipids are arranged in a bilayer, with the hydrophobic lipid tails at the centre and the hydrophilic/polar lipid head groups oriented towards the solvent. The lipids in the bilayer have a repertoire of functions, including energy storage, compartmentalisation of solvent, and signal transduction. Fluctuating nano-assemblies of lipid and protein form signalling platforms in discrete parts of the cell membrane [5, 6].

One refinement to this model includes the contribution of peripheral membrane proteins (PMPs) (fig.1.1). PMPs are soluble proteins that bind transiently to the cell membrane, generally to modulate signalling events [7]. By binding to the cell membrane, PMPs can cause physico-chemical changes which propagate a signal. Errors in these processes can have great effects on downstream processes, and therefore PMPs are implicated in many diseases; this highlights the need to investigate, understand, and modulate their mode of action.

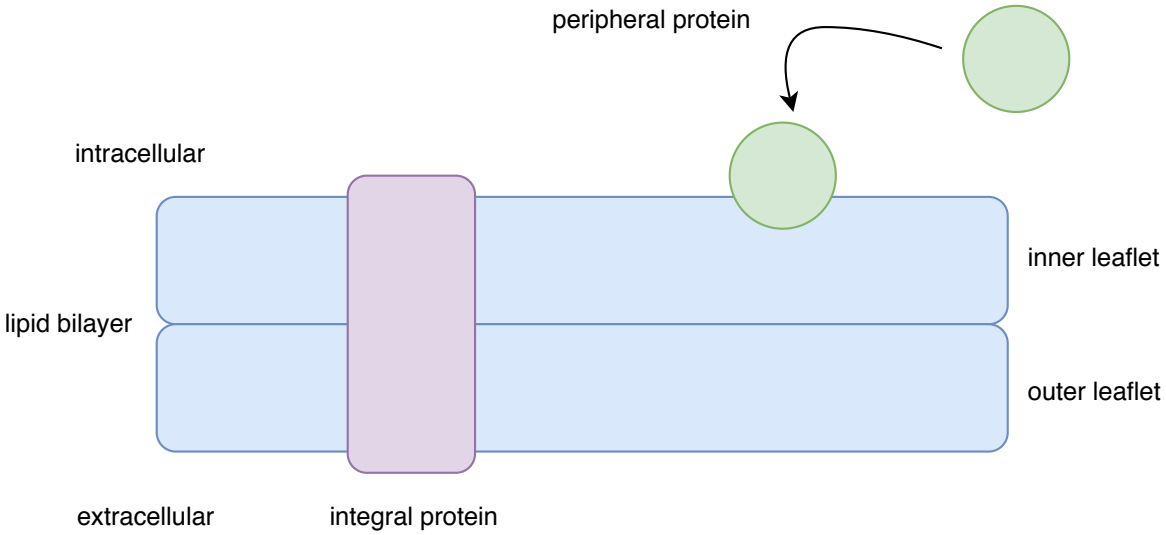


Fig. 1.1: The lipid bilayer with integral and peripheral membrane proteins. Blue indicates a lipid bilayer, with inner and outer leaflets corresponding to the intracellular and extracellular leaflets. Integral membrane proteins (purple) span the membrane and include channels and transporters. Peripheral proteins (green) exist within the cytosol but transiently and reversibly bind to the membrane surface via electrostatic and other non-covalent forces. [8]

Many PMPs recognise and bind to the cell membrane with exquisite spatiotemporal specificity. The binding sites of PMPs are often poorly understood and identifying a mode of action is challenging. Some simple structural motifs that contribute to binding are overrepresented in PMPs compared to other types of proteins [9]. There are inherent limitations in the classical biophysical techniques employed to study these proteins and therefore this thesis uses computational approaches, namely molecular dynamics simulations and mathematical/statistical modelling, to further elucidate these systems. An introduction to these concepts is presented in the following sections.

1.1.1 Lipids

The lipid composition of the bilayer includes phospholipids, glycerolipids, and sphingolipids [10]. This thesis is concerned with mammalian cells. There are over 1000 types of phospholipid molecules in cell membranes [11] and whilst precise composition

depends on the cell type, all mammalian cells posess phsphatidylcholine (PC) phosphatidylethanolamine (PE), sphingolipids and cholesterol in varying concentrations. PC, PE, and sphingolipids are zwitterionic, but the cell membrane also contains anionic lipids. Anionic lipids are generally found in the inner cytosolic membrane and are believed to constitute between 10-20% of total composition. Lipidomics has emerged as the study of the lipid components and interactions in cells and new analytical techniques address the difficulty in ascertaining the lipid composition [12].

Phosphatidylserine (PS), at around 2-10%, is an important anionic lipid component [11]. It contributes to cell signalling processes by recruiting proteins to the membrane through electrostatic interactions. Other anionic lipids are also present, and whilst their concentration may be lower, they make important contributions to cell signalling. One important family of signalling lipids are the phosphatidylyinositol (PI) derivatives; when these are phosphorylated, they are known as phosphatidylinositol phosphates (PIPs) or phosphoinositides. PI has an inositol (cyclohexanehexol) headgroup, a six-carbon ring that can be phosphorylated by various enzymes at the hydroxyl sites to produce various derivatives (PIPs) with different physical properties and functions.

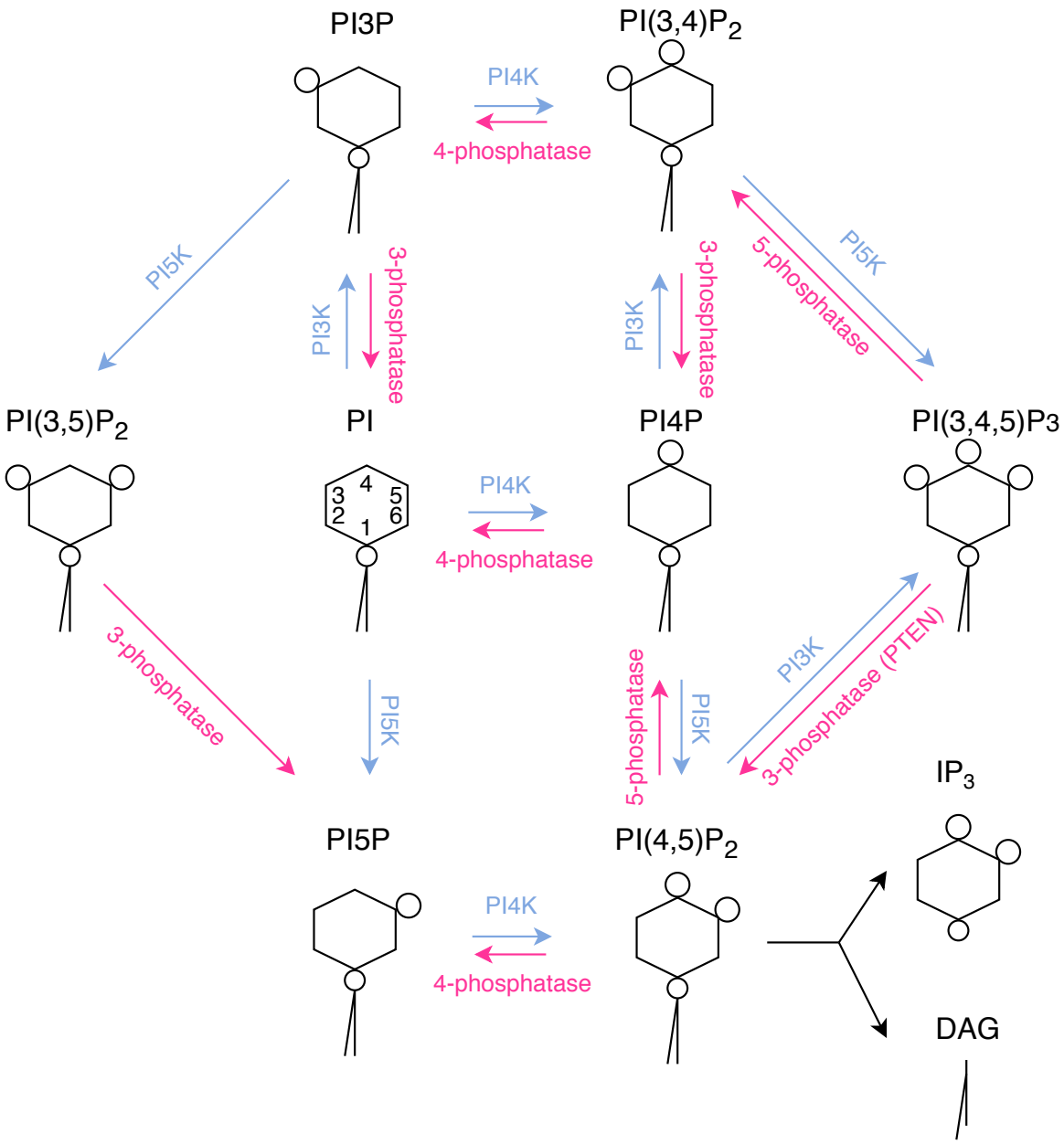


Fig. 1.2: Interconversion of PIPs. Representation of PIPs showing the 6-carbon ring (PI) with positions labelled. Kinases are shown in blue, phosphatases in pink. Each PIP species is labelled with the position of phosphates on the head group indicated by circles. Adapted from Bishé *et al.* 2012.

PI can be phosphorylated at the 3′, 4′, and/or 5′ carbon position on the inositol ring. The singly phosphorylated PI(4)P and doubly phosphorylated PI(4,5)P_2 are the most abundant of these derivatives [13]. PIP_2 constitutes around 1% of the total

lipid composition [14]. It has important roles in regulating cell proliferation, survival, and apoptosis, in addition to regulating membrane trafficking and and cytoskeletal reorganisation [15]. In calcium signalling, PIP_2 is hydrolysed by phospholipase C (PLC) to produce soluble $I(1,4,5)P_3$ and diacylglycerol (DAG). IP_3 goes on to trigger release of Ca2+ by binding to IP_3 receptors in the endoplasmic reticulum. DAG activates protein kinase C (PKC) on binding, which in turn promotes downstream signalling.

The discovery of phosphatidylinositol 3-kinase (PI3-K) introduced the idea of lipid products functioning as second messengers by recruiting PMPs to the cell membrane. Since then, a number of different PIP-binding proteins have been identified. It is now well established that a number of PMPs require a degree of non-specific interaction between basic residues of the protein and clusters of anionic phospholipids in the membrane [16]. The PMP landscape is described in the next section.

1.1.2 Peripheral Membrane Proteins

The first PIP binding domain identified was the plekstrin homology (PH) domain [17]. This domain is found in the plekstrin protein, but has since been identified in many other PMPs and is important for signal transduction through binding PIP_2 [18, 19]. The PH domain is around 100 amino acids in length and is present in many kinases, phospholipase C isoforms, and GTPases [20]. It has an open seven-stranded β-barrel structure, followed by an α-helix at the C-terminus. The domain has two binding sites: the canonical and the alternative. Each is flanked by loops that exhibit high sequence variation, which encodes a range of selectivity and specificity. Even domains that recognise the same ligand often use different structural elements for this ligand determination [21].

Since the discovery of PH domains, a number of different PMPs have been identified

[7]. These include conserved region-1 (C1) domains, C2 (named after the second homology region in PKC [7])domains, Phox homology (PPX) domains, FYVE (Fab1, YOTB, Vac1, EEA1) domains, and PROPPINs. Some of these domains are unspecific in their binding; others are highly specific for PIPs. Some domains, such as the BAR and ENTH families, induce or recognise membrane curvature. It follows therefore that the spatiotemporal regulation of PIPs, and in particular PIP_2, is essential for proper signalling. However, this regulation was poorly understood into the early 2000s and the evidence, for example for microdomains (also referred to as pools or rafts), remains challenging and sometimes contradictory [22]. Fig.1.3 summarises the PMP domain family and their substrates.

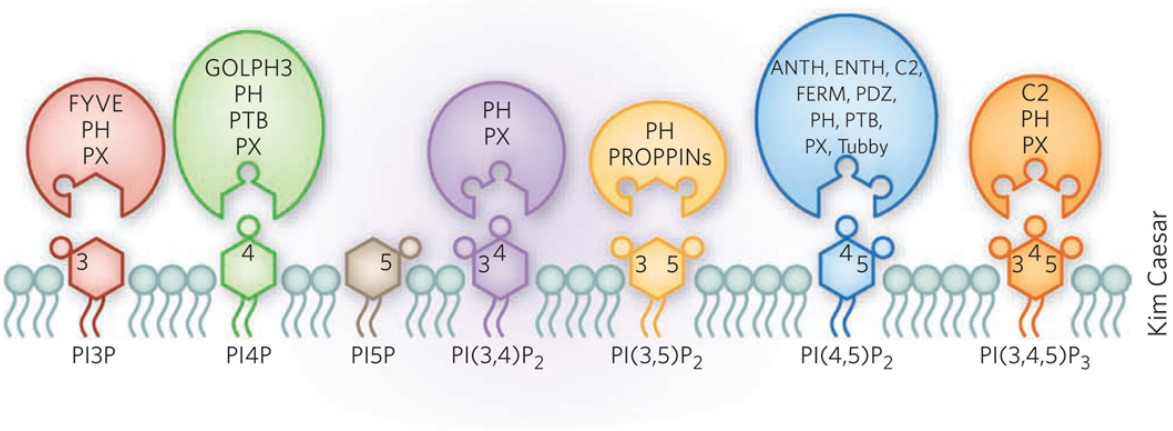


Fig. 1.3: PMP domains and their lipid substrates. Reproduced with permission from Macmillan Publishers Ltd: Nature Chemical Biology, Kutateladze T.G. **6**, 507-513 (2010), ©2020.

The phosphatidylinositol phosphate kinases (PIPKs), the study of which forms a large part of this thesis, are another class of PMPs. They catalyse the formation of PIP_2 from various PIP substrates through a number of pathways [23]. PIP5K1 consists of three isoforms: A, B, and C (sometimes referred to as α, β, γ) [24]. They have not received much attention until recently compared to other PMPs due to a lack of structural data. However, interest in PIPKs has been revived since the publication

of two X-ray crystal structures of the type I PIP5K1A kinase [25, 26]. PIP5K1A phosphorylates PI4P to form PIP_2 and therefore plays an integral role in cell signalling processes. In particular, it is implicated in prostate cancer [27].

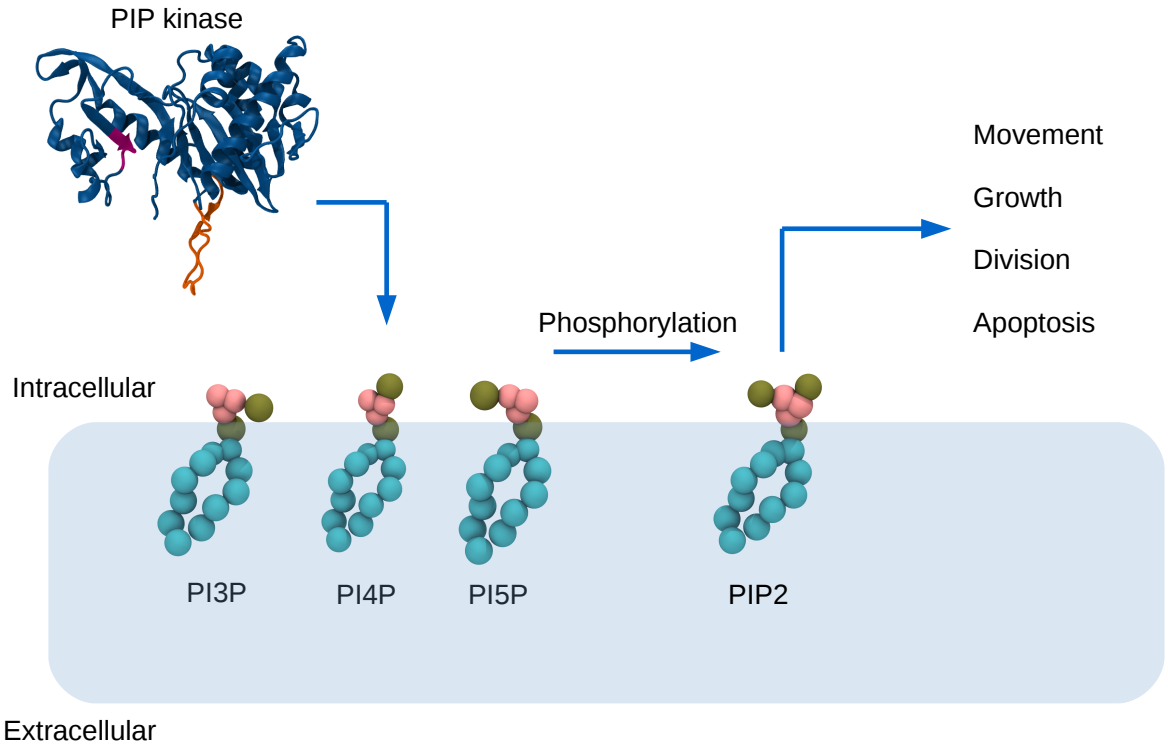


Fig. 1.4: PIP kinase signalling. Graphical representation of a PIP kinase and CG representations of PIPs, showing the conversion to PIP_2. Blue indicates the relative position of the membrane with intracellular and extracellular sides labelled. The PIP kinase, in this case PIP5K1A, is shown in blue (with its active site highlighted in pink and its activation loop in orange). PIP5K1A phosphorylates PI4P (tan bead) to form PIP_2, which regulates a number of signalling processes.

Another set of PMPs which have emerged as having great importance are proteins associated with neurological pathologies. One chapter of this thesis is concerned with the lipid-binding characteristics of α-synuclein, which associates with synaptic vesicle membranes. It is concentrated in the brain, but is also present in smaller concentrations in cardiac, muscle, and gut tissue. The normal function of α-synuclein is still unclear. However, it is linked to Parkinson's disease, and it is thought that one

of the ways it contributes is through aberrant oligomerisation on the cell membrane. This is thought to lead to neuronal death [28]. Evidence from mouse models showed that inoculation with synthetic α-synuclein fibrils triggered pathology, suggesting the α-synucelin can cause further disease progression by spreading to healthy cells [29]. Disease progression appears to spread through anatomically connected regions of the brain, suggesting that pathology is spread trans-synaptically [30, 31]. The sequence of α-synuclein is highly conserved across different organisms [32] and there are at least three isoforms in humans [33]. The most prominent is α-synuclein-140, the full-length transcript, which consists of three major regions:

- the N-terminus (residues 1-60), which contains four 11-amino acid motifs containing a conserved hexameric motif (KTKEGV);
- the central region (residues 61-95), contains the non-amyloid component (NAC) region which is prone to aggregation [34];
- the C-terminus (residues 96-140), which contains many acidic residues and prolines as well as three highly-conserved tyrosine residues.

A number of biophysical techniques have been used to characterise the α-synuclein structure, which is disordered under physiological conditions [35]. Intrinsically isordered proteins (IDPs) are a relatively new class of proteins that do not posess typical secondary structure and are highly dynamic and flexible entities. The sequence of α-synuclein is typical of IDPs and number of bioinformatics approaches predicted that α-synuclein is disordered [36–40]. Eliezer *et al.* used NMR spectroscopy to characterise the structures of α-synuclein in solution and when bound to lipid vesicles and micelles [41]. They showed that WT α-synuclein is unfolded in solution, but that the N-terminus adopts a helical conformation when bound to a membrane. The C-terminus remains free and unfolded. Toxicity is thought to arise from fibrillar structure

formation that triggers aggregation. X-ray and electron diffraction of α-synuclein filaments showed that the protein adopts an amyloid-like cross-β formation [42]. Recent advances in cryo-EM have enabled a number of studies which have determined the fibril structure [43–45]. The structural conformations of α-synuclein are investigated in chapter 5 of this thesis.

The two systems investigated in this thesis (PIP5K1A and α-synuclein) were chosen because they are both peripheral proteins that are involved in signalling, but their structure and function are currently less well understood than other systems outlined above. They are important targets in cancer and Parkinson's disease respectively, and therefore understanding their mechanism of binding to the membrane is of great importance. Structurally PIP5K1A is mostly globular in structure with short disordered regions, whereas α-synuclein transitions between being fully disordered and a partially folded helical conformation with the C-terminus remaining disordered. The PIP5K1A provided a the opportunity to investigate a shorter disordered region in the context of the membrane, and the methods developed in this thesis were then extended to α-synuclein. Studying these two systems posed a number of challenges because of their structural disorder. When simulating such systems it is important to consider carefully the technical approaches used; these include choice of force field (which appropriately parameterises disordered proteins), and assessment of convergence (because the energy landscape can be difficult to fully explore) [46]. A more detailed discussion of disorder follows in the next section.

1.1.3 The Role of Disordered Regions

In the previous sections, the idea of PMPs frequently having loops or disordered regions was introduced, and here follows a short discussion of protein disorder and its influence in lipid-binding PMPs.

IDPs and intrinsically disordered protein regions (IDPRs) are characterised by conformational flexibility and play various important roles in biological processes. They lack a well-defined tertiary structure. At the level of the amino acid sequence, IDPs are known to be enriched in disorder-promoting residues (ASP, MET, LYS, ARG, SER, GLN, PRO, and GLU) and depleted in order-promoting residues (CYS, TRP, TYR, ILE, PHE, VAL, LEU, HIS, THR, and ASN) [47]. This reduced sequence space gives rise to greater structural complexity in terms of conformational heterogeneity. In fact it is often sufficient to use only the global absolute charge and mean hydropathy to differentiate between ordered and disordered proteins [48]. However, IDPs exist as a spectrum of fully to partially folding proteins at one end with completely unfoldable proteins at the other; this is reflected in the free energy landscape, which is relatively flat and simple [49]. This results in IDPs being extremely sensitive to environmental conditions which moderate the landscape.

Reductionist approaches are slowly being complemented by systems-wide approaches. An example of this was a systematic quantification of the relationship between order/disorder and common single nucleotide polymorphism (SNP) variants. SNPs were found to be enriched in disordered regions, showing that structurally constrained regions are less tolerant to common variants, so instead they cluster in regions of disorder. However, pathogenic mutations are concentrated in ordered regions [50].

The purpose of protein disorder remains elusive; recently it has been shown that lipid-binding IDP/IDPRs comprise 15% of the DisProt database [51] in another early systematic approach. This is considered to be a conservative estimate due to the lack of experimental data. This suggests that lipid-binding, and therefore lipid metabolism, is an important aspect of the function of disorder in the cell. In addition, many pathologies are associated with lipid-binding by IDP/IDPRs. For example, variation in the bilayer pentetration depth affects nucleation propensity for amyloid fibril formation

[52], as illustrated by the synucleins which aggregate rapidly when exposed to long chain polyunsaturated fatty acids [53]. Furthermore, Tau fibrillisation is induced *in vitro* by large anionic surfaces, and it is speculated that anionic membranes may serve this function *in vivo* [54].

1.1.4 Protein-lipid Interactions

PMPs are characterised by their ability to interact reversibly with the membrane, and this requires the recognition of and binding to lipids. These protein-lipid interactions have been extensively investigated experimentally and computationally for many integral membrane proteins, including GPCRs, tyrosine kinases, phospholipases, and phosphatidylinositol-3-kinases [55]. However, PMPs present a particular challenge and so they have not received as much attention. One of the reasons for this is the heterogeneity of the system which means standard experimental techniques are limited.

Many different experimental techniques have been used to investigate protein-lipid interactions in integral membrane proteins. Lipid interaction sites can be identified with X-ray crystallography. Bacteriorhodopsin which was crystallised with 18 tightly-bound lipid chains, and found that trimers are mediated by lipids [56]. Similarly, protein-lipid interactions in secondary transporters [57] and voltage-dependent K^+ channels [58] have been studied with X-ray crystallography. However, these studies cannot account for the transient nature of some interactions and present challenges when dealing with detergents and expression. More recent developments include cryo-EM structures, which provide a high-resolution 3D view of proteins such as TRPV1 channels [59] and the ABC transporter [60]. Moreover, fluorescence spectroscopy can give some indication of function related to lipid composition [61] as demonstrated with the voltage-dependent anion channel (VDAC). Solid-state NMR can also be used to probe interactions and has been used for GPCRs and K^+ channels [62].

Peripheral membrane proteins have remained less well-studied, although PH domains have received more attention over recent years. The binding sites for lipids in PMPs have been determined through a number of biophysical methods. For example the PH domain of ceramide transfer protein (CERT) was crystallised with a sulfate hydrogen-bonded into the canonical ligand binding pocket [63]. This binding was shown by NMR to mimic PI4P binding, and that this binding results in structural rearrangements in the pocket, suggesting that this flexibility results in low affinity. The authors postulate that this could explain the role of CERT in shuttling between the golgi and endoplasmic reticulum, and therefore requiring only short residence times at each membrane. Another X-ray crystallography study of PX-PH tandem domains, whilst it did not observe lipid density directly, did observe a degree of flexibility which the authors postulate may be necessary for recognising different membrane environments with respect to curvature [64]. X-ray diffraction has been used to determine the structure of the PMP P2, a fatty acid binding protein, that binds phospholipid membranes; the structure revealed an internal hydrogen bonding network that determines the orientation of a bound fatty acid [65]. X-ray diffraction can also be combined with X-ray inferometry/holography to determine the orientation of yeast cytochrome C tethered to monolayers. Early X-ray studies of the FYVE domain Vps27p revealed lattice contacts that showed how a PIP molecule binds in a pocket flanked by an (R/K)HHCR motif. X-ray crystallography was also used to characterise the PIP binding site of PIP5K1A [26] and the authors speculated on how structural features conferred specificity. NMR has also been used to study the FYVE domains; endosome antigen 1 (EEA1) was studied to determine its orientation and micelle penetration. Interestingly, the study identified conformational changes in the membrane insertion loop that suggested its further penetration into the non-polar micelle core [66].

A popular biochemical experiment to determine PIP binding is the use of PIP

strips, where membranes containing different species of phospholipid are spotted with GST-tagged protein and measured with anti-GST antibodies; this can give a semi-quantitative measurement of specificity experimentally and has been used for PH domains such as Ipl and Tih1 [67]. Recently, the lipid-binding properties of 91 PH domains were quantified by liposome microarray-based assay [68]. Firstly, the study revealed that cancer mutations resulted in discrete binding affinity changes. Secondly, cooperativity was demonstrated to act as a key mechanism for membrane recruitment. Cooperativity was defined as when the binding affinity to a liposome containing two signalling lipids was greater than the sum of binding affinities to liposomes containing the individual lipid. The study revealed cooperativity between plasma membrane PIPs and phosphorylated sphingoid long-chain bases, phosphatidylserine (PS), and ceramides. This implies that membrane recognition is dependent on different lipids together, and that this cooperative effect partly explains the spatio-temporal specificity of PMPs. This sensitivity to membrane context was also demonstrated in three PS-receptors of the T-cell Immunoglobulin and Mucin containing (TIM) protein family [69]. Liposomes prepared with varying levels of PS were used to investigate effect the of PS surface density on TIM and Hill coefficients were used to show that binding is cooperative, and that this cooperativity cannot be explained just by multiple PS binding sites. Overall, these efforts to systematically quantify binding properties for whole lipid-binding protein families is now coming to the fore [70] and this will surely further advance the understanding of PMP-lipid interactions which have so far lagged behind integral membrane proteins.

Other approaches include quantitative spatial imaging and this has been applied to Ras proteins [71]. While many early studies of PMPs focused on electrostatics [72], this study demonstrated through a combination of fluorescence imaging and atomistic simulations that binding to the membrane is not only dependent on electrostatics, and

that conformational changes in binding regions have an important role to play. Binding anchors are a common feature of the Ras family; the synergy between their binding and different membrane characteristics (not just electrostatics) is only now beginning to be appreciated. It appears that there are a number of underlying biophysical characteristics of the membrane yet to be uncovered; this was demonstrated using molecular field theory calculations for N-Ras [73].

Where limitations in these methods arise, computational methods can be used to build models that can inform further experimental investigation, for example molecular dynamics simulation studies have been combined with experimental approaches (vesicle pull-down experiments) to investigate mutations in the EphA2 receptor ectodomain and how they reduce lipid binding [74, 75].

1.1.5 Previous Computational Studies of PMPs

This section outlines key papers which contributed to the computational study of PMPs. First I discuss bioinformatics approaches to investigate peripheral membrane proteins and their interactions with the membrane and then continuum electrostatics calculations and MD simulations.

Bioinformatics

A support vector machine (SVM) model was used to identify and classify peripheral membrane proteins from their sequences [76]. This approach was an early application of machine learning protocols to the PMP field and achieved an accuracy of 93.7% for identifying PMPs. The protocol was experimentally validated by first identifying which of four C2 domains with 50% sequence similarity bound to the membrane, and then verifying this with surface plasmon resonance studies. This was followed by a semi-supervised learning technique for genome-wide sequence-based prediction of

PMPs with accuracy up to 95% [77].

Two online databases with particular relevance to PMP prediction exist: MeTaDoR (membrane targeting domains resource) and the Orientation of Proteins in Membranes (OPM) [78]. MeTaDoR integrates sequence, structure, and membrane binding modes. It also includes MD resources, including parameters for more obscure lipids and an online prediction server. OPM was initially developed for transmembrane proteins but was expanded to PMPs [79] in 2012. However, the database only includes structures that can be computationally predicted rather than all membrane-associated proteins found in the PDB. This is because PMPs may adopt more than one conformation on the membrane (for example in PTEN) and many of the membrane-anchoring regions are unresolved in structures. The PPM server was integrated into OPM to discriminate PMPs and TM domains based on membrane penetration, transfer energy, and detection of one or two boundary planes (an abstract representation of where the phospholipid head groups sit). The server does not include NMR models with poorly-resolved loops, theoretical models, or models with missing membrane-anchor fragments, so the database is by no means exhaustive. Importantly, studies of the OPM database concluded that hydrophobic interactions are the main determinants of binding orientation at the membrane, and that whilst electrostatic interactions may have increased importance at long range, the picture becomes more complicated when it comes to short range interactions.

Other contributions from bioinformatics studies include models based on identifying hydrophobic protrusions and the concept of 'co-insertability' [9] which strongly discriminate membrane-binding and non-membrane-binding proteins. A combined bioinformatics and manual annotation approach was used to determined that the cytoplasmic inner membrane PMP proteome of E. *coli* cromprises about 19% of the overall proteome, rising to 25% of the expressed proteome in LB medium to mid-log

phase[80].

Continuum Electrostatics Calculations

In order to investigate the contribution of electrostatic interactions, continuum electrostatics claculations have been used to investigated the orientaton of PMP domains on the membrane. These include studies of the FYVE domain which reveal that PIPs may act as a masking 'switch', enabling deeper insertion into the membrane [81]. The binding properties of PLC homology model structures were characterised using continuum electrostatics calculations [82] and showed how a common structural fold scaffold can support a range of surface features that affect biophysical functions. However, both of the approaches outlined above are inherently limited due to their consideration of PMPs as static structures, even though the dynamic behaviour of these proteins is clearly important for understanding their function.

Molecular Dynamics Simulations

Here the use of molecular dynamics simulations comes into its own; a computational method that offers a time-resolved perspective of the dynamic nature of biomolecular systems. MD simulations may be used as a 'computational microscope' through which to view mechanistic details of PMP membrane-binding and behaviour.

Simulations of PMPs are somewhat more challenging than those of integral membrane proteins because extended simulation timescales are required for PMP binding events. This is because of the need to address some important factors in membrane recognition and binding: diffusion through and reorientation in the solution phase before binding, redistribution of lipids, conformational change, and membrane penetration. Additionally, the size of the simulation box is likely to be larger because of the need to position the PMP sufficiently far away from the membrane such that it is

able to tumble in solution. However, coarse-grained (CG) simulation approaches have been used to overcome these limitations (and their methods are discussed more fully in the methods section). Recent studies have probed these interactions with free energy calculations on GRP1 PH domains by use of umbrella sampling [83]. Free energy calculations can calculate the strength of lipid interaction and investigate the interaction process by potential mean force (PMF) calculations. Multiscale simulations have been used to investigate how the Dok7 PH domain interacts with PIP_2 and identifies canonical and alternative binding sites [84]. A number of atomistic simulation studies have been conducted on BAR [85], C2 [86], FYVE [81], PX [87], and PH domains [88, 89].

Analysing MD simulations can prove challenging and requires appropriate choice of methods to extract useful and valid information. One approach for this is the use of Markov state models (MSMs) which are used in this thesis. An introduction follows in the next section, with a more detailed explanation of the theory in the methods.

1.2 Markov State Models

Making sense of simulation data in a statistically valid and robust way is challenging; one approach to overcome this is the use of Markov state models (MSMs). The motivations for building Markov State models to analyse MD simulations are threefold [90]:

- To quantitatively predict experimental data of various forms from simulation trajectories;
- To handle input data efficiently;
- To build models that are easily understood by humans.

The MSM itself is essentially an $n \times n$ transition probability matrix, which gives the probability of transitioning from the current state into another, where n is the total number of states (fig.1.5). The theory of MSM analysis is given in the methods section. Here I present a short overview of the major advances in MSM theory as applied to MD simulations.

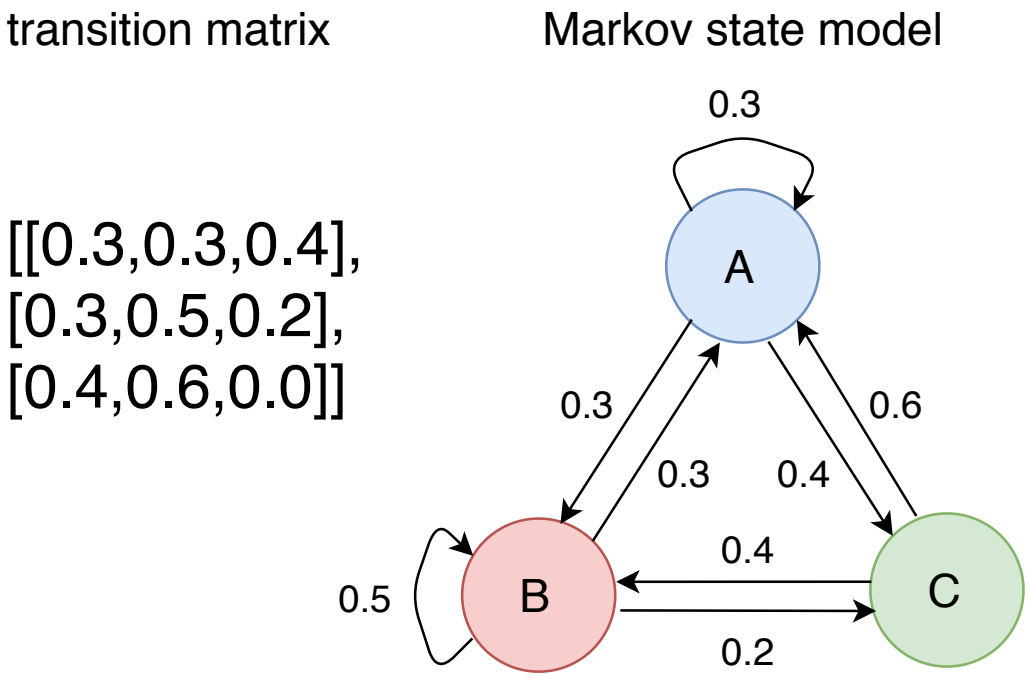


Fig. 1.5: The transition matrix and the Markov state model. Graphical representation of a transition matrix and the related MSM. The transition matrix is a matrix of probability of transition from the current (row) state to the subsequent (column) state. The MSM is labelled with the probabilities from the transition matrix. Note that state C has a 0% probability of transitioning back to state C, so there is no arrow indicating a transition back to state C. The sum of all rows is 1. The Markov property states that the probability of transition is dependent only on the current state (as shown by the probabilities).

1.2.1 Estimation and Validation of Markov State Models

MSMs are based on the Markov assumption that the conditional probability distribution of state x_{t+1} is based only on x_t. Simply put, the future states depend only on the current state, not the past. Markov processes are therefore memory-less. Brownian motion is a famous example of a Markov process.

The application of MSMs to MD systems began with the work of Schütte *et al.* [91], where they presented the formal application of the transfer operator (which is just a probability of moving from state x_t to x_{t+1}) to dynamical systems. Importantly they state that the system conformations should be defined by metastable states, as found

by approximating 'almost invariant' sets (which can be thought of as a dynamically isolated region) using the eigendecomposition of the transition matrix. This is as opposed to using geometric definitions for conformations.

The Markovianity of dynamics is assessed by checking that the processes identified by decomposition are constant beyond the smallest lag time τ. The validation of this method was presented by Swope *et al.* [92], noting that it is possible that folding behaviour is not Markovian at smaller timescales, but that Markovian behaviour can be recovered by choosing a larger tau timescale. This is the basis for implied timescale analysis, which is a special case of the Chapman-Kolmogorov test [93]. Additionally in a companion paper, Swope *et al.* [94]. notes that different folding mechanisms of a β-hairpin were uncovered when state definitions were varied.

Error arises from two sources when building MSMs [95]: firstly from poor state decomposition, and secondly from finite sampling of transitions. This has led to recent advances in 'adaptive sampling' where trajectories are analysed and then used to 're-seed' the sampling landscape. This enables areas of low sampling to be explored more fully.

There is now greater agreement about state decomposition for proteins thanks to work by Noé *et al.* [96], which identified internal degrees of freedom as superior to 'global descriptors' such as contact analysis. These factors highlight the difficulty of producing and validating MSMs, and it is often noted by practitioners of the method that choosing parameters is non-trivial and that application of the method requires a degree of experience [97]. Open challenges in using MSMs for protein simulation remain in choosing the optimal clustering method or distance metric that optimally reflects a kinetic distance metric. Further to this, a more quantitative and automated approach to parameter selection (e.g. for lag time) would be very useful. Finally, coarse-graining approaches vary in their accuracy and new methods to address this

would be very useful.

1.2.2 Software Advances & Application to Protein Simulation

Several theoretical advances throughout the 2000s enabled the use of MSMs for analysis of MD simulations. The first version of MSMBuilder, a python module for MSM analysis, was released in 2009 [98]. This was followed by EMMA (Emma's Markov Model Analysis) in 2012 [99], which was written in Java. Both programs have similar functionalities, but MSMBuilder was initially optimized for fast RMSD clustering of very large datasets, whereas EMMA 1.3 was implemented with greater statistical validation utility and was more focused on smaller MSMs. EMMA was superseded by PyEmma in 2015, which was implemented in python. MSMBuilder and PyEmma have been updated since and continue to have slightly different foci. PyEmma still places greater emphasis on statistical testing, whereas MSMBuilder was historically optimised for larger datasets and had more accessible documentation. MSMBuilder has the advantage of a visualisation plugin known as MSMExplorer[100].

MSMBuilder was first applied to the villin headiece (HP-35 NleNle), a small fast folding protein, and demonstrated that the model provided accurate structure prediction and that the longest timescale events corresponded to folding. They recommended the use of pairwise RMSD for featurisation and clustering the output with the k-centers algorithm. This led to several further key investigations, including the first model of *ab initio* protein folding using Folding@home simulations of the NTL9 protein [101]. Following this, MSMs were used to show that native states have hub-like behaviour by capturing folding on a 10 ms timescale of the λ-repressor [102].

MSMs have been used frequently in the study of protein folding and dynamics. They have been used for systems where the events along a protein folding pathway are of interest, such as the villin headpiece [103, 104]. They have also been used to

study conformational changes that facilitate lipid transport (opsin [105]) or signalling (Src kinases [106]).

1.3 Thesis Organisation

This thesis describes computational studies carried out on two model peripheral membrane protein systems (PIP5K1A and α-synuclein) and the associated methods development. This work arises from previous studies from the Sansom group. This thesis seeks to address four related questions.

- How can multiscale MD simulations be used to investigate a heretofore neglected PMP system, PIP5K1A?
- Can MSMs be used to investigate the disordered regions of the PIP5K1A system?
- Can the methods developed to investigate the PIP5K1A system be applied to another system of particular pharmaceutical interest?
- How do large-scale simulations inform cell signalling events?

These four questions are addressed in four chapters in this thesis.

The thesis is organised as follows. Chapter 2 contains a summary of the theoretical background of methods used in the thesis. Chapter 3 concerns a multiscale simulation study of PIP5K1A in its monomeric and dimeric form. A new structural study was published in 2015 at the very beginning of the project, and was therefore chosen as a new avenue for exploration with computational methods. The PIP5K1A is an attractive system for investigation because of the implications of its aberrant signalling. Chapter 4 builds on the findings of Chapter 3, where atomistic simulations and MSMs are used to probe in detail the behaviour of the disordered activation loop. This work was carried out in order to see if MSMs could be used to inform its structure

and function. Chapter 5 applies the methods developed throughout the project to a new system (α-synuclein) and was carried out as part of an industrial placement with pharmaceutical company UCB. A full understanding of this considerably more complex system has eluded researchers for years and therefore provided a challenging, yet interesting, system for further study. There is a wealth of experimental, and to a lesser extent computational, literature available. α-synuclein is an important clinical target and therefore represents an important avenue for study. Finally, Chapter 6 takes a macroscale view of the PIP5K1A system and comprises studies on simulations of larger membrane patches. This work was carried out in order to investigate the relationship between lipid composition and curvature and its relationship to PMP-membrane binding.

The novelty of this thesis lies in its combination of the study of challenging biological systems with computational and statistical methods at varied time and length scales. More precisely, to my knowledge at the time of writing, there has been no previous study of regions of PMP disorder with MSMs that take into account the membrane environment. It is hoped that the results from this work will further inform future work into complex PMP systems. Indeed, the work on α-synuclein described in this thesis has led to the development of new research projects. Additionally, a large amount of effort into developing new methods for studying these systems was undertaken and it is hoped that these methods may be implemented by other researchers to further investigate other important systems.

Chapter 2

Theory and Methods

"[Computational biology] has become a very large field and it has always in some ways been the poor sister, or the ugly sister, to experimental biology."

Michael Levitt

Computational approaches are now employed across the natural sciences to great success. This chapter provides a brief technical introduction to the underlying theory of molecular dynamics simulations, Markov state models, and machine learning approaches used throughout this thesis. It also outlines practical considerations of these methods and how they are implemented to best effect.

2.1 Classical Molecular Dynamics Simulations

Numerous biophysical techniques were developed over the 20th century and they gave unprecedented insight into the molecular basis of life. However, there is no escaping that these systems often elude our best efforts to describe their behaviour in time and

space. Techniques such as X-ray crystallography provide an inherently static picture, whereas other approaches such as NMR spectroscopy may be limited in their resolution capabilities in challenging environments. It is necessary to use techniques which aim to capture these details at the appropriate spatiotemporal resolution. Computational methods are therefore a useful approach, and this thesis is concerned primarily with the use of molecular dynamics (MD) simulation to answer biological questions. The aim of MD simulation is therefore to derive a trajectory using Newton's equations of motion that describes the movements of all atoms within a system over time. MD trajectories may then be subjected to a plethora of analyses. Like any biophysical technique, MD simulations come with their own limitations and present their own challenges in interpretation.

I begin with the mathematical basis for classical MD simulation. Here, 'classical' refers to the fact that we assume that the Born-Oppenheimer approximation holds and that we ignore the electronic degrees of freedom. Mathematically, the Born-Oppenheimer approximation states that the nuclear kinetic energy terms in the Hamiltonian operator for a given system may be neglected in solving for the electronic wavefunctions and energies. This is because the mass of the nuclei is greater than that of electrons by more than 1000 times; therefore the acceleration of the atomic nuclei is negligible compared to the acceleration of electrons. It is therefore justified to represent atoms in terms of a single nuclear Cartesian coordinate [107].

Consider a system of N atoms. Given the coordinates and momenta of all particles at time t, we can use Newton's equations of motion to give the coordinates and momenta at a later time $t + \Delta t$. These are given by the set of 3N coupled second

order differential equations:

$$\begin{aligned} \mathbf{F}_i &= m_i \ddot{\mathbf{r}}_i \\ &= -\nabla_i V_N(\mathbf{r}^N), \end{aligned} \tag{2.1}$$

where: F_i is the force on the i^{th} atom, m is the mass of the i^{th} atom, and $\ddot{\mathbf{r}}_i$ is the second derivative of the atomic position with respect to time, *i.e.* the acceleration. $V_N(\mathbf{r}^N)$ is the potential describing the interaction of the atoms, and the derivative of this potential gives F_i. Atom trajectories cannot be solved analytically because the potential is continuous; the set of differential equations shown above must be solved numerically using finite difference methods.

2.1.1 Integrating the Equations of Motion

Here follows a brief mathematical introduction to two methods for solving the equations of motion: Verlet [108] and leap-frog [109].

2.1.1.1 Verlet algorithm

Consider the coordinates r_i of atom i at time t. Performing a Taylor expansion forwards and backwards in time about r_i yields the following expression for the forward expansion:

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \Delta t \dot{\mathbf{r}}_i(t) + \frac{\Delta t^2}{2!} \ddot{\mathbf{r}}_i(t) + \mathcal{O}(\Delta t^3). \tag{2.2}$$

The corresponding expression for the backward expansion is:

$$\mathbf{r}_i(t - \Delta t) = \mathbf{r}_i(t) - \Delta t \dot{\mathbf{r}}_i(t) + \frac{\Delta t^2}{2!} \ddot{\mathbf{r}}_i(t) - \mathcal{O}(\Delta t^3). \tag{2.3}$$

We add together these two expressions to give an estimate for the particle coordinates at time $t + \Delta t$:

$$\mathbf{r}_i(t + \Delta t) \approx -\mathbf{r}_i(t - \Delta t) + 2\mathbf{r}_i(t) + \frac{\Delta t^2}{m_i}\mathbf{F}_i(t). \tag{2.4}$$

We subtract Equation 2.3 from Equation 2.2 to yield an estimate for the atom velocity $\dot{r}_i$ at time t:

$$\dot{\mathbf{r}}_i(t) \approx \frac{1}{2\Delta t}[\mathbf{r}_i(t + \Delta t) - \mathbf{r}_i(t - \Delta t)]. \tag{2.5}$$

This yields a finite difference method to solve Newton's equations of motion and calculate the atomic trajectories over time.

2.1.1.2 Leap-frog algorithm

The leap-frog algorithm evaluates velocities at half-integer time steps and uses these velocities to calculate the new positions.

As in Frenkel and Smit [110] the velocities at half-integer time steps are defined as follows:

$$\dot{\mathbf{r}}\left(t - \frac{\Delta t}{2}\right) = \frac{\mathbf{r}_i(t) - \mathbf{r}_i(t - \Delta t)}{\Delta t} \text{ and } \dot{\mathbf{r}}\left(t + \frac{\Delta t}{2}\right) = \frac{\mathbf{r}_i(t + \Delta t) - \mathbf{r}_i(t)}{\Delta t}. \tag{2.6}$$

We rearrange the right hand expression of Equation 2.6 to yield the new positions:

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \dot{\mathbf{r}}_i\left(t + \frac{\Delta t}{2}\right)\Delta t. \tag{2.7}$$

We equate Equation 2.4 with Equation 2.7 to extract an expression for the veloci-

ties:

$$-\mathbf{r}_i(t-\Delta t)+2\mathbf{r}_i(t)+\frac{\Delta t^2}{m_i}\mathbf{F}_i(t)=\mathbf{r}_i(t)+\dot{\mathbf{r}}_i\Big(t+\frac{\Delta t}{2}\Big)\Delta t. \tag{2.8}$$

We then collect terms, rearrange, and simplify to:

$$\dot{\mathbf{r}}_i\Big(t+\frac{\Delta t}{2}\Big)\Delta t=-\mathbf{r}_i(t-\Delta t)+\mathbf{r}_i(t)+\frac{\Delta t^2}{m_i}\mathbf{F}_i(t). \tag{2.9}$$

We then substitute the expression of the left hand side of Equation 2.6 to provide an expression for the velocities:

$$\dot{\mathbf{r}}_i\Big(t+\frac{\Delta t}{2}\Big)=\dot{\mathbf{r}}_i\Big(t-\frac{\Delta t}{2}\Big)+\frac{\Delta t}{m_i}\mathbf{F}_i(t), \tag{2.10}$$

arriving at a suitable integrator for carrying out MD simulations. The leap frog algorithm is used for all simulations in this thesis. For an outline of a typical MD workflow, see Fig. 2.1. Before the numerical integration is carried out, the forces must first be computed, and this process is outlined in the next section.

Fig. 2.1 shows a typical MD simulation program workflow. Before the equations of motion are computed, it is first necessary to compute the forces.

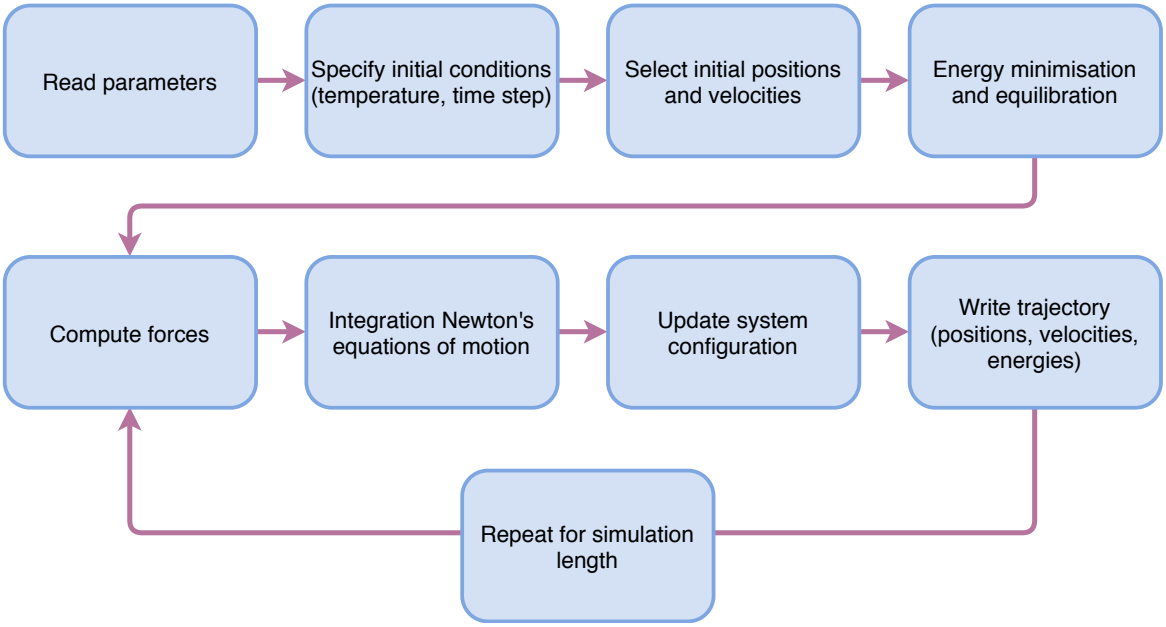


Fig. 2.1: Flowchart summarising the MD algorithm.

2.1.2 Computing the Forces

Computing the forces acting between all atoms or particles is often the most computationally intensive part of the simulation process. Atoms interact via a series of potentials classed as either bonded or non-bonded interactions. The forces are computed by calculating the forces between all non-bonded atom pairs and all the forces due to bonded interactions. The overall potential energy function is the sum of non-bonded and bonded interactions:

$$V = V_{non-bonded} + V_{bonded}. \tag{2.11}$$

Non-bonded Interactions

Non-bonded interactions act between atoms that are not covalently linked and are usually divided into two types: van der Waals interactions and electrostatic interac-

tions.

The Lennard-Jones potential approximates the van der Waals interaction between pairs of atoms. It is a mathematically simple model which results in its popularity in computational simulations, even where more accurate potentials could be used.

According to the short range LJ potential, pairs of atoms (i, j) interact as follows:

$$V_{LJ}(r_{ij}) = 4\epsilon_{ij}\left(\left[\frac{\sigma_{ij}}{r_{ij}}\right]^{12} - \left[\frac{\sigma_{ij}}{r_{ij}}\right]^{6}\right) \tag{2.12}$$

Where: ϵ_{ij} determines the depth of the potential well, σ_{ij} is the separation at which the potential is zero, and r_{ij} is the distance between the atoms.

The long range electrostatic interactions between pairs of charged atoms are modelled by the Coulomb potential:

$$V_C(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.13}$$

Where: q_i and q_j represent the charges on the two atoms, ϵ_0 is the permittivity of free space (an ideal physical constant, which is the value of absolute dielectric permittivity of a classical vacuum), and r_{ij} is the separation between atoms.

Bonded Interactions

Bonded interactions restrain the motion of covalently linked particles and consist of two-, three-, or four-atom potentials corresponding to bond stretching, bond angle, and bond torsion.

Bond stretching is represented by an harmonic potential:

$$V_b(r_{ij}) = \frac{1}{2}k^b_{ij}(r_{ij} - b_{ij})^2, \tag{2.14}$$

where b_{ij} is the reference bond length, k_{ij} is the spring constant of the bond.

The bond angle between a triplet of atoms (i, j, k) is represented by an harmonic potential:

$$V_a(\theta_{ijk}) = \frac{1}{2}k^\theta_{ijk}(\theta_{ijk} - \theta^0_{ijk})^2, \tag{2.15}$$

where $\theta_{i,j,k}$ is the bond angle and $\theta^0_{i,j,k}$ is the reference bond angle.

Angular vibrations are represented by a trigonometric function:

$$V_a(\theta_{ijk}) = \frac{1}{2}k^\theta_{ijk}\Big[\cos(\theta_{ijk}) - \cos(\theta^0_{ijk})\Big]^2. \tag{2.16}$$

Proper dihedral angles or bond torsions are described by dihedral potentials between quadruplets (*i,j,k,l*) of four atoms. These are implemented through the following potential:

$$V_d(\Phi_{ijkl}) = k_\phi(1 + \cos(n\Phi - \Phi_s)). \tag{2.17}$$

An alternative function for proper dihedrals is a Ryckaert-Bellemans function:

$$V_{RB}(\Phi_{ijkl}) = \sum_{n=0}^{5} C_n(\cos(\theta - \pi))^n. \tag{2.18}$$

Improper dihedral potentials are used to limit bending in planar groups and is

given thusly:

$$V_{id} = \xi_{ijkl} = \frac{1}{2} k_\xi (\xi_{ijkl} - \xi_0)^2, \tag{2.19}$$

where ξ_0 indicates the current improper dihedral and ξ_{ijkl} indicates the reference dihedral.

The time step length chosen for MD simulations is limited by molecular oscillations of high frequency and low amplitude. The time step must be short enough to monitor these fluctuations. Constraint algorithms are therefore used to overcome this limitation and allow a longer time step to be used, which increases computational efficiency. The LINCS algorithm is a common method employed in MD simulations [111].

2.1.3 Force Fields

The MD force field comprises two distinct components to describe particle interactions: the potential functions used to generate potential energies and forces (as described above), and the parameters used in these equations. These parameters are derived either from quantum chemical calculations [112] or fitting to spectroscopic data [113, 114].

There are three types of force fields, listed in order of increasing computational efficiency:

- all atom (AA) - parameters are provided for every single atom in the system
- united atom (UA) - parameters are provided for all atoms except non-polar hydrogens.
- coarse-grained (CG) - a representation where a small number of atoms are grouped as particles.

The force field families available in the GROMACS [115] simulation engine are: the Assisted Model Building with Energy Refinement (AMBER) force field [116, 117]; the Chemistry at HARvard Molecular Mechanics (CHARMM) force field [118–120]; the GROningen MOlecular Simulation (GROMOS) force field [121, 122]; the MARTINI CG force field [123]; and the Optimised Potentials for Liquid Simulations (OPLS) force field [124].

Three force fields were used in this thesis. The MARTINI force field version 2.1 was used for CG simulations of the PIP5K1A monomer and dimer in chapters 3 and 5 and for CG simulations of α-synuclein in chapter 6. The gromos53a6 [121] force field was used for all atomistic simulations of PIP5K1A. The GROMOS force field is based on free energies of hydration and apolar solvation.

The charmm36 force field was used for atomistic simulations of α-synuclein in chapter 5 [125]. It should be noted that since the start of the project an updated version of the CHARMM forcefield, charmm36m [126], has been released. This forcefield has been improved for intrinsically disordered regions by reducing the bias towards α-helix sampling; however there is good agreement between experimental and simulation data with regard to secondary structure characteristics in chapter 5 with the use of charmm36. Therefore we consider this still to be a valid approach to simulating α-synuclein and was the best choice given available backmapping strategies (see section 2.1.8 and methods chapter 5).

2.1.4 Periodic Boundary Conditions

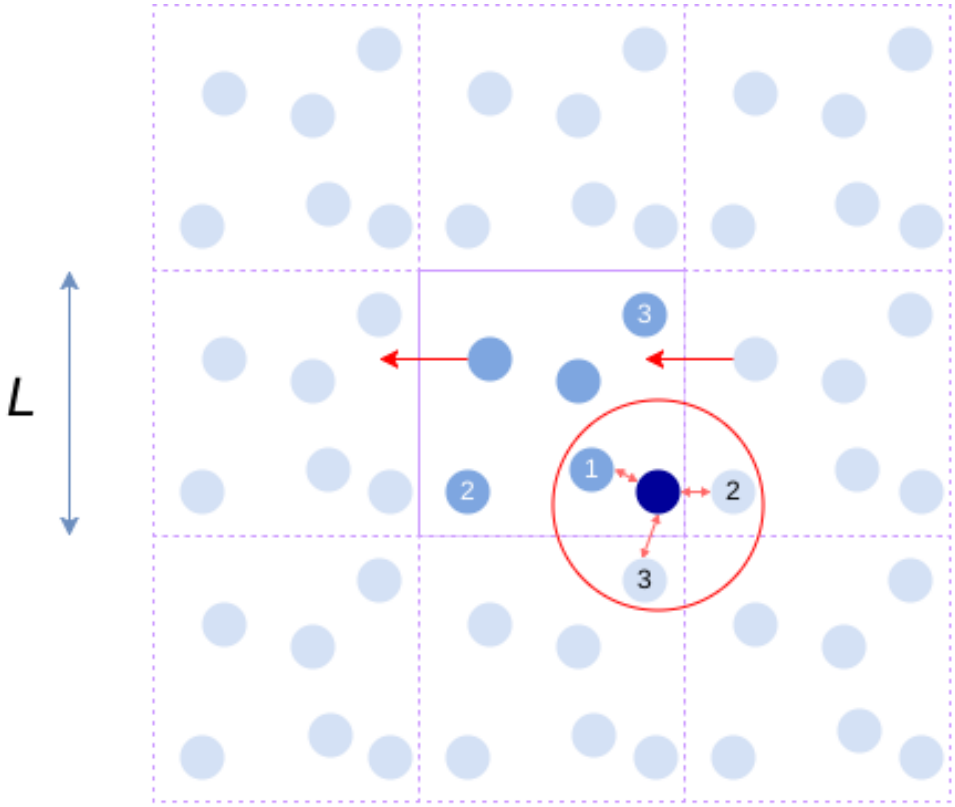


Fig. 2.2: Schematic representation of periodic boundary conditions in 2D. The simulation box is represented by the central square (solid line, particles in medium blue). The red arrow shows how a particle approaching and crossing the boundary reemerges at the opposite side at the same velocity. Interacting pairs of particles as determined by the minimum image convention are numbered. The cutoff distance is shown as a red circle.

Periodic boundary conditions (PBC) are used to approximate physical behaviour more appropriately by treating systems as a periodic lattice of cells. This represents the system as if it were in a homogenous bulk phase. Each molecule interacts with other molecules in its own periodic cell, molecules in its periodic image, and the periodic images of all other molecules. The disadvantage of this approach is that it then requires an infinite sum of all interactions; this is addressed by using appropriate interaction cutoffs, as described in the next section.

2.1.5 Treatment of Electrostatic Interactions

The non-bonded interactions impose a computational challenge when calculating the potential energy of the system. In order to avoid the situation outlined above, where PBC otherwise results in an infinite sum, the non-bonded interaction potentials are modified with cutoffs, typically of 1 nm. This is appropriate for the Lennard-Jones potential (see equation 2.12) which decays quickly because the distance term is of the order r^{-6}. As the derivative of this potential needs to be calculated, and a hard cutoff in the function would complicate this, it is common practice to shift and truncate this potential.

Generally, potentials should not be truncated unless the function decays faster than r^{-3} because of the formulation of the tail correction to a truncated energy potential (see Frenkel & Smit, 2002, p.31 [110]). We notice that the Coulomb potential decays as r^{-1} and therefore realise that the interaction is non-negligible even at large particle separations. We therefore make use of the Particle Mesh Ewald (PME) method in order to avoid inaccuracies in the force calculation.

The total Coulomb potential of a charge q_i located at the position $\mathbf{r}_i$ in the unit cell is expressed as:

$$V_C(\mathbf{r}_i) = q_i \Phi_{tot}(\mathbf{r}_i) = \frac{q_i}{4\pi\varepsilon_0\varepsilon_r} \sum_n {\sum_j}' \frac{q_j}{|\mathbf{r}_i - n\mathbf{r}_j|}, \tag{2.20}$$

where $\Phi_{tot}(\mathbf{r}_i)$ is the total electrostatic potential at $\mathbf{r}_i$, with the sums over j and n representing contributions from all the charged particles in the unit cell and their periodic copies respectively. The term $|\mathbf{r}_i - n\mathbf{r}_j|$ represents the distance between the charge q_i and q_j without taking into account the minimum image convention. The prime over the second sum indicates that for $i = j$ the term $n = 0$ should be omitted - meaning self-interaction by charge q_i is prevented, but that it does still interact with its replica images. Initially Ewald developed this method, now nearly a century ago,

for calculating electrostatic energies in ionic crystals. However, its relevance as to the assumption of periodicity in MD simulations is more tenuous and requires the consideration of the simulation cell as a unit cell repeated infinitely in all directions, and this is represented in 2D in Fig 2.2. Nevertheless, the method is superior to using hard cutoffs.

This sum above is only conditionally convergent, meaning that it does not converge absolutely (the absolute value of the sum is not finite). This means that the sum cannot be accurately evaluated. Particle mesh Ewald (PME) is a popular method for handling the electrostatic interactions, and it makes the conditionally convergent sum easier to evaluate by splitting the Coulomb energy into two parts. Then, the Poisson equation (a PDE commonly used to describe the potential field caused by a charge or mass density distribution) is solved by distributing the charges over a mesh. This conveniently discretises the problem. In order to tackle the sum above, the Coulomb energy can be partitioned by use of the following identity:

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1 - f(r)}{r}. \tag{2.21}$$

This is the efficient summation technique devised by Ewald that, given a judicious choice of f, results in two rapidly converging sums. This represents *adding* and *subtracting* a suitable charge distribution around the position $\mathbf{r}_j$ of each point charge q_j. If we consider a Gaussian charge density $\rho_j(r) = q_j e^{-(\alpha r)^2}$, solving Poisson's equation for electrostatics:

$$\nabla^2 \Phi_j = -\frac{\rho_j}{\epsilon_0 \epsilon_r}, \tag{2.22}$$

where $\nabla^2 \Phi_j$ is the Laplace operator (a differential operator given by the divergence of a gradient of a function on Euclidian space), leads to the corresponding electrostatic

potential:

$$\Phi_j(\mathbf{r}_i) = \frac{1}{4\pi\epsilon_0\epsilon_r}\frac{q_j\,\mathrm{erf}(\alpha|\mathbf{r}_i-\mathbf{r}_j|}{|\mathbf{r}_i-\mathbf{r}_j|}, \tag{2.23}$$

where erf() indicates the error function defined by:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t, \tag{2.24}$$

and it therefore follows that, taking into account the respective potential contributions due to the addition and subtraction of such Gaussian charge distributions around the q_j charges, the total electrostatic potential described in eq. 2.23. can be expressed as:

$$\begin{aligned}\Phi_j(\mathbf{r}_i) &= \frac{q_i}{4\pi\varepsilon_0\varepsilon_r}\sum_n{\sum_j}'\left[\frac{q_j}{|\mathbf{r}_i-n\mathbf{r}_j|}+\frac{q_j\,\mathrm{erf}(\alpha|\mathbf{r}_i-n\mathbf{r}_j|)}{|\mathbf{r}_i-n\mathbf{r}_j|}-\frac{q_j\,\mathrm{erf}(\alpha|\mathbf{r}_i-n\mathbf{r}_j|)}{|\mathbf{r}_i-n\mathbf{r}_j|}\right]\\ &= \frac{q_i}{4\pi\varepsilon_0\varepsilon_r}\sum_n{\sum_j}'\left[\frac{q_j\,\mathrm{erf}(\alpha|\mathbf{r}_i-n\mathbf{r}_j|)}{|\mathbf{r}_i-n\mathbf{r}_j|}+\frac{q_j\,\mathrm{erfc}(\alpha|\mathbf{r}_i-n\mathbf{r}_j|)}{|\mathbf{r}_i-n\mathbf{r}_j|}\right]\end{aligned}, \tag{2.25}$$

where $\mathrm{erfc}(x) = 1-\mathrm{erf}(x)$, the complementary error function. Using Ewald technique, the total electrostatic potential can be calculated as the sum of two potentials:

$$\Phi_{tot}(\mathbf{r}_i) = \Phi_{dir}(\mathbf{r}_i) + \Phi_{rec}(\mathbf{r}_i), \tag{2.26}$$

where Φ_{dir} and Φ_{rec} are respectively defined by:

$$\Phi_{dir}(\mathbf{r}_i) = \frac{q_i}{4\pi\varepsilon_0\varepsilon_r}\sum_n{\sum_j}'\frac{q_j\,\mathrm{erfc}(\alpha|\mathbf{r}_i-n\mathbf{r}_j|)}{|\mathbf{r}_i-n\mathbf{r}_j|}, \tag{2.27}$$

and:

$$\Phi_{rec}(\mathbf{r}_i) = \frac{q_i}{4\pi\varepsilon_0\varepsilon_r}\sum_n{\sum_j}'\frac{q_j\,\mathrm{erf}(\alpha|\mathbf{r}_i-n\mathbf{r}_j|)}{|\mathbf{r}_i-n\mathbf{r}_j|}. \tag{2.28}$$

The erfc(x) converges rapidly towards 0, and therefore the Φ_{dir} potential is obtained by direct calculation of the sum in eq. 2.27. This requires a cutoff of approximately 1 nm. This aproach has rendered the influence of charges not in the direct vicinity of q_i negligible due to the erfc($\alpha|\mathbf{r}_i - n\mathbf{r}_j|$) term. However, the sum in eq. 2.28 is slowly convergent as the erf() function rapidly converges to 1, and therefore the Φ_{rec} potential cannot be calculated directly. However, by noting that the Φ_{rec} potential corresponds to the electrostatic potential created by the Gaussian charge distributions ρ_j that were added to the system, and recalling the periodic nature of the system, Fourier transforms can be used to solve the corresponding Poisson equation in reciprocal Fourier space [127]. This can be expressed as

$$\Phi_{rec}(\mathbf{r}_i) = \frac{q_i}{4\pi\varepsilon_0\varepsilon_r} \sum_k \frac{\tilde{\rho}(k)}{|k^2|} e^{i\mathbf{k}\cdot\mathbf{r}_i}, \tag{2.29}$$

where the sum over k indicates a sum over reciprocal space vectors and $\tilde{\rho}(k)$ corresponds to the Fourier coefficient of the Gaussian charge distributions added to the system. This can be expressed as:

$$\tilde{\rho}(k) = S(k)e^{-(\frac{k}{4\alpha})^2} = \left(\sum_j q_j e^{-i\mathbf{k}\cdot\mathbf{r}_i}\right) e^{-(\frac{k}{4\alpha})^2}, \tag{2.30}$$

where $S(k)$ represents the structure factor of the system and is calculated over the charges present in the unit cell only. The terms $\frac{1}{k^2}$ and e^{-k^2} result in the fast convergence of the reciprocal space sum towards Φ_{rec}. The speed of convergence of the direct and reciprocal sums depend on the width of the Gaussian charge distribution, which is controlled by the α parameter.

Various Ewald summation schemes and their performances are presented in [128]. In summary, the PME technique provides an efficient and accurate technique that interpolates the distribution of charges q_j onto a mesh, which enables the use of

fast Fourier transform (FFT) algorithms to calculate the structure factor $S(k)$ [129]. Ewald summation in its traditional formulation is an $O(N^2)$ algorithm, whilst FFT approaches reduce this complexity to $O(N \log(N))$. It is important to note that historically the errors associated with the truncation of potentials have been thought to be negligible; however some studies suggest that the influence of the cutoff is greater than previously suspected and caution naturally must be exercised when interpreting results [130].

2.1.6 Temperature and Pressure Coupling

A molecular dynamics simulation by default consists of N particles moving within constant volume V. The energy, E, of that system is also constant. Therefore we refer to this statistical ensemble of states as the NVE (or microcanonical) ensemble. However, when simulating biological systems, it is often more desirable to simulate systems at relevant temperatures, T, and pressures, P. Consequently we couple the system to a thermostat to control T, yielding the canonical or NVT ensemble. We can also couple the system to a barostat, yielding the isothermal-isobaric or NPT ensemble.

The Berendsen thermostat [131] is a commonly used thermostat. It is based on the equipartition theorem in classical statistical mechanics. The equipartition theorem states that the total kinetic energy of a system is shared equally among all of its constituent parts, meaning that even though the kinetic energy of a molecule can fluctuate wildly, its *average* energy can be calculated at any temperature. At thermal equilibrium the relationship between the temperature T and the velocity v of a particle α with mass m is given as:

$$\langle v_\alpha^2 \rangle = \frac{k_B T}{m}. \tag{2.31}$$

, where k_B is the Boltzmann constant. This relationship can be used to monitor the temperature fluctuations over time:

$$k_B T(t) = \frac{1}{N} \sum_{i=1}^{N} m v_{\alpha,i}^2(t). \tag{2.32}$$

Because it is not possible to set the initial temperature accurately, the velocities need to be rescaled by a constant factor in order to adjust the temperature. The Berendsen thermostat achieves this by scaling accordingly:

$$\lambda = \left[1 + \frac{\Delta t}{\tau_t} \left(\frac{T_0}{T(t)} - 1 \right) \right]^{\frac{1}{2}} \tag{2.33}$$

λ is the velocity scaling parameter and is determined by the reference temperature T_0, the instantaneous temperature $T(t)$, and the coupling parameter τ_T (which can be thought of as regulating the strength of coupling). Overall, the role of the thermostat is to ensure that the average temperature of the system is correct, and that fluctuations in the kinetic energy are of the correct size. More recently if has been shown that in fact the Berendsen thermostat does not reproduce the canonical ensemble, but rather what is called a weak coupling ensemble, which lies somewhere between the microcanonical and canonical ensembles [132]. This thesis uses the velocity rescaling thermostat which is an extension of the Berendsen thermostat and enforces the correct distribution for the kinetic energy [133].

Similarly, pressure control is achieved through algorithms that almost always are based on methods derived by Andersen [134]. Andersen's derivation is lengthy and as this is not a methodological thesis, readers are directed to the original paper. A common implementation is the Berendsen barostat [131]. Suffice to say that the method is based on allowing volume to fluctuate around an average value, which is determined by the imposed pressure, P, and the internal stresses in the system. Deviations from the reference pressure are controlled by a compressibility matrix β.

It also allows rescaling of the box vector lengths and can lead to important changes the in volume and shape of the box.

Further work by Parrinello and Rahman [135] extended Andersen's methods to account for changes in shape as well as volume, and the reader is referred for more detail to the discussion by Nosé and Klein [136]. The Parrinello-Rahman barostat was used in this thesis.

2.1.7 Energy Minimisation

Before production simulation can begin, an energy minimisation (EM) step is carried out in order to ensure that the system has no steric clashes or inappropriate geometry, which would result in intractably high forces. The energy potential described above results in a complex energy hypersurface. This is particularly important when a starting configuration is generated by artificial positioning of molecules or when large systems are created by replicating smaller systems. Energy minimisation is usually carried out by the steepest descent algorithm. This is a 'first derivative' approach, as it exploits the derivative of the energy potential in order to find the path down over a given number of iterations to the nearest local energy minimum by line search [137].

2.1.8 Coarse-grained Molecular Dynamics Simulations

The set of equations described above represent a generic approach to simulating particles and their interactions within a molecular system. This was developed based on atomistic models, where each atom is represented explicitly. However, reduced resolution approaches can be used by using particles that represent more than one atom. The same mathematical formalism can be used to describe interactions between these coarse-grained (CG) particles. Because the number of overall particles is reduced, consequently the number of interactions is also reduced, and hence the number of

calculations that must be carried out computationally is fewer. This enables simulation approaches to approach longer time- and larger length-scales. This provides greater insight into biologically relevant timescales, which are often on the order of μ- to m-seconds and are out of reach of current atomistic simulation regimes.

The Martini forcefield is used for simulating CG systems in this thesis [123]. The philosophy of the Martini coarse-graining approach aims for a broad representation of molecules with wide applicability for different systems without the need for reparameterisation [138]. This force field was originally derived from extensible calibration against thermodynamic data, in particular oil/water-partitioning coefficients [123]. The forcefield was developed initially for lipids, and later extended to peptides [139], DNA [140], and RNA [141]. The protein parameters were updated in 2013 in order to correct for some minor shortcomings in the orginal model; these included improving the partitioning free energies of the proline, phenylalanine, and tryptophan side chains, a more realistic representation of interactions between oppositely charged residues by way of an off-center charge model, parameterisation of polarised beads, and some adjustment of bonded terms to improve the length of standard α-helices.

An approximately four-to-one mapping of heavy atoms to CG particles is employed, as shown in fig.2.3. The Martini model employs four types of CG particles which capture the chemical features of the underlying structure: polar (P), apolar (C), nonpolar (N), and charged (Q). Hydrophilic chemical groups are represened with polar particles, and hydrophobic groups are represented by apolar particles. Intermediate types are used for mixed groups. Q particles are used for groups with full (or almost full) charges, such as phosphates. Further detail is captured using subcategories for each particle type; for types Q and N, hydrogen bonding capability is distinguished by letter codes (d = donor, a = acceptor, da = both donor and acceptor, 0 = neither). For types P and C, polarity is indicated by a number between 1 (low) and 5 (high). Ring

particles are based on a 3:1 or 2:1 mapping for additional necessary resolution; these particles are indicated by the prefix S. For amino acids, sidechains were parameterised by calculating free energies of partitioning between oil/aqueous phases in a two phase butane/water system and comparing to experiment. It should be noted that four water molecules are described by a single particle type P4, and ions are represented by particle type Q, which includes their hydration shell. Periodic potentials are used to restrain dihedral angles between backbone particles; this means that Martini does not enable the study of conformational changes of secondary structure. An elastic network is applied to prevent the unfolding of tertiary structure [142]. Due to this limitation, this thesis employs a multiscale approach in order to backmap CG structures to an atomistic representation for further simulation and is described further below. Further simulation details are provided in the appropriate results section where required.

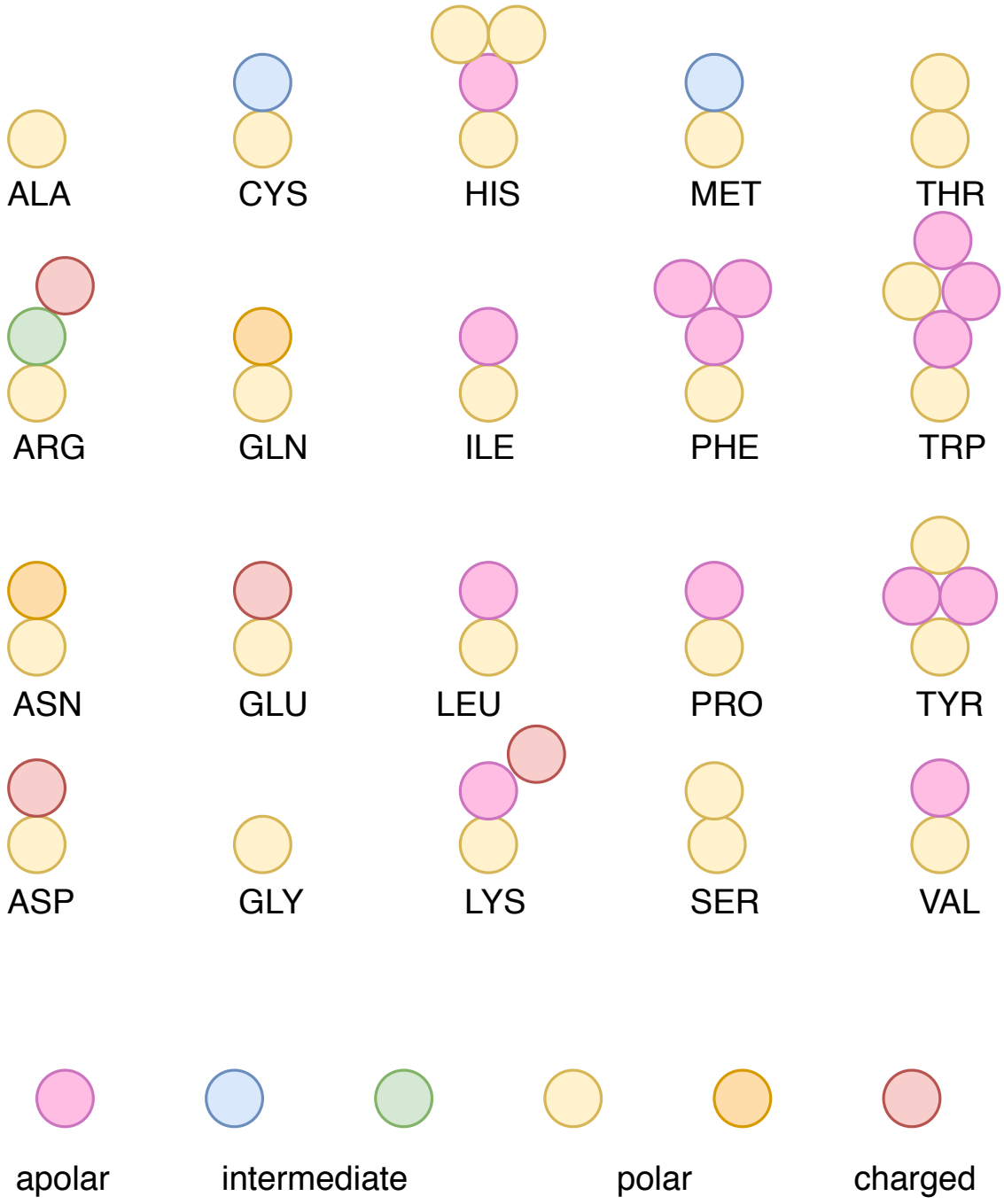


Fig. 2.3: Martini model CG representations of amino acid residues. Particle beads are coloured according to polarity and charge.

In the CG simulation regime, the Lennard-Jones and Coulomb interactions are shifted and truncated. This increases computational efficiency, but as described pre-

viously results in artefacts.

All simulations were carried out in GROMACS [115, 143, 144]. Details of parameters for each experimental setup are provided in the relevant methods section for each results chapter.

Multiscale Simulations

As outlined in the description of the Martini model above, the CG simulation regime is unable to capture large scale or secondary structure transitions in proteins due to use of an elastic network model. This thesis therefore makes extensive use of 'backmapping' software, in this case the CG2AT approach [145]. CG2AT employs a fragment-based fitting approach in order to convert a CG simulation frame to an atomistic representation. This approach has the advantage of capturing initial dynamics quickly though CG simulation, and then, by judicious choice of conversion time, further investigating system dynamics in more detail. Further details are described in the methods section of each results chapter.

2.1.9 Tools for Analysing Molecular Dynamics Simulations

Software

The GROMACS package has a number of functions for processing and analysis of simulation trajectories. Where these methods have been used, they are described in the requisite methods section in each results chapter, and some common algorithms have been described in detail below. However, there are limitations on the types of analysis that can be carried out with GROMACS tools and this requires that other analysis tools be used. This thesis makes extensive use of the Python programming language [146] and relevant modules are tabulated below.

Module	Application
MDAnalysis [147, 148]	trajectory processing & analysis
MDTraj [149]	trajectory processing & analysis
numpy [150]	array calculations, statistics
scipy [151]	curve fitting
pandas [152]	data structures, statistics
ipython/jupyter [153]	data analysis
scikit-learn [154]	clustering
matplotlib [155]	visualisation
seaborn [155]	visualisation
MSMBuilder [156, 157]	MSMs
MSMExplorer [100]	MSM visualisation
PyEMMA [158]	MSMs

Snapshots for visualisation were generated in Visual Molecular Dynamics (VMD) [159]. All of the code used for processing, analysis, and visualisation is available at https://github.com/sbtaamos, except where embargo applies; in this case, a representative protocol is provided.

GROMACS tools

Some commonly used GROMACS tools are outlined below.

The root mean square deviation (RMSD) of a protein's globular fold is calculated by a least squares fit of the structure at time $t = t_2$ compared to a reference structure at time $t = t_1$. This yields an average distance between a collection of atoms at t_2 *vs.*

t_1.

$$\mathrm{RMSD}(t_2, t_1) = \left[\frac{1}{M} \sum_{i=1}^{N} m_i |\mathbf{r}_i(t_2) - \mathbf{r}_i(t_1)|^2 \right]^{\frac{1}{2}}, \tag{2.34}$$

where $M = \sum_{i=1}^{N}$ is the total mass and $\mathbf{r_i(t)}$ is the position of atom i at time t.

Whereas the RMSD is a useful tool to examine the overall change in a fold over a simulation, the root mean square fluctuatation (RMSF) describes the average fluctuations of an amino acid residue over the course of a trajectory:

$$\mathrm{RMSF} = \left[\frac{1}{(t_2 - t_1)} \sum_{t=t_1}^{t=t_2} |\mathbf{r}_i(t) - \bar{\mathbf{r}}_i)|^2 \right]^{\frac{1}{2}}, \tag{2.35}$$

where $\bar{\mathbf{r}}_i$ indicates the the time-averaged position of an atom over the time interval $(t_2 - t_1)$.

The radial distribution function (RDF) is used to assess the structure surrounding a given atom, and is frequently used in this thesis to quantify lipid clustering. It is computed thusly:

$$\begin{aligned} g_{AB}(r) &= \frac{\langle \rho_B(r) \rangle}{\langle \rho_B \rangle_{local}} \\ &= \frac{1}{\langle \rho_B \rangle_{local}} \frac{1}{N_A} \sum_{i \in A}^{N_A} \sum_{j \in B}^{N_B} \frac{\delta(r_{ij} - r)}{4\pi r^2}, \end{aligned} \tag{2.36}$$

where $\langle \rho_B \rangle$ is the density of atom type B at distance r around atom A, while $\langle \rho_B \rangle_{local}$ is the time-averaged density of all atoms of type B around atom A up to a defined radius, r_{max}, usually half the length of the simulation box.

Other forms of specific analyses are described in the relevant methods section for each results chapter.

2.1.10 Limitations of Molecular Simulations

Like any experimental technique, MD simulations have their own set of limitations which must be carefully considered [160]. Firstly, classical MD simulations are unsuitable where quantum effects are important for understanding a system. Hybrid quantum mechanics/molecular mechanics (QMMM) simulations have been used to study some systems and may provide further insights in the future. Bonds are not made and broken in MD simulations and this has implications for the protonation states of titratable amino acid residues [161]. Further to this, the representation of electronic polarisation is limited, and polarisable force fields are still in early developments and are always computationally expensive [162].

Force fields are a common source of debate and even though there has been substantial improvement over the years, they are inherently approximate in nature [163]. As with all techniques, *in silico* experiments should be planned appropriately with parameters chosen according to the system of interest. Additionally simulation replicate numbers should be chosen such that the study results are statistically valid. There is debate as to whether it is better to run many short simulations or fewer long simulations in order to best capture processes. There is some evidence that the former approach is better for avoiding false positives [164], but this must be balanced with the approach that attempts to capture events on a longer timescale where resources are more limited.

An intrinsic limitation lies in the fact that even the longest simulations in the literature are at best on the order of milliseconds in length; whilst this has the advantage of providing exquisite temporal resolution, many biochemical processes are on timescales that are still out of reach. Advances in hardware have greatly helped in this respect [165], as have enhanced sampling regimes (such as replica exchange [166] or metadynamics [167]) and more advanced statistical protocols such as Markov state

models. This thesis makes extensive use of MSMs and these are described in the next section.

Related to this is the question of ergodicity. The ergodic hypothesis states that the microcanonical phase space ensemble average is equal to the time average over a simulation trajectory. This axiom of statistical mechanics means that we can equate simulation time averages to experimental ensemble averages. This raises the question of convergence and can be a challenging problem to address [168], particularly with respect to MSMs.

Finally, the availability and quality of structural data is an important factor in simulating biomolecules. A simulation must be based on an accurate structure, or the results are meaningless.

2.2 Markov State Models

Markov state models (MSMs) are a powerful approach for studying high-dimensional MD simulation trajectories [97], and offers a statistically robust set of methods for studying system dynamics over time. MSM analysis is now available to the simulation community via a number of open source software packages.

2.2.1 Theory of Markov State Models

A Markov chain is a stochastic model that describes a sequence of possible events, in which the current state depends on only the state of the previous event. This 'memorylessness' is a characteristic of all Markov processes, named after the Russian mathematician Andrey Markov, best known for his work on stochastic processes.

In 1983 Zwanzig outlines that a continuous time random walk can be used to treat migration of a classical dynamical system between regions of configuration space

[169]. The system is memoryless, meaning that it does not 'remember' how it got to its current state, and that therefore has no effect on progression to the next state. An MSM represents a master equation framework, and this means that the MSM can be used to describe the entire dynamics of a system. The MSM itself is an $n \times n$ matrix, which is often a transition probability matrix, where the entire configuration space has been divided into n states. We can assess the progression of a dynamical system by observing the state of a system at two points separated by the lag time τ. It is necessary to choose a lag time that is long enough to be Markovian, as described above, but that is also short enough to resolve system dynamics appropriately.

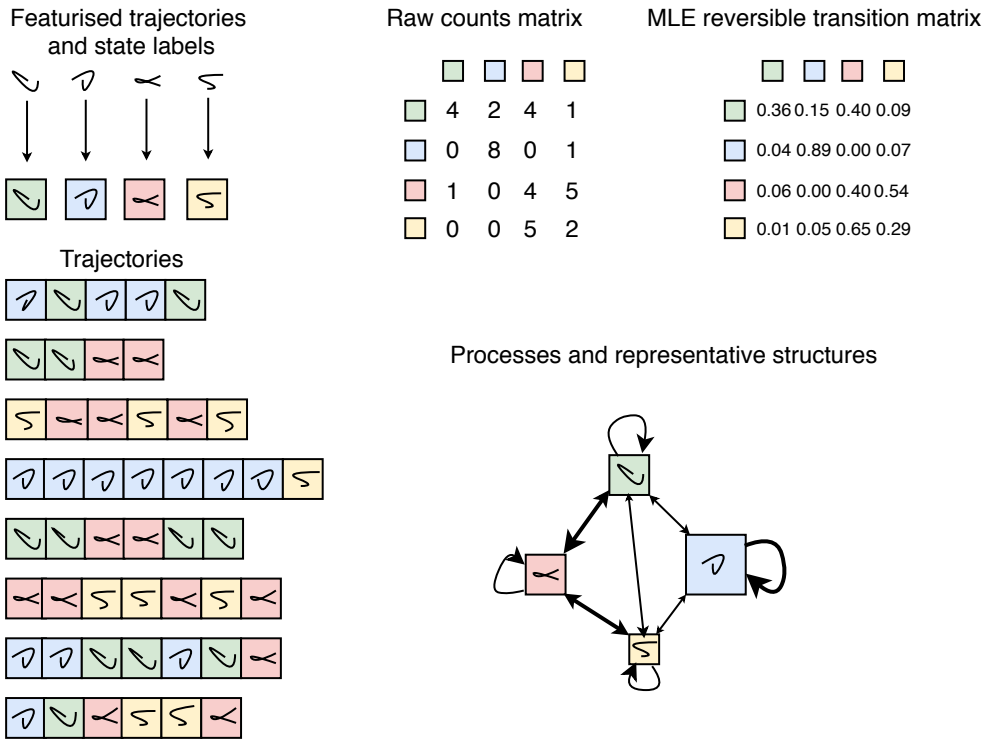


Fig. 2.4: Visual outline of the construction of Markov state models from MD simulation trajectories. The trajectories are featurised according to some metric e.g. RMSD, dihedrals and state labels assigned. Trajectory frames are then assigned to a state, and the transition counts between states are represented by the raw counts matrix. The MLE is used to calculate a reversible transition matrix. Eigendecomposition of this matrix can be used to deduce the dominant processes and calculate timescales for transitions between states. State centre features can then be reverse-transformed to produce representative structures. Adapted from Husic & Pande, 2018 [97].

This is summarised in fig.2.4 where the trajectories are assigned state labels (indicated here by the colours) and then used to calculate the transition counts matrix, from each row-indexed state into each column-indexed state. This is then used to calculate the reversible transition matrix. The first eigenvector of this matrix is used to calculate the adjusted state populations. This can be converted to free energies by use of statistical mechanics, and the transition probabilities can be used to construct kinetic schemes of possible pathways that might describe the process of interest.

The transition matrix provides a model for the system and yields a number of convenient properties. Given that a system is in thermodynamic equilibrium (every transition happens and equal number of times forward and backward), symmetric (with respect to the equilibrium distribution), ergodic (meaning any state can be reached from any other state given enough time), and aperiodic (the equilibrium distribution is independent of starting conformation), we can obtain a wealth of information about the system through eigendecomposition of the transition matrix:

- The eigenvectors corresponding to each eigenvalue have n elements corresponding to each of the the n states, and the magnitudes and signs of these elements explain which states are contributing to the process (as defined by the eigenvalue).
- The highest eigenvalue is 1, corresponding to the equilibrium distribution.
- All other eigenvalues are less than 1; positive eigenvalues correspond to decay processes, and negative eigenvalues correspond to oscillations (although in practice the latter are not physically meaningful).
- The positive eigenvalues can be converted to physically meaningful timescales by recalling the lag time τ.

An excellent full mathematical description of MSM theory and construction is given by Bowman *et al.* [170] and the reader is referred to this book for further details.

2.2.2 MSM workflow

The practical steps construction of MSMs is outlined in fig.2.5.

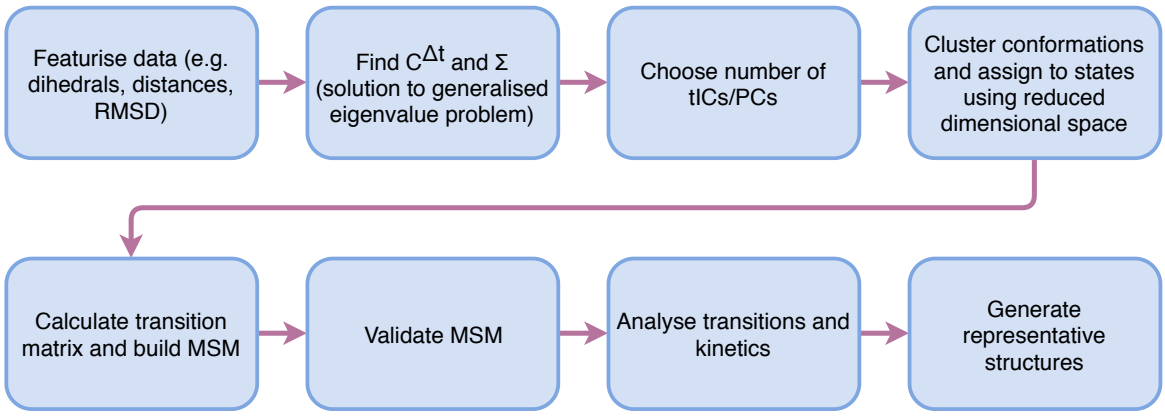


Fig. 2.5: Visual outline of the MSM workflow. The flowchart shows the critical steps for performing MSM analysis. Firstly, data is featurised from raw coordinates into a suitable collective variable, often angles or distances. This is followed by performing tICA analysis which yields a solution to the generalised eigenvalue problem. The number of tICs is chosen, and then this reduced dimensionality space is used to cluster conformations. From this the transition matrix is calculated. The MSM can be validated using different approaches and analysed. Ultimately in this thesis, the goal is to generate representative structure which give a picture of how proteins behave in their respective binding events.

Fig.2.5 alludes to dimensionality reduction for MSMs; in this thesis this is carried out using time-lagged independent component analysis [97]. The method for MSMs was introduced independently in two studies [171, 172] in 2013. tICs serve as reaction coordinates for the system and encode kinetic similarity in the states (as opposed to just using structural similarity as a proxy for kinetic similarity). Considering a high dimensional simulation trajectory $\mathbf{X}(t) \in \mathbb{R}^d$, tICA analysis aims to find projection vectors that maximise their autocorrelation function [157]. We state this generalised eigenvalue problem (which is closely related to the PCA eigenvalue problem (see section 2.3.1) but tICA is more appropriate for capturing slow transitions):

$$C^{(\Delta t)} v = \lambda \sum v, \tag{2.37}$$

where $C^{(\Delta t)}v$ is the time lag correlation matrix defined by:

$$C_{ij}^{(\Delta t)}v = \mathbb{E}\Big[X_i(t)X_j(t+\Delta t)\Big]. \tag{2.38}$$

tICA can be thought of as interpreting the slowest motions in a trajectory (such as a folding event, for example). It is important to remember that with any dimensionality reduction, there is the possibility of throwing out important information, and tICA sometimes has the drawback of not representing faster timescales well [157].

2.3 Machine Learning Techniques

With the ever increasing amount of data produced in the sciences, there is a need for techniques that are able to process and infer from very large datasets. Machine learning is a field closely aligned with statistics but differs in its terminology and emphasis. The aim is to build mathematical models derived from sample data (a 'training set') that are able to make predictions on a 'test set' without explicit programming to do so [173].

Broadly, machine learning algorithms are divided into two approaches: supervised and unsupervised learning [174]. Supervised learning exists primarily to solve classification problems through approaches such as linear regression, logistic regression, support vector machines, and naive Bayes etc. The process takes an input object (often, but not always, a vector) and produces a desired output value. We can formulate this as the goal being to map $\mathbf{x}$ to outputs y where $y \in \{1, ..., C\}$, where C is the number of classes. However, supervised learning is of limited use where labelled datasets are not available, and has limited use within this thesis.

Unsupervised learning has broad applications to data analysis as in contrast to supervised learning its purpose is to explore the data and look for meaningful patterns.

We are not told the output for each input. Instead, we formulate this as a density estimation problem $p(x_i|\theta)$. This is in contrast to $p(y_i|x_i, \theta)$ as in a supervised learning example. Supervised learning is conditional (on $\mathbf{x}$) whereas unsupervised learning is unconditional.

Unsupervised learning arguably has wider applications, and certainly within this thesis has been used in many parts of the data analysis process. Here follows a brief description of the major unsupervised learning algorithms employed.

2.3.1 Principal Component Analysis

Principal component analysis (PCA) is a statistical technique for dimensionality reduction and exploratory data analysis. It was invented by Pearson in 1901 [175], and then independently proposed and named by Hotelling in 1933 [176]. A thorough mathematical treatment of the PCA theorem and proof is given by Murphy [174]. Here follows a short mathematical statement of the PCA problem and a graphical illustration for clarity.

Consider a 2-dimensional dataset of randomly generated uniformly distributed numbers, as shown in Fig. 2.6. By eye, it is clear that there is some kind of linear relationship between the X and Y variables. In a linear regression problem (i.e. a supervised approach) we would attempt to predict the Y values from the X values. In this case, PCA attempts to quantify this relationship by finding the principal components in the data, and the explained variance. These quantities are visualised in Fig.2.6, with the direction of the vector indicating the principal component, and the squared length of the vector indicating the explained variance.

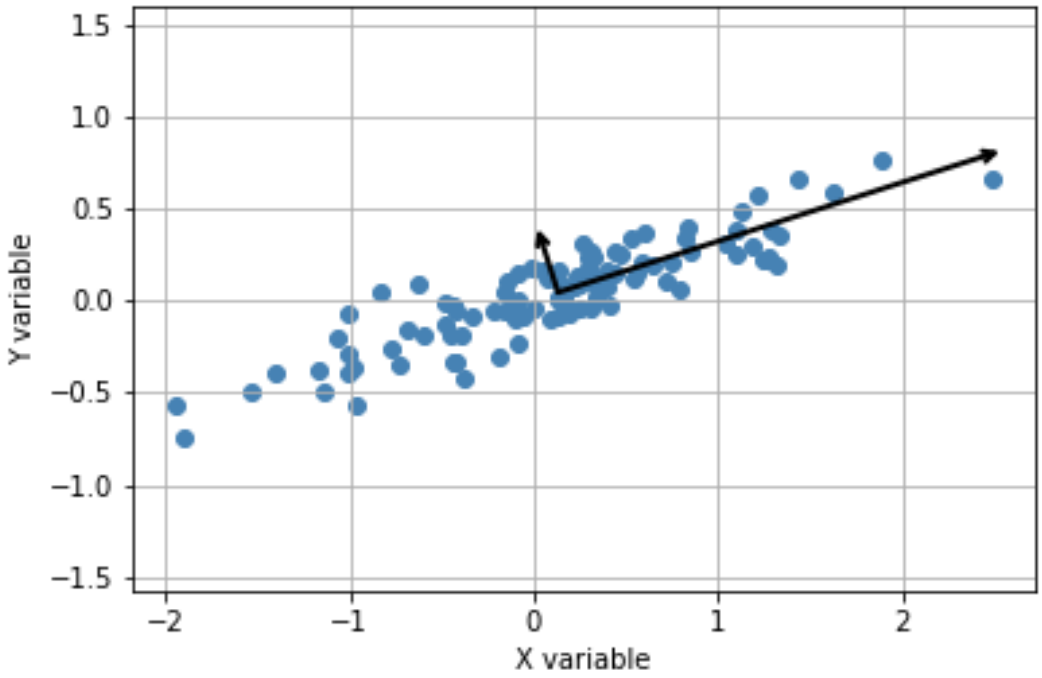


Fig. 2.6: Vectors indicating the first 2 principal components of 2D dataset. The direction of the vector indicates the principal component, and the length of the vector indicates the explained variance.

We can see that the first principal component describes the greatest variance of the dataset, and we can think of that as an indication of how 'important' that axis is in describing the dataset. In very large, high-dimensional datasets, it can be useful to reduce the dimensionality by transforming the data in order to make a dataset more tractable whilst still preserving maximal variance.

The mathematical formalism is briefly described as follows [177]:

A data matrix, $\mathbf{X}$, can be transformed as defined by a set of p-dimensional vectors of weights $\mathbf{w}_{(k)} = (w_1, \ldots, w_p)_{(k)}$. These weights map each row vector $\mathbf{x}_{(i)}$ of $\mathbf{X}$ to a new vector of principal component scores $\mathbf{t}_{(i)} = (t_1, \ldots, t_l)_{(i)}$. These scores are given

by:

$$t_{k(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(k)} \qquad \text{for} \qquad i = 1, \ldots, n \qquad k = 1, \ldots, l, \tag{2.39}$$

which shows the transformation of row vector $\mathbf{x}_{(i)}$ is transformed by the weights $\mathbf{w}_{(k)}$. Each of the individual variables $t_1, \ldots, t_l$ of t successively inherit the maximum possible variance.

The first weight vector must satisfy:

$$\begin{aligned} \mathbf{w}_{(1)} &= \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (t_1)^2_{(i)} \right\} \\ &= \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w} \right)^2 \right\} \end{aligned} \tag{2.40}$$

where the argmax operation determines the input point where the output of the function is maximised. Written in matrix form, this gives:

$$\begin{aligned} \mathbf{w}_{(1)} &= \arg\max_{\|\mathbf{w}\|=1} \{ \|\mathbf{X}\mathbf{w}\|^2 \} \\ &= \arg\max_{\|\mathbf{w}\|=1} \left\{ \mathbf{w}^T \mathbf{X}^\mathbf{T} \mathbf{X} \mathbf{w} \right\} \end{aligned} \tag{2.41}$$

An important result arises out of the fact that given w is defined to be a unit vector, it also satisfies

$$w_{(1)} = \arg\max \left\{ \frac{\mathbf{w}^T \mathbf{X}^\mathbf{T} \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\} \tag{2.42}$$

and therefore the maximisable quantity is a Rayleigh quotient; the maximum possible value is therefore the largest eigenvalue of the matrix, with w as the corresponding eigenvector. Further components are then found by subtracting the first principal

components from $\mathbf{X}$ and then finding the weight vector to find the maximum variance from the new matrix. For further details the reader is referred to the work of Jolliffe [177].

This shows how information can be conveyed by using PCA to transform datasets. This becomes very useful when dealing with high-dimensional datasets, and is employed throughout this thesis. Simulation data is high dimensional; a trajectory may carry information about angles, coordinates, and orientations, and deciding which components of the dataset are most important for analysis is an exceedingly useful approach.

2.3.2 Clustering

Clustering is a process by which data points are grouped together according to some algorithm in order to infer meaning from the data. This meaning usually arises out of some kind of classification of the data. Cluster analysis is used frequently throughout this thesis for data analysis and interpretation, and this section introduces the main algorithms used. All clustering was implemented in python using the scikit-learn module, except in chapter 4 where the R programming language [178] is used in order to make use of the marginal histogram plotting function.

The scikit-learn python module enables the use of many different clustering algorithms; this thesis uses KMeans, MiniBatchKMeans, DBSCAN, and Gaussian Mixture Modelling. Figure 2.8 presents a visual introduction to the performance of each algorithm on different datasets.

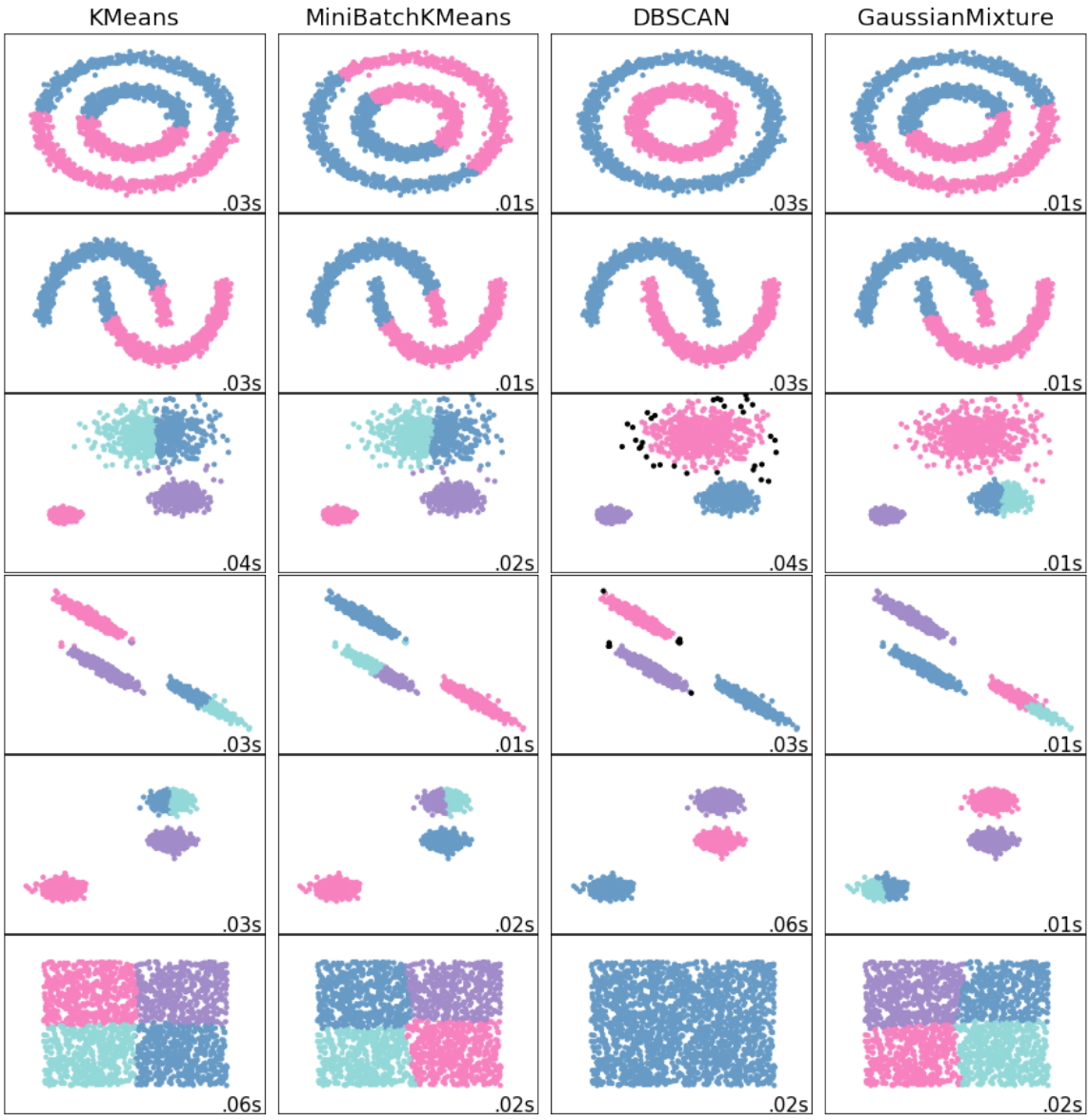


Fig. 2.7: Comparison of clustering algorithms used in the thesis. The scikit-learn python module was used to randomly generate data samples and apply the four clustering algorithms. Colour indicates cluster assignment. The time for clustering to run on 2000 data points is indicated in the bottom right of each square. Outlier points, where found by the clusterer, are shown in black. Adapted from the scikit-learn reference manual [154]

K-Means is widely used in data analysis and was introduced by MacQueen in the 1960s [179]. It remains a popular clustering algorithm due to its ease of implementation and computational efficiency. K-Means attempts to group data points that minimised

the squared Euclidean distance, given by:

$$d(x,y)^2 = \sum_{j=1}^{m}(x_j - y_j)^2 = \| x - y \|_2^2, \tag{2.43}$$

where j is the jth dimension of the sample points x and y.

K-Means attempts to minimise the cluster inertia factor, given by the Sum of Squared Errors withing the clustering context:

$$\text{SSE} = \sum_{i=1}^{n}\sum_{j=1}^{m} w^{(i,j)} \| x^{(i)} - \mu^{(j)} \|_2^2, \tag{2.44}$$

where $\mu^{(j)}$ is the centroid for cluster j, and $w^{(i,j)}$ is 1 if sample $x^{(i)}$ is in cluster j and 0 otherwise.

Important considerations when using K-Means are that the output from a dataset will change whenever the algorithm is run because the initial centroids are set randomly. Additionally, due to the reliance on Euclidian distance, it is an unsuitable algorithm when dealing with non-spherical clusters (and this has important implications in the thesis). Fig. 2.5 illustrates the performance of K-Means on different sample datasets. Note that the algorithm is fairly good at identifying spherical clusters, but performs poorly on other datasets. K-Means also requires the specification of the number of clusters, and so identifies 4 clusters even when it may not be a suitable number to cluster the data.

MiniBatch K-Means is a faster implementation of the K-Means algorithm and is suitable for very large datasets. However, its performance at cluster boundaries can be less accurate, and this is shown in the comparison.

Density-Based Spatial Clustering of Applications with Noise, of DBSCAN, is another clustering used in this thesis. DBSCAN clusters data by a user-specified number of minimum points (min) required to form a cluster. a point p is a core point if $\geq$

min points are within a within a defined radius ϵ of point p. A point p is a border point if it falls within radius ϵ of a core point, but has $\leq min$ number of minimum points as neighbors. All other points are considered noise. DBSCAN has the considerable advantage that we do not need to specify the number of clusters. In this thesis, DBSCAN was used for time-resolved clustering of points in large model membranes. Given that the number of clusters changes over time, DBSCAN was chosen for this task. It also has the advantage of identifying clusters that are not circular. In Fig. 2.5, it can be seen that DBSCAN performs well across a number of datasets.

Gaussian Mixture Models (GMM) assume that all samples are generated from a number of Gaussian distributions with as-yet unknown parameters. GMM is a soft clustering technique, which means it assigns a probability that a given point belongs to a given cluster. It follows an expectation-maximisation algorithm. Whilst GMM has the advantage of offering a probabilistic perspective on the clustering, it is sensitive to initial values and tends to converge at local minima.

Suffice to say that the choice of clustering algorithm depends very much on the dataset and cannot be applied blindly.

Chapter 3

Membrane Recognition & binding by PIP5K1A

"One never notices what has been done; one can only see what remains to be done."

Marie Curie

3.1 Introduction

The interaction of membrane lipids and proteins plays a key role in signaling within and between cells in multicellular organisms. The cell membrane comprises multiple lipid species of varying chemical composition that are subject to spatiotemporal regulation [55] for cell signaling. A wide range of cellular functions, including signaling and trafficking, are dependent on lipid-dependent binding of peripheral membrane proteins to the plasma membrane [7]. Thus, peripheral membrane proteins may carry out e.g., lipid modifications, activation of small GTPases, or co-localization and recruitment with interacting partner proteins. The binding targets of peripheral membrane pro-

teins are often anionic phospholipids [180], with considerable variety in the structure and binding mechanisms which enables control of their interactions.

Phosphatidylinositol phosphates (PIPs) are one of the major targets for directing binding of peripheral membrane proteins [181]. They constitute a small fraction of the lipid membrane but are essential for cell signaling [182]. For example, processes regulated by phosphatidylinositol-4,5-bisphosphate (PIP_2) include endocytosis, exocytosis, cell motility, cell adhesion, and signal transduction [13]. Furthermore it is becoming clear that several ion channels [183] and receptors [184] are regulated by PIP_2. It is important that these processes are tightly regulated, and therefore it is unsurprising that lipid-metabolizing enzymes are of biomedical interest. For example, the 3-phosphatase PTEN enzyme, which catalyzes the hydrolysis of phosphatidylinositol-3,4,5-trisphosphate ($PI(3,4,5)P_3$ or PIP_3) to form PIP_2, is mutated in many cancers [185].

PIP kinases regulate the activation of various signaling pathways by catalyzing the formation of PIP_2 lipids from various lipid substrates [23]. There are three known families of PIP kinases categorized as types I, II, and III. Type I and type II PIP kinases phosphorylate phosphatidylinositol-4-phosphate (PI4P) and phosphatidylinositol-5-phosphate, respectively, to form PIP_2. For example, PIP5K1A catalyzes the phosphorylation of PI4P to form PIP_2, thus playing an important role in many cell signaling processes. PIP5K1A is also implicated in e.g., prostate cancer [27], and the PIP5K1 family of kinases is regulated by interactions with other signaling proteins such as Talin [186], Rac1 [187], ARF6 [188], and ARF1 [189].

The crystal structure of PIP5K1A (Fig.3.1A) was initially determined at 3.3 Å resolution [25]. PIP5K1A participates in Wnt signaling, and increased levels of the kinase contribute to increased cancer cell invasion, proliferation, and survival [190]. Comparisons with the structure of homologous kinase PIP4K2B (PDB: 1BO1) [191]

and other kinases [26] have provided insights into likely mechanisms of catalysis and selectivity of the PIPKs.

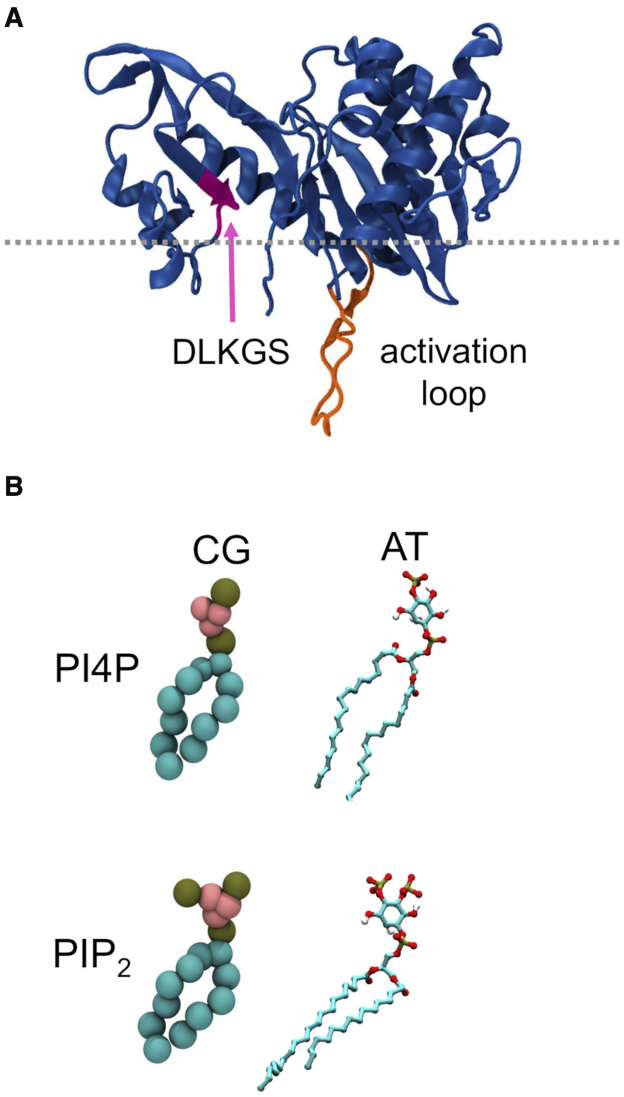


Fig. 3.1: PIP5K1A Protein, Substrate and Product. (A) Structure of the PIP5K1A monomer (PDB: 4TZ7) with the DLKGS motif (which contributes a key lysine K238 to the active site) and the unstructured activation loop (PDB residues 378–415) labeled. The secondary structural features of the backbone atoms are in blue, the DLKGS motif in magenta, and the activation loop is in orange. The grey dotted line indicates the approximate position of the lipid membrane as revealed by subsequent simulations. (B) Coarse-grain (CG) and atomistic (AT) representations of the phosphatidylinositol phosphates PI4P and PIP_2 (the substrate and product respectively of PIP5K1A). The tan CG particles represent phosphate groups, the pink particles represent the inositol ring, and the cyan particles represent the lipid tails. The coarse-grain representations are created using the MARTINI [139] force field, which uses a 4:1 mapping of non-hydrogen atoms on to coarse-grained particles. In the AT representation, tan particles represent phosphorus atoms, red represent oxygen atoms, cyan represent carbon atoms, and white represent hydrogen atoms.

The structure of the zebrafish PIP5K1A catalytic domain reveals a fold similar to that of protein kinases with a PIPK-specific subdomain that contains a "DLKGS" sequence motif. The extended DLKGS motif (DLK238GSxxxR244) is thought to form a "PIP-binding motif" [26]. There is also an unstructured activation loop (also seen in PIP4K2B) that is also thought to play a role in PIP binding and selectivity. In the crystal, PIP5K1A is seen to form a side-to-side dimer, and biophysical and biochemical evidence suggest that this dimer forms in solution and is required for full catalytic activity resolution [25].

The crystal structure of PIP5K1A did not directly reveal any substrate or lipid interactions. An NMR study [192] has suggested a role of the activation loop as a molecular sensor for lipid interactions. Molecular dynamics (MD) simulations, which have been widely used to explore the interactions of both integral [193] and peripheral proteins [194] with membranes and their lipids, provide an opportunity to explore the dynamic interactions of PIPK with a PIP-containing lipid bilayer. In particular, we use multiscale MD simulations to reveal the interactions of PIP5K1A with PI4P in a bilayer, and the nature of the interactions of the PIPK dimer with the membrane surface.

3.2 Methods

Simulation Details

A table of simulations performed in this chapter is presented below:

Simulation	Bilayer[a]	Duration (μs)	N
CG			
PIP5K1A/PC	PC	1	5
PIP5K1A/PC/PS	PC/PS	1	5
PIP5K1A/PC/PS/PI4P	PC/PS/PI4P	2	25
PIP5K1A/PC/PS/PIP_2	PC/PS/PIP_2	1	5
PIP5K1A/PC/PS/PI4P/PIP_2	PC/PS/PI4P/PIP_2	1	5
PIP5K1A/PC-dimer	PC	1	5
PIP5K1A/PC/PS-dimer	PC/PS	1	5
PIP5K1A/PC/PS/PI4P-dimer	PC/PS/PI4P	1–3[b]	2 and 3[b]
PIP5K1A/PC/PS/PI4P-dimer-flat	PC/PS/PI4P	1	5
PIP5K1A/PC/PS/PI4P-dimer-NMA	PC/PS/PI4P	1	5
PIP5K1A/PC/PS/PI4P-dimer-restrained-1PIP	PC/PS/PI4P	0.1	3
PIP5K1A/PC/PS/PI4P-dimer-restrained-2PIP	PC/PS/PI4P	0.1	3
PIP5K1A/PC/PS/PI4P-dimer-restrained-3PIP	PC/PS/PI4P	0.1	3
PIP5K1A/PC/PS/PI4P-dimer-ATconf	PC/PS/PI4P	1	3
AT			
PIP5K1A/PC/PS/PI4P-dimer	PC/PS/PI4P	0.10, 0.15, and 1.00	1 each

Table 3.1: Table of large-scale simulations in simple symmetric composition bilayers. [a] Bilayer composition: 75% PC, 20% PS, and 5% PI4P (80%, 20% for PC/PS bilayers). [b] Of the five initial simulations of 1 μs duration, three were extended to 3 μs.

CG simulations were performed using the MARTINI 2.1 force field [123, 139] with a 20 fs time step. Particle coordinates were written out every 0.5 ns. Coulombic interactions were shifted to zero between 0 and 1.2 nm. Lennard-Jones interactions were shifted to zero between 0.9 and 1.2 nm. The nearest neighbour list was updated every 10 steps. A Berendsen thermostat [131] (coupling constant 1ps) and barostat (coupling constant 1 ps, compressibility 5x10^{-6} bar^{-1}) were used to maintain temperature at 323 K and pressure at 1 bar. The LINCS algorithm [111] was used to constrain bond lengths.

Atomistic simulations were performed using the GROMOS 43a1 [195] and 53a6 [121] force fields with a 2 fs time step. Particle coordinates were written out every 20 ps. Lennard-Jones interactions were shifted to zero between 0.9 and 1.2 nm. Long-range electrostatic interactions were treated using the particle-mesh Ewald method (PME) using default parameters pme-order = 4 and ewald-rtol = 10^{-5}, fourierspacing = 0.12. PME was shifted from 0 to 1 nm [129]. The nearest neighbour list was updated every

10 steps. A V-rescale thermostat (coupling constant 1ps) and Parrinello-Rahman [135] barostat (coupling constant 1ps, compressibility 5x10-6 bar-1) were used to maintain the temperature and pressure. The LINCS algorithm was used to constrain bond lengths.

Modelling

The crystal structure of PIP5K1A was acquired from the Protein Data Bank (PDB: 4TZ7). Both were deposited as dimers so for monomer simulations the PDB file was amended to include only the first monomer. There are four regions missing from the PIP5K1A structure: the initial sequence of residues (residues 1-55), a short turn (residues 154-156) between the 3 and 4, the insert (residues 310-356), and the activation loop (residues 386-401). For this investigation only the short turn and the activation loop were modelled in the structure, using MODELLER 9.9 [196]. The same regions were also removed in the dimer structures. In the numbering scheme used in figures, simulation residues 1 to 253 correspond to PDB structure residues 57 to 309. Simulation residues 254 to 324 correspond to PDB structure residues 356 to 426.

Membrane Binding Simulations

Coarse-grained simulations of PIP5K1A monomers were performed using a 7.5 x 7.5 nm^2 area bilayer of the relevant membrane composition. Lipids were exchanged into a preformed 100% POPC bilayer using a local script [197]. The final lipid percentages were 75% POPC, 20% POPS, and 5% PIP. The kinase centre of mass was positioned 8 nm away from the bilayer centre of mass in a random orientation. The box (7 x 7 x 20 nm^3) was solvated and sodium and chloride ions added to a concentration of 0.15 M. The dimer simulations were performed in a similar manner but with a 15 x

15 x 24 nm3 bilayer. CG simulations were performed using GROMACS 4.6.5 [143]. Energy minimisation was carried out via steepest descent and the system equilibrated for 5 ns with protein backbone particles restrained. For each model, 25 repeats were performed with different initial velocities.

Atomistic simulations were performed by obtaining a snapshot from CG simulation and conversion via CG2AT, a fragment-based approach [145]. AT-MD simulations carried out using the GROMACS 5.1 software: the system was equilibrated for 1.5 ns with the backbone atoms of the protein restrained, then a production run of 100-1000 ns carried out.

Analysis

VMD [159] was used for simulation visualisation. Graphs were generated in matplotlib [155]. For membrane binding, The GROMACS g_dist command was used to obtain the distance between the protein centre of mass and the bilayer centre of mass over time. For protein-membrane contacts, The GROMACS g_mindist command was first used to obtain a list of distances of every residue to the nearest PIP phosphate group in each simulation frame. The total number of contacts was counted and normalised to the residue with the largest number of contacts, defining a 'contact' as a minimum distance of less than 0.7 nm (0.35 nm in the atomistic simulations). For distance-orientation analysis, a reference 'productively bound' structure was selected from simulations in which the kinase catalytic site faces the bilayer headgroups. Frames from each simulations trajectory were fitted in the xy plane (parallel to the membrane surface). The rotation matrix of these xy-aligned frames relative to the reference structure was then obtained using the GROMACS g_rotmat tool, and the R_{ZZ} component, giving rotation from the z axis (the membrane normal), was recorded. Distances were obtained using the GROMACS g_dist tool, taking the z component of

distance between the protein and lipid centers-of-mass. Corresponding distance and R_{ZZ} values were plotted using a locally written script. Progress of the membrane association simulations was monitored as the distance of the kinase centre of mass from the bilayer centre vs. time. The distance was averaged over all simulations in the ensemble (for ensemble sizes N = 5 or 25, see Table 3.1). Exponential fits to binding curves were performed using the Scipy curve fit function.

3.3 Results & Discussion

3.3.1 Binding of PIP5K1A to a Substrate-Containing Lipid Bilayer

CG simulations were performed to explore the mode of interaction of the PIP5K1A monomer with a lipid bilayer. The protein was initially positioned away from the bilayer (Figure 2A) and multiple replicates (N = 5–25) of a 2 μs simulation performed (Table 3.1, see methods) in which the protein diffused within the aqueous phase before encountering and binding to the bilayer (Figures 3.2B and 3.3). We performed multiple shorter simulations rather than a single long simulation, as recent studies [164] suggest that this may provide more effective sampling.

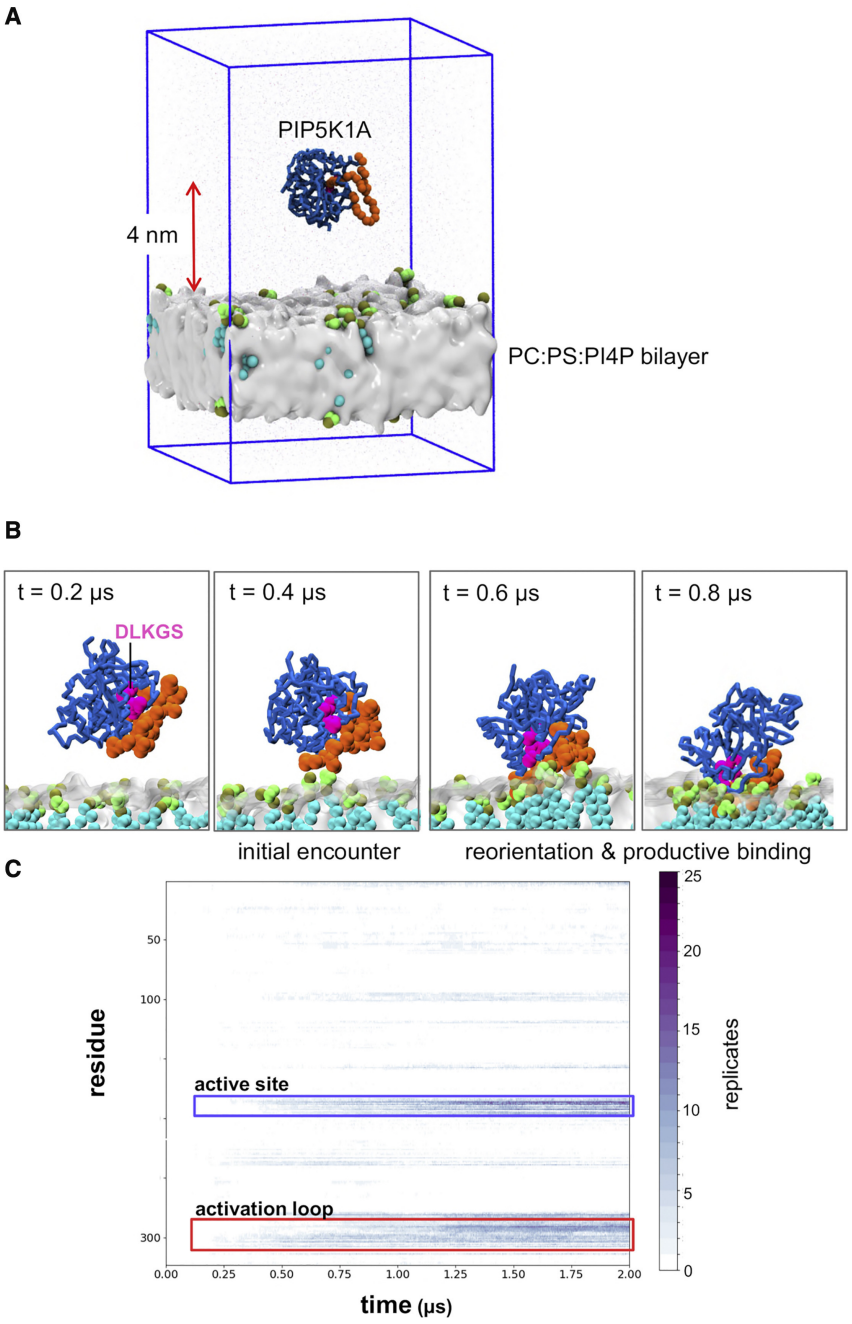


Fig. 3.2: PIP5K1A/Membrane Encounter Simulations. A: The initial simulation setup. The CG model of the PIP5K1A monomer was placed such that its center of mass was 4 nm away from the surface of the lipid bilayer. The activation loop of the kinase is in orange. The blue box represents the periodic boundaries. The lipid bilayer (PC and PS) is shown in gray with PI4P molecules in green and cyan. The blue box represents the boundaries of the simulation box. B: An example of the progress of a CG simulation (PIP5K1A/PC/PS/PI4P; Table 1) in which the PIP5K1A molecule, initially displaced from the membrane, first encounters the bilayer. The PIP5K1A monomer first encounters the membrane, making contacts with PI4P molecules via the activation loop (orange). The PIP5K1A molecule then reorients so that both the activation loop and the DLGKS motif (pink) contact the PI4P molecules (green/cyan). C: Time course of protein contacts to PI4P from the PIP5K1A/PC/PS/PI4P (Table 1) simulation The activation loop (red box) makes initial contact with the membrane and is followed by reorientation and "productive binding" of DLKGS motif (which is close to the active site; blue box).

Simulations were performed with a bilayer containing the substrate (PI4P) as well as control simulations in which either a zwitterionic (PC) or anionic (PC/PS) bilayer without any PIP molecules present was used (see Table 1). In these simulations the protein is able to tumble in solution before encountering the membrane (Fig.3.3. In the PC/PS/PI4P simulations the protein encountered the bilayer typically within the first 0.5 μs before then reorienting and binding in a "productive" mode in which the DLKGS motif and the activation loop interacted with PI4P molecules at the membrane surface. In contrast, PIP5K1A did not bind to PC only or to anionic PC/PS membranes (Fig.3.3). Instead, in the PC/PS control simulations the kinase tumbles randomly in solution making only occasionally transient contacts with PS molecules. Recent studies of, e.g., Ras family proteins, suggest that lipid specificity is encoded by defined structures rather than just electrostatic interactions [71], and that specific interactions with PIP_2 molecules in the membrane mediate G-Ras function [198]. This supports the view that the activation loop conformation is the main determinant of lipid specificity [15] and that this effect is captured by the coarse-grained model (Fig.3.3).

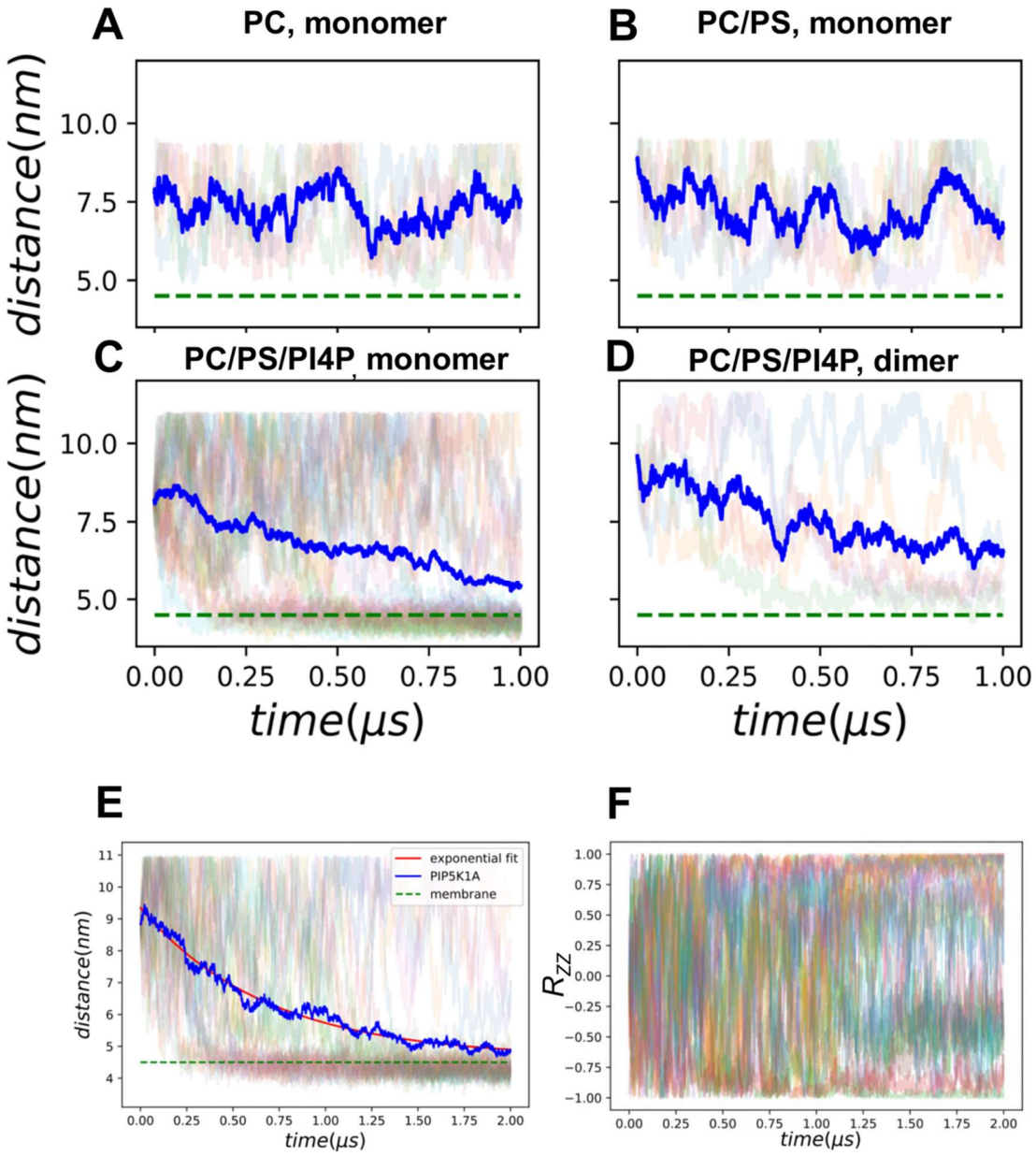


Fig. 3.3: PIP5K1A/Membrane encounter controls. Progress of the membrane association simulations shown as the distance of the kinase centre of mass from the bilayer centre vs. time, averaged over all simulations in the ensemble (see Table 3.1), for A: PIP5K1A (monomer)/PC; B: PIP5K1A (monomer)/PC/PS; C: PIP5K1A (monomer)/PC/PS/PI4P; and D: PIP5K1A (dimer)/PC/PS/PI4P simulations. The broken horizontal green line indicates the distance at which the kinase would be bound to the bilayer surface. The pale coloured lines represent the distance vs. time data for the individual simulations within each ensemble. E: PIP5K1A (monomer)/PC/PS/PI4P, 2μs simulations. The average distance of the kinase centre of mass was fitted with a single exponential model (red curve) with a decay time of 0.7 μs. F: R_{zz} of the individual PIP5K1A monomers over time (shown via a different coloured line for each monomer), demonstrating that the protein tumbles in solution before interacting with the bilayer. The R_{ZZ} component describes the zz component of the rotational matrix of the protein and describes the rotation from the z axis (the membrane normal).

The PC:PS:PI4P simulations were analyzed in terms of the contacts formed to PI4P molecules by the (monomeric) kinase (Fig.3.4A.), and of the orientation of the protein molecule relative to the bilayer to establish how the kinase interacts with the membrane.

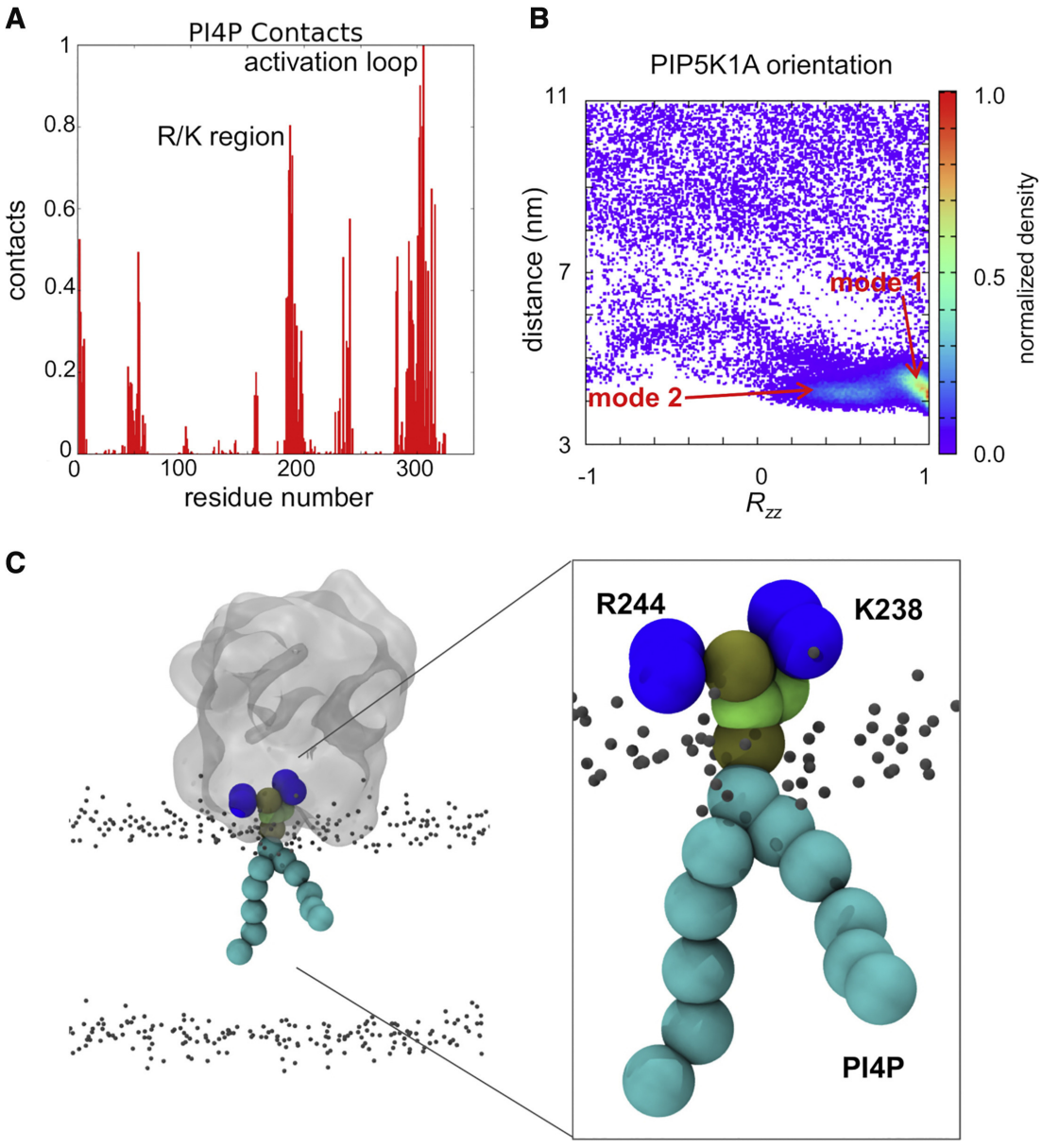


Fig. 3.4: Membrane Contacts of PIP5K1A. A: Contacts to PI4P for the membrane bound monomeric kinase. The normalized number of PI4P contacts averaged across all PIP5K1A/PC/PS/PI4P (Table 1) CG simulations is shown as a function of simulation residue number. (Simulation residues 1–253 correspond to PDB structure residues 57–309 and simulation residues 254–324 correspond to PDB structure residues 356–426.) Contacts are concentrated in two regions, an arginine/lysine-rich region correspond to the β8-α4c loop, which follows the DLKGS motif (PDB residues 243–253; simulation residues 187–197) and the activation loop (PDB residues 378–415; simulation residues 276–313). B: Density map of the distance from an orientation (given as the R_{zz} of the orientation matrix; see main text for details) relative to the bilayer of the kinase averaged over the five PIP5K1A/PC/PS/PI4P CG simulations. Two major modes of interactions are seen: mode 1 is further from the membrane, with the activation loop interacting with PIP molecules, whereas in mode 2 the kinase monomer is bound to the membrane with the DLKGS motif and the active site facing toward the surface of the bilayer. C: Snapshot from the end of a PIP5K1A/PC/PS/PI4P simulation showing the protein (grey; K238 and R244 in blue) bound (mode 2) to a PI4P molecule in a lipid bilayer (indicated by the phosphate particles shown as small gray spheres). The inset shows the PI4P head group (tan and green particles) positioned between residues K238 and R244 (in blue).

The PIP5K1A molecule made contacts with PI4P via two main regions: an arginine / lysine-rich region corresponding to the β8-α4c loop which is immediately after the DLKGS motif (PDB residues 243–253; which are simulation residues 187–197 in Fig.3.4) and the activation loop (PDB residues 378–415, simulation residues 276–313). A distance/orientation density map (averaged over the 25 repeat simulations - Figs.3.4B and 3.5) revealed the orientation of the kinase as a function of distance from the membrane. Fig.3.5. shows the controls orientations maps and demonstrates that this binding mode is not seen in membranes lacking PIPs.

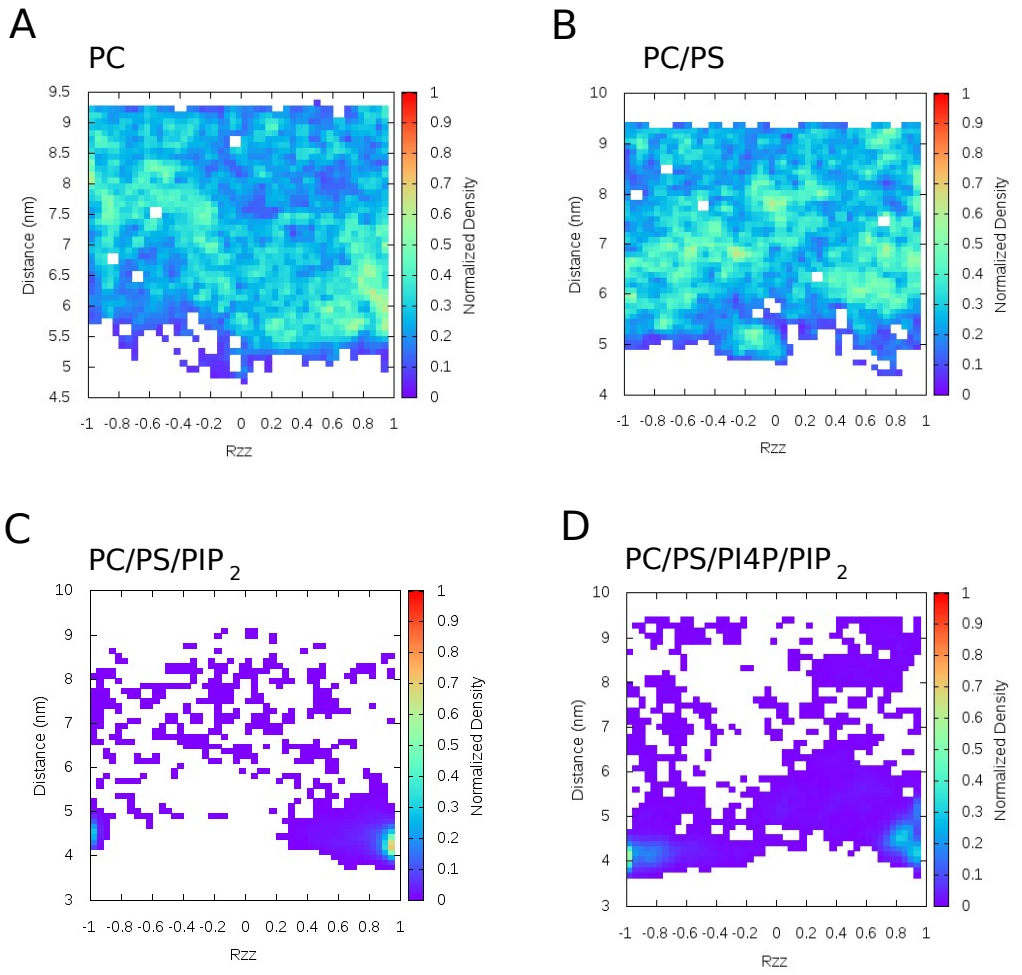


Fig. 3.5: Orientation density maps. A: Orientation density map for PC control; B: Orientation density map for PC/PS; C: Orientation density map for PC/PS/PIP_2; D: Orientation density map for PC/PS/PIP_2/PI4P mixed. Both control simulations (i.e. A and B) show a random distribution of orientations, and therefore no specific orientations with respect to the bilayer

Two modes of interaction were seen in the interaction with PI4P. In mode 1 the activation loop formed the main interaction with the membrane. In mode 2 (seen following reorientation of the protein on the membrane surface), contacts are also formed by the arginine/lysine-rich region noted above. This yields a model of the PIP5K1A binding in a preferred orientation for catalysis with the active site and

nearby residues (including D236, K238, and R244; simulation residues 180, 182, and 188) in close contact with lipid head groups.

The activation loop is a primary determinant of substrate specificity [15]. Our simulations suggest that the activation loop (which is absent from the crystal structure and so was modeled as a flexible region) plays an important role in the binding of the kinase. It has been suggested that there is an interplay between the phosphate binding site of the substrate of the phosphorylation site (residues K238 and R244) and the activation loop [26]. Our simulations indicate that K238 and R244 form major contacts to PI4P and that the activation loop retains its dynamic behavior throughout the binding process, interacting with multiple PIP molecules. Thus our results support the suggestion that the DLK238GSxxxR244 sequence (conserved among the PIPKs) corresponds to a PIP-binding motif [26].

It has also been proposed that PIP_2, the product of the phosphorylation reaction, may compete for interaction with the activation loop and therefore act as inherent feedback mechanism in the kinases [15]. In simulations of the interaction of PIP5K1A with a PIP_2-containing bilayer (Figure S3A; Table 1), we observe that it does bind to PIP_2, with a similar overall distribution of contacts. PI4P showed a slight preference for interacting with the activation loop over the active site, whereas PIP_2 showed a slight preference for the active site over the activation loop when comparing normalized numbers of contacts. The contacts made by the PIP_2 are shifted toward the N-terminus of the activation loop compared with the PI4P interactions. This suggests there may be a complex interplay of substrate and product determining the interactions and conformation of the activation loop at the surface of the bilayer, which could in principle be explored in simulations of PIP5K1A bound to membranes with both PI4P and PIP_2 present.

3.3.2 Dimer Binding

Having established a detailed model of recognition of a PI4P-containing bilayer by monomeric kinase we extended our simulations to explore the mode of binding of the dimer observed in the crystal structure of PIP5K1A. It is clear from biophysical and biochemical data that the side-by-side dimer forms in solution and that dimerization is required for full catalytic activity [25]. However, simulations of the X-ray structure of the dimer (PIP5K1A/PC/PS/PI4P-dimer; see Table 3.1; Fig.3.7A for the structure) showed that the dimer makes contacts with PI4P molecules in the membrane through either one or other of the two active sites (as defined by the DLKGS motif) on a similar timescale to that observed in comparable simulations of the monomeric kinase (Fig.3.2C and D) but not through both sites of the dimer at the same time.

To test if this was due to a limitation of the simulation duration we extended the simulations to 3 μs (Table 1). However, binding of both active sites to the membrane at the same time was not observed in any of these extended simulations. An example contact profile is shown in fig.3.6.

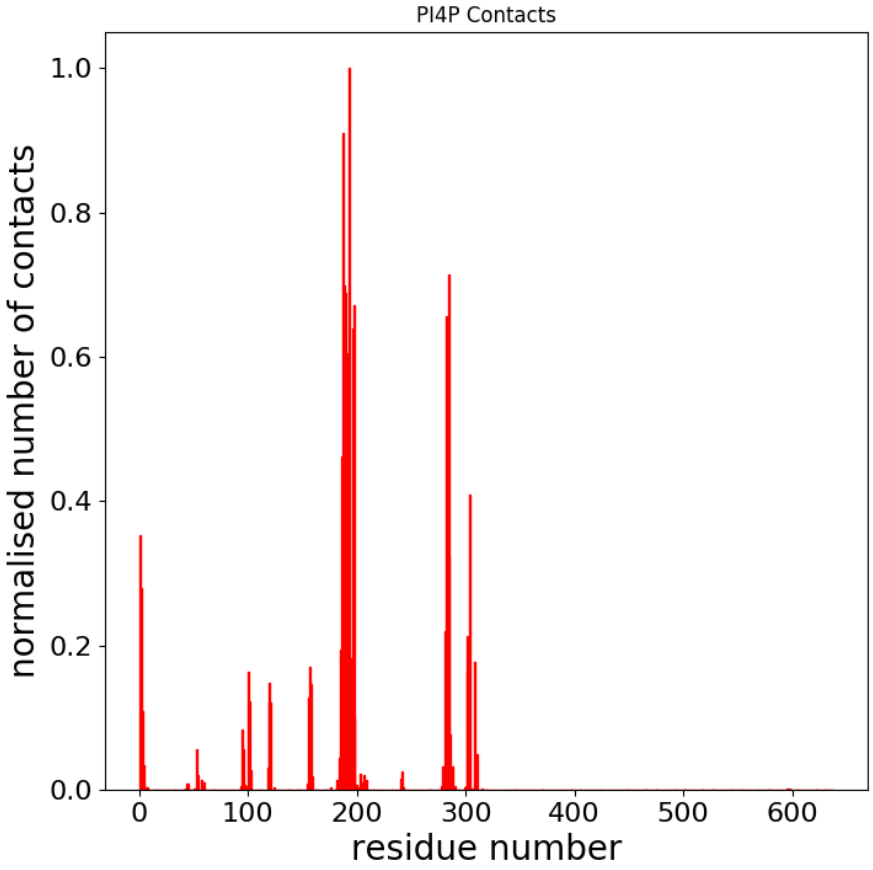


Fig. 3.6: PIP5K1A dimer-PI4P contact profile. Contacts between PIP5K1A residues and PI4P lipids. One subunit dominates the contact profile.

Simulations were therefore performed with the dimer pre-positioned flat (with active sites at equal distance from the membrane surface) and just 1 nm above the bilayer (PIP5K1A/PC/PS/PI4P-dimer-flat; Table 1). In these simulations one active site quickly (within 0.1 μs) formed contacts with PI4P in the membrane, but simultaneous PI4P binding by both sites of the dimer was not observed. We reasoned this may be due to a need for some degree of flexibility in the dimer greater than that observed in the CG simulations (in which an elastic network is imposed to preserve the tertiary, but not quaternary, structure of the protein). We thus used normal mode analy-

sis (via the elNémo server [199]) to estimate the low-frequency normal modes of the dimer. From the models thus generated we selected the most "flattened" conformation (fig.3.7A) of the dimer and repeated our CG simulations of interactions with a bilayer using this as a starting model (simulation PIP5K1A/PC/PS/PI4P-dimer-NMA; see Table 1).

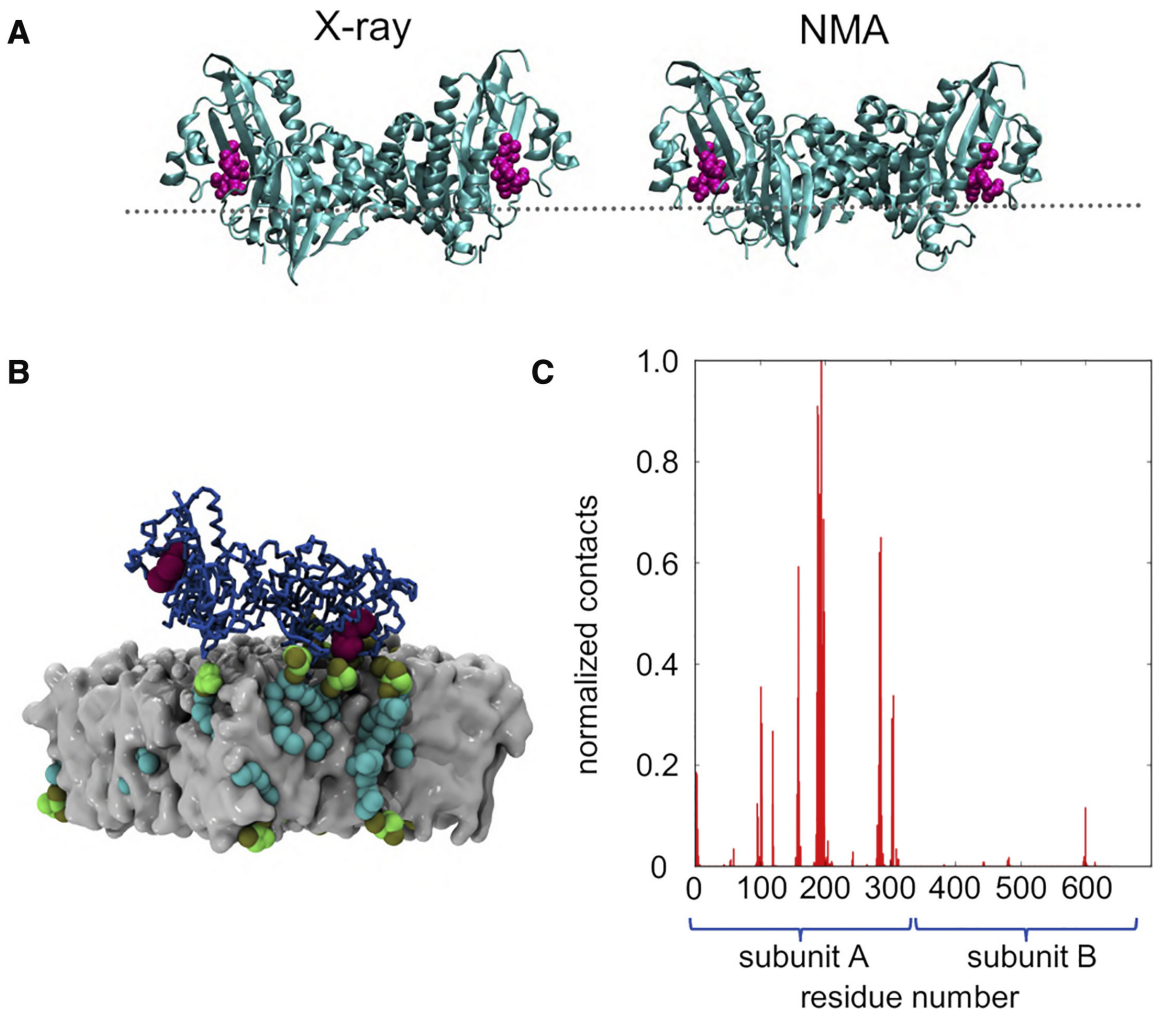


Fig. 3.7: PIP5K1A dimer at a membrane surface. A: PIP5K1A dimer, comparing the original PDB structure (X-ray) and a model based on a low-frequency mode produced by normal modes analysis (NMA). The DLKGS motif is in pink and the approximate location of the bilayer surface (as defined by lipid head groups) is shown as a broken gray line. B: Snapshot from the PIP5K1A/PC/PS/PI4P-dimer-NMA simulation of the interaction of the dimer with a PI4P-containing membrane. It can be seen that only one subunit of the dimer forms extensive contacts to the PI4P molecules and the bilayer. While the activation loop of the unbound monomer makes some negligible contacts with PIPs in the membrane, the dimer remains bound at one site only. C: Normalized number of contacts between PIP5K1A NMA dimer residues and PI4P particles.

In this case only two out of five simulations resulted in the dimer forming extended contacts with the membrane, and only one of these bound to the membrane for an extended period (fig.3.7B). In this case, only one active site bound to a PI4P molecule

in the membrane, although minimal and transient contacts made with PI4P by residues of the other monomer (fig.3.7C). These contacts of the second subunit were in the activation loop, which has sufficient flexibility and length to make occasional contacts with the membrane.

To further test whether the dimer could bind to the membrane while interacting with two PI4P molecules, one bound to each monomer, we explored an initial configuration in which this was "forced" to be the case. Thus, a series of simulations (PIP5K1A/PC/PS/PI4P-dimer-restrained-1PIP, -2PIP, and -3PIP; see Table 1) were performed in which the dimer was initially positioned in a "flat" configuration on the membrane and in which distance restraints were imposed between the lysine residue of the DLKGS motif of the PI4P binding site (see above) and one, two, or three adjacent PI4P molecules in the membrane (fig.3.8).

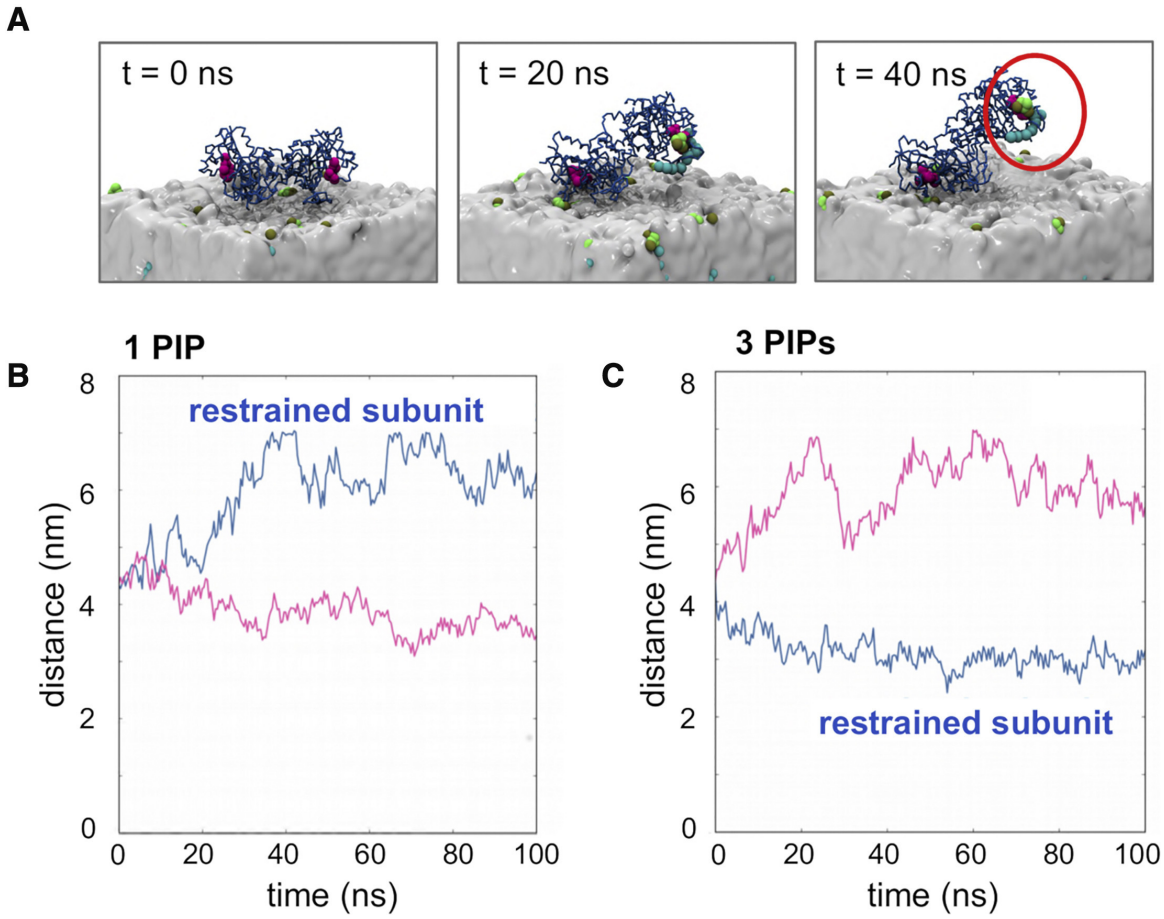


Fig. 3.8: PIP5K1A Dimer Interactions with PI4P at a Membrane Surface. (A) The PIP5K1A dimer binds to the PI4P membrane at either of the two active sites, but not both simultaneously, such that contacts are made at one site exclusively. The snapshots show progression of a PIP5K1A/PC/PS/PI4P-dimer-restrained-1PIP (Table 1) simulation performed with the dimer initially positioned flat on the surface of the membrane and restraints placed between the active site lysine residue (PDB residue 328; simulation residue 182) and a single PI4P molecule (the restraint was between the COMs of the residue and the PI4P head group COM). Further simulations were performed with corresponding restraints to 2 or 3 PI4P molecules (see Table 1). The restrained lipid (in cyan and green) is extracted from the membrane.

In none of these simulations did the dimer adopt a stable configuration in which both subunits remained at the membrane surface. Thus, in the PIP5K1A/PC/PS/PI4P-dimer-restrained-1PIP simulations the unrestrained subunit remained at the membrane surface, and the restrained subunit moved away, while extracting the PI4P

molecule out of the bilayer. In the PIP5K1A/PC/PS/PI4P-dimer-restrained-3PIP simulations, the restrained subunit remained bound to the membrane, while the unrestrained subunit dissociated from the bilayer surface (Figure 3.8B). Together, the various simulations may suggest that the dimer of PIP5K1A is unlikely to bind to a membrane via two PI4P molecules, one bound to each active site.

3.3.3 The bound dimer

To explore in more detail the interactions of the bound dimer, an atomistic model of the bound dimer was generated from the endpoint of the first replicate of the PIP5K1A-PC/PS/PI4P CG simulation. This was used to initiate three short atomistic simulations (0.1, 0.15, and 0.25 μs) of the bound dimer, of which the 0.25 μs simulation was extended to 1.0 μs.

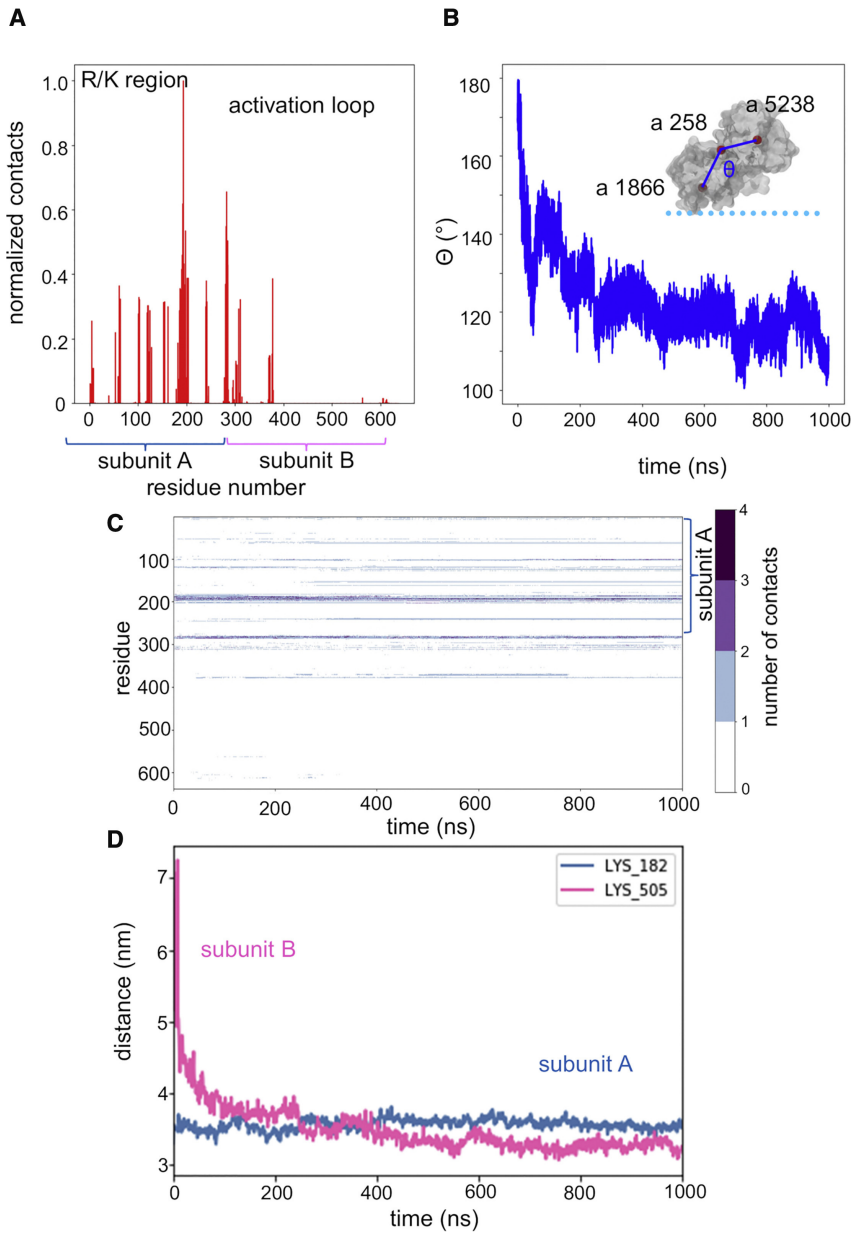


Fig. 3.9: Atomistic Simulations of the membrane bound-PIP5K1A Dimer. A: Normalized number of contacts between PIP5K1A residues and PI4P molecules. Distribution of contacts remains similar. Notably in this replicate, the middle of the activation loop has folded up out of the membrane with the end residues making contacts with PI4P. B: Angle metric (θ) describes the between two subunits in the dimer over time. The inset shows a schematic of the definition of θ (defined by atoms 1866, 358, 5238 in residues K182, V29, K505). The angle decreases from 180° at the start of simulation (corresponding closely to the crystallographic dimer) to 140° where it remains for most of the simulation with transient periods around 110°. C: Contacts as defined by a 0.35 nm cutoff between the residue COM and the PI4P head group COM show PIPs collecting near the active site (DLKGS, simulation residues 180–184), while PIP contacts with the activation loop (simulation residues 280–300) remain fairly constant. With the exception of a small number of contacts made with PI4P by the activation loop in subunit B at the start of the simulation, there are no major contacts made by subunit B with PI4P. D: The dimer was observed to bend at the dimerization interface site to bring the second monomer 4 nm closer to the membrane as shown by subunit B moving from a distance of 7 nm away from the membrane COM to 3 nm in the second half of the simulation. Taken with the angle metric (B), this reveals that the dimer tilts bringing subunit B to the membrane COM, but does not form contacts with PI4P.

The simulations confirmed the dimer conformation as making contacts with the membrane at one site only (Figure 3.9A). The pattern of PI4P contacts for the "bound" subunit was as in the CG simulations, with key contacts formed by the DLK238GSxxxR244 motif and by the activation loop. Dynamic fluctuations are seen in these contacts (Figure 3.9C), especially for the activation loop, which shows fluctuations in its interaction with PI4Ps. On the longer (μs) timescale the dimer tilts toward the membrane bringing subunit B into closer contact with the membrane COM, but this does not result in substantive contacts with PI4P. In support of this we determined the number of PI4P molecules within 3.5 nm of the two active sites of the dimer (noting that the approximate radius of the monomer is 2.5 nm).

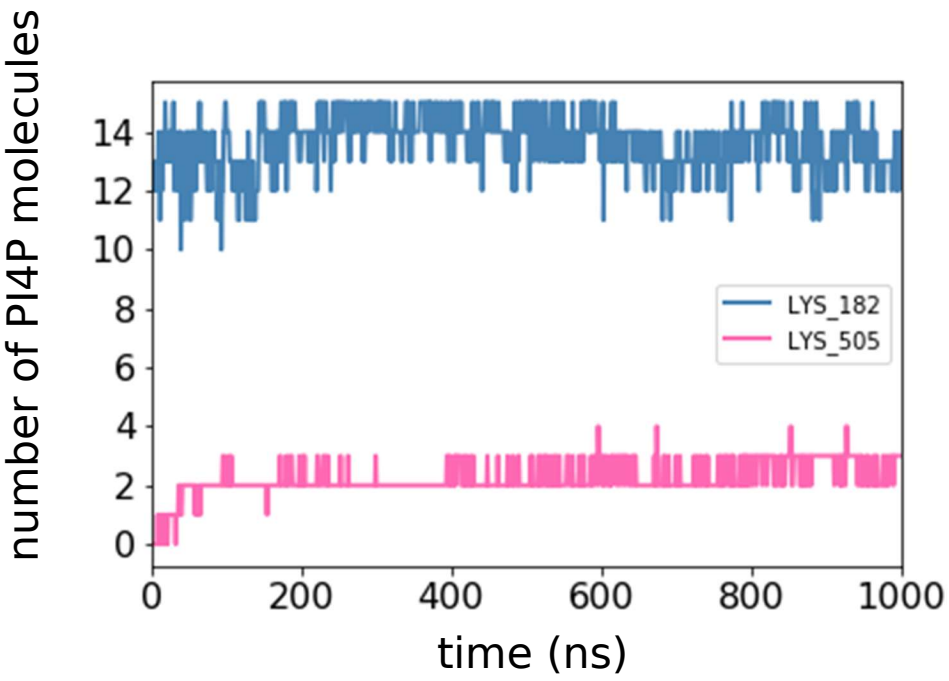


Fig. 3.10: PIPs around the bound dimer. Number of PI4P molecules within 3.5 nm of the active site catalytic residue over the course of the 1 atomistic simulation of the kinase dimer (PIP5K1A/PC/PS/PI4P-dimer). There are 12-14 PI4Ps clustered around the bound monomer, and around 2-4 PIPs near the active site of the other (unbound) monomer

While there are more PI4P molecules near the bound monomer we also see that

there are around 3–5 PI4Ps within 3.5 nm of the active site of the unbound monomer. This number remains fairly constant throughout the simulation. So, it seems that, while PI4P molecules are "within range" of the unbound monomer they do not interact directly with its active site. Thus, as in the CG simulations, substantive interactions with PI4P were only seen for one subunit of the dimer.

The atomistic simulations indicated a degree of flexibility in the PIP5K1A dimer: the dimer underwent hinge-bending motion at the subunit:subunit interface as demonstrated by tracking the Cα root-mean-square deviations of the individual subunits versus the dimer as a whole and via a simple angle metric (fig.3.9B). This allowed both subunits to approach the membrane (fig.3.9D), although binding of PI4P to the second subunit was not observed on the timescale of the atomistic simulations (fig.3.9C). Thus, over the time course of the simulation (fig.3.9D), the kinase tilts such that the unbound active site moves closer to the membrane, whereas the PI4P-bound monomer moves a little away from the membrane. This suggests that the dimer may adopt a flattened conformation on the bilayer but that this state is transient. This is seen toward the end of the 1.0 μs simulation, in which the dimer fluctuates between 100° and 120°. However, no contacts with the PI4P lipids are observed for subunit B (fig.3.9C), even though several PI4P molecules were within range of the subunit (fig.3.8). The flattened conformation from the atomistic simulation was converted to CG and run for 3 $\times$ 1 μs to see if PIPs could bind at the catalytic site at the same time (fig.3.10).

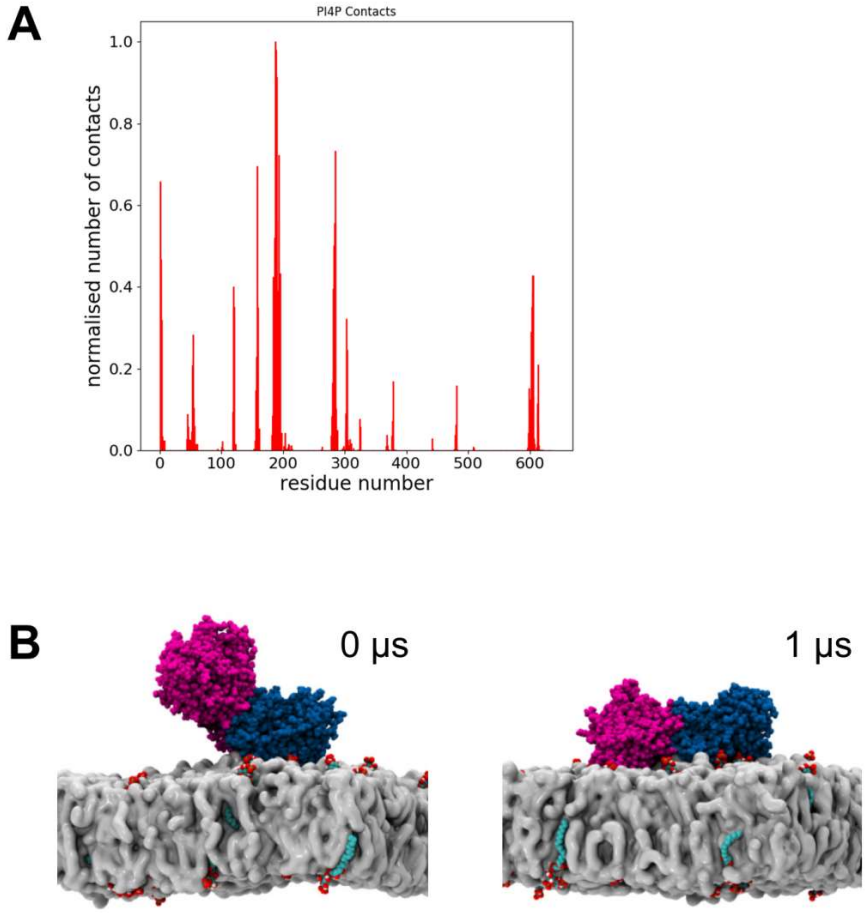


Fig. 3.11: Bound dimer controls. A: An atomistic PIP5K1A kinase dimer + bilayer system from the AT PIP5K1A/PC/PS/PI4P-dimer simulation (see Table 1) was converted to CG and run for 3 x 1 μs (simulation PIP5K1A/PC/PS/PI4P-dimer-ATconf) to determine whether PI4P could bind to both catalytic sites at the same time. B: Snapshots of the tilted (0 μs) and flattened orientation of the dimer at the bilayer surface from the AT PIP5K1A/PC/PS/PI4P-dimer simulation.

The bound monomer still dominated the contact profile, although some contacts were observed in the activation loop region of the unbound monomer. However, there were still no contacts made with the catalytic lysine residue of the unbound monomer throughout the simulation. The catalytic K238 residue makes between zero

and two hydrogen bonds with PI4P molecules intermittently throughout the simulation (fig.3.12).

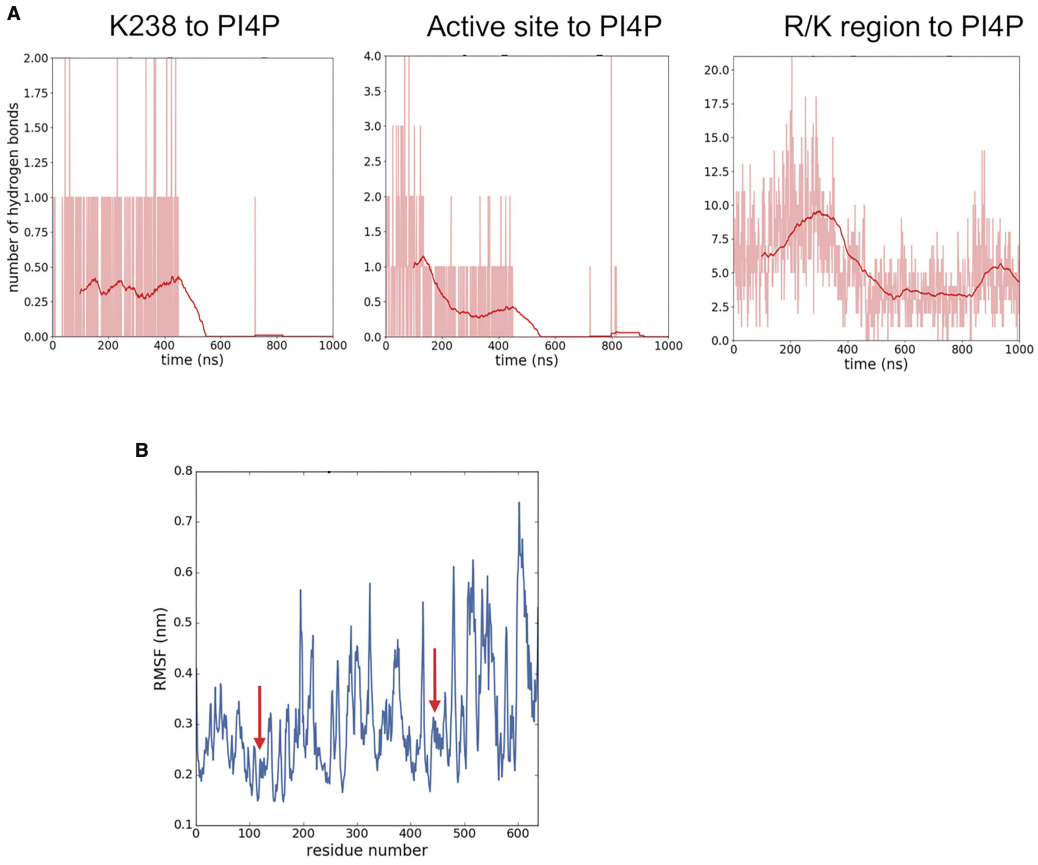


Fig. 3.12: Hydrogen bonding and RMSF in the PIP5K1A dimer system. A: Hydrogen bonding interactions to PI4P for an atomistic simulation of the PIP5K1A dimer bound to a PC/PS/PI4P membrane (see Table 1). Hydrogen bonding (shown as instantaneous numbers in pink and as moving averages in red) for a PIP5K1A dimer bound to a PC/PS/PI4P membrane simulated over 1 μs. The K238 residue makes between 0 and 2 hydrogen bonds with PIPs in the membrane. Other residues in the active site also contribute to hydrogen bonding. Residues in the R/K-rich region are involved in an extensive hydrogen bonding network which shows small fluctuations over time but remains broadly between 5 and 10 hydrogen bonds over the course of the simulation while the dimer is bending. B: RMSF as a function of residues number for a PIP5K1A dimer bound to a PC/PS/PI4P membrane. The arrows show the regions corresponding to the C-helix that show reduced mobility.

In addition, up to six hydrogen bonds are made between active site residues and PI4P molecules. The R/K-rich region, adjacent to the catalytic site, makes multiple

hydrogen bonds with PI4P molecules. In the second half of the simulation during which the dimer tilts, a concurrent loss of hydrogen bonding between PI4P molecules and the active site of subunit A is observed. In addition, the root-mean-square fluctuation (RMSF) of the Cα atoms was investigated, revealing that the regions with the lowest and least variable RMSF (0.2–0.25 nm) correspond to the regulatory C-helices. Experimental evidence suggests that membrane or substrate binding does not explain increased dimeric enzyme activity. The C-helix is an element known to be crucial for kinase regulation and it has been postulated that its tight packing in the dimer interface results in conformational changes in the monomer that result in increased kinase activity [25]. A lower level of structural variation in these regions (fig.3.12) of an otherwise dynamic protein would be consistent with this region of the structure being important for regulation and it is striking that this is revealed in the simulations.

Dimerization has been proposed [25] to increase the binding affinity of the kinase to the membrane by doubling the basic surface area that is able to make contacts with the acidic lipid head groups in the membrane, in addition to creating a flat surface that is complementary to the plane of the membrane. It has been shown in liposome flotation assays that the monomeric and dimeric binding is similar, and therefore that dimerization does not cause higher catalytic activity through membrane binding at both sites, but rather increases activity by keeping the kinase in a productive orientation. From the simulation results it would appear that the dimeric PIP5K1A kinase does not bind lipid substrates concurrently at both sites, which is consistent with an explanation that is based around key structural features keeping the kinase in a productive orientation as indicated by the RMSF profiles of the atomistic dimer simulations.

3.4 Conclusions

Our simulations provide a dynamic model of the productive orientation of PIP5K1A at the surface of a PI4P-containing lipid bilayer. One monomer (fig.3.13) is bound to a substrate molecule with further PI4P molecules clustered about the "footprint" of the dimer on the membrane.

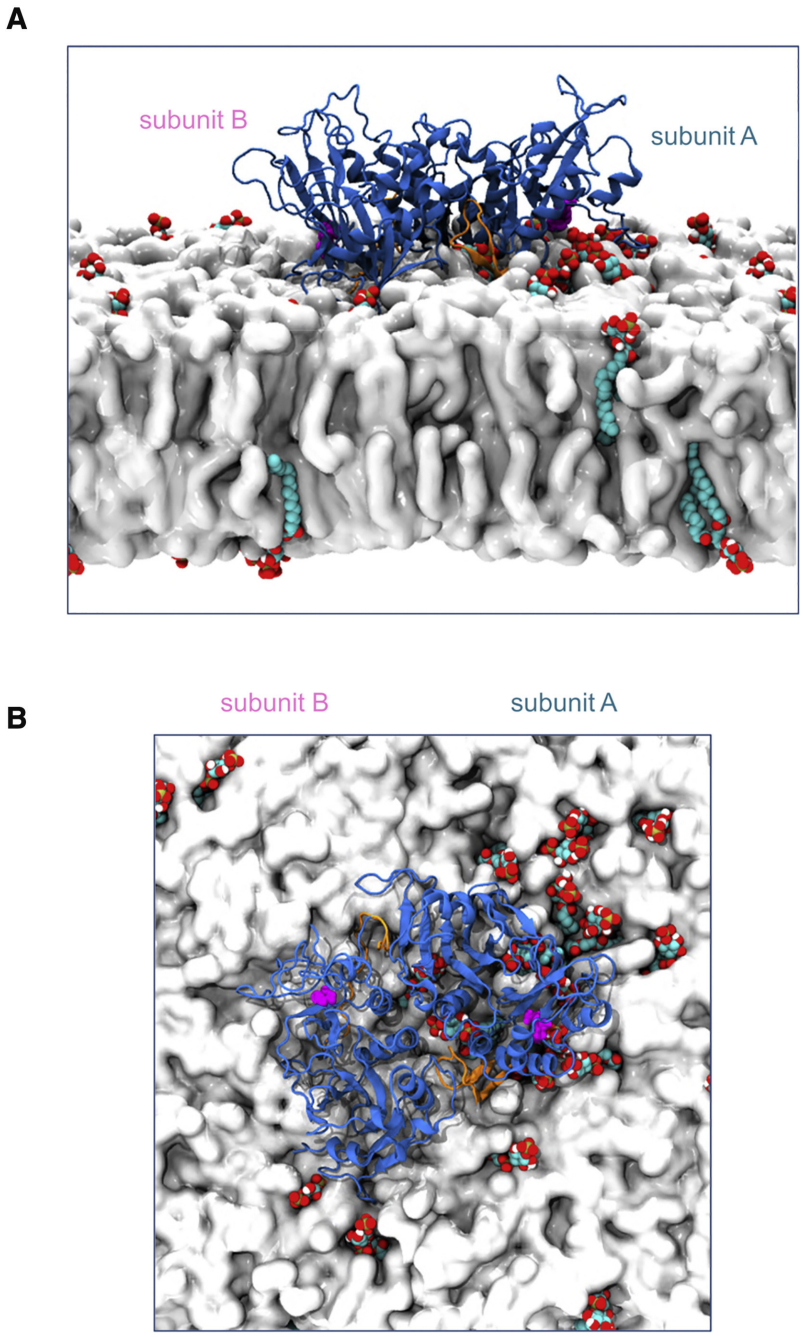


Fig. 3.13: Final Configuration of the Membrane-Bound Dimer. A: Snapshot of the dimer bound to the membrane after 1 μs of atomistic simulation. The kinase is shown in blue with the catalytic motif highlighted in magenta and the activation loops highlighted in orange. PI4P lipids are shown in red, white, and cyan and PC/PS lipids are shown in white for clarity. A number of PI4P molecules can be seen clustering around the bound monomer's catalytic site, whereas the unbound monomer remains close to the membrane but not bound to any PI4P molecules. B: Snapshot of the bound dimer viewed from above. Multiple PIPs are seen clustered around the bound subunit A with no PI4P molecules in contact with subunit B.

It is clear from our analysis of the interactions of PIP5K1A with the bilayer that: (1) the DLK238GSxxxR244 sequence (the PIP-binding motif of Muftuoglu *et al.*, 2016) and the activation loop form the main contacts of the enzyme with PI4P molecules in the membrane; (2) the K238 and R244 residues form the binding site for the 4-phosphate of PI4P; and (3) although the dimeric protein can sit more or less flat on the bilayer surface, it can only interact tightly with one PI4P molecule at a time.

While we may speculate that substrate binding occurs one site at a time, our studies that show no simultaneous binding across any simulations. It is quite possible that on longer (i.e., >1 μs timescales) one monomer could detach from the membrane and the dimer (stochastically) switch to binding via the other monomer. This could in turn result in a piston-like mechanism whereby dissociation of one subunit allows substrate to diffuse in/product to diffuse out. Such a mechanism might be advantageous if nano clusters of PI4P become locally exhausted due to enzyme activity. Furthermore, while the dimeric protein does not appear to bind planar membranes at both sites simultaneously, binding of the dimer to curved membranes might enable both active sites to interact. This will be investigated further in future simulations of large membranes with more complex lipid compositions (and multiple copies of bound dimer) which can undergo dynamic fluctuations in curvature [197, 200].

Questions remain around the dynamic behavior of the activation loop. An NMR study [192] of an isolated peptide commensurate with the PIP5K1A activation loop indicated that the unstructured region folded into an amphipathic helix in DHPC micelles and provided evidence that the activation loop functions as a membrane sensor. However, this study did not address the structure of the activation loop in a more complex membrane lipid composition, and/or within the large protein structure bound to a membrane (although this is addressed in chapter 4).

Limitations of the current study may include the convergence of the dimer sim-

ulations, and whether or not enough time is allowed for the flattening of the dimer onto the bilayer. However, given that, in the simulations, the dimer flattens and then relaxes, we consider that the limit of this flattening has been reached. In the 1 μs atomistic simulation this is observed twice, with no concurrent binding of PI4P to the catalytic site of the second subunit observed. Taken together with the CG simulations, the results of our multiscale simulation approach suggest that the dimeric protein cannot bind PI4P molecules at both active sites, while the lipids are in a planar membrane and the protein adopts the crystallographic dimer conformation. For the dimer to bind PI4P at both sites at the same time, our results suggest that this would require either significant distortion of the bilayer and/or a significant change in the dimer interface from the crystal structure. This is relevant to our more general understanding of the interactions of complex peripheral proteins with membranes, revealing the need to take account dynamic of protein/membrane interactions.

Finally, while the MARTINI2 model has been suggested to be too "sticky" to model protein-protein interactions (see [201, 202] vs. [203], there is a growing body of evidence that suggests simulated protein-lipid interactions reproduce experimental results reasonably well, both within ([193, 204–206] and at the surface [83, 207] of membranes. There is scope for future studies of atomistic free energy calculations to compare the relative strengths of PI4P and PIP_2 binding and also the contribution of residues in the activation loop to specificity. As noted above, it will also be of interest simulate larger coarse-grained membranes that are more complex (i.e., *in vivo* mimetic) in their lipid composition in order to investigate the influence of cholesterol and other lipids on dynamic bilayer fluctuations and PIP5K1A interactions. In order to ensure that these calculations produce representative data, it is imperative that more and better systematic data collection is carried out for experimental PMP systems. This will enable improved parameterisation of force fields for PMPs, and will enable better

comparison to experimental data.

Overall, we have shown that PIP5K1A can bind to the cell membrane as either a monomer or a dimer. PIP5K1A binds in a productive orientation for catalysis, such that the activation loop binds to PIP molecules and aids in the orientation of the kinase. The lipid head groups are therefore positioned close to the catalytic site for phosphorylation. Atomistic simulations reveal details of the catalytic conformation of the dimer for catalysis and the role of individual residues. In particular, we have shown that the activation loop of the PIP5K1A kinase leads the recognition of and binding to the membrane, and that a simple model of simultaneous binding of the dimer catalytic sites to the membrane is not adequate to explain how the dimeric kinase increases rates of catalysis. In the future, it could be of interest to extend these studies to bilayers that model the asymmetric lipid composition between the two leaflets of the bilayer [208], and at different concentrations corresponding to pre- and post-stimulus PIP levels in order to capture some of the complexity of signaling events.

Chapter 4

Markov State Models of the PIP5K1A Activation Loop

"The worst thing you can do is to completely solve a problem."

Daniel Kleitman

This chapter comprises work following on from chapter 3. I investigate the activation loop of PIP5K1A in greater detail and apply Markov state model (MSM) analysis to atomistic simulations of the initial binding event between PIP5K1A and a model membrane with the aim of understanding how the activation loop confers specificity.

4.1 Introduction

Peripheral membrane proteins bind to the cell membrane and modify its physical properties to convey intracellular messages. Many of these proteins contain regions of disorder that are important for their specific function, but their dynamic behaviour is difficult to capture on experimental timescales and characterise experimentally. Type I phosphatidylinositol phosphate kinase A (PIP5K1A) is one such example and catalyses

lipid phosphorylation in the cell membrane. PIP5K1A contains a disordered activation loop which confers lipid specificity (fig.4.1) [15].

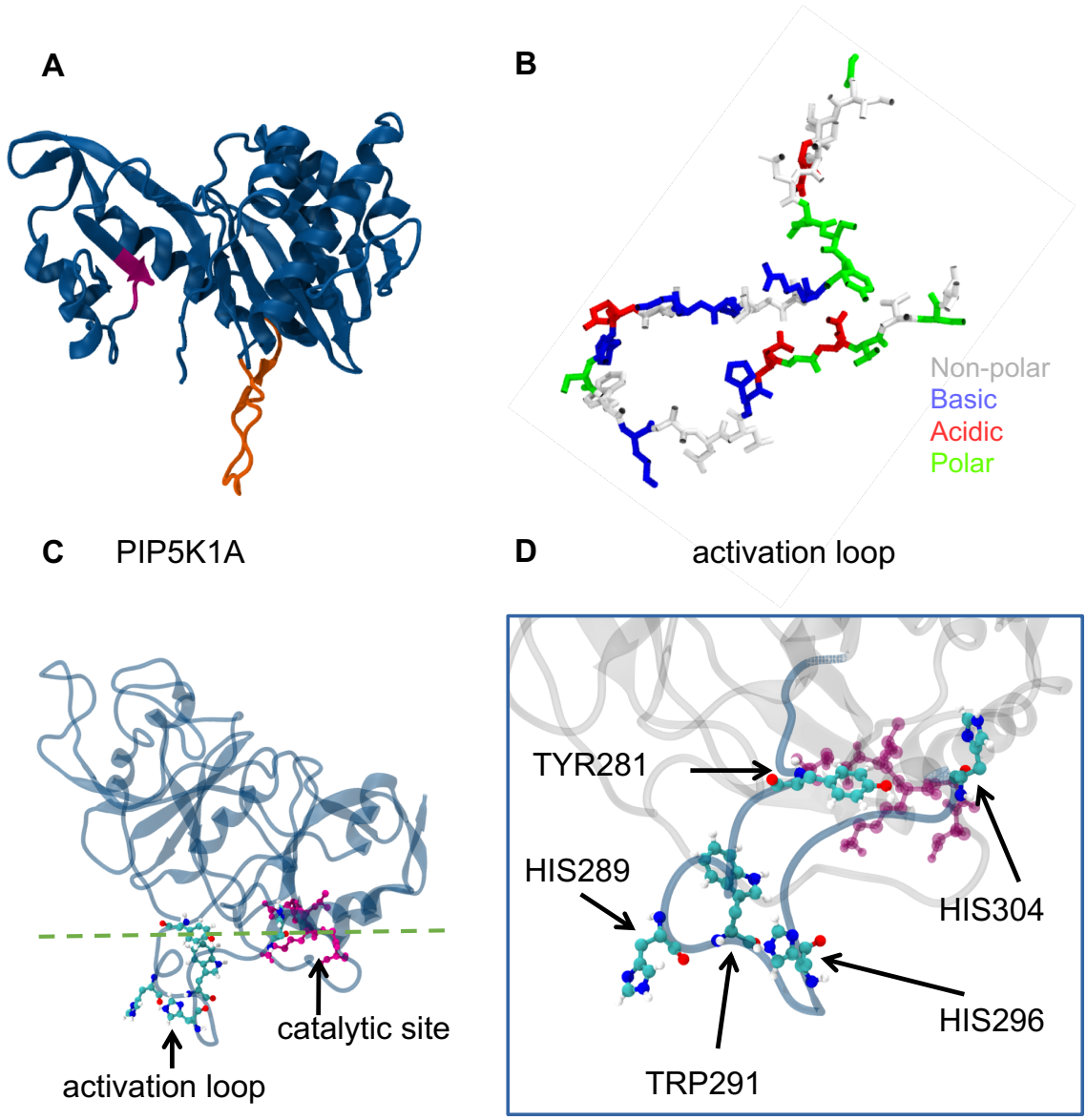


Fig. 4.1: Structure of the PIP5K1A kinase monomer. A) Secondary structure representation of the crystal structure (PDB: 4TZ7) of PIP5K1A with catalytic site residues DLKGS in magenta. B) Representation of the activation loop with non-polar, basic, acidic, and polar residues highlighted. C) Representation of the kinase with aromatic residues on the loop highlighted and the approximate position of phosphate headgroups of the membrane shown as a green dashed line. Active site residues shown in magenta. D) Aromatic residues of the activation loop highlighted and the active site residues shown in magenta.

The study of conformational changes in proteins by MD simulations remains challenging due to the timescales on which transitions occur. These can be in the region of microseconds; for large systems this is currently at the cutting edge of resources. It is often not feasible to simulate for long enough to capture these events or to sample the conformational landscape well enough to be confident in attaining convergence. MSMs can address these problems by either extracting information pertinent to longer timescales than the constituent simulations or by guiding new simulations by measuring the areas of greatest variance and 'reseeding' simulations for further runs in these regions, known as adaptive sampling [209]. Ultimately the generation of an MSM aligns with that of the wider aim of computational simulation; that is to accurately make quantitative predictions that are experimentally useful and relevant, and shed light where experimental methods reach their limitations [210].

In this study I use the disordered activation loop of the phosphatidylinositol phosphate-5-kinase A (PIP5K1A) to investigate how a disordered region [192] may exert its functional effects through structural conformations. PIP5K1A binds to phosphatidyl inositol phosphate (PIP) lipids in the membrane in order to catalyse their phosphorylation [211]; this in turn modifies the physical properties of the membrane and therefore enables the propagation of signals within the cell. A recent NMR study of the PIP5K1A activation loop suggested that the loop forms an α-helix structure in the membrane; however this study was carried out with 1,2-dihexanoyl-sn-glycero-3-phosphocholine (DHPC) micelles which is not the physiological environment for the protein [192]. The study by Liu *et al.* used an isolated activation loop peptide of 21 residues. The loop adopted two rapidly exchanging conformations; unstructured and partially folded, and they postulate that the hydrophobic surface of this amphipathic helix is essential for membrane sensing. This signalling is upstream of many important signalling cascades [212]. Errors in this signal propagation results in aberrant

cell processes such as cell growth, propagation, and death. Consequently, PIP5K1A mutations are strongly implicated in colon and breast cancers [211]. More broadly, there is a tendency to ignore disordered regions in structural biology as they are less amenable to various experimental and computational techniques [213]. I carried out this work to see if MSM methods could be used to investigate these systems and eventually be used to uncover why mutations cause aberrant signalling from a structural perspective.

4.2 Methods

Atomistic simulations were performed using the GROMOS 53a6 force field [122] with a 2 fs time step. Particle coordinates were written out every 20 ps. Lennard-Jones interactions were shifted to zero between 0.9 and 1.2 nm. Long-range electrostatic interactions were treated using the particle-mesh Ewald method (PME) [129]. The nearest neighbour list was updated every 10 steps. A V-rescale thermostat (coupling constant 1 ps) and Parrinello-Rahman [135] barostat (coupling constant 1 ps, compressibility 5x10^{-6} bar^{-1}) were used to maintain the temperature at 323 K and 1 atm pressure. The LINCS algorithm was used to constrain bond lengths[111].

Modelling

The crystal structure of PIP5K1A was acquired from the Protein Data Bank (PDB: 4TZ7). The structure was deposited as a dimer so for monomer simulations the PDB file was amended to include only the first monomer. There are four regions missing from the PIP5K1A structure: the initial sequence of residues (residues 1-55), a short turn (residues 154-156) between the β3 and β4 elecments, the insert (residues 310-356), and the activation loop (residues 386-401). For this investigation only the short

turn and the activation loop were modelled in the structure, using MODELLER [196]. In the numbering scheme used in figures, simulation residues 1 to 253 correspond to PDB structure residues 57 to 309. Simulation residues 254 to 324 correspond to PDB structure residues 356 to 426.

Membrane binding simulations

Atomistic bound simulations were performed by obtaining a snapshot from CG simulations in chapter 3, chosen from simulations where the kinase is in its productive bound orientation with PI4P in contact with the catalytic site. Conversion was carried out at the point of initial contact with CG2AT [145]. AT-MD simulations were carried out using the GROMACS 5.1 software: the system was equilibrated for 1.5 ns with the backbone atoms of the protein restrained (force constant = 1000) to enable the membrane to equilibrate without interference from the protein, then production runs of 5 x 100 ns carried out.

Atomistic initial binding simulations were performed by obtaining a snapshot from CG simulations in chapter 3, chosen from simulations where the kinase makes contact with the membrane. Three initial confirmations were chosen and conversion was carried out at the point of initial contact with CG2AT. Initial contact was defined as the first frame where an activation loop residue was observed to be in contact with a phosphate headgroup of PI4P by visual inspection. Simulations were run as in the paragraph above, for a total of 7.5 μs aggregate simulation.

Analysis

VMD [159] was used for simulation visualisation. Graphs were generated in matplotlib and R [178]. Analysis and Markov state model protocols were written using the MSMBuilder [157] and MSMExplorer [100] packages in Python. MSM parameters are

described in the text. Secondary structure analysis was carried out with DSSP [214].

4.3 Results & Discussion

Firstly I consider the fully bound PIP5K1A monomer in contact with PIPs in the membrane. These simulations arose from previous work in chapter 3 with coarse-grained simulations that show PIP5K1A binding to the membrane in accordance with X-ray crystallography data and provide the background and rationale for Markov state modelling [215]. In these simulations, previous coarse-grained systems were converted to atomistic representations and simulated for 5x100 ns. The results are presented in figs.4.1-4.

Then I consider the initial binding event of the activation loop of the PIP5K1A and its analysis by Markov state modelling. The PIP5K1A protein was simulated from its initial point of contact between the activation loop and the membrane (75% POPC, 20% POPS, 5% PI4P). Three initial conformations (fig.4.5) were derived from previous coarse-grained simulations. Five replicates of each 500 ns simulation were performed. This resulted in an aggregate simulation time of 7.5 μs. These results are presented in figs.4.5-13.

4.3.1 The activation loop in the membrane

The PIP5K1A was simulated in its bound state in order to see where PIP lipids bind and to see the structure of the disordered activation loop (fig.4.2).

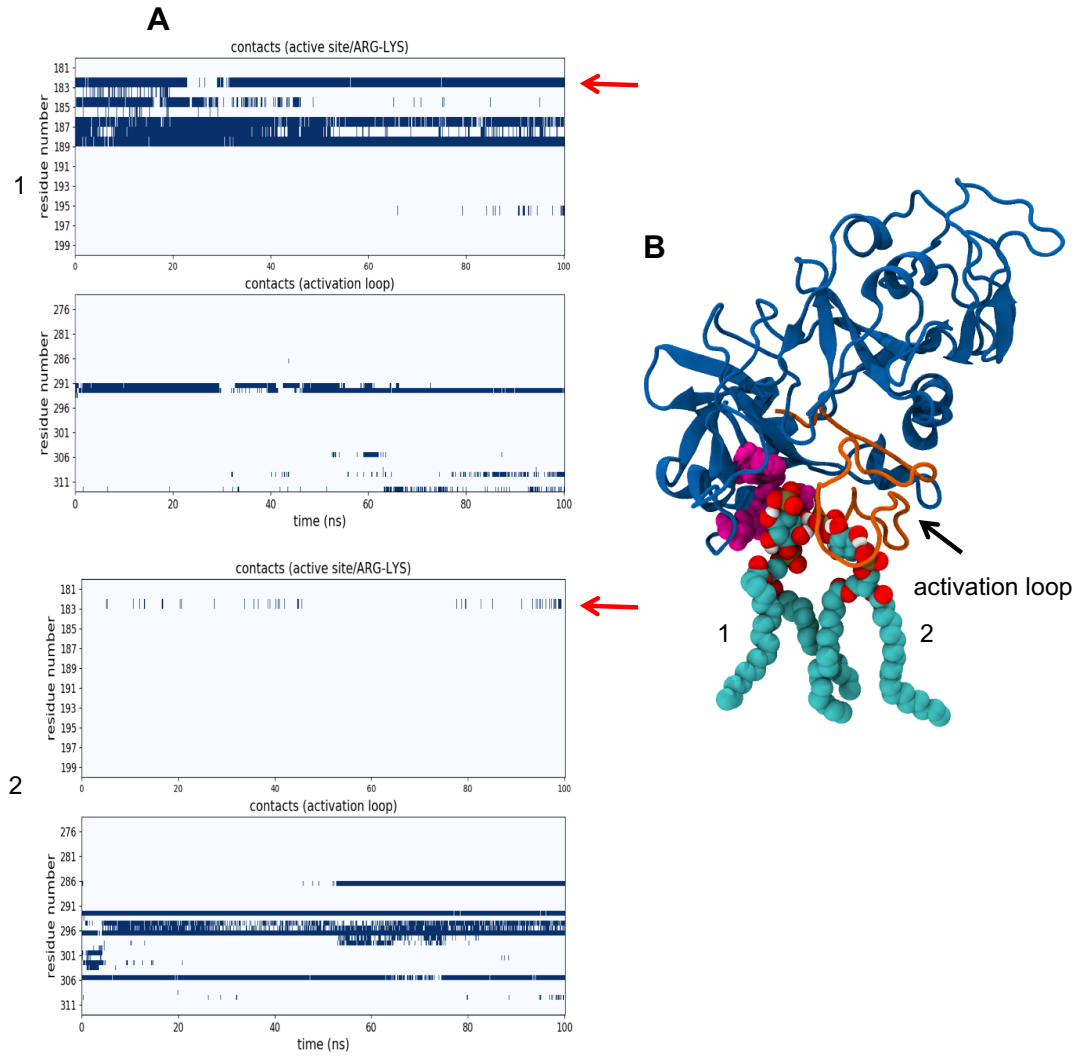


Fig. 4.2: PIP5K1A-PI4P contacts. A) Lipid-residue contacts over time for lipids 1 and 2. Red arrows indicate contacts between PIP phosphate head groups and the catalytic residue 182. B) Graphical representation of the lipid residue interactions. PIP lipids are shown in green and red, the catalytic site in magenta, the activation loop in orange, and the PIP5K1A structure in blue.

The PIP5K1A was found to bind PI4P in the catalytic site (fig.4.2A) with contacts being made between residues 180-200 and PIP lipids. In this representative trajectory, contacts were observed for $> 90\%$ of simulation time with simulation residue 182 which is proposed to catalyse the phosphorylation of the PIP phosphate head group. Both lipids 1 and 2 also exhibit contacts made with the activation loop whilst bound; the

activation loop contacts exhibit some constant residue contacts and others that are intermittent or formed and broken over the course of the trajectory. This indicates that the loop residues may have dynamic structural behaviour in the membrane. The loop was also consistently observed to have a structure which curves around lipids in the membrane (fig.4.2B) whilst they occupy the catalytic site.

In order to investigate the activation loop structure more closely, the conformation over time was clustered as a function of root mean squared deviation (RMSD) vs the distance between residues TYR281-HIS296 (fig.4.3A). The choice for this type of clustering arose from NMR studies of p53 [216], another disordered protein, where conformations were distinguished by clustering the RMSD vs. distances between aromatic residues in order to observe conformational switching.

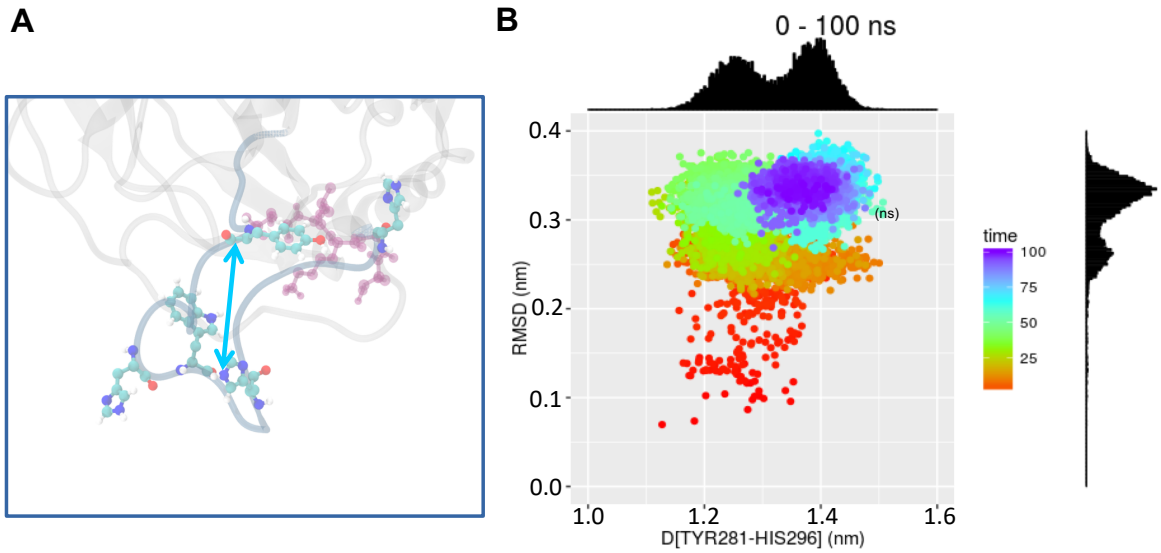


Fig. 4.3: Clustering conformations of the activation loop. A) Graphical representation of the activation loop. The distance between Tyr281 and His296 is indicated with a blue arrow. B) Plot showing Cα RMSD of the activation loop vs Distance [TYR281-HIS296] over time. Marginal histograms indicate density.

The clustering indicates that the loop can be clustered into different conformations based on a relatively simple metric (fig.4.3B). Some structural change in the first 25 ns is to be expected due to limitations of equilibration simulations, but the density

profiles show small changes that are distinct in the latter 75 ns. This was observed consistently across simulations for different distance pairings (data not shown).

I also investigated the penetration into the bilayer by the activation loop. Fig.4.4 shows residue Cα-membrane COM distances and demonstrates which residues penetrate most into the membrane.

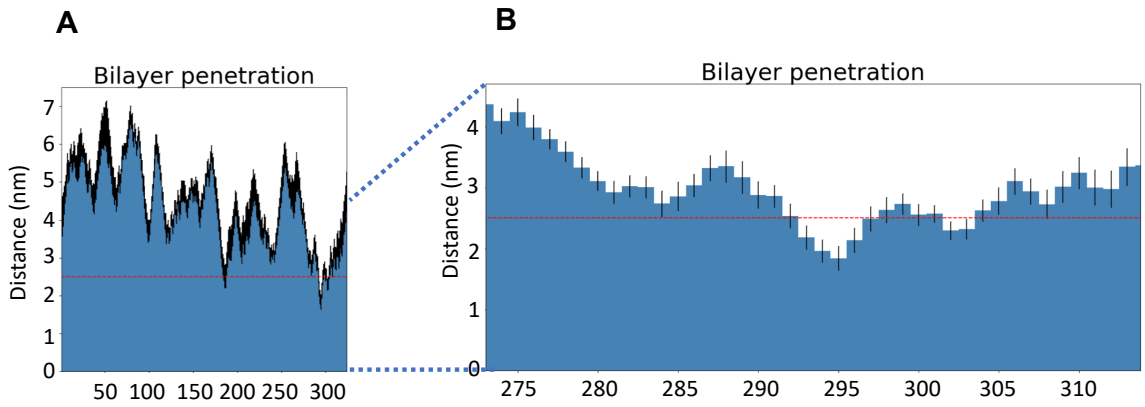


Fig. 4.4: Penetration of activation loop residues into the membrane. Red line indicates where an amino acid is level with the phosphate head groups. Black lines indicate standard error. A) Distance between residue and membrane centre of mass for all PIP5K1A residues in one representative trajectory. B) Distance between residue and membrane centre of mass for activation loop residues only. note that the region has been presented on a smaller axis for clarity.

Fig.4.4A. shows that the regions closest to the membrane are the catalytic site (residues 180-184) and the activation loop. Fig.4.4B. shows the activation loop residues positioned in the membrane (residue 292-304). Residues LYS292, ALA293, LEU294, LEU295, 296HIS, ASP297, SER302 and VAL303 penetrate most deeply into the membrane.

Given that the activation loop has been postulated to undergo subtle conformational changes that enable function [192], these results provided a first indication that a structural study of the loop was warranted.

4.3.2 Lipid contacts show the activation loop leads binding and makes diverse lipid contacts

In order to simulate and analyse the initial binding event of the activation loop, 3 different initial conformations (fig.4.5) were selected from previous coarse-grained simulations, converted to an atomistic representation, and simulated for 5 x 500 ns, leading to 7.5 μs of total simulation time. The trajectories were analysed using the coordinates of the 21 residues of the loop (residues 286-306) that make the most contacts with lipids (see fig.4.2B).

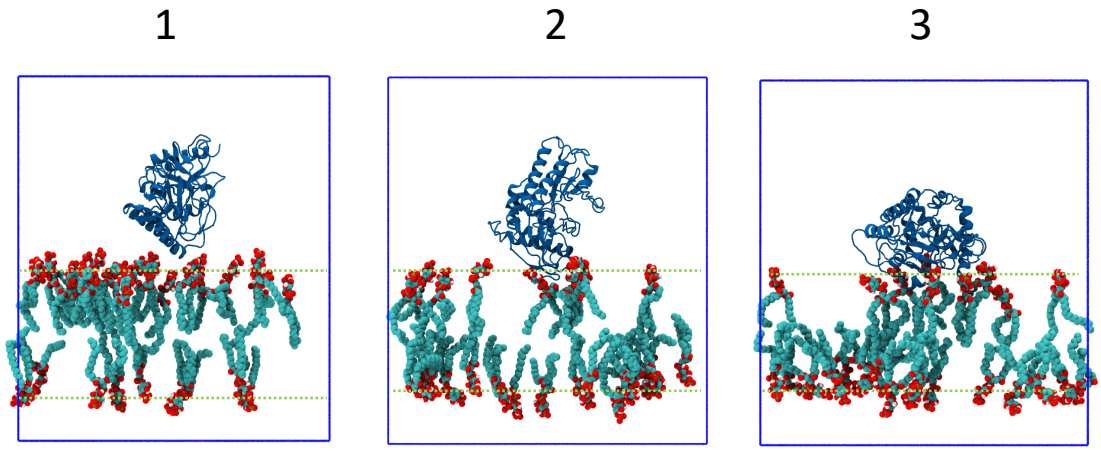


Fig. 4.5: Graphical representation of the three starting structures for simulation. Start conformations were derived from coarse-grained simulations and converted to an atomistic representation at the point of initial contact with PIPs in the membrane. Solvent, ions, and bulk lipid omitted for clarity. PIP5K1A shown in blue, PIPs in green (hydrocarbon tail) and red (phosphate head groups).

I analysed the contact profile for each set of replicates and present the data as both raw counts of the number of lipids in contact with each residue and the average (fig.4.6).

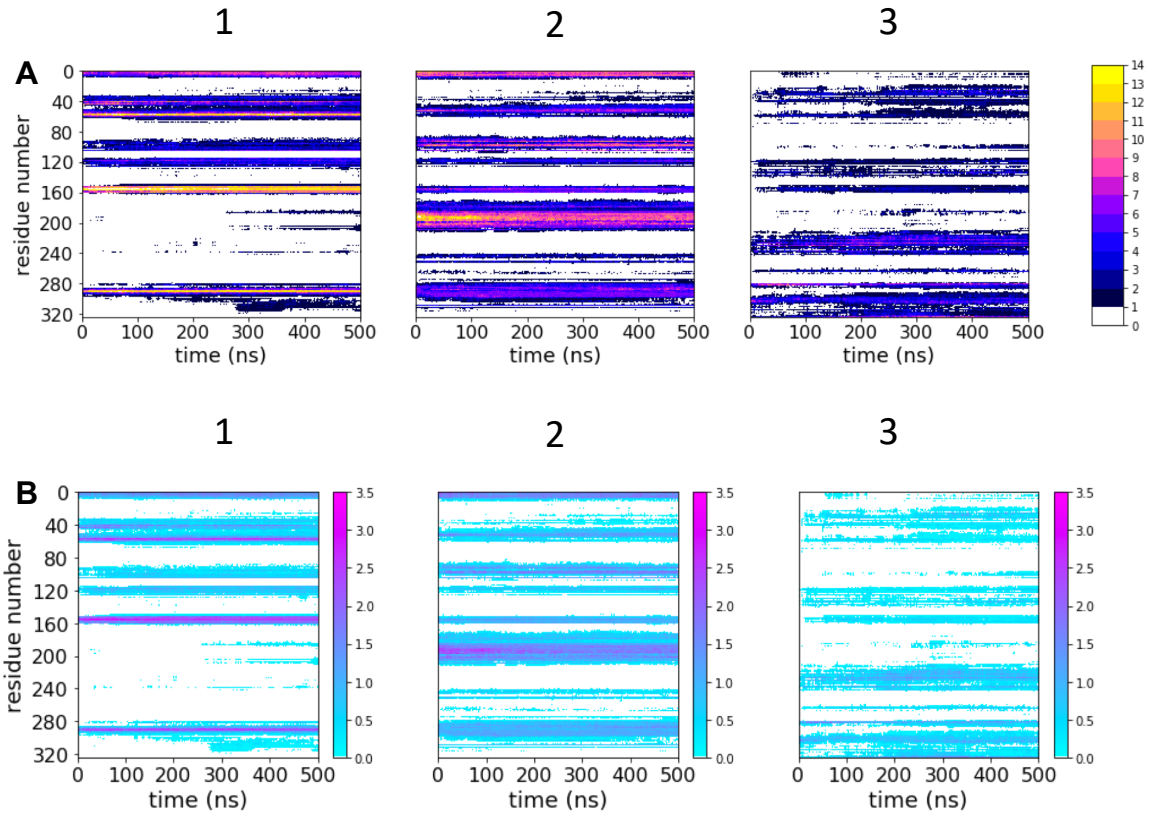


Fig. 4.6: PI4P-residue contacts over time for PIP5K1A. A) Raw contacts over all simulations. B) Average number of contacts per simulation. All conformations show a large number of dynamic contacts with the activation loop. Conformation 2 shows early contacts with the active site (simulation residue 180); these are not developed until later on in conformations 1 and 3.

All conformations show PIPs in contact with residue 182, at the catalytic site, albeit with some variation in the number of PIPs across the 3 conformations. A large number of contacts are made over time with the activation loop. Whilst these contacts appear to be quite dynamic, it appears to show the region containing residues 286-306 is very important for binding. Interestingly, whilst conformation 2 shows a very large number of contacts between the active site and PI4P lipids, these contacts are not established until around 250 ns in conformations 1 and 3. Particularly in conformation 1, it seems that contacts at the active site are correlated with the formation of contacts in the activation loop. It suggests that the region around residue 160 may play a role

in binding and opens up the possiblity of a second interaction site.

I also analysed the RMSD over time of all replicates in each conformer in fig.4.7. to establish the dynamics of the activation loop over time.

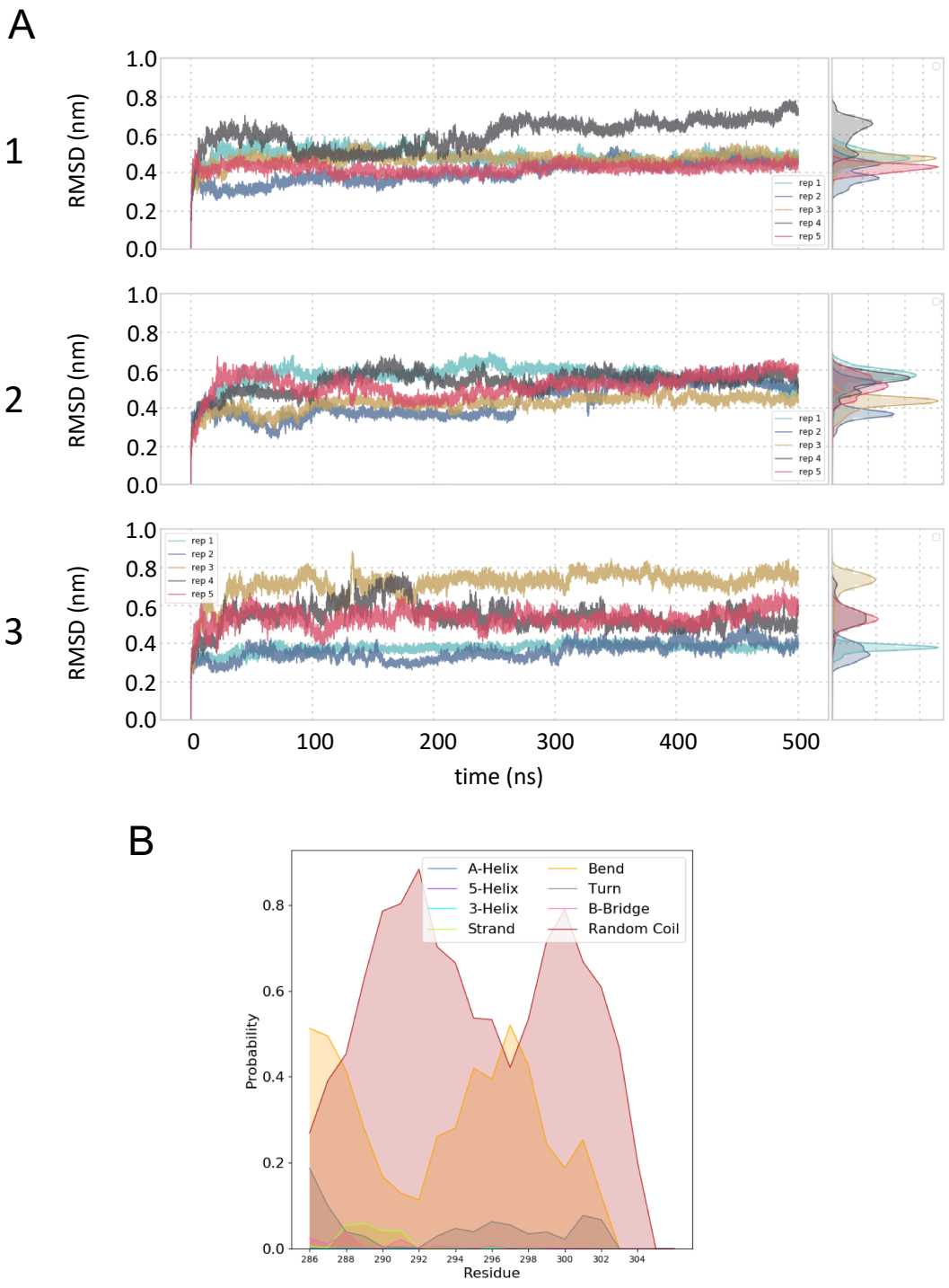


Fig. 4.7: Initial binding RMSD and secondary structure calculations. A: Cα RMSDs of the loop over time for all replicates of each initial conformation. Colours indicate the replicate in each conformation dataset. Numbers indicate the starting conformation referred to in fig.4.5. B: DSSP secondary structure analysis of all trajectories. The averages secondary structure is presented for each residue. Random coil (red) and bend (yellow) dominates the structure.

The RMSDs show that there is general agreement with some outliers, which indicates that the loop structures in the simulations demonstrate some variability. It appears that most structures are broadly stable after 100 ns, but that there is some conformational change in some replicates that continues throughout the simulation (e.g. conformation 1, replicate 4). Interestingly the profiles in the loop are not identical and appear to have different timescales of binding. This suggests that the binding contact pattern is dependent on the initial states of membrane recognition.

The secondary structure analysis shows a strong preference for random coil, and in some cases, turn. In the simulations there is no folding into an α-helix structure. This is in contravention of the study by Liu *et al.* [192] which proposes a helical structure in the membrane. This result highlights the importance of performing a study in a physiological environment with lipids which are relevant to the structural specificity of the loop.

4.3.3 tICA shows the loop switches between two slowly equilibrating populations

MD simulation trajectories contain a lot of noisy information and it is therefore useful to define a set of collective variables with which to determine conformational states. MD simulations also permit a system to tumble randomly in solution, and whilst this is not as much of a concern for the activation loop system, it is still useful to remove rotational motion by featurisation. Featurisation reduces the raw Cartesian coordinates of the trajectory to a dataset that provides a meaningful model. Commonly used features are dihedral angles or atom-atom distances [97]. In order to select the best features for dimensionality reduction, I trialled 4 feature sets. I performed tICA analysis using 2 tICs. tICA is a dimensionality reduction method that finds high-autocorrelation linear combinations of the input degrees of freedom (see methods).

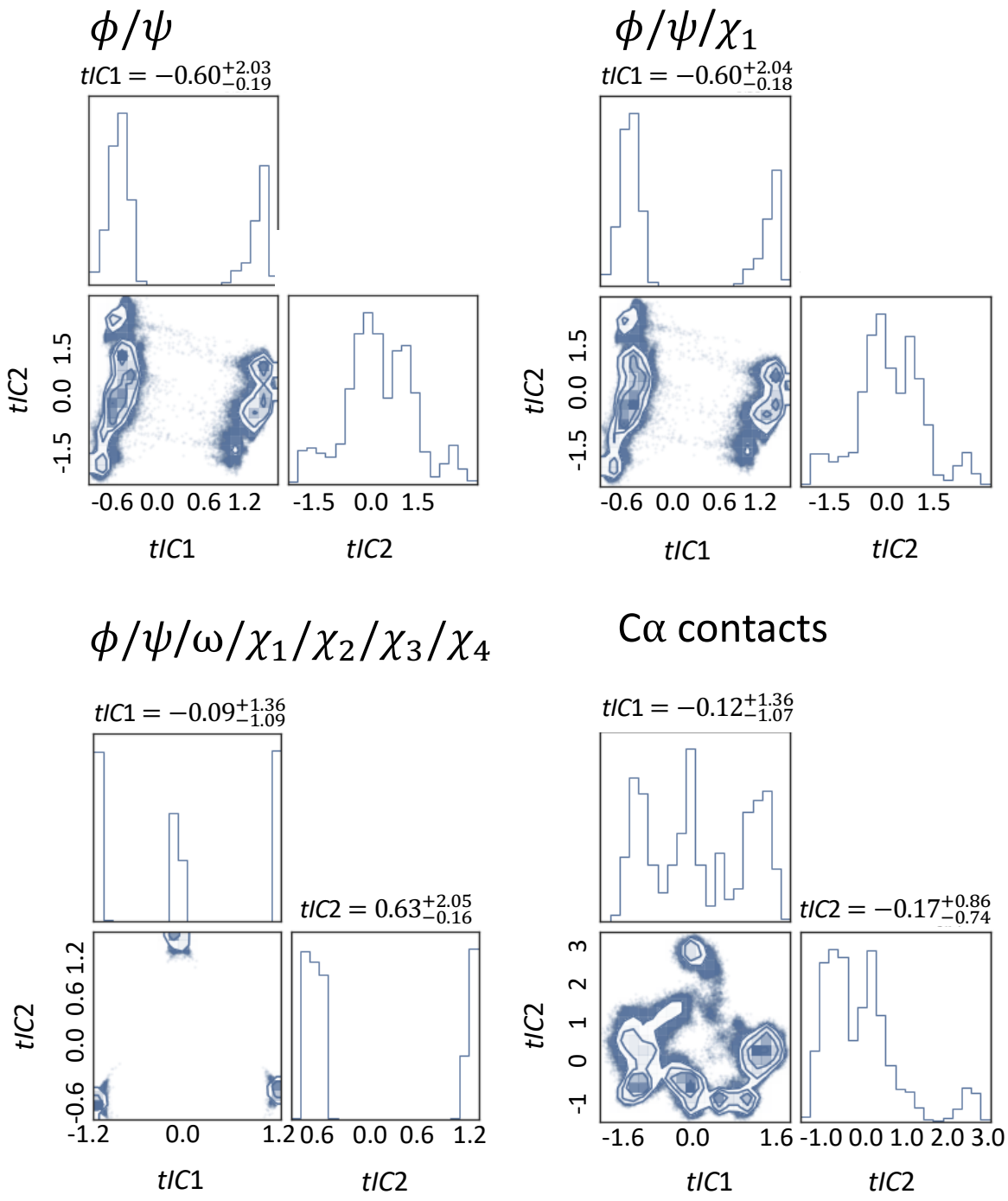


Fig. 4.8: Feature selection. The dataset was featurised on four feature sets as labelled above. The corner plots show the top two tICs as 1D histograms and their 2D projection. The mean and upper/lower quantiles for each tIC are shown above their respective 1D histogram. Featurising on ϕ, ψ or ϕ, ψ, χ_1) angles yields broadly similar landscapes and tIC values. Featurising on $\phi, \psi, \omega, \chi_{1-4}$ angles yields three states which are too dissimilar for building an MSM because there is no reversible sampling of states (see 1D histogram). Cα contacts yields a broader, more homogeneous landscape.

The corner plots show that using all dihedral angles (ϕ, ψ, ω, χ) results in three states which are too dissimilar to construct an MSM. Using Cα contacts gives a landscape which is quite homogeneous. However, featurisation based on a reduced set of angles yields a two state landscape which may be interesting to investigate further. The feature selection plots show that using only ϕ and ψ angles gives a similar landscape to using ϕ, ψ, and χ_1 angles, but given that we are interested in the positioning of the side chains and wish to generate representative structures that encapsulate these positions, I proceeded with featurisation based on ϕ, ψ, and χ_1 angles.

The dimensionality of the featurised dataset was reduced using tICA using a 2.5 ns tICA lag time (fig.9) and clustered using the MiniBatchKMeans algorithm into 50 microstates. There is currently limited agreement in the literature on the best choice of algorithm or number of states [97] (although some scoring approaches are being developed [157]), but in this study this combination showed the least error when calculating timescales. Additionally, increasing the discretisation (number of microstates) from 50 to 100 or 200 appeared not to alter timescales so I consider 50 microstates to be sufficient.

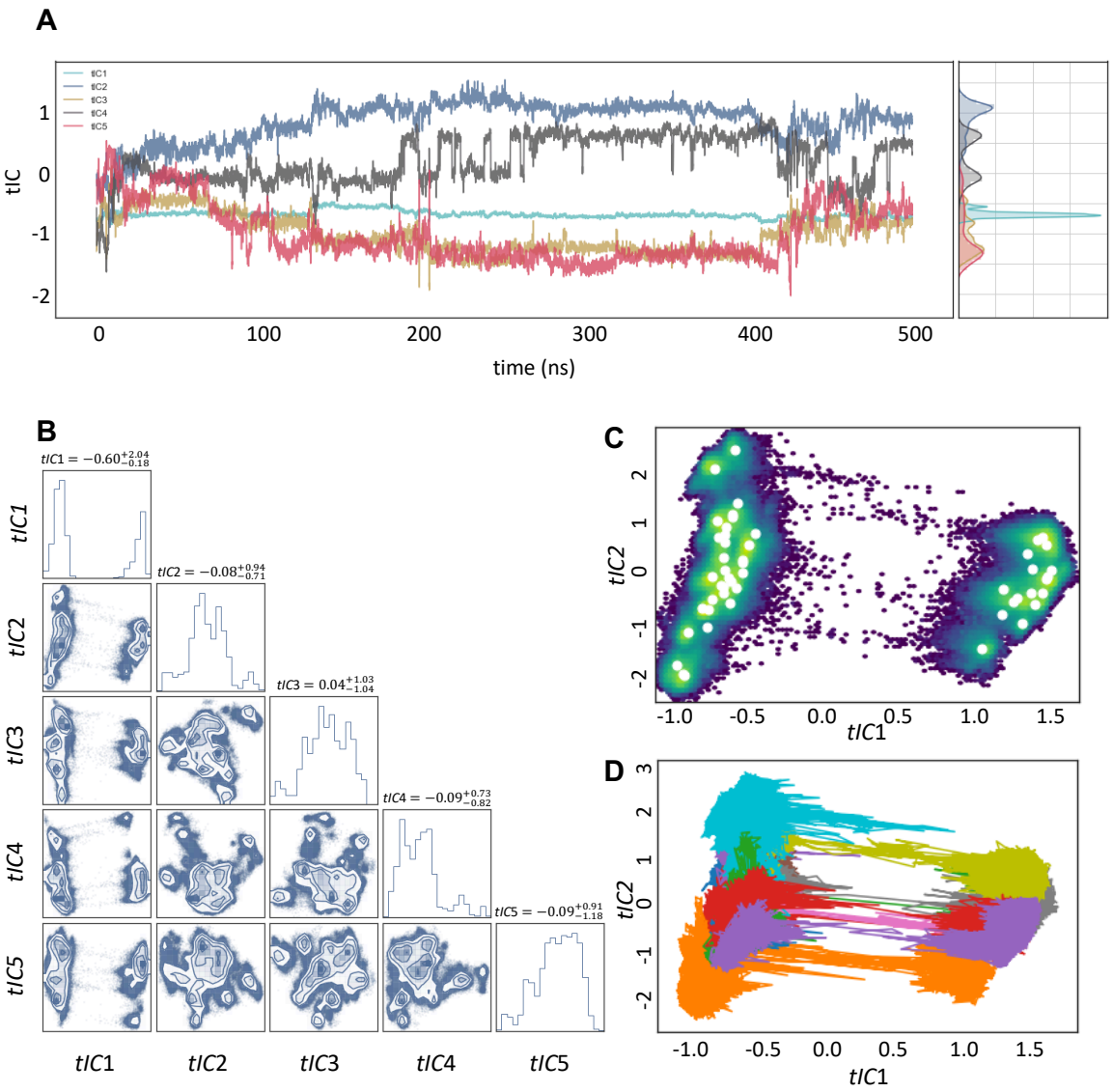


Fig. 4.9: Dimensionality reduction. A) Trace plot of the 5 tICs and their marginal histograms for one representative trajectory. Green=tIC1, blue=tIC2, yellow=tIC3, grey=tIC4, red=tIC5. The plot shows that the different tICs have resolved the transitionas as discrete jumps. B) Corner plot and histograms of the 5 tICs for all trajectories. Mean tIC value and upper/lower quantiles indicated above respective 1D histograms. C) 2D projection of tICs 1 and 2. Colour indicates density. White circles indicate cluster centres of microstates as determined by MiniBatchKMeans. D) All tICA trajectories plotted over time on the tIC1 vs. tIC2 projection. The colours indicate different trajectories. Three trajectories (orange, pink, yellow) appear to indicate reversible transition.

tIC 1 shows two slowly equilibrating states as indicated by the distribution (fig.4.9B). tIC2 was chosen for projection with tIC1 due to its 1D histogram indicating a sepa-

ration of populations and lower tICs being less likely to represent noise.

MiniBatchKMeans clustering (fig.4.9C) shows a good coverage of all regions of the projection. It reveals two areas of high density which will have slowly equilibrating populations (which correspond to structural conformations). This dimensionality reduction is used for constructing the Markov state model. I also show the individual coverage of each replicate in tICA space (fig.4.9D) which shows a number of replicates crossing into the right hand state. This gives confidence that the tICA space is sampled well enough to build a model.

In Fig.4.10, I show the individual sampling of each trajectory over time. Numbers 1,2,3 indicate the initial conformation as shown in fig.4.5.

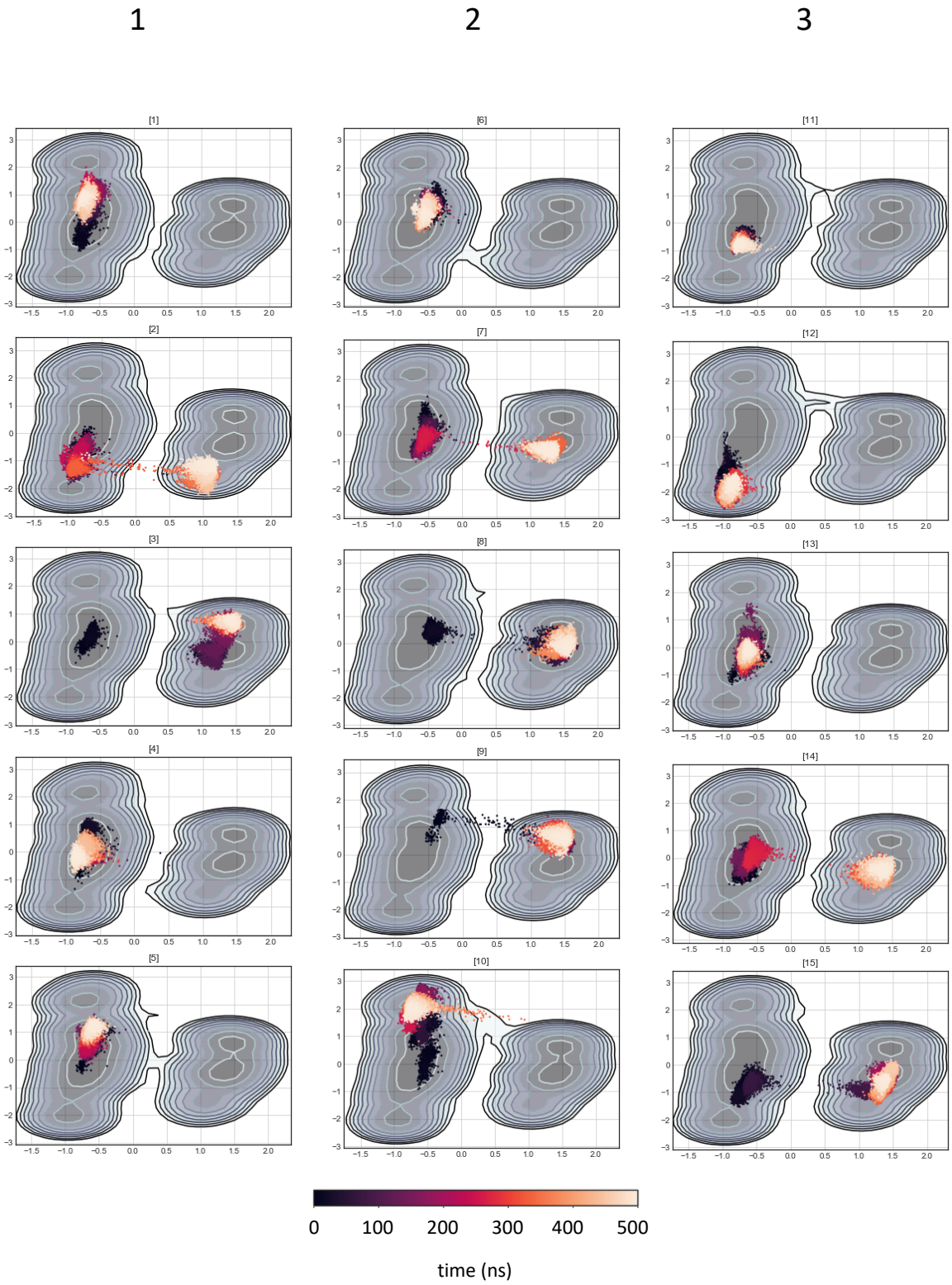


Fig. 4.10: All trajectories in tICA space over time. The tICA landscape is represented as a grey pseudo-free energy landscape with individual frames of each trajectory represented as points and coloured by time. Numbers 1,2,3 indicate the initial conformation as shown in fig.4.5.

This analysis shows around 40% of trajectories cross into the right state from the start conformation. It also indicates the sampling of the left state.

4.3.4 Markov state models of the activation loop

MSMs for MD data are derived from calculating a transition matrix between conformational states. These microstates were calculated from the MiniBatchKMeans clustering algorithm in the previous figure. I initially calculated a number of MSMs at different lag times in order to optimise lag time selection (fig.4.11A). I turned off ergodic trimming in order to maximise the amount of data used for the MSM and therefore reversibility was calculated by direct transpose (as is more appropriate when ergodic trimming is turned off - see MSMBuilder manual [157]).

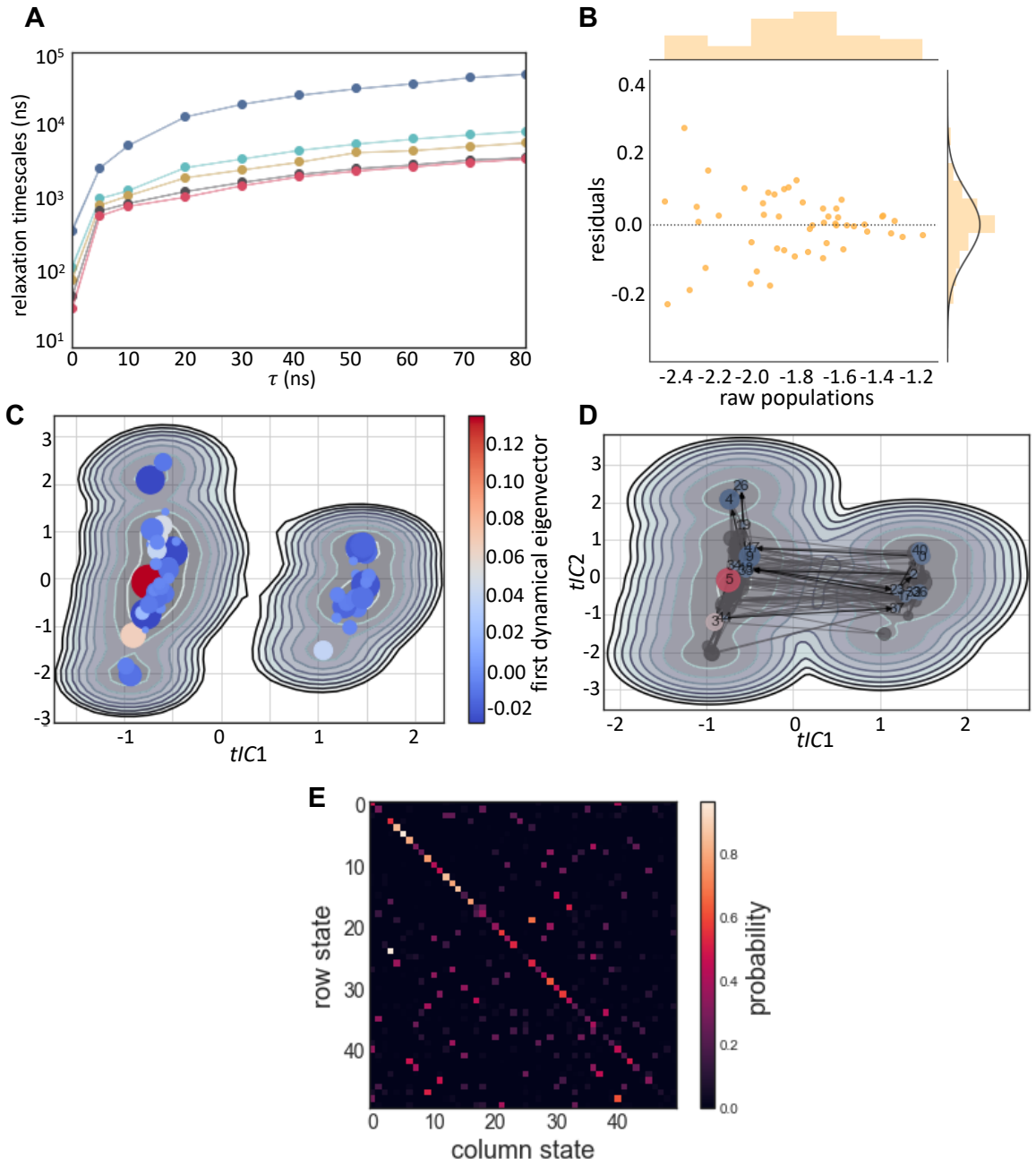


Fig. 4.11: MSM of the PIP5K1A activation loop. A) Timescales of 5 processes derived from Markov state models constructed using lag times indicated on the x-axis. B) Population residuals derived from the MSM at 6000 frames (60 ns). C) Microstate centers are shown as circles on the pseudo-energy landscape, sized according to state population and coloured according to the first dynamical eigenvector. The slowest process is a transition from red states to blue. D) Markov state model of the initial binding of the PIP5K1A activation loop to a PIP-containing membrane. The slowest process is from red (state 0) to blue (state 10). These are projected onto a pseudo-energy landscape derived from the tICA projection. Nodes of the Markov State model with the top three dominant flux pathways shown as bold black arrows. Minor states and transitions between them are indicated in grey. E) The transition matrix represented as a heatmap, with transitions coloured according to probability. The diagonal shows that most states transition back to themselves, but other high probability transitions are observed.

The MSMs were constructed on 50 microstates (fig.4.9A). Kinetic analysis of relaxation times reveals a minimum lag time of 60 ns for constructing MSMs due to each process approaching its maximum value (fig.4.11B). This kinetic analysis also reveals a low error associated with the processes and that 4 important kinetic processes exist (there is some overlap at the shorter timescales which indicates the MSM classifies identical processes as being different).

An MSM was constructed using 60 ns lag time. The raw population residuals (fig.4.11B) indicate that the residuals are decorrelated and reasonable sampling has been achieved. It is important (as always) to consider in the interpretation of the MSM that there may be areas which are less well sampled; these might lend themselves to enhanced sampling regimes. However, taken with the low error and behaviour of the extracted timescales, I can proceed in the knowledge that the areas that are sampled are probably close to convergence.

The resultant MSM shows that the slowest progression from red to dark blue (fig.4.11C). This can be interpreted as a gross structural motion.

I then analysed the dominant fluxes between states in order to investigate the progression of the activation loop structure in the initial binding event (fig.4.11D). I plotted the top three transition pathways. The transition pathway model reveals a dominant process from state 5, into the right state, which either remains there or transitions back to the left state and up to a state at the top of the left state (n.b. states are numbered arbitrarily). The larger sizes of these circles indicate they are the largest populations, presumably the start and end conformations. This is projected on a pseudo energy landscape which indicates low energy regions around state 5, 26, 40, and 4. There is also a transition through state 3 which is well populated.

Finally, in order to provide some structural meaning to the MSMs, I produced representative structures.

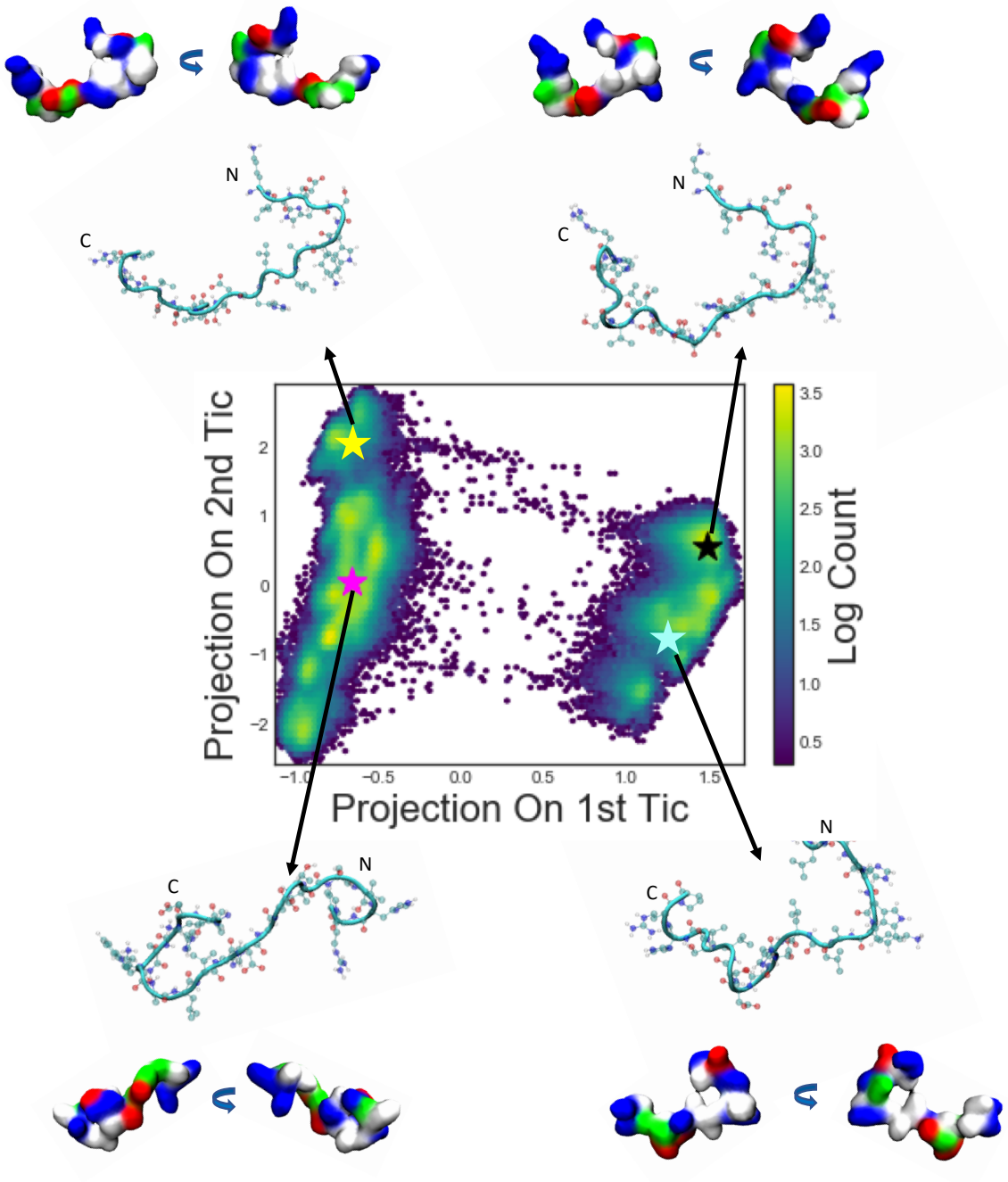


Fig. 4.12: 2D projection of tICs 1 and 2 with representative structures of the loop. Each star indicates a region of high density in tICA space, and tICA space was reverse-transformed to produce representative structures of the activation loop Representative structures are shown with a ribbon representation of the peptide backbone to indicate overall geometry, with individual residues represented as translucent CPK particles. Atoms are coloured according to atom type; green = C, blue = N, red = O, white = H. Surface representations are displayed above and below the atomistic representations; these are coloured according to residue type as in fig.4.1 (white = non-polar, blue = basic, red = acidic, and green = polar residues). A 180° rotation is also shown.

The two separated regions can be interpreted as the loop transitioning between two dominant conformational pathways on binding. There are many transitions between the two major states and a high degree of interconnectedness within each major state; this indicates that the loop can transition through a number of kinetically and structurally similar conformations. This is an important result as this is a commonly described feature of intrinsically disordered proteins, and it is interesting to see it revealed in an MSM.

The representative structures extracted from the states show a number of large structural changes with regard to the positioning of the side chains. Interestingly no states show α-helix structure (see fig.4.7B), as has been postulated in DHPC membranes [192]. The start state (magenta) shows a flatter conformation, whereas the others show a more extended structure (presumably for penetration into the bilayer). The region penetrating most deeply into the membrane appears to adopt different orientations with regard to its larger residues; it is also interesting that there are two small regions of sequential acidic/polar residues, one thought to be placed higher in the membrane (towards the N-terminus) and one that penetrates quite deeply (C-terminal end). This region at the C-terminus is flanked by basic residues. Additionally, the loop in this region is dominated by a large number of basic residues. It suggests that the large conformational changes in this region, combined with the distrubution of charges, may act to assist lipid shuttling in and out of the active site; it would be useful in the future to characterise which region of the lipids each loop residue makes contact with to unpack this further.

Finally, I produced a representative binding event using PCA filtering.

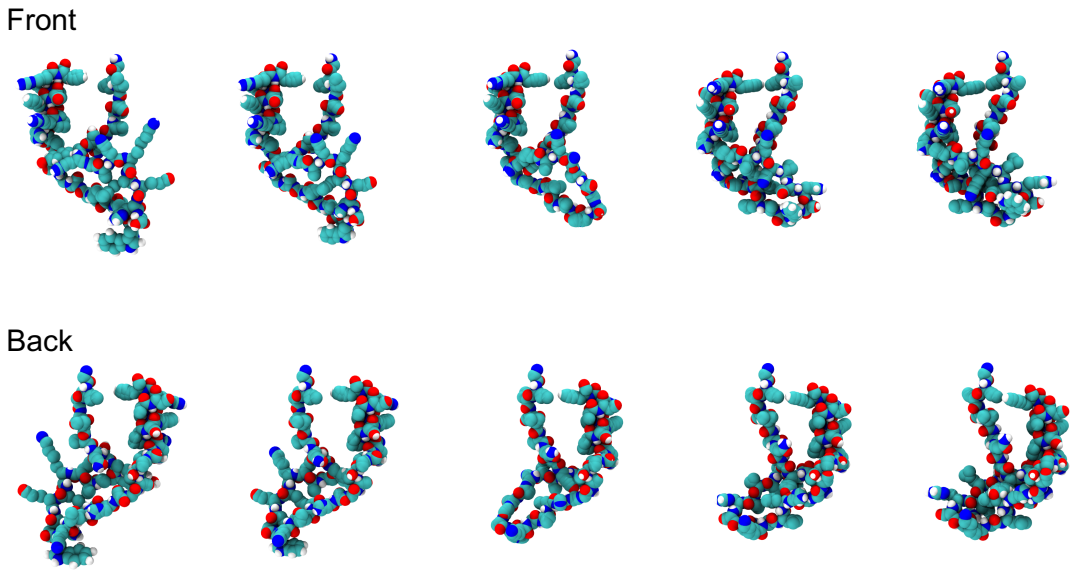


Fig. 4.13: Graphical representation of an example folding trajectory in PCA space viewed from front and back. Atom colouring indicates atom type: green = C, red = O, blue = N, white = H.

Identifying functionally relevant motions by eye is challenging because I observe both small local fluctuations and large global motions occurring simultaneously. PCA can make it easier to understand these motions because it filters the large collective motions from the small local fluctuations. Large collective motions occur stochastically, and so identifying smooth transitions is challenging; one solution to this is to interpolate between the extreme conformations (fig.4.13). PCA reveals a folding of the loop into the state observed in fig.4.3B and corroborated by the representative structures of fig.4.12. It also shows a number of large collective motions of large amino acid residues; in particular large aromatic residues in the middle of the loop which were used to cluster conformations in the preliminary analysis in fig.4.4B. These certainly warrant further investigation in order to identify which are the key residues in the binding event and if these residues bind in a particular order (given the dominant flux pathway, this seems to be a reasonable hypothesis). Taken with the penetration

analysis and the contact profiles, residues 291-303 make the greatest contribution to binding. This region is dominated by large non-polar and basic residues, followed by alternating acidic and polar residues. I postulate that there is some interplay between residue size and charge that may assist in the recognition event, which is then followed by structural reorientation to bring the PIPs into contact with the catalytic site. The acidic residues may contribute to the autoinhibitory mechanism as they are very close to the catalytic site and may induce some kind of structural change which filters the catalytic product (PIP2) away from the active site due to its increased charge density.

More broadly, the finding that the activation loop can transition into many intermediate states is consistent with behaviour observed for other intrinsically disordered proteins [217–219]. Disordered regions require extreme conformational variability in order to carry out their functions in heterogeneous environments. As the PIP5K1A approaches the membrane, the loop must be able to explore the membrane landscape in order to recognise and identify PIPs, and then adopt a conformation that orients the rest of the protein onto the membrane in the appropriate conformation for catalysis. Where Liu *et al.* [192] postulated a folding into an α-helical structure, we appear not to see this. DHPC micelles overestimate α-helix propensity and it seems unlikely that this disordered region would maintain a static helical structure in the membrane if part of its function is to assist in lipid shuttling. I therefore consider this study an important extension to our understanding of how the loop behaves in a physiological environment.

It follows that whilst there may be a dominant preferred pathway for this process, it needs to be flexible enough in order to manage this recruitment at different PIP concentrations and at different directions of approach or at different membrane curvatures. The loop has also been postulated to have an autoinhibitory role [15]; by being sensitive to the increased charge of its product lipid (which is more negatively charged

after phosphorylation) the loop would reorient to release the PIP from the catalytic site. Longer simulations in mixed membranes with a greater diversity of lipid species may help us understand this problem in terms of the exquisite structural sensitivity of the disordered loop.

4.4 Conclusion

This chapter sets out to provide proof of concept of the use of MSMs for MD simulation data of intrinsically disordered regions interacting with cell membranes. The motivation for this is that disordered regions are poorly understood; the MSM protocol provides a method for fast analysis of the simulations to extract data about important conformational states and the transitions between them.

I showed that the activation loop makes contacts with PIPs in the membrane that result in a productive binding conformation with the PI4P substrate positioned in the active site. This conformation was previously demonstrated in 2016 by X-ray crystallography where PI4P was identified in the active site [26] and coarse-grained simulations [215]. An NMR study has also shown that this region is important for membrane recognition [192]. This provided a rationale for applying the MSM protocol to simulation data. I showed that an MSM can be constructed to investigate an initial binding event of a disordered region to a membrane; this is the first demonstration of this to our knowledge. I showed that the MSM can be used to derive useful structural information about the initial binding pathway, and that this can be combined with PCA to investigate gross structural changes that may have important functions.

One of the limitations of this study is the amount of time the systems were simulated for. Other MSM studies of a similar nature have been on much smaller systems and therefore have total simulation times in the 20-40 μs range, compared to 7.5 μs used here. The challenge of simulating this system lies in the fact that

the water box needs to be large enough to accommodate both a globular protein and a model membrane with sufficient clearance between the periodic boundaries, whereas other studies tend to simulate one or the other with subsequently smaller box sizes (and fewer calculations). However, this study still provides useful results even at shorter sampling times, and has resulted in the development of a protocol (https://github.com/sbtaamos/MSM_analysis) that can be supplied with simulation data and run with very little user input (aside from lag time and feature selection). It could also be used to initiate enhanced sampling simulations in future in order to ensure convergence in less-sampled regions. Another limitation in the field is that there is currently little experimental data to validate these findings. It is hoped that the increased interest in the PIP kinases over the last two years will result in new studies which focus on the structure-function interplay of the loop and provide data that can be fed back into the model for refinement.

In the immediate future I intend to do further work on optimising the parameters for MSM construction and use bootstrapping to assess convergence in more detail. The simulations also lend themselves to enhanced sampling regimes, such as adaptive sampling. More generally, I anticipate that the structural information and method development derived from this study and others can be used to investigate regions of interest to drug design that hitherto have been elusive.

Lipid kinases are implicated in many cancers as a result of their influence on cell processes and studies in mice and human tissue samples suggest inhibition of PIP5K1A as a target for treatment. These studies not only contribute to a greater physical understanding of signalling processes and provide new tools for further study of peripheral proteins in general, but may also aid in the development of clinical therapies.

Chapter 5

Computational Studies of α-synuclein

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

John von Neumann

5.1 Introduction

α-synuclein is a protein implicated in a group of neurodegenerative disorders, including Parkinson's disease and Lewy body dementia [220], but its function in the healthy brain is currently unknown [221]. It is now well established that pathology arises from the misfolded, aggregated α-synuclein on the lipid membrane [222]. Consequently, interactions between α-synuclein and lipids appear to be important for when and how the protein misfolds and aggregates, and moreover how α-synuclein consequently causes cellular damage.

Fig.5.1 shows a graphical representation of the amino acid sequence: the N-terminus

from 1-60, the NAC region from 61-95, with the C-terminus remaining disordered. α-synuclein exists in the cytosol as an intrinsically disordered monomer [35, 223], but folds into an α-helix when it associates with lipid membranes [224] (as shown by NMR studies on SDS micelles, including PDB structure 1xq8). Fig.5.1 outlines the structure; helix 1, from residue 1-38, followed by an interhelical region, followed by helix 2 from residue 46 to 95, and the disordered C-terminus from residue 96 onwards. Lipid association is thought to be important for its function in regulating synaptic vesicles, where it has been shown to be essential for SNARE complex assembly at the presynaptic membrane [225, 226]. It has been postulated to have a wide range of functions, including neuronal differentiation [226] and suppression of apoptosis [227]. The N-terminus and the NAC region both contain seven conserved KTKEGV motifs (although the sixth and seventh are not fully conserved) which are thought to have a significant role in maintaining the native unfolded status [228]. The disordered C-terminus extends from residue 95-140 and contains 10 glutamate and 5 aspartate residues. Consequently, this region is highly negatively charged.

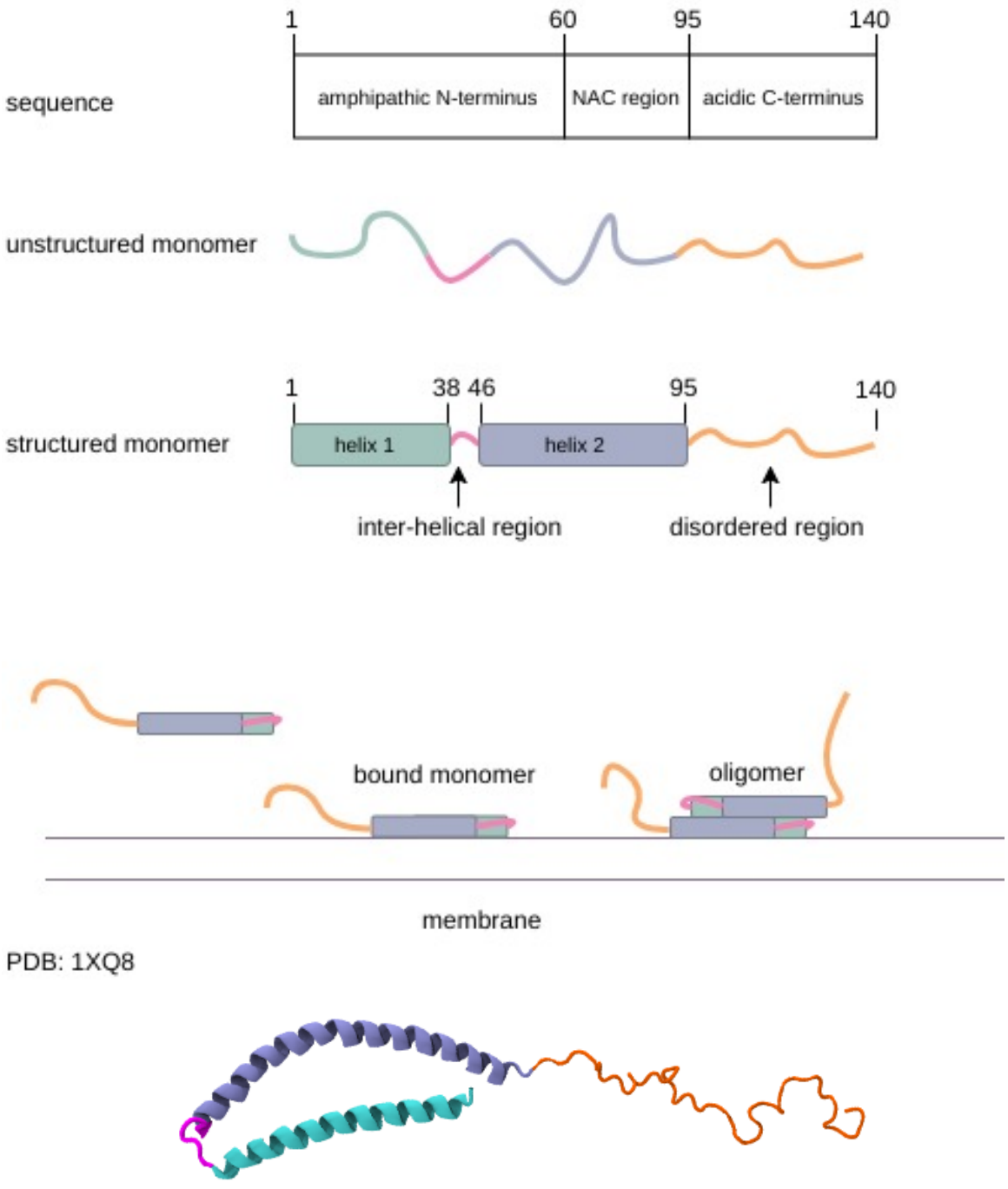


Fig. 5.1: The sequence and structure of α-synuclein. The α-synuclein sequence is 140 residues in length and is broadly divided into three regions: the amphipathic N-terminus, the hydrophobic NAC region, and the acidic C-terminus. α-synuclein is unstructured in solution but undergoes some folding to form a helical structure when binding to membranes. α-synuclein binds to the membrane and can oligomerize with other monomers. PDB structure 1XQ8 is visualised and coloured according to the diagrams.

The NAC region was identified in plaques in Alzheimer's disease; it has been shown to self-aggregate and also to seed β-sheet aggregation. It is, therefore, thought that the NAC region drives pathology. This aggregation results in a transition from α-helical structure to β-sheet fibrils that cause damage to the cell membrane. The amyloid fibril structure was recently solved by cryo-EM [43, 44]. It consists of two protofilaments intertwining along an approximate 2_1 screw axis, and features β-strand architecture with Greek key-like topology. The dimer interface of the fibril is home to four of the six early onset familial mutations for Parkinson's disease: H50Q, G51D, and A53T/E. One (E46K) is located at the interhelical region and appears to have a role in stabilising the protofilament. These mutations alter the structure of the fibrils [44].

The relevance of lipid interactions to the function of α-synuclein is demonstrated by its structural rearrangement upon vesicle binding. Further to this, it has been shown that lipid binding and structural rearrangement is disrupted by pathogenic mutations [229]. However, the picture is complicated by contradictory results, particularly with regard to mechanism, lipid class preference, and activation of downstream pathways [230].

There is some evidence that oligomeric intermediates, which only form on the membrane, are more toxic than the mature fibrils [231–234]. In this context, it becomes important to understand the initial binding event of α-synuclein to the lipid membrane, and then the consequent structural rearrangements that take place in the monomeric and oligomeric forms. Furthermore, the modulation of membrane binding as a therapeutic strategy remains an open question [235]. Therefore in this chapter I seek to to investigate this initial binding process with simulations and Markov state models. This chapter also contains NMR chemical shift data contributed by colleagues at UCB.

5.2 Methods

Simulation details

The NMR structure of α-synuclein was acquired from the Protein Data Bank (PDB: 1XQ8) and converted to a CG representation using in house scripts. CG simulations of α-synuclein were performed using a 10 x 10 nm^2 area bilayer of POPG or a lipid mixture of DOPC/DOPE/DOPS in a 2/5/3 ratio. I chose to use systems with anionic lipids because α synuclein has been shown to bind selectively to anionic lipids, and because experimental collaborators were using these compositions; this is a commonly used lipid composition for other studies (see [236, 237]). Bilayers were built using INSANE [238]. The monomer was positioned 4 nm away from the bilayer surface in an orientation from which the monomer is thought to approach. The box was solvated and sodium and chloride ions added to a concentration of 0.15 M. CG simulations were performed using GROMACS 5.1 [143]. Energy minimisation was carried out via steepest descent and the system equilibrated for 5 ns with protein backbone particles restrained. Productions simulations were run without restraints for 1 or 2 μs. For each model, 10 replicates were performed with different initial velocities.

CG simulations were performed using the MARTINI 2.1 force field [123, 139] with a 20 fs time step. Particle coordinates were written out every 0.5 ns. Coulombic interactions were shifted to zero between 0 and 1.2 nm. Lennard-Jones interactions were shifted to zero between 0.9 and 1.2 nm. The nearest neighbour list was updated every 10 steps. The Berendsen thermostat [131] (coupling constant 1 ps) and barostat (coupling constant 1 ps, compressibility $5x10^{-6}$ bar^{-1}) were used to maintain temperature at 323 K and pressure at 1bar. The LINCS algorithm [111] was used to constrain bond lengths.

AT simulations were performed using the Charmm36 force field [125] with a 2

fs time step. Particle coordinates were written out every 20 ps. Lennard-Jones interactions were shifted to zero between 0.9 and 1.2 nm. Long-range electrostatic interactions were treated using the particle-mesh Ewald method (PME) using default parameters pme-order = 4 and ewald-rtol = 10^{-5}, fourierspacing = 0.12. PME was shifted from 0 to 1 nm [129]. The nearest neighbour list was updated every 10 steps. A velocity-rescale thermostat (coupling constant 1 ps) and Parrinello-Rahman [135] barostat (coupling constant 1 ps, compressibility 5x10^{-6} bar^{-1}) were used to maintain the temperature and pressure. The LINCS algorithm was used to constrain bond lengths.

AT simulations were performed by obtaining a snapshot from CG simulation and conversion via CG2AT, a fragment-based approach [145]. AT-MD simulations were performed using GROMACS 5.1/2018. The system was equilibrated for 1.5 ns with the backbone atoms of the protein restrained, then a production run of 100-250 ns carried out.

Analysis

VMD [159] was used for simulation visualisation. Graphs were generated in matplotlib [155] and seaborn.

Contact analysis was carried out with in house python scripts using a cutoff of 0.7 nm for CG and 1.0 nm for atomistic from the Cα carbon atom. PCA and correlation analysis were carried out using standard libraries in matplotlib/seaborn [155] and scikit-learn [154]. Ratios of lipids in binding events were calculated using numpy. The gromacs DSSP library was used to calculate secondary structure propensity and plotted using a script distributed by the Bevan lab under the GNU General Public License [239].

Markov state modelling was carried out in MSMBuilder [157]. The system was

featurised on backbone contacts and scaled using StandardScaler. The number of microstates was set at 50 for calculation and implied timescales plotted to assess convergence. The system was clustered into macrostates (this is further explained in the context of the results) and the two lowest energy states were used to extract representative structures which were visualised with VMD.

5.3 Results & Discussion

In order to investigate the interactions of α-synuclein with membranes, I performed a number of CG and AA simulations as shown in Table 1. I adopted a multiscale approach where initial simulations were conducted using the CG Martini model and then backmapped to AT resolution. Simulation details are given in the methods section.

CG/AA	lipids	time (ns)	n	total time (μs)	description
CG	POPG	1000	10	10	monomer
CG	POPG	1000	4x10	40	decreasing helicity
CG	DOPC/DOPE/DOPS 2:5:3	2000	10	20	mixed lipid
AA	POPG	100-250	3x10	5.5	initial binding

Table 5.1: Table of α-synuclein simulations.

5.3.1 CG simulations show that the initial binding sequence is led by the inter-helical region

CG simulations were set up by positioning the α-synuclein monomer 4 nm away from a POPG membrane surface (fig.5.2A). The protein was observed binding to the membrane.

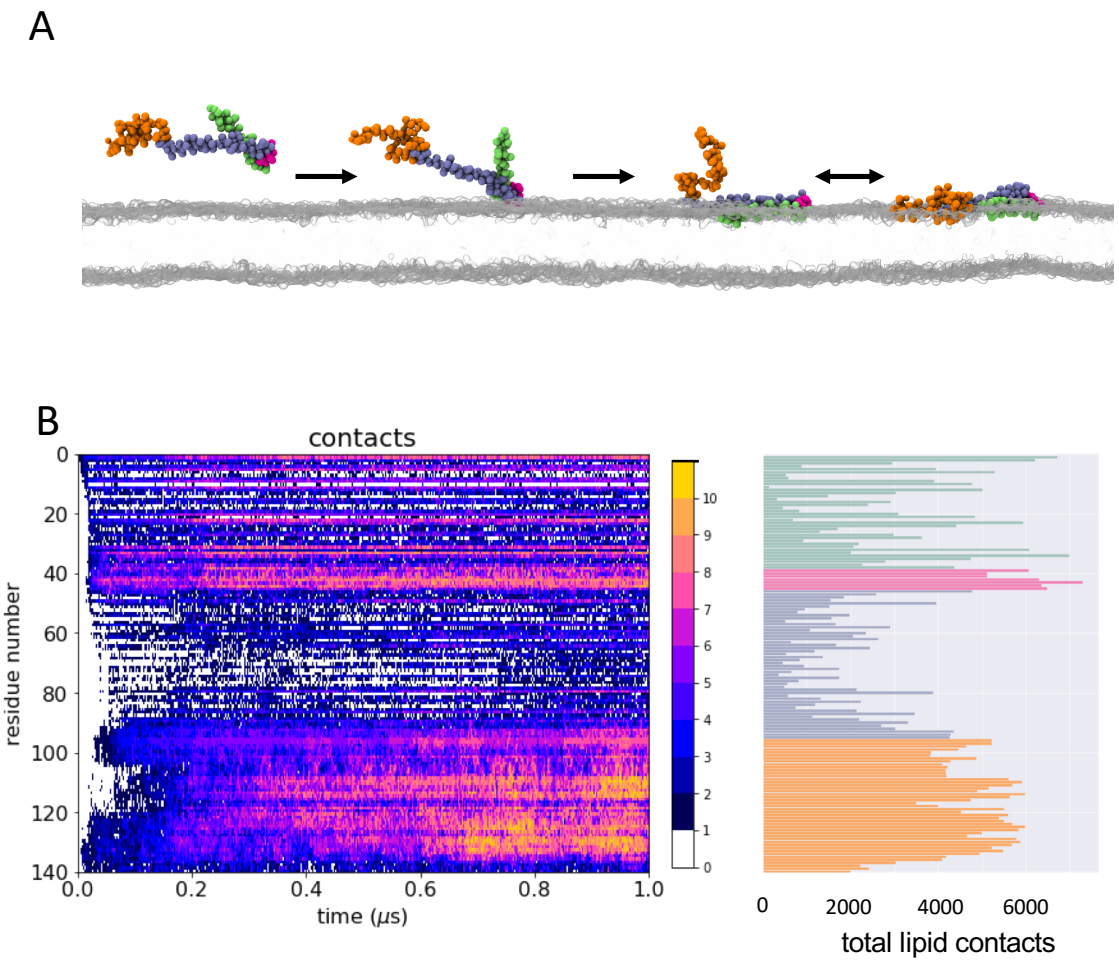


Fig. 5.2: Coarse grained simulations of α-synuclein with a POPG membrane. A: The monomeric α-synuclein is positioned distal to the membrane. 10 replicates of 1 μs were performed and the α-synuclein was observed binding to the membrane. Colours indicate helix 1 (green), the interhelical region (pink), helix 2 (purple), and the disordered C-terminus (orange). B: Total contacts with POPG for each residue of α-synuclein across 10 simulation replicates. A residue is considered in contact if the residue backbone particle is within 0.7 nm of any POPG particle.

Simulations of the α-synuclein (helical from 1-100) show that the interhelical region makes the first contacts with a POPG membrane (fig.5.2B). 50% of simulations showed binding at the interhelical region within 100 ns, and all simulations showed binding within 1 μs. Residues in the N-terminus which made many contacts with POPG residues were residues 1, 21, 23, 32 and 34. Helix 2 makes few contacts with the POPG membrane, although residues 50 and 80 contribute to the overall binding profile as the

simulation progresses. The disordered C terminus makes a number of contacts with the membrane as the simulation progresses.

In order to further analyse the POPG binding profile, we conducted correlation analysis of the contacts which shows strong positive correlation in contacts between the disordered tail and the interhelical region and r79-80 (fig.5.3A).

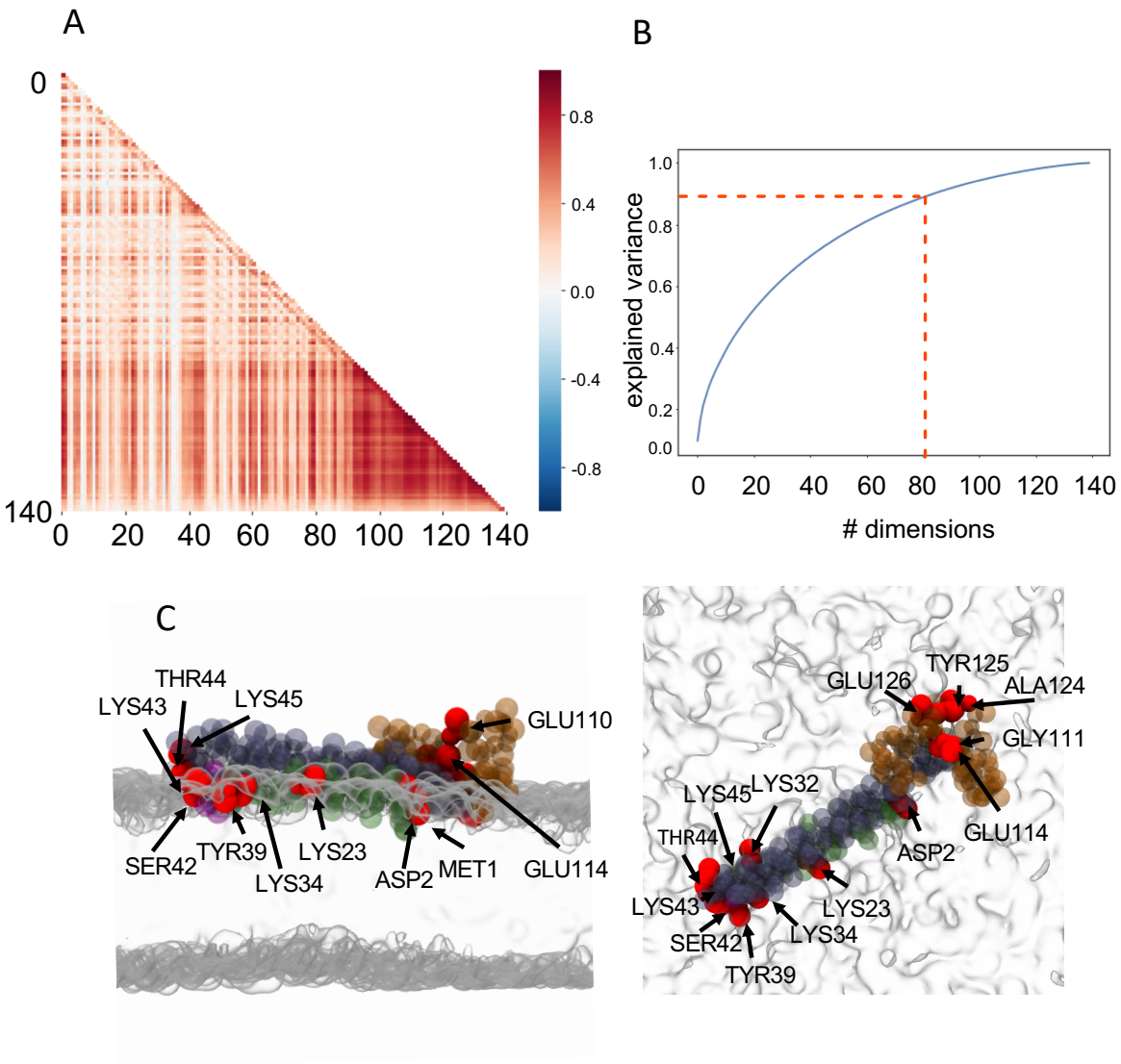


Fig. 5.3: Coarse grained simulations of α-synuclein with a POPG membrane identify important residue contacts. A: Correlation analysis of the residue-POPG contacts reveals strong correlation between contacts in the interhelical region and the disordered C-terminus. B: PCA shows that 95% of the variance in contacts is explained by 80 residues. C: Side and top views of the monomeric -synuclein bound to the POPG membrane. Red particles indicate the top 14 residues making contact with POPG lipids.

PCA shows that 80 residues explain 95% of the variance in binding (fig.5.3B). Fig.5.3C shows a representative bound structure of α-synuclein with the top 14 residues making contacts (from fig.5.2B) highlighted in red. 4 of these residues are in the interhelical region (residues 39, 42, 43, 44, 45). We observe that helix 1 is in close

contact with the membrane, with major contributions to the contact binding profile from residues 1, 2, and 32. Interestingly, we observe that helix 2 (containing the NAC region) is folded back over the top of helix 1.

To establish how variation in the secondary structure of α-synuclein might contribute to the binding profile, we modelled the α-synuclein monomer with increasing degrees of helicity: residues 1-20, 1-40, 1-60, and 1-80, in order to compare with the simulations derived from the 1xq8 PDB structure in fig.5.2A. Again, 10 replicates for each helix composition were run for 1 μs. Further details are given in the methods.

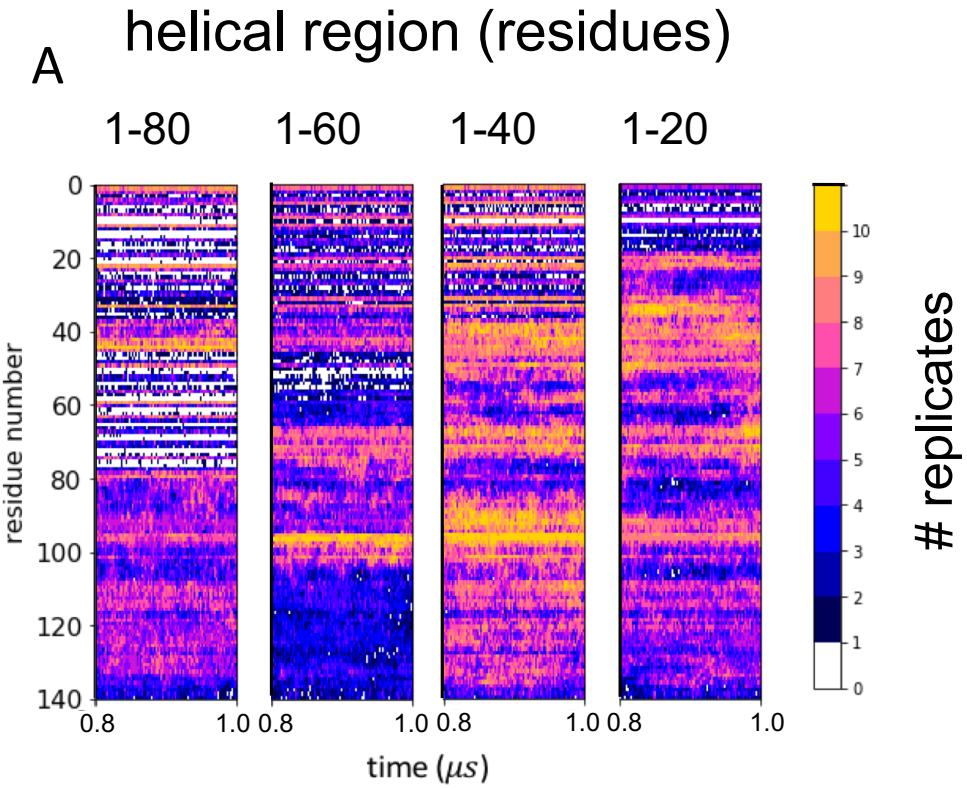


Fig. 5.4: Coarse grained simulations of α-synuclein with a POPG membrane with varying degrees of helicity. Each conformation was simulated for 10 replicates of 1 μs. The colours indicate the number of replicates across the ensemble making contacts at that time point. Only the last 0.2 μs of each set of simulations are shown. The contact profiles show that the contribution from the inter-helical region remains a common feature of binding, but abrogation of helicity results in a greater contribution from residues 45-100.

Fig.5.4 shows that as overall helicity of α-synuclein decreases, we observe a greater

contribution to the contact profile from helix 2. The large contribution from the interhelical region remains unchanged. Residues 1, 20, 79 and 80 continue to make many contacts with the membrane regardless of structure. Overall the results show that abolishing helix structure increases the binding contribution from helix 2, presumably because it is now free to make contacts with the membrane instead of being folded over helix-1 (as in fig.5.3C). Additionally, the interhelical region appears to dominate the contact profile even when helicity is reduced.

5.3.2 CG simulations in mixed lipid bilayers indicates time-resolved lipid preference

To establish the binding of the α-synuclein monomer to a more complex model membrane, we performed CG simulations of α-synuclein binding to a mixed DOPE/DOPS/DOPC bilayer in a 5:3:2 ratio. Fig.5.5 shows binding led by the interhelical region, but with a reduced contribution from the C-terminus (Fig.5.5A). Residue 45 makes the most contacts with the membrane, followed by residues 39 and 46.

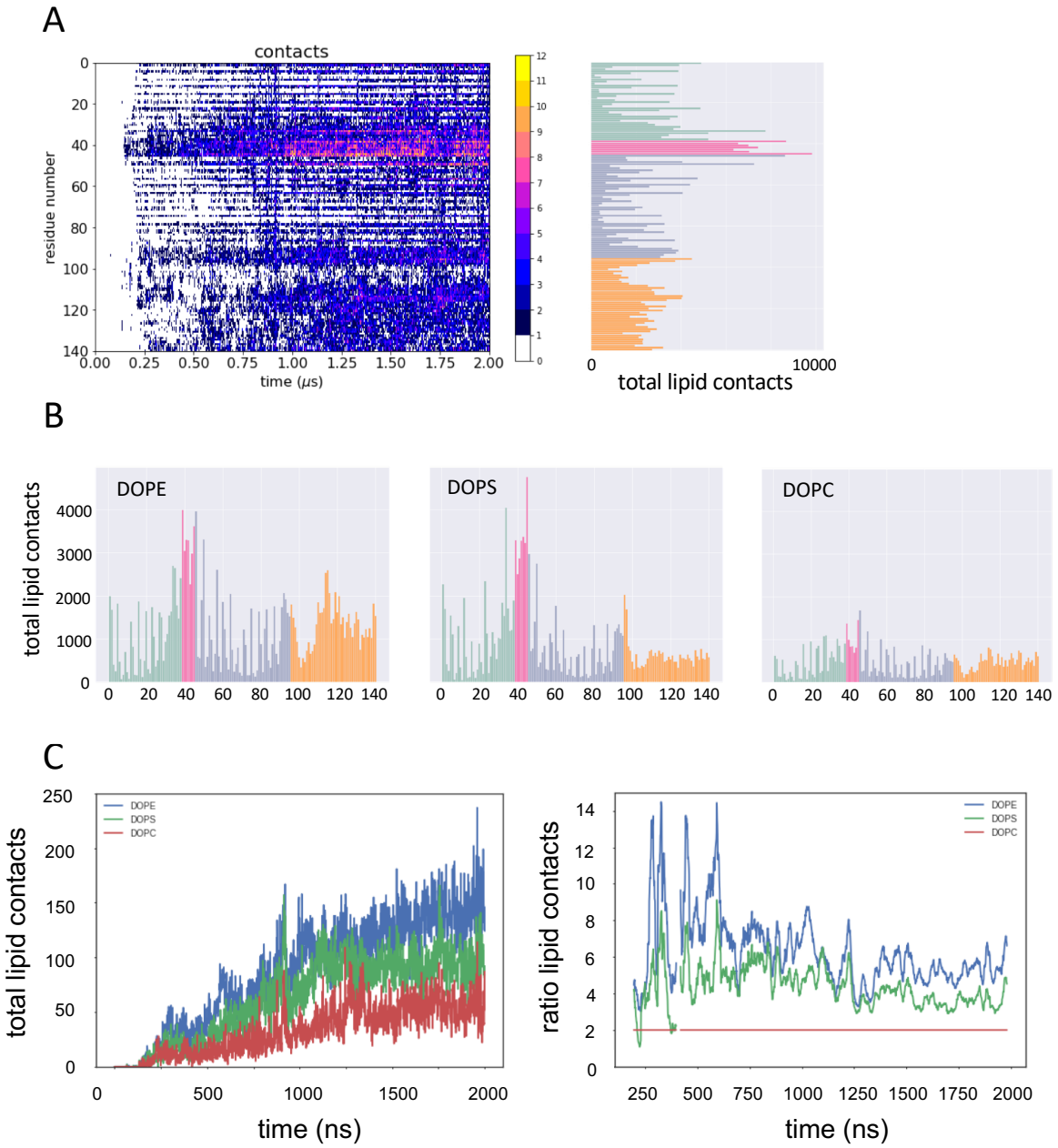


Fig. 5.5: CG simulations of the interaction of α-synuclein with a DOPC/DOPE/DOPS 2:5:3 membrane. Each conformation was simulated for 10 replicates of 2 μs. A: Contacts for with all lipids for each residue of -synuclein over the course of 10 simulation replicates. Colours on the histogram indicate structural regions. A residue is considered in contact if the residue backbone particle is within 0.7 nm of a lipid particle. B: Histograms of contacts over all 10 replicates for each lipid species. C: Total lipid contacts for each species shown over time (see legend), and this data is also expressed as ratios in D. These ratios have been presented as ratios with respect to the DOPC concentration (red) set to 2 for easy comparison with the bulk 2:5:3 composition.

Fig.5.5B shows the individual contribution of each lipid species to the contact profile. Residue 46 makes more contacts with DOPS than DOPE. DOPS also makes a large number of contacts with residue 34. The number of contacts made by the inter-helical region with DOPE and DOPS appear to be similar, despite DOPS being at a lower concentration in the membrane. The disordered C-terminus however appears to interact preferentially with DOPE over DOPS.

Fig.5.5C shows the total number of contacts made with each species over all replicates. In order to see if α-synuclein shows a preference for any lipid species, these totals are expressed as a ratio (fig.5.5B) over a moving average of 20 frames (2 ns). In the initial stages of interaction, there is a preference for DOPE and DOPS over DOPC. However, this ratio approaches the bulk 5:3:2 composition over the course of all simulations. Simulations show a preference for DOPE and DOPS in the initial interaction (fig.5.5C). However, density map analysis RDFs show no specific hotspots for lipid interactions (Fig.5.6).

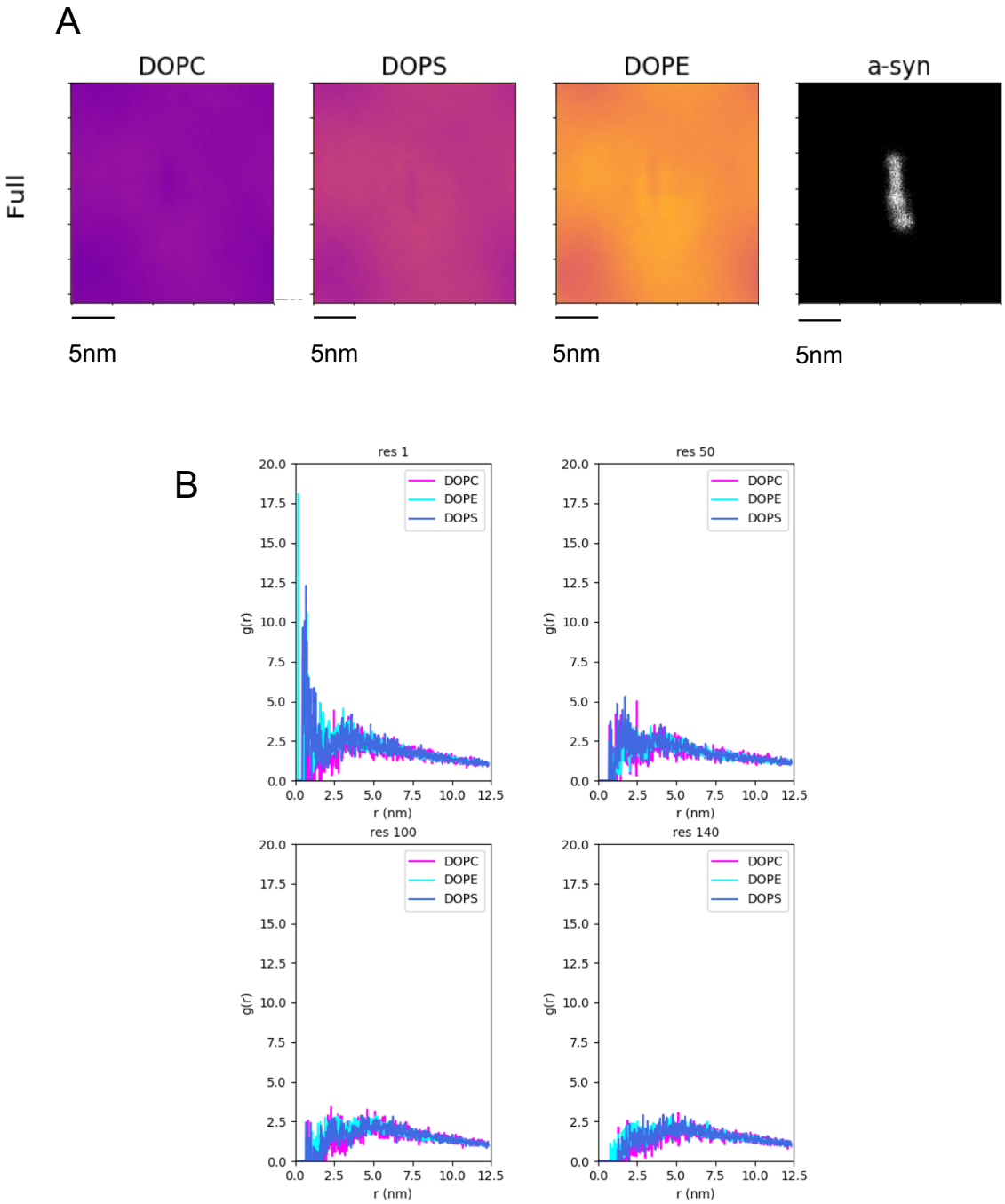


Fig. 5.6: Coarse grained simulations of α-synuclein with a mixed DOP membrane. A: Averaged density map for lipids. All replicates were fitted to the interhelical structure. Purple indicates low density, and yellow indicates high density for the lipids. No hotspots of density for any lipids can be seen. The position of the fitted α-synuclein structure is shown in the black box, where white indicates particle density. Here the long helical structure of helix 1 with helix 2 folded back over can be clearly seen. B: RDFs of lipids around residues 1, 50, 100, 140 for one trajectory where colours indicate lipid species. Whilst there is increased density immediately around residue 1 (known to make many contacts with the membrane, see fig.5.3) and the $g(r)$ value for DOPE appears to be quite high, the other residues appear not to show a particular preference.

We have shown which residues are important for α-synuclein binding to POPG and a mixed DOP membrane, and that the helicity of the protein greatly influences the binding pattern. However, it seems that residues 1, 20, 79 and 80 are important for binding regardless of helicity and lipid composition. Other important contacts are in the region of helix-1 and the interhelical region. A number of known mutations are found in this region, including A30P and E46K [229].

5.3.3 Atomistic simulations show great variation in the initial binding sequence and corroborate experimental data

To verify the CG simulation findings, and to further explore the structural changes that take place on binding to the membrane, we simulated 3 conformations of α-synuclein binding to the membrane. We chose 3 conformations from the CG simulations and backmapped them to AT resolution using CG2AT [145]. We then simulated 10 replicates of each conformation for 100-250 ns, resulting in an aggregated simulation time of 5.5 μs.

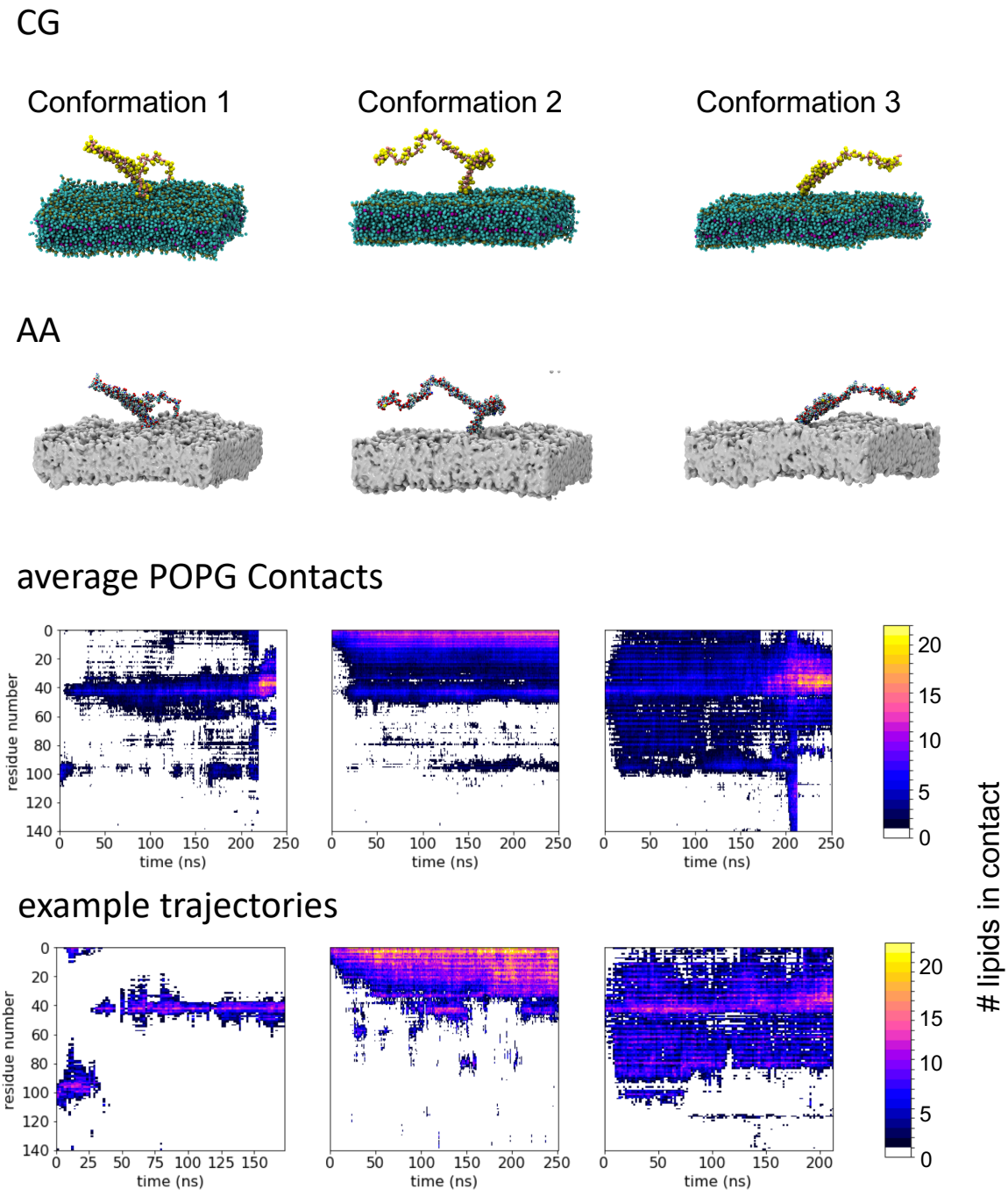


Fig. 5.7: Atomistic simulations of α-synuclein with a POPG membrane. Three initial binding poses from the CG simulations were chosen to capture initial binding from the C-terminus (left), N-terminus (centre) and inter-helical region (right) and converted to atomistic resolution. Each conformation was simulated for 100-250 ns across 10 replicates, resulting in an aggregated simulation time of 5.5 μs. The number of lipids within a 1 nm radius of the Cα atom of each residue was counted at each time point. The total number of lipids in contact with each residue across all replicates is shown with each residue over time. Note that not all simulations were the same length, so later time steps are averaged over fewer than 10 trajectories. Three example individual trajectories are shown from each starting conformation in the dataset.

The number of lipids within a 1 nm radius of the Cα atom of each residue was counted for each simulation frame and is shown in fig.5.7. (N.B. this cutoff was chosen in line with other studies [240] and is thought to give a general indication of lipids in proximity for simplicity; optimisations to this are discussed in future work). Contact analysis shows the interhelical region generally binds first, with the disordered C-terminus making transient contact with the membrane (fig.5.6). Where the C-terminus is initially in contact, this is followed by reorientation to bring the interhelical region close to the membrane. An interesting result is that the contact profile and bilayer penetration is very variable across the ensemble; it appears that the α-synuclein monomer prefers to bind initially through the interhelical region, but that there is some flexibility in the binding pose.

In order to assess structural changes of the α-synuclein monomer on the POPG membrane, we carried out secondary structure analysis using DSSP (see fig.5.8).

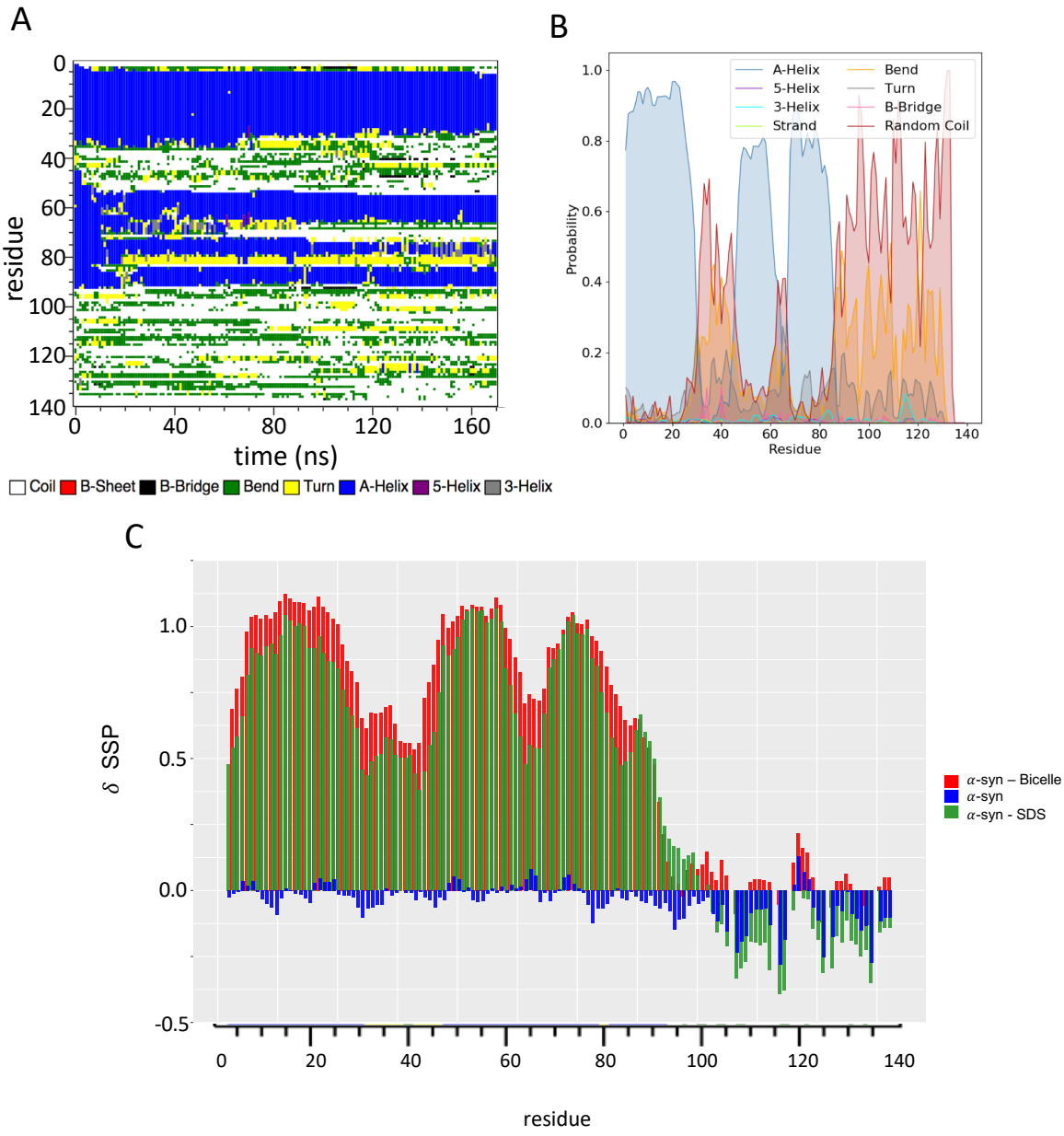


Fig. 5.8: Secondary structure analysis of the α-synuclein monomer. A: DSSP analysis over time for one representative trajectory from conformation 1. The disruption of helical structure in helix 2 is observed. B: Probability of secondary structure for each residue for all simulations of conformation 1. α-helix propensity, shown in blue, is clearly reduced in favour of random coil between residues 60-70. C: Chemical shift data provided by colleagues at UCB. The chemical shift indexing [241, 242] shows field-shifted atoms in the region between residues 60-75, indicating a change in propensity to α-helical structure.

We frequently observed a break in helix 2 over the course of the simulation (fig.5.8A).

A representative trajectory shows helix 2 (residues 46-95) breaks between residues 60-85. Fig.7B shows the probability of all structural elements over 10 replicates. It shows that 60-70 have a greater probability of existing as random coil than α-helix. This data supports NMR chemical shift data (fig.5.8C) which suggests a break in helix 2 in the same region.

We also observed 4 trajectories where DSSP identified β-sheet structure in some residues, mostly in helix 2 and around the interhelical region (fig.5.9A).

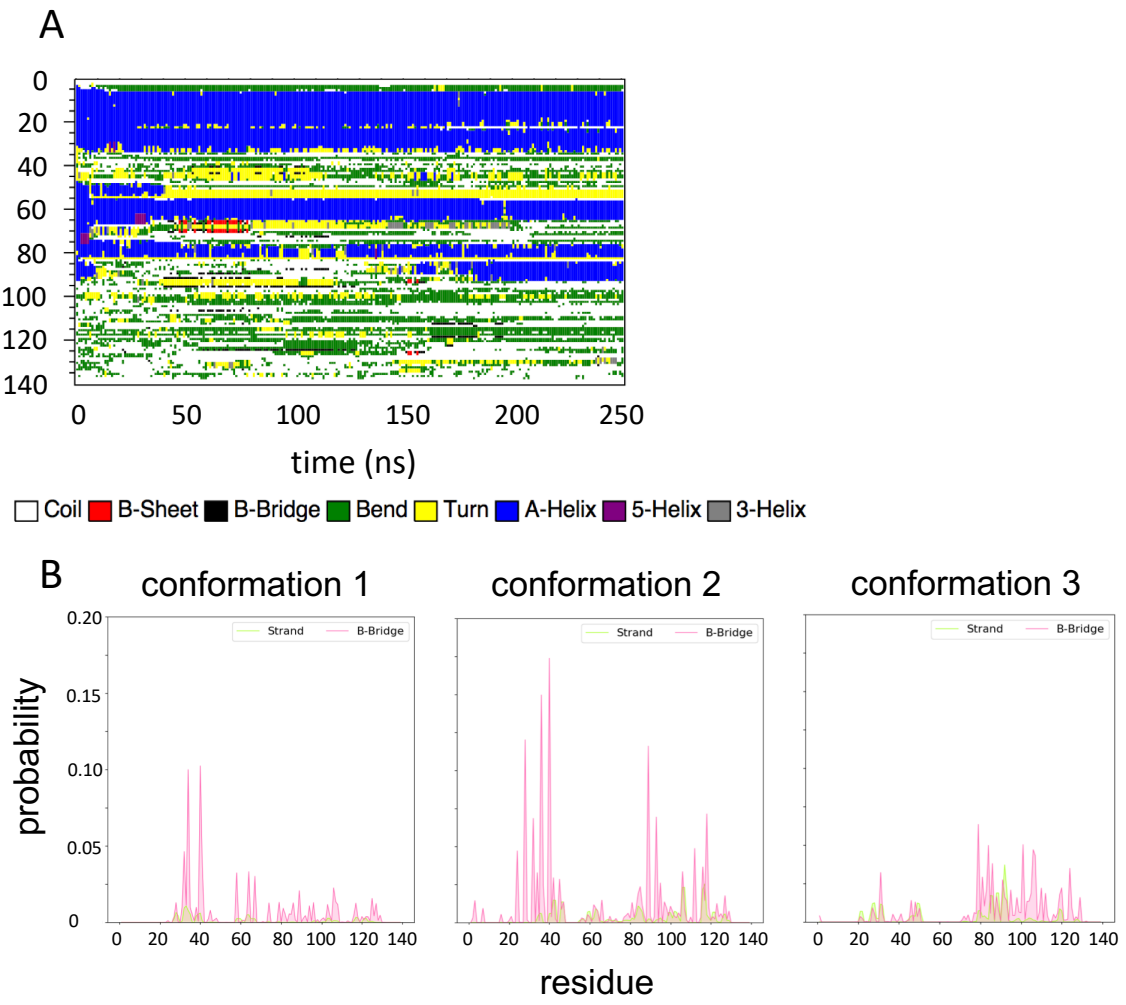


Fig. 5.9: Secondary structure analysis of the -synuclein monomer binding to a POPG membrane. A: DSSP analysis for one trajectory from conformation 2. The trajectories show β-sheet structure emerging around the interhelical region and in the NAC region. B: Probability of β-strand and β-bridge assignment across the three conformations. β-bridge is shown in pink and β-sheet is shown in green.

Fig.5.9B shows the probability of β-strand and β-bridge for each conformer dataset. All conformer datasets show a probability of β-structure of around 1-17% around the interhelical region, the NAC region, and in the C-terminus. Interestingly, AT simulations of α-synuclein binding to a POPG membrane shows hints of β-structure formation without the use of any adaptive or biased sampling. Taken with the CG results which show helix 2 laid over the top of helix 1, we postulate that β-structure is seeded

in the inter-helical region or NAC region, which initiates hydrophobic aggregation.

Fig.5.10 demonstrates the intramolecular contacts within the bound α-synuclein as analysed by AT simulations which are represented as heat maps.

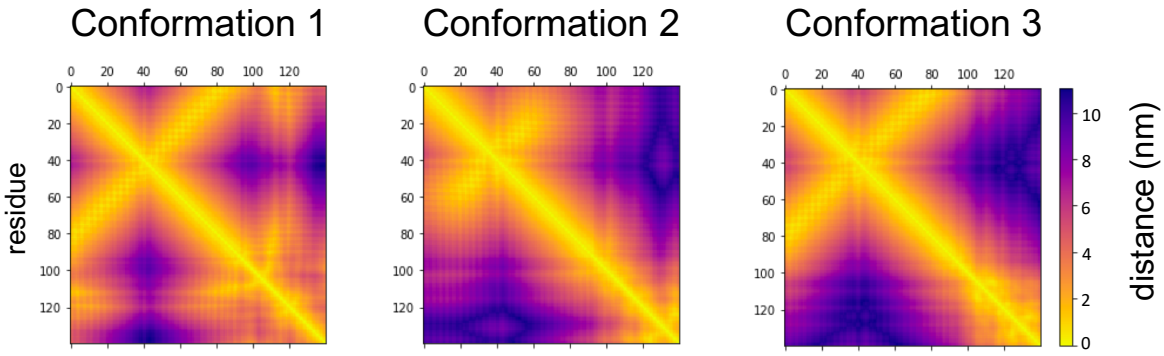


Fig. 5.10: Residue-residue distances of the α-synuclein monomer binding to a POPG membrane. A: Averaged residue-residue distances for three trajectories from conformations 1, 2, and 3. Yellow indicates residues are close to each other.

In all simulations, the diagonal away from the autocorrelation shows the helix 1 and 2 lying adjacent to each other; interestingly there is some loss of this linkage in conformation 2, where the simulations show the α-synuclein monomer binding to the POPG membrane with the N-terminus first. We suggest that this mode of binding somehow abrogates the structure of helix 1, and this is supported by the lipid contact profiles.

5.3.4 Markov state models provide a view for understanding the initial binding of a-synuclein

To further investigate the structural changes of the α-synuclein monomer on the POPG membrane, we constructed Markov state models using MSMBuilder [157]. Firstly, we attempted to construct an MSM based on all replicates of the three conformations, but tICA analysis (see chapter 2) of states shows that the three different replicates of the full structure cannot be analysed together on dihedrals only (fig.5.11); their

states are too dissimilar such that an MSM cannot identify transitions between them due to the lack of ergodicity. This may however imply that the dihedral angles of the α-synuclein when bound to the membrane depend on their mode of binding, which is an interesting finding in itself.

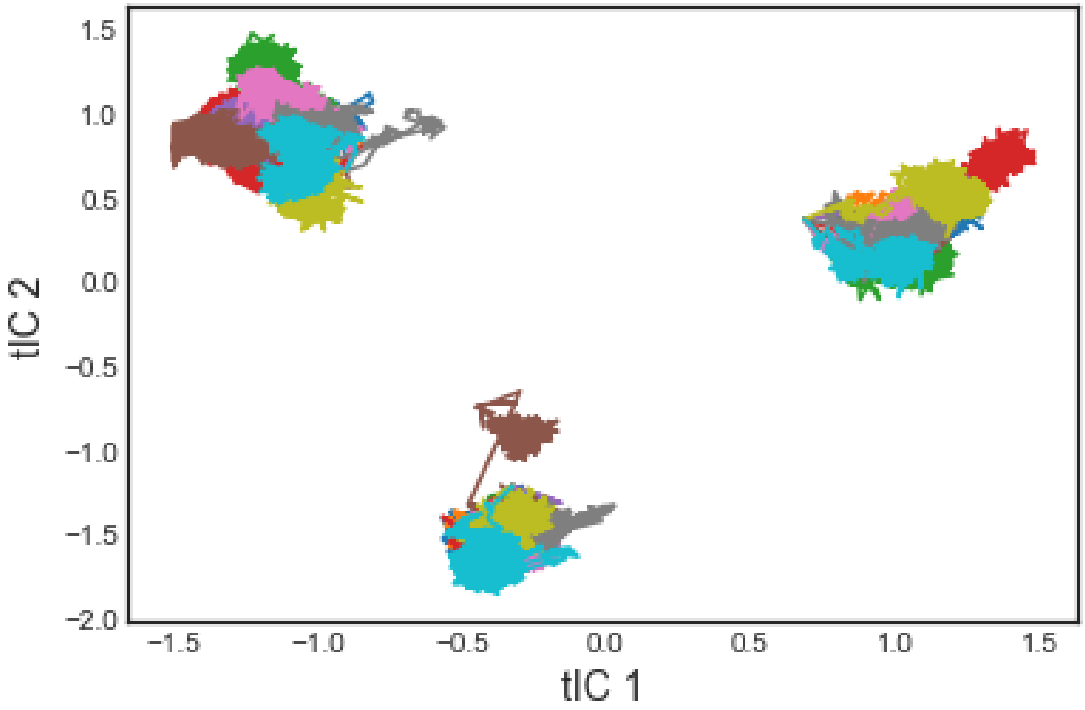


Fig. 5.11: tICA analysis of all atomistic simulations based on dihedral featurisation. The dataset is shown projected onto the first two tICs. Colours indicate the contribution of individual replicates. The tICA analysis identifies three regions, corresponding to the three initial conformations. That there is no crossover between these three regions means that it is not possible to build an MSM over all states in the dataset because there are no transitions between them.

To formulate a simpler problem, I constructed an MSM based on only residues 60-70. This was chosen because we wanted to see if an MSM would capture a break in the helix, as indicated by figs.5.8-9. We constructed an MSM featurised on the Cα contacts of the residues (we initially trialled dihedral angles, but using Cα contacts gave better convergence in the implied timescales analysis). We plotted the implied

timescales to check for convergence and chose an MSM lag time of 2 ns for model construction (fig.5.12A).

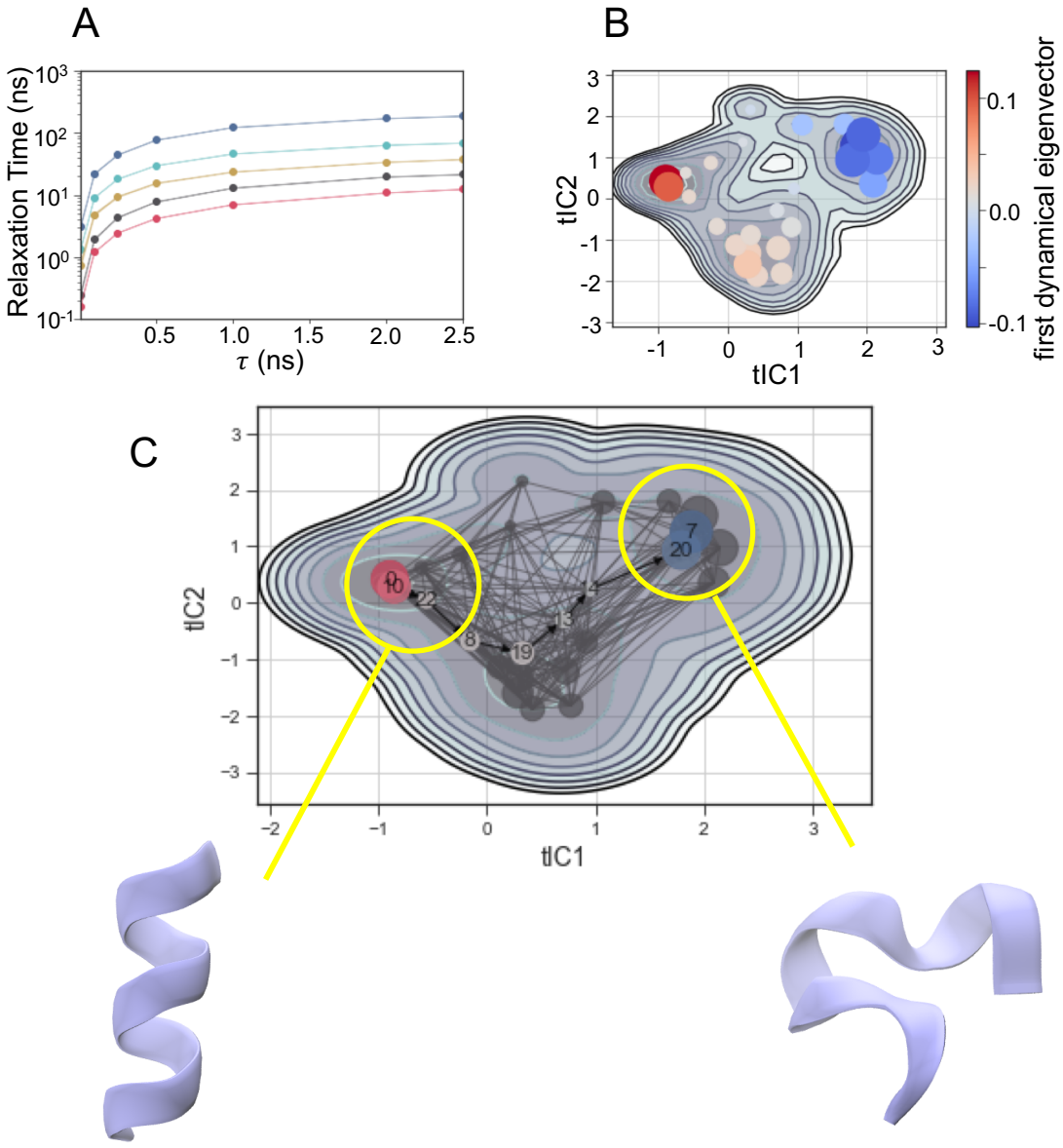


Fig. 5.12: Markov state model of residues 60-70. A: Implied timescales show processes converge at a lag time of 200 frames (2ns). B: Slowest motion as represented by the 1st dynamical eigenvector from red states to blue. C: MSM microstate transitions and representative structures. Yellow circles indicate the two lowest energy macrostate centres from which the representative structures are extracted.

Fig.5.12B shows the clustered states coloured by the first eigenvector, which represents the slowest transition from red states to blue. Finally we clustered these microstates into six macrostates. An appropriate number of macrostates is one more than the number of timescales above the major gap [98]; in this case the timescales would indicate choosing either two or three. However, the energy landscape indicates there may be more than three (see positioning of microstates in fig.5.12B) and it has been recommended to err on the side of having too many microstates rather than too few [98]; therefore I arrived at five microstates from examining the pseudo-free energy landscape. I generated representative coordinates from the two lowest energy states and visualised them in VMD. The MSM captures a transition from a helical structure to a helix which breaks and folds back on itself. We present the atomistic representations of these structures in fig.5.13 with residues labelled. A large conformational change in this region is demonstrated in line with the NMR shift data.

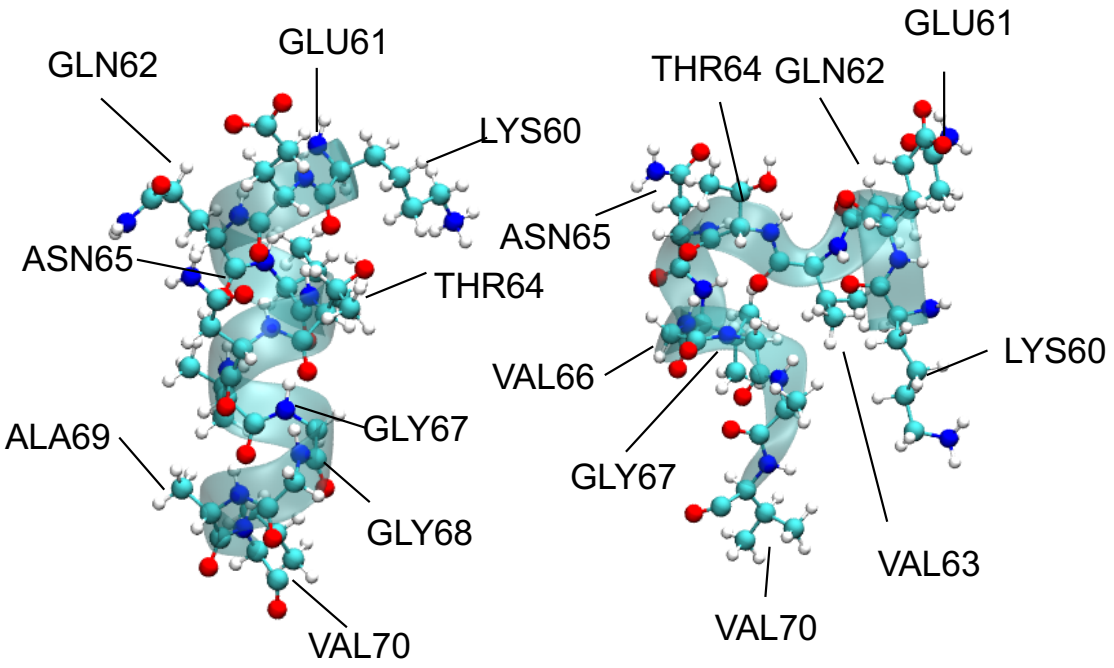


Fig. 5.13: Model of the two structures derived from the MSM. A translucent cartoon representation of the secondary structure is overlaid with CPK representation of the residues. Colours indicate the atom type: cyan = C, blue = N, red = O, white = H. The first structure shows the α-helical structure of residues 60-70. This is generated from a region that is structurally close to the conformation at the start of the trajectory so helical structure is maintained. The second structure shows a large break in the helix, where the structure begins to turn back on itself. Residues are labelled.

This turn structure is interesting because the amyloid structures solved by cryo-EM feature a number of turning regions [43–45]; it is an attractive hypothesis that this initial turn may seed the initial amyloid formation event by either transmitting conformational changes to the region containing the pathogenic mutations. I have mapped the region of residue 60-70 which demonstrates a turn in the MSM onto the PDB structures 1XQ8 [243] and 6FLT [45] as shown in fig.5.14. The region used in the MSM which results in a turn structure is clearly shown also to have a turning structure in the cryo-EM structure of the α-synuclein fibril unit. I suggest that this helix break shown in the MSM and NMR data may propagate further structural

changes that expose residues upstream and downstream of these positions to other monomers present on the membrane. This could seed the formation of the protofibril and its subsequent extension to full toxic fibril. A number of important mutations are upstream of this region and these residues fall in the interface between two monomers when in fibril formation; it may be that the helix break in wild-type α-synuclein may be less harmful than in the pathogenic form. One might speculate that the physical properties of the mutated residues would result in more favourable interactions with other monomers and therefore greater fibrillisation.

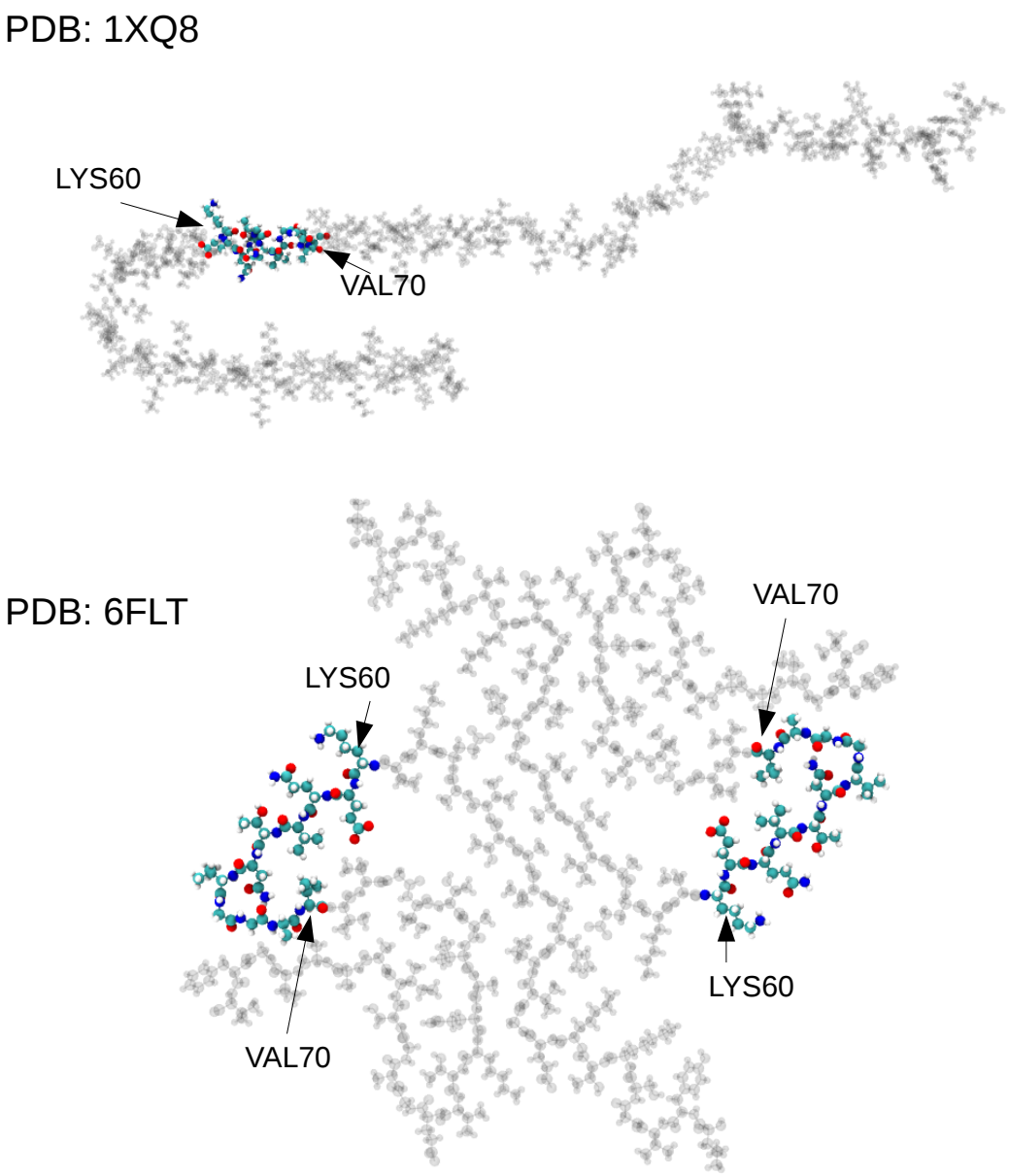


Fig. 5.14: Residues of the MSM mapped back onto PDB structures 1XQ8 and 6FLT. A translucent representation of the structure is overlaid with ball and stick representation of the residues 60-70. Colours indicate the atom type: cyan = C, blue = N, red = O, white = H. The first structure shows the α-helical structure of residues 60-70 in the context of the micelle-bound human α-synuclein as solved by NMR (PDB: 1XQ8). The second structure shows the residues 60-70 in the context of the cryo-EM structure of the α-synuclein fibrils. The structure shows two α-synuclein monomers forming a flat folded motif; in the cryo-EM structure, multiple copies of this motif are stacked on top of each other, forming the fibril. The mapping shows that the region found to turn in the MSM is also found to be a turning region in the cryo-EM structure.

This section has demonstrated that it is still challenging to build converged MSMs for large systems; in this case, the irreversibility of the binding process means that the system is not ergodic. Where some approaches have trimmed states in order to ensure that calculations are converged [244], this can result in loss of information and misleading models (see [156] for a discussion regarding reversibility and detailed balance). However, we have been able to formulate a simpler reduced problem (that of the helix 2 break) and this has been well addressed by the MSM.

More broadly, in Sirur *et al.* [245] the PEVK segment of titin (another protein with large disordered segments) was observed to transition from partially misfolded intermediates to partially folded intermediates without transitioning through a fully unfolded state. This insight may be useful in thinking about α-synuclein because it would seem that a more simplified view, where folded and misfolded states are thought to be reached from the fully unfolded state, would be unlikely in a membrane bound state. A common theory in the amyloid field is that fibril aggregation is initiated by hydrophobic collapse into disordered aggregates followed by β-sheet transition. However, MSMs of amyloid β protein have shown that this appears not to be the case [246], and that the hydrophobic collapse is a consequence of overly high peptide concentrations and overstabilisation of inter-protein interactions due to some forcefield parameters.

In this study we appear to see a break in the helix 2 which transforms into a turning structure. A recent study [247] presented dihedral angle analysis in the region of helix 2, albeit focused on residues 70-80, where simulations also showed peptide plane flipping in the region, which provides an interesting parallel to this study and suggests the MSM should be extended to include these residues. There do not appear to be any transitions directly from the starting state to this break structure, so it would seem that there is some kind of pathway (indeed as shown in Fig.5.12). In the context

of Sengupta *et al.* [246], it remains an open question of whether or not the β-structure formation is indeed 'off pathway', and more investigations will need to be done into the MSM to fully establish how the pathway contributes to β-structure formation. Finally, a recent study showed that a small molecule SynuClean-D interacts primarily with residues in the region of residues 55-73, over the region we have built the model, to abrograte fibril formation and the progression of PD in animal models [248].

Further work related to this study could include further optimisation of the contact distance cutoff to account for sidechains rather than Cα atoms to give a more nuanced picture; this approach currently may overestimate the number of lipids within the radius of the small amino acid residues and underestimate the contribution from lipids around residues with very large sidechains. However, the data thus far appears to be consistent with experimental data and has provided useful information. Furthermore, forcefields are always a source of debate. We consider the use of the CHARMM36 forcefield, which was the most appropriate available in the CG2AT method, remains appropriate in modelling a peptide with both ordered and disordered regions. Additionally, McKiernan *et al.* have shown previously that AMBER ff99SB-ILDN and CHARMM22* have a tendency to overstabilise the native state in MSMs of CLN025 β-hairpin formation. Future work may include a comparison of these forcefields.

A limitation of this study has been that it has not been possible to build converged MSMs of the whole α-synuclein structure. This could be addressed in two ways: firstly by implementing adaptive sampling to 'reseed' simulations in less well explored areas, and secondly to implement methods that do not assume ergodicity. This makes sense as we can expect some large conformational transitions from the start of initial binding that may not necessarily be sampled again. An approach which has taken advantage of this is in work carried out on titin misfolding by Sirur *et al* [245]. Finally we have not here implemented any optimized parameter selection or cross-validation, and

this would be useful to do in future [249]; however this can be more challenging to implement in MSMBuilder than PyEmma.

5.4 Conclusion

We have performed a range of CG and AA simulations to investigate the membrane-binding properties of α-synuclein. We have shown through CG simulations the important contacts for binding and how these contacts are affected by protein helicity and lipid composition. We have performed a large ensemble of simulations of α-synuclein binding to the membrane and confirmed the finding that it is usually the interhelical region that leads binding. However, there is great variablility in contact progression between simulations depending on the orientation of the initial encounter. We have also used DSSP to calculate secondary structure propensities and shown that there are small propensities for β structure in a number of important regions. Taken together with the secondary structural studies and the NMR chemical shift data, we have been able to show how helix 2 of α-synuclein breaks and can begin to form β-structure elements.

The Markov state models have been used to establish a converged model that is able to capture this transition and generate representative structures. These structures are able to provide a model of how this helix changes structure and this may help to explain how toxic aggregation via formation of oligomers happens. We would extend this study by implementing some more advanced MSM methods and longer simulations to address convergence of the systems, and also perform atomistic simulations of more than one monomer. This may begin to give some insight of how toxic oligomers form. Ultimately these structural insights may help with the design and development of novel therapeutic approaches.

Chapter 6

Large-scale Simulations of PIP5K1A in Mixed Lipid Membranes

"All models are wrong but some are useful."

George Box

6.1 Introduction

Simulations of protein-membrane systems can be viewed as computational microscopes [250], and as such are able to investigate structural phenomena with necessary spatiotemporal resolutions. Over the previous chapters, we considered atomistic or small-scale CG systems. Ultimately however, we want to be able to understand the larger time and lengthscales of cellular systems and therefore we need to build large-scale models of these processes.

CG simulations of membrane proteins in large lipid membranes have been used to

investigate many cell processes. One advance in this respect is that there are now CG parameters available for many different lipid species and this enables the simulation of more complex membranes that are closer in composition to physiological systems [208]. It can often be challenging to ascertain the most appropriate lipid composition for a given system from lipidomics [251] and sometimes a choice must be made between simulating what is physiologically most appropriate or choosing a composition for accurate comparison with experimental studies. However, more studies have been carried out which demonstrate the use of complex membranes in large-scale membrane protein simulations [197, 252].

Various physical phenomena have been investigated by large-scale simulations for transmembrane-proteins. Studies have shown that lipids move in concerted motion, and that these correlated motions can extend up to 10 nm [253]. Additionally, protein crowding appears to have a large effect on diffusion and lateral mobility [201, 254]. There is a complex interplay between lipid composition and membrane curvature, and together these may act as regulatory mechanisms [255, 256].

Another advantage of this large-scale approach is that it enables simulations to approach experimental time- and length-scales. Advances in simulation methodologies and hardware have complemented the advances in super-resolution microscopy techniques. For example, stimulated emission depletion (STED) far-field fluorescence nanoscopy can be used to detect that sphingolipids and glycosylphosphatidylinositol-anchored proteins are transiently associated in cholesterol-mediated molecular complexes [257]. Another study showed the use of super-resolution scanning STED microscopy with scanning fluorescence correlation spectroscopy to identify transient interaction hotspots for sphingomyelin analogues [258]. Microscopy approaches have also been used to great effect more specifically in synaptic membranes, especially the organisation of AMPA receptors and PSD95 (a scaffolding protein) [259]. These find-

ings have advanced the conceptual steps of neurosynaptic transmission and function. The experimental and simulation techniques can be applied to *in vitro* and *in silico* data in tandem to result in elegant studies which offer the best of both worlds. Sometimes both *in vitro* and *in silico* data can be collected in such a way that they can be analysed with the same analysis program as shown in Rassam *et al.*, 2015 [260].

GPCRs have been well-studied through the large scale simulation approach, including rhodopsin [261] and opioid receptors [262]. Simulations of the β_2-adrenergic receptor showed that ordered arrays of the receptor are dependent on hydrophobic mismatch, and that this pattern of mismatch is very different to that of the β_1-adrenergic receptor, a finding that is commensurate with experimental studies. This implies that the different spatial organisation of these proteins may have some mechanistic effect. Studies have also shown the importance of protein-lipid interactions on GPCR organisation in the membrane [263]. Interestingly, a number of studies have also shown the effect of clustering on membrane curvature, including simulations of the ATP-synthase dimer [264]. Combined with electron microscopy, Davies *et al.* showed that ATP-synthase dimers are found along the highly curved christae ridges, and showed through simulations that the dimer induced plastic curvature in the membrane; they also propose that the curvature is a result of the elastic energy of the membrane rather than direct contacts.

Peripheral membrane proteins have been less well studied with large-scale simulations thus far, due to the requirement for larger box size and increased calculations, and due to many PMPs being less well studied. Some early studies of the lipidated membrane-targeting motif of H-Ras were shown to form exchanging clusters of four to ten molecules [265]. Simulations of N-Ras demonstrated that the protein interacts preferentially with the interface between ordered and disordered lipid domains, and that this reduces the line tension between domains. The limitations of this study include

that the C-terminus structure is poorly understood (and this is where the membrane-associating anchors are found) and that the membrane contained doubly-unsaturated lipid tails. This highlights the difficulty with simulating PMP systems as often the models are based on a large number of assumptions and resources often do not permit exhaustive simulations. It has also been established that Bin/Amphiphysin/Rvs (BAR) domains sense, generate, and regulate curvature [266], and multiscale simulations have been used to show how F-BAR proteins form regular lattices in line with electron microscopy data. Simulations also conversely showed that increased membrane surface tension alters the geometry of protein association [267], and pose the question of how these proteins depend on their physical environment [268]. These studies highlight the complex interplay between generation, recognition, and regulation of curvature by proteins. Finally, a CG simulation of 48 α-synuclein monomers (residues 1-100 only) showed that the membrane was remodelled within 40 ns of simulation time, and speculated that different partition depths in different membrane mixtures had a large effect on membrane remodelling [269]. However, these simulations did not include the full length of α-synuclein and modelled the N-terminus as an extended helix, rather than two antiparallel helices.

One of the logistical limitations of large-scale simulations is the amount of data they produce; the simulations carried out for this chapter alone produced > 6 Tb of data in coordinate files alone. New methods have been developed to analyse and visualise these systems, but new challenges arise when using tools developed on planar membranes as these often assume a flat bilayer; for example, when calculating distances, often only x and y coordinates may have been sufficient to describe a planar system but this breaks down when dealing with high degrees of curvature.

This chapter comprises work carried out on large-scale simulations of the PIP5K1A kinase. I aimed to investigate the relationship between lipid composition and PIP5K1A

binding, and to investigate if curvature and clustering play any part in this process (as was speculated in chapter 3). This chapter offers a perspective on the overall signalling event. Although the conclusions are more speculative, they offer interesting directions for future study. The main contribution of this chapter is the effective use of HDBSCAN [270] for investigating lipid clustering, and the finding of large degrees of curvature in the complex membrane mixture. Both of these findings represent important starting points for further understanding of PMP signalling events.

6.2 Methods

Monomer simulations were set up by taking the initial starting conformations used in chapter 1 for the CG binding simulations and replicating the starting structure into a 5x5 grid using the GROMACS genconf tool. The simulations were energy minimised for 10000 steps, followed by solvation and addition of ions at 150 mM NaCl, and then equilibrated for 1 ns with position restraints placed on a central atom of the PIP5K1A molecule. This enabled the proteins to assume a random orientation with regard to the membrane without being able to approach it. The whole system was then simulated for 10 μs using a 20 fs timestep.

Dimer simulations were initially set up by generating a large membrane using the INSANE method [238]. Dimeric PIP5K1A structures were taken from studies in chapter 1 and replicated using genconf into a 3x3 grid. The coordinate files were concatenated. The protein/lipid system was energy minimised for 10000 steps before solvation and addition of ions at 150 mM NaCl. The system was again equilibrated for 1ns with position restraints placed on a central atom of the PIP5K1A molecule.

Electrostatic forces were calculated using a reaction field potential with a cutoff at 1.2 nm. van der Waals interactions were calculated between all beads separated by less than 1.2 nm, with the potential switched from 0.9 nm. The Verlet cutoff scheme

was used. The temperature of the system was maintained using a velocity rescale thermostat with a time constant of 1 ps. All simulations were run at 323 K with pressure at 1 bar using a semi-isotropic Berendsen barostat with a time constant of 2 ps and a compressibility of 3×10^{-4} bar^{-1}. Coordinates were written out every 200–600 ps.

All simulations were run on the ARCHER supercomputing service (https://www.archer.ac.uk).

Simulation analysis was carried out using in house scripts written in python (see www.github.com/sbtaamos). Cutoffs for the monomer simulations were established by measuring the COM distance of two monomers in the dimeric form and determined to be 5.5 nm. Cutoffs were tested for optimal distance (data not shown).

Lipid clustering was determined using a python implementation of the HDBSCAN algorithm (see www.github.com/sbtaamos) using a minimum cluster size of 5. Only lipids with phosphate headgroups within 4.5 nm of the catalytic site (simulation residue 182) were counted.

6.3 Results & Discussion

We performed MD simulations of 25 PIP5K1A monomers positioned 4 nm away from a model membrane (see fig.6.1A). We compared 3 different membrane compositions: POPC/POPS/PI4P, POPC/POPS/PIP2, and POPC/POPS/PI4P/PIP2. Each simulation box was 65 x 65 x 23 nm in size and contained approximately 840000 particles. Each system contained approximately 14000 lipids. Two replicates of each system were simulated for 10 μs, resulting in an aggregate simulation time of 60 μs. A table of simulations performed is shown below:

composition	monomers	lipid ratios	time (μs)	n
POPC:POPS:PI4P	25	75:20:5	10	2
POPC:POPS:PIP2	25	75:20:5	10	2
POPC:POPS:PI4P:PIP2	25	70:20:5:5	10	2

Table 6.1: Table of large-scale PIP5K1A monomer simulations in simple symmetric composition bilayers.

6.3.1 Contact analysis identifies dimerisation sites in line with experimental data

Fig.6.1A shows an initial simulation setup conformation with the PIP5K1A monomers positioned 4 nm away from a membrane containing PIPs. These simulations were run to see how PIP5K1A interacted with the membrane in a simulated signalling event. Initial inspection of the trajectories revealed that the PIP5K1A monomers clustered on the membrane (fig.6.1B), forming dimers and higher order structures.

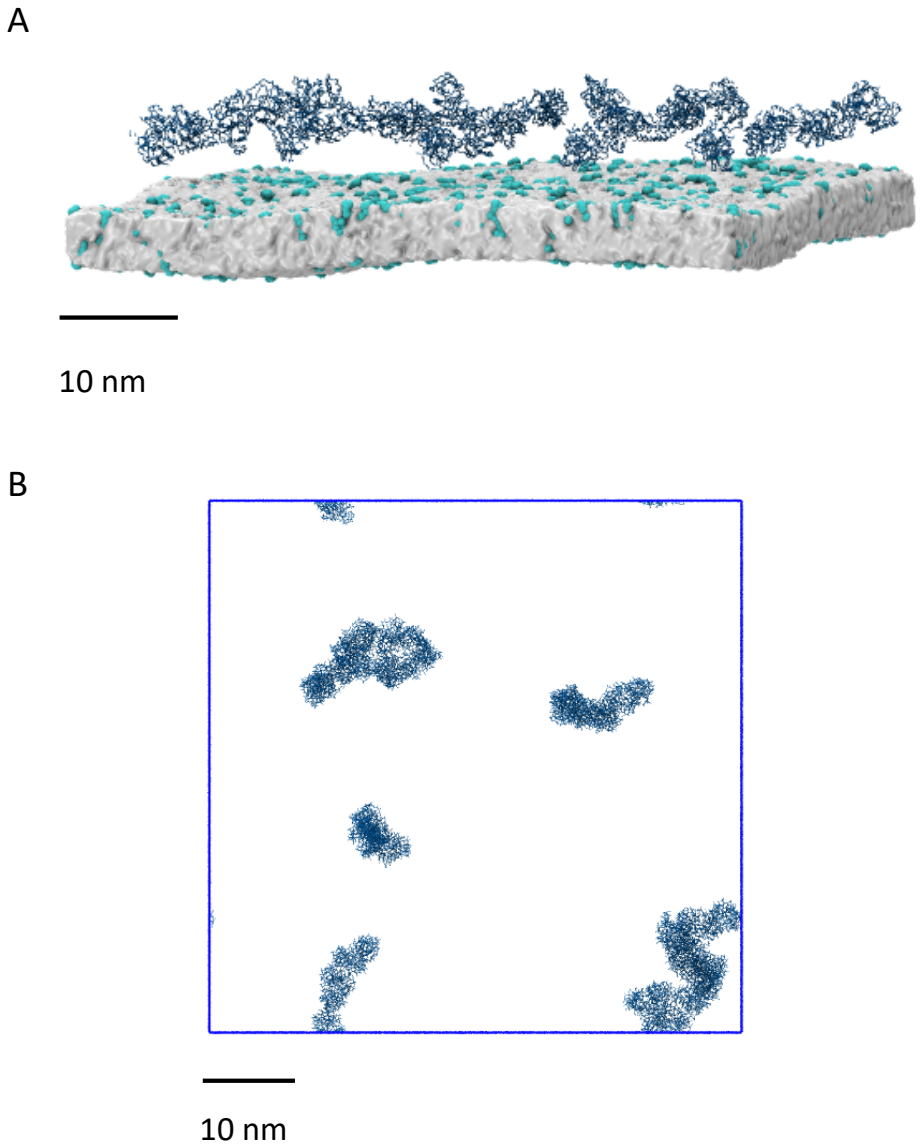


Fig. 6.1: Simulation setup and protein clustering. A: Initial simulation setup for an example trajectory (PIP5K1A-POPC/POPS/PI4P). Kinases are shown in dark blue. The membrane is shown as white (POPC and POPS) with PIPs highlighted in cyan. Monomers were positioned 4 nm away from the membrane surface and rotated randomly during equilibration. Solvent omitted for clarity. B: Top-down view of the final protein coordinates. The kinases are visualised in dark blue and the simulation box is shown as a blue line. Five clusters of kinases are seen of various sizes. Solvent and lipids omitted for clarity.

Next I investigated the protein residue-residue contacts in the PIP5K1A over the

simulation to see how the kinases made contact with each other. These interactions for one representative trajectory are shown in fig.6.2.

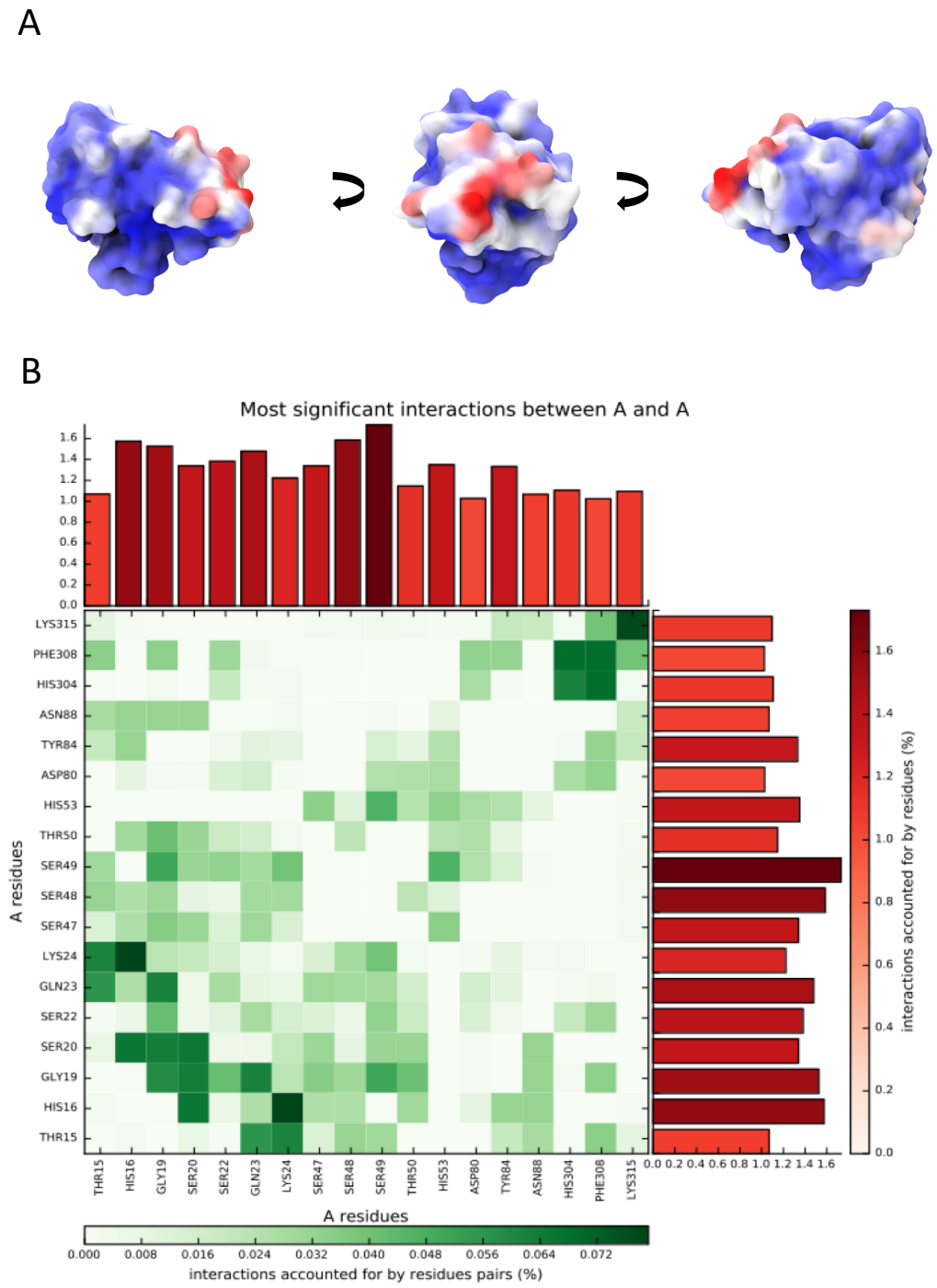


Fig. 6.2: Protein-protein interactions for dimers in one of the PI4P only replicates. A: PIP5K1A structure coloured by residue contact frequency. Red indicates region of more contacts. The activation loop can be seen protruding out from the bottom of the kinase in dark blue, showing it is not involved in dimerisation. Contacts are focused around B: Plot of most significant interactions between residues. The marginal histograms show the percentage of interactions accounted for by each residue. The green plot shows residue-pair interactions and the colour shows the percentage of interactions accounted for by the residue pairs. In this trajectory, SER49 makes 1.6% of interactions.

The top 18 residue interactions are shown in Fig.6.2. SER49 appears to make the most number of contacts with other residues, followed by SER48, GLY19 and HIS18 (see marginal histograms). These residues each account for $> 1.5\%$ of contacts observed. Additionally, the most important residue pair interactions are between LYS24 and HIS16, and LYS315-LYS315. By colouring a structure of PIP5K1A by this contact frequency we notice that the most contacts (red) are made in the region identified in the dimer interface of the PDB structure 4TZ7 [25]. I therefore consider that the model has accurately represented the physiological binding conformation.

6.3.2 The dimerisation interface is independent of lipid composition

To check that the lipid composition of the simulation did not affect the monomer-monomer binding pose, I analysed the contacts of all residues across the ensemble of simulations.

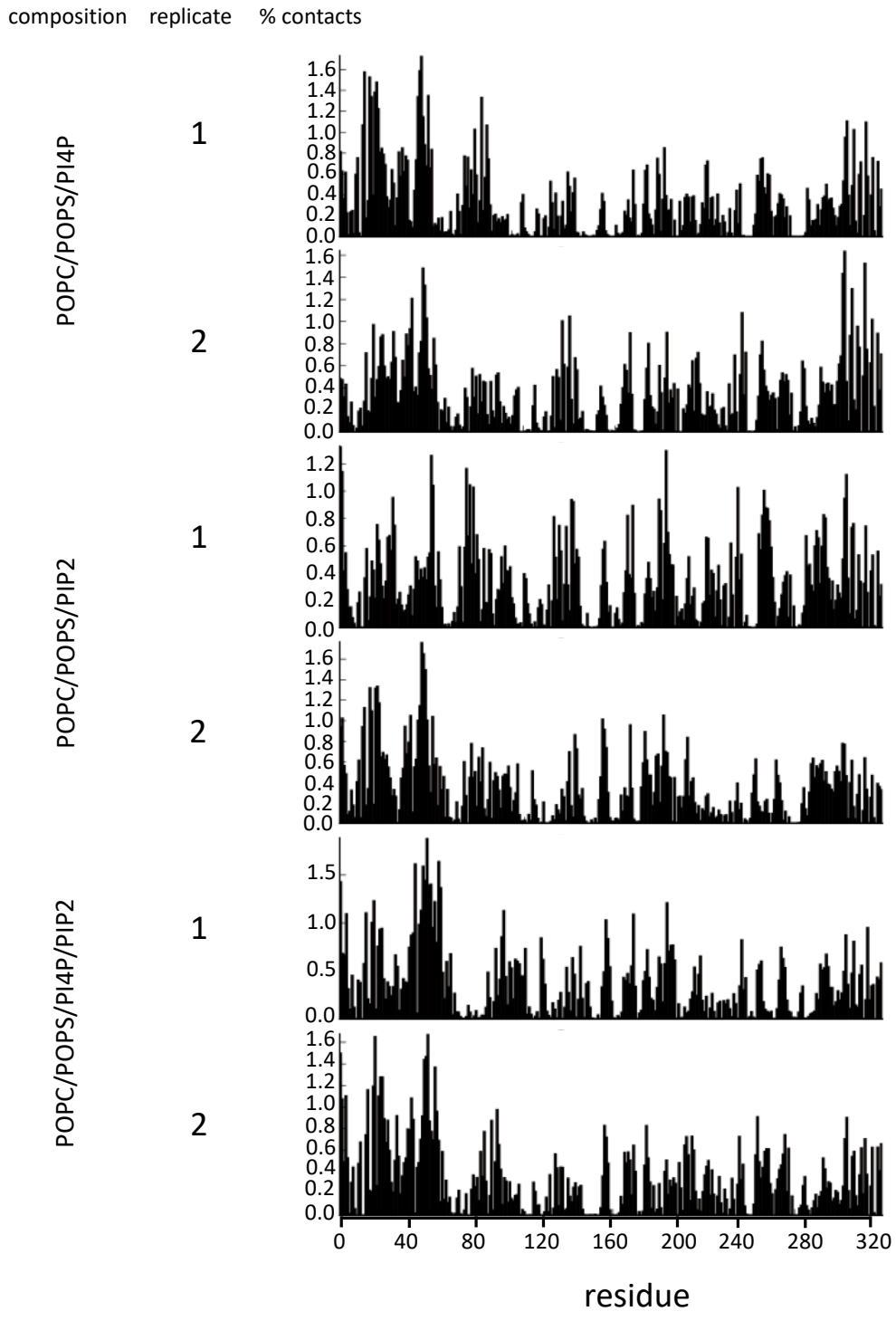
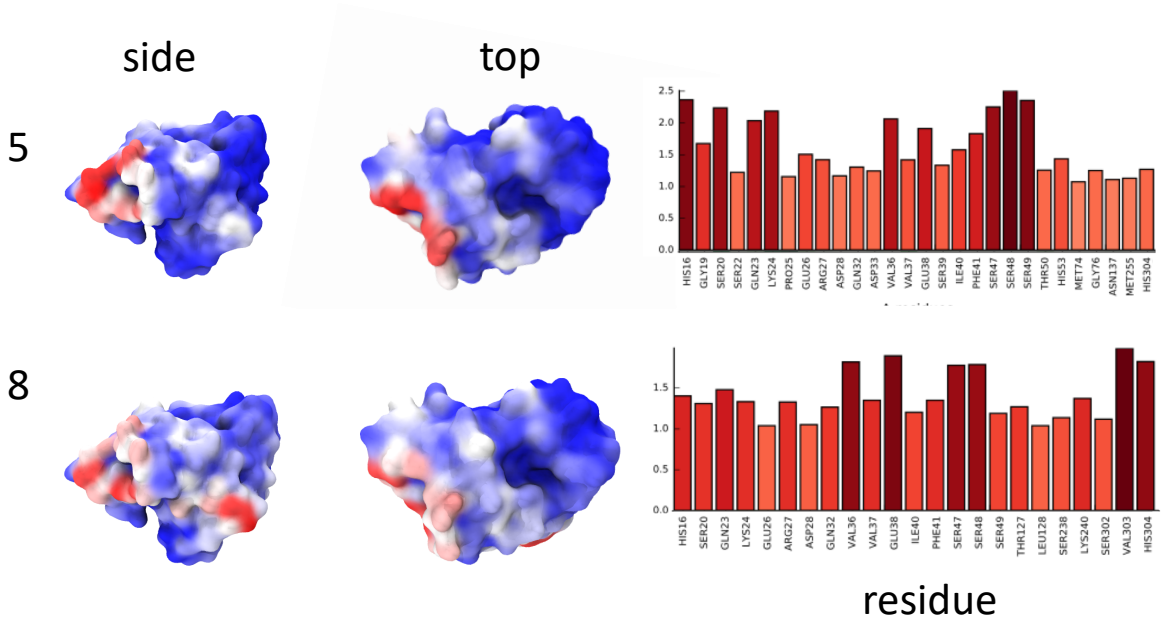


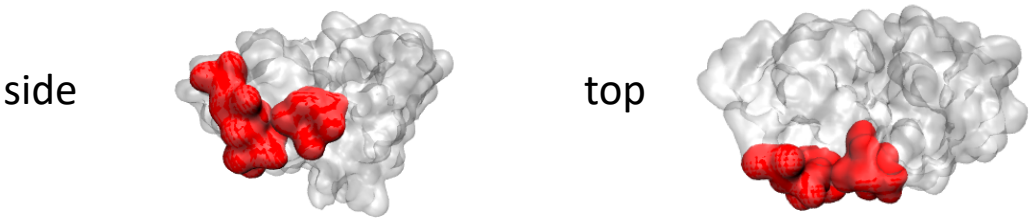
Fig. 6.3: Protein contacts made by each residue across all replicates. The residues are shown as a percentage of the total interactions they account for. Replicate number and PIP composition is labelled. The profiles appear to be similar across all trajectories, indicating that simulations show convergence and that the dimer composition does not affect the dimerisation interface.

The most number of interactions are around residues 15-30 and residues 45-60. There is also a contribution from the region residues 80-100. There are very few interactions recorded in the disordered activation loop region (residues 285-305), although there are peaks in the C-terminus (residues 313-324). There are almost none in the revision around the active site (simulation residues 180-184) as it is a cleft. There is good agreement between replicates and the histograms show that the overall binding interaction is independent of lipid composition. There is also good agreement across the ensemble as to the proportion of interactions that each residue accounts for and this increases confidence that the simulations are reproducible.

I wanted to compare the dimer contacts from the simulation to the X-ray crystallographic structure (PDB: 4TZ7) [25]. We found that higher order structures more accurately captured the crystallographic dimer interface contacts. Fig.6.4 shows the contact maps from structures of clusters of 5 or 8 monomers with side and top views of the structures coloured by contact frequency.



PDB 4TZ7



Crystallographic Dimer

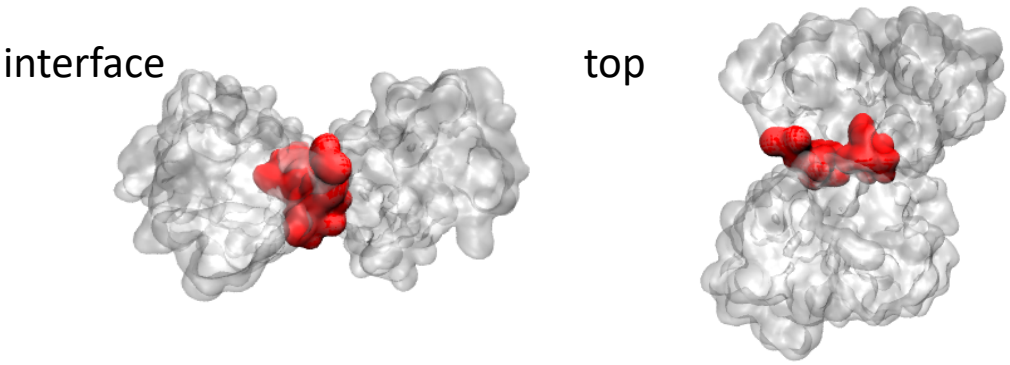


Fig. 6.4: Comparison between simulation and structural data of dimeric interfaces. The % of interactions accounted for each residue is shown by histograms for 5 and 8 monomers clustered together. Below this are shown structures of the crystallographic monomer with interface residues highlighted in red. The full crystallographic dimer is also shown with the interfacial residues of one monomer coloured red.

The crystallographic dimer interface residues are mainly in the region of simulation residues 25-35 and 131-138. These residues were not significantly present in contact histograms of clusters of only two monomers, and other residues were found to make more contacts. However, where 5 or 8 monomers are deemed to be clustering together, it appears that these regions make more of a contribution to the binding interface. This may be as a result of sampling through the simulation, where clusters of 5 or 8 monomers (i.e. 2 dimers and a monomer, or 4 dimers) are more likely to record the contribution of important residues, whereas the analysis may only capture transient dimers and therefore the contact profile indicates a transient state that does not quite lead to full binding. That said, there is good general agreement about the regions of contacts when coloured on the structure, and especially when compared to the crystallographic structure. It has often been suggested that the Martini forcefield is 'too sticky' [201] and that interactions are not always meaningful, but it is encouraging that the simulations indicate contacts broadly in the region where they would be expected. Another factor is that Martini, as a CG forcefield, cannot always capture the full geometry and charge distribution of the side chains, but the simulation results here indicate that the general interaction is demonstrated. Given the conclusions of results chapter 1, where it was postulated that the crystallographic dimer is unable to bind to planar membranes, the CG simulations may even suggest a slightly altered interface pattern that could be more suitable for binding, but this is merely speculative.

6.3.3 Cluster analysis shows time-progression of protein dimerisation and oligomerisation

In the previous section I identified that the PIP5K1A clusters and identified the important residues for binding. To establish how the PIP5K1A monomer dimerises over time, in house python scripts were used to count how many kinases were in dimers

or oligomers over the course of the simulation. Fig.6.5 shows that dimers and trimers begin to form quickly over the course of the simulation.

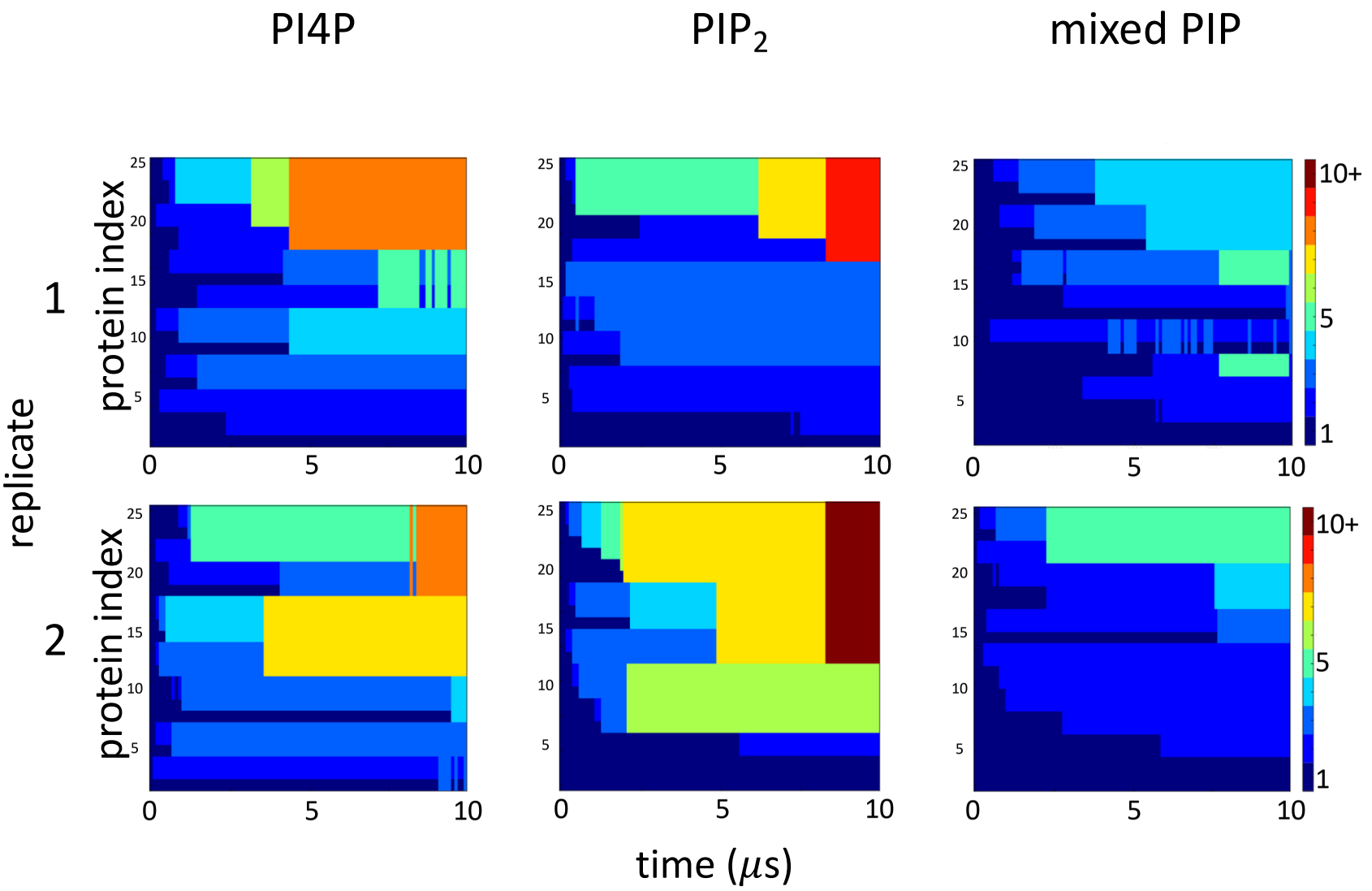


Fig. 6.5: Protein clustering over all replicates in 2D. Protein indicates the protein index. The colour indicates what size cluster the kinase is in. The colour bars indicate the size of the cluster from 0 to >10 monomers. Replicates and PIP species present are labelled (mixed=PI4P and PIP_2).

Clusters of 4 kinase monomers appear in under 1 μs. In the PI4P simulations, clusters of eight kinases are observed from 5 μs and 8 μs. In the PIP_2 simulations, a similar pattern is observed, and one replicates shows two clusters of seven kinases forming a large cluster of 14 monomers. Interestingly, in the mixed PIP simulations the kinase clustering appears to progress more slowly in both replicates despite otherwise identical conditions (fig.6.6). There is a higher charge density on the membrane as both PIPs are present at 5% resulting in an overall 10% lipid bilayer, so it may be that the kinases recognise and bind to the membrane more quickly, and diffuse more slowly on the membrane, resulting in fewer oligomers. Interestingly around 20-50% of monomers remain monomeric until well over 5 μs. Most importantly, it is important that the combined PIP membrane appears to have an effect on membrane binding, and that this is not unexpected in the context of PIP-binding cooperativity [68, 83], composition [271] or membrane curvature [272]. It is clear that more research in this area is important and these types of studies may help to relate the underlying biophysical properties to each other [273]. Previously spatial imaging analysis has mostly been applied to integral membrane proteins. It would be interesting to perform fluorescence microscopy experiments to compare with these simulations; particularly with regard to the size of clusters formed and whether or not they are transient. This study does show some transience of interactions in lower order (e.g. two to three monomers) oligomer interactions but it does indicate that large clusters of many monomers are more stable. Protein crowding experiments would be useful here to determine the effecting of concentration on interactions, as this has been demonstrated to be important in the dynamic organisation of ion channels in simulation [274].

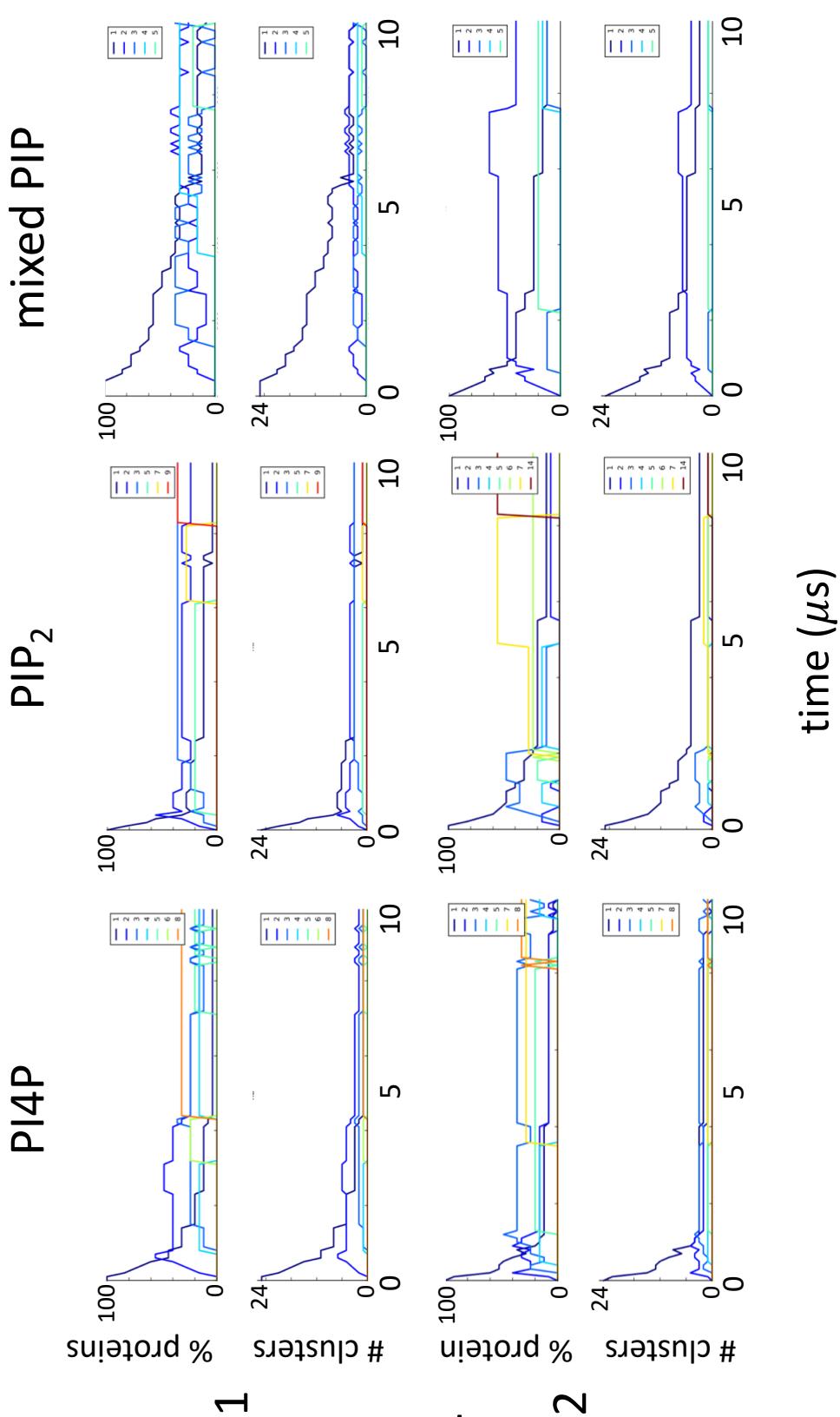


Fig. 6.6: Protein clustering over all replicates in 1D. The colour bars indicate the size of the cluster from 0 to >10 monomers. The 1D plots represent both the raw counts of monomers in the cluster and also the % of proteins in the cluster. PIP mixtures and replicates are labelled.

6.3.4 Cluster and diffusion analysis show lipids cluster in simulated signalling events

Having investigated protein clustering, I wanted to investigate if there was any concurreny PIP or other lipid clustering. First I performed density analysis to see if and how PIPs clustered in the membrane.

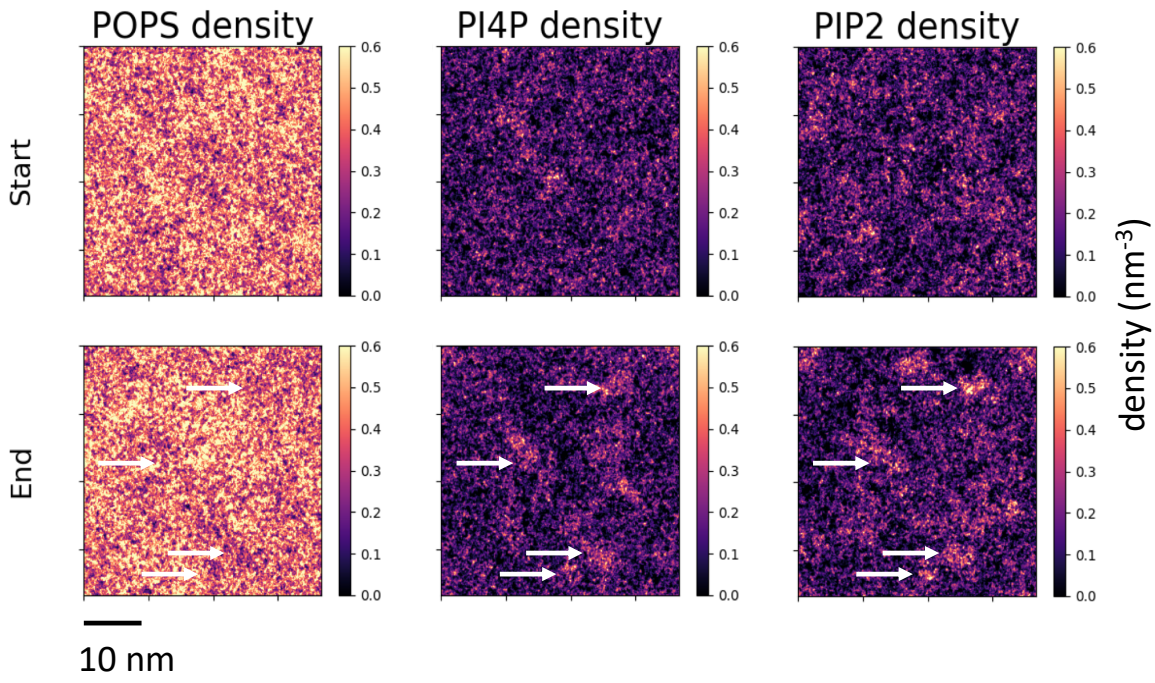


Fig. 6.7: 2D density analysis of PIPs in the membrane. Start and end density maps for POPS, PI4P, and PIP_2 for one of the mixed PIP simulations. Density is indicated by the colour bar. Yellow indicates regions of higher density. White arrows have been used to highlight areas of increased PIP density (middle and right) and the concurrent decreased density of POPS. The arrows show that both PIPs form clusters that are either adjacent or overlapping.

I used the densmap tool in GROMACS to calculate the density of PIPs over 0-1 μs (start) and 9-10 μs (end). We observed increased density of PI4P and PIP_2; these appear to be co-localised. We show the POPS density for reference and show that POPS density appears to be depleted where PIP density is increased (white arrows,

fig.6.7).

In order to see if this PIP clustering was correlated with PIP5K1A binding, I performed RDF analysis around the protein and calculated the lipid diffusion coefficients. I present the average RDF over the final μs (9-10) of simulation) in fig.6.8.

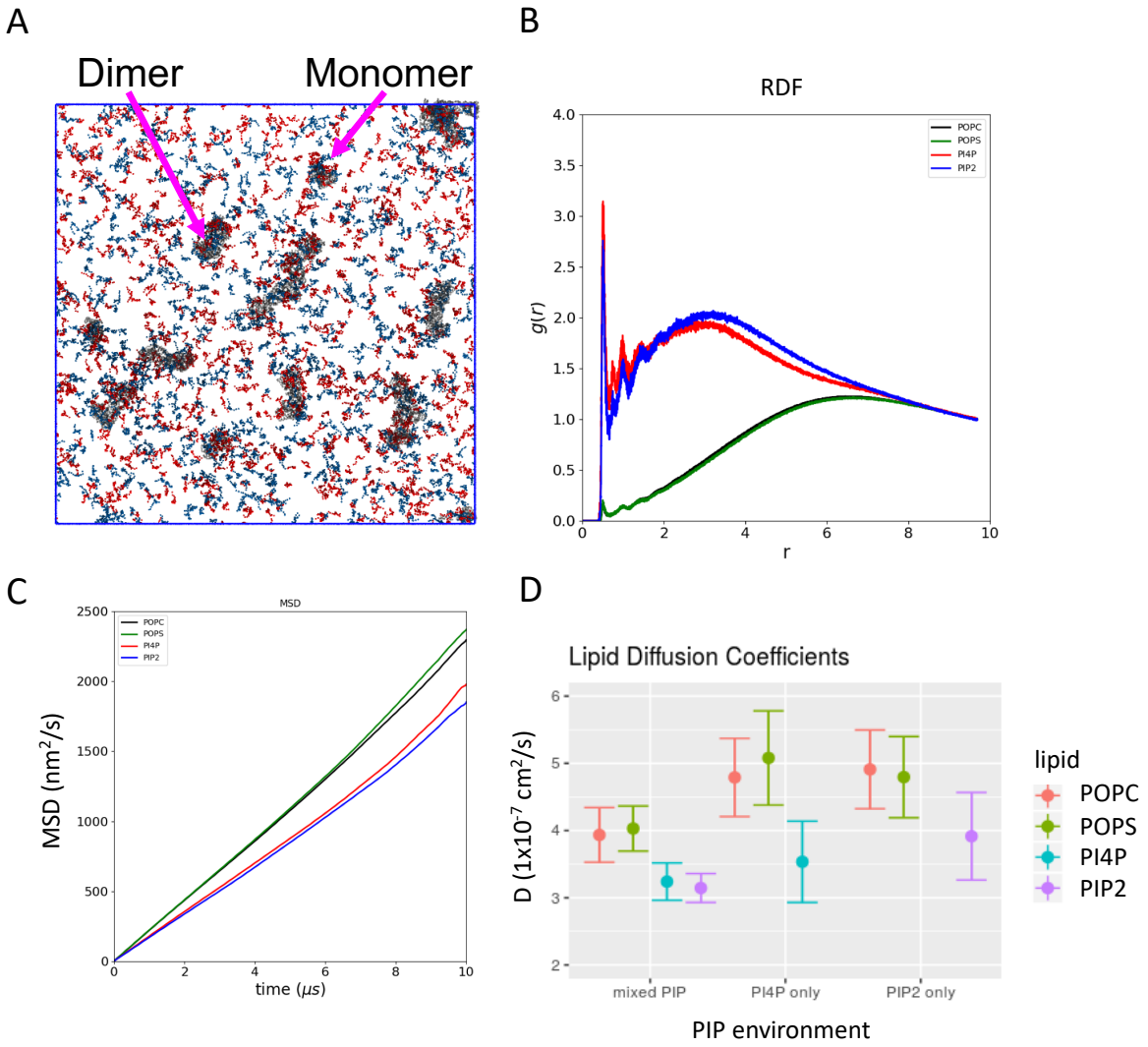


Fig. 6.8: Lipid diffusion clustering in mixed PIP simulations. A: Snapshot of the simulation box showing PIP_2 in blue, PI4P in red, and protein in grey. Bulk lipids and solvent are omitted for clarity. Example monomers and dimers are indicated by the arrows. B: Radial distribution function for lipids in one mixed PIP simulation. C: Mean squared displacement for lipids over time. D: Average lipid diffusion coefficients for all replicates.

The RDF shows an emphatic preference for PIPs over bulk lipid underneath ($r = 0$)

the protein. Interestingly it also shows an additional 3 peaks from 0.1-2 nm, indicating 3 annular shells underneath the protein. There appears to be a slight preference for PI4P closer to the protein (<2 nm) and beyond this a slight preference for PIP_2. It is interesting that the CG model is able to capture this detail; we expect the PIP5K1A to bind PI4P preferentially as its substrate lipid, and to show less preference for its product, PIP_2. Lipid diffusion coefficients also show reduced diffusion of the PIPs in the different lipid environments. Interestingly the standard deviation of lipid diffusion coefficients is much smaller in the mixed lipid simulations. The diffusion of PIP_2 only seems to be reduced significantly when mixed with PI4P. We also notice that the diffusion of the bulk lipid is reduced more with mixed PIP than with either PI4P or PIP_2 alone.

This interplay between the lipids is interesting because it has been proposed that PI4P and PIP_2 are independent lipid determinants of membrane identity [275], where it was suggested that PI4P makes an additional contribution to plasma membrane requirements for polyvalent anionic lipids, and not just as a precursor to PIP_2. This leaves PIP_2 to undergo rapid turnover without disturbing the defining electrostatic properties of the plasma membrane. Whilst the differences in diffusion shown here are slight, they perhaps indicate that this independence is not quite so clear cut, and gives support to the hypothesis that clustering of PIPs is a diffusion-driven process [276]; for example, BAR domains have been shown to halt PIP diffusion in membrane-binding events [277]. It also supports experimental findings that membrane-bound peptides sequester PIPs but not monovalent (POPS) lipids [278]. Moreover, PH domains have been shown to exhibit temporally fluctuating diffusivity resulting from changing interactions between the protein and PIP molecules [279]. Taken with the results presented here, it suggests that this complexity has an important role to play in the overall signalling landscape.

6.3.5 HDBSCAN shows lipid clusters forming over time

To investigate how lipids cluster in the membrane over time, we used a python implementation of the HDBSCAN algorithm to identify lipid clusters at each time step and investigate their size in the mixed PIP simulation.

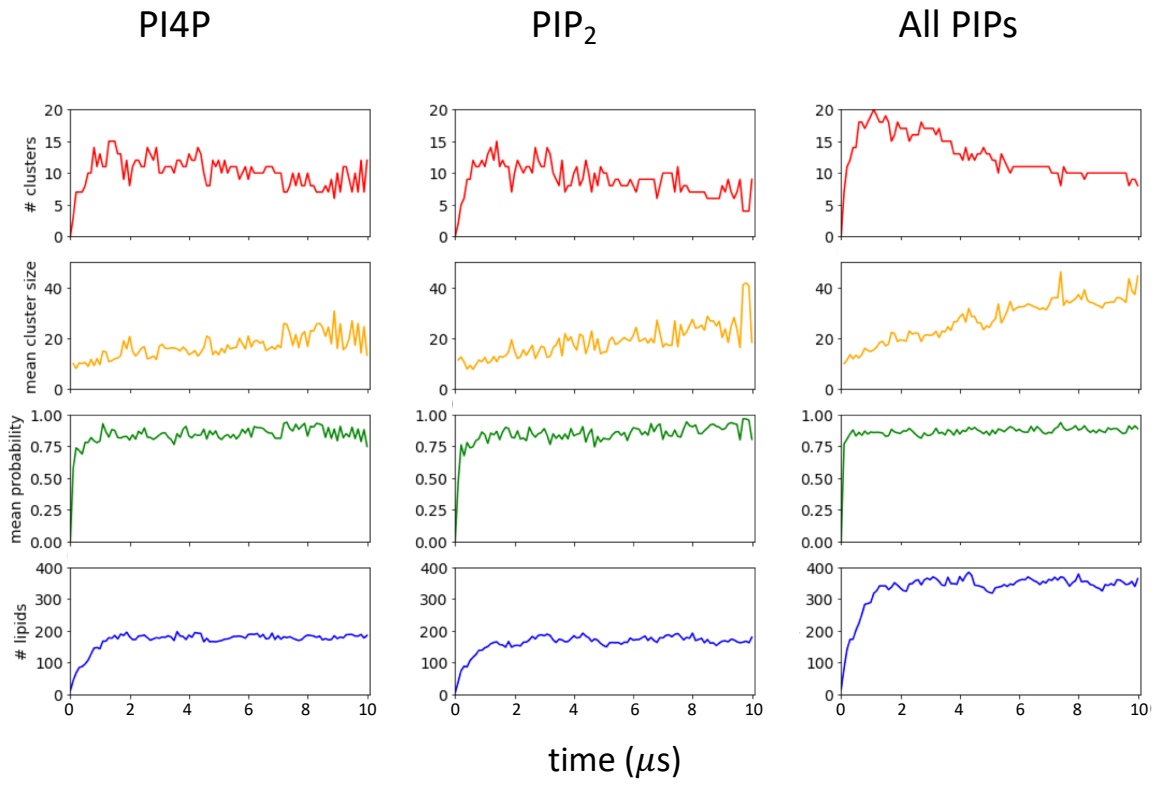


Fig. 6.9: Lipid clustering over the course of one mixed PIP simulation. The red plot shows the number of clusters identified in the simulation over time. This number increases quickly through the first 2 μs and then slowly decreases as the mean cluster size increases and clusters merge. The yellow plot shows the mean cluster size for both single PIPs gradually increases from 10 to >20. The green plot shows the mean probability of assignment; this indicates how confidently the DBSCAN algorithm has assigned a lipid to its designated cluster. DBSCAN converges at around 80-90% so we can assume that the clustering is reasonable. The blue plot shows the total number of lipids involved in clusters under kinases.

The number of clusters increases over the first 2 μs and appears to remain constant at this point. The mean number of lipids in each cluster appears to be the same for both PIPs, and converges around 20 lipids. Towards the end of the simulation, smaller

clusters appear to coalesce with the mean cluster size increasing and the number of clusters decreasing. By performing this analysis on all PIPs, we show that double the number of lipids are found in clusters, but that the number of clusters is similar to the single PIP simulations. This indicates that the clusters are mixed and not clusters of single PIPs. This corroborates the RDF data which finds both PI4P and PIP_2 in the annular shells, albeit with PIP_2 at a lower probability. The probability of assignment increases over the course of the simulation, which indicates the algorithm has assigned lipids with greater confidence. Finally, the total number of lipids involved in clustering fluctuates between 150 and 200 for both species. Overall it seems there is no great difference between the species in their clustering behaviour.

Unsupervised learning as yet has not been greatly used for studying lipid clustering; there is one instance of nonnegative matrix factorisation applied to phase separation at time of writing [280]. This section demonstrates that the HDBSCAN algorithm can be used to quickly and easily derive metrics from simulations of lipid clustering that may further inform our understanding of signalling events.

6.3.6 Simulations of dimeric PIP5K1A with a complex membrane indicate a binding preference for curved regions

Having shown in the previous section the formation of dimers and higher order structures from large-scale simulations of multiple copies of monomeric PIP5K1A, I wanted to investigate the behaviour of dimeric PIP5K1A. The idea for these simulations arose out of the findings of chapter 3, where it is postulated that membrane curvature may be an important aspect of PMP binding. We also wanted to establish the binding behaviour of the PIP5K1A when the lipid composition is closer to that of physiological cell membranes.

A table of simulations is given below:

composition (EC)	composition (IC)	dimers	lipid ratios (EC)	lipid ratios (IC)	time (μs)	n
POPC:POPE:DPSM:DPG3:CHOL	POPC:POPE:POPS:PIP_2:PI4P:CHOL	9	40:10:15:10:25	10:40:15:5:5:25	10	1
POPC:POPE:DPSM:DPG3:CHOL	POPC:POPE:POPS:PIP_2:PI4P:CHOL	0	40:10:15:10:25	10:40:15:5:5:25	10	1

Table 6.2: Table of large-scale PIP5K1A dimer simulations in complex asymmetric composition bilayers.

Simulations of the PIP5K1A dimer approaching a large complex membrane were performed. I chose a complex asymmetric composition that has been previously studied with simulation [197] to mimic the plasma membrane, but modified the 10% PIP_2 concentration to include PI4P yielding a 5% composition for each. The dimer simulation box was 71 x 71 x 36 nm and contained over 1.6 million particles. A control simulation of just the bilayer was also performed, with dimensions 71 x 71 x 20 nm. Both bilayers contained approximately 33,000 lipids and were run for 10 μs each.

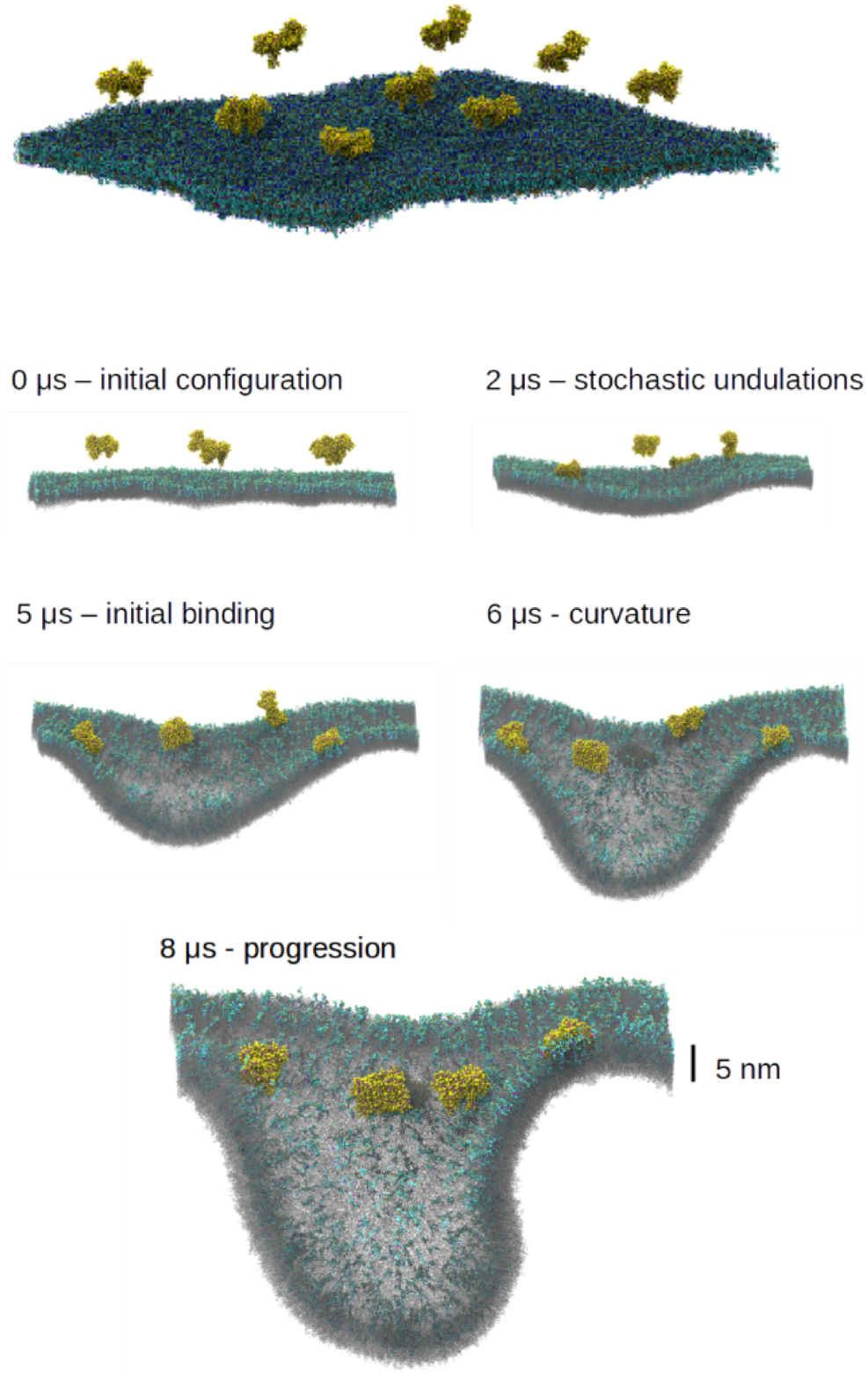


Fig. 6.10: Snapshots showing progression of a simulation of 9 PIP5K1A dimers with a complex asymmetric lipid membrane. The 9 dimers are shown in yellow, initially positioned distal to the membrane on the 'intracellular' composition side. The membrane is shown in grey with PIPs on the inner leaflet highlighted in cyan.

Fig.6.10. shows the time-progression of this simulation. Unfortunately, the dimers did not make contact with each other throughout the simulation; this is possibly because they diffuse through solution more slowly than the monomer. However this means that we were unable to quantify protein clustering as for the monomer simulations. This is possibly because the dimers are slightly more spaced, and further simulations could be modified in their setup to reflect this.

However, what is striking about this simulation is the level of curvature observed in the lipid membrane, and the concurrent localisation of protein to the regions of highest curvature. Fig.6.8. shows stochastic undulations for the first 2-4 μs, accompanied by some binding events of the dimer. At 5 μs, we observe a reasonable degree of membrane curvature and that most of the dimers are bound. Beyond this point, we see the membrane strongly increase the degree of curvature. At 8 μs, we observe dimers bound to the regions of highest curvature (both concave and convex) and we see PIPs clustered in these regions. It has previously been shown in simulations that lipid composition is coupled to local curvature [197]. At this point, the membrane appears to be forming a 'bud' and has crossed the periodic boundaries; here the membrane is so curved that it has begun to interact with dimers bound to the other side of itself, and so we only consider the first 8 μs here. We also ran a control simulation with no kinase present, and observed a similar effect (data not shown). We observe clustering of PIPs underneath the dimers. Five dimers have bound to the 'correct' intracellular membrane side. Four dimers have crossed the periodic boundaries to bind, incorrectly, to the extracellular side of the membrane. We have discounted these in our analysis.

Although it is unwise to draw any conclusions from this simulation as yet, we report it as it seems to demonstrate that there is some interaction between lipid composition, curvature, and PMP binding. We suggested in our early studies of the PIP5K1A dimer that whilst it cannot bind to planar membranes, it may well require binding to be at

areas of membrane curvature. We postulate that it would be interesting to simulate the system with the same membrane in a larger box size and possibly introduce repulsive potentials to ensure the dimers do not cross the periodic boundary to bind to the extracellular leaflet. The need for more complex representations of cell signalling events is required for a fuller understanding of cell signalling and this chapter provides proof of concept for this approach.

6.4 Conclusions and Future Directions

This chapter comprises some more exploratory studies that were carried out to investigate the behaviour of the PIP5K1A kinase in a model signalling event. We showed that lipids cluster in these signalling events, and that these clusters are positioned directly under the protein. We also showed that the kinases cluster in the simulation events and form dimers in ways that are analogous to the crystal structures. We showed that the progression of this clustering over time and that in mixed lipid bilayers with higher charge density this clustering appears to be slower, perhaps due to monomers binding to the membrane more quickly and experiencing more limited diffusion. We would suggest performing these simulations with varying concentrations of PIPs in order to establish if there is an effect on the speed of binding. We also demonstrated the use of the HDBSCAN algorithm for investigating lipid clustering in membranes and that it can be used to establish useful information about cluster size and probability of assignment. We would suggest that there is much scope for this technique, particularly in simulations with more complex bilayers, and that it would be interesting to quantify the relationship between lipid clustering and PMP binding. Finally, we ran two simulations of a complex membrane with and without the dimeric kinase structure. In these simulations we observed a high degree of curvature, and that the kinases appear to bind to regions of high curvature, as postulated in chapter

3. However these results are quite speculative and the development of new tools is necessary to analyse these bilayers. We suggest that these simulations are analysed with methods that account for curvature, for example using Helfrich-Canham theory to relate equilibrium fluctuations of a membrane to elastic properties [200].

Overall this chapter shows that there is great scope for simulating PMP signalling events, and that membrane curvature may play an important part in these phenomena.

Chapter 7

Conclusions & Future Directions

"It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment."

Carl Friedrich Gauss

This thesis has attempted to answer four questions related to the study of PMPs:

- How do multiscale MD simulations give insight to a heretofore neglected PMP system, PIP5K1A?
- Can MSMs be used to investigate the disordered regions of the PIP5K1A system?
- Can the methods developed to investigate the PIP5K1A system be applied to another system of particular biomedical interest?
- How do large-scale simulations inform cell signalling events?

I have attempted to answer these questions through computational methods, including MD simulations and MSMs. Here I review the main findings and speculate as to possible future directions of study.

7.1 Membrane recognition by peripheral membrane proteins

In chapter 3, I presented a multiscale simulation study of the PIP5K1A kinase to investigate the characteristics of membrane recognition and binding.

We found that PIP5K1A binds to PIP-containing membranes (and not membranes without PIP) and that this interaction is led by a disordered region, the activation loop. This was an interesting finding because previous studies have not investigated the loop and became the basis of work presented in chapter 4. This initial interaction is followed by reorientation and binding. We showed that the CG simulations recapitulate experimental crystallography data by showing that PI4P binds in the active site and makes contact with the catalytic lysine residue. It is flanked in this position by K238 and R244.

This study also explored the mechanism of dimer binding. PIP5K1A was crystallised as a dimer, and dimerisation was proposed to increase the rate of catalysis by either binding to the membrane at both active sites simultaneously, or by some kind of allosteric effect which stabilises the binding conformation. The study indicated that the latter is more likely, due to the inability of CG dimers based on the crystal structure being unable to bind to the membrane with both active sites. The study indicates, through analysis of the particle fluctuations, that increased tension in the C-helix may contribute to an allosteric effect. However, this study was conducted in planar membranes only and I propose that membrane curvature may have an effect on binding. To this end, future studies could be carried out in curved membranes and in more complex membrane compositions (indeed as I have included in chapter 6, although these results are currently speculative). This may have implications for the general mechanism of PIP binding, as curvature is not yet a large consideration in the

model.

7.2 Markov state models of the PIP5K1A activation loop

I studied the PIP5K1A activation loop in further detail using atomistic simulations and MSMs after the findings presented in chapter 3. I carried out this study because disordered regions in PMPs are not well understood, but their functional activity could have large implications for pharmaceutical design.

I investigated if it was possible to cluster loop conformations based on two collective variables, in this case RMSD and residue distance. I found that this was possible, which is interesting because it perhaps points to a general property of conformational switching based around large aromatic residues in disordered regions. I also analysed lipid contacts to identify the important lipid contacts in the activation loop. Further to this, I performed MSM analysis based on the ϕ, ψ, and χ_1 angles of the activation loop and found that the loop occupies two major states during the initial binding event. 40% of trajectories transitioned to a second state, distinct from the state occupied at the start of the simulation. I was able to build robust MSMs based on 7.5 μs of aggregate simulation and generate representative structures for the loop. I also investigated the pathways the loop follows in the binding event.

It would be useful to refine these models further and perhaps incorporate lipid contact formation into the model. This may help answer the question of whether or not the activation loop performs a shuttling function once in the membrane to manoeuvere PIPs in and out of the catalytic site. It would be useful especially to perform these simulations in bilayers of mixed PI4P and PIP_2 composition on long timescales to address this.

More generally, it would be useful to apply this protocol to other activation loops in the PIP kinase family to identify if there are differences between them from an MSM perspective. The protocol can be used to investigate other PMPs containing loops and yield structural and functional insight.

7.3 Insights into the structure and function of α-synuclein

I sought to apply the methods developed in chapters 3 and 4 to a new system of interest. I conducted a multiscale simulation study with MSM analysis of the α-synuclein monomer in contact with lipid membranes. I sought to investigate if the methods developed could be used to yield greater insight into this system, as the membrane recognition and binding by α-synuclein appears to be important for its function but still poorly understood.

I conducted CG simulations of α-synuclein in contact with POPG membranes at varying degrees of helicity. I found that the inter-helical region leads the binding interaction, and characterised the important residues for binding. I found that helix-1 makes the greatest number of contacts with the membrane, with helix-2 lying over the top of helix-1. This is in line with other previous experimental studies [237]. I also found that decreasing the helicity of α-synuclein increases the binding contribution from the NAC helix.

Through simulations of α-synuclein with mixed DOP lipid membranes, it appears that there are not specific binding hotspots for particular DOP species, but that α-synuclein interacts preferentially with DOPE in the initial stages of interaction. However, I showed that upon full binding the lipid distribution appears to be close to the bulk distribution. It would be useful and interesting to simulate α-synuclein

with more complex membranes that more closely mimic the composition of the brain. It is anticipated that advances in lipidomics may assist in designing physiologically relevant membranes.

I performed atomistic simulations of the initial binding event of α-synuclein from three initial conformations. I showed that the contact profile progression varies widely between replicates; this is interesting because it hints at the great conformational variability of the protein and therefore why it has been so difficult to study. I used MSMs to study the NAC region of α-synuclein and found that this region undergoes a break in the helical structure. This is particularly interesting because this is in line with NMR data presented in this chapter which was obtained by collaborators. It is also in line with a study that shows a small molecule binding to this region is able to stop progression of Parkinson's in a disease model [248]. We were able to produce representative structures that indicate this part of the helix forms a turn and exposes a number of hydrophobic residues. Given that this helix-2 lies above helix-1 and is exposed, it may seed aggregation events when other monomers make contact with the structure.

Two main avenues for further exploration arise out of this: firstly, it would be useful to perform MSM analysis on simulations of α-synuclein with known mutations, and perhaps also with small molecule ligands of interest. Secondly, further study with many α-synuclein monomers would be able to address how these proteins aggregate and cause disease. Furthermore, taken together with the conclusions from chapters 3 and 4, I speculate that further examination of other disordered PMP systems may yield parallels in their behaviour which inform our understanding in the wider context of disorder and its contribution to membrane recognition.

7.4 Simulation of cell signalling events

Finally, I performed large-scale CG simulations of PIP5K1A monomers and dimers with large membranes of varying composition. I performed this study to investigate how large-scale simulations could give insight into the more general signalling event in which the PIP5K1A plays an important role.

I found that these simulations show dimerisation of the PIP5K1A monomers in line with the X-ray crystallographic structure and identify important residue-residue contacts.

I also investigated the lipid clustering and found that PIPs cluster underneath the PIP kinases. There appears to be some interesting interplay between the product and substrate lipid in mixed lipid composition simulations which certainly warrants further investigation. I also applied an unsupervised learning technique to investigate lipid clustering over time, which revealed average cluster sizes and the time progression of these events.

I investigated the dimeric structure of PIP5K1A in contact with a large membrane of a complex composition reminiscent of the cell membrane. Although these results are preliminary, I identified a large degree of curvature which I alluded to in chapter 3; interestingly I found the dimer does indeed bind to regions of high curvature, but that these regions are convex rather than concave (which would more likely facilitate dual binding of the PIP5K1A dimer active sites). Nonetheless it offers a natural area for further investigation; the simulation of large complex membranes with multiple PMPs, possibly of different species. PIP5K1A is also known to interact with intracellular components of the Wnt signalling pathway [25] and it may be interesting to perform studies to investigate this. Addressing curvature in membrane simulations, particularly of asymmetric membranes, is challenging and new methods would need to be developed to overcome this.

7.5 The frontiers of biomolecular simulation

In conclusion, this thesis set out to answer four related questions with regard to the membrane recognition and binding of PMPs. Through the use of MD simulations and MSMs I have uncovered new insights into the mechanisms of PIP5K1A and α-synuclein which may be important when studying disease processes or designing new clinical therapies for the diseases which they cause. A natural progression from this point would be a large-scale serial analysis of PMPs in their native lipid environments, and the application of MSM analysis to their disordered regions to facilitate comparison between them. This not only would inform our perspective of PMP binding mechanisms, but also the more general protein disorder field.

A common theme of this thesis is the need for multiscale simulations which are capable of handling long timescales and large length-scales, and the appropriate use of statistical methods to generate useful insight which robust yet is comprehensible to humans. One cannot ignore the pace of current computational advances; at the time of writing, machine learning approaches are very fashionable even if the insights they currently yield are somewhat limited (at least in the field of biophysics). However it seems that as the field progresses in this regard we may see great advances in the processing of relevant data that will surely have great influence in the simulation community. Related to this are more general advances in hardware [281] and simulation which will enable the analysis of ever larger systems. To help us understand the less well understood corners of biophysical processes, it may one day be possible for us to view the whole cell through the (computational) microscope, as Hooke did in 1665.

References

[1] J. Lombard, "Once upon a time the cell membranes: 175 years of cell boundary research.," *Biology Direct*, vol. 9, p. 32, 2014.

[2] G. Quincke, "Ueber periodische Ausbreitung an Flüssigkeitsoberflächen und dadurch hervorgerufene Bewegungserscheinungen," *Annalen der Physik und Chemie*, vol. 271, no. 12, pp. 580–642, 1888.

[3] S. J. Singer and G. L. Nicolson, "The Fluid Mosaic Model of the Structure of Cell Membranes," *Science*, vol. 175, no. 4023, pp. 720–731, 1972.

[4] K. Jacobson, E. D. Sheets, and R. Simson, "Revisiting the fluid mosaic model of membranes.," *Science*, vol. 268, no. 5216, pp. 1441–2, 1995.

[5] D. Lingwood and K. Simons, "Lipid Rafts As a Membrane-Organizing Principle," *Science*, vol. 327, no. 5961, pp. 46–50, 2010.

[6] Ü. Coskun and K. Simons, "Cell Membranes: The Lipid Perspective," *Structure*, vol. 19, no. 11, pp. 1543–1548, 2011.

[7] M. A. Lemmon, "Membrane recognition by phospholipid-binding domains.," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 2, pp. 99–111, 2008.

[8] N. Schwierz, S. Krysiak, T. Hugel, and M. Zacharias, "Mechanism of Reversible Peptide–Bilayer Attachment: Combined Simulation and Experimental Single-Molecule Study," *Langmuir*, vol. 32, no. 3, pp. 810–821, 2016.

[9] E. Fuglebakk and N. Reuter, "A model for hydrophobic protrusions on peripheral membrane proteins," *PLOS Computational Biology*, vol. 14, no. 7, p. e1006325, 2018.

[10] D. Casares, P. V. Escribá, C. A. Rosselló, D. Casares, P. V. Escribá, and C. A. Rosselló, "Membrane Lipid Composition: Effect on Membrane and Organelle Structure, Function and Compartmentalization and Therapeutic Avenues," *International Journal of Molecular Sciences*, vol. 20, no. 9, p. 2167, 2019.

[11] J. E. Vance and R. Steenbergen, "Metabolism and functions of phosphatidylserine," *Progress in Lipid Research*, vol. 44, no. 4, pp. 207–234, 2005.

[12] M. R. Wenk, "The emerging field of lipidomics," *Nature Reviews Drug Discovery*, vol. 4, no. 7, pp. 594–610, 2005.

[13] S. McLaughlin, J. Wang, A. Gambhir, and D. Murray, "PIP2 and Proteins: Interactions, Organization, and Information Flow," *Annual Review of Biophysics and Biomolecular Structure*, vol. 31, no. 1, pp. 151–175, 2002.

[14] S. McLaughlin and D. Murray, "Plasma membrane phosphoinositide organization by protein electrostatics," *Nature*, vol. 438, no. 7068, pp. 605–611, 2005.

[15] J. Kunz, M. P. Wilson, M. Kisseleva, J. H. Hurley, P. W. Majerus, and R. A. Anderson, "The Activation Loop of Phosphatidylinositol Phosphate Kinases Determines Signaling Specificity," *Molecular Cell*, vol. 5, no. 1, pp. 1–11, 2000.

[16] A. Mulgrew-Nesbitt, K. Diraviyam, J. Wang, S. Singh, P. Murray, Z. Li, L. Rogers, N. Mirkovic, and D. Murray, "The role of electrostatics in protein–membrane interactions," *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, vol. 1761, no. 8, pp. 812–826, 2006.

[17] B. J. Mayer, R. Ren, K. L. Clark, and D. Baltimore, "A putative modular domain present in diverse signaling proteins," *Cell*, vol. 73, no. 4, pp. 629–630, 1993.

[18] J. E. Harlan, P. J. Hajduk, H. S. Yoon, and S. W. Fesik, "Pleckstrin homology domains bind to phosphatidylinositol-4,5-bisphosphate," *Nature*, vol. 371, no. 6493, pp. 168–170, 1994.

[19] R. J. Haslam, H. B. Koide, and B. A. Hemmings, "Pleckstrin domain homology," *Nature*, vol. 363, no. 6427, pp. 309–310, 1993.

[20] A. Musacchio, T. Gibson, P. Rice, J. Thompson, and M. Saraste, "The PH domain: a common piece in the structural patchwork of signalling proteins," *Trends in Biochemical Sciences*, vol. 18, no. 9, pp. 343–348, 1993.

[21] J. P. DiNitto and D. G. Lambright, "Membrane and juxtamembrane targeting by PH and PTB domains," *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, vol. 1761, no. 8, pp. 850–867, 2006.

[22] N. Gamper and M. S. Shapiro, "Target-specific PIP(2) signalling: how might it work?," *The Journal of Physiology*, vol. 582, no. Pt 3, pp. 967–75, 2007.

[23] M. Heath, C.M., Stahl, P.D., Barbieri, "Lipid kinases play crucial and multiple roles in membrane trafficking and signaling," *Histology and Histopathology*, vol. 18, no. 3, pp. 989–998, 2003.

[24] P. A. O. Weernink, K. Meletiadis, S. Hommeltenberg, M. Hinz, H. Ishihara, M. Schmidt, and K. H. Jakobs, "Activation of type I phosphatidylinositol 4-phosphate 5-kinase isoforms by the Rho GTPases, RhoA, Rac1, and Cdc42.," *The Journal of Biological Chemistry*, vol. 279, no. 9, pp. 7840–9, 2004.

[25] J. Hu, Q. Yuan, X. Kang, Y. Qin, L. Li, Y. Ha, and D. Wu, "Resolution of structure of PIP5K1A reveals molecular mechanism for its regulation by dimerization and dishevelled," *Nat Commun*, vol. 6, 2015.

[26] Y. Muftuoglu, Y. Xue, X. Gao, D. Wu, and Y. Ha, "Mechanism of substrate specificity of phosphatidylinositol phosphate kinases.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 31, pp. 8711–6, 2016.

[27] J. M. Drake and J. Huang, "PIP5K1α inhibition as a therapeutic strategy for prostate cancer.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 35, pp. 12578–9, 2014.

[28] L. Stefanis, "α-Synuclein in Parkinson's disease.," *Cold Spring Harbor Perspectives in Medicine*, vol. 2, no. 2, p. a009399, 2012.

[29] A. Recasens, B. Dehay, J. Bové, I. Carballo-Carbajal, S. Dovero, A. Pérez-Villalba, P.-O. Fernagut, J. Blesa, A. Parent, C. Perier, I. Fariñas, J. A. Obeso, E. Bezard, and M. Vila, "Lewy body extracts from Parkinson disease brains trigger α-synuclein pathology and neurodegeneration in mice and monkeys," *Annals of Neurology*, vol. 75, no. 3, pp. 351–362, 2014.

[30] H. Braak, K. D. Tredici, U. Rüb, R. A. de Vos, E. N. Jansen Steur, and E. Braak, "Staging of brain pathology related to sporadic Parkinson's disease," *Neurobiology of Aging*, vol. 24, no. 2, pp. 197–211, 2003.

[31] K. Del Tredici and H. Braak, "Review: Sporadic Parkinson's disease: development and distribution of α -synuclein pathology," *Neuropathology and Applied Neurobiology*, vol. 42, no. 1, pp. 33–50, 2016.

[32] D. F. Clayton and J. M. George, "The synucleins: a family of proteins involved in synaptic function, plasticity, neurodegeneration and disease," *Trends in Neurosciences*, vol. 21, no. 6, pp. 249–254, 1998.

[33] K. Beyer, "α-Synuclein structure, postttranslational modification and alternative splicing as aggregation enhancers," *Acta Neuropathologica*, vol. 112, no. 3, pp. 237–251, 2006.

[34] K. Uéda, H. Fukushima, E. Masliah, Y. Xia, A. Iwai, M. Yoshimoto, D. A. Otero, J. Kondo, Y. Ihara, and T. Saitoh, "Molecular cloning of cDNA encoding an unrecognized component of amyloid in Alzheimer disease.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 90, no. 23, pp. 11282–6, 1993.

[35] P. H. Weinreb, W. Zhen, A. W. Poon, K. A. Conway, and P. T. Lansbury, "NACP, A Protein Implicated in Alzheimer's Disease and Learning, Is Natively Unfolded," *Biochemistry*, vol. 35, no. 43, pp. 13709–13715, 1996.

[36] Li, Romero, Rani, Dunker, and Obradovic, "Predicting Protein Disorder for N-, C-, and Internal Regions.," *Genome Informatics*, vol. 10, pp. 30–40, 1999.

[37] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, "Sequence complexity of disordered protein," *Proteins: Structure, Function, and Genetics*, vol. 42, no. 1, pp. 38–48, 2001.

[38] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, and A. K. Dunker, "Exploiting heterogeneous sequence properties improves prediction of protein disorder," *Proteins: Structure, Function, and Bioinformatics*, vol. 61, no. S7, pp. 176–182, 2005.

[39] Z. R. Yang, R. Thomson, P. McNeil, and R. M. Esnouf, "RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins," *Bioinformatics*, vol. 21, no. 16, pp. 3369–3376, 2005.

[40] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon, "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content," *Bioinformatics*, vol. 21, no. 16, pp. 3433–3434, 2005.

[41] D. Eliezer, E. Kutluay, R. Bussell, and G. Browne, "Conformational properties of α-synuclein in its free and lipid-associated states," *Journal of Molecular Biology*, vol. 307, no. 4, pp. 1061–1073, 2001.

[42] L. C. Serpell, J. Berriman, R. Jakes, M. Goedert, and R. A. Crowther, "Fiber diffraction of synthetic α-synuclein filaments shows amyloid-like cross-β conformation," *Proceedings of the National Academy of Sciences*, vol. 97, no. 9, pp. 4897–4902, 2000.

[43] Y. Li, C. Zhao, F. Luo, Z. Liu, X. Gui, Z. Luo, X. Zhang, D. Li, C. Liu, and X. Li, "Amyloid fibril structure of α-synuclein determined by cryo-electron microscopy," *Cell Research*, vol. 28, no. 9, pp. 897–903, 2018.

[44] B. Li, P. Ge, K. A. Murray, P. Sheth, M. Zhang, G. Nair, M. R. Sawaya, W. S. Shin, D. R. Boyer, S. Ye, D. S. Eisenberg, Z. H. Zhou, and L. Jiang, "Cryo-EM of full-length α-synuclein reveals fibril polymorphs with a common structural kernel," *Nature Communications*, vol. 9, no. 1, p. 3609, 2018.

[45] R. Guerrero-Ferreira, N. M. Taylor, D. Mona, P. Ringler, M. E. Lauer, R. Riek, M. Britschgi, and H. Stahlberg, "Cryo-EM structure of alpha-synuclein fibrils," *eLife*, vol. 7, 2018.

[46] P. Das, S. Matysiak, and J. Mittal, "Looking at the Disordered Proteins through the Computational Microscope.," *ACS central science*, vol. 4, no. 5, pp. 534–542, 2018.

[47] P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky, and A. K. Dunker, "Intrinsic Disorder and Functional Proteomics," *Biophysical Journal*, vol. 92, no. 5, pp. 1439–1456, 2007.

[48] C. J. Oldfield, Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky, and A. K. Dunker, "Comparing and Combining Predictors of Mostly Disordered Proteins," *Biochemistry*, vol. 44, no. 6, pp. 1989–2000, 2005.

[49] C. K. Fisher and C. M. Stultz, "Constructing ensembles for intrinsically disordered proteins," *Current Opinion in Structural Biology*, vol. 21, no. 3, pp. 426–431, 2011.

[50] H.-C. Lu, S. S. Chung, A. Fornili, and F. Fraternali, "Anatomy of protein disorder, flexibility and disease-related mutations," *Frontiers in Molecular Biosciences*, vol. 2, p. 47, 2015.

[51] E. Deryusheva, E. Nemashkalova, M. Galloux, C. Richard, J. Eléouët, D. Kovacs, K. Belle, P. Tompa, V. Uversky, and S. Permyakov, "Does Intrinsic Disorder in Proteins Favor Their Interaction with Lipids?," *Proteomics*, vol. 19, no. 6, p. 1800098, 2019.

[52] G. P. Gorbenko, "The role of lipid–protein interactions in amyloid-type protein fibril formation," *Chemistry and Physics of Lipids*, vol. 141, no. 1-2, pp. 72–82, 2006.

[53] R. J. Perrin, W. S. Woods, D. F. Clayton, and J. M. George, "Exposure to long chain polyunsaturated fatty acids triggers rapid multimerization of synucleins.," *The Journal of Biological Chemistry*, vol. 276, no. 45, pp. 41958–62, 2001.

[54] C. N. Chirita, M. Necula, and J. Kuret, "Anionic micelles and vesicles induce tau fibrillization in vitro.," *The Journal of Biological Chemistry*, vol. 278, no. 28, pp. 25644–50, 2003.

[55] A. Z. Fernandis and M. R. Wenk, "Membrane lipids as signaling molecules," *Current Opinion in Lipidology*, vol. 18, no. 2, pp. 121–128, 2007.

[56] H. Luecke, B. Schobert, H.-T. Richter, J.-P. Cartailler, and J. K. Lanyi, "Structure of bacteriorhodopsin at 1.55 Å resolution," *Journal of Molecular Biology*, vol. 291, no. 4, pp. 899–911, 1999.

[57] C. Koshy and C. Ziegler, "Structural insights into functional lipid–protein interactions in secondary transporters," *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1850, no. 3, pp. 476–487, 2015.

[58] S. B. Long, X. Tao, E. B. Campbell, and R. MacKinnon, "Atomic structure of a voltage-dependent K+ channel in a lipid membrane-like environment," *Nature*, vol. 450, no. 7168, pp. 376–382, 2007.

[59] Y. Gao, E. Cao, D. Julius, and Y. Cheng, "TRPV1 structures in nanodiscs reveal mechanisms of ligand and lipid action," *Nature*, vol. 534, no. 7607, pp. 347–351, 2016.

[60] S. M. Jackson, I. Manolaridis, J. Kowal, M. Zechner, N. M. I. Taylor, M. Bause, S. Bauer, R. Bartholomaeus, G. Bernhardt, B. Koenig, A. Buschauer, H. Stahlberg, K.-H. Altmann, and K. P. Locher, "Structural basis of small-molecule inhibition of human multidrug transporter ABCG2," *Nature Structural & Molecular Biology*, vol. 25, no. 4, pp. 333–340, 2018.

[61] V. Betaneli, E. P. Petrov, and P. Schwille, "The Role of Lipids in VDAC Oligomerization," *Biophysical Journal*, vol. 102, no. 3, pp. 523–531, 2012.

[62] D. Huster, "Solid-state NMR spectroscopy to study protein–lipid interactions," *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, vol. 1841, no. 8, pp. 1146–1160, 2014.

[63] J. Prashek, T. Truong, and X. Yao, "Crystal Structure of the Pleckstrin Homology Domain from the Ceramide Transfer Protein: Implications for Conformational Change upon Ligand Binding," *PLoS ONE*, vol. 8, no. 11, p. e79590, 2013.

[64] I. Ali, S. Eu, D. Koch, N. Bleimling, R. S. Goody, and M. P. Müller, "Structure of the tandem PX-PH domains of Bem3 from Saccharomyces cerevisiae.," *Acta Crystallographica. Section F, Structural biology communications*, vol. 74, no. Pt 5, pp. 315–321, 2018.

[65] S. Ruskamo, R. P. Yadav, S. Sharma, M. Lehtimäki, S. Laulumaa, S. Aggarwal, M. Simons, J. Bürck, A. S. Ulrich, A. H. Juffer, I. Kursula, P. Kursula, and IUCr, "Atomic resolution view into the structure–function relationships of the human myelin peripheral membrane protein P2," *Acta Crystallographica Section D Biological Crystallography*, vol. 70, no. 1, pp. 165–176, 2014.

[66] R. Brunecky, S. Lee, P. W. Rzepecki, M. Overduin, G. D. Prestwich, and T. G. Kutateladze, Andrei G. Kutateladze, "Investigation of the Binding Geometry of a Peripheral Membrane Protein," *Biochemistry*, vol. 44, no. 49, pp. 16064–71, 2005.

[67] A. Saxena, P. Morozov, D. Frank, R. Musalo, M. A. Lemmon, E. Y. Skolnik, and B. Tycko, "Phosphoinositide binding by the pleckstrin homology domains of Ipl and Tih1.," *The Journal of Biological Chemistry*, vol. 277, no. 51, pp. 49935–44, 2002.

[68] I. Vonkova, A.-E. Saliba, S. Deghou, K. Anand, S. Ceschia, T. Doerks, A. Galih, K. G. Kugler, K. Maeda, V. Rybin, V. van Noort, J. Ellenberg, P. Bork, and A.-C. Gavin, "Lipid Cooperativity as a General Membrane-Recruitment Principle for PH Domains," *Cell Reports*, vol. 12, no. 9, pp. 1519–1530, 2015.

[69] D. Kerr, G. T. Tietjen, Z. Gong, E. Tajkhorshid, E. J. Adams, and K. Y. C. Lee, "Sensitivity of peripheral membrane proteins to the membrane context: A case study of phosphatidylserine and the TIM proteins," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1860, no. 10, pp. 2126–2133, 2018.

[70] A.-E. Saliba, I. Vonkova, and A.-C. Gavin, "The systematic analysis of protein–lipid interactions comes of age," *Nature Reviews Molecular Cell Biology*, vol. 16, no. 12, pp. 753–761, 2015.

[71] Y. Zhou, P. Prakash, H. Liang, K.-J. Cho, A. A. Gorfe, and J. F. Hancock, "Lipid-Sorting Specificity Encoded in K-Ras Membrane Anchor Regulates Signal Output," *Cell*, vol. 168, no. 1-2, pp. 239–251.e16, 2017.

[72] S. McLaughlin, "The Electrostatic Properties of Membranes," *Annual Review of Biophysics and Biophysical Chemistry*, vol. 18, no. 1, pp. 113–136, 1989.

[73] J. B. Larsen, C. Kennard, S. L. Pedersen, K. J. Jensen, M. J. Uline, N. S. Hatzakis, and D. Stamou, "Membrane Curvature and Lipid Composition Synergize To Regulate N-Ras Anchor Recruitment," *Biophysical Journal*, vol. 113, no. 6, pp. 1269–1279, 2017.

[74] M. Chavent, E. Seiradake, E. Y. Jones, and M. S. P. Sansom, "Structures of the EphA2 Receptor at the Membrane: Role of Lipid Interactions," *Structure*, vol. 24, no. 2, pp. 337–347, 2016.

[75] M. Chavent, D. Karia, A. C. Kalli, J. Domański, A. L. Duncan, G. Hedger, P. J. Stansfeld, E. Seiradake, E. Y. Jones, and M. S. Sansom, "Interactions of the EphA2 Kinase Domain with PIPs in Membranes: Implications for Receptor Function," *Structure*, vol. 26, no. 7, pp. 1025–1034.e2, 2018.

[76] N. Bhardwaj, R. V. Stahelin, R. E. Langlois, W. Cho, and H. Lu, "Structural bioinformatics prediction of membrane-binding proteins.," *Journal of Molecular Biology*, vol. 359, no. 2, pp. 486–95, 2006.

[77] N. Bhardwaj, M. Gerstein, and H. Lu, "Genome-wide sequence-based prediction of peripheral proteins using a novel semi-supervised learning technique," *BMC Bioinformatics*, vol. 11, no. Suppl 1, p. S6, 2010.

[78] M. A. Lomize, A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg, "OPM: Orientations of Proteins in Membranes database," *Bioinformatics*, vol. 22, no. 5, pp. 623–625, 2006.

[79] M. A. Lomize, I. D. Pogozheva, H. Joo, H. I. Mosberg, and A. L. Lomize, "OPM database and PPM web server: resources for positioning of proteins in membranes.," *Nucleic Acids Research*, vol. 40, no. Database issue, pp. D370–6, 2012.

[80] M. Papanastasiou, G. Orfanoudaki, M. Koukaki, N. Kountourakis, M. F. Sardis, M. Aivaliotis, S. Karamanou, and A. Economou, "The Escherichia coli Peripheral Inner Membrane Proteome," *Molecular & Cellular Proteomics*, vol. 12, no. 3, pp. 599–610, 2013.

[81] K. Diraviyam, R. V. Stahelin, W. Cho, and D. Murray, "Computer modeling of the membrane interaction of FYVE domains.," *Journal of Molecular Biology*, vol. 328, no. 3, pp. 721–36, 2003.

[82] S. M. Singh and D. Murray, "Molecular modeling of the membrane targeting of phospholipase C pleckstrin homology domains," *Protein Science*, vol. 12, no. 9, pp. 1934–1953, 2003.

[83] F. B. Naughton, A. C. Kalli, and M. S. P. Sansom, "Association of Peripheral Membrane Proteins with Membranes: Free Energy of Binding of GRP1 PH Domain with Phosphatidylinositol Phosphate-Containing Model Bilayers," *The Journal of Physical Chemistry Letters*, vol. 7, no. 7, pp. 1219–1224, 2016.

[84] A. Buyan, A. C. Kalli, and M. S. P. Sansom, "Multiscale Simulations Suggest a Mechanism for the Association of the Dok7 PH Domain with PIP-Containing Membranes," *PLOS Computational Biology*, vol. 12, no. 7, p. e1005028, 2016.

[85] H. Cui, G. S. Ayton, and G. A. Voth, "Membrane Binding by the Endophilin N-BAR Domain," *Biophysical Journal*, vol. 97, no. 10, pp. 2746–2753, 2009.

[86] S. Jaud, D. J. Tobias, J. J. Falke, and S. H. White, "Self-Induced Docking Site of a Deeply Embedded Peripheral Membrane Protein," *Biophysical Journal*, vol. 92, no. 2, pp. 517–524, 2007.

[87] E. Psachoulia and M. S. P. Sansom, "PX- and FYVE-Mediated Interactions with Membranes: Simulation Studies," *Biochemistry*, vol. 48, no. 23, pp. 5090–5095, 2009.

[88] C. N. Lumb, J. He, Y. Xue, P. J. Stansfeld, R. V. Stahelin, T. G. Kutateladze, and M. S. Sansom, "Biophysical and Computational Studies of Membrane Penetration by the GRP1 Pleckstrin Homology Domain," *Structure*, vol. 19, no. 9, pp. 1338–1346, 2011.

[89] E. Psachoulia and M. S. P. Sansom, "Interactions of the Pleckstrin Homology Domain with Phosphatidylinositol Phosphate and Membranes: Characterization via Molecular Dynamics Simulations †," *Biochemistry*, vol. 47, no. 14, pp. 4211–4220, 2008.

[90] V. S. Pande, K. Beauchamp, and G. R. Bowman, "Everything you wanted to know about Markov State Models but were afraid to ask," *Methods*, vol. 52, no. 1, pp. 99–105, 2010.

[91] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, "A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo," *Journal of Computational Physics*, vol. 151, no. 1, pp. 146–168, 1999.

[92] William C. Swope, J. W. Pitera, and F. Suits, "Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory," *Journal of Physical Chemistry B*, 2004.

[93] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," *The Journal of Chemical Physics*, vol. 134, no. 17, p. 174105, 2011.

[94] William C. Swope, J. W. Pitera, F. Suits, Mike Pitman, Maria Eleftheriou, Blake G. Fitch, Robert S. Germain, Aleksandr Rayshubski, T. J. C. Ward, Y. Zhestkov, and R. Zhou, "Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 2. Example Applications to Alanine Dipeptide and a β-Hairpin Peptide†," *Journal of Physical Chemistry B*, 2004.

[95] N. Singhal and V. S. Pande, "Error analysis and efficient sampling in Markovian state models for molecular dynamics," *The Journal of Chemical Physics*, vol. 123, no. 20, p. 204909, 2005.

[96] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, "Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states," *The Journal of Chemical Physics*, vol. 126, no. 15, p. 155102, 2007.

[97] B. E. Husic and V. S. Pande, "Markov State Models: From an Art to a Science," *Journal of the American Chemical Society*, vol. 140, no. 7, pp. 2386–2396, 2018.

[98] G. R. Bowman, X. Huang, and V. S. Pande, "Using generalized ensemble simulations and Markov state models to identify conformational states.," *Methods*, vol. 49, no. 2, pp. 197–201, 2009.

[99] M. Senne, B. Trendelkamp-Schroer, A. S. Mey, C. Schütte, and F. Noé, "EMMA: A Software Package for Markov Model Building and Analysis," *Journal of Chemical Theory and Computation*, vol. 8, no. 7, pp. 2223–2238, 2012.

[100] C. X. Hernández, M. P. Harrigan, M. M. Sultan, and V. S. Pande, "MSMExplorer: Data Visualizations for Biomolecular Dynamics," *Journal of Open Source Software*, vol. 2, no. 12, p. 188, 2017.

[101] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, "Molecular Simulation of ab initio Protein Folding for a Millisecond Folder NTL9(139)," *Journal of the American Chemical Society*, vol. 132, no. 5, pp. 1526–1528, 2010.

[102] G. R. Bowman, V. A. Voelz, and V. S. Pande, "Atomistic Folding Simulations of the Five-Helix Bundle Protein λ 685," *Journal of the American Chemical Society*, vol. 133, no. 4, pp. 664–667, 2011.

[103] D. L. Ensign, P. M. Kasson, and V. S. Pande, "Heterogeneity Even at the Speed Limit of Folding: Large-scale Molecular Dynamics Study of a Fast-folding Variant of the Villin Headpiece," *Journal of Molecular Biology*, vol. 374, no. 3, pp. 806–816, 2007.

[104] K. A. Beauchamp, D. L. Ensign, R. Das, and V. S. Pande, "Quantitative comparison of villin headpiece subdomain simulations and triplet-triplet energy transfer experiments.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 31, pp. 12734–9, 2011.

[105] G. Morra, A. M. Razavi, K. Pandey, H. Weinstein, A. K. Menon, and G. Khelashvili, "Mechanisms of Lipid Scrambling by the G Protein-Coupled Receptor Opsin," *Structure*, vol. 26, no. 2, pp. 356–367.e3, 2018.

[106] D. Shukla, Y. Meng, B. Roux, and V. S. Pande, "Activation pathway of Src kinase reveals intermediate states as targets for drug design," *Nature Communications*, vol. 5, no. 1, p. 3397, 2014.

[107] S. A. Adcock and J. A. McCammon, "Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins," *Chemical Reviews*, vol. 106, no. 5, pp. 1589–1615, 2006.

[108] L. Verlet, "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules," *Physical Review*, vol. 159, no. 1, pp. 98–103, 1967.

[109] R. W. Hockney and J. W. Eastwood, *Computer simulation using particles*. A. Hilger, 1988.

[110] D. Frenkel and B. Smit, *Understanding molecular simulation : from algorithms to applications*. 2002.

[111] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A linear constraint solver for molecular simulations," *Journal of Computational Chemistry*, vol. 18, no. 12, pp. 1463–1472, 1997.

[112] M. G. Martin and J. I. Siepmann, "Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes," 1998.

[113] N. L. Allinger, Y. H. Yuh, and J. H. Lii, "Molecular mechanics. The MM3 force field for hydrocarbons. 1," *Journal of the American Chemical Society*, vol. 111, no. 23, pp. 8551–8566, 1989.

[114] D. Dubbeldam, S. Calero, T. J. H. Vlugt, R. Krishna, T. L. M. Maesen, and B. Smit, *The journal of physical chemistry. B.*, vol. 108. 2004.

[115] H. Berendsen, D. van der Spoel, and R. van Drunen, "GROMACS: A message-passing parallel molecular dynamics implementation," *Computer Physics Communications*, vol. 91, no. 1-3, pp. 43–56, 1995.

[116] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The Amber biomolecular simulation programs," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1668–1688, 2005.

[117] J. W. Ponder and D. A. Case, "Force fields for protein simulations.," *Advances in Protein Chemistry*, vol. 66, pp. 27–85, 2003.

[118] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, D. York, and M. Karplus, "CHARMM: the biomolecular simulation program.," *Journal of computational chemistry*, vol. 30, no. 10, pp. 1545–614, 2009.

[119] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, no. 2, pp. 187–217, 1983.

[120] B. R. Gelin and M. Karplus, "Sidechain torsional potentials and motion of amino acids in porteins: bovine pancreatic trypsin inhibitor.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 72, no. 6, pp. 2002–6, 1975.

[121] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1656–1676, 2004.

[122] C. Oostenbrink, A. E. Thereza, A. Soares, N. F. A. Van Der, V. Ae, and W. F. Van Gunsteren, "Validation of the 53A6 GROMOS force field," *Biophysical Journal*, 2004.

[123] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, "The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations," *Journal of Physical Chemistry B*, vol. 111, no. 27, pp. 7812–7824, 2007.

[124] W. L. Jorgensen and J. Tirado-Rives, "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin," *Journal of the American Chemical Society*, vol. 110, no. 6, pp. 1657–1666, 1988.

[125] J. Huang and A. D. MacKerell, "CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data," *Journal of Computational Chemistry*, vol. 34, no. 25, pp. 2135–2145, 2013.

[126] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell, "CHARMM36m: an improved force field for folded and intrinsically disordered proteins," *Nature Methods*, vol. 14, no. 1, pp. 71–73, 2017.

[127] C. Kittel, *Introduction to solide state physics*. John Wiley and Sons Inc., 8 ed., 2004.

[128] A. Y. Toukmaji and J. A. Board, "Ewald summation techniques in perspective: a survey," *Computer Physics Communications*, vol. 95, no. 2-3, pp. 73–92, 1996.

[129] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N log(N) method for Ewald sums in large systems," *The Journal of Chemical Physics*, vol. 98, no. 12, pp. 10089–10092, 1993.

[130] M. Fitzner, L. Joly, M. Ma, G. C. Sosso, A. Zen, and A. Michaelides, "Communication: Truncated non-bonded potentials can yield unphysical behavior in molecular dynamics simulations of interfaces," *The Journal of Chemical Physics*, vol. 147, p. 121102, 2017.

[131] H. J. Berendsen, J. P. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of Chemical Physics*, vol. 81, no. 8, pp. 3684–3690, 1984.

[132] T. Morishita, "Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath," *The Journal of Chemical Physics*, vol. 113, no. 8, pp. 2976–2982, 2000.

[133] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *The Journal of Chemical Physics*, vol. 126, no. 1, p. 014101, 2007.

[134] H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature," *The Journal of Chemical Physics*, vol. 72, no. 4, pp. 2384–2393, 1980.

[135] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *Journal of Applied Physics*, vol. 52, no. 12, pp. 7182–7190, 1981.

[136] S. Nosé and M. Klein, "Constant pressure molecular dynamics for molecular systems," *Molecular Physics*, vol. 50, no. 5, pp. 1055–1076, 1983.

[137] A. R. Leach, *Molecular modelling : principles and applications*. Harlow, England: Pearson, Prentice Hall, 2nd ed., 2001.

[138] S. J. Marrink and D. P. Tieleman, "Perspective on the Martini model.," *Chemical Society Reviews*, vol. 42, no. 16, pp. 6801–22, 2013.

[139] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, "The MARTINI Coarse-Grained Force Field: Extension to Proteins," *Journal of Chemical Theory and Computation*, vol. 4, no. 5, pp. 819–834, 2008.

[140] J. J. Uusitalo, H. I. Ingólfsson, P. Akhshi, D. P. Tieleman, and S. J. Marrink, "Martini Coarse-Grained Force Field: Extension to DNA," *Journal of Chemical Theory and Computation*, vol. 11, no. 8, pp. 3932–3945, 2015.

[141] J. J. Uusitalo, H. I. Ingólfsson, S. J. Marrink, and I. Faustino, "Martini Coarse-Grained Force Field: Extension to RNA," *Biophysical Journal*, vol. 113, no. 2, pp. 246–256, 2017.

[142] X. Periole, M. Cavalli, S.-J. Marrink, and M. A. Ceruso, "Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition," *Journal of Chemical Theory and Computation*, vol. 5, no. 9, pp. 2531–2543, 2009.

[143] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, 2008.

[144] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1-2, pp. 19–25, 2015.

[145] P. J. Stansfeld and M. S. Sansom, "From Coarse Grained to Atomistic: A Serial Multiscale Approach to Membrane Protein Simulations," *Journal of Chemical Theory and Computation*, vol. 7, no. 4, pp. 1157–1166, 2011.

[146] T. E. Oliphant, "Python for Scientific Computing," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 10–20, 2007.

[147] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, "MDAnalysis: A toolkit for the analysis of molecular dynamics simulations," *Journal of Computational Chemistry*, vol. 32, no. 10, pp. 2319–2327, 2011.

[148] R. Gowers, M. Linke, J. Barnoud, T. Reddy, M. Melo, S. Seyler, J. Domański, D. Dotson, S. Buchoux, I. Kenney, and O. Beckstein, "MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations," in *Proceedings of the 15th Python in Science Conference*, pp. 98–105, 2016.

[149] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories," *Biophysical Journal*, vol. 109, no. 8, pp. 1528–1532, 2015.

[150] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[151] E. Jones, T. Oliphant, and P. Peterson, "SciPy: Open Source Scientific Tools for Python," 2001.

[152] W. McKinney, "Data Structures for Statistical Computing in Python," 2010.

[153] F. Perez and B. E. Granger, "IPython: A System for Interactive Scientific Computing," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 21–29, 2007.

[154] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[155] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[156] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, "MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale.," *Journal of Chemical Theory and Computation*, vol. 7, no. 10, pp. 3412–3419, 2011.

[157] M. P. Harrigan, M. M. Sultan, C. X. Hernández, B. E. Husic, P. Eastman, C. R. Schwantes, K. A. Beauchamp, R. T. McGibbon, and V. S. Pande, "MSMBuilder: Statistical Models for Biomolecular Dynamics," *Biophysical Journal*, vol. 112, no. 1, pp. 10–15, 2017.

[158] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, "PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models," *Journal of Chemical Theory and Computation*, vol. 11, no. 11, pp. 5525–5542, 2015.

[159] W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[160] S. A. Hollingsworth and R. O. Dror, "Molecular Dynamics Simulation for All," *Neuron*, vol. 99, no. 6, pp. 1129–1143, 2018.

[161] G. B. Goh, B. S. Hulbert, H. Zhou, and C. L. Brooks, "Constant pH molecular dynamics of proteins in explicit solvent with proton tautomerism," *Proteins: Structure, Function, and Bioinformatics*, vol. 82, no. 7, pp. 1319–1331, 2014.

[162] A. Warshel, Mitsunori Kato, , and A. V. Pisliakov, "Polarizable Force Fields: History, Test Cases, and Prospects," *Journal of Chemical Theory and Computation*, vol. 3, no. 6, pp. 2034–2045, 2007.

[163] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Systematic Validation of Protein Force Fields against Experimental Data," *PLoS ONE*, vol. 7, no. 2, p. e32131, 2012.

[164] B. Knapp, L. Ospina, and C. M. Deane, "Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas," *Journal of Chemical Theory and Computation*, vol. 14, no. 12, pp. 6127–6138, 2018.

[165] D. E. Shaw, J. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L.-S. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y.-H. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. B. Schafer, N. Siddique, C. W. Snyder, J. Spengler, P. T. P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang, and C. Young, "Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer," in *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 41–53, IEEE, 2014.

[166] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical Physics Letters*, vol. 314, no. 1-2, pp. 141–151, 1999.

[167] A. Laio and F. L. Gervasio, "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science," *Reports on Progress in Physics*, vol. 71, no. 12, p. 126601, 2008.

[168] L. Sawle and K. Ghosh, "Convergence of Molecular Dynamics Simulation of Protein Native States: Feasibility vs Self-Consistency Dilemma," *Journal of Chemical Theory and Computation*, vol. 12, no. 2, pp. 861–869, 2016.

[169] R. Zwanzig, "From classical dynamics to continuous time random walks," *Journal of Statistical Physics*, vol. 30, no. 2, pp. 255–262, 1983.

[170] G. R. Bowman, V. S. Pande, and F. Noé, eds., *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, vol. 797 of *Advances in Experimental Medicine and Biology*. Dordrecht: Springer Netherlands, 2014.

[171] C. R. Schwantes and V. S. Pande, "Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9.," *Journal of Chemical Theory and Computation*, vol. 9, no. 4, pp. 2000–2009, 2013.

[172] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," *The Journal of Chemical Physics*, vol. 139, no. 1, p. 015102, 2013.

[173] J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane, "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming," in *Artificial Intelligence in Design*, pp. 151–170, Dordrecht: Springer Netherlands, 1996.

[174] K. Murphy, *Machine Learning, a Probabilistic Perspective*. Cambridge: MIT Press, 2012.

[175] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[176] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[177] M. E. Timmerman, "Principal Component Analysis," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 1082–1083, 2003.

[178] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[179] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, (Berkeley, California), pp. 281–297, University of California Press, 1967.

[180] R. V. Stahelin, J. L. Scott, and C. T. Frick, "Cellular and molecular interactions of phosphoinositides and peripheral proteins," *Chemistry and Physics of Lipids*, vol. 182, pp. 3–18, 2014.

[181] T. G. Kutateladze, "Translation of the phosphoinositide code by PI effectors," *Nature Chemical Biology*, vol. 6, no. 7, pp. 507–513, 2010.

[182] G. Di Paolo and P. De Camilli, "Phosphoinositides in cell regulation and membrane dynamics," *Nature*, vol. 443, no. 7112, pp. 651–657, 2006.

[183] S. B. Hansen, "Lipid agonism: The PIP2 paradigm of ligand-gated ion channels," *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, vol. 1851, no. 5, pp. 620–628, 2015.

[184] I. E. Michailidis, R. Rusinova, A. Georgakopoulos, Y. Chen, R. Iyengar, N. K. Robakis, D. E. Logothetis, and L. Baki, "Phosphatidylinositol-4,5-bisphosphate regulates epidermal growth factor receptor activation," *European Journal of Physiology*, vol. 461, no. 3, pp. 387–397, 2011.

[185] M. J. Wishart and J. E. Dixon, "PTEN and myotubularin phosphatases: from 3-phosphoinositide dephosphorylation to disease," *Trends in Cell Biology*, vol. 12, no. 12, pp. 579–585, 2002.

[186] G. Di Paolo, L. Pellegrini, K. Letinic, G. Cestra, R. Zoncu, S. Voronov, S. Chang, J. Guo, M. R. Wenk, and P. De Camilli, "Recruitment and regulation of phosphatidylinositol phosphate kinase type 1γ by the FERM domain of talin," *Nature*, vol. 420, no. 6911, pp. 85–89, 2002.

[187] W.-T. Chao, A. C. Daquinag, F. Ashcroft, and J. Kunz, "Type I PIPK-alpha regulates directed cell migration by modulating Rac1 plasma membrane targeting and activation.," *The Journal of Cell Biology*, vol. 190, no. 2, pp. 247–62, 2010.

[188] A. Honda, M. Nogami, T. Yokozeki, M. Yamazaki, H. Nakamura, H. Watanabe, K. Kawamoto, K. Nakayama, A. J. Morris, M. A. Frohman, and Y. Kanaho, "Phosphatidylinositol 4-Phosphate 5-Kinase α Is a Downstream Effector of the Small G Protein ARF6 in Membrane Ruffle Formation," *Cell*, vol. 99, no. 5, pp. 521–532, 1999.

[189] D. H. Jones, J. B. Morris, C. P. Morgan, H. Kondo, R. F. Irvine, and S. Cockcroft, "Type I phosphatidylinositol 4-phosphate 5-kinase directly interacts with ADP-ribosylation factor 1 and is responsible for phosphatidylinositol 4,5-bisphosphate synthesis in the golgi compartment.," *The Journal of Biological Chemistry*, vol. 275, no. 18, pp. 13962–6, 2000.

[190] J. Semenas, A. Hedblom, R. R. Miftakhova, M. Sarwar, R. Larsson, L. Shcherbina, M. E. Johansson, P. Härkönen, O. Sterner, and J. L. Persson, "The role of PI3K/AKT-related PIP5K1α and the discovery of its selective inhibitor for treatment of advanced prostate cancer.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 35, pp. E3689–98, 2014.

[191] V. D. Rao, S. Misra, I. V. Boronenkov, R. A. Anderson, and J. H. Hurley, "Structure of Type IIβ Phosphatidylinositol Phosphate Kinase: A Protein Kinase Fold Flattened for Interfacial Phosphorylation," *Cell*, vol. 94, no. 6, pp. 829–839, 1998.

[192] A. Liu, D. Sui, D. Wu, and J. Hu, "The activation loop of PIP5K functions as a membrane sensor essential for lipid substrate processing," *Science Advances*, vol. 2, no. 11, 2016.

[193] G. Hedger and M. S. Sansom, "Lipid interaction sites on channels, transporters and receptors: Recent insights from molecular dynamics simulations," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1858, no. 10, pp. 2390–2400, 2016.

[194] A. C. Kalli and M. S. P. Sansom, "Interactions of peripheral proteins with model membranes as viewed by molecular dynamics simulations.," *Biochemical Society Transactions*, vol. 42, no. 5, pp. 1418–24, 2014.

[195] Walter R. P. Scott, Philippe H. Hünenberger, Ilario G. Tironi, Alan E. Mark, Salomon R. Billeter, Jens Fennen, Andrew E. Torda, Thomas Huber, Peter Krüger, , and W. F. van Gunsteren, "The GROMOS Biomolecular Simulation

Program Package," *Journal of Physical Chemistry A*, vol. 103, no. 19, pp. 3596–3607, 1999.

[196] A. Fiser and A. Šali, "Modeller: Generation and Refinement of Homology-Based Protein Structure Models," *Methods in Enzymology*, vol. 374, pp. 461–491, 2003.

[197] H. Koldsø, D. Shorthouse, J. Hélie, and M. S. P. Sansom, "Lipid Clustering Correlates with Membrane Curvature as Revealed by Molecular Simulations of Complex Lipid Bilayers," *PLoS Computational Biology*, vol. 10, no. 10, p. e1003911, 2014.

[198] S. Cao, S. Chung, S. Kim, Z. Li, D. Manor, and M. Buck, "K-Ras G-domain binding with signaling lipid phosphatidylinositol (4,5)-phosphate (PIP2): membrane association, protein orientation, and function.," *The Journal of Biological Chemistry*, vol. 294, no. 17, pp. 7068–7084, 2019.

[199] K. Suhre and Y.-H. Sanejouand, "On the potential of normal-mode analysis for solving difficult molecular-replacement problems," *Acta Crystallographica*, pp. 60–796, 2004.

[200] P. W. Fowler, J. Hélie, A. Duncan, M. Chavent, H. Koldsø, and M. S. P. Sansom, "Membrane stiffness is modified by integral membrane proteins," *Soft Matter*, vol. 12, no. 37, pp. 7792–7803, 2016.

[201] M. Javanainen, H. Hammaren, L. Monticelli, J.-H. Jeon, M. S. Miettinen, H. Martinez-Seara, R. Metzler, and I. Vattulainen, "Anomalous and normal diffusion of proteins and lipids in crowded lipid membranes," *Faraday Discuss.*, vol. 161, no. 0, pp. 397–417, 2013.

[202] A. C. Stark, C. T. Andrews, and A. H. Elcock, "Toward optimized potential functions for protein-protein interactions in aqueous solutions: osmotic second virial coefficient calculations using the MARTINI coarse-grained force field.," *Journal of Chemical Theory and Computation*, vol. 9, no. 9, 2013.

[203] J. Domański, G. Hedger, R. B. Best, P. J. Stansfeld, and M. S. P. Sansom, "Convergence and Sampling in Determining Free Energy Landscapes for Membrane Protein Association," *The Journal of Physical Chemistry B*, vol. 121, no. 15, pp. 3364–3375, 2017.

[204] C. Arnarez, S. J. Marrink, and X. Periole, "Identification of cardiolipin binding sites on cytochrome c oxidase at the entrance of proton channels," *Scientific Reports*, vol. 3, no. 1, p. 1263, 2013.

[205] C. Arnarez, J.-P. Mazat, J. Elezgaray, S.-J. Marrink, and X. Periole, "Evidence for Cardiolipin Binding Sites on the Membrane-Exposed Surface of the Cytochrome bc 1," *Journal of the American Chemical Society*, vol. 135, no. 8, pp. 3112–3120, 2013.

[206] M. R. Schmidt, P. J. Stansfeld, S. J. Tucker, and M. S. P. Sansom, "Simulation-Based Prediction of Phosphatidylinositol 4,5-Bisphosphate Binding to an Ion Channel," *Biochemistry*, vol. 52, no. 2, pp. 279–281, 2013.

[207] F. B. Naughton, A. C. Kalli, and M. S. Sansom, "Structural and Energetic Details of the Binding of PTEN to Phosphatidylinositol Phosphate-Containing Membranes through Molecular Simulations," *Biophysical Journal*, vol. 114, no. 3, p. 277a, 2018.

[208] H. I. Ingólfsson, M. N. Melo, F. J. van Eerden, C. Arnarez, C. A. Lopez, T. A. Wassenaar, X. Periole, A. H. de Vries, D. P. Tieleman, and S. J. Marrink, "Lipid Organization of the Plasma Membrane," *Journal of the American Chemical Society*, vol. 136, no. 41, pp. 14554–14559, 2014.

[209] V. S. Pande, K. Beauchamp, and G. R. Bowman, "Everything you wanted to know about Markov State Models but were afraid to ask," *Methods*, vol. 52, no. 1, pp. 99–105, 2010.

[210] J. D. Chodera and F. Noé, "Markov state models of biomolecular conformational dynamics.," *Current Opinion in Structural Biology*, vol. 25, pp. 135–44, 2014.

[211] M. Schramp, A. Hedman, W. Li, X. Tan, and R. Anderson, "PIP kinases from the cell membrane to the nucleus.," *Sub-cellular Biochemistry*, vol. 58, pp. 25–59, 2012.

[212] C. Gao and Y.-G. Chen, "Dishevelled: The hub of Wnt signaling.," *Cellular Signalling*, vol. 22, no. 5, pp. 717–27, 2010.

[213] P. Robustelli, S. Piana, and D. E. Shaw, "Developing a molecular dynamics force field for both folded and disordered protein states.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 21, pp. E4758–E4766, 2018.

[214] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[215] S.-B. T. Amos, A. C. Kalli, J. Shi, and M. S. Sansom, "Membrane Recognition and Binding by the Phosphatidylinositol Phosphate Kinase PIP5K1A: A Multiscale Simulation Study," *Structure*, vol. 27, no. 8, pp. 1336–1346.e2, 2019.

[216] K. D. Eichenbaum, Y. RodrÃguez, M. Mezei, and R. Osman, "The energetics of the acetylation switch in p53-mediated transcriptional activation," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 2, pp. 447–456, 2010.

[217] T. Guo, W. Noble, and D. P. Hanger, "Roles of tau protein in health and disease," *Acta Neuropathologica*, vol. 133, no. 5, pp. 665–704, 2017.

[218] P. E. Wright and H. J. Dyson, "Intrinsically disordered proteins in cellular signalling and regulation.," *Nature Reviews Molecular cell biology*, vol. 16, no. 1, pp. 18–29, 2015.

[219] D. Shukla, C. X. Hernández, J. K. Weber, and V. S. Pande, "Markov State Models Provide Insights into Dynamic Modulation of Protein Function," *Accounts of Chemical Research*, vol. 48, no. 2, pp. 414–422, 2015.

[220] M. Goedert, R. Jakes, and M. G. Spillantini, "The Synucleinopathies: Twenty Years On," *Journal of Parkinson's Disease*, vol. 7, no. s1, pp. S51–S69, 2017.

[221] M. H. Polymeropoulos, C. Lavedan, E. Leroy, S. E. Ide, A. Dehejia, A. Dutra, B. Pike, H. Root, J. Rubenstein, R. Boyer, E. S. Stenroos, S. Chandrasekharappa, A. Athanassiadou, T. Papapetropoulos, W. G. Johnson, A. M. Lazzarini, R. C. Duvoisin, G. Di Iorio, L. I. Golbe, and R. L. Nussbaum, "Mutation in the alpha-synuclein gene identified in families with Parkinson's disease.," *Science*, vol. 276, no. 5321, pp. 2045–7, 1997.

[222] C. L. Ugalde, D. I. Finkelstein, V. A. Lawson, and A. F. Hill, "Pathogenic mechanisms of prion protein, amyloid-β and α-synuclein misfolding: the prion concept and neurotoxicity of protein oligomers," *Journal of Neurochemistry*, vol. 139, no. 2, pp. 162–180, 2016.

[223] B. Fauvet, M. K. Mbefo, M.-B. Fares, C. Desobry, S. Michael, M. T. Ardah, E. Tsika, P. Coune, M. Prudent, N. Lion, D. Eliezer, D. J. Moore, B. Schneider, P. Aebischer, O. M. El-Agnaf, E. Masliah, and H. A. Lashuel, "α-Synuclein in Central Nervous System and from Erythrocytes, Mammalian Cells, and Escherichia coli Exists Predominantly as Disordered Monomer," *Journal of Biological Chemistry*, vol. 287, no. 19, pp. 15345–15364, 2012.

[224] T. Bartels, J. G. Choi, and D. J. Selkoe, "α-Synuclein occurs physiologically as a helically folded tetramer that resists aggregation," *Nature*, vol. 477, no. 7362, pp. 107–110, 2011.

[225] J. Burre, M. Sharma, T. Tsetsenis, V. Buchman, M. R. Etherton, and T. C. Sudhof, "-Synuclein Promotes SNARE-Complex Assembly in Vivo and in Vitro," *Science*, vol. 329, no. 5999, pp. 1663–1667, 2010.

[226] L. M. Oliveira, L. J. Falomir-Lockhart, M. G. Botelho, K. H. Lin, P. Wales, J. C. Koch, E. Gerhardt, H. Taschenberger, T. F. Outeiro, P. Lingor, B. Schüle, D. J. Arndt-Jovin, and T. M. Jovin, "Elevated α-synuclein caused by SNCA gene triplication impairs neuronal differentiation and maturation in Parkinson's patient-derived induced pluripotent stem cells," *Cell Death and Disease*, vol. 6, no. 11, pp. e1994–e1994, 2015.

[227] H. Jin, A. Kanthasamy, A. Ghosh, Y. Yang, V. Anantharam, and A. G. Kanthasamy, "-Synuclein Negatively Regulates Protein Kinase C Expression to Suppress Apoptosis in Dopaminergic Neurons by Reducing p300 Histone Acetyltransferase Activity," *Journal of Neuroscience*, vol. 31, no. 6, pp. 2035–2051, 2011.

[228] K. Sode, S. Ochiai, N. Kobayashi, and E. Usuzaka, "Effect of reparation of repeat sequences in the human alpha-synuclein on fibrillation ability.," *International Journal of Biological Sciences*, vol. 3, no. 1, pp. 1–7, 2006.

[229] P. Flagmeier, G. Meisl, M. Vendruscolo, T. P. J. Knowles, C. M. Dobson, A. K. Buell, and C. Galvagnion, "Mutations associated with familial Parkinson's disease alter the initiation and amplification steps of α-synuclein aggregation.,"

Proceedings of the National Academy of Sciences of the United States of America, vol. 113, no. 37, pp. 10328–33, 2016.

[230] A. Leitao, A. Bhumkar, D. J. B. Hunter, Y. Gambin, and E. Sierecki, "Unveiling a Selective Mechanism for the Inhibition of α-Synuclein Aggregation by β-Synuclein.," *International Journal of Molecular Sciences*, vol. 19, no. 2, 2018.

[231] N. Bengoa-Vergniory, R. F. Roberts, R. Wade-Martins, and J. Alegre-Abarrategui, "Alpha-synuclein oligomers: a new hope.," *Acta Neuropathologica*, vol. 134, no. 6, pp. 819–838, 2017.

[232] B. Winner, R. Jappelli, S. K. Maji, P. A. Desplats, L. Boyer, S. Aigner, C. Hetzer, T. Loher, M. Vilar, S. Campioni, C. Tzitzilonis, A. Soragni, S. Jessberger, H. Mira, A. Consiglio, E. Pham, E. Masliah, F. H. Gage, and R. Riek, "In vivo demonstration that α-synuclein oligomers are toxic," *Proceedings of the National Academy of Sciences*, vol. 108, no. 10, pp. 4194–4199, 2011.

[233] S. W. Chen, S. Drakulic, E. Deas, M. Ouberai, F. A. Aprile, R. Arranz, S. Ness, C. Roodveldt, T. Guilliams, E. J. De-Genst, D. Klenerman, N. W. Wood, T. P. J. Knowles, C. Alfonso, G. Rivas, A. Y. Abramov, J. M. Valpuesta, C. M. Dobson, and N. Cremades, "Structural characterization of toxic oligomers that are kinetically trapped during α-synuclein fibril formation.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 16, pp. E1994–2003, 2015.

[234] T. F. Outeiro, P. Putcha, J. E. Tetzlaff, R. Spoelgen, M. Koker, F. Carvalho, B. T. Hyman, and P. J. McLean, "Formation of toxic oligomeric alpha-synuclein species in living cells.," *PloS one*, vol. 3, no. 4, p. e1867, 2008.

[235] A. Pineda and J. Burré, "Modulating membrane binding of α-synuclein as a therapeutic strategy," *Proceedings of the National Academy of Sciences*, vol. 114, no. 6, pp. 1223–1225, 2017.

[236] M. Perni, C. Galvagnion, A. Maltsev, G. Meisl, M. B. D. Müller, P. K. Challa, J. B. Kirkegaard, P. Flagmeier, S. I. A. Cohen, R. Cascella, S. W. Chen, R. Limbocker, P. Sormanni, G. T. Heller, F. A. Aprile, N. Cremades, C. Cecchi, F. Chiti, E. A. A. Nollen, T. P. J. Knowles, M. Vendruscolo, A. Bax, M. Zasloff, and C. M. Dobson, "A natural product inhibits the initiation of α-synuclein aggregation and suppresses its toxicity.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 6, pp. E1009–E1017, 2017.

[237] G. Fusco, A. De Simone, P. Arosio, M. Vendruscolo, G. Veglia, and C. M. Dobson, "Structural Ensembles of Membrane-bound α-Synuclein Reveal the Molecular Determinants of Synaptic Vesicle Affinity.," *Scientific Reports*, vol. 6, p. 27125, 2016.

[238] T. A. Wassenaar, H. I. Ingólfsson, R. A. Böckmann, D. P. Tieleman, and S. J. Marrink, "Computational Lipidomics with insane : A Versatile Tool for Generating Custom Membranes for Molecular Simulations," *Journal of Chemical Theory and Computation*, vol. 11, no. 5, pp. 2144–2155, 2015.

[239] J. Lemkul, "Scripts: http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/scripts.html," 2007.

[240] A. C. Kalli, M. S. P. Sansom, and R. A. F. Reithmeier, "Molecular Dynamics Simulations of the Bacterial UraA H+-Uracil Symporter in Lipid Bilayers Reveal a Closed State and a Selective Interaction with Cardiolipin," *PLOS Computational Biology*, vol. 11, no. 3, p. e1004123, 2015.

[241] D. S. Wishart, B. D. Sykes, and F. M. Richards, "The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy," *Biochemistry*, vol. 31, no. 6, pp. 1647–1651, 1992.

[242] D. Wishart and B. Sykes, "The 13C Chemical-Shift Index: A simple method for the identification of protein secondary structure using 13C chemical-shift data," *Journal of Biomolecular NMR*, vol. 4, no. 2, pp. 171–180, 1994.

[243] T. S. Ulmer, A. Bax, N. B. Cole, and R. L. Nussbaum, "Structure and dynamics of micelle-bound human alpha-synuclein.," *The Journal of Biological Chemistry*, vol. 280, no. 10, pp. 9595–603, 2005.

[244] C. Liao, V. May, and J. Li, "Assessment of Conformational State Transitions of Class B GPCRs Using Molecular Dynamics.," *Methods in Mol*, vol. 1947, pp. 3–19, 2019.

[245] A. Sirur, D. De Sancho, and R. B. Best, "Markov state models of protein misfolding," *The Journal of Chemical Physics*, vol. 144, no. 7, p. 075101, 2016.

[246] U. Sengupta, M. Carballo-Pacheco, and B. Strodel, "Automated Markov state models for molecular dynamics simulations of aggregation and self-assembly," *The Journal of Chemical Physics*, vol. 150, no. 11, p. 115101, 2019.

[247] A. Balupuri, K.-E. Choi, and N. S. Kang, "Computational insights into the role of α-strand/sheet in aggregation of α-synuclein," *Scientific Reports*, vol. 9, no. 1, p. 59, 2019.

[248] J. Pujols, S. Peña-Díaz, D. F. Lázaro, F. Peccati, F. Pinheiro, D. González, A. Carija, S. Navarro, M. Conde-Giménez, J. García, S. Guardiola, E. Giralt, X. Salvatella, J. Sancho, M. Sodupe, T. F. Outeiro, E. Dalfó, and S. Ventura, "Small molecule inhibits α-synuclein aggregation, disrupts amyloid fibrils, and prevents degeneration of dopaminergic neurons.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 41, pp. 10481–10486, 2018.

[249] B. E. Husic, R. T. McGibbon, M. M. Sultan, and V. S. Pande, "Optimized parameter selection reveals trends in Markov state models for protein folding," *The Journal of Chemical Physics*, vol. 145, no. 19, p. 194103, 2016.

[250] H. I. Ingólfsson, C. Arnarez, X. Periole, and S. J. Marrink, "Computational 'microscopy' of cellular membranes.," *Journal of Cell Science*, vol. 129, no. 2, pp. 257–68, 2016.

[251] G. van Meer, "Cellular lipidomics," *The EMBO Journal*, vol. 24, no. 18, pp. 3159–3165, 2005.

[252] F. J. van Eerden, D. H. de Jong, A. H. de Vries, T. A. Wassenaar, and S. J. Marrink, "Characterization of thylakoid lipid membranes from cyanobacteria and higher plants by molecular dynamics simulations," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1848, no. 6, pp. 1319–1330, 2015.

[253] T. Apajalahti, P. Niemelä, P. N. Govindan, M. S. Miettinen, E. Salonen, S.-J. Marrink, and I. Vattulainen, "Concerted diffusion of lipids in raft-like membranes," *Faraday Discuss.*, vol. 144, no. 0, pp. 411–430, 2010.

[254] J. E. Goose and M. S. P. Sansom, "Reduced Lateral Mobility of Lipids and Proteins in Crowded Membranes," *PLoS Computational Biology*, vol. 9, no. 4, p. e1003033, 2013.

[255] L. Iversen, S. Mathiasen, J. B. Larsen, and D. Stamou, "Membrane curvature bends the laws of physics and chemistry," *Nature Chemical Biology*, vol. 11, no. 11, pp. 822–825, 2015.

[256] G. Guigas and M. Weiss, "Effects of protein crowding on membrane systems," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1858, no. 10, pp. 2441–2450, 2016.

[257] C. Eggeling, C. Ringemann, R. Medda, G. Schwarzmann, K. Sandhoff, S. Polyakova, V. N. Belov, B. Hein, C. von Middendorff, A. Schönle, and S. W. Hell, "Direct observation of the nanoscale dynamics of membrane lipids in a living cell," *Nature*, vol. 457, no. 7233, pp. 1159–1162, 2009.

[258] A. Honigmann, V. Mueller, H. Ta, A. Schoenle, E. Sezgin, S. W. Hell, and C. Eggeling, "Scanning STED-FCS reveals spatiotemporal heterogeneity of lipid interaction in the plasma membrane of living cells," *Nature Communications*, vol. 5, no. 1, p. 5412, 2014.

[259] E. Hosy, C. Butler, and J.-B. Sibarita, "Organization and dynamics of AMPA receptors inside synapses—nano-organization of AMPA receptors and main synaptic scaffolding proteins revealed by super-resolution imaging," *Current Opinion in Chemical Biology*, vol. 20, pp. 120–126, 2014.

[260] P. Rassam, N. A. Copeland, O. Birkholz, C. Tóth, M. Chavent, A. L. Duncan, S. J. Cross, N. G. Housden, R. Kaminska, U. Seger, D. M. Quinn, T. J. Garrod, M. S. P. Sansom, J. Piehler, C. G. Baumann, and C. Kleanthous, "Supramolecular assemblies underpin turnover of outer membrane proteins in bacteria," *Nature*, vol. 523, no. 7560, pp. 333–336, 2015.

[261] X. Periole, A. M. Knepp, T. P. Sakmar, S. J. Marrink, and T. Huber, "Structural Determinants of the Supramolecular Organization of G Protein-Coupled Receptors in Bilayers," *Journal of the American Chemical Society*, vol. 134, no. 26, pp. 10959–10965, 2012.

[262] D. Provasi, M. B. Boz, J. M. Johnston, and M. Filizola, "Preferred Supramolecular Organization and Dimer Interfaces of Opioid Receptors from Simulated Self-Association," *PLOS Computational Biology*, vol. 11, no. 3, p. e1004148, 2015.

[263] H. Koldsø and M. S. P. Sansom, "Organization and Dynamics of Receptor Proteins in a Plasma Membrane," *Journal of the American Chemical Society*, vol. 137, no. 46, pp. 14694–14704, 2015.

[264] K. M. Davies, C. Anselmi, I. Wittig, J. D. Faraldo-Gómez, and W. Kühlbrandt, "Structure of the yeast F1Fo-ATP synthase dimer and its role in shaping the mitochondrial cristae.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 34, pp. 13602–7, 2012.

[265] L. Janosi, Z. Li, J. F. Hancock, and A. A. Gorfe, "Organization, dynamics, and segregation of Ras nanoclusters in membrane domains.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 21, pp. 8097–102, 2012.

[266] C. Mim and V. M. Unger, "Membrane curvature and its generation by BAR proteins.," *Trends in Biochemical Sciences*, vol. 37, no. 12, pp. 526–33, 2012.

[267] M. Simunovic and G. A. Voth, "Membrane tension controls the assembly of curvature-generating proteins," *Nature Communications*, vol. 6, no. 1, p. 7219, 2015.

[268] M. Simunovic, G. A. Voth, A. Callan-Jones, and P. Bassereau, "When Physics Takes Over: BAR Proteins and Membrane Curvature," *Trends in Cell Biology*, vol. 25, no. 12, pp. 780–792, 2015.

[269] A. West, B. E. Brummel, A. R. Braun, E. Rhoades, and J. N. Sachs, "Membrane remodeling and mechanics: Experiments and simulations of α-Synuclein," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1858, no. 7, pp. 1594–1609, 2016.

[270] L. Mcinnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, 2017.

[271] E. Jefferys, M. S. P. Sansom, and P. W. Fowler, "NRas slows the rate at which a model lipid bilayer phase separates.," *Faraday discussions*, vol. 169, pp. 209–23, 2014.

[272] H. Alimohamadi and P. Rangamani, "Modeling Membrane Curvature Generation due to MembraneProtein Interactions.," *Biomolecules*, vol. 8, no. 4, 2018.

[273] L. Paparelli, N. Corthout, B. Pavie, W. Annaert, and S. Munck, "Analyzing Protein Clusters on the Plasma Membrane: Application of Spatial Statistical Analysis Methods on Super-Resolution Microscopy Images," in *Advances in Anatomy, Embryology, and Cell Biology*, vol. 219, pp. 95–122, 2016.

[274] A. L. Duncan, T. Reddy, H. Koldsø, J. Hélie, P. W. Fowler, M. Chavent, and M. S. P. Sansom, "Protein crowding and lipid complexity influence the nanoscale dynamic organization of ion channels in cell membranes," *Scientific Reports*, vol. 7, no. 1, p. 16647, 2017.

[275] G. R. V. Hammond, M. J. Fischer, K. E. Anderson, J. Holdich, A. Koteci, T. Balla, and R. F. Irvine, "PI4P and PI(4,5)P2 are essential but independent lipid determinants of membrane identity.," *Science*, vol. 337, no. 6095, pp. 727–30, 2012.

[276] L. Picas, F. Gaits-Iacovoni, and B. Goud, "The emerging role of phosphoinositide clustering in intracellular trafficking and signal transduction.," *F1000Research*, vol. 5, 2016.

[277] H. Zhao, A. Michelot, E. V. Koskela, V. Tkach, D. Stamou, D. G. Drubin, and P. Lappalainen, "Membrane-Sculpting BAR Domains Generate Stable Lipid Microdomains," *Cell Reports*, vol. 4, no. 6, pp. 1213–1223, 2013.

[278] U. Golebiewska, A. Gambhir, G. Hangyás-Mihályné, I. Zaitseva, J. Rädler, and S. McLaughlin, "Membrane-Bound Basic Peptides Sequester Multivalent (PIP2), but Not Monovalent (PS), Acidic Lipids," *Biophysical Journal*, vol. 91, no. 2, pp. 588–599, 2006.

[279] E. Yamamoto, T. Akimoto, A. C. Kalli, K. Yasuoka, and M. S. P. Sansom, "Dynamic interactions between a membrane binding protein and lipids induce fluctuating diffusivity," *Science Advances*, vol. 3, no. 1, p. e1601871, 2017.

[280] C. A. López, V. V. Vesselinov, S. Gnanakaran, and B. S. Alexandrov, "Unsupervised Machine Learning for Analysis of Coexisting Lipid Phases and Domain Growth in Biological Membranes," *ArXiV*, 2019.

[281] D. E. Shaw, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, R. O. Dror, S. Piana, Y. Shan, B. Towles, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, and B. Batson, "Millisecond-scale molecular dynamics simulations on Anton," in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09*, (New York, New York, USA), p. 1, ACM Press, 2009.