

Tools and resources for molecular simulations of integral membrane proteins



Thomas D. Newport
St Cross College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2017

Acknowledgements

My thanks firstly go to supervisors Mark Sansom and Phill Stansfeld, for their guidance and support over the past three years, and to the EPSRC's Systems Biology DTC for their funding.

Many thanks to those who made SBCB such an enjoyable place to work, in particular, Amanda, Anna, Fiona, Heidi, Laura, Maria, Nick, Ole, Shanlin, Teresa and Vishal. Thanks for the tea breaks, lunch breaks, pub trips, baking and craft sessions.

I would also like to thank my colleagues at Oxford Nanopore Technologies Ltd. for helping me to acquire the technological know-how to build MemProtMD, as well as a much-needed change of scene (and subsequent gainful employment). Special thanks go to my managers Richard Carter and Angelos Molfetas, fellow citizens of Row 6, and all those who appreciate a proper lunch break.

Finally, thanks go to Alice, Rachel and Mihail for keeping me more or less grounded in a world beyond my work, my fiancée Tonya for her love and support through the highs and lows of the past years, and my parents Jim and Denise, and sister Emma, for a lifetime of encouragement and inspiration.

Abstract

Integral Membrane Proteins (IMPs) are an important and scientifically interesting class of protein which span the lipid bilayer surrounding cells, cell compartments and many viruses. Molecular Dynamics (MD) simulation has revealed intimate and often highly specific relationships between membrane lipids and IMPs critical to many metabolic and signalling pathways. Meanwhile, the use of Coarse-Grained (CG) MD techniques has extended capabilities of biomolecular simulation to larger proteins over longer time periods. Several tools and resources for biomolecular simulations of IMPs are presented here, as well as two MD studies of specific IMPs. The previously developed MemProtMD pipeline automates the setup of MD simulations of IMPs; major extensions to this are presented here with the MemProtMD database and web server, automating the analysis of IMP simulations. The results of this can be viewed using the MemProtMD web server, an interactive, searchable online resource containing data from simulations of over 3000 experimentally determined IMP structures in explicit lipid bilayers. Using data from analysis of the entire MemProtMD database, MemProtMetrics has been developed to automate identification and orientation of IMP structures from Protein DataBank (PDB) depositions. This is shown to effectively predict membrane protein orientations seen in MD simulations. A tool for identification and classification of membrane lipids is also described, and used to identify over 500 IMPs structures with resolved lipids. CGMD simulations have also been used to assess dependence on side-chain ionisation state of interactions between lipids and two IMPs observed in mass spectrometry experiments. The simulations reveal similar trends to those seen in experiments. Finally, using multi-scale simulations, and through the development of a novel method for altering membrane composition in MD simulations, lipid-specific scramblase activity was shown for a novel structure of the TMEM16K scramblase IMP.

Publications

The following publications will discuss work presented in this thesis.

MemProtMD: an international resource for membrane protein structures. **Thomas D. Newport**, Mark S.P. Sansom and Phillip J. Stansfeld. In preparation for eLife. (MemProtMD database, MemProtMD web server and MemProtMetrics described in Chapters 2 and 3).

Lipid binding attenuates channel closure of the outer membrane protein OmpF. Ildir Liko, Matteo T. Degiacomi, Sejeong Lee, **Thomas D. Newport**, Joseph Gault, Eamonn Reading, Jonathan T.S. Hopper, Nicholas G. Housden, Paul White, Matthew Colledge, Altin Sula, Bonnie A. Wallace, Colin Kleanthous, Phillip Stansfeld, Hagan Bayley, Justin L.P. Benesch, Timothy M. Allison and Carol V. Robinson. In preparation for PNAS. Chapter (Coarse-grained simulations of AmtB and OmpF described in Chapter 4).

Structure of the TMEM16K Ca^{2+} activated lipid scramblase (working title). Simon Bushell, Ashley Pike, **Thomas D. Newport**, Phillip J. Stansfeld, Liz Carpenter and others. (Multi-scale simulations of TMEM16K described in Chapter 4).

Statement of Originality

Except where indicated otherwise, all text, images and illustrations printed in this thesis are my own original work. Some images and illustrations have been submitted to academic journals as parts of resulting publications. No text has been submitted elsewhere as part of any other publication or examination.

Work described in this thesis was completed by me between October 2014 and January 2018, except where otherwise stated. The MemProtMD database builds on earlier work undertaken by Dr Phillip Stansfeld, who built the MemProtMD simulation setup pipeline, which included some limited analysis tools and web visualisations. Work described in Chapter 2 and Chapter 4 uses the MemProtMD simulation setup pipeline. The original MemProtMD analysis and visualisation tools are not used in whole or in part for any work described in this thesis.

Contents

1	Introduction	1
1.1	Interactions of lipids with integral membrane proteins	3
1.1.1	Capabilities and limitations of methods for characterisation of lipid interactions with integral membrane proteins	6
1.1.2	Evidence for lipid interactions with integral membrane proteins from structural and simulation studies	8
1.1.3	Distribution of amino acids in lipid bilayers	17
1.2	Computational methods for the study of integral membrane proteins .	19
1.2.1	Biomolecular simulation	20
1.2.2	Prediction of integral membrane protein structure	22
1.2.3	Online tools and web databases	25
1.3	Thesis overview	28
2	MemProtMD: A resource for membrane protein structures	29
2.1	Abstract	29
2.2	Introduction	30
2.3	Materials and Methods	32
2.3.1	System setup and simulation	32
2.3.2	Characterising bilayer properties	36
2.3.3	Efficient calculation of lipid and solvent contacts with protein residues	38
2.3.4	Representations of membrane protein topology	39
2.3.5	Static files	42
2.3.6	Acquisition and storage of metadata	44
2.3.7	Automated classification of IMPs based on simulation results .	46
2.3.8	Ensemble analyses	47
2.3.9	Data storage	49
2.3.10	Application structure	52

2.3.11	Web application	54
2.4	Results	54
2.4.1	The MemProtMD web application	54
2.4.2	Ensemble and collection analyses	67
2.4.3	The MemProtMD API	75
2.4.4	The Story of Mr Membrane Protein	78
2.5	Conclusions	78
3	Automated identification of protein and lipid membrane components in PDB depositions	82
3.1	Abstract	82
3.2	Introduction	83
3.3	MemProtMetrics: Knowledge-based identification of IMP structures in PDB depositions	85
3.3.1	Materials and Methods	85
3.3.2	Results	94
3.3.3	Estimation of leaflet surfaces for IMP structures	96
3.3.4	Benchmarking accuracy of MemProtMetrics, OPM and MEMEM-BED by reference to MemProtMD simulation results	98
3.3.5	Cross-validation and fit bias	103
3.4	Fragment-based identification of lipid bilayer components in PDB depositions	107
3.4.1	Materials and Methods	107
3.4.2	Results	115
3.5	Conclusions	118
4	Simulations of IMP structures in mixed-lipid bilayers	122
4.1	Abstract	122
4.2	Introduction	123
4.3	Materials and methods	126
4.3.1	Coarse-grained molecular dynamics simulations of AmtB and OmpF	126
4.3.2	Alchex: automated alteration of membrane composition	129
4.3.3	Multi-scale molecular dynamics simulations of a novel TMEM16K scramblase structure	134
4.4	Results	135
4.4.1	Ion mode dependence of lipid binding to AmtB and OmpF	135

4.4.2	Alchex: Automated exchange of simulated molecules	136
4.4.3	Multi-scale molecular dynamics simulations of a novel TMEM16K scramblase structure	138
4.5	Conclusions	144
4.5.1	Ion-mode dependence of lipid binding to AmtB and OmpF . .	144
4.5.2	Alchex: automated alteration of membrane composition	144
4.5.3	Multi-scale molecular dynamics simulations of a novel TMEM16K scramblase structure	146
5	Conclusions	148
5.1	Results	148
5.2	Future prospects	149
5.2.1	Better use of PDB data	149
5.2.2	Integration of specialised analyses	151
5.2.3	Extension of MemProtMD to mixed lipid systems	151
5.2.4	Automated setup of large systems	153
5.2.5	Extension of MemProtMD to lipoproteins and peripheral mem- brane proteins	153
5.2.6	Improving transparency and repeatability of MemProtMD . .	153
5.2.7	MD simulation of IMPs in the ionised gas phase	155
5.2.8	Free energy landscape of TMEM16 flip-flops	155
5.3	Closing remarks	157
	References	159

List of Figures

1.1	Anatomy of common membrane lipids	4
1.2	Experimentally determined G-Protein Coupled Receptors (GPCRs) structures	10
1.3	Experimentally determined Receptor Tyrosine Kinases (RTKs) structures	12
1.4	Experimentally determined cys-loop receptor structures	13
1.5	Experimentally determined potassium inward rectifier structures . . .	14
1.6	Experimentally determined mechanosensitive ion channel structures .	15
1.7	Experimentally determined flippase and scramblase structures	16
1.8	Amino acids in a membrane by TAB	18
1.9	CG and atomistic representations of a novel TMEM16K structure . .	21
1.10	WebGL visualisations using NGL Viewer, PV and LiteMol	27
2.1	MemProtMD data processing overview	33
2.2	Validation of MemProtMD data	35
2.3	Pre-rendered images produced by MemProtMD	43
2.4	MemProtMD infrastructure overview	51
2.5	Home page of the MemProtMD web application	55
2.6	Examples of the Quick Search component in use	55
2.7	Example of a PDB reference viewed using the MemProtMD web application	57
2.8	Examples of 3D molecular visualisations using the MemProtMD web application	59
2.9	Examples of protein chain summaries	60
2.10	Examples of secondary structure topology visualisations	61
2.11	MemProtMD sequence viewer	62
2.12	A sub-chain selection across multiple connected visualisations	64
2.13	MemProtMD text search tool	65
2.14	MemProtMD author profile	66

2.15	Global usage of the MemProtMD web application	68
2.16	Growth in numbers of IMP structures since 1985	69
2.17	Experimental methods used to determine IMP structures	70
2.18	Histogram showing resulting membrane thickness from simulations of α -helical and β -barrel IMP structures	71
2.19	Per-residue contact scales derived from MemProtMD	73
2.20	Histograms showing occurrence of amino acid types along the mem- brane normal	74
2.21	Multi-sequence alignment of all GPCR structures in the MemProtMD database	76
2.22	Multi-sequence alignment of all Gram-negative porin structures in the MemProtMD database	77
2.23	Still from Mr Membrane Protein animation	78
3.1	Example of α -shape generation from a two-dimensional point cloud .	87
3.2	α -shape generation from a protein structure	88
3.3	Smoothing of a surface metric across a protein mesh	90
3.4	Illustration of protein pore and cavity inner surface identification . .	91
3.5	Illustration of objective function calculation	93
3.6	MemProtMetrics objective function for IMP and non-IMP structures	95
3.7	Examples of leaflet surface estimation for IMP structures	97
3.8	Scatter plot showing angular errors from MemProtMetrics and MEMEM- BED	99
3.9	Scatter plot showing angular errors from MemProtMetrics and OPM	100
3.10	Cumulative error frequency comparison of MemProtMetrics and MEMEM- BED	101
3.11	Cumulative error frequency comparison of MemProtMetrics and OPM	102
3.12	Cumulative error frequency for training and test sets	104
3.13	MemProtMetrics and MEMEMBED error by SCOP fold	105
3.14	MemProtMetrics error by number of proteins	106
3.15	Classification of membrane ligands based on a steroid nucleus	109
3.16	Minimum spanning tree construction for a cardiolipin molecule	109
3.17	Examples of common connectors for branched fatty acid-type mem- brane lipids	111
3.18	Examples of common head-groups for branched fatty acid-type mem- brane lipids	111

3.19	Branched fatty acid-optimised model applied to three lipid molecules	112
3.20	Classifications applied to molecules by the fatty acid-optimised model	113
3.21	Examples of membrane lipids identified in PDB structures	116
3.22	Structures of cholesterol and phosphatidylcholine lipids annotated using contacts found in PDB depositions	119
4.1	Protonation of the MARTINI forcefield	128
4.2	Illustration of the Alchex molecule replacement method	131
4.3	Impact of Alchex lipid replacement on bilayer thickness	133
4.4	Coarse-grained representations of AmtB and OmpF in lipid bilayers .	136
4.5	Effects of protonation on lipid binding to AmtB and OmpF	137
4.6	Carbon- α Root-mean-square deviation (RMSD) of a TMEM16K structure during atomistic simulations	139
4.7	Lipid contacts, secondary structure displacement from bilayer centre and Root-mean-square fluctuation (RMSF) along the TMEM16K amino acid sequence	140
4.8	Surface and cartoon representations of TMEM16K coloured according to protein-lipid contacts	141
4.9	Flip-flopping of lipids between leaflets facilitated by a simulated TMEM16K structure	142
4.10	Calcium ions present in the TMEM16K crystal structure	143
4.11	Contacts between bound calcium ions and amino acid side-chains . .	145
5.1	The extended MemProtMD pipeline	150
5.2	Annular thickness of simulated pure DPPC and mixed simulations . .	152
5.3	Automated setup of large systems using MemProtMD	154
5.4	Repeatable, transparent research through continuous integration and archiving	156
5.5	Determining the energy landscape of a TMEM16 lipid flip-flop	157

List of Tables

2.1	Tags assigned by MemProtMD to MARTINI beads	39
2.2	Groups of beads used to calculate MemProtMD contact data	40
2.3	Static files available from MemProtMD	42
2.4	Fields used to perform text searches on the MemProtMD database . .	52
3.1	Protein amino acid groupings derived from MemProtMD	115
3.2	Lipid types found in PDB Ligand Expo and MemProtMD structures	117
4.1	pK_a of standard amino acids	127

Acronyms and Abbreviations

ANN Artificial Neural Network

API Application Programming Interface

CCM Cholesterol Consensus Motif

CD Continuous Delivery

CDP-DAG CDP-diacylglycerol

CDS CDP-diacylglycerol synthase

CG Coarse-Grained

CGSA Coarse-Grained Self-Assembly

CI Continuous Integration

COBYLA Constrained Optimization by Linear Approximation

CRAC Cholesterol Recognition/interaction Amino acid Consensus

Cryo-EM Cryo-Electron Microscopy

DAG Diacylglycerol

DLPG 1,2-Dilauroyl-sn-glycero-3-phosphorylcholine

DLPG 1,2-Dilauroyl-sn-glycero-3-phosphorylglycerol

DOI Digital Object Identifier

DPG Diphosphatidylglycerol

DPPC dipalmitoylphosphatidylcholine

DXPC 1,2-Dilignoceroyl-sn-glycero-3-phosphorylcholine

EGFR Epidermal Growth Factor Receptor

EPR Electron Paramagnetic Resonance

ER Endoplasmic Reticulum

ESI Electrospray Ionization

GABA_AR γ -aminobutyric acid receptor

GO Gene Ontology

GPCR G-Protein Coupled Receptor

GUI Graphical User Interface

HMM Hidden Markov Model

HTML HyperText Markup Language

HTTP HyperText Type Protocol

IMP Integral Membrane Protein

JPEG Joint Photographic Experts Group

JSON JavaScript Object Notation

Kir Potassium (K⁺) Inward Rectifier

LCP Lipidic Cubic Phase

LPS Lipopolysaccharide

MALDI Matrix Assisted Laser Desorption Ionization

MCS Maximum Common Subgraph

MD Molecular Dynamics

mmCIF Macromolecular Crystallographic Information File

MMTF Macromolecular Transmission Format

mpstruc Membrane Proteins of Known Structure

MSA Multiple Sequence Alignment

nAChR Nicotinic Acetylcholine Receptor

NFS Network File System

NMR Nuclear Magnetic Resonance

OPM Orientations of Proteins in Membranes

PA Phosphatidic Acid

PC Phosphatidylcholine

PDB Protein DataBank

PDBe Protein DataBank in Europe

PDBj Protein DataBank in Japan

PDBTM Protein DataBank of Transmembrane Proteins

PE Phosphatidylethanolamine

Pfam Protein Families

PG Phosphatidylglycerol

PI Phosphatidylinositol

PIP Phosphatidylinositol phosphate

PIP₂ Phosphatidylinositol 4,5-bisphosphate

PIP₃ Phosphatidylinositol (3,4,5)-trisphosphate

POPC 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine

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

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

POPS 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoserine

PPA Pyrophosphatidic Acid

PS Phosphatidylserine

RAID Redundant Array of Independent Disks

RCSB Research Collaboratory for Structural Bioinformatics

RDF Radial Distribution Function

REST Representational State Transfer

RMSD Root-mean-square deviation

RMSF Root-mean-square fluctuation

RTK Receptor Tyrosine Kinase

SVM Support Vector Machine

TCDB Transporter Classification DataBase

UniProt Universal Protein Resource

URI Uniform Resource Identifier

VMD Visual Molecular Dynamics

XML Extensible Markup Language

YAML YAML Ain't Markup Language

Chapter 1

Introduction

The boundaries of cells, organelles, vesicles and viruses are, with few exceptions, defined by a 3-4 nm thick membrane composed primarily of lipid and protein molecules, and held together by the intense hydrophobicity of the membrane core. Proteins with domains natively inhabiting this hydrophobic core, termed Integral Membrane Proteins (IMPs), are uniquely important to a great many biological processes, often acting as gatekeepers for signals or chemicals passing through the membrane. Indeed, IMPs account for as much as 30% of sequenced genomes [1], and are estimated to account for over half of druggable targets [2].

In spite of this, insights into the structure and function of IMPs have long remained elusive. To an IMP, the highly specific and anisotropic bilayer structure formed by lipid head-groups, acyl tails, water and ions in and around a biological membrane acts as a solvent. Interactions with this environment, and, in some cases, highly specific lipid species, are required to maintain structural integrity and function of a great many IMPs, rendering their study inaccessible to the majority of biophysical assays and established methods for protein structure determination.

By the time Johann Deisenhofer and colleagues determined the first IMP structure in 1985 [3], 27 years of soluble protein structure determination had already produced over 200 soluble protein structures. Since then, however, sustained advances in methods available for IMP structure determination such as the development of the Lipidic Cubic Phase (LCP) for X-ray crystallography, and, more recently, Nuclear Magnetic Resonance (NMR) spectroscopy and Cryo-Electron Microscopy (Cryo-EM), have resulted in the determination of more than 3300 high-resolution IMP structures [by combination of 4, 5, see Chapter 3], although this still amounts to fewer than 3% of all protein structures deposited in the Protein DataBank (PDB).

In parallel with developments to structure determination methods, many insights into IMP structure and function have been facilitated by development of a wealth of

methods and resources for *in silico* studies of IMP using bioinformatics approaches, modelling and simulation [6]. The use of biomolecular simulation methods such as Molecular Dynamics (MD) has revealed a complex and intimate relationship between IMPs and the membrane environment, from stabilisation and regulation by interactions with specific chemical groups of other membrane lipid and protein components to responses to bulk bilayer properties.

Although methods for IMP structure determination are yielding an ever-increasing number of high-resolution structures, many IMP structures remain elusive. With a better understanding of IMP-lipid interactions, as well as the general principles governing protein structure in lipid bilayers, development of accurate homology modelling methods for IMPs such as MEDELLER [7] has expanded the range of proteins accessible to computational studies. The pre-existence of a high-resolution structure is no longer a pre-requisite for biomolecular simulation, especially where structures of closely related proteins are available [8, 9, e.g.].

A major revelation of early IMP structure determination was that, in spite of the great diversity of IMP function and extramembrane structure, the transmembrane domain of an IMP follows one of two basic structural paradigms. IMP transmembrane domains are immediately recognisable, consisting either of one or more α -helices, or a β -barrel structure, composed of many spiralling β -strands. The majority of IMPs have an α -helical bundle topology, with β -barrel IMPs confined to the outer membranes of bacterial cells or endosymbiont-derived organelles such as mitochondria and chloroplasts. It seems likely that these conserved structural elements are due to both a shared evolutionary history [10], and the constraints imposed by the unique hydropathy landscape of the lipid bilayer [11].

Within these structural constraints, however, IMPs carry out a vast range of functions critical to the overwhelming majority of metabolic and signalling pathways. IMPs are responsible for the active and passive transport of a vast number of small soluble molecules, such as water, ions, saccharides and even gasses such as ammonia. As receptors, IMPs are involved in signal transduction cascades, passing on and amplifying signalling between or within cells, recognising other cells, or detecting the identity and location of endogenous or exogenous substances. Several IMPs also act as enzymes, playing critical roles in photosynthesis and both aerobic and anaerobic respiration. IMPs also include key components of the immune system, flagellar motors, toxins, cytoskeletal anchors and intercellular junctions. Several types of ion channels, receptors and transporters are discussed below with relation to their interactions with membrane lipids.

1.1 Interactions of lipids with integral membrane proteins

Key to the understanding of IMP structure and function is an understanding of their interactions with lipid molecules of the membranes they inhabit. Structures of over 40,000 unique lipid species are currently represented in the LIPID MAPS database [12] (<http://www.lipidmaps.org>), with around 1000 unique lipid species currently thought to be represented in the lipidome of a single cell [13]. A great many of these lipids are used to form bilayers, with various compositions and structures, which surround cells, cell compartments and vesicles. Lipidomics studies have revealed a great diversity of membrane lipid compositions, differing between cell compartments and leaflets (reviewed by van Meers [14] and online via Membrane Protein Composition Atlas <http://opm.phar.umich.edu/atlas.php>). Leaflet or microdomain localisation of several lipid species is known to play key signalling roles, with many lipids able to recruit or activate specific protein effectors [15]. The misregulation or exploitation of membrane lipidomes has extensively documented links to both infectious [e.g. 16, 17, 18] and non-infectious diseases [e.g. 19, 20, 21], therefore a better understanding of lipid signalling and interactions with proteins provides a great opportunity for the development of a vast array of new or improved therapeutics targeted to IMPs or membrane lipids [22].

Membrane lipids interacting with IMPs may be categorised according to the dynamics of these interactions, and time spent in contact with specific sites of the protein (residence time) [23]. **Bulk** lipids diffuse freely within the membrane, rarely interacting with amino-acid side-chains, and modulating protein function primarily through bilayer structural properties such as tension and thickness [e.g. 24, 25]. Significantly less mobile than bulk lipids, **Annular** lipids are found near protein transmembrane region, forming one or more [26] shells, similar to hydration shells formed around extramembrane domains by water and ions. Annular lipids may occupy specific binding sites on the protein surface, although these are typically dynamic and short lived. Frequent exchange occurs between bulk and annular lipids, with experimentally determined transition times from bulk to annular phase of around 100 ns under physiological conditions [27]. **Non-annular** lipids exchange with annular and bulk lipids at vastly slower rates, and form long-lived, high-affinity interactions at specific sites, often buried within α -helices or between subunits of an IMP.

A growing number of IMPs are now known to be structurally or functionally modulated by the composition or mechanical properties of their native lipid bilayer. In

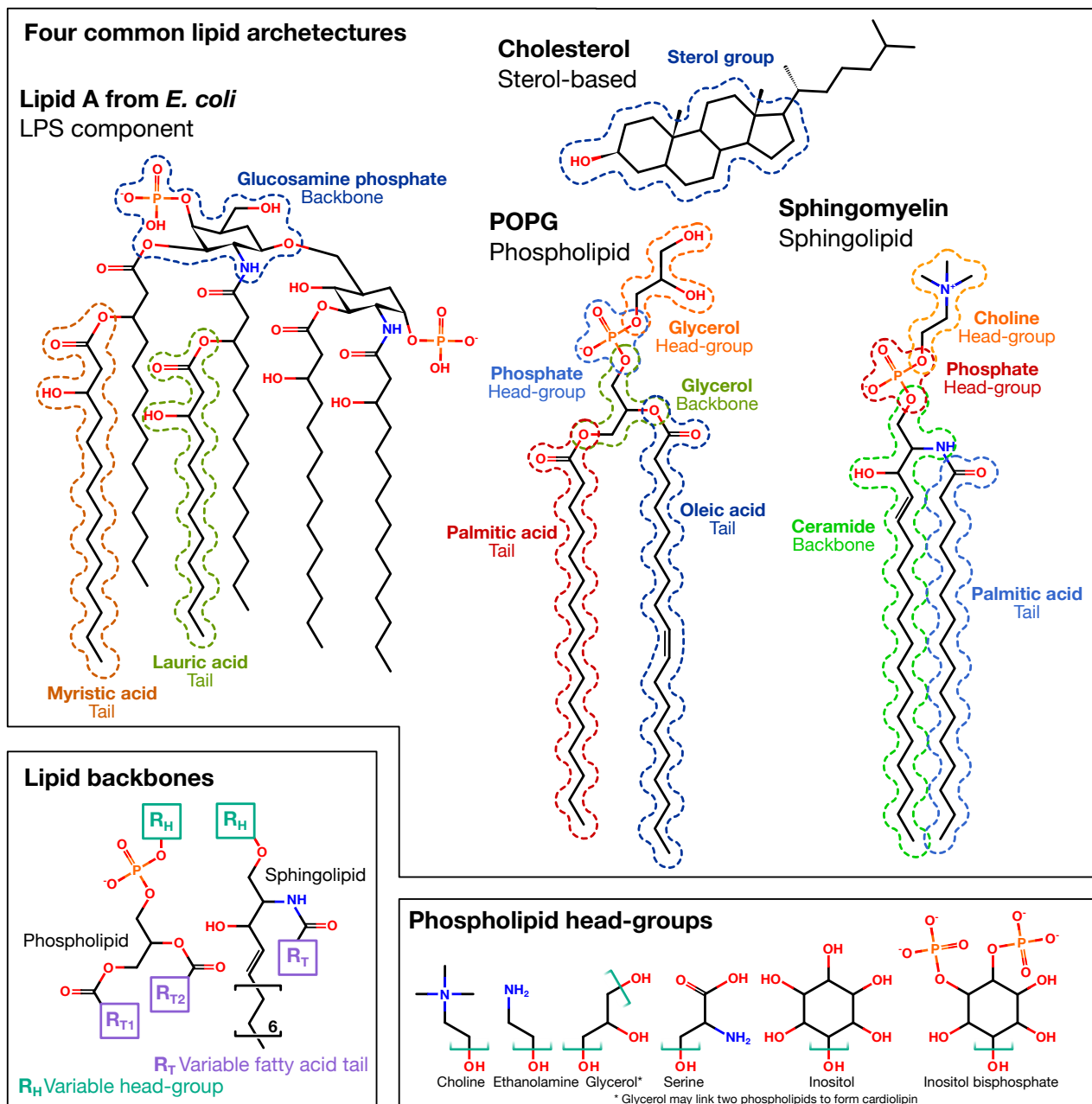


Figure 1.1: Anatomy of common membrane lipids

Top - Chemical structures of four membrane lipids (Lipid A from *E. coli*, cholesterol, 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoglycerol (POPG) and sphingomyelin) are shown with several head, tail and backbone chemical groups labelled. Bottom - phospholipid and sphingolipid backbones are shown (left), as well as several common phospholipid head-groups (right).

many cases, the presence or absence of a particular lipid species may modulate aspects of protein function. For example, many ion channels, such as Potassium (K^+) Inward Rectifier (Kir) channels [28], are known to be activated by Phosphatidylinositol 4,5-bisphosphate (PIP_2) [29], whilst the function of many G-Protein Coupled Receptors (GPCRs) is known to be modulated by cholesterol [30]. In other cases, modulation of protein function depends more on local mechanical properties of the lipid bilayer. Several proteins ion channels are able to open or close in response to changes in membrane tension, often as a protective measure during osmotic shock; these include the bacterial channel MscS [31], and potassium channels TREK-1 [32] and TREK-2 [33]. Local bilayer thickness has also been shown to affect protein function, as with Ca^{2+} ATPase, for which activation has shown to depend on lipid tail length [25], although it is not known whether this dependence is physiologically relevant.

In some cases, lipid molecules may act as native cofactors or substrates in a reaction catalysed by an IMP, as is the case for IMPs which form part of phospholipid biosynthetic pathways. Phosphatidate cytidyltransferase, also known as CDP-diacylglycerol synthase (CDS), is an IMP which catalyses the synthesis of CDP-diacylglycerol (CDP-DAG), an important precursor of Phosphatidylinositol (PI), Phosphatidylglycerol (PG) and Phosphatidylcholine (PC) lipids, from Phosphatidic Acid (PA) lipids [34], and is necessary for the transduction of PI-mediated signalling cascades [35]. In common with many enzymes, CDS is regulated, in part, by feedback inhibition, with binding of CDP-DAG shown to inhibit enzymatic activity [36]. Many other enzymes forming part of phospholipid biosynthetic pathways are IMPs, including Phosphatidylinositol phosphate (PIP)-synthase, which catalyses the synthesis of PIP from CDP-DAG and myo-inositol [37], and several members of the DAG-kinase family of enzymes, which phosphorylate Diacylglycerol (DAG) to produce PA [38].

Within membranes of cells or cell compartments, the bilayer composition exposed to each IMP is far from homogeneous. It has long been known that the composition of many membrane types varies between the two leaflets of a lipid bilayer [39]. IMPs modulate membrane asymmetry by active (in the case of lipid flippases and floppases) or passive (in the case of lipid scramblases) translocation of lipids between leaflets of the lipid bilayer (known as lipid flip-flopping). Members of the TMEM16 protein family are associated with the breakdown of bilayer asymmetry and Phosphatidylserine (PS) signalling by exposure of PS lipids in the extracellular leaflet [40]. Specificity for PS appears to vary within the TMEM16 family [41].

Membrane proteins and lipids can be found organised into microdomains [42]; regions of cell membrane between 30 nm and 750 nm in diameter [43] with distinct

protein and lipid compositions. These include planar lipid rafts and invaginations of the lipid bilayer known as caveolae [44]. In most documented cases, formation of a microdomains is associated with enrichment of cholesterol and sphingolipids relative to typical bilayer composition [45], however evidence suggests the existence of microdomains formed of other lipid types, such as PS and PC [46]. Microdomain enrichment of various lipids, including PA [47], PIP₂ [48] and PS [49], plays a vital role in many signalling pathways by recruitment or activation of protein effectors.

1.1.1 Capabilities and limitations of methods for characterisation of lipid interactions with integral membrane proteins

In most cases, the structure and function of an IMP are dependent on the presence of a lipid bilayer, complicating a number of experimental approaches which cause disruption this structure [50]. Characterisation of IMP-lipid interactions remains challenging, however, through the combination of various experimental and computational approaches summarised below, it has been possible to characterise lipid interactions of several IMPs in some considerable detail [6].

The presence of co-purified lipids in experimentally determined high-resolution molecular IMP structures has revealed many higher-affinity binding modes of lipids to IMPs. X-ray crystallography has provided the majority of high-resolution molecular structures of IMPs. The development of techniques for *in meso* crystallisation, such as LCP crystallization [51], has allowed for protein crystals to form in the presence of membrane lipids. As a result, many hundreds of IMP structures have been resolved in the presence of lipids [52]. X-ray crystallographic studies of IMPs are limited, however, by the challenging task of both purifying and crystallising sufficient amounts of IMPs.

Recent development of two electron microscopy techniques for IMPs structure determination, single-particle Cryo-EM [53] and 2D-electron crystallography [54], which require the purification of substantially lower amounts of protein than X-ray crystallography, have resulted in several high-resolution molecular structures of IMPs with co-purified lipids. 2D-electron crystallography, which uses 2-dimensional crystal packing of IMPs reconstituted into a lipid bilayer, can be used to resolve lower-affinity lipid binding, such as the annular lipids of an aquaporin [55]. Where structures are not amenable to crystallographic methods, single-particle Cryo-EM techniques can be used to determine high-resolution structures for IMPs, often with higher-affinity

bound lipids originating from the expression host [e.g. 56], without requiring large amounts of protein or formation of crystals.

Whilst X-ray crystallography and electron microscopy can reveal sites of lipids binding to IMPs at high resolution, these methods alone cannot be used to examine the dynamics of these interactions. NMR spectroscopy, however, can be used to reveal both structure and dynamics of protein structures and their binding to molecules including membrane lipids. Solution-NMR is limited in application to IMPs by the need to solubilise proteins, requiring selection of a small protein and the use of detergents [57]. With the development of solid-state NMR, however, NMR spectra can be obtained from IMPs reconstituted in lipid bilayers, allowing for the study of dynamics of lipid interactions with larger IMP structures under more physiological conditions [Reviewed in 58].

In addition to structure determination methods such as X-ray crystallography, electron microscopy and NMR, a range of other biophysical assays may be used to characterise interactions between IMPs and membrane lipids [Reviewed in 59]. Amongst these, native mass spectrometry has recently emerged as an increasingly powerful method, allowing not only the quantification of lipid binding [60], but its effects on protein oligomeric state [61] and protein stability [62] to be quantified. Little is understood, however, of the impact of the ionised, gas-phase environment of a mass spectrometry experiment on IMPs and their interaction with membrane lipids.

Working in combination with experimental methods, biomolecular simulations can examine dynamics of IMP-lipid interactions at a number of temporal and spatial scales, from detailed atomistic MD simulations to more abstract Coarse-Grained (CG) MD and even continuum representations [63]. By reconstituting an experimentally determined or computationally predicted IMP structure in an explicit lipid bilayer, MD simulations may capture events such as lipid binding and protein conformational change, giving mechanistic insights into the function or regulation of a protein. Many lipid binding events observed in MD simulations show good agreement with evidence from experimental studies. A more general discussion of the application of biomolecular simulation techniques to IMPs may be found in Section 1.2.1.

MD techniques are highly amenable to the development of high-throughput studies of large ensembles of proteins and lipids. Large numbers of MD simulations of IMPs in explicit lipid bilayers of various compositions may be set-up automatically using tools such as MemProtMD [5], HTMD [64] and the CHARMM-GUI [65]. This has facilitated simulations of ensembles of IMP ranging from an entire protein family [e.g.

66] to the over 3000 IMPs currently represented by simulations in the MemProtMD Database (see Chapter 2).

Experimental approaches for characterisation of IMP-lipid interactions have so far been limited to low-throughput studies of individual or small numbers of IMPs, due to the inherent difficulties of working with IMPs. Given the vast number of IMPs, and importance of their interactions with lipids, development of high-throughput, -omics-type approaches for *in-vitro* and *in-vivo* characterisation of IMP-lipid interactions is currently a high priority [59].

1.1.2 Evidence for lipid interactions with integral membrane proteins from structural and simulation studies

The following examples are extensive and intended to be representative of the variety of IMPs and their interactions with membrane lipids, with a particular focus on IMP families discussed in Chapters 2-4, but an exhaustive list would be beyond the scope of this introduction. More examples of IMPs-lipid interactions may be found reviewed from an MD perspective by Chavent *et al.* [63], and Hedger and Sansom [6], and from a mass spectrometry perspective by Barrera *et al.* [60].

GPCRs

The GPCR family of receptors each contain a highly recognisable topology consisting of seven transmembrane α -helices, and are a major class of IMP receptor, accounting for up to 40% of prescription drug targets [67]. GPCRs are known to be functionally modulated by interactions with a variety of lipid species. Of these, the best characterised, and perhaps most important interaction, is that of cholesterol. Experimental studies have been shown key roles for cholesterol in stabilising GPCRs [68], promoting oligomerisation [69], and altering activity of native ligands [70]. From results of experimental, structural and MD studies, it is becoming increasingly apparent that cholesterol acts both indirectly, by altering bulk bilayer structure, and directly, by binding to specific sites on GPCRs.

In comparison to most other membrane lipids, cholesterol varies little in structure or conformation. It is therefore possible to generalise the properties of a cholesterol binding site on an IMPs and thereby identify likely sites in other protein structures. Cholesterol Recognition/interaction Amino acid Consensus (CRAC) [25] binding sites follow a linear pattern from C to N-termini consisting of an apolar leucine or valine residue, located between one and five residues from an aromatic tyrosine residue,

located in turn between one and five residues from a basic lysine or arginine residue. The CARC motif [71] is a variation on CRAC which runs from N to C-termini and allows for tyrosine or phenylalanine in the central position. The Cholesterol Consensus Motif (CCM) is a stricter and more complex variation on these motifs based on binding of cholesterol to Class A GPCRs [72], finding likely cholesterol binding sites in 44% of structures.

Determination of high-resolution GPCR structures with bound cholesterol, as well as unbiased MD simulations of GPCRs in the presence of cholesterol, has revealed the existence of several direct binding sites. All-atom MD simulations of the A_{2A} adenosine receptor were used to follow cholesterol molecules in a cholesterol/1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) bilayer over 800 ns [73], identifying three sites which formed interactions with cholesterol which, although comparatively weak, persisted for the length of the simulation. Subsequent publication of a high-resolution structure revealed three co-crystallised cholesterol molecules, of which one corresponded to a site seen in MD simulations [74].

A series of 80 μ s CG simulations of the Serotonin_{1A} receptor in a mixed POPC/cholesterol bilayer identified several sites in both leaflets which preferentially (although weakly) bound cholesterol [75], coinciding in one case with a CRAC and in another with a CCM. Similar simulations of the closely related Serotonin_{2A} receptor, however, showed an entirely different set of binding sites [76]; it is unclear whether this difference is due to differences in protein structures or limitations of the CG methodology.

The oligomeric state of GPCRs is known to modulate receptor function and activation [69], and studies of several GPCRs have shown oligomerisation to be modulated by cholesterol. GPCR oligomerisation appears to be driven by hydrophobic mismatch, with energetically favourable protein-protein contacts shielding helices with greater mismatch from the lipid bilayer [77]. A recent study used CG simulations including multiple β -adrenergic receptors to measure hydrophobic mismatch of two β -adrenergic receptors [78]. Addition of cholesterol was found to substantially alter hydrophobic mismatch of several transmembrane helices; in a pure POPC bilayer TM4 showed the greatest degree of mismatch, followed by TM1 and TM5, however with the addition of cholesterol, hydrophobic mismatch of TM1 increased, almost doubling in the case of the β_2 -adrenergic receptor, whilst hydrophobic mismatch of TM4 was reduced. A suggested cause was bilayer thickening due to increased order of phospholipid acyl tails induced by cholesterol. Various interfaces were formed in simulations using a pure POPC bilayer, including TM1-TM1, TM4-TM5, and TM5-TM5; in each case, protein-protein interactions were found to reduce hydrophobic mismatch.

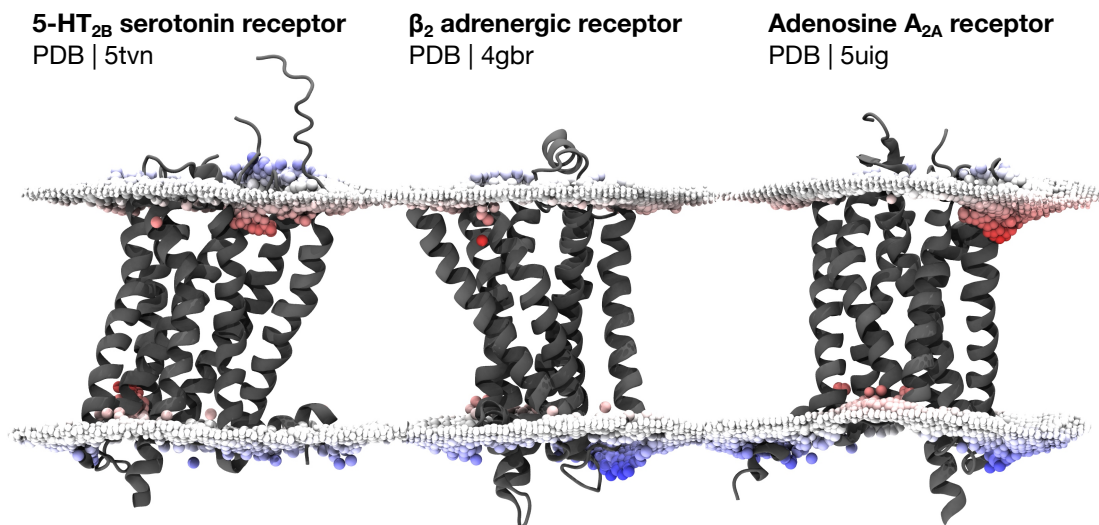


Figure 1.2: Experimentally determined GPCRs structures

Structures are shown in cartoon format with bilayer distortion data from the MemProtMD database shown as spheres, scaled by distance from the protein. Leaflet thinning and thickening is shown in red and blue respectively.

The cholesterol-dependence of the β_2 -adrenergic receptor was further studied using a series of CG simulations with two protein molecules and varied amounts of cholesterol [79]. In the absence of cholesterol, protein dimerisation occurred via the TM4-TM5 interface, however, when cholesterol was present, the TM4-TM5 interface was destabilised, and the protein dimer reformed via an interface between TM1 and TM2. Cholesterol was found to disrupt the TM4-TM5 interface by binding to a CCM on TM4, a site where cholesterol binding is seen in crystal structures. It would appear, therefore, that cholesterol affects oligomerisation of β -adrenergic both indirectly, through effects on hydrophobic mismatch, and directly, by competing with protein-protein interactions at the TM4 helix.

Cholesterol is far from the only lipid type with known interactions with GPCRs. As is the case with several GPCRs, function of the β_2 -adrenergic receptor is known to be modulated by anionic PG lipids, which promote ligand binding and activation of the receptor [80]. A series of all-atom MD simulations of the β_2 -adrenergic receptor in a POPC/POPG bilayer found that after around 1 μ s, a POPG lipid entered a pocket in the cytosolic side of the protein [81]. Interactions between the lipid head-group and protein were found to disrupt several long-range electrostatic interactions, destabilising the inactive form of the protein. This suggests a likely mechanism by which PG lipids may promote ligand binding and receptor activation.

Receptor tyrosine kinases

Receptor Tyrosine Kinases (RTKs) are another class of receptors known to be modulated by lipid interactions. The activation of RTKs such as Epidermal Growth Factor Receptor (EGFR) begins with ligand binding, which promotes dimerisation of two protein monomers. As a result of dimerisation, several tyrosine residues become phosphorylated, creating a docking site which binds proteins involved in downstream signalling [82]. Experimental studies of the human EGFR, have shown receptor activity to be downregulated by the presence of the GM3 ganglioside (a sphingolipid) in the outer leaflet [83], and upregulated by the presence of PIP₂ [84] in the inner leaflet. RTK monomers consist of a single pass transmembrane helix, attached by a juxtamembrane region at the inner leaflet to a kinase domain, and to a ligand-binding ectodomain in the outer leaflet. Structures have been determined for several individual parts of RTKs, including NMR structures of transmembrane helix dimers, however to date a high-resolution structure of a complete, intact RTK has not been determined.

Simulations of partial RTK structures have revealed some interesting interactions with membrane lipids. CG simulations of structures or homology models of the transmembrane and juxtamembrane region of all 58 human RTK monomers in bilayers containing PC, PS and PIP₂ found a conserved clustering of PIP₂ around the juxtamembrane region [66]. Further simulations of the transmembrane and juxtamembrane region of a human EGFR dimer found clustering around the protein of GM3 in the outer leaflet and PIP₂ in the inner leaflet, and examined the free energy landscape of binding of PIP₂, GM3, PS and PC lipids to the dimer, finding that free energy of PIP₂ binding to a site on the inner leaflet side of the dimer was substantially lower than for other lipids [85]. Alterations of the ionisation of PIP₂, and *in-silico* mutagenesis of charged residues of the EGFR extramembrane region were found to increase free energy associated with PIP₂ binding, suggesting a substantial role for electrostatic forces in this binding.

Cys-loop receptors

Cys-loop receptors are pentameric ligand-gated ion channels which open upon binding to a variety of neurotransmitters. Amongst these are the γ -aminobutyric acid receptor (GABA_AR), which is sensitive to membrane cholesterol content [86], and the Nicotinic Acetylcholine Receptor (nAChR), which is sensitive to membrane cholesterol content, thickness and the presence of anionic lipids [87]. An all-atom MD simulation

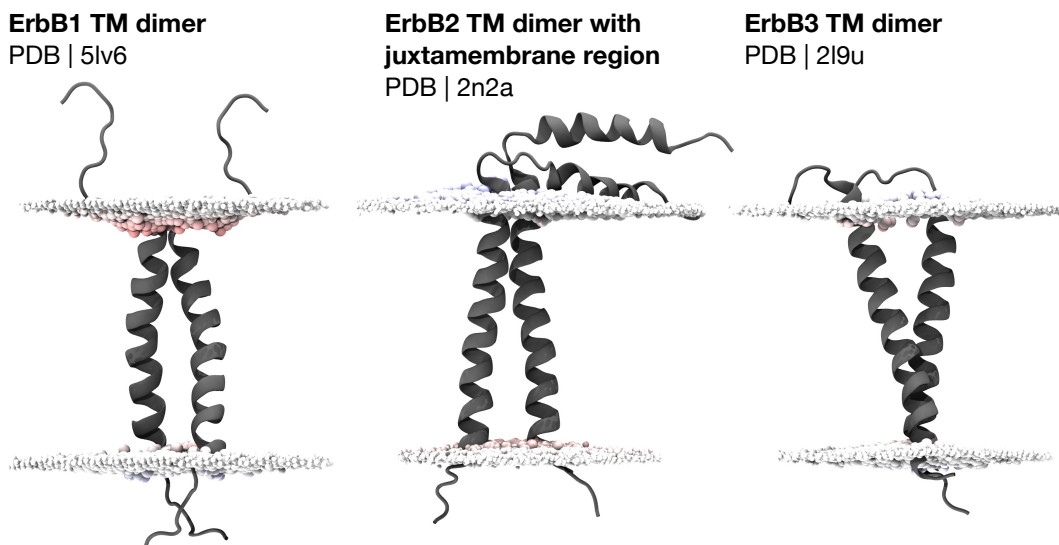


Figure 1.3: Experimentally determined RTKs structures

Structures are shown in cartoon format with bilayer distortion data from the Mem-ProtMD database shown as spheres, scaled by distance from the protein. Leaflet thinning and thickening is shown in red and blue respectively.

approach has used a Cryo-EM structure of nAChR to reveal a possible mechanism for cholesterol sensitivity [88]. In the absence of cholesterol, water entered several gaps between sparsely packed transmembrane helices, resulting in the collapse of the channel. Docking identified 15 potential binding sites for cholesterol between these helices, which, when occupied by cholesterol in subsequent simulations, resulted in expulsion of water and maintenance of channel structure. It seems likely therefore that binding of cholesterol is vital to the structural integrity of nAChR.

Two computational studies have provided insights into the mechanism of GABA_AR cholesterol sensitivity, also using a combined docking and MD approach, based on homology models of GABA_AR structures. In the first case [9], docking was used to find a likely cholesterol binding site on the GABA_AR monomer, and thus five sites on the intact pentamer. Over 200 ns of all-atom simulation, docked cholesterol changed in orientation, either remaining *in situ*, or in some cases exiting the site and rebinding in a different orientation, but always remaining closely associated with identified sites. Similar binding sites were found, although with cholesterol oriented differently, in a second set of all-atom MD simulations [89], in this case using binding of a hydrophobic drug to a related molecule to predict cholesterol binding. The differences between simulations of GABA_AR in the presence and absence of cholesterol were also shown to be less drastic than in the case of nAChR, with cholesterol binding inducing a

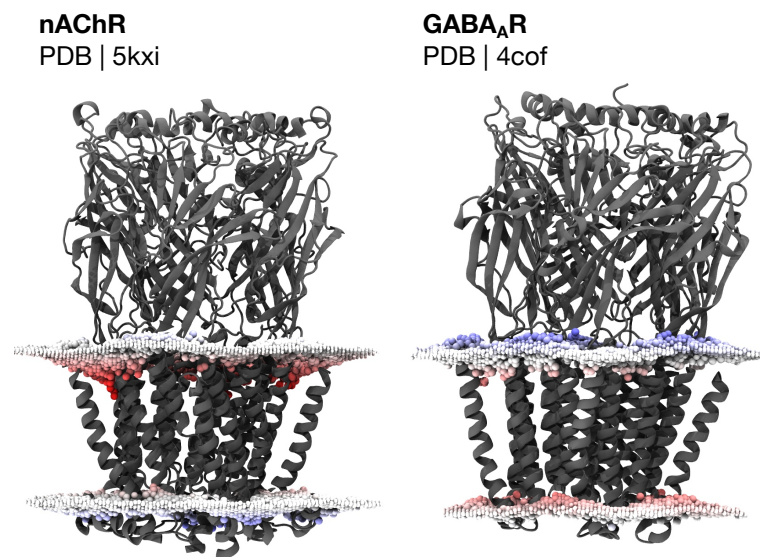


Figure 1.4: Experimentally determined cys-loop receptor structures

Structures are shown in cartoon format with bilayer distortion data from the MemProtMD database shown as spheres, scaled by distance from the protein. Leaflet thinning and thickening is shown in red and blue respectively.

small but significant opening of the channel pore.

Potassium (K^+) inward rectifier channels

Kir channels are found in both eukaryotic and bacterial membranes, allowing selective passage (with a bias towards import into the cell) of K^+ ions and thus regulating the resting membrane potential of the cell. In the case of the Kir2 family of Kir channels and many members of the Kir1 family, membranous PIP_2 is a well characterised ligand required for full activation of the channel [90]. The binding of PIP_2 to Kir channels was investigated using MD simulations, using two available structures for KirBac1.1 and a chimæra of Kir3.1 and KirBac1.3, as well as a homology model of Kir6.2 [8]. During 5 μs simulations in POPG bilayers, four PIP_2 molecules added to the bilayer bound to one of four binding sites located on the cytosolic side of each subunit interface, and remained bound for the duration of the simulation.

Subsequently, high-resolution X-ray crystal structures of Kir2.2 were published with both PIP_2 and Pyrophosphatidic Acid (PPA) lipids co-crystallised [91]. Both lipids were found to bind to similar sites to those observed in simulation, by interactions between acyl tails and a non-specific region of the Kir transmembrane domain. Lipids differed, however, in their interaction with the Kir cytoplasmic domain, with PIP_2 forming strong interactions whilst PPA failed to bind. Binding of PIP_2 induced

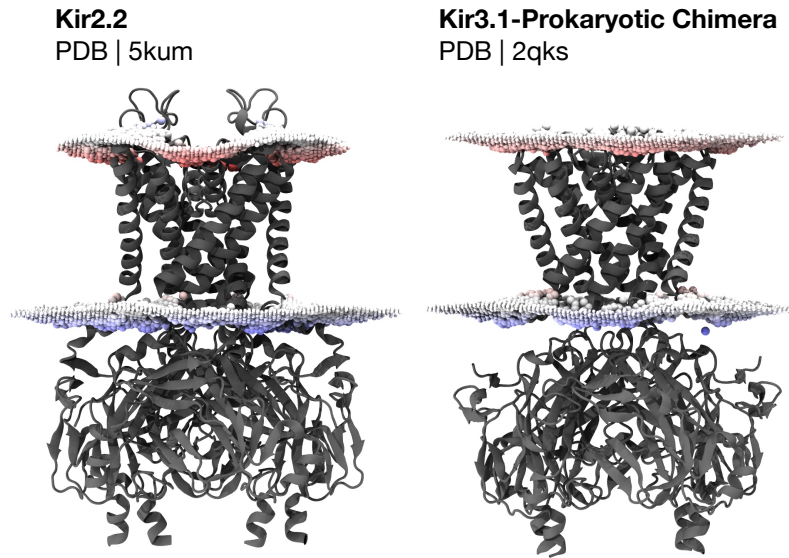


Figure 1.5: Experimentally determined potassium inward rectifier structures

Structures are shown in cartoon format with bilayer distortion data from the MemProtMD database shown as spheres, scaled by distance from the protein. Leaflet thinning and thickening is shown in red and blue respectively.

major conformational change, with a 0.6 nm movement of the cytoplasmic domain and opening of helices of the transmembrane domain, suggesting a likely mechanism for activation of the channel.

Mechanosensitive ion channels

Several ion channels are known to open in response to membrane tension, and a combination of structural and MD studies have revealed possible mechanisms for mechanosensitivity of the potassium ion channel TREK-2 and non-specific bacterial channel MscS. In the case of MscS, CG MD simulations found binding of lipid tails to pockets located near the cytosolic leaflet [92]. Where MscS was in an open conformation, fewer lipids bound in pockets. It was therefore suggested that increased lipid occupancy in pockets due to lower bilayer tension may result in channel closing. A recent all-atom MD simulation study of the TREK-2 mechanosensitive potassium channel showed that by introducing tension to the surrounding lipid bilayer, an experimentally determined TREK-2 structure in the closed state could be made to occupy a conformational state highly similar to an experimentally determined structure in the open state [93], showing a direct relationship between membrane tension and channel state.

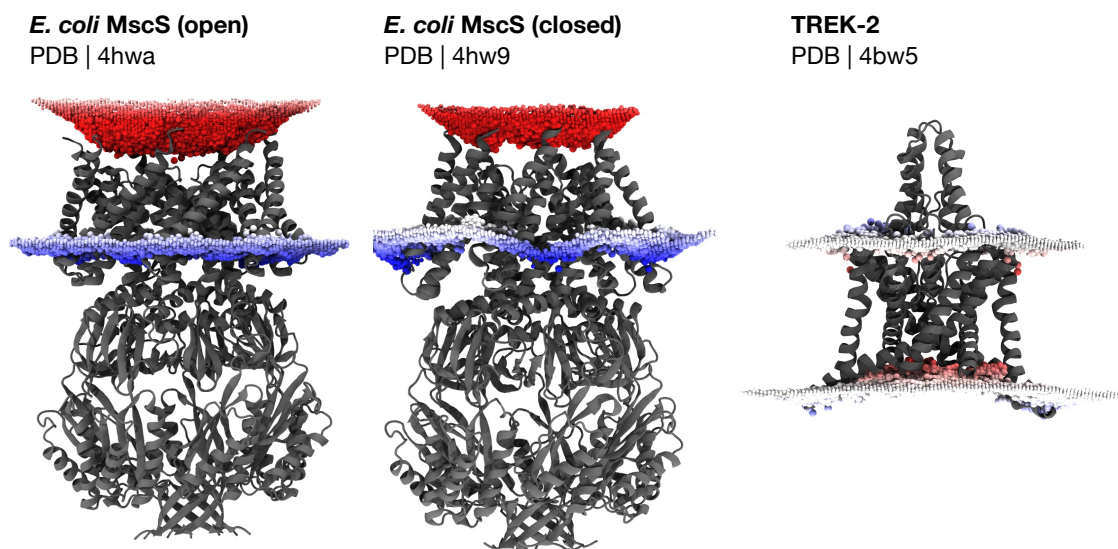


Figure 1.6: Experimentally determined mechanosensitive ion channel structures

Structures are shown in cartoon format with bilayer distortion data from the MemProtMD database shown as spheres, scaled by distance from the protein. Leaflet thinning and thickening is shown in red and blue respectively.

The TMEM16 (anoctamin) family

The TMEM16 family are dimeric IMPs of which many are involved in calcium-dependent passive lipid translocation, known as scramblase activity. It was initially thought that all TMEM16 proteins were calcium-dependent anion channels possessing eight transmembrane α -helices, hence the alternative name of anoctamins (**an**ion/**octa**helical), however it is now known that many TMEM16 proteins show no ion channel activity and instead function primarily as lipid scramblases [94]. Lipid scramblase activity varies in rate and specificity between TMEM16 proteins, with many TMEM16 proteins showing specificity for PS lipids [41]. In particular, TMEM16F plays a major role in triggering cell-cell signalling by the presentation of PS lipids on the membrane outer leaflet [95].

Determination of a high-resolution structure for nhTMEM16 [96], a TMEM16 protein from *Nectria haematococca*, by X-ray crystallography, has given some structural insight into the mechanism for TMEM16 lipid scramblase activity. The structure showed a groove formed by three transmembrane α -helices with a large number of polar residues, unlikely to make energetically favourable contacts with lipid tails, but potentially allowing head-groups to slide through while tails remained in the bilayer core. This model of lipid translocation is often referred to as the "credit card" model,

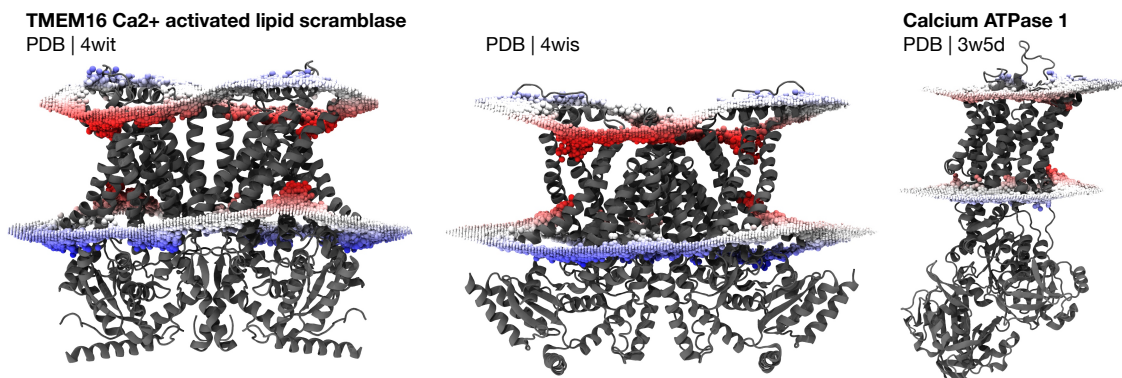


Figure 1.7: Experimentally determined flippase and scramblase structures
Structures are shown in cartoon format with bilayer distortion data from the MemProtMD database shown as spheres, scaled by distance from the protein. Leaflet thinning and thickening is shown in red and blue respectively.

by comparison to the sliding of the magnetic strip of a credit card through the groove of an older magnetic-type card payment terminal. This model has subsequently been validated by CG and all-atom MD simulations, which shown lipid flip-flops enabled by head-groups passing through this translocation groove [97].

It seems likely that this lipid translocation groove is also involved in ion transport, and two conflicting theories seek to explain ion channel activity of TMEM16 proteins [96]. One theory holds that, for TMEM16 proteins with ion channel activity, dimerisation occurs along a different interface, enclosing the lipid translocation groove in the centre of the new dimer, whilst the other holds that dimerisation does not change, and the groove is merely repurposed for ion channel activity. The determination by Cryo-EM of a low-resolution structure of the ion channel TMEM16A [94], a protein with little scramblase activity [41], would seem to support the latter theory, since dimerisation is the same as that seen in nhTMEM16. A further development of this theory is the possibility that the lipid head-groups and anions may pass through the same channel in some dual-function scramblase/anion channel TMEM16 proteins [98], which seems plausible given the calcium dependence of both processes.

The solute carrier family

The solute carrier family contains a great many transporter proteins which transport ions, gasses and small organic solutes across the bilayer. In addition to many lipid-regulated transporter proteins, the solute carrier family contains flippase and floppase transporters which are able to actively and selectively transport lipids between bilayer leaflets. Cardiolipin is linked to stabilisation of both the ADP/ATP carrier

and UraA H⁺-uracil symporter, with simulation studies finding strong binding sites for cardiolipin on membrane-exposed surfaces of both proteins, and X-ray crystallographic structures finding co-crystallised cardiolipin molecules in three structures of the ADP/ATP carrier [Reviewed by 6].

P4-ATPases are members of the solute carrier family which are flippases or floppases, and actively transport phospholipids between leaflets. Proteins vary in lipid specificity, with, for example, ATP10A involved in transport of PC lipids [99], whilst ATP8A2 is primarily involved in transport of PS lipids [100]. As with many proteins involved in maintaining or altering membrane lipid composition, and PS localisation in particular [49], mutations or deficiencies in P4-ATPases are associated with severe neuronal degradation [101]. MD simulation based on homology modelling of the PS flippase ATP8A2 revealed a similar "credit card" mechanism to that seen in nhTMEM16, with lipid headgroups moving along a groove between transmembrane helices [102]. Amino acid side-chains projecting into this groove are thought to actively propel head-groups through the groove upon dephosphorylation of ATP, as well as select for specific lipids. A recent study used directed evolution to engineer sphingomyelin flippase activity in a PC flippase (PC contains similar acyl tails and head-group, differing only in backbone). Using homology modelling of the engineered protein, a single amino acid residue at the cytosolic exit of the translocation groove was identified as conferring this specificity [103].

1.1.3 Distribution of amino acids in lipid bilayers

IMPs exist in a chemically diverse environment which creates a specific role for each amino acid, determined largely by the chemical properties of its side-chain. Some of these are summarised in Figure 1.8. This results in patterns of enrichment and depletion of each amino acid at different points within the bilayer. In some cases, the length and flexibility of an amino acid side-chain can allow optimal location of chemical groups through snorkelling [104], where side-chains of charged residues such as arginine and lysine located nearer the membrane core are able to reach out to the membrane-water interface [105], or anti-snorkelling, where side-chains of more hydrophobic residues such as phenylalanine are able to reach in towards the hydrophobic core [106].

The hydrophobic environment created in the membrane core favours hydrophobic amino acids, particularly phenylalanine, isoleucine, leucine and valine, which are found frequently in IMP structures at sites located deep in the lipid bilayer [107, 108]. Flanking the core of the lipid bilayer, and clustering around lipid carbonyl groups,

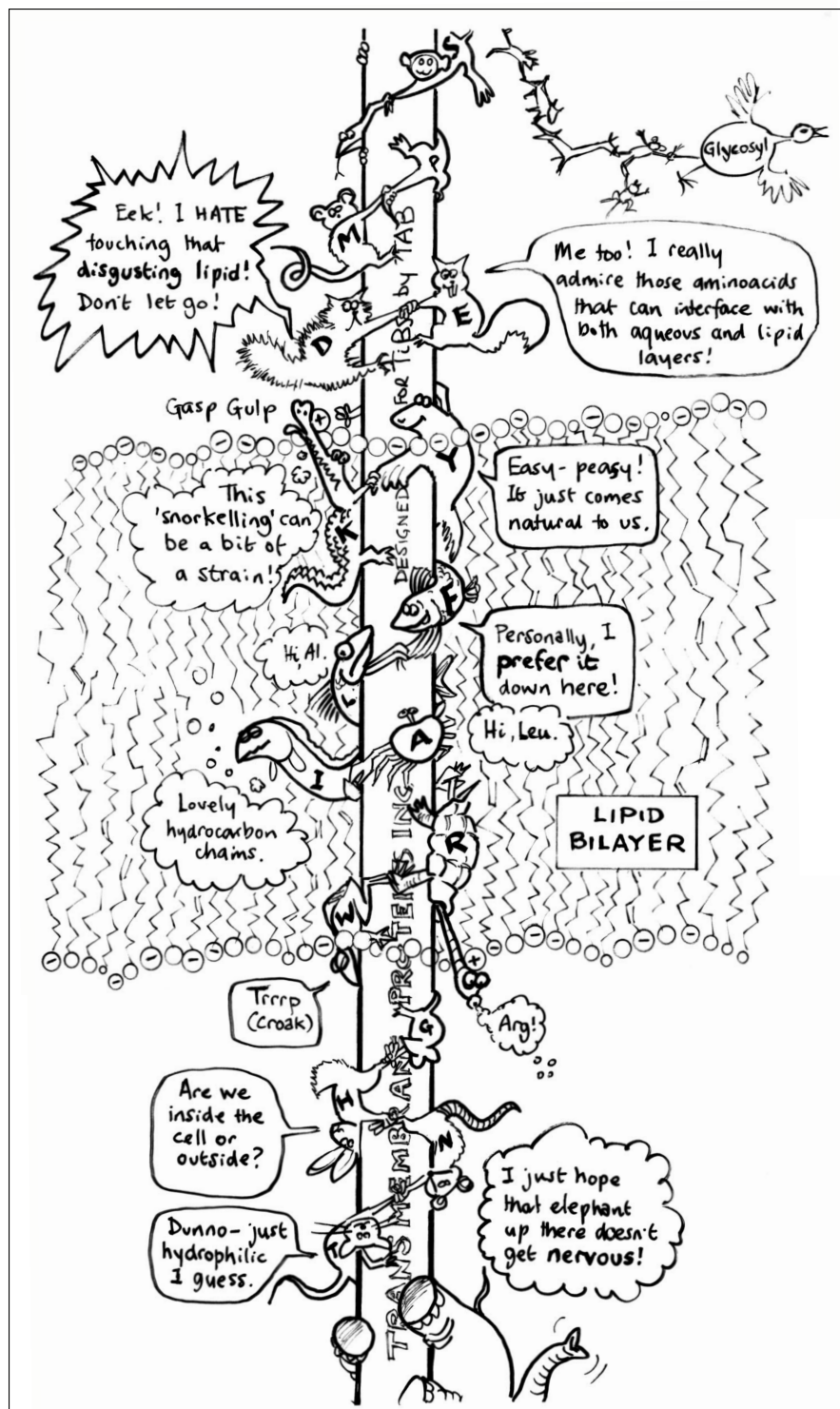


Figure 1.8: Amino acids in a membrane by TAB

Cartoon illustration of a single pass transmembrane α -helix, demonstrating many principles of amino acid membrane distribution, including snorkelling of arginine and lysine, interface localisation of tyrosine and tryptophan and the positive inside rule. Reprinted from Killian and Von Heijne 2000 [105].

Drawn for Trends in Biochemical Sciences by TAB. © 2000, Elsevier Science Ltd.

a ring of aromatic residues is typically found [109]. In α -helical IMP structures this region is particularly marked by enrichment of tyrosine and tryptophan, likely due to their amphipathic natures [109]. The distribution of phenylalanine residues appears to differ between α -helical and β -barrel membrane proteins, being clustered around the lipid carbonyl groups of β -barrel proteins but found throughout the hydrophobic core of α -helical proteins [110].

Charged residues interact more readily with water and are rarely found in the membrane core. Several studies of IMP sequences and structures have found that, whilst acidic and negatively charged residues are distributed evenly between cytosolic and extracellular locations, positively charged, basic residues lysine and arginine are found more commonly on the cytosolic side. This effect, known as the positive-inside rule [111], can be seen across many different species [112] and protein topologies [113]. The causes of this effect remain unclear; in most cases the resulting difference in net charge should be favoured by the transmembrane electrical potential, however examples exist of proteins following a positive-inside rule where this potential is reversed [114]. Alternatively, the positive-inside rule may originate with the N-terminal signal peptide, or may promote correct insertion of the protein by the Sec translocon [115]. It has also been suggested that the increased propensity at the protein N-terminus for conformations of arginine and lysine which permit snorkelling may contribute to this effect [104].

In the case of α -helical IMPs, the distribution of some amino acids may be explained by their specific effects on the structure and packing of α -helices. Studies of individual membrane spanning α -helices have found enrichment of asparagine and serine at the helix N-terminus [113, 109]. These residues are thought to stabilise the initiation of α -helices by forming hydrogen bonds with the protein backbone. Glycine is also found enriched in transmembrane segments of membrane-spanning α -helices [109], likely due to its high helix packing value [116].

1.2 Computational methods for the study of integral membrane proteins

The complexities of working with IMPs in an experimental context, coupled with development of powerful algorithms and high-performance computing has resulted in an increasing focus on computational methods for examining IMP structure and function. The use of web applications and web servers, meanwhile, has reduced technological

barriers to the use of many of these methods. A paucity of high-resolution IMP structures has driven an emphasis on methods for knowledge-based prediction of protein secondary, tertiary and quaternary structure. Using predicted or experimentally determined IMP structures, biomolecular simulations can then reveal interactions and dynamics of an IMP (or many IMPs) in a near-native setting over time.

1.2.1 Biomolecular simulation

In general terms, biomolecular simulation takes experimentally determined or predicted macromolecular structures and, by reference to various (often in the order of tens of thousands or higher) physical constraints, models the evolution of a system over time or an energy landscape. Methods applied to IMPs vary greatly in both applicable time and length scales.

Most biomolecular simulations of IMPs, however, use either atomistic or Coarse-Grained (CG) Molecular Dynamics (MD). MD simulations consider a system as a set of particles with coordinates, and a set of constraints and potential energy functions (the forcefield) used to calculate forces on particles. These forces comprise bonded interactions; potential energies due to the length, angle, dihedral and improper angles of a particular bond or series of bonds, and non-bonded interactions; potentials arising from Vand der Waals and electrostatic interactions. With each step of the simulation, acceleration resulting from these forces is calculated for each particle and applied to the particle velocity and position [117]. Time steps for atomistic simulations are typically around two femtoseconds, whilst for CG simulations around 20 femtoseconds is used [118].

A great many MD forcefields are available, each applying greater or lesser levels of simplification and efficiency. Atomistic forcefields may be termed all-atom, where all atoms are explicitly simulated as MD particles, or united-atom, where many (and often all) hydrogen atoms are implicit. CG forcefields such as the MARTINI forcefield [119, 118, 120] introduce a further simplification, using MD particles (often referred to as beads) which represent three to four heavy (non-hydrogen) atoms. Simplifying the system even further, the Dry Martini method [121] simulates large systems without explicit water molecules, further reducing the number of particles in the system.

Multi-scale simulation approaches combine the benefits of multiple forcefields operating at various levels of simplification. Mostly these methods apply consecutive simulations with various levels of representation, in most cases switching between CG and atomistic representation using CG2AT [122] or Backward [123]; an example of this is shown in Figure 1.9. Adaptive resolution simulation methods can be

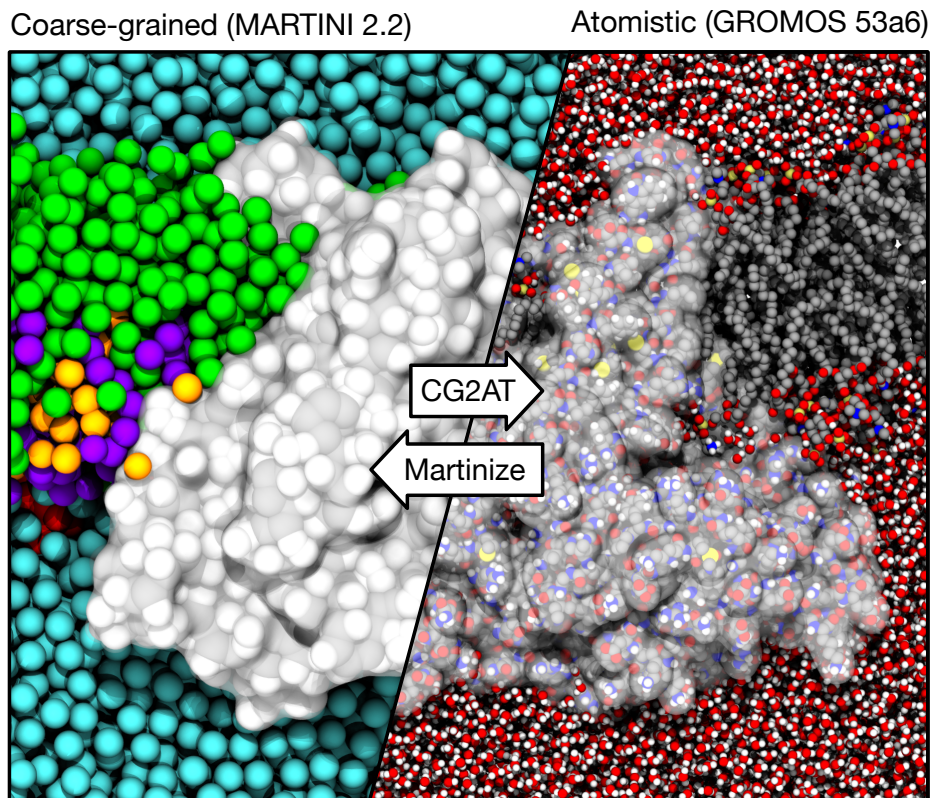


Figure 1.9: CG and atomistic representations of a novel TMEM16K structure

Mixed-lipid models of a novel TMEM16K structure are shown represented by the MARTINI 2.2 CG forcefield [119] (left) and the GROMOS 53a6 atomistic forcefield [125] (right). CG2AT [122] is used to convert from CG to atomistic representation, whilst Martinize [118] is used to convert from atomistic to CG representation. For CG simulation, water beads are shown in cyan, protein backbone and side-chain in black and white respectively, and PIP2, POPS, POPC and POPE in red, orange, green and purple respectively. Atomistic simulation is coloured according to the CPK standard.

used to concurrently simulate different domains of a simulation box at different representations, but are presently limited by the difficulty of calculating bonded and non-bonded inter-representational interactions. Since the vast majority of solvent does not make contact with the protein, an adaptive resolution approach can be used to simplify solvent-solvent interactions using CG or implicit bulk water, which is converted to all-atom representation within a set distance of the protein [124], restricting inter-representational interactions to those between solvent particles.

Beyond MD simulations, there exist several more approximate biomolecular simulation approaches which may be applied to IMPs. Many MD studies of IMP are concerned with protein-protein interactions or specific binding of lipid or drug molecules,

and so are well suited to the use of docking methods such as AutoDock Vina [126], which can suggest likely sites for protein-ligand or protein-protein interactions with limited conformational flexibility. With some exceptions, such as the MemDock α -helical protein-protein docking tool [127], docking approaches do not take the membrane environment into account, and so require refinement by CG or atomistic MD in an explicit lipid bilayer to find an optimal binding pose. Monte Carlo simulations use a stochastic method to sample a broader energy landscape over longer time periods than MD simulations, and may be applied to smaller IMP-lipid systems [e.g. 128]. The development of mesoscale simulation approaches on distance scales comparable to current super-resolution microscopy techniques (reviewed by Chavent [63]) may soon bridge the gap between biomolecular simulation and imaging of IMPs.

1.2.2 Prediction of integral membrane protein structure

High-resolution experimentally determined structures are available for only a tiny fraction of the many millions of predicted IMPs, however, thanks to genomics and proteomics sequencing approaches, IMP amino acid sequences are considerably easier to determine, with many already represented in sequence databases. Using these sequences, along with current knowledge of IMP structural biology, several methods can be used to infer protein structure from sequence.

Several algorithms can be used to predict membrane topology of IMPs, using similar sequence-based approaches to those used by secondary structure prediction algorithms such as JPred [129] or PSIPRED [130]. Early methods for prediction of α -helical IMP topology simply relied on identification of regions containing hydrophobic residues [e.g. 131], but later methods improved on this by considering transitions between multiple states along the sequence length. The first of these was Toppred [132], which used the positive inside rule [133] to model transitions between three states corresponding to cytosolic extramembrane, membrane and exoplasmic extramembrane domains. MEMSAT [134] introduced further improvements to this approach, using dynamic programming to ensure the resulting topology represented a globally optimal set of state transitions.

Further advances in prediction of α -helical membrane topology came about with the use of Hidden Markov Models (HMMs) and Artificial Neural Networks (ANNs), both methods which are highly applicable to determining an optimal set of state transitions for a set of sequential data. The first of such methods was PHDhtm

[135], which used an ANN to assign one of three (exoplasmic, membrane, cytoplasmic) states to a multi-sequence alignment. This was soon followed by two methods, TMHMM [136] and HMMTOP [137] which each used an HMM to increase the range of states considered by the model to seven and five respectively. A further development is the integration of an ANN with evolutionary information to identify conserved sites, as used by PRODIV-TMHMM [138] and MEMSAT3 [139]. More recent methods introduce a more complex set of states, which allow for re-entrant or membrane dipping α -helices. These include OCTOPUS [140], which combines an ANN used to score residue hydropathy with an HMM used to predict transition between states, and MEMSAT-SVM [141], which is also able to identify the cleaved signal peptide, and uses evolutionary information interpreted by supervised learning approach implemented using Support Vector Machines (SVMs). These methods may then be combined using consensus methods such CCTOP [142] and TOPCONS [143], which use HMMs to predict IMP topology from the outputs of several other topology prediction algorithms.

A separate set of methods can be used to predict β -barrel IMP topology, using similar approaches to those used to predict α -helical topology. Identification of β -barrels in sequences is more difficult than with α -helices since they are more reliant on long-range interactions, and so β -barrel topology prediction typically finds the most likely arrangement of β -strands on the assumption that an input sequence contains a β -barrel. These include HMM approaches such as HHomp [144], TMB-HMM [145] and BOCTOPUS [146] and approaches using ANNs such as TMBPRO [147].

Key to the accuracy of such scales is the scoring of residue hydropathy. This task is far from simple, since, as discussed in Section 1.1.3, amino acids interact very differently with each part of the lipid bilayer. A linear hydropathy scale for residue types will necessarily be an extreme simplification; for example, intermediate preference for solvent or bilayer core does not imply strong preference for an intermediate environment. The choice of linear hydropathy scale is therefore essential when predicting likely interactions of a particular residue. Peters and Elofsson [148] recently evaluated twelve such scales, derived from either experimental measurements or analysis of surface residues of IMP structures, and found that scales derived from analysis of existing transmembrane α -helices performed best when predicting IMP topology.

Predicting the orientation of a three-dimensional IMP structure in a lipid bilayer shares many of the same challenges faced when predicting IMP topology from sequence. Knowledge of the properties of amino acids in a lipid bilayer is key to

approaches such as TMDET [149] or MEMEMBED [150]. Several of these methods are discussed in more detail in Chapter 3.

Where an IMP protein sequence shares homology with sequences of already determined structures, it may be possible to predict its three-dimensional structure based on sequence alone, using a structure prediction method such as homology modelling. This approach uses a Multiple Sequence Alignment (MSA) to identify template sequences amongst proteins with known structures, from which the three-dimensional structure of a target sequence may be modelled. Homology modelling approaches developed for globular proteins, such as SwissModel [151], HHPred [152] or the Phyre2 fold recognition server [153] can be used to build accurate models of IMPs [154], however several IMP-specific optimisations can deliver substantial improvements.

One key step in the homology modelling process is finding homologous structures using MSA. Substitution rates vary substantially between membrane-spanning and non-membrane-spanning residues, and so MSAs for IMPs are substantially improved by using a membrane-specific substitution table. The first of such tables was the JTT table [155], followed by the PHAT table [156] used by the PRALINETM MSA tool [157]. MSAs required a separate algorithm to identify transmembrane regions to decide which substitution table to use. A particularly effective variation on this approach is the SToMP tables [158], as used by the Membrane Protein Threader (MP-T) MSA software [159]. MP-T uses iMembrane [160] to partition sequences into one of 34 distinct environments, 26 in the membrane and 8 in solution, based on secondary structure and predicted contacts with membrane lipids. A substitution table is determined for each environment, substantially improving MSAs where sequence identity is low (10-25%). Other methods for building MSAs for IMPs include PSI/TM-Coffee [161], AlignME [162] and TM-Align [163].

The Memoir web server [164] is an IMP-specific homology pipeline. Memoir combines multiple IMP-optimised methods, using MP-T [159] used to build MSAs with environment-specific substitution models. The membrane-specific modelling tool MEDELLER [7] is then used to build a core model of the protein, and loops are then modelled using FREAD [165], a highly accurate knowledge-based loop modelling method which is equally applicable to globular proteins. Using these methods, Memoir is capable of predicting IMP structures more accurately than SwissModel or HHPred [164].

A long-standing problem in protein structure prediction is the prediction of long-range contacts, where residues which are structurally in close proximity occur at disparate sequence locations, as is commonly the case with multi-pass or oligomeric

proteins. One emerging method is the use of long-distance correlation of mutations in MSAs to identify co-evolution of a protein-protein interface [166]. The MetaPSICOV server [167] uses several co-evolutionary methods, along with more traditional neural network predictions, to predict long-range contacts. The power of the MetaPSICOV approach was demonstrated during CASP11 by the CONSIP2 server [168], based on the MetaPSICOV method, which achieved unprecedented accuracy in prediction of intra-molecular residue-residue contacts [169].

1.2.3 Online tools and web databases

An increasingly common means to distribute computational methods or resources is through web applications, which have a number of benefits for both developers and users. To end users, web applications are instantly available, without requiring new software, configuring machines or purchasing compatible hardware, and consume only minimal local resources. For software developers, the development cycle is sped up since only a single, known host system is targeted, and once released impact may be easily quantified using standard telemetry software. Restrictive licensing of third party software may also complicate redistribution to end users, whilst permitting use on a single web server.

Web databases provide instant access to a searchable catalogue of data, which can be downloaded and used locally. The backbones of many computational tools for structural biology are provided by the Universal Protein Resource (UniProt) database family [170], a set of online databases storing annotated amino acid sequences, and the PDB network [171], comprising the Protein DataBank in Europe (PDBe) [172], Protein DataBank in Japan (PDBj) [173] and Research Collaboratory for Structural Bioinformatics (RCSB) PDB [174], storing experimentally determined three-dimensional structures. Several other databases provide further annotations of these sequences and structures, such as Gene Ontology (GO) [175], which stores semantic, evidence-tagged annotations of protein structure, function and expression, and Protein Families (Pfam) [176], which groups UniProt database accessions into families with shared ancestry and sequence identity.

Several specialised web databases focus on classification of membrane proteins and IMPs in particular. Membrane Proteins of Known Structure (mpstruc) [177] is an expert-curated database storing a hierarchy of membrane protein structures organised according to various structural and functional groupings. Transporter Classification DataBase (TCDB) [178] provides a similar hierarchy, built using expert-guided sequence alignments, for membrane transport proteins. Further specialisms exist, such

as the GPCRdb [179], a database of GPCR sequences and structures which also provides GPCR-specific annotations, and the now retired LGICdb [180], which formerly provided annotations of ligand-gated ion channels.

The question of IMP topology is addressed by several web databases. Orientations of Proteins in Membranes (OPM) [4] and Protein DataBank of Transmembrane Proteins (PDBTM) [181] contain IMP subsets of PDB structures, oriented relative to simplified lipid bilayers. Other databases expand IMP topologies to sequence-based predictions and non-structural experimental evidence. For example, TOPDB [182] and ExTopoDB [183] store evidence from over 2000 mostly non-structural experiments to determine IMP topology using approaches such as gene fusion, post-translational modification and epitope mapping. TOPDOM [184] relies less on experimental evidence and uses the CCTOP algorithm [142] to predict IMP topology, as well as identify domains and motifs with predictable sub-cellular locations.

The ubiquity of reliable, high-speed internet access, coupled with standardisation of web technologies for the development of rich interactive Graphical User Interfaces (GUIs) has enabled distribution of several computational methods as web applications. Many previously discussed tools, such as Memoir [164], Phyre2 [153] and MetaPSICOV [167] are distributed as web applications. CHARMM-GUI [185] is a complex web application providing a front-end to the CHARMM MD software [186], which can be used to set up a variety of simulation types, including IMP simulations in complex, mixed-lipid bilayers [65].

Several web applications have focused on exploration of annotated protein sequences and MSAs. Jalview and JalviewLite [187], for example, provides web-based and local GUIs to several software tools focused on building and annotating MSAs such as jpred [129] or PSI-Coffee [161], as well as annotations from UniProt [170] and Pfam [176]. The PSIPRED Protein Analysis Workbench [188], meanwhile, integrates several sequence-based prediction methods including the PSIPRED secondary structure prediction algorithm [130] and MEMSAT3 membrane helix prediction algorithm [139]. To support online GUI development, BioJS [189] is a project which provides modular visualisation components for client-side rendering of many standard diagrams, such as plots of MSAs or sequence logos.

The implementation of WebGL standards [190] by most web browsers has enabled intensive three-dimensional visualisations of protein structure within web applications without installation of additional software. PV [191] is a simple, lightweight viewer used by SwissModel [151] and RCSB PDB [174] to display cartoon or space-filling representations of protein structure. NGL Viewer [192] introduces several additional

WebGL visualisations of PDB | 4wis Ca^{2+} activated lipid scramblase

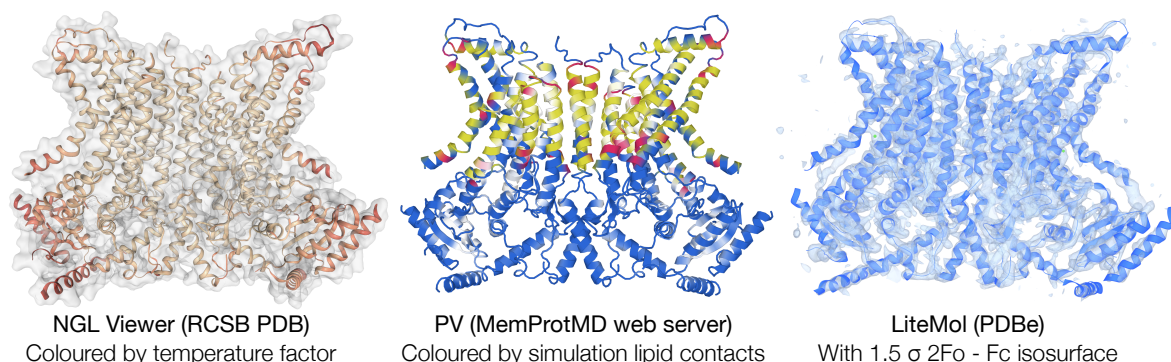


Figure 1.10: WebGL visualisations using NGL Viewer, PV and LiteMol
Images of a structure of the Ca^{2+} activated lipid scramblase (PDB:4wis) rendered by three WebGL protein viewers are shown. From left to right: cartoon representation coloured by temperature factors, overlaid with surface representation, rendered by the RCSB PDB web application using NGL Viewer [192]; cartoon representation coloured by simulation lipid contacts (blue: solvent, red: lipid head-groups, yellow: lipid tails) rendered by the MemProtMD web application using PV [191]; cartoon representation overlaid with a 1.5 σ 2Fo-Fc isosurface rendered by the PDBe web application using LiteMol (<http://webchemdev.ncbr.muni.cz/Litemol/>).

capabilities, such as the ability to play back simulation trajectories and estimate protein surfaces. LiteMol (<http://webchemdev.ncbr.muni.cz/Litemol/>) extends these capabilities even further with the ability to view electron density isosurfaces from both X-ray crystallography and Cryo-EM experiments.

The use of web applications and web servers has many pitfalls, however, in terms of reproducibility, reliability, transparency and compatibility with other software tools when compared to freely distributed, locally installed software. Web server software is often released as closed-source, is often modified, sometimes without clear version numbering, and may become temporarily or permanently unavailable with failure or maintenance of host machines or networks. Where distribution of source code and executable binaries is precluded by software licensing, transparent version control is essential to ensure research is transparent and reproducible. The web server paradigm also complicates the development of high-throughput software tools such as bioinformatics pipelines, with the majority of web servers requiring considerable expertise in computer networking to integrate with other applications. A well documented Application Programming Interface (API) and usage policy can substantially reduce barriers to integration with other applications.

1.3 Thesis overview

The general aim of my DPhil project was to develop methods and resources to assist in the computational study of IMPs structure and function, with a particular focus on examining the interactions of IMPs and membrane lipids. The results of this project are described in Chapters 2-4.

In Chapter 2, I describe the development and capabilities of the MemProtMD web server and database, a free web resource providing raw data and analysis of CG MD simulations of over 3000 IMPs structures in explicit lipid bilayers. I show how analyses of individual proteins may be combined using sequence alignments to reveal conserved contacts between IMPs and lipids across protein families, and across all current IMP structures to reveal general principles of IMP-lipid interactions, as well as trends in structure determination.

Chapter 3 concerns the development of tools for the automated identification and orientation of IMP structures from amongst over 130,000 proteins structures deposited in the PDB, using data drawn from ensemble analyses of the entire MemProtMD database presented in Chapter 2. Also described in Chapter 3 is the development of a novel method for identification and classification of membrane lipids from molecular structures, and application of this method to compile the most complete list of IMP-lipid interactions in PDB structures so far assembled.

Applications of MD simulations to studies of individual IMPs in mixed-lipid bilayers are explored in Chapter 4. I also describe the development of Alchex (Alchemical Exchange), an open-source tool for automated alteration of bilayer compositions in MD simulations.

Finally, in Chapter 5, I discuss the impact and significance of resources, methods and findings produced by work described in Chapters 2-4, as well as potential for future research based on these results.

Chapter 2

MemProtMD: A resource for membrane protein structures

2.1 Abstract

Integral Membrane Proteins (IMPs) are a pharmacologically important and scientifically interesting class of proteins for which an increasing number of molecular structures have been determined. Interactions between specific amino acid residues of the protein and components of the lipid bilayer play a vital role in the structure and function of IMPs, however, the majority of experimentally determined structures do not include lipid molecules. Using the MemProtMD pipeline [5], membrane protein structures can be embedded *in silico* in a lipid bilayer using a Coarse-Grained Self-Assembly (CGSA) approach and multi-scale Molecular Dynamics (MD) simulations.

As extensions to the existing MemProtMD simulation setup pipeline, a pipeline has been set up to automatically analyse the results of membrane protein simulations, with results populating a database and web application. The MemProtMD database and web application can be used to view analyses of individual proteins in a simulated membrane environment, as well as ensemble analyses combining results of multiple simulations to answer wider questions concerning the structure, function and evolution of IMPs.

The MemProtMD web application is an online resource, currently with over 900 users worldwide, providing access to findings from a constantly growing database of simulations of over 3000 IMPs structures. It can be accessed free of charge at <http://memprotmd.bioch.ox.ac.uk>.

2.2 Introduction

In the 32 years since the publication of the first high-resolution Integral Membrane Protein (IMP) structure [3], new IMP structures have been determined at an exponential rate [193], sustained by continuing improvements to existing methods of structure determination, producing a wealth of structural information for a growing number of individual proteins and protein families. As of July 2017, over 3000 IMP structures have been determined, 388 in 2016 alone. IMPs are embedded in lipid bilayers, interacting with the hydrophilic solvent and lipid head-groups as well as hydrophobic lipid tails. Both specific and non-specific interactions between IMPs and bilayer lipids are vital to IMP structure and function, and whilst some individual lipids can be resolved in a small number of IMP structures [52], most are resolved in the total absence of lipids.

Several methods exist to predict membrane protein orientation and bilayer thickness based on the distribution of surface amino acids. Chemical properties of each amino acid side-chain make certain bilayer locations more or less energetically favourable, with hydrophobic residues occurring more towards the core of the bilayer, aromatic residues clustering around the lipid carbonyl groups, and charged and polar residues mostly excluded from the hydrophobic core of the bilayer [108, 105, 110, 109]. Amino acids with large, flexible side-chains can be found at otherwise unfavourable locations, with the side-chain either snorkelling out of the bilayer core (e.g. arginine and lysine) or anti-snorkelling into it (e.g. phenylalanine) [106, 105]. A diverse range of proteins appear to follow a positive-inside rule, where positively charged residues appear more frequently on the cytosolic side than the extramembrane side [111].

Experimental studies have found the thickness of an unperturbed dipalmitoylphosphatidylcholine (DPPC) bilayer (measured as the distance between phosphate groups of each bilayer) to be around 4 nm [194], however bilayer thickness is thought to vary between different biological membranes, depending on the constituent lipids, with the bacterial and mitochondrial outer membrane estimated to be 0.7 nm thinner than those of a plasma membrane, based on hydrophobic thickness seen in protein structures [110].

Using statistical models of membrane thickness and amino acid distribution in the membrane, methods such as MEMEMBED [150] can be used to view an IMP in a simplified slab representation of a lipid bilayer, with each leaflet indicated by a single plane. A similar method is Orientations of Proteins in Membranes (OPM) [4], the results of which are provided for every membrane protein through a web server,

addressing a growing need for a resource to provide information about the lipid bilayer missing from solved structures.

Whilst simplified or planar models do go some way to capturing the bulk properties of a lipid bilayer, many facets of the interaction between protein and membrane remain obscured by this simple approximation. Lipids often form highly specific interactions with sites on the surface of IMPs, which have been shown to regulate protein function [195], promote dimerisation [196, 61] or catalyse chemical reactions [197]. IMPs also induce changes in the lipid bilayer, interacting with an annular shell of lipids in direct contact with amino acid side-chains and inducing complex distortions in the lipid bilayer which often modulate protein function [198] and may alter cell physiology [199]. In some cases, the protein may have even larger effects on the bilayer, such as transporting lipids between leaflets, known as flip-flopping [200].

The evolutionary impact of the relationship between membrane localisation and evolutionary rates remains unclear. It has been found that IMPs generally evolve faster than cytosolic proteins and slower than secreted proteins, likely due to higher evolutionary pressures from the extracellular environment, particularly for single-cell organisms [201]. However, it has also been shown that whilst the extracellular domains of IMPs evolve faster than soluble proteins [202], transmembrane domains alone, and especially buried residues of α -helical IMPs evolve at a substantially slower rate than soluble proteins [11, 203]. Several families of IMPs exist with conserved structure and function, and for one such family, the aquaporins, computational studies have shown that protein-lipid interactions are likely to be similarly conserved [204].

Molecular Dynamics (MD) simulations can be used to computationally reconstitute IMPs in a lipid bilayer at either atomistic representation, where all atoms are explicitly simulated, or Coarse-Grained (CG) representation, where atoms are joined into single particles representing 3-4 heavy (non-hydrogen) atoms. Through this simplification, CG representations, such as the MARTINI forcefield [118, 119, 120], are able to simulate larger systems over longer time-scales, albeit with considerable simplification, and are often used to study IMPs. In many cases a multi-scale approach can be used, converting MD simulations from CG representation to an all-atom representation using tools such as Backward [123] or CG2AT [122].

To set up, run and interpret MD simulations of IMPs, however, is a complex task which requires specialised tools and knowledge at every stage. Several tools exist to simplify the first stages of this process, where an MD forcefield is applied to a protein structure, individual lipid molecules are inserted to form a bilayer structure, and water molecules and ions are inserted to solvate the system, including HTMD

[64], the MARTINI Insane tool [205] and the CHARMM-GUI membrane builder [65]. The MemProtMD Pipeline is one such method which [5] uses Coarse-Grained Self-Assembly (CGSA) simulations, where a 1 μ s CG simulation is used to allow initially randomly placed membrane lipids to self-assemble to form a lipid bilayer, followed by equilibration.

Presented here are several major extensions to the MemProtMD pipeline, which further simplify the setting up, running and interpretation MD simulations of IMPs. An automated analysis pipeline has been developed to characterise bilayer properties, protein-lipid interactions and membrane topology of simulations set up using the MemProtMD Pipeline. Results of analyses of individual simulations are combined in several regularly updated ensemble analyses, which provide insights into the generalised properties of membrane proteins and lipid bilayers. Simulations of over 3000 IMP structures have analysed and deposited in the MemProtMD database, which can be viewed using the MemProtMD web application at <http://memprotmd.bioch.ox.ac.uk>.

2.3 Materials and Methods

MemProtMD comprises four major components. The **MemProtMD pipeline**, developed by Stansfeld *et al.* [5], inserts an IMP structure into a lipid bilayer using a multi-scale simulation approach. The resulting simulation trajectories are analysed using the **MemProtMD analysis application**, and data is then stored in the **MemProtMD database**. The MemProtMD analysis application also performs a variety of ensemble analyses. The **MemProtMD web server and web application** provide a Graphical User Interface (GUI) and Application Programming Interface (API) which can be used to browse and search the results of various analyses. An overview of the systems used to maintain, store and provide the MemProtMD database and web application can be seen in Figure 2.1.

2.3.1 System setup and simulation

The MemProtMD Pipeline

IMP structures and simulations appearing in the MemProtMD database and web application are derived from experimentally determined structures deposited in the PDB. Coarse-grained and atomistic simulations based on protein structures are set up and run using the MemProtMD pipeline [5]. The protein is first converted from

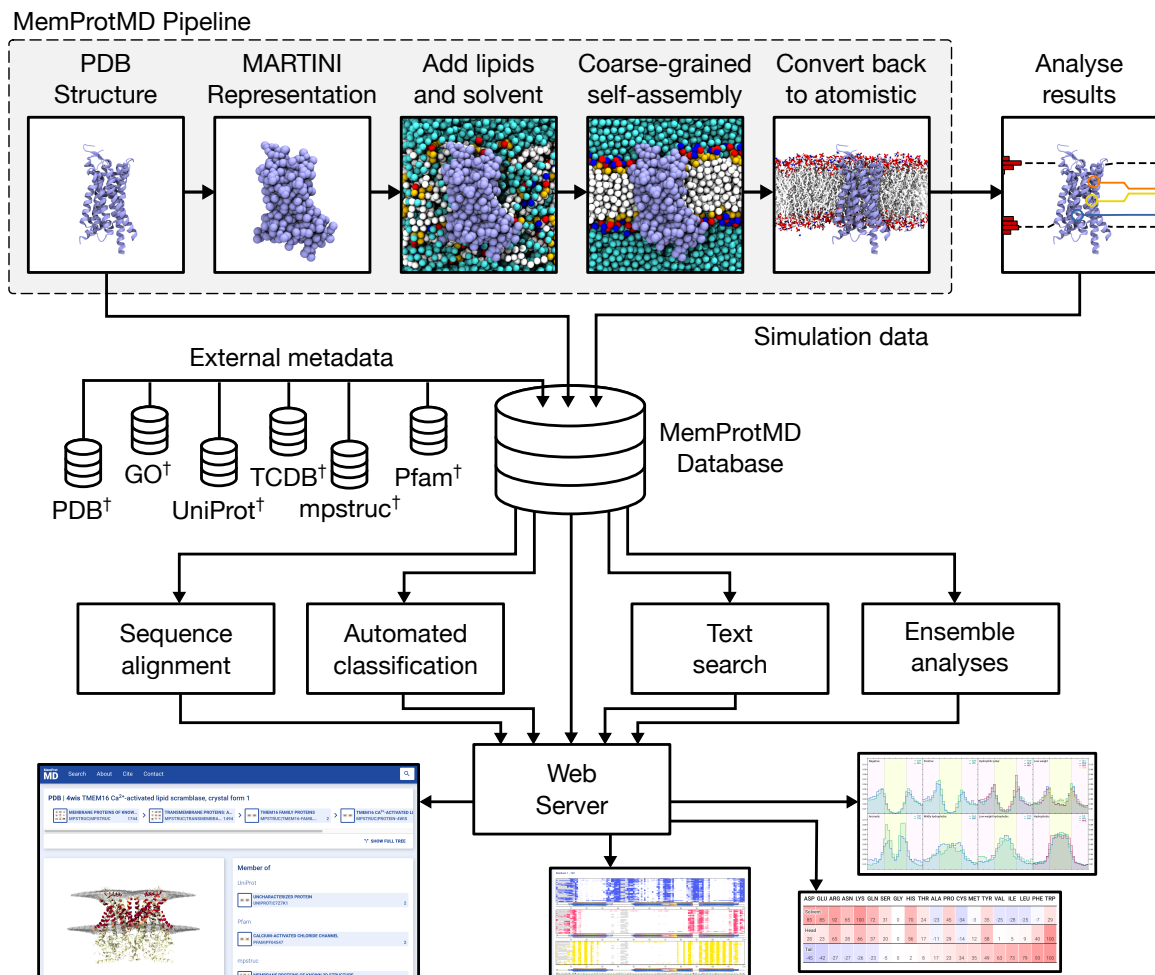


Figure 2.1: MemProtMD data processing overview

IMP structures identified in the Protein DataBank (PDB) are inserted into a coarse-grained lipid bilayer using CGSA which is then converted to atomistic representation using the MemProtMD pipeline. Metadata and annotations from several external databases are loaded and deposited along with results of simulation analysis into the MemProtMD database. These data are then used to provide sequence alignments, automated classifications, text search tools and ensemble analyses. Analysis data can be accessed through the MemProtMD web server, and is rendered using the MemProtMD web application.

The MemProtMD pipeline (grey dashed box) was developed by Stansfeld *et al.* [5]. Data sources indicated with a (†) represent external databases.

atomistic to MARTINI coarse-grained representation [118, 119, 120]. Lipids are then packed randomly into a slab near the predicted transmembrane region of the protein. Water and ion beads are added and a 200 ns simulation is run, during which lipid molecules self-assemble to form a bilayer. A further 800 ns of coarse-grained simulation is then performed. The system is then converted to atomistic representation using CG2AT [122].

Job Status Monitoring

Once identified in the PDB, a protein passes through several simulation and analysis processes before appearing in the MemProtMD database and web application. Progress on current jobs is monitored and a job status as well as the time of last status change is recorded and can be viewed on the MemProtMD web app's status page at <http://memprotmd.bioch.ox.ac.uk/status>. An identified protein initially receives an "Identified" status, which changes to "Simulating" once a corresponding simulation can be found currently in progress. Once this simulation is completed, it is marked as "Converting" until atomistic simulations are completed. The status is then marked as "Analysing" until various analyses are completed, whereupon the status changes to "Done".

Validation of simulation results

The MemProtMD pipeline processes a diverse set of inputs using complex software which may not always behave as expected. Simulation data must be validated before analyses are performed and results deposited into the MemProtMD database. Final frames of coarse-grained and atomistic simulations are stored as PDB files, and validated using the PDB parser provided by MDAnalysis [206]. The number of protein residues in the starting structure, coarse-grained final frame and atomistic final frame is counted and must match. The simulation trajectory file is then validated, ensuring it is a valid file and contains the expected number of frames. Finally, vertical distribution of protein and lipid atoms is calculated for 20 frames sampled from the full length of simulation to ensure the simulation contains two leaflets and that the protein is embedded in the membrane. If any tests are failed, analysis is stopped and details of errors added to a log. Tests are periodically re-run and analysis resumes as soon as the error is fixed, typically after some manual inspection.

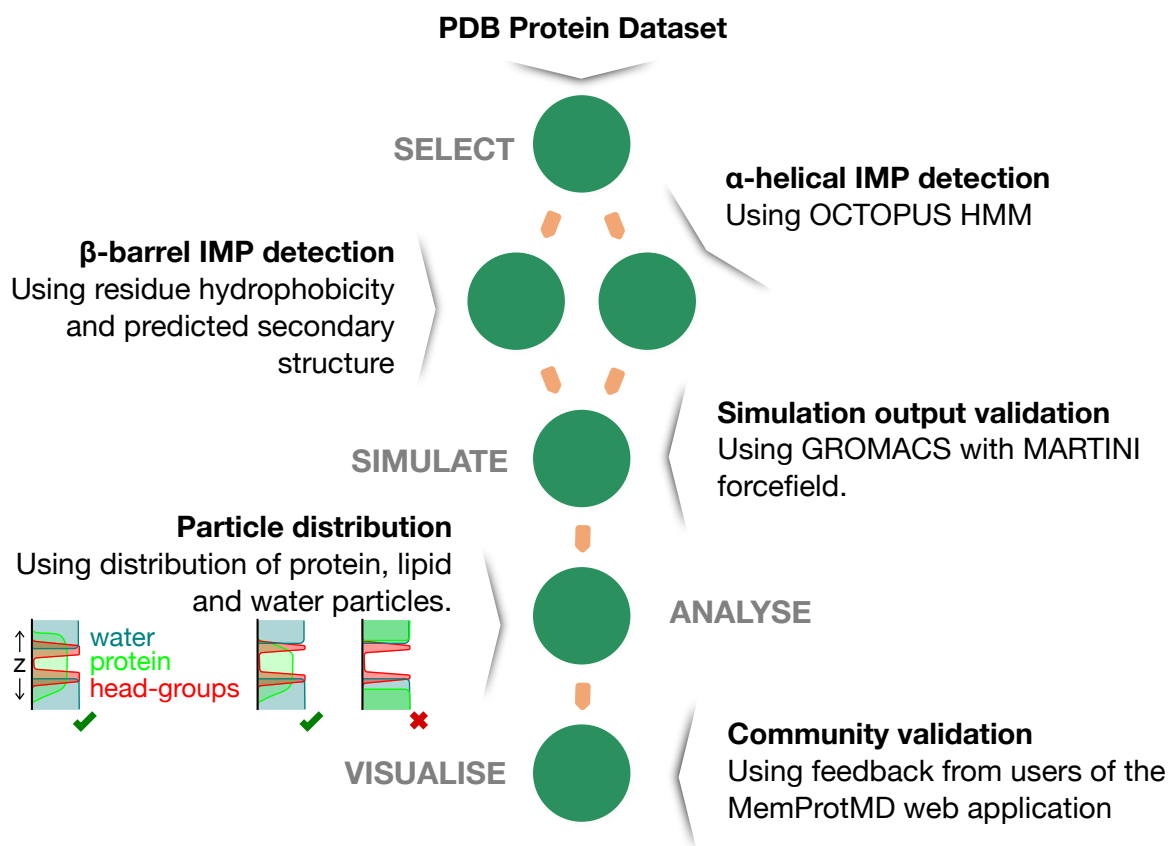


Figure 2.2: Validation of MemProtMD data

The validation steps applied to protein simulation data in the MemProtMD database are shown. Initial identification of membrane proteins uses a system developed by Dr Phillip Stansfeld; α -helical are selected using OCTOPUS [140], while β -barrel proteins are selected using hydrophobicity and secondary structure [5]. Simulation results from GROMACS [117] are automatically validated, ensuring no abnormal forces or violation of constraints set by the MARTINI forcefield [118]. The distribution of water, protein and lipid head-group atoms is then calculated, allowing automatic detection of proteins not embedded in the bilayer or water ingress into the bilayer. Remaining errors are identified by web application users.

2.3.2 Characterising bilayer properties

Identification of leaflets and flip-flop behaviour

The CGSA simulation process produces a planar lipid bilayer around each protein. Lipid bilayers are composed of two leaflets, which are in turn composed of individual lipid molecules. Spontaneous movement of a freely diffusing lipid between leaflets is energetically unfavourable and therefore rare, however, many proteins are able to carry lipids between leaflets, and this "flip-flop" behaviour is seen in MemProtMD simulations of the TMEM16 lipid scramblases PDB:4wis and PDB:4wit.

For each simulation frame, non-protein molecules are assigned a topological class; they may belong to either the upper or lower leaflet, belong to neither leaflet (in the case of lipids deeply buried in the protein structure), be in the process of flipping either from the upper leaflet to lower leaflet or lower leaflet to upper leaflet, or be part of the solvent.

To identify membrane leaflets, clustering is performed on the phosphate (PO4) beads of lipid molecules. For each frame, a graph is constructed with each phosphate bead represented by a single node. Edges are inserted where pairs of phosphate beads are less than 1.2 nm apart. NetworkX [207] is then used to find the two largest connected components of the graph, which are labelled as upper and lower leaflets according to the vertical coordinate of their respective centres of geometry. In some cases, an error occurs and data from the last successful frame are used instead. Errors are usually caused when flip-flopping lipids are able to connect the upper and lower leaflets, resulting in a connected component spanning both leaflets.

Resulting leaflet assignments contain a great deal of noise, and a post-processing step is required to produce accurate results and identify flip-flop behaviour. For each lipid, a sliding window of up to 10 ns before and 10 ns after the current frame is considered. If the lipid belongs to a particular leaflet for at least 70% of frames within that window it is assigned to that leaflet, otherwise it is assigned to neither leaflet. If a lipid is assigned to one leaflet, then is unassigned for several frames, then is assigned to a different leaflet, it is labelled as flipping during the unassigned period. If the lipid is neither flipping nor assigned to a leaflet, but is both in contact with the protein and within the vertical range of leaflet-assigned lipids, it is labelled as buried. Other beads, such as those representing water and ions, are labelled as solvent. To reduce storage size whilst allowing data to be quickly loaded into memory, frame-by-frame labels are delta-encoded and stored in YAML Ain't Markup Language (YAML) format.

Characterising leaflet surfaces

Lipids in planar bilayers rarely conform to simple plane geometry; leaflets may become thinner or thicker in proximity to the protein, or distort around specific domains. Once lipid molecules are assigned to leaflets, detailed surfaces are calculated for each leaflet which can reveal these distortions. For each leaflet, the mean vertical position of phosphate beads is calculated for each cell within a 2-dimensional 0.1 nm grid, resulting in a mesh showing distortions in the membrane surface. Protein atoms within 0.6 nm of a grid cell are also recorded and used to overlay the mesh with an atomistic version of the protein.

Characterising annular shells of lipids

The presence of a protein in the lipid bilayer results in the formation of a highly ordered annular shell of lipid molecules in contact with the protein, surrounded increasingly less ordered shells of lipids. Lipid molecules are usually rapidly exchanged between annular shells. Lipid annular shells are identified using the Radial Distribution Function (RDF) of lipid phosphate groups around the protein, which contains progressively less extreme peaks and troughs of lipid density radiating out from the protein (in a manner similar to damped oscillation). The first peak is taken as the centre of the first annular shell, and the first trough as the boundary between first and second annular shells. Further peaks and troughs are interpreted as annular shells until the distance between adjacent troughs is less than 0.4 nm or greater than 0.8 nm.

Determining annular and bulk bilayer properties

Membrane properties vary greatly between lipids far away from the protein (bulk) and those in direct contact with it (annular). In close proximity to protein side-chains, the membrane may become thicker or thinner, or become raised or lowered relative to bulk. These changes may reveal the native membrane in which the protein occurs, or be characteristic of proteins such as pore-forming toxins. Bulk membrane lipids are defined as those which fall outside of an identified annular shell, or are at least 2 nm from the nearest protein residue, whilst annular membrane lipids are defined as those which are closer to the protein than the edge of the first annular shell. Leaflet surface meshes are used to determine the mean vertical coordinate for each leaflet in both annular and bulk regions. The thickness and vertical coordinate of the membrane centre are calculated respectively as the difference and mean of vertical coordinates

of leaflets in each region. Thickness and vertical centre is stored for bulk and annular regions, as well as any change in thickness and vertical centre.

Determining per-residue bilayer properties

Extrema of bilayer distortions are likely to occur around specific residues of the protein, and the vertical coordinate of the membrane surface near each residue is estimated. Residues less than 0.6 nm from the bilayer surface are assigned the vertical coordinate of the nearest surface point. Coordinates from residues further from the surface are interpolated using a Gaussian kernel.

2.3.3 Efficient calculation of lipid and solvent contacts with protein residues

Characterising contacts made between parts of the protein and lipid or solvents can be helpful when studying the relationship between protein and membrane. Contact times are calculated for each residue in a simulated protein as the proportion of time for which it is within 0.6 nm of beads representing parts of the lipid or solvent. Contacts are calculated for groups of beads, rather than between individual beads, e.g. all the Choline beads of a molecule of DPPC in contact with a particular residue of the protein. Lipid and solvent beads defined in the MARTINI forcefield each receive a number of tags which are later used to construct groups. These tags are detailed in Table 2.1. Groups for which contacts are calculated incorporate these tags, as well as the topological feature the bead belongs to (upper leaflet, lower leaflet, flipping between leaflets or solvent) and the bead residue name. These are detailed in Table 2.2.

KDTree-based contact identification

An efficient method to identify members of a set of points within a cutoff distance of another set of points is to use a KDTree. KDTrees divide an n-dimensional volume into increasingly small cells, represented as a tree structure, branching at each cell division. KDTrees for protein and non-protein atoms are constructed using an algorithm developed by Maneewongvatana and Mount [208], and contacts within 0.6 nm are determined for each frame of each simulation using the Scientific Python (SciPy) library [209].

Table 2.1: Tags assigned by MemProtMD to MARTINI beads

Tag	Beads
Solvent	W, NA+, CL-
Head	NC3, PO4, GL1, GL2
Tail	C1A, C2A, C3A, C4A, C1B, C2B, C3B, C4B
Water	W
Sodium	NA+
Chloride	CL-
Choline	NC3
Phosphate	PO4
Glycerol	GL1, GL2
Tail_1	C1A, C1B
Tail_2	C2A, C2B
Tail_3	C3A, C3B
Tail_4	C4A, C4B

Aggregation of contact data

Initial determination of contacts produces pairs of protein and non-protein beads within 0.6 nm of each other. Aggregation is then performed to identify contacts between entire residues of the protein and classes of lipid and solvent beads. For solvent and lipid molecules, a mapping between individual beads and groups (as defined in Table 2.2) is constructed for each frame (since in some cases beads may transition between leaflets over time). Protein beads are mapped to their corresponding residue. These mappings are then used to identify contacts between each group of lipid or solvent beads and each protein residue. Once contacts are identified for each frame, binary contact data is time-averaged to determine the proportion of frames for which a contact is made.

2.3.4 Representations of membrane protein topology

2D basis-spline abstractions of α -helices and β -strands

Each IMP structure can be viewed as a collection of membrane-spanning secondary structure elements. In the case of α -helical proteins, this will consist of at least one membrane-spanning α -helix, whilst in a β -barrel protein this will consist of several membrane-spanning β -strands. Cartoon representations of membrane protein topology often depict the membrane as a slab with secondary structure elements of the protein passing through it, connected by loops. In order to display a similar diagram,

Table 2.2: Groups of beads used to calculate MemProtMD contact data

Topological Feature	Residue Name	Bead Tag
Upper leaflet	DPPC	Choline
Upper leaflet	DPPC	Phosphate
Upper leaflet	DPPC	Glycerol
Upper leaflet	DPPC	Tail_1
Upper leaflet	DPPC	Tail_2
Upper leaflet	DPPC	Tail_3
Upper leaflet	DPPC	Tail_4
Lower leaflet	DPPC	Choline
Lower leaflet	DPPC	Phosphate
Lower leaflet	DPPC	Glycerol
Lower leaflet	DPPC	Tail_1
Lower leaflet	DPPC	Tail_2
Lower leaflet	DPPC	Tail_3
Lower leaflet	DPPC	Tail_4
Flipping between leaflets	DPPC	Choline
Flipping between leaflets	DPPC	Phosphate
Flipping between leaflets	DPPC	Glycerol
Flipping between leaflets	DPPC	Tail_1
Flipping between leaflets	DPPC	Tail_2
Flipping between leaflets	DPPC	Tail_3
Flipping between leaflets	DPPC	Tail_4
Solvent	W	Water
Solvent	ION	Chloride
Solvent	ION	Sodium
(Any)	(Any)	Solvent
(Any)	(Any)	Head
(Any)	(Any)	Tail
Upper	(Any)	(Any)
Lower	(Any)	(Any)
Flipping	(Any)	(Any)
Solvent	(Any)	(Any)
(Any)	ION	(Any)
(Any)	DPPC	(Any)
(Any)	W	(Any)

the protein is summarised as a series of basis splines which can be used to draw a 2D topology cartoon by the MemProtMD web application.

DSSP [210, 211] is first used to assign secondary structure to each residue. Single-residue breaks in α -helices and β -strands are filled in, and a smoothed 3D backbone trace is then calculated, either in the case of β -strands by convolution with a Gaussian kernel or in the case of an α -helix by finding the helix core using the register of individual helices. Three-dimensional points are then reduced to two dimensions, taking the membrane normal as the vertical coordinate and cumulative displacement in two axes perpendicular to the membrane normal as the horizontal coordinate. A basis spline is then fitted for the set of two-dimensional points making up each secondary structure element, using a Monte-Carlo method which optimises and adds new basis nodes until the mean distance between original and fitted points is less than 0.1 nm. In most cases, three nodes are sufficient to describe each secondary structure element.

A YAML file is saved for each structure, containing a list of secondary structure elements, their corresponding basis spline nodes, their angle and the complexity of curvature (number of basis spline nodes required for an accurate fit). This file can be used in conjunction with the annular and bulk properties of the bilayer to draw a cartoon representation of membrane topology.

Per-chain summary of protein topology

The number and type of membrane spanning secondary structure elements present in each chain of a protein structure provides a quick topological overview, and may be used to identify particular types of membrane protein (for example, the seven transmembrane α -helices characteristic of a G-Protein Coupled Receptor (GPCR) structure). For each chain, the number of α -helical and β -barrel membrane spanning elements is determined. The backbone trace of the protein is first smoothed using a Gaussian kernel. The residues of the protein falling between the upper and lower leaflets is then split into membrane-spanning elements using abrupt changes in direction relative to the membrane normal. Each membrane-spanning element containing at least 70% α -helical or β -strand residues is classified as α -helical or β -stranded respectively. Counts are then saved in a YAML file.

Table 2.3: Static files available from MemProtMD

Listed files are available to download for any completed simulation. The tilde ($\sim$) may be replaced with the name of any individual simulation.

Filename	Description
$\sim$ -ready.pdb	All-atom representation, no lipids or solvent
$\sim$ -atomistic.pdb	Final frame of atomistic simulation
$\sim$ -coarsegrained.pdb	Final frame of coarse-grained simulation
$\sim$ -distortions.pdb	All-atom representation with grid of dummy atoms along bilayer surfaces
$\sim$ -head-contacts.pdb	All-atom representation with lipid head-group contact frequency stored in temperature factors
$\sim$ -tail-contacts.pdb	All-atom representation with lipid tail contact frequency stored in temperature factors
$\sim$ -solvent-contacts.pdb	All-atom representation with solvent contact frequency stored in temperature factors
$\sim$ -atomistic-simulation.zip	Zip archive containing files required to run further atomistic simulations
$\sim$ -coarsegrained-simulation.zip	Zip archive containing files required to run further coarse-grained simulations
$\sim$ -by-resid.csv	Table of residue properties (contacts, coordinates etc.)
$\sim$ -thinning.latest.yml	Annular and bulk membrane thicknesses and vertical coordinates

2.3.5 Static files

The results of many analyses are stored in the MemProtMD database and can be retrieved using a MongoDB JavaScript Object Notation (JSON) query. In some cases, PDB-formatted structures, tabular data and the results of analyses can be downloaded as a static file. Details of these files can be found in Table 2.3. Before users are offered the opportunity to download each file, the MemProtMD web application checks that the file exists and reports the file size to the user.

Pre-rendered images

Whilst structures can be interactively visualised online, this typically takes several seconds to initialise and so a collection of pre-rendered images are also created which

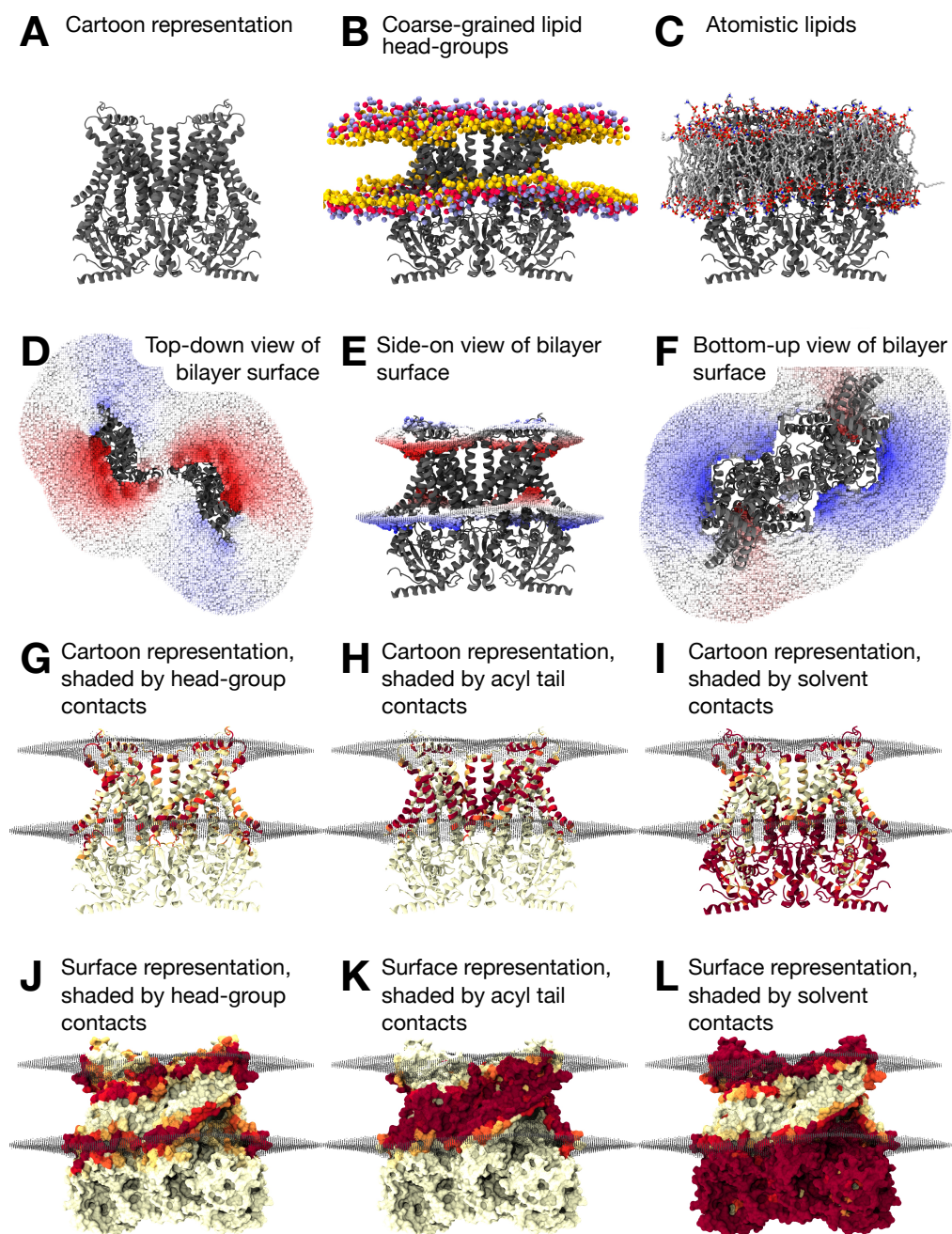


Figure 2.3: Pre-rendered images produced by MemProtMD

A number of pre-rendered images for a TMEM16 scramblase protein (PDB-4wis) are shown. **A** The protein is shown aligned to the membrane normal in cartoon representation. Further images show the addition of **B** coarse-grained lipid head-groups and **C** atomistic lipids. The protein is shown with a mean bilayer surface in **D** top-down, **E** side-on and **F** bottom-up orientation relative to the bilayer. Areas of membrane thinning are shown in red, whilst areas of membrane thickening are shown in blue. The protein is shown shaded according to contacts with **G,J** lipid head groups, **H,K** lipid tails and **I,L** solvent molecules in both **G,H,I** cartoon and **JKL** surface representations.

can be downloaded and viewed instantly. Images are rendered using Visual Molecular Dynamics (VMD), controlled using the pexpect Python library. Images include representations of coarse-grained and atomistic lipids, membrane surfaces and per-residue contacts with solvent, lipid head-groups and lipid tails. A variety of resolutions are produced to allow either quick loading of small thumbnails or detailed viewing of full-resolution images. Examples of pre-rendered images are shown in Figure 2.3.

2.3.6 Acquisition and storage of metadata

The MemProtMD web application uses metadata drawn from several external online sources to improve search and browsing experiences. In order to handle references to multiple databases, MemProtMD records use a compound key comprised of the database name (e.g. PDB) and the accession within the database (e.g. 4wis). Where a single key is required these are joined by the vertical pipe character (e.g. PDB|4wis). Each record keeps a note of the license under which content has been reused, as well as the original authors and a link to source material. Any two records may be connected using relationships, regardless of original source. Relationships between records consist of a type such as "child/subset of", "parent/superset of", "sibling of", "collection containing", "contained by collection", as well as optional parameters. Altering parameters of the parent and child relationships can be used to great grandparent/grandchild relationships using the "level" parameter, or alternatively an automatically assigned "contained by collection" relationship may also include a "confidence" parameter.

Protein DataBank (PDB)

The Protein DataBank (PDB) [171] is a database of protein structures provided by an international network of websites comprising the Research Collaboratory for Structural Bioinformatics (RCSB) [174], Protein DataBank in Europe (PDBe) [172] and Protein DataBank in Japan (PDBj) [173]. Each experimentally determined protein structure is deposited in the PDB as a record containing metadata indicating the nature of the experiment and type of protein, at least one set of atomic coordinates, and a reference to a journal article where the structure was first published.

PDB accessions are the primary references for MemProtMD simulations; the MemProtMD pipeline [5] aims to perform a single CGSA simulation for the most biologically relevant assembly of each IMP structure deposited in the PDB. Using a Python function, metadata for each PDB accession is downloaded from the RCSB website

in Macromolecular Crystallographic Information File (mmCIF) format. From this, the deposition title, keywords list, author list and publication Digital Object Identifier (DOI) are extracted. A series of requests are then made to a Representational State Transfer (REST) API provided by the RCSB website to download details of the ligands and protein sequences contained in the structure.

PDB deposition titles often contain details of experimental protocols in addition to the name of the protein, and so an additional title is constructed to be the most widely used name for the protein. For example, the PDB deposition **1rh5** is assigned the title *The structure of a protein conducting channel*, but would be better described as SecYE β translocon. In order to assign a simplified title, UniProt and mpstruc databases are first queried for titles of records referring to the PDB deposition. If a title which is shorter and lexicographically similar to the structure title (assessed by computing Levenshtein distance [212]) it is used. If no suitable alternative can be found, a Regular Expression is used to trim commonly used words or phrases such as "the * structure of", "in complex with **", "bound to **", "at * resolution" (the * character used here as a wildcard).

Universal Protein Resource (UniProt)

The Universal Protein Resource (UniProt) [170] stores information about protein sequences, including annotations such as (where available) organism lineage, genomic location, subcellular location and topology. Although in some cases mutations, insertions and deletions may have been performed, all protein structures deposited in the PDB reference at least one UniProt sequence.

UniProt sequences represented in MemProtMD are identified through multiple Extensible Markup Language (XML) requests to the RCSB REST API, which establish a per-amino-acid mapping between PDB structures and UniProt sequences. UniProt records are then downloaded in XML format and a title, keywords, organism, subcellular location and any functional annotations are then extracted. Mappings to Protein Families (Pfam) and Gene Ontology (GO) accessions are also extracted and linked to both the UniProt accession and referenced PDB accession.

Protein Families (Pfam)

Protein Families (Pfam) [176] groups protein sequences into families based on sequence homology. Each Pfam family contains a number of UniProt protein sequences, and Pfam accessions are mapped to UniProt sequences when UniProt metadata is

downloaded. A title is extracted from a document downloaded from the Pfam website in HyperText Markup Language (HTML) format.

Membrane Proteins of Known Structure (mpstruc)

Membrane Proteins of Known Structure (mpstruc) [177] is an expert-curated tree structure of membrane protein structures and families, with annotations including names of proteins, mutations and ligands. The tree is downloaded in XML format and parsed to extract titles of proteins and child/parent relationships. These relationships are added to each record and can be viewed online. Each PDB structure categorised by mpstruc is added as a leaf of the mpstruc tree.

Transporter Classification DataBase (TCDB)

The Transporter Classification DataBase (TCDB) [178] is a similarly expert-curated tree focusing on transporter proteins. The tree is downloaded in HTML format from the TCDB website and parsed to produce a set of parent/child relationships, and a mapping from TCDB accessions to PDB accessions is downloaded as a tab-delimited text file. Each node within the TCDB tree becomes an accession in the MemProtMD database.

2.3.7 Automated classification of IMPs based on simulation results

Whilst several classification systems for IMPs already exist, the simulation data available through MemProtMD provides novel scope for more comprehensive and accurate classification. Classification parameters for each group of proteins are defined in YAML document, which contains a title, description, parent collection (if applicable), and a number of conditions with positive or negative scores. Each matched condition increases or decreases a structure's score, depending on the condition. Conditions may query the topology of the protein, numbers of lipid contacting residues or membrane surface parameters, as well as structure metadata and membership of external classification systems. Any structure scoring over 100 is added to the collection, and the score is added as a parameter of the relationship between structure and collection.

2.3.8 Ensemble analyses

The results of analyses performed on individual protein structures and simulation trajectories can be combined to produce a set of ensemble analyses, from which more

general conclusions can be drawn about IMP structure and function. Ensemble analyses are updated weekly and can be browsed at <http://memprotmd.bioch.ox.ac.uk/stats>.

Collection statistics

For each automatic collection, each external database and each external database accession referencing more than 30 structures, a set of simple summary statistics are calculated. These include the experimental methods used to determine structures, annular thickness of the simulated membrane, date of publication and resolution of structures determined by X-ray crystallography. These can reveal trends resulting from the development of new structure determination methods, or show the general properties of a group of proteins. Interactive histograms and bar charts are used to display these data online; clicking a bar creates a new database query returning matching proteins. The scope of these statistics is limited to IMPs by filtering out any database accession which does not contain at least ten residues making contact with membrane lipid tails for over 30% of simulation time.

Development of per-residue contact scales

The unique chemical properties of each of the twenty naturally occurring amino acids result in variation in the frequency of contacts between amino acid types and different parts of the membrane and solvent. This can be calculated for every single protein in the MemProtMD database and used to rank amino acids by frequency of contacts with specific parts of the membrane or solvent. This can in turn be used to draw conclusions about the chemical properties of each amino acid in a membrane environment.

Once lipid and solvent contacts with protein residues are calculated for an individual protein, each residue is added to a large pool of residues, from which the mean frequency of contacts with different parts of the membrane and solvent can be calculated. First, residues which are in contact with neither solvent or lipid beads for over 70 % of simulation time are excluded. Mean contact frequency between each amino acid type and each lipid bead type is then calculated based on the remaining residues. Mean frequencies are then normalised for consistency with similar scales derived experimentally by Monera *et Al.* [213] such that Glycine becomes 0 and the highest frequency becomes 100.

Determining amino acid residue occupancy along the membrane normal

An alternative way to examine the chemical properties of amino acid types in a lipid bilayer environment is to calculate the frequency at which each amino acid occurs at different positions along the membrane normal. In order to calculate this, all analysed amino acid residues in the MemProtMD database which form contacts with lipid or solvent molecules for at least 30% of simulation time are considered. Residues forming the inner surface of membrane pores are removed, and displacement from the membrane centre is calculated for each remaining residue. To account for variation in membrane thickness between proteins, displacement values occurring between the phosphate beads of each leaflet are normalised to produce a bilayer thickness of exactly 4 nm. Residues are assigned displacement bins of 0.4 nm width, and the frequency of each amino acid type within each bin is calculated.

Aggregation of simulation results using sequence homology

Contacts observed between lipid and solvent molecules and simulated proteins should be reproduced by structures of the same protein, and may be conserved in structures of proteins with some sequence homology. In order to examine this, lipid and solvent contacts and position relative to the membrane measured during simulations are overlaid on a sequence alignment for each UniProt and Pfam accession. Each wild-type residue in a simulated protein is mapped to a position in a UniProt protein sequence, and in the case of UniProt alignments properties are simply aggregated using this mapping. Sequence alignments for Pfam accessions are produced by first mapping positions within UniProt sequences to a position in a Multiple Sequence Alignment (MSA). Promals [214] is used in combination with TM-align [163] to produce a multi-sequence alignment which takes protein structure and transmembrane sequences into account. The output from Promals is used to produce mappings from UniProt sequence positions to Pfam MSA positions, which in turn can be used to map residues of simulated proteins to Pfam MSA positions. The final alignment shows consensus amino-acid sequence, contacts with lipid headgroups, contacts with lipid tails, contacts with solvent, depth in the membrane and secondary structure.

2.3.9 Data storage

Data models

A number of data models are defined, representing different types of entities for which the MemProtMD database stores data. Data are stored on a backed up Network File System (NFS) volume, with a folder for each model containing subfolders for each instance of that model. Presently data models exist for: a CGSA simulation of an IMP; an external database; a connected social media account; an ensemble analysis and a static page on the MemProtMD website. For each entity a lifecycle is defined, which schedules a number of operations to be performed when some conditions are met. This may include publication of updated metadata, completion of a simulation or an update to one or more methods used to analyse simulations.

Data which can be readily represented as either a dictionary or list are stored either in YAML format (where the data can be safely edited by hand if needed) or in JSON format (where editing data by hand is inadvisable). Tabular data, such as per-residue contact data, is stored in Comma Separated Values (CSV) format. Protein structures are stored in PDB format and images in Joint Photographic Experts Group (JPEG) format. All of these formats can be read, written and sent over HyperText Type Protocol (HTTP) efficiently by both Python and JavaScript

Data validation

Data are drawn from diverse sources and methods, and so must be routinely validated to ensure referential integrity. Where accessions in external databases have been linked through relationships, these are checked regularly to ensure no circular relationships have been created, and that an appropriate reciprocal relationship exists (for example, if A is a child of B then B must be a parent of A). Accessions in external databases are also checked to ensure at least one completed CGSA simulation is referenced. UniProt accessions may occasionally be joined, split or deprecated; in these cases several API queries to the UniProt website are used to update UniProt data.

Data storage in a document-oriented database

To facilitate faster querying, data are also replicated in a MongoDB document-oriented database. MongoDB stores data in JSON format and does not use any

kind of schema. This allows for changes in data structure such as adding or removing fields with less overhead than a relational database. Documents are grouped into three collections storing, respectively, documents relating to simulations, references to external databases and structure authors. These may then be queried using a JSON formatted set of search conditions, set of fields to return (projection) and optional sorting instructions.

The MemProtMD database contains data which may be helpful when building other bioinformatics applications, and using MongoDB allows external users to directly query the underlying database with minimal restrictions. Queries are first searched for MongoDB commands; any query containing a command which has not been explicitly permitted is rejected. If the query is valid JSON and contains only permitted commands, it is passed on to the database and results are returned. Each query is timed and API users are subject to a quota, reset every 15 minutes, which limits time spent responding to API calls from external sources.

Database indexes

MongoDB indexes store small amounts of data from each document in memory to enable faster searching and sorting. Single-field indexes simply store the value of a particular field for each document, whilst text indexes store text drawn from a number of fields and can be used to score documents according to their probable relevance to a word or phrase. Commonly queried fields, such as (where present) the accession (or ID), database, title, authors, release date, resolution and experimental method each receive a single field key.

For references to other databases, a text index is maintained to provide a simple text search. Fields are weighted by their likely relevance to improve the order in which results appear (see Table 2.4); the accession is given a high value (100), ensuring that searches which exactly match an accession will return the record with the corresponding accession first. Other fields are weighted according to the likely significance of matching a search term in that field. Text search results are also weighted by the number of other simulations linked to a reference; this ensures that a search for the string "GPCRs" will return the GPCR family ahead of individual GPCR proteins.

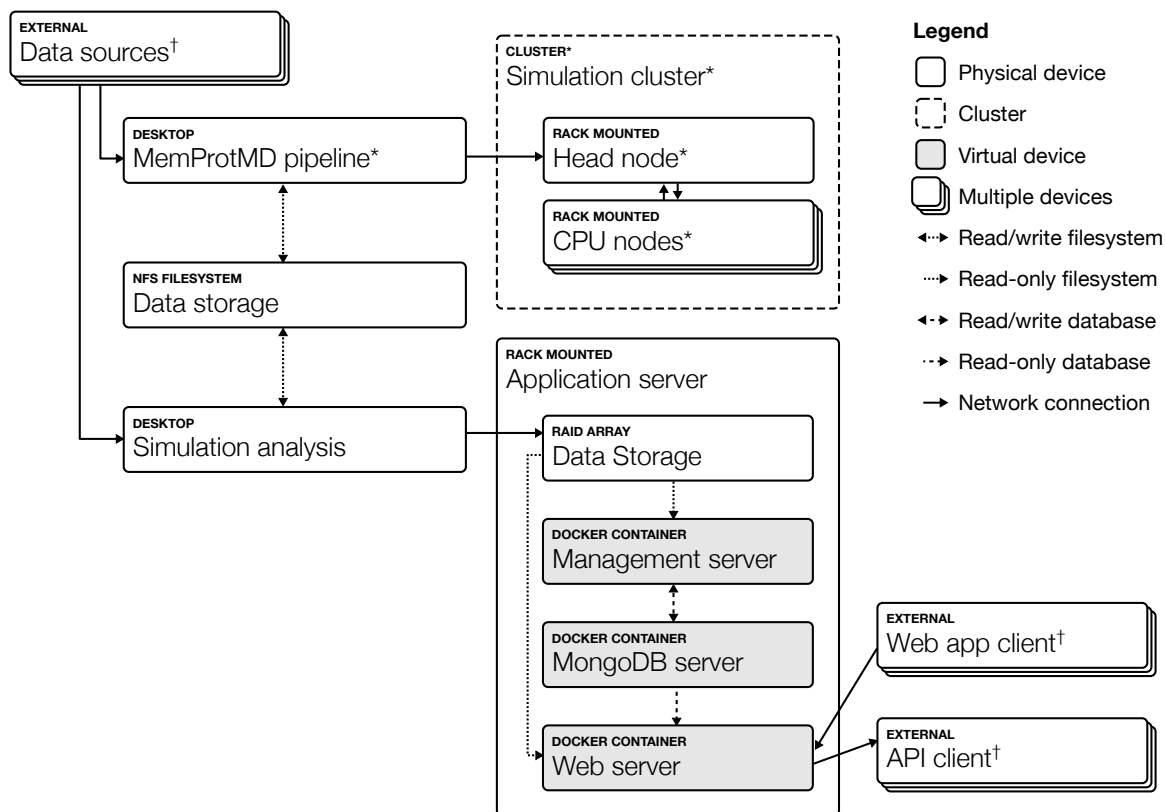


Figure 2.4: MemProtMD infrastructure overview

An overview of the physical and virtual devices used by the MemProtMD pipeline. Data are loaded from PDB structures by a desktop computer operating the MemProtMD pipeline. Once simulation setup is complete, files are uploaded to a cluster to be run. Completed simulations are stored on an NFS volume and then analysed using a Python application on another desktop computer. Once analysis is completed, data are uploaded to a Redundant Array of Independent Disks (RAID) device attached to a rack-mounted server. A Docker container running a Python application keeps a MongoDB database (also running in a Docker container) updated. A final Docker container runs a Node.JS web server which serves web application and API clients over HTTP.

Devices indicated with an asterisk (*) are part of the MemProtMD pipeline and were configured by Stansfeld *et al.* [5]. Devices indicated with a (†) are operated externally.

Table 2.4: Fields used to perform text searches on the MemProtMD database

Field name	Weight	Content
accession	100	Accession/ID of the record in the source database
title	50	Title of the record - may have been edited in the case of PDB records.
subtitle	50	Original title (if not the same as title) or alternative title.
keyword_string	6	Comma-separated list of keywords assigned by the original database.
organism.common_name	4	Common name of the organism where protein is found.
organism.scientific_name	4	Scientific name of the organism where the protein is found.
auto_keywords	1	Keywords which have been added automatically but not checked.

2.3.10 Application structure

Infrastructure overview

An overview of the physical and virtual devices used by MemProtMD is shown in Figure 2.4. Infrastructure providing the MemProtMD simulation analysis application, database and web server build on previously existing infrastructure set up to run the MemProtMD pipeline[5].

Analysis application

The MemProtMD analysis application is written in Python and automatically performs a variety of functions. In some cases, actions are performed according to a set schedule. These include querying external databases for updated metadata, checking relationships to ensure referential integrity, checking the status of jobs currently in progress and copying new data to the RAID volume on the application server. Ensemble analyses are also scheduled to update periodically, typically running once per week.

Analysis methods may be packaged as jobs associated with particular data models, and the analysis application will ensure that jobs are performed on instances of those models in the correct order. Each analysis method is assigned a name, version, required inputs and list of dependencies (other jobs which must be completed for

that instance before starting). The analysis application performs each job as soon as all dependencies are completed and all inputs are present. The job name, job version, parameters and input files are combined and used to calculate a hash value, which can be quickly checked to see if the job should be re-run. If an updated version of an analysis method becomes available, or inputs change, a job is re-run, triggering the re-running of all dependent jobs.

Once a set of simulation analyses are available to view online, the analysis application posts a publicly visible message (tweet) using the Twitter social networking platform. Users following the MemProtMD Twitter feed can be notified of newly deposited IMP structures, and frequently discuss and share these with other users. In order to create these tweets, the database is queried for new simulation entries every 15 minutes. Each new entry is checked to ensure it is fully available online before a tweet is posted using the Twitter API. A randomly selected pre-rendered image is attached to each tweet.

Containerised application server

The application server is a rack-mounted device which provides the MemProtMD web server and database. Since it is accessible through the internet, it is isolated from all other machines except via an SSH connection from the machine running the analysis application. Data are copied by the analysis application to a RAID array attached to the application server. A Python application updates a MongoDB database when new data becomes available, and a Node.JS web server makes a web application, files and results of database queries available through HTTP requests.

To facilitate rapid deployment on new hardware, the Docker containerisation system is used. Docker containers are similar to virtual machines, but avoid much of the performance overheads encountered in traditional virtualisation by sharing the host machine's kernel. Docker containers are defined programmatically, and can be distributed either as code or as a system image built from that code. Access to the network interfaces, devices and volumes attached to the host machine can be permitted or restricted as required.

The Python management application, Node.JS web server and MongoDB database server are each distributed in an individual docker container. Using docker-compose, management, database and web server images can be built, networked, given appropriate access to the resources on the host machine and started with a single command. All dependency installation is handled by the individual containers, so any host on which Docker has been installed can begin to serve the MemProtMD database and

web application in a matter of seconds. If the host machine is rebooted, Docker images are set up to resume operation as soon as all required volumes are mounted and network interfaces are available.

2.3.11 Web application

The MemProtMD web application provides a user-friendly interface to browse, search and view the results of simulation analyses and ensemble analyses, using the MemProtMD API to load data and query MemProtMD web server. The application stores data locally using the Redux JavaScript library. Data is transformed into an HTML document using the ReactJS Javascript library. In addition to ReactJS, several visualisations use the D3.JS (Data-Driven Documents) [215] JavaScript library where complex geometry or standard chart elements are required. 3D molecular visualisations are drawn using PV [191], a JavaScript and WebGL library.

Source code is developed in JavaScript using the ECMAScript 6 standard, which is packaged into a single file containing highly compressed JavaScript conforming to the ECMAScript 5 standard, which is supported by the majority of modern web browsers. User interface and page changes are handled entirely on the client (unlike traditional web applications where page changes required a new request to be made to the server), and requests are made only when data not found in the local Redux store is required.

In order to quantify usage of the MemProtMD web application, Google Analytics is used to track user interactions. Each action and page view is recorded and can be used to determine how users are engaging with the application.

2.4 Results

The MemProtMD database and web application are available at <http://memprotmd.bioch.ox.ac.uk>, providing an API and GUI which provide access to detailed analysis of simulations of over 3000 IMPs. Since releasing the final version of the MemProtMD web application at the beginning of 2017, MemProtMD has developed a global network of over 900 users from both academic and industrial institutions.

2.4.1 The MemProtMD web application

The MemProtMD web application at <http://memprotmd.bioch.ox.ac.uk> provides a GUI to search and browse data derived from MemProtMD simulations, and is the

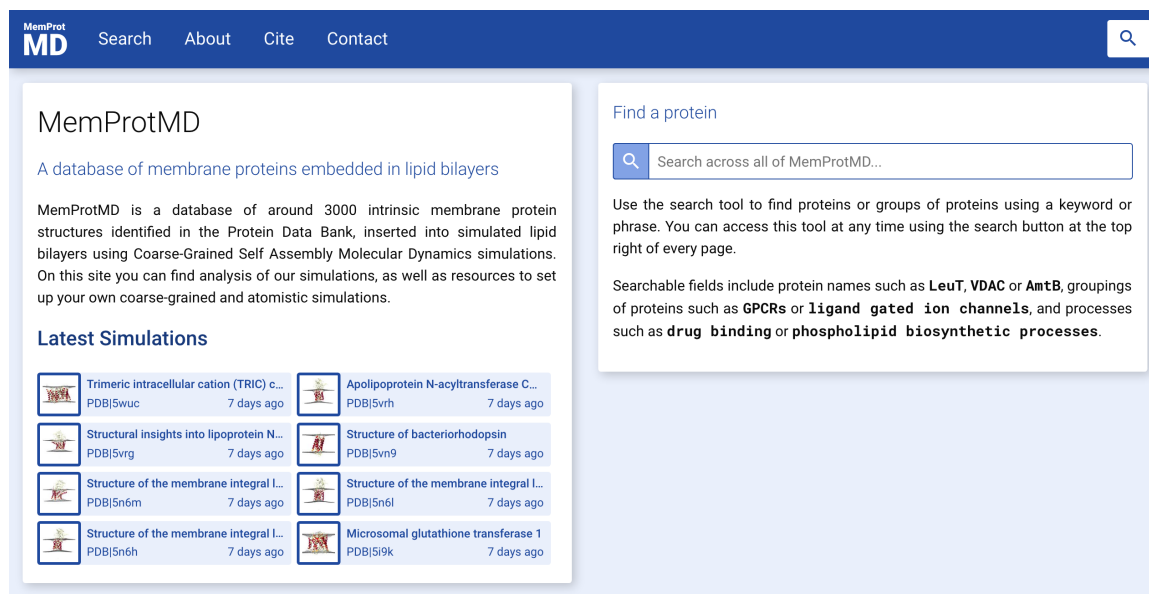


Figure 2.5: Home page of the MemProtMD web application

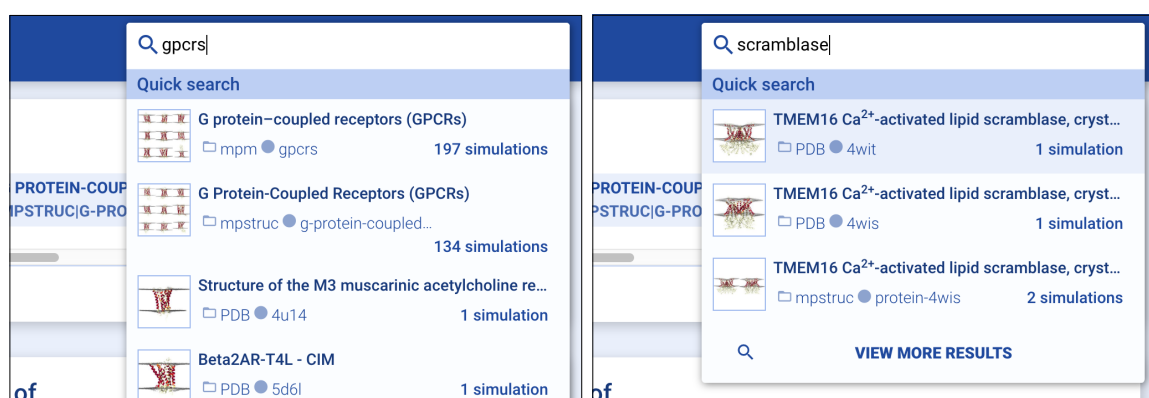


Figure 2.6: Examples of the Quick Search component in use

interface through which most users access the data. The MemProtMD API is used to populate the MemProtMD web application with results of MemProtMD database queries. The home page, shown in Figure 2.5, contains a list of the most recently completed IMP simulations and an in-page search. A set of links can be used to access commonly viewed IMP types, such as GPCRs, ion channels, outer membrane proteins and aquaporins. The title bar at the top of the application is always shown, and provides links to the site home page, search tools, about page, instructions for citing data and contact details. A quick-search button appears on the right-hand side, and expands to provide a search box which displays live results and suggestions, shown in Figure 2.6.

Once a user has selected an external database reference, a new page with further details is shown. As an example, the page for PDB:4wis, a TMEM16 scramblase, can be seen in Figure 2.7. If the reference is part of a tree structure, a panel at the top shows ancestor nodes, and can be expanded to show the current tree in full. This can be seen toward the top of Figure 2.7, where the mpstruc tree is shown. A montage showing data from up to nine referenced simulations (although in the case of PDB:4wis only a single simulation is available) is shown above a summary of metadata, set of links to external sites, and an attached journal article if applicable. A second panel shows links to related references which vary according to the type of reference being displayed, and may show collection members, parent or child nodes or collections to which a reference belongs. Following any of these links will display data for the related reference.

Much of the content displayed on this page has been produced by other authors and is used under a variety of licenses. Where this is the case, it is important that original authors are properly credited, and the terms under which content is licensed are made clear to MemProtMD users. Where content is imported from an external source, full details are attached to database records, including the names of authors, type of license, a link to full terms of the license and a link to the licensed content in its original form. A notice can then be displayed in the MemProtMD web application, giving proper attribution to the original authors.

Visualising simulation data

Each database reference is linked to at least one simulation record, which may be selected to view detailed simulation analysis. If only one simulation is available it is automatically selected, otherwise a simulation can be manually selected. Once a simulation is selected, data is loaded into panels in the page below.

MemProtMD

Search About Cite Contact

PDB | 4wis TMEM16 Ca²⁺-activated lipid scramblase, crystal form 1

TRANSMEMBRANE PROTEINS: ALPHA-HELICAL

MPSTRUC|TRANSMEMBRANE-PROTEINS-ALPHA-HELICAL

1494

TMEM16 FAMILY PROTEINS

MPSTRUC|TMEM16-FAMILY-PROTEINS

2

TMEM16 CA²⁺-ACTIVATED LIPID SCRAMBLASE, CRYSTAL FORM 1

MPSTRUC|PROTEIN-4WIS

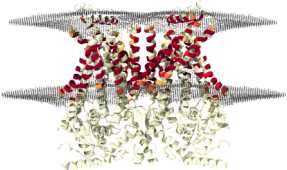
2

TMEM16 CA²⁺-ACTIVATED LIPID SCRAMBLASE, CRYSTAL FORM 1

PDB|4WIS

1

SHOW FULL TREE



PDB | 4wis TMEM16 Ca²⁺-activated lipid scramblase, crystal form 1

Crystal structure of the lipid scramblase nhTMEM16 in crystal form 1

Keywords

membrane protein, transport protein, calcium activation, nech7, lipid transport, c7z7k1

Published

11/12/2014

Resolution

3.3 Å

Method

X-RAY DIFFRACTION

Ligands

CALCIUM ION

Organism

Nectria haematococca (strain 77-13-4 / ATCC MYA-4622 / FGSC 9596 / MPVJ)

VIEW ON PDBe

VIEW ON PDBj

VIEW ON RCSB PDB

Original publication

Title

X-ray structure of a calcium-activated TMEM16 lipid scramblase.

Journal

Nature 2014

DOI

10.1038/nature13984

Authors

Lim, N.K., Duerst, A., Brunner, J.D., Schenck, S., Dutzler, R.

READ ARTICLE FROM NATURE

Member of

UniProt

UNCHARACTERIZED PROTEIN

UNIPROT|C7Z7K1

2

Pfam

CALCIUM-ACTIVATED CHLORIDE CHANNEL

PFAM|PF04547

2

mpstruc

MEMBRANE PROTEINS OF KNOWN 3D STRUCTURE

MPSTRUC|MPSTRUC

1744

TRANSMEMBRANE PROTEINS: ALPHA-HELICAL

MPSTRUC|TRANSMEMBRANE-PROTEINS-ALPHA-HELICAL

1494

TMEM16 FAMILY PROTEINS

MPSTRUC|TMEM16-FAMILY-PROTEINS

2

TMEM16 CA²⁺-ACTIVATED LIPID SCRAMBLASE, CRYSTAL FORM 1

MPSTRUC|PROTEIN-4WIS

2

TCDB

CHANNELS/PORES

TCDB|1

681

A-TYPE CHANNELS

TCDB|1.A

441

VOLTAGE-GATED ION CHANNEL (VIC) SUPERFAMILY

TCDB|1.A.17.1

2

THE CALCIUM-DEPENDENT CHLORIDE CHANNEL (CA-CLC) FAMILY

TCDB|1.A.17

2

mpm

MEMBRANE PROTEINS

MPM|MEMBRANE_PROTEINS

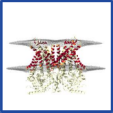
2891

ALPHA HELICAL TRANSMEMBRANE PROTEINS

MPM|ALPHA_HELICAL

2575

Simulations available



4WIS_DEFAULT_DPPC

© Data provided by Research Collaboration for Structural Biology, Lim, N.K., Duerst, A., Brunner, J.D., Schenck, S., Dutzler, R., provided free of copyright via the Protein DataBank. [View original data](#) or [read license terms](#).

Figure 2.7: Example of a PDB reference viewed using the MemProtMD web application

57

Interactive 3D protein visualisation

Interactive 3D visualisations of protein structures can be displayed in any modern web browser using a 3D molecular visualisation component. Users may select from a range of protein and lipid representations through a menu, and the protein can be shaded according to the chemical properties of residues or contacts with various groups of lipid and solvent beads. The orientation of the protein may be interactively changed, or aligned to one of six preset viewing angles. If only part of the protein is selected through the user interface, the colours of the remaining parts of the protein are desaturated to emphasise the selected residues. Some examples of the variety of visualisation options available are shown in Figure 2.8. The PV JavaScript library is used [191] to render molecular data interactively in a web browser without the need for any additional software.

Chain summary view

Protein structures deposited in the PDB are divided into lettered chains, typically assigning one unique, continuous protein sequence to each chain. When viewing a protein in the MemProtMD web application, individual chains can be selected and viewed in more detail using the chain summary tool. Two examples of this can be seen in Figure 2.9. A simplified representation of each chain shows a violin plot of the density of all protein atoms, α -helical protein atoms and β -strand atoms along the membrane normal. This allows chains to be quickly and intuitively identified based on their secondary structure and location relative to the membrane.

Once a chain selected, several attributes of the chain can be seen. The chain is classified as either membrane spanning (where head-groups of both leaflets interact with the chain), membrane-embedded (where the chain interacts with head-groups of one leaflet but only tails of another leaflet), single leaflet (where the chain interacts with a single leaflet), membrane touching (where the chain interacts with lipid head-groups only), or membrane detached (where the chain does not interact with any membrane lipids). The number of transmembrane α -helices and β -strands is shown next, followed by the number of residues making up the chain. A bar appears below, showing the amino acid sequence of the chain coloured using the scheme currently used by the 3D protein viewer, and allowing a subset of residues within the chain to be selected if desired.

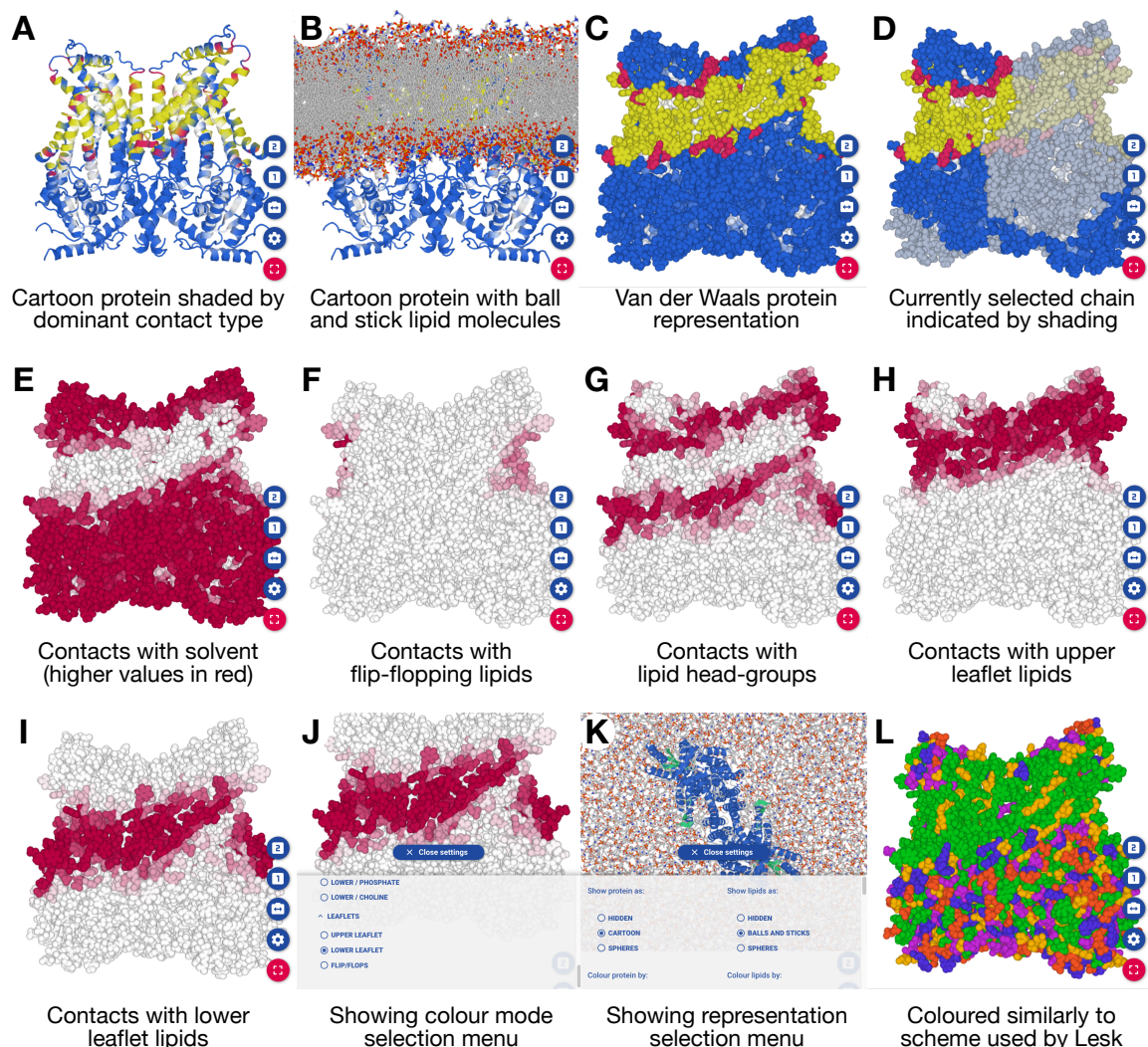
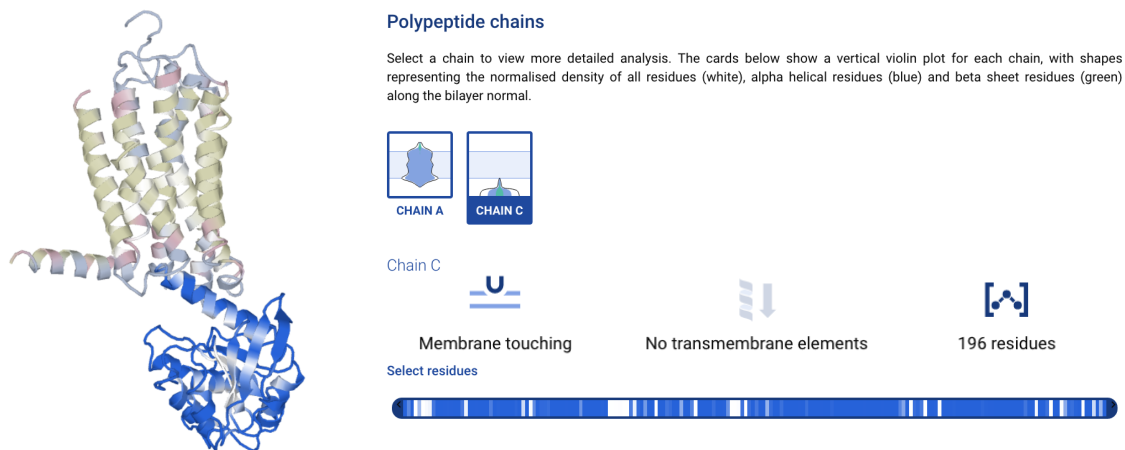


Figure 2.8: Examples of 3D molecular visualisations using the MemProtMD web application

The 3D molecular visualisation web application component is shown in a number of states for a TMEM16 scramblase protein (PDB:4wis). In **A-D** the protein is shown coloured by dominant contact type; blue for contacts with solvent, red for contacts with lipid head-groups and yellow for contacts with lipid tails. **A** shows the protein in cartoon representation; in **B** lipids are added in ball and stick representation. **C** and **D** show the protein in Van der Waals representation. In **D** the protein is shaded to indicate a selected chain.

E-J show the protein shaded according to contacts with a specific group of solvent molecules; **E** - solvent, **F** - flip-flopping lipids, **G** - contacts with lipid head groups, **H** contacts with upper leaflet lipids, **I** - contacts with lower leaflet lipids. The group used for shading can be selected from a menu shown in **J**, and representation can be selected using the menu shown in **K**. Several additional colouring options are available, including the scheme used in Lesk [216], shown in **L**.

PDB | 5g53 A_{2A} adenosine receptor with bound engineered G protein



PDB | 1bcc Cytochrome bc₁

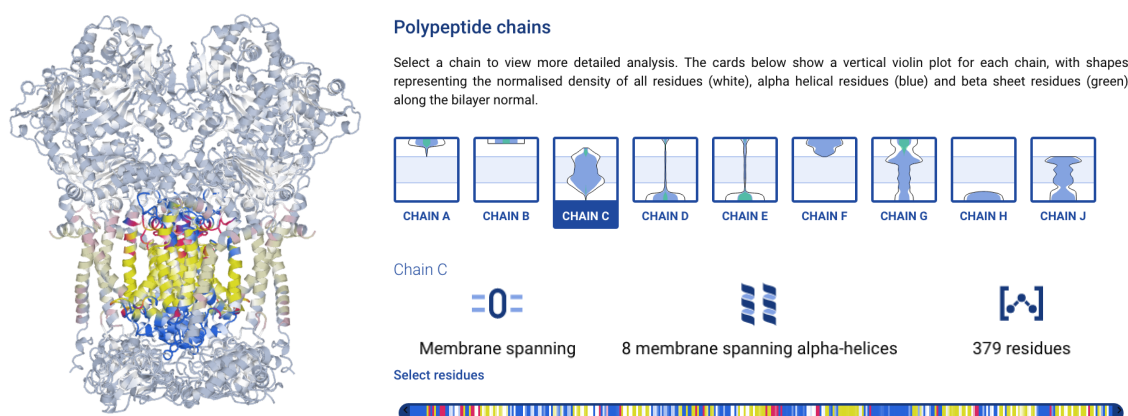
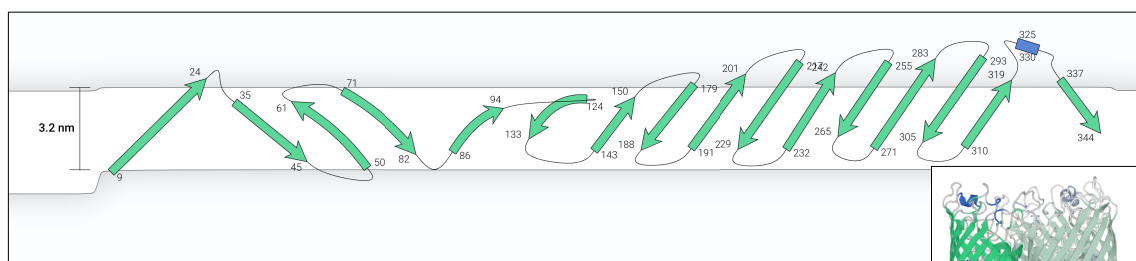


Figure 2.9: Examples of protein chain summaries

Two examples of chain summaries produced by the MemProtMD web application are shown for PDB:5g53 and PDB:1bcc, with one chain selected in each case. On the left, an image is shown indicating the selected chain in a 3D representation of the protein. Each chain is represented by a box showing the lipid bilayer as a rectangle shaded in blue, overlaid with a violin plot showing the distribution of all protein atoms (white envelope), atoms of α -helical residues (blue) and atoms of β -strand residues (green). For the selected chain, several attributes are shown, including whether the chain is embedded in the membrane, the number and type of secondary structure elements, and number of residues. A bar below shows the amino acid sequence, using the same colour scheme as the 3D view. Position and scale of some elements has been altered for clarity and to improve legibility in print.

PDB | 2j1n OmpC Osmoporin



PDB | 5g53 A_{2A} adenosine receptor with bound engineered G protein

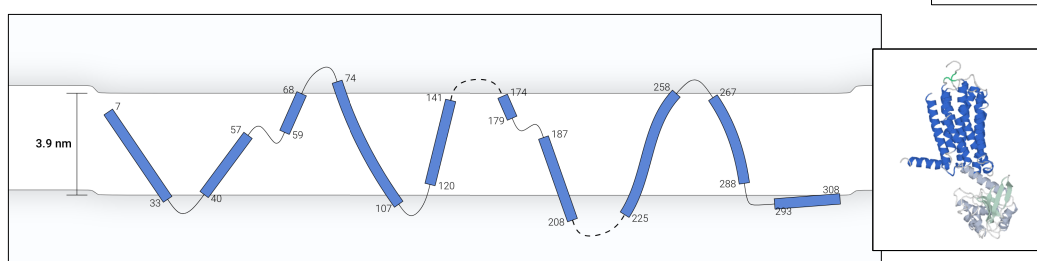


Figure 2.10: Examples of secondary structure topology visualisations

Two examples of secondary structure topology visualisations produced by the MemProtMD web application are shown for PDB:5g53 and PDB:2j1n. α -helices are shown in blue whilst β -strands are shown in green. The curvature and angle of the secondary structure element relative to the membrane normal is preserved. Breaks in the protein sequence are indicated by dashed lines. Start and end residue numbers are shown by small labels. An inset image shows a 3D representation of the same topological features.

Topology visualisation

The MemProtMD web application is able to show protein topology relative to the membrane in a simplified 2D representation, preserving curvature and tilt angles of individual secondary structure elements. Example topologies of a β -barrel membrane protein and α -helical IMP are shown in Figure 2.10. The secondary structure topology plot can be used to quickly identify unusual secondary structure elements, such as kinked or bent α -helices or β -strands with an unusual tilt. The curvature of the sixth helix of the GPCR shown in Figure 2.10 can be easily identified by simple visual inspection. Where β -strands form a β -sheet, their direction is alternated to indicate interactions between adjacent strands. This can be seen in the outer membrane protein in Figure 2.10, where several β -strands interact with adjacent β -strands to form several β -sheets.

Chain A, Residues 430 - 547

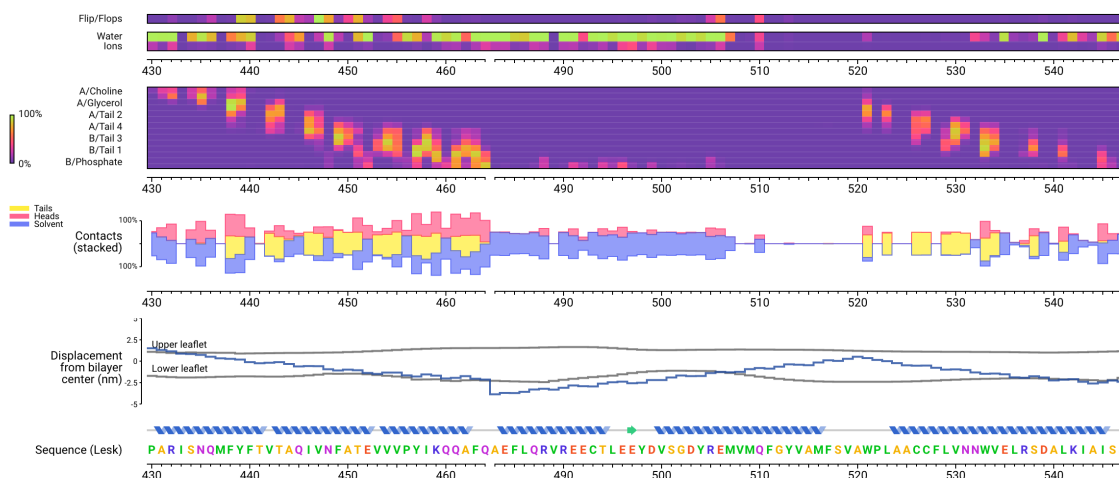


Figure 2.11: MemProtMD sequence viewer

A screenshot of the MemProtMD sequence viewer tool, showing residues 430-547 of PDB:4wis, a TMEM16 scramblase. The amino acid sequence runs left to right, indicated by numbered scale bars. Breaks between amino acids are shown as gaps in scale bars. Several visualisations are shown; from top to bottom: heat map showing contacts with lipids moving between leaflets; heat map showing contacts with water and ions; heat map showing contacts with each lipid bead type, from choline bead of the upper leaflet (top) to choline bead of the lower leaflet (bottom); stacked plot showing contacts with lipid head-groups, lipid tails and solvent; topology plot showing displacement of protein, upper leaflet and lower leaflet relative to the bilayer centre; secondary structure schematic showing alpha helices in blue and beta strands in green; and an amino acid sequence coloured using a scheme found in Lesk [216].

Sequence visualisation

Several metrics are calculated for each amino acid residue in a protein, such as contacts with lipid and solvent beads, position relative to the bilayer centre and secondary structure. Once a protein chain has been selected, these can be viewed using the MemProtMD web application, which displays a number of visualisations along the amino acid sequence of the protein. An example of this can be seen in Figure 2.11, where a short section of the sequence of a TMEM16 scramblase (PDB:4wis) is shown. The sequence viewer allows for easy identification of residues making contact with solvent and lipid molecules, as well as residues contributing to the flipping of lipids between the upper and lower leaflets. Deformations of the lipid bilayer induced by the protein can also be seen, such as the thinning of the membrane induced around the 500th residue.

Visualisations of simulation data produced by the MemProtMD web application

are interactive and connected, and chain or residue selections made using one visualisation will be reflected in other visualisations. Using multiple, connected visualisations of simulation data can help to build a more complete understanding of the structural basis for a biochemical phenomenon. An example of this can be seen in Figure 2.12, where several residues interacting with lipid molecules flipping between leaflets are selected using the sequence viewer. These residues are mostly α -helical, and when viewed using the topology viewer can be seen to be broken into three α -helices with increasingly steep angles relative to the membrane normal. The same residues can then be seen in a 3D cartoon view of the entire protein, coloured according to contacts with lipids moving from the upper to lower leaflet, and clearly identifying a groove between two alpha-helices which makes frequent contacts with lipids moving between leaflets.

Search facilities

The MemProtMD web application provides several options to search for specific proteins or groups of proteins amongst the over 8000 external database references in the MemProtMD database. The Quick search tool, shown in Figure 2.6, is always shown at the top of the MemProtMD web application and is a simple and intuitive way to find a protein of interest using text matching. The more advanced text search tool, accessed through the quick search tool and shown in Figure 2.13, shows more results and allows results to be filtered to include or exclude specific external databases.

An advanced search tool is also provided for users needing to perform more complex searches. This provides an online editor which can be used to construct a MongoDB query which can then be executed directly through the MemProtMD API. Queries have full access to all attributes stored in MongoDB documents, including membrane thickness, external annotations and numbers of simulations. The query can be saved for later use, downloaded as a Python, Bash or JavaScript script, or be performed directly in the MemProtMD web application, with results shown as a table, JSON document or CSV document.

Structure author profiles

Each protein structure deposited in the PDB records the authors responsible for determining the protein structure. The MemProtMD web application provides a profile for each structure author, including a timeline of recently published structures

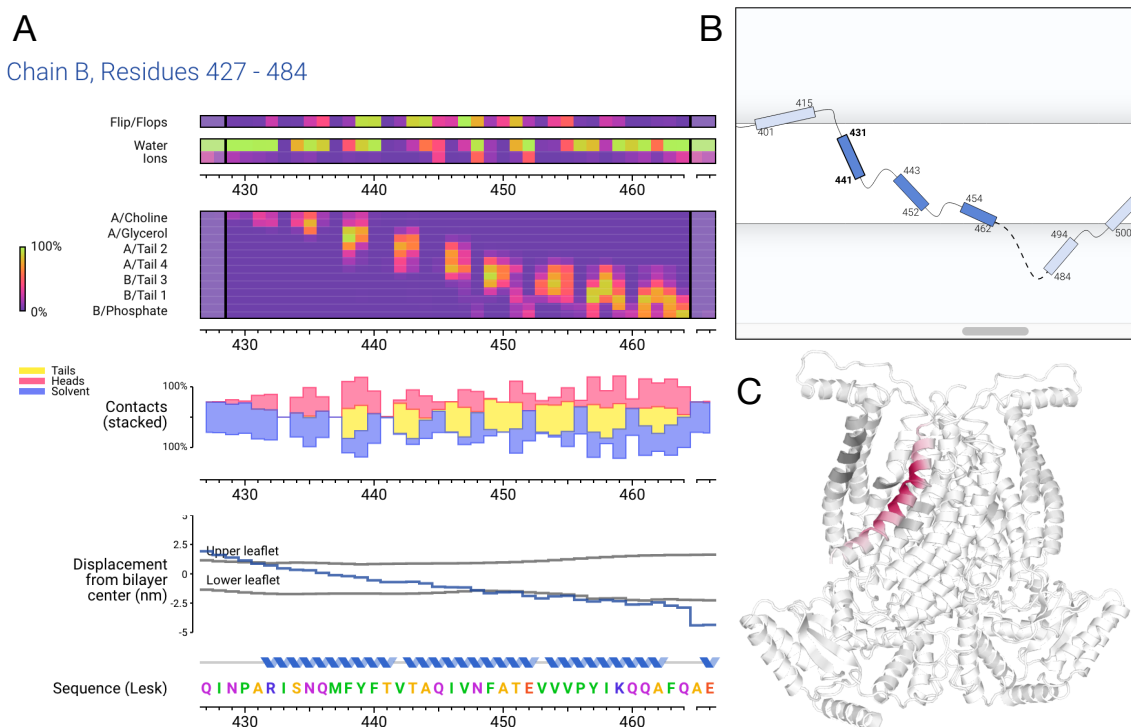


Figure 2.12: A sub-chain selection across multiple connected visualisations
Three visualisations produced by the MemProtMD web application are shown, with residues 429-464 of chain B of a TMEM16 scramblase (PDB:4wis) selected. In **A**, the sequence viewer can be seen, with selected residues indicated by vertical black lines and shading. In **B**, the topology of the protein is shown, with the selected residues indicated by darker shading. In **C**, the protein is shown in a 3D cartoon representation, shaded according to contacts with flipping lipid molecules from dark (many contacts) to light (few contacts). Selected residues are coloured red. This selection can be viewed online at http://memprotmd.bioch.ox.ac.uk/_ref/PDB/4wis/_sim/4wis_default_dppc/Chain.B/Residue.429-464/.

MemProtMD
Search
About
Cite
Contact

Text Search

This tool can be used to search for records across MemProtMD which match specific words or phrases. Currently text search works on whole words only - **lipo** will not match **lipoprotein**. You can match an entire phrase by putting it "in quotes", or exclude certain terms by using the **-negation** **-symbol**. Metadata used for the search is sourced from the MemProtMD classification system as well as six external sources.

Sources

☐ PDB
☐ UNIPROT
☐ MPSTRUC
☐ PFAM
☒ GO
☐ TCDB
☒ MEMPROTMD

Click to include or exclude a data source. More information can be viewed by hovering over the database name (touch users may need to tap and hold). **These rules will be applied next time you click the search button.**

Searching for **lipid** in GO and mpm.

	Resolution	Organism	Published	Method	Simulations
	Response to lipid				7
	http://memprotmd.bioch.ox.ac.uk/_ref/GO/GO:0033993/				
	Lipid catabolic process				11
	http://memprotmd.bioch.ox.ac.uk/_ref/GO/GO:0016042/				
	Lipid A biosynthetic process				4
	http://memprotmd.bioch.ox.ac.uk/_ref/GO/GO:0009245/				
	Lipid A metabolic process				1
	http://memprotmd.bioch.ox.ac.uk/_ref/GO/GO:0046493/				
	Negative regulation of lipid storage				6
	http://memprotmd.bioch.ox.ac.uk/_ref/GO/GO:0010888/				

Figure 2.13: MemProtMD text search tool

A search for records matching the word "lipid" is shown, with results filtered to include only GO records and MemProtMD records.

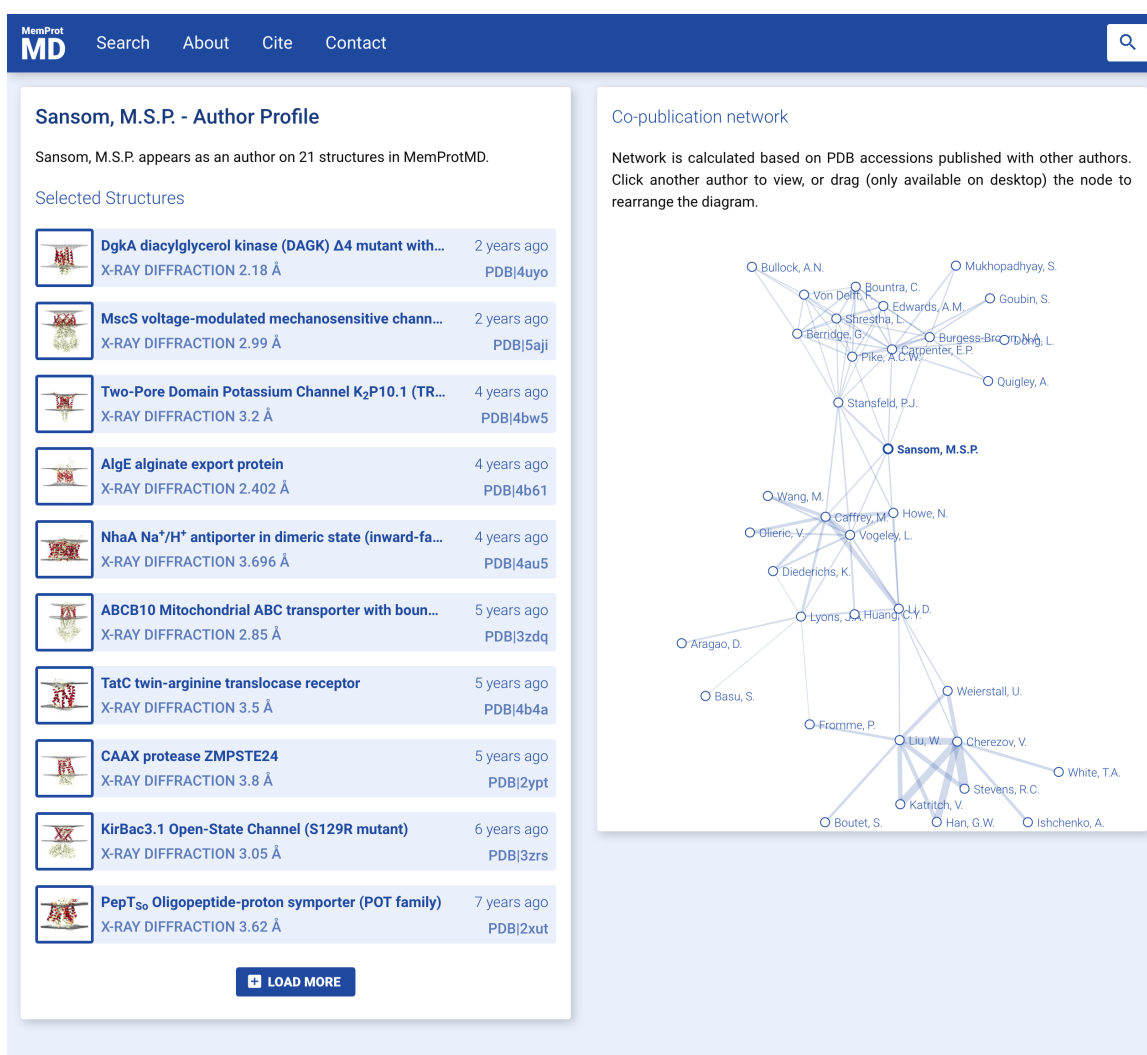


Figure 2.14: MemProtMD author profile

An example of a MemProtMD author profile page for Mark S.P. Sansom. The left-hand panel shows a timeline of recently published structures, whilst the right-hand panel shows a graph of frequent co-authors.

and a co-publication diagram showing frequent collaborators. An example author profile can be seen in Figure 2.14.

The co-publication diagram shows a graph with a node for each author and edges scored using a matrix containing the number of times any two authors have published a structure together. The layout is optimised using a Verlet integrator implemented in D3.js [215]. A repulsive electrostatic force is applied between nodes, and an attractive force along edges, proportional to the number of co-authored publishers. After a few seconds of simulation, the graph layout reveals clusters of frequent collaborators, which can be clicked to view a different author.

Usage statistics

The MemProtMD web application has a growing international network of both academic and industrial users. The global distribution of users, and variety of institutions making regular use of MemProtMD resources is shown in Figure 2.15. As of July 2017, over 900 unique users have viewed pages on the MemProtMD web application, of which 56% have been retained as regular users. Over 10% of users during this period accessed the site through a smartphone or tablet device. On average, between 90 and 120 usage sessions, each lasting an average of 7 minutes, are recorded each week.

Through a semi-automated Twitter account, MemProtMD is able to reach 350 followers, with 25% of users finding the MemProtMD web application through social media. A search-engine optimised page design has resulted in another 25% of users acquired from search engines. Remaining traffic comes either through direct entry of a Uniform Resource Identifier (URI) (40%) or through links from other online resources (10%).

2.4.2 Ensemble and collection analyses

Counting transmembrane database accessions

As structure determination methods continually improve, IMP structures are being determined at an increasing rate. To quantify this, the MemProtMD pipeline determines the number of structures determined for each month since 1985, when the first IMP structure was determined. As can be seen in Figure 2.16, the number of IMPs with determined structures continues to rise at a near-exponential rate. Represented within these structures are an increasing number of unique proteins and protein families, although these are increasing less rapidly. As of July 2017, the MemProtMD

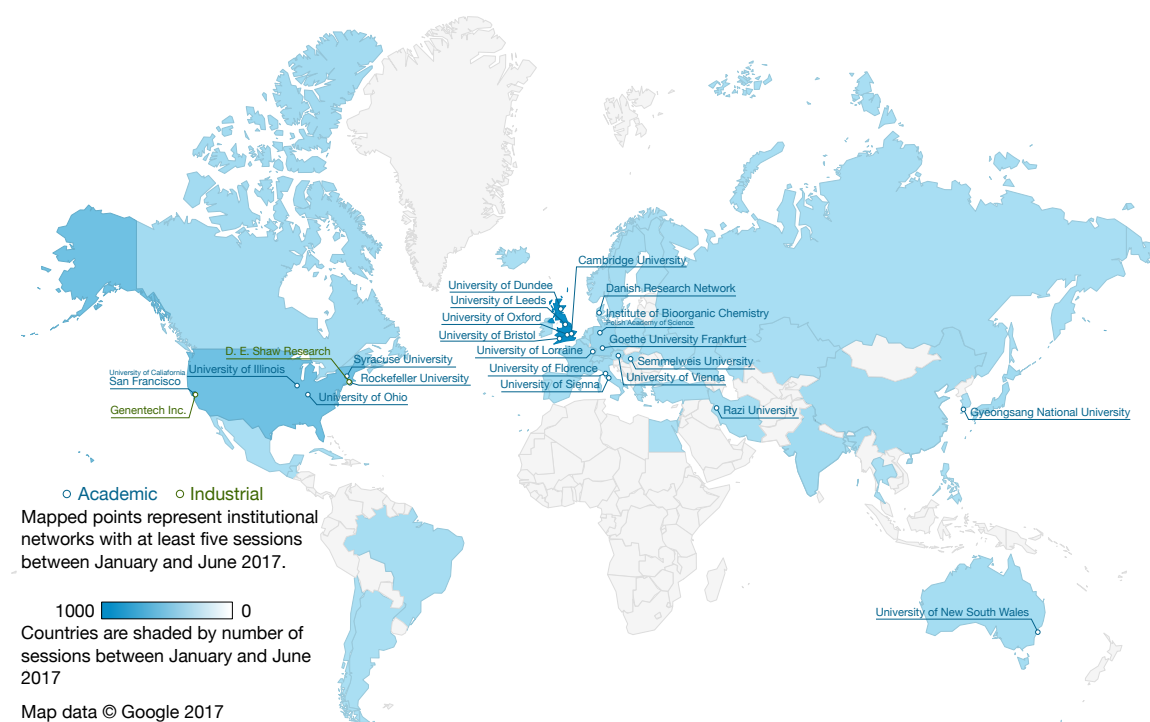


Figure 2.15: Global usage of the MemProtMD web application

Countries are shown shaded from white to blue according to the number of web application sessions detected by Google Analytics between January and June 2017. Institutional networks making at least five independent sessions are mapped, showing the variety of institutional and academic organisations making regular use of the MemProtMD web application.

Map data is supplied by Google, ©Alphabet 2017.

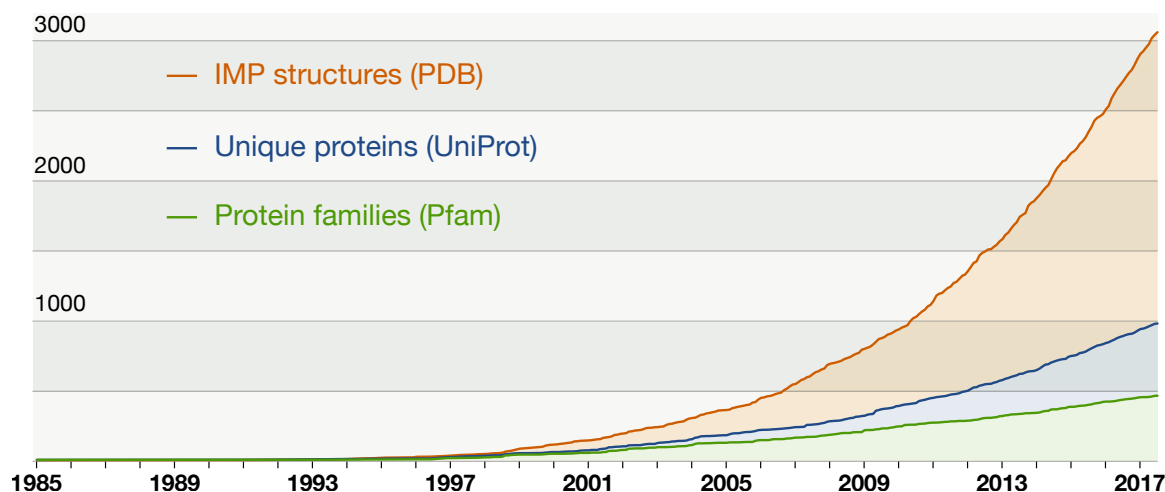


Figure 2.16: Growth in numbers of IMP structures since 1985

The cumulative number of accessions included in the MemProtMD database is plotted for each month since 1985 for PDB, UniProt and Pfam databases. Sustained growth can be seen in the number of IMP structures, as well as the number of unique IMPs and IMP families for which at least one structure has been determined.

UniProt and Pfam accessions which do not make contact with membrane lipid tails are excluded.

Figure is based on an automatically generated graphic at <http://memprotmd.bioch.ox.ac.uk/stats> and is extended to dates prior to the establishment of the PDB.

database contains 3056 IMP structures, 977 IMP unique protein sequences, and 461 IMP families. On average, each unique protein sequence is represented by 3.1 structures, and each protein family is represented by 4.6 structures.

Represented amongst the transmembrane regions of MemProtMD proteins are 28 unique protein folds identified by the SCOP database [217]. Most common amongst these are single transmembrane helices (SCOP f.23), which account for 35% of simulated proteins, followed by GPCRs (SCOP f.13) at 13% and β -barrel proteins at 10%. 90% of MemProtMD proteins are represented by the ten most common folds.

It can be seen in Figure 2.17 that X-ray crystallography remains the predominant experimental method used for IMP structure determination, being used to determine 87% of IMP structures between 1989 and 2016. Solution Nuclear Magnetic Resonance (NMR) is the second most widely used method, accounting for just 6% of solved structures. Since 2011, the number of structures solved using electron microscopy (including Cryo-Electron Microscopy (Cryo-EM)) has risen sharply, from 1.5% of IMP structures (3 structures) in 2011 to 11% (42 structures) in 2016.

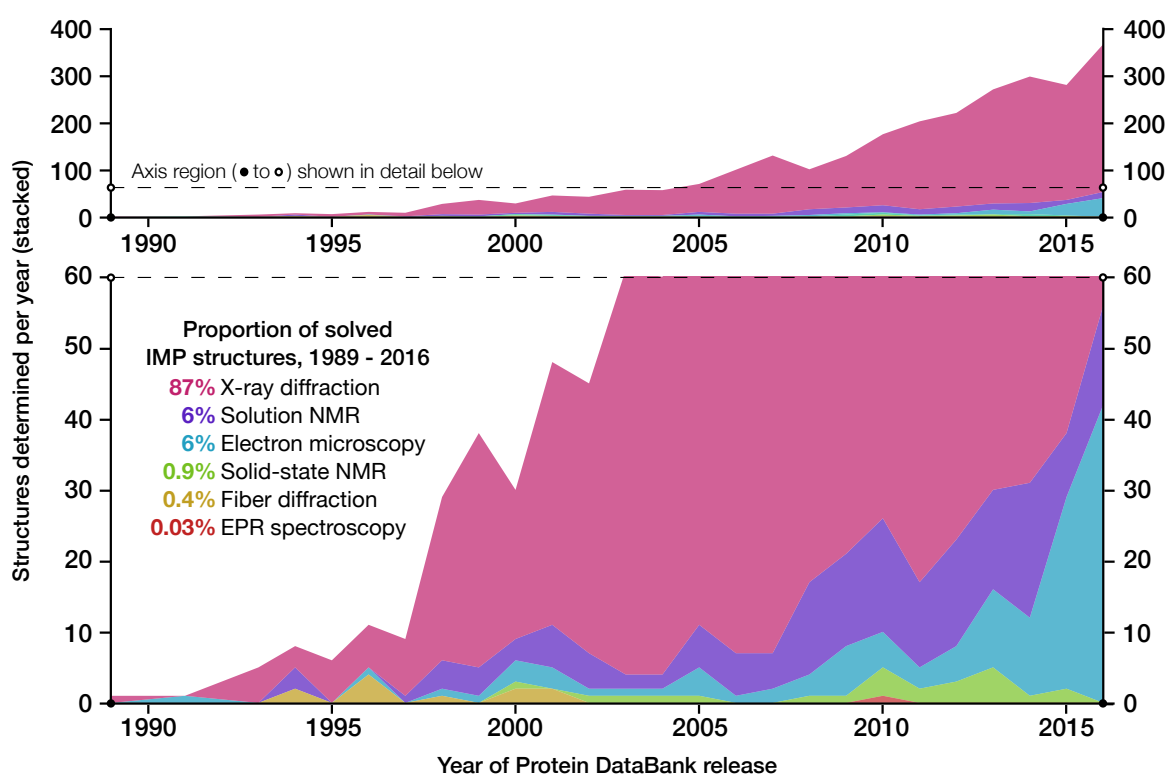


Figure 2.17: Experimental methods used to determine IMP structures
Stacked area plot showing the number of IMP structures deposited annually in the PDB. To better show less frequently used methods, the plot is shown once (above) with the full y-axis range and once (below) with limited y-axis range. The percentages of structures published before 2016 using each method are also tabulated.

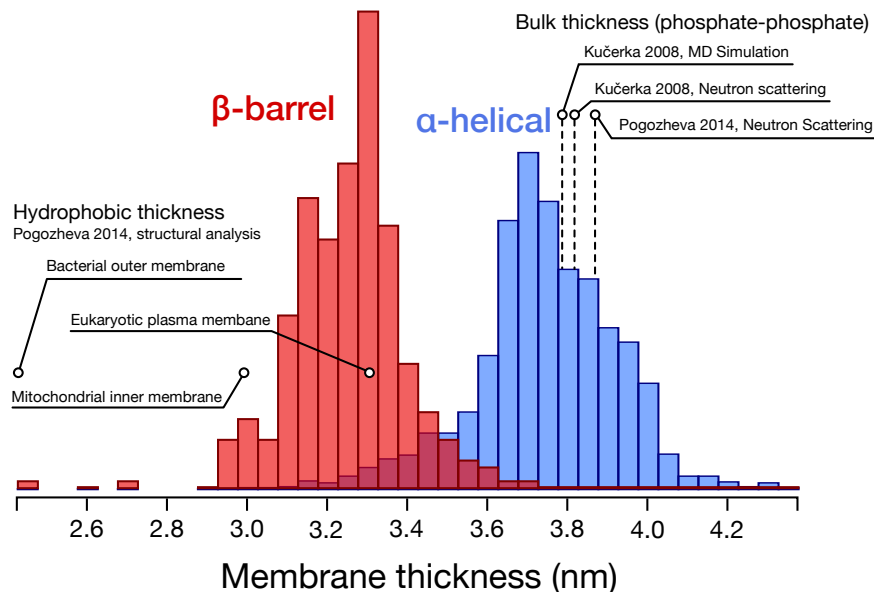


Figure 2.18: Histogram showing resulting membrane thickness from simulations of α -helical and β -barrel IMP structures

The thickness of membranes formed during simulations of 2592 α -helical IMP structures and 340 β -barrel IMP structures, as measured by the distance between phosphate beads of annular lipids along the membrane normal. For comparison, membrane thickness measured as phosphate-phosphate distance and hydrophobic thickness by other experimental and simulation studies is also shown. Hydrophobic thickness is generally 0.5 nm less than phosphate-phosphate distance.

Thickness of simulated membranes around α -helical and β -barrel proteins

Physiological membrane thickness varies greatly both within and between different membranes. The bacterial outer membrane contains mostly β -barrel proteins, and has a considerably thinner hydrophobic core than other membrane types. A histogram showing the distribution of membrane thickness, measured as distance between annular phosphates of upper and lower leaflets, separately for α -helical and β -barrel proteins is shown in Figure 2.18. A pronounced difference can be seen between simulations of α -helical IMP structures, mostly forming membranes between 3.6 nm and 4.0 nm in thickness, and simulations of β -barrel IMP structures, mostly forming membranes between 3.1 nm and 3.4 nm in thickness.

Per-residue contact scales

To examine the different chemical properties of amino acids in the membrane environment, contact frequencies are aggregated and normalised between all simulations in the MemProtMD database. The results are shown in Figure 2.19. In 2.19A, contacts are shown grouped by parts of the membrane; lipid tails, lipid head-groups and the solvent, and amino acids are sorted according to contacts with lipid tails. Although the frequency of amino acids contact with lipid tail beads inversely correlates to some degree with frequency of water contacts, the relationship is not a simple inverse, with amphipathic aromatic residues tyrosine and tryptophan forming more contacts with water beads than would be expected from lipid tail contacts alone. This suggests that whilst hydropathy plays a significant role in determining interactions between amino acids and lipid molecules, other attributes must be taken into account for a complete understanding of interactions between residues of IMPs and membrane lipids.

A more detailed breakdown of protein-lipid and protein-solvent interactions can be seen in 2.19B. The hydrophobic residues phenylalanine, isoleucine, leucine can clearly be seen making more contacts with the core of the membrane, represented by the Tail 3 and Tail 4 labels. Moving from Tail 2 to Glycerol, aromatic residues tyrosine and tryptophan become increasingly enriched, whilst from the Phosphate to Water label more charged, hydrophilic residues such as arginine, lysine, glutamate and aspartate make more contacts.

Occurrence of amino acid types along the membrane normal

The effects of chemical properties of different amino acids can also be shown by changes in frequency of amino acids along the membrane normal. This is calculated across all proteins in the MemProtMD database and shown in Figure 2.20. Here, the hydrophobic residues phenylalanine, isoleucine and leucine can be seen strongly associated with the hydrophobic core of the membrane, with a sharp cutoff occurring at the membrane interface. The interface association of aromatic residues tryptophan and tyrosine can be seen with clear peaks at the edge of the hydrophobic core. Closer to the solvent, peaks can be seen for positively and negatively charged residues. Although a slightly elevated peak can be seen for positive residues around the inner leaflet, seeming to support the positive inside rule, this is likely due to the use of the rule to provide an initial orientation and so cannot be taken as independent evidence of this rule.

A	ASP	GLU	ARG	ASN	LYS	GLN	SER	GLY	HIS	THR	ALA	PRO	CYS	MET	TYR	VAL	ILE	LEU	PHE	TRP
Solvent	85	85	92	65	100	72	31	0	70	23	-23	45	-34	-3	35	-25	-28	-25	-7	29
Head	28	23	65	28	66	37	20	0	56	17	-11	29	-14	12	58	1	5	9	40	100
Tail	-45	-42	-27	-27	-26	-23	-5	0	2	7	17	22	35	35	49	63	73	79	93	100

B	ASP	GLU	ARG	LYS	ASN	GLN	SER	GLY	HIS	THR	ALA	PRO	CYS	MET	TYR	VAL	ILE	LEU	PHE	TRP
Chloride	0	3	95	100	20	29	14	0	33	9	-5	8	-6	2	9	-6	-6	-6	-2	5
Sodium	97	100	4	6	22	28	14	0	38	11	-7	9	-10	-3	8	-8	-9	-8	-3	5
Water	85	84	91	100	65	72	31	0	69	23	-22	45	-33	-3	34	-24	-27	-25	-6	28
Choline	98	89	66	61	53	76	30	0	100	19	-15	30	-21	4	52	-17	-16	-14	14	85
Phosphate	11	7	100	100	34	42	23	0	63	17	-13	16	-14	8	40	-12	-11	-10	11	71
Glycerol	-5	-6	46	44	11	15	13	0	39	15	-4	22	-1	18	57	12	17	20	48	100
Tail 1	-16	-15	-1	1	-4	0	5	0	21	10	5	27	14	26	56	34	41	45	70	100
Tail 2	-19	-19	-15	-14	-13	-11	-1	0	4	5	8	22	24	25	47	50	58	62	94	100
Tail 3	-23	-22	-21	-20	-17	-16	-3	0	-5	4	10	14	31	21	32	56	63	68	100	77
Tail 4	-29	-28	-27	-27	-22	-21	-5	0	-9	4	13	10	36	22	27	60	68	73	100	67

C	ASP	GLU	ARG	LYS	ASN	GLN	SER	GLY	HIS	THR	ALA	PRO	CYS	MET	TYR	VAL	ILE	LEU	PHE	TRP
Monera 1999	-55	64	-14	-23	-28	-10	-5	0	-31	13	41	*	63	74	13	63	99	97	100	97
Moon 2011	3.0	1.6	3.7	5.4	3.5	3.0	1.8	1.7	4.8	1.8	0.0	-1.5	0.5	-0.8	-1.1	-0.8	-1.6	-1.8	-2.2	-0.4

* Proline data unavailable for pH 7

Figure 2.19: Per-residue contact scales derived from MemProtMD
 Contacts between each amino acid type and selected components of a DPPC bilayer are shown, linearly normalised such that Glycine receives a value of 0 and the maximum value becomes 100. Table cells are shaded from blue (-100) to red (100). In **A**, contacts are shown for all solvent beads, all head-group beads and all tail beads, whilst in **B**, contacts are further broken down to show individual bead types. Experimental scales reported in Monera [213] and Moon [218] are shown below in panel **C**.

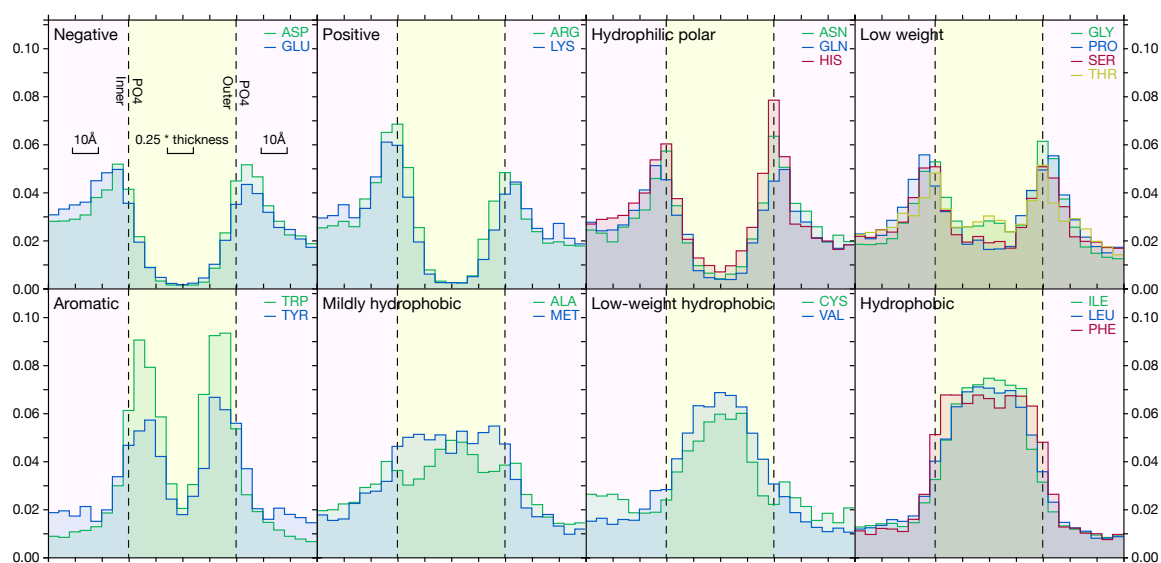


Figure 2.20: Histograms showing occurrence of amino acid types along the membrane normal

Histograms show the frequency of each amino acid with respect to position along the membrane normal. Amino acids are clustered into eight groups based on similarity. The transmembrane region is shaded in yellow, with the mean position of annular phosphate groups represented by dashed lines. In order to account for variations in membrane thickness and allow frequencies to be aggregated, the membrane normal coordinate is adjusted to ensure phosphate beads are always separated by 4 nm.

Aggregation of contacts through sequence homology

The relationships between the sequence of a protein and its structural and functional properties are often highly conserved among homologous proteins. Where multiple simulations exist for a unique protein (as assigned by UniProt) or family of proteins (as assigned by Pfam), sequence homology can be used to compare the results of simulation analyses between structures. The MemProtMD database contains MSAs for each protein or family of proteins represented by two or more structures. Figure 2.21 shows one such alignment produced for the Pfam classification of GPCRs. A high level of conservation can be seen in both the contacts formed between the protein and different parts of the membrane, and in the secondary structure of GPCRs, with the characteristic seven transmembrane helices clearly visible.

Another example can be seen in Figure 2.22, where contacts from simulations of 53 β -barrel Gram-negative porin proteins can be seen. In addition to contacts with lipid head-groups and tails, contacts with water molecules can also be seen. The alignment shows a secondary structure composed of highly conserved transmembrane β -strands, as well as some less conserved α -helices. Contacts with solvent and lipid beads are again shown to be highly conserved between protein family members.

2.4.3 The MemProtMD API

Accessing the MemProtMD database

Data generated from the MemProtMD pipeline and subsequent analyses are stored in the MemProtMD database, which consists of an indexed MongoDB database and a collection of binary files stored on disk. The MemProtMD API (<http://memprotmd.bioch.ox.ac.uk/api/>) allows various applications to query the database and retrieve data from either files or MongoDB documents. Database queries can range from retrieving data about a single simulation, an external database reference such as a UniProt identifier or a particular structure author to searching for simulations matching a set of complex criteria including membrane thickness, size of protein or text search.

Usage statistics

Between January 2017 and June 2017, over 2000 requests were made to the MemProtMD API, not including those made by the MemProtMD web application. API

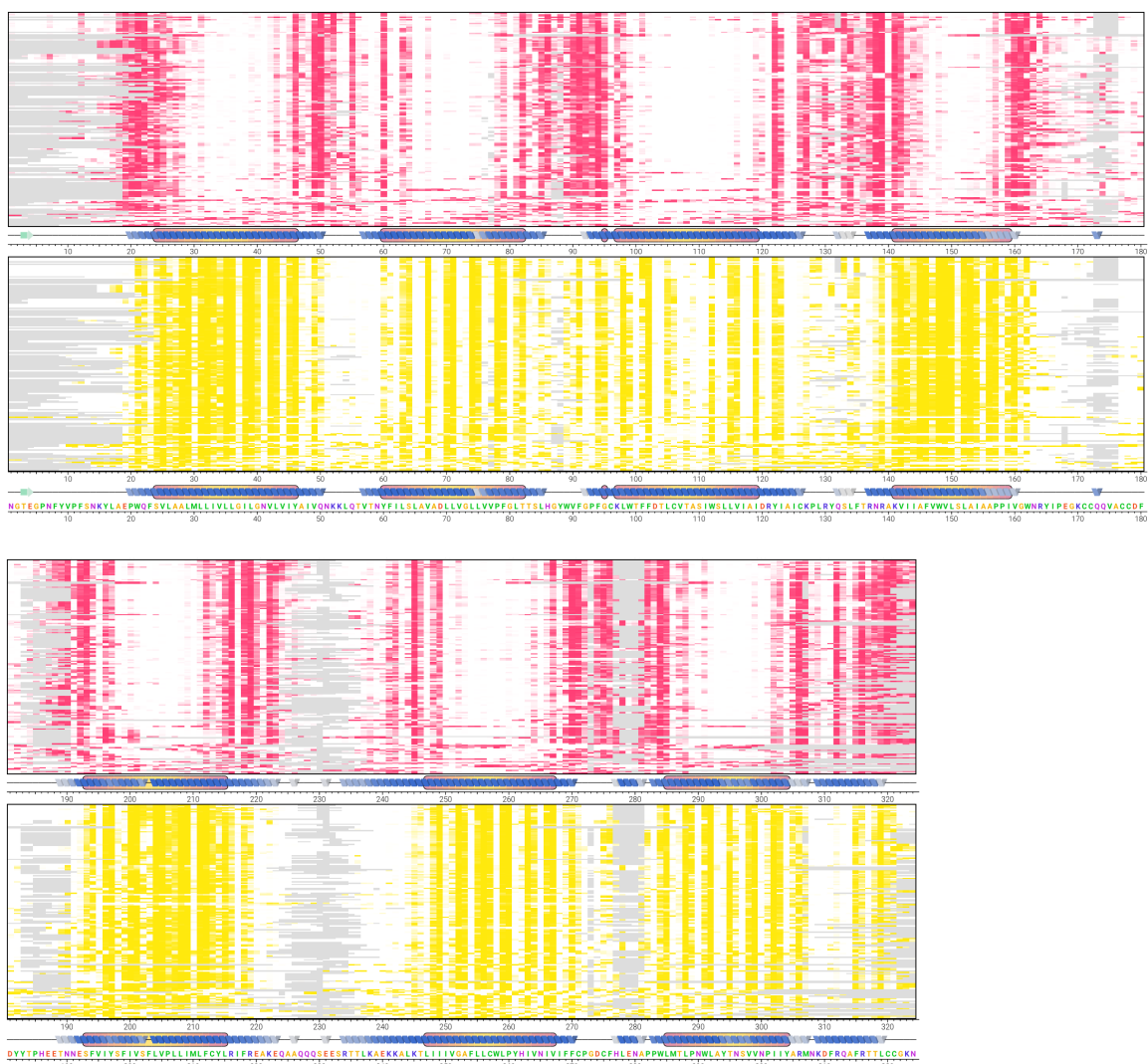


Figure 2.21: Multi-sequence alignment of all GPCR structures in the MemProtMD database

Sequences from 172 simulations of GPCR structures are aligned, and contacts with lipid head-groups (red) and lipid tails (yellow) are plotted for each aligned residue. Grey cells indicate gaps in the multi-sequence alignment. A representation of secondary structure is shown below, coloured darker where secondary structure is more conserved. Transmembrane regions are contained in boxes, shaded from red (less buried) to yellow (more buried). A consensus sequence is shown below.



Figure 2.22: Multi-sequence alignment of all Gram-negative porin structures in the MemProtMD database

Sequences from 53 simulations of Gram-negative porin structures are aligned. Contacts with water (blue), lipid head-groups (red) and lipid tails (yellow) are shown. Secondary structure plots show beta strands (green) and alpha helices (blue) coloured darker where secondary structure is more conserved. Transmembrane regions are indicated by a box, shaded from red (less buried) to yellow (more buried), and a consensus sequence is shown below.

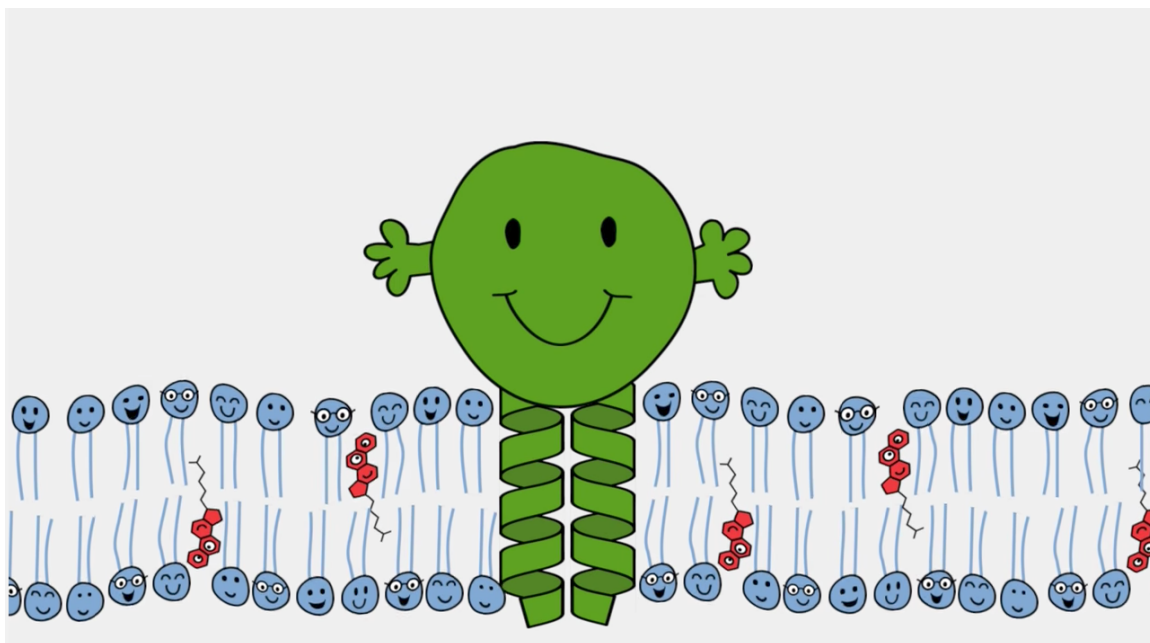


Figure 2.23: Still from Mr Membrane Protein animation

A still from the animated video Mr Membrane Protein (<https://www.youtube.com/watch?v=UTSXEMiA-9s>), showing an abstract representation of an IMP in a mixed-species bilayer containing cholesterol.

usage is currently sporadic and consists of large, infrequent batch operations, often performing around 100 requests in under an hour.

2.4.4 The Story of Mr Membrane Protein

Mr Membrane Protein (<https://www.youtube.com/watch?v=UTSXEMiA-9s>) is a short video produced to explain the motivation behind MemProtMD in an accessible and entertaining style. The video uses a mixture of 2D and 3D animation techniques to introduce an α -helical IMP and outlines the processes of X-ray crystallography and CGSA simulation, in a style similar to the Roger Hargreaves "Mr Men" series. The video has received over 3000 views since its publication in June 2015. A still from Mr Membrane Protein is shown in Figure 2.23

2.5 Conclusions

The MemProtMD database and web application is now established as a vital resource for IMP structural biology, contextualising an unprecedented number of experimentally derived IMP structures in a realistic bilayer environment, and providing access to numerous analysis tools which can be used to address a range of scientific questions.

Through an intuitive and powerful user interface, the MemProtMD web application can be used to find IMPs of interest, examine protein-lipid interactions and view lipid bilayer structure using an array of novel visualisation tools. Outside of the MemProtMD web application, data held in the MemProtMD Database has numerous applications, and considerable amounts of data have been accessed through the MemProtMD API.

Since Stephen White’s initial observations in 2004 [193], the exponential rise in the number of IMP structures has been sustained and, as seen in Figure 2.16, has been accompanied by a steady increase in the number of unique proteins and protein families for which structures exist. Although X-ray crystallography remains the dominant method for structure determination, rapid growth in the number of Cryo-EM structures seen in Figure 2.17 is likely to continue, producing an increasing number of structures for large IMP complexes. Solution and solid-state NMR remain important methods for IMP structure determination, especially when addressing questions of protein dynamics, however use of both methods is growing more slowly than either X-ray crystallography or Cryo-EM.

As the amount and diversity of structural IMP data increases, the MemProtMD pipeline, database and web application will continue to provide analyses of individual IMPs in lipid bilayers, whilst also providing constantly updating analyses of families of membrane proteins, as well as of membrane proteins as a whole. Automatically created sequence alignments for protein families reveal conserved protein-lipid interactions as well as protein topology, and this can be seen in Figure 2.22 for Gram-negative porins, and in Figure 2.21 for GPCRs, with both families forming sets of highly conserved contacts with membrane lipids, similar to the highly conserved lipid contacts observed in simulations of aquaporins [204].

Ensemble analyses of both protein-lipid interactions and protein structures reveal properties of amino acids which support previously published studies (e.g. [105, 108, 109, 110]). The affinity of each type of amino acid for different parts of the lipid bilayer cannot be explained by a simple hydropathy gradient, and this can be seen clearly in Figure 2.19, where amino acids are scored by likelihood of contacting each part of the membrane. Tail 3 and Tail 4, representing the bilayer core, mostly form contacts with phenylalanine, leucine and isoleucine residues. Moving from Tail 2 to the glycerol bead (representing lipid carbonyls) aromatic residues tyrosine and tryptophan form an increasing number of contacts, after which charged and polar residues form the majority of contacts with the choline and solvent beads.

Figure 2.20 more clearly shows the spatial distribution of these residues, revealing some more subtle trends. Compared to aspartate and glutamate, arginine and lysine form peaks closer to the bilayer core, suggesting that larger side-chains are enabling arginine and lysine snorkelling away from the bilayer core [105]. The flatter, wider peak seen in the distribution of phenylalanine may also point to anti-snorkelling of phenylalanine side-chains towards the membrane core [106]. The distribution of arginine and lysine would appear to support a positive-inside rule, however, since this rule is used to provide initial orientations of proteins, this cannot be taken independently as evidence for a positive-inside rule.

Annular membrane thickness varies greatly between different proteins and membrane types, and in Figure 2.18 this variation can be clearly seen for both α -helical and β -barrel IMPs. The annular shell of lipids around a β -barrel protein typically forms a membrane around 0.55 nm thinner than that of an α -helical membrane protein, having a median thickness of 3.25 and 3.8 nm respectively. This variation is expected given the lesser hydrophobic thickness of bacterial and mitochondrial outer membranes in which β -barrel proteins mostly reside. Although these values are similar to those previously reported in experimental [194] and computational [110] studies, the median thickness of an α -helical IMP is around 0.2 nm thinner than that of an idealised DPPC bilayer, likely due to slight bilayer thinning induced by the protein. Differences in thickness between α -helical and β -barrel proteins are smaller than might be expected from measures of hydrophobic thickness, and this may be due to hydrophobic mismatch of DPPC lipid tails compared to shorter tails seen in the bacterial outer membrane.

Following the success of the MemProtMD pipeline, database and web application, several expansions and improvements have been proposed for future development. At present, the majority of data is drawn from a 1000 ns coarse-grained simulations in a pure DPPC bilayer, which in future will be supplemented by coarse-grained simulations in a mixed-species bilayer, with a more physiologically accurate composition selected to best match the protein’s native bilayer. With increased computational power, it may be possible to perform longer atomistic simulations, potentially including binding studies of some relevant ligands.

The analyses performed on IMPs by the MemProtMD analysis application are currently designed to be applicable to all IMPs, however, several analysis approaches exist which are particularly informative for specific classes of IMPs. Channels and pores contain a central cavity through which water and other molecules may pass, and tools such as HOLE [219] can be used to examine the residues forming this cavity

and identify gates and bottlenecks. Where appropriate, integration of the findings of such tools into the MemProtMD database and web application would give a more complete overview of the many structures of channels and pores currently available. Additionally, pipelines for simulation and analysis of several types membrane associated protein, including peripheral membrane proteins and lipoproteins, could be integrated into MemProtMD using existing infrastructure and software.

A great deal of effort in setting up the MemProtMD pipeline, analysis application, database and web application has been spent designing systems which compensate for erroneous, inconsistent or incomplete data held in external databases, and several of these are yet to be addressed. Many examples exist in PDB depositions of incorrect author-submitted biological assemblies, some with incorrect oligomeric states and some with subunits arranged according to biologically unrealistic crystal packing. Metadata is also frequently incomplete or inadequate; IMP structures in the PDB are often not labelled as such, and in the case of most IMPs it is difficult to automatically determine native membrane type from either PDB or UniProt data.

The enthusiastic reception received so far by the MemProtMD database and web application demonstrates the increasing popularity and utility of molecular simulation as a tool to understand the structure and function of IMPs. With an ever-increasing number of new IMP structures published each year, MemProtMD becomes an increasingly important resource by providing instant access to data derived from individual simulations and ensemble analyses, as well as encouraging further multi-scale simulation work through provision of IMP structures in pre-equilibrated bilayers at both coarse-grained and atomistic representation. Future developments will see MemProtMD becoming an increasingly mature service, with a greater variety of types of membrane protein, simulation and analysis.

Chapter 3

Automated identification of protein and lipid membrane components in PDB depositions

3.1 Abstract

Availability of high-resolution Integral Membrane Protein (IMP) structures in the Protein DataBank (PDB) has shed new light on the structure and function of IMPs, as well as interactions between IMPs and membrane lipids. Many approaches for studying IMPs rely on accurate orientation of proteins relative to a bilayer, however presently, the orientation of IMP deposited in the PDB is unreliable, and, whilst several databases have been developed with the aim of including every IMP structure, no complete classification exists, complicating the development of methods reliant on a large ensemble of IMPs.

MemProtMetrics has been developed both to orient proteins to a lipid bilayer and identify IMP structures in the PDB, by global optimisation of a knowledge-based objective function based on analysis of over 3000 simulations of IMPs in MemProtMD. Orientations produced by this method are in good agreement with both simulations of proteins in explicit bilayers and MEMEMBED and Orientations of Proteins in Membranes (OPM) orientations, although OPM orientations are generally in better agreement with simulation data. Using a threshold value, the MemProtMetrics objective function can be used to identify IMP structures in the PDB.

In addition, a method has been developed to identify and classify membrane ligands identified in PDB structures using a fragment-based approach. Using this method, 570 IMP structures bound to a total of 3449 membrane lipids have been identified. Lipid molecules are then annotated using protein-lipid interactions seen in

PDB structures.

3.2 Introduction

Over 3000 high-resolution Integral Membrane Protein (IMP) structures, representing more than 1000 unique proteins have been deposited in the Protein DataBank (PDB), revealing unprecedented insights into IMP structure and function, particularly in understanding interactions between proteins and membrane lipids. Using this wealth of structures, methods can be developed which rely on the existence of a large ensemble of IMP structures. Protein structures deposited in the PDB do not currently include any form of classification, and with IMP structures representing fewer than 3% of PDB depositions, the task of assembling an ensemble of IMP structures is far from trivial.

Several databases presently exist to address this problem. Orientations of Proteins in Membranes (OPM) [4] is a database of proteins oriented in simplified lipid bilayers, of which 2930 structures are identified as IMPs¹. The Research Collaboratory for Structural Bioinformatics (RCSB) PDB currently uses homology with OPM structures to categorise IMPs and monotopic membrane proteins. Unlike OPM, Membrane Proteins of Known Structure (mpstruc) [177] is a hand-curated database of IMP and monotopic membrane proteins, currently containing 2183 protein structures¹ categorised individually using expert knowledge. The MemProtMD Database (see Chapter 2) is currently the most complete collection of IMP structures, containing 3,052 IMP structures¹, verified by simulation in explicit lipid bilayers. However, the MemProtMD database is far from complete, with OPM containing 273 IMP structures not included in the MemProtMD database¹. The MemProtMD database also contains 412 PDB structures not included in OPM¹, with the two databases jointly containing 3342 IMP structures¹.

When embedded in a lipid bilayer, IMPs occupy a specific orientation, with a tilt relative to the lipid bilayer largely suggested to be determined by hydrophobic mismatch between side-chains of protein surface residues and membrane lipids [220, 221]. Experimental evidence of protein orientation is scarce, however tilt can be estimated from some solid-state Nuclear Magnetic Resonance (NMR) studies of IMPs [222, 223]. Although many structures are oriented relative to a lipid bilayer, IMP orientation is not recorded as standard in PDB depositions and so orientation seen in PDB coordinates cannot be relied on.

¹All counts accurate as of September 10th 2017.

Several tools presently exist to computationally estimate the orientation of IMPs in lipid bilayers using a variety of methods. TMDET [149] considers the solvent-accessible surface area of an input protein, and labels each exposed residue as either hydrophobic or hydrophilic. Protein orientation is then optimised using both the protein secondary structure and exposed hydrophobic surface, which are also used to predict whether a structure is an IMP. OPM [4] optimises protein orientation using parameters from Electron Paramagnetic Resonance (EPR) studies of lipid bilayer interface polarity, solvation energies in water and decadiene, and ionisation energies for charged residues. Knowledge-based methods also exist which apply a statistical potential derived from existing structures in the PDB to assign energy of bilayer insertion to a given orientation. These include *E_Z-3D* [224], which uses a potential generated from 76 α -helical and 35 β -barrel IMPs, and MEMEMBED [150], which uses a potential generated from a homology-reduced set of 183 α -helical and 37 β -barrel IMPs, taken from structures with at least a 3.5Å resolution.

IMPs constantly interact with membrane lipids; these interactions may be short-lived and non-specific, or highly specific. In many cases, specific lipid interactions have been shown to modulate protein function [62, 225, 226]. Although most structure determination methods result in the removal of the majority of membrane lipids, some lipids can be seen modelled in many experimentally determined IMP structures in the PDB [52]. These are likely to represent protein-lipid interactions which are of a sufficiently high affinity to survive displacement by detergents, and specific enough to produce an interpretable electron density.

Interactions between IMPs and membrane lipids in PDB depositions are of interest, both for validation of specific protein-lipid interactions seen in Molecular Dynamics (MD) simulations and in terms of the general properties of lipid bilayers. However, no dataset currently exists identifying protein-lipid interactions in PDB structures. The PDB Ligand Expo [227] contains chemical information for all standard PDB ligands, however does not store a commonly used name, or classification key which could be used to identify membrane lipids. COMPLIG [228] attempts to cluster standard PDB ligands based on graph topology, however performs poorly for membrane lipids. The problem is further complicated by the presence of non-standard lipid ligands, constructed from bonded fragments of standard ligands, precluding an approach which relies on classification of standard ligands.

Described here are methods to detect and process, respectively, IMP structures, and membrane lipid compounds deposited in the PDB. Using parameters derived from analysis of over 3000 simulations of IMP structures in explicit lipid bilayers in

the MemProtMD database, the MemProtMetrics tool can be used to both identify IMP structures, and predict the likely orientation and positions of leaflets of the lipid bilayer, using an approach which combines the strengths of a surface-based approach used by TMDET with those of a knowledge-based approach such as MEMEMBED. A second method has been developed to identify and categorise membrane lipids in PDB structures using a fragment-based approach. Once identified, experimentally determined contacts between IMPs and membrane lipids can be analysed and used to annotate lipid structures.

3.3 MemProtMetrics: Knowledge-based identification of IMP structures in PDB depositions

3.3.1 Materials and Methods

Acquisition and normalisation of protein structures

New protein structures are initially either downloaded from the RCSB PDB in Macromolecular Transmission Format (MMTF) format, a highly compressed structure for efficiently accessing macromolecular structure data, or loaded from the local filesystem in PDB or MMTF format. From this file, a set of entities are loaded, which may be protein chains, nucleic acid chains or ligands and are initially used to construct the asymmetric unit of the structure. In many cases the arrangements of entities in an asymmetric unit is not biologically relevant, and so, where possible, structures are loaded for both the asymmetric unit and biological assemblies. To isolate the protein and remove small peptide ligands, entities which contain mostly non-protein residues, or fewer than five residues, are removed.

In addition to the 20 amino acids of the universal genetic code, PDB structures occasionally contain rarer naturally occurring proteinogenic and non-proteinogenic amino acids such as selenocysteine, pyrrolysine, hydroxyproline and selenomethionine, or synthetic amino acids such as fluorotyrosine or aminosuccinimide. Protein-lipid and protein-solvent contact data from MemProtMD is only available for the 20 most common amino acids, so rarer amino acids are treated as their most similar commonly occurring counterpart by querying the PDB Ligand Expo (formerly PDB Ligand Depot) [227], which assigns a "nonstandard parent ID" to many rarer amino acids. In cases where an amino acid is ambiguous, unknown, or not assigned a parent id by the PDB Ligand Expo, it is treated as alanine, although this rarely happens.

”Preflight” checks of structure suitability

Once the protein has been loaded and normalised, several simple checks are performed to screen out inappropriately large proteins (such as virus capsid structures) or inappropriately small proteins. Proteins with more than 200000 atoms or 20000 residues are rejected as too large, while proteins with a longest dimension less than 2 nm are rejected as too small. Proteins with fewer than ten solvent-exposed hydrophobic residues are also rejected.

Finite element representation of approximate protein surfaces

A finite element approach is used to efficiently represent an approximation of the internal structure and solvent-exposed surface of a protein. In order to generate a mesh, protein atoms are first considered as a static point cloud. Three-dimensional Delaunay triangulation of this point cloud is performed using QHull [229], resulting in an optimised triangle mesh, bordered by the convex hull of the point cloud. Where a set of three vertices share three edges, this is termed a **face**, and where a set of four vertices share six edges (forming a tetrahedron), this is termed a **3-simplex**. To approximate a solvent-accessible surface, any face with an edge longer than the parameter α is removed. For protein structures, a value of $\alpha = 0.8$ nm produces reasonably detailed surfaces for both complete protein structures and structures containing only backbone or carbon- α atoms.

Faces of the resulting mesh may be either termed **surface**, if the face is part of a single 3-simplex or **internal**, where the face belongs to two 3-simplices. Edges are termed surface edges if they belong to at least one surface face, or internal if all faces to which they belong are internal. A surface mesh, representing the surface edges and faces of the resulting α -shape can then be extracted. A simplified two-dimensional depiction of this process is shown in Figure 3.1, and several three-dimensional examples are shown in Figure 3.2.

Protein surface parametrisation using MemProtMD

Once a surface mesh is generated, a metric ($\rho_{mem} - \rho_{wat}$) is calculated for each vertex which ranges from -1 (likely to contact water) to 1 (likely to contact the membrane core). Contact scales resulting from ensemble analyses of MemProtMD simulations are used to calculate, for the 20 most commonly occurring amino acids, ρ_{mem} ; from 0 (unlikely to contact lipid acyl tails) to 1 (likely to contact lipid acyl tails), and ρ_{wat} ; from 0 (unlikely to contact water) to 1 (likely to contact water). $\rho_{mem} - \rho_{wat}$ is first

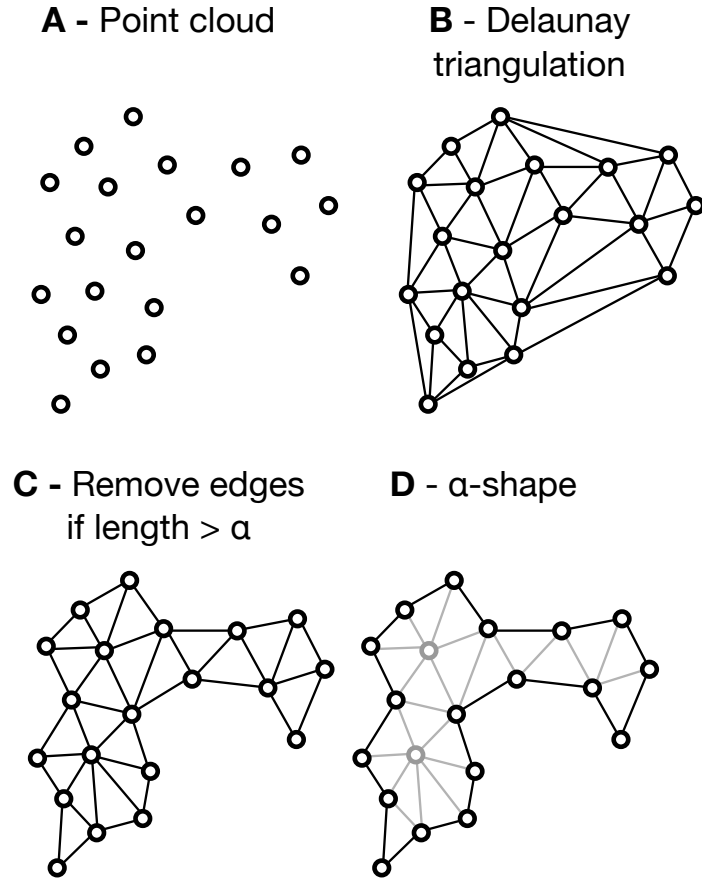


Figure 3.1: Example of α -shape generation from a two-dimensional point cloud

Starting with a two-dimensional point cloud (**A**), Delaunay triangulation is first performed to construct edges between points (**B**). From this an α -shape can be derived by removing edges longer than a specified length (α) (**C**). Edges and nodes of the resulting mesh can then be labelled as internal or external, revealing the boundaries of the α -shape (**D**).

This approach may be applied to three-dimensional point clouds by removing any faces from the Delaunay triangulation mesh with edges longer than α .

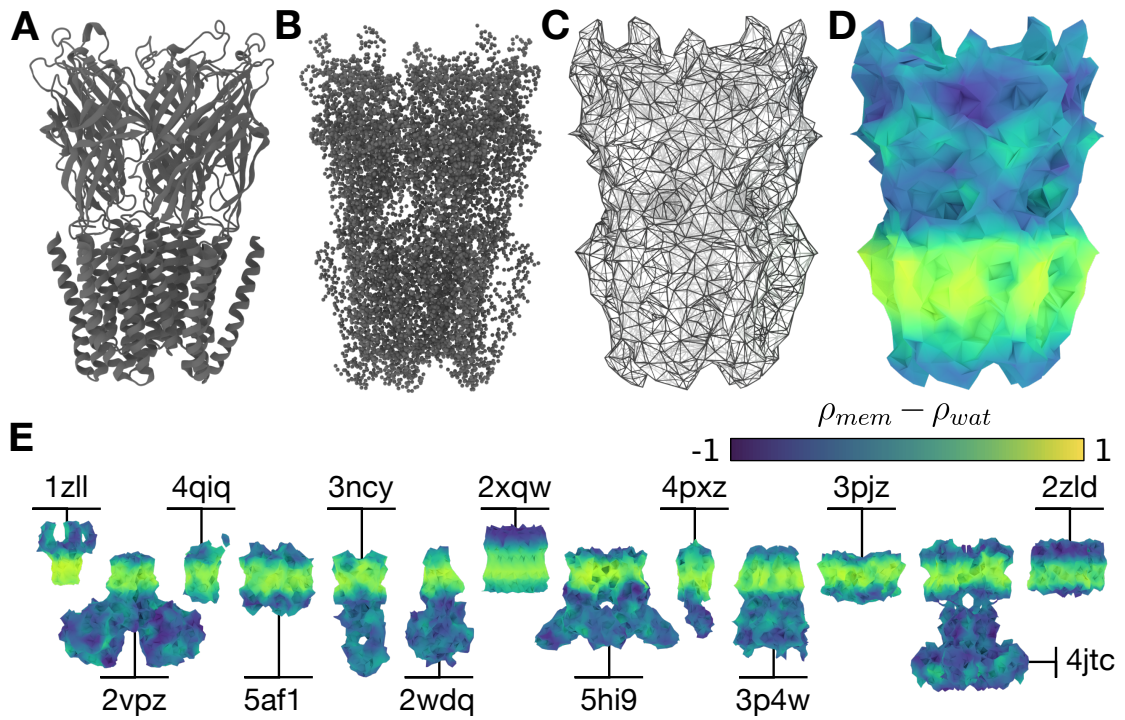


Figure 3.2: α -shape generation from a protein structure

The protein structure PDB:3lsv (shown in cartoon representation **A**) is first considered as a point cloud (**B**). Delaunay triangulation is performed, and faces with edges longer than α are removed, creating an α -shape (**C**). The surface is then parametrised with the metric $\rho_{mem} - \rho_{wat}$ (**D**), calculated for each residue and then smoothed across the protein surface. Examples of parametrised surfaces of several other proteins are also shown (**E**).

calculated for each surface vertex based on the residue the vertex atom belongs to. Smoothing is then applied, using edges of the surface mesh to iteratively assign to each vertex a weighted average of its immediate neighbours' values $\rho_{mem} - \rho_{wat}$ and its own value. The results of this smoothing process can be seen in Figure 3.3.

During development of MemProtMetrics, several other metrics were investigated in addition to $\rho_{mem} - \rho_{wat}$. These included simpler measures such as likelihood of interacting with lipid tail-groups, water or ions, as well as more complex metrics integrating multiple variables such as likely coordinate along the membrane normal. Whilst simpler metrics than $\rho_{mem} - \rho_{wat}$ introduced a greater degree of random error, more complex metrics tended to bias parametrisation in favour of multi-pass α -helical IMPs. $\rho_{mem} - \rho_{wat}$ was eventually selected as the parameter which effectively balanced random and systematic error.

Global optimisation of bilayer orientation

Once a protein surface is parametrised, the most likely protein orientation relative to the lipid bilayer is determined. Global optimisation is performed, sampling two rotational angles to find the orientation resulting in the greatest mean value of $\rho_{mem} - \rho_{wat}$ for a band of membrane-exposed surface around the protein. This first requires identification of potentially membrane-exposed surface, followed by calculating mean values of $\rho_{mem} - \rho_{wat}$ for the protein surface in a series of overlapping intervals spaced evenly along the proposed membrane normal.

Identification of pore and cavity inner surfaces

To identify potentially membrane-exposed surfaces, inner surfaces of pores and cavities must first be identified and excluded. These surfaces are typically not exposed to membrane lipids, and are likely to have an entirely different amino acid composition and consequently different value of $\rho_{mem} - \rho_{wat}$ to nearby lipid-exposed surfaces. Lipid diffusion in a lipid bilayer is anisotropic, being mostly constrained by leaflet surfaces, and which surfaces are membrane-exposed may differ depending on the orientation of a protein.

To identify pore and cavity inner vertices for a proposed membrane normal, the protein is considered as a set of overlapping slabs of $2 \times \alpha$ (1.6 nm) thickness, spaced evenly along the proposed membrane normal. Each slab contains a subset of the protein α -shape mesh, containing both surface and internal edges and faces. A further subset of this mesh may be termed a **surface-connected component** if any two

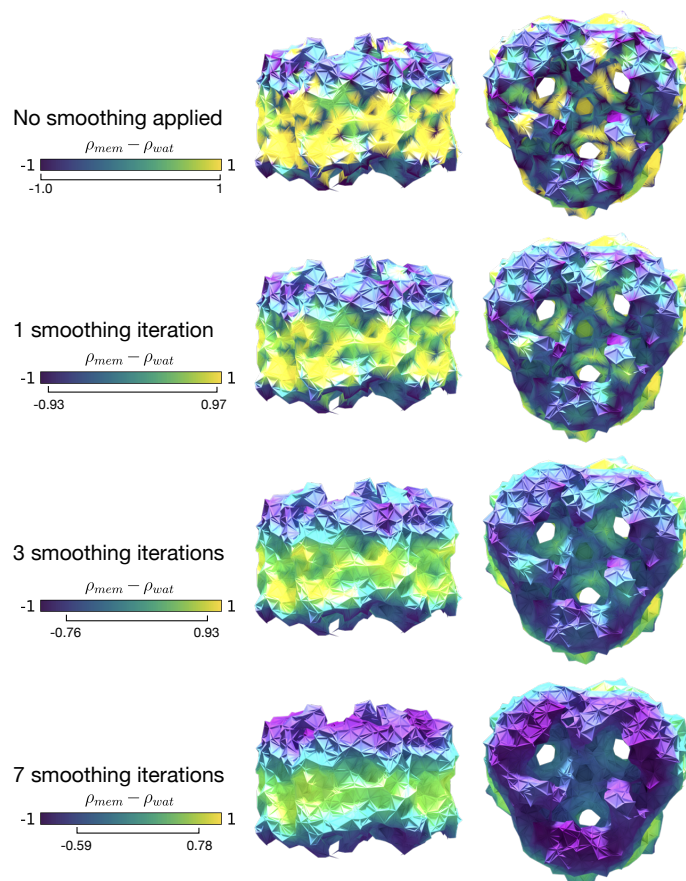


Figure 3.3: Smoothing of a surface metric across a protein mesh
 $\rho_{mem} - \rho_{wat}$ is initially calculated for each surface atom based on the residue it belongs to. A smoothing function assigns to each vertex a new value, calculated as a weighted mean of its own value and the values of other vertices with which it shares an edge. The outcome of one, three and seven iterations of the weighted smoothing function is shown. Indicated below scale bars are the maximum and minimum values of $\rho_{mem} - \rho_{wat}$ for the corresponding surface.

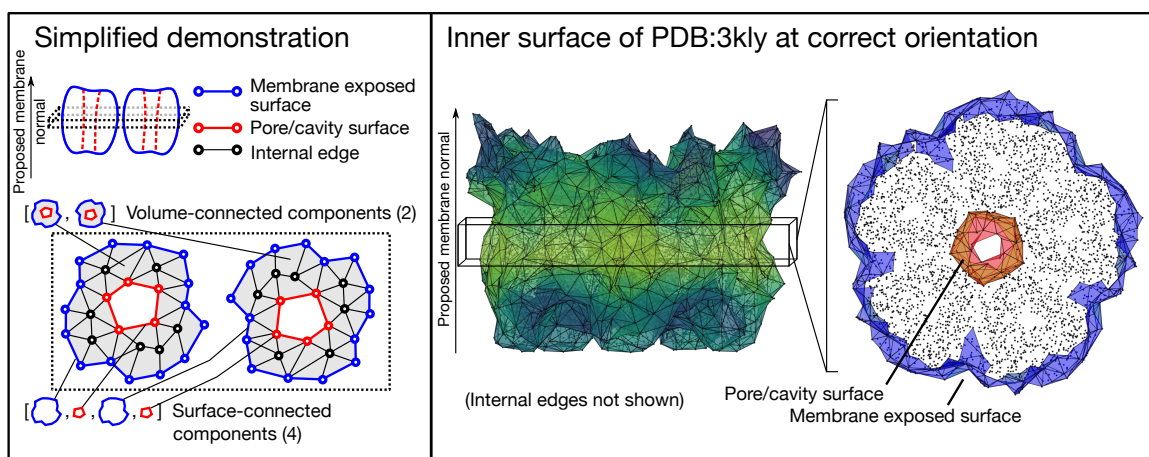


Figure 3.4: Illustration of protein pore and cavity inner surface identification

Two illustrations show the process of identifying a protein pore or cavity inner surface. On the left, a simplified demonstration considers a protein dimer, with each monomer containing a pore running parallel to the membrane normal (shown in red). Distance between dimers is exaggerated to demonstrate the case where more than one volume-connected component exists. Volume and surface-connected components are considered for a slab positioned perpendicular to the membrane normal, shown below. For each of two volume-connected components, it can be seen that the larger surface-connected component represents the likely membrane-exposed surface (blue), whilst the smaller surface connected component represents a pore inner surface. On the right, an example is shown using real data from the pentameric channel protein PDB:3kly, with the membrane exposed surface shown in blue and pore inner surface in red. For clarity, only the vertices of the internal mesh are shown as black dots. It is interesting to note the five regions of sparsity visible amongst the vertices of the internal mesh which represent additional pores within the protein. Due to the large value of α (0.8 nm), the surface mesh is too coarse to capture these pores however, if using a lower value of α such as 0.4 nm, these pores would be represented in the surface mesh and similarly identified.

vertices are connected by a path consisting entirely of surface edges, and no vertex is connected by surface edges to an additional vertex within the slab. A subset of this mesh may be termed a **volume-connected component** if any two vertices are connected by a path consisting of any edge type, and no vertex is connected to any additional vertices by any edge type. Each volume-connected component will contain at least one surface-connected component.

For a given 1nm slab, the largest surface-connected component of each volume-connected component is labelled as membrane-exposed, whilst all other surface-connected components are labelled as a pore or cavity inner surface. A visual demonstration of this may be seen for both simplified and realistic meshes in Figure 3.4. Since slabs overlap, each vertex will be labelled multiple times, and so will ultimately be assigned whichever label (membrane-facing or pore/cavity inner surface) is most commonly assigned.

Objective function calculation

Once pore and cavity inner surfaces are identified, the mean value of $\rho_{mem} - \rho_{wat}$ is calculated for all membrane-exposed vertices in a series of overlapping 2.2 nm slabs which form bins along the proposed membrane normal. Bin values are smoothed to remove noise using a Gaussian kernel, and the maximum bin value is taken as the value of the objective function, $\max(\rho_{mem} - \rho_{wat})$. An illustration of this process may be seen in Figure 3.5. The highest and lowest smoothed values are also recorded, along with the position along the proposed membrane normal of the highest scoring bin. This is taken as the most likely membrane centre and used to estimate leaflet positions if once the objective function is maximised.

Global optimisation

Global optimisation of the two rotational angles is performed to maximise the objective function using the Basin Hopping algorithm [230], an iterative method which performs multiple local optimisations with perturbations to initial values. Local optimisation is performed using Constrained Optimization by Linear Approximation (COBYLA) [231]. 200 Basin Hopping iterations are performed, each consisting of 1000 COBYLA iterations. Pore and cavity inner surface identification is only performed if a previously cached result is unavailable for a pair of angles within 20° of the proposed orientation.

Once a global maximum is found, the orientation and likely positions of membrane leaflets are recorded, along with the maximum of the objective function $\max(\rho_{mem} -$

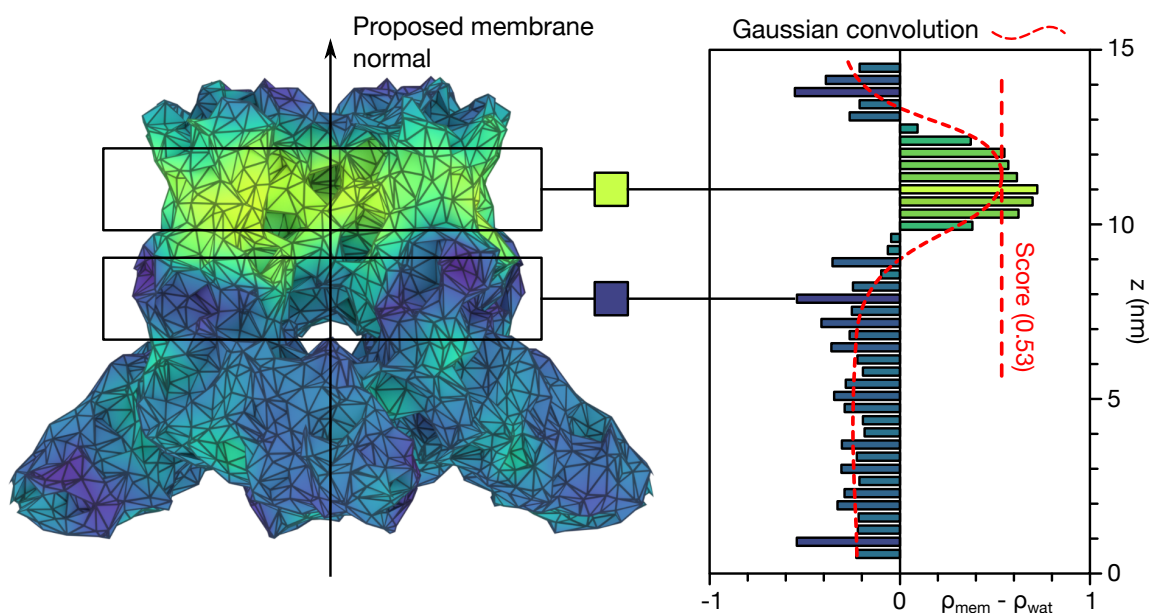


Figure 3.5: Illustration of objective function calculation

Calculation of the MemProtMetrics objective function for the optimal orientation of the TRPV2 channel PDB:5hi9 is shown. The parametrised protein surface is considered as a series of overlapping 2.2 nm thick slabs spaced along the protein normal. Two slabs, representing the maximum and minimum mean values of $\rho_{mem} - \rho_{wat}$ are shown. The mean value of $\rho_{mem} - \rho_{wat}$ is plotted against position along the membrane normal for each slab, and a dashed red line shows the results of convolution of these values with a Gaussian kernel.

ρ_{wat}). Protein structures which produce $\max(\rho_{mem} - \rho_{wat})$ above a threshold value of around 0.2 may be considered likely membrane proteins.

Benchmarking of MemProtMetrics, OPM and MEMEMBED by reference to simulation data

For each protein currently in the MemProtMD database, the atomistic structure produced by CG2AT based on the final frame of a 1 μ s Coarse-Grained (CG) simulation in the presence of an explicit bilayer was assumed to be in the correct orientation. Protein coordinates from this frame were rotated using a randomised rotation matrix and output to a file in PDB format. Alignment was performed by both MemProtMetrics and MEMEMBED using default parameters, except in the case of β -barrel proteins, where MEMEMBED was run using supplied parameters for β -barrel proteins. Structures aligned by OPM were downloaded from <http://opm.phar.umich.edu/pdb/tmprot.tar>. The angles between the assumed correct membrane normal and the membrane normals calculated respectively by MemProtMetrics, OPM and MEMEMBED were measured and used to benchmark the accuracy of each method. When comparing OPM and MemProtMetrics, in cases where an OPM-aligned structure was unavailable or differed in assembly, the structure was omitted from analysis.

Implementation

MemProtMD metrics is written in Python, and uses QHull [229] for mesh generation, NetworkX [207] for mesh storage and SciPy [209] to perform local and global optimisation. Results of protein scoring are stored using SQLite (<https://www.sqlite.org/>).

3.3.2 Results

As of September 10th 2017, the PDB contains a total of 133,397 structures, of which the vast majority (130,233) contain at least one protein molecule. 80,000 protein structures were randomly selected and assessed using MemProtMetrics. The distribution of objective function values can be seen in Figures 3.6A and 3.6B, from which it can be seen that $\max(\rho_{mem} - \rho_{wat})$ differs substantially between structures included in the MemProtMD database and those which are not. Known IMP structures in the MemProtMD database account for just 2% of PDB structures, and so even a method for IMP identification with an extremely low false positive rate will produce a large number of false positives (type I error).

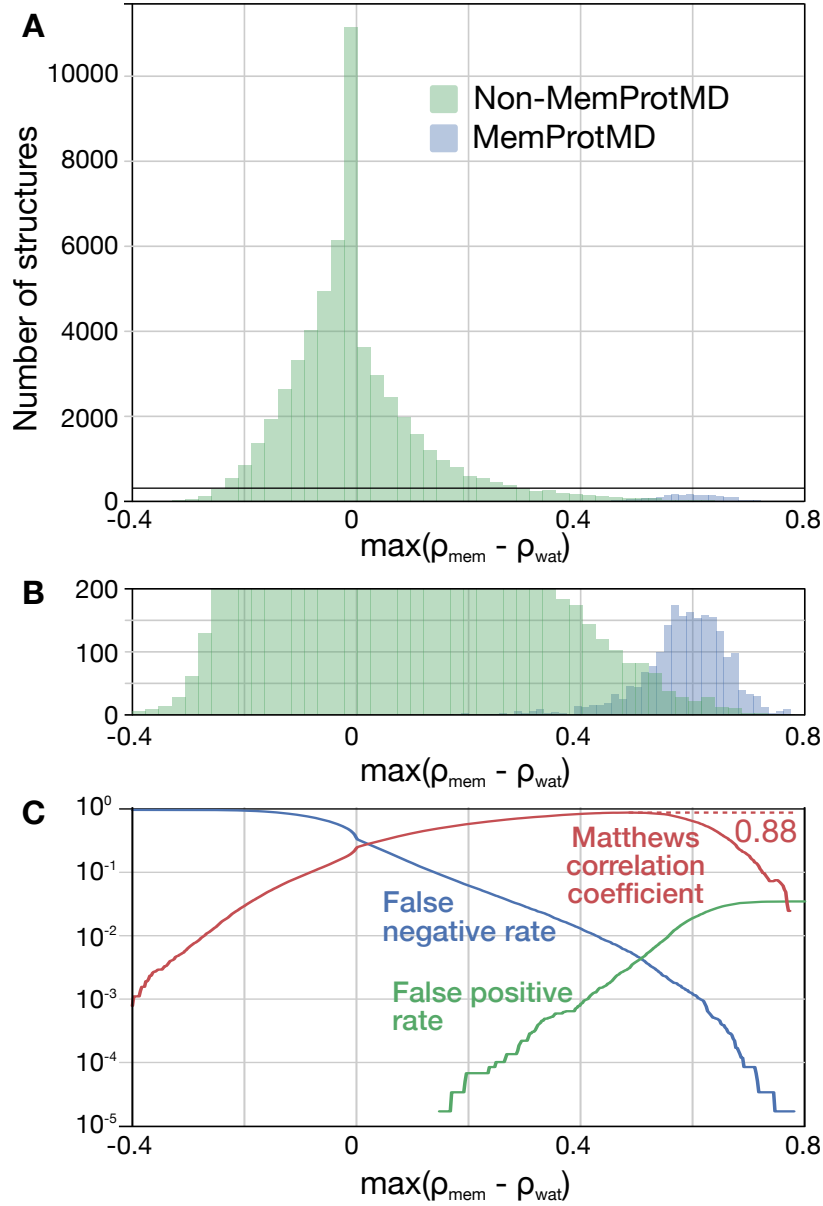


Figure 3.6: MemProtMetrics objective function for IMP and non-IMP structures

(A) The distribution of $\max(\rho_{\text{mem}} - \rho_{\text{wat}})$ is shown for 80,000 randomly selected proteins from the PDB, divided according to membership to the MemProtMD database of IMP. (B) The 0-200 y-axis range of this plot is repeated to show the distribution of $\rho_{\text{mem}} - \rho_{\text{wat}}$ for IMPs. (C) for cutoff values of $\max(\rho_{\text{mem}} - \rho_{\text{wat}})$ used to separate IMP and non-IMP structures between -0.4 and 0.8, the false positive rate, false negative rate and Matthews correlation coefficient are shown on a log scale.

The effect of using various threshold values of $\max(\rho_{mem} - \rho_{wat})$ to identify IMP structures is shown in Figure 3.6C. False positive rate (type I error) and false negative rate (type II error) are calculated respectively as $\frac{[fp]}{N}$ and $\frac{[fn]}{N}$, where N is the total number of structures, and fp and fn are respectively the number of non-IMPs above the threshold and number of IMPs below the threshold. Matthews correlation coefficient is a measure of the quality of a binary classification, taking into account both type I and type II error, and is calculated as $\frac{[tp] \times [tn] - [fp] \times [fn]}{\sqrt{([tp] + [fp])([tp] + [fn])([tn] + [fp])([tn] + [fn])}}$, where tp and tn are respectively the number of IMPs above the threshold and number of non-IMPs below the threshold. The Matthews correlation coefficient may be interpreted similarly to a Pearson product-moment correlation coefficient, with a value of 1 indicating perfect correlation.

Matthews correlation coefficient reaches a maximum of 0.88 at around $\max(\rho_{mem} - \rho_{wat}) = 0.5$, where the false positive rate and false negative rate are equal at 0.005, representing around 400 IMP structures classified as non-IMP structures and vice-versa. For this analysis the MemProtMD database is assumed to be a complete list of IMP structures, however manual inspection of 200 structures with greatest $\max(\rho_{mem} - \rho_{wat})$ not included in the MemProtMD database found 64 IMP structures, and so the false positive rate shown here is likely to be overestimated.

3.3.3 Estimation of leaflet surfaces for IMP structures

In addition to identifying likely IMP structures in PDB depositions, MemProtMetrics can suggest a likely membrane normal and estimate upper and lower leaflet surfaces. Three examples of this can be seen in Figure 3.7. In the cases of PDB:3nya and PDB:3aod, the estimated surfaces are close to the leaflet surfaces calculated from MD simulations. In the case of PDB:2nq7, however, the estimated leaflet surfaces are offset by approximately 20° from the simulated leaflet surfaces. This is likely due to a limitation of the MemProtMetrics approach when applied to single-pass transmembrane α -helices which are tilted relative to the membrane normal, resulting from the global optimisation process, which tends to favour placing single-pass α -helices parallel to the membrane normal, due to the scoring function not applying a penalty for likely membrane interacting surfaces appearing outside the proposed membrane. When this penalty is introduced, performance for larger α -helical proteins, β -barrel proteins and some single-pass α -helical proteins is negatively affected.

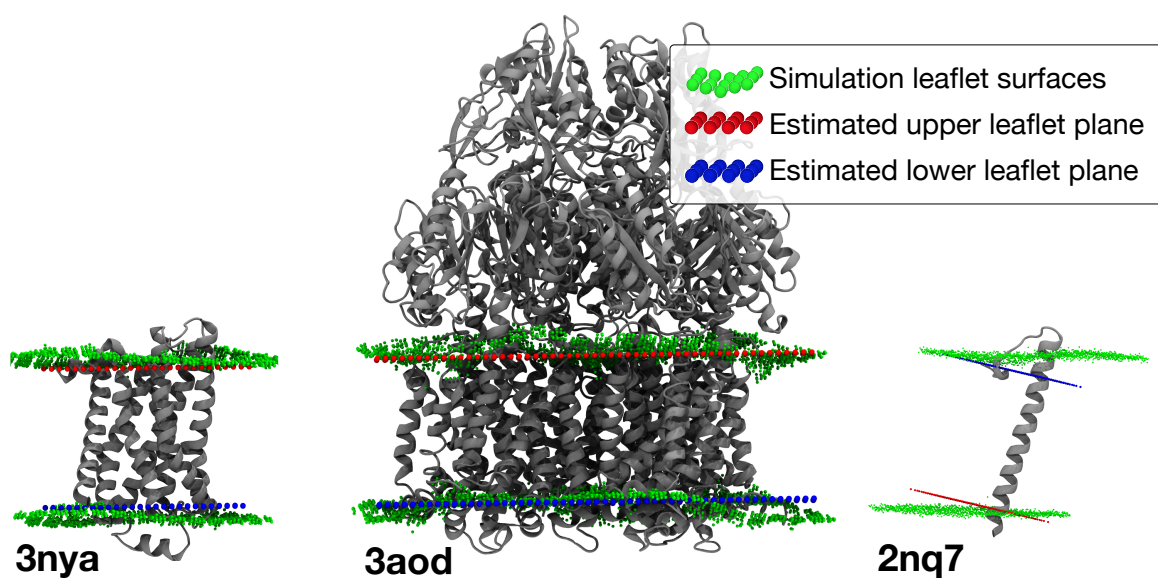


Figure 3.7: Examples of leaflet surface estimation for IMP structures
 Structures are shown overlaid with accurate leaflet surfaces downloaded from the MemProtMD Database and estimated leaflet surfaces (green) determined by MemProtMD metrics (red and blue). Proteins are aligned according to final orientation after 1 μ s CG simulations in the MemProtMD Database. Structures for human β_2 adrenergic receptor PDB:3nya, multidrug exporter AcrB PDB:3aod and type I human methionine aminopeptidase PDB:2nq7 are shown.

3.3.4 Benchmarking accuracy of MemProtMetrics, OPM and MEMEBED by reference to MemProtMD simulation results

Membrane alignments performed using MemProtMetrics and MEMEBED, as well as 1956 OPM-aligned structures with directly comparable assemblies, were compared to membrane normals obtained from 3052 MemProtMD simulations performed with an explicit bilayer. In Figures 3.8 and 3.9, membrane proteins are shown scattered according to the difference in angle between estimation and simulation produced by MemProtMD and MEMEBED or OPM respectively.

From benchmarking of MemProtMetrics and MEMEBED shown in Figure 3.8, it can be seen that error for both methods in the vast majority of structures is less than 10° , however several structures are aligned with errors greater than 20° . Whilst some structures are oriented erroneously by both methods, in most cases, where one method produces a large error, the other will produce a substantially smaller error; MemProtMetrics appears to perform more poorly with large α -helical structures, whilst MEMEBED appears to perform more poorly with β -barrel structures. Protein size appears to have little effect on accuracy of orientation, although both MEMEBED and MemProtMetrics often perform poorly for single-pass α -helical IMPs.

Benchmarking of MemProtMetrics and OPM shown in Figure 3.9 similarly shows errors below 10° for the majority of structures. Distribution of points, as well as the more flattened kernel density plot suggests that OPM generally orients structures more accurately and with fewer large errors, however there are still several structures for which MemProtMD metrics achieves a superior orientation, and several structures which neither method is able to align correctly. There does not seem to be any correlation between accuracy and protein secondary structure or size for either method.

A more direct comparison between accuracy of orientation produced by MEMEBED, OPM and MemProtMetrics is made in Figures 3.10 and 3.11, showing the number of structures aligned as accurately or more accurately than ϑ from 0° to 30° . Comparing MEMEBED with MemProtMetrics in Figure 3.10, it can be seen that up to 20° , there are a substantial number of structures for which only one method is able to orient a structure accurately; for 598 structures only the MemProtMetrics orientation is within 5° of the correct orientation, and for 556 structures only the MEMEBED orientation is within 5° of the correct orientation. For most values of ϑ , more structures are oriented correctly by MemProtMD metrics alone than by

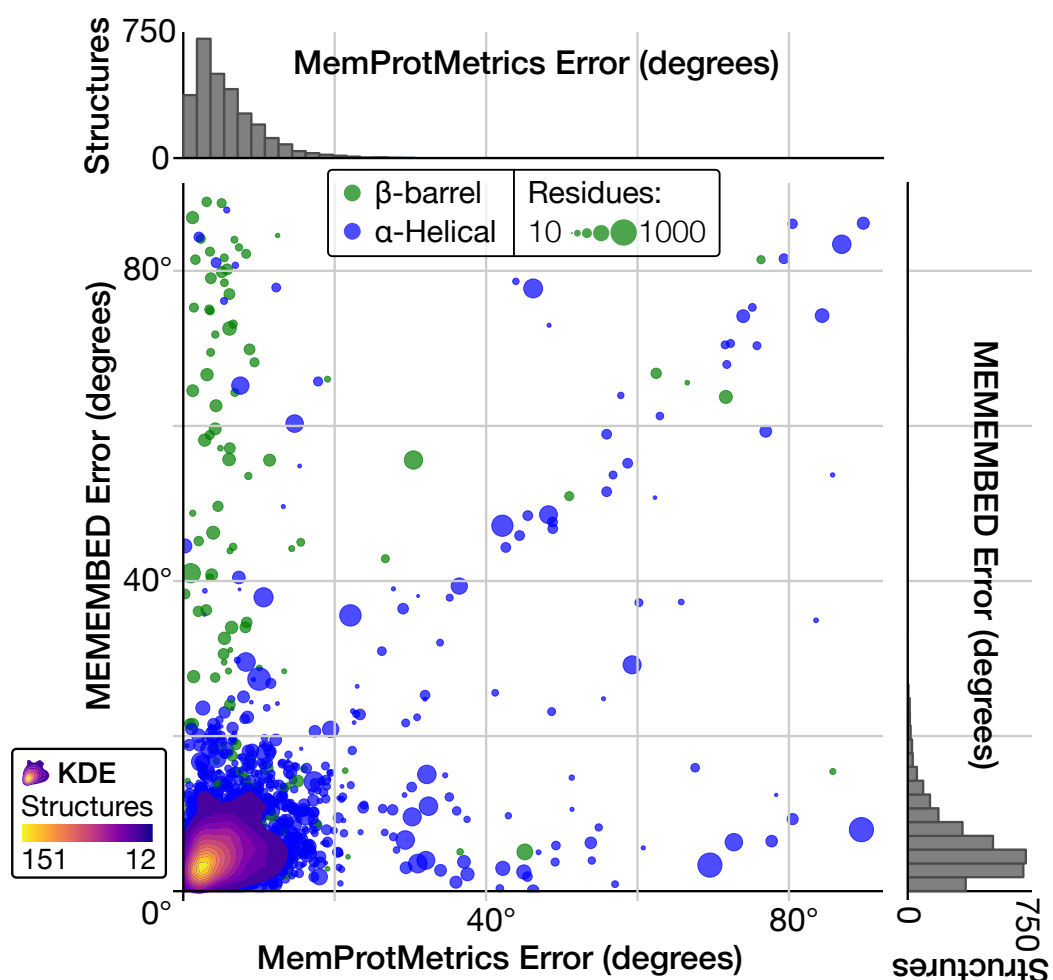


Figure 3.8: Scatter plot showing angular errors from MemProtMetrics and MEMEBED

Scatter plot of MemProtMetrics error against MEMEBED error. Where points become indistinguishable, a kernel density estimate is shown, shaded from blue (low density) to yellow (high density). Scattered points are coloured blue for α -helical proteins and green for β -barrel proteins, and are sized according to number of residues. Flanking the scatter plot are histograms showing the frequency of angular errors for each method.

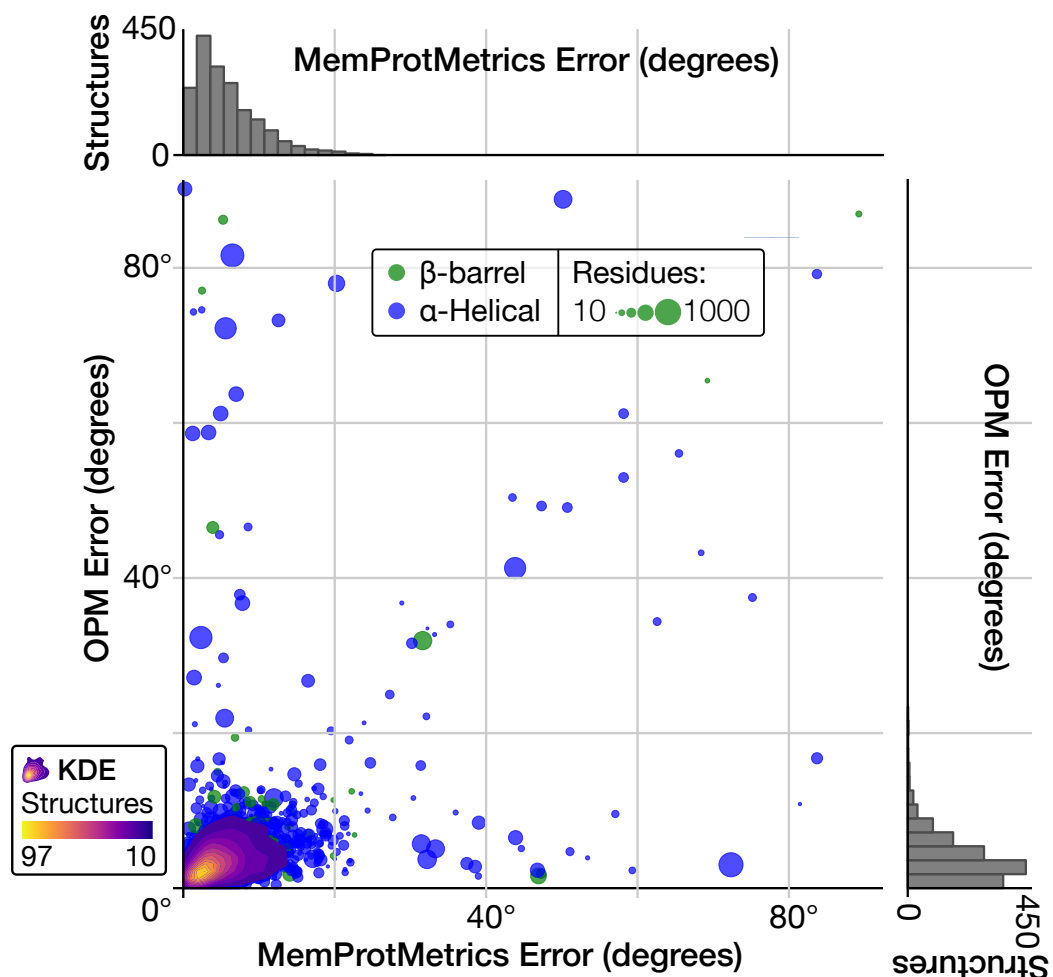


Figure 3.9: Scatter plot showing angular errors from MemProtMetrics and OPM

Scatter plot of MemProtMetrics error against OPM error. Where points become indistinguishable, a kernel density estimate is shown, shaded from blue (low density) to yellow (high density). Scattered points are coloured blue for α -helical proteins and green for β -barrel proteins, and are sized according to number of residues. Flanking the scatter plot are histograms showing the frequency of angular errors for each method.

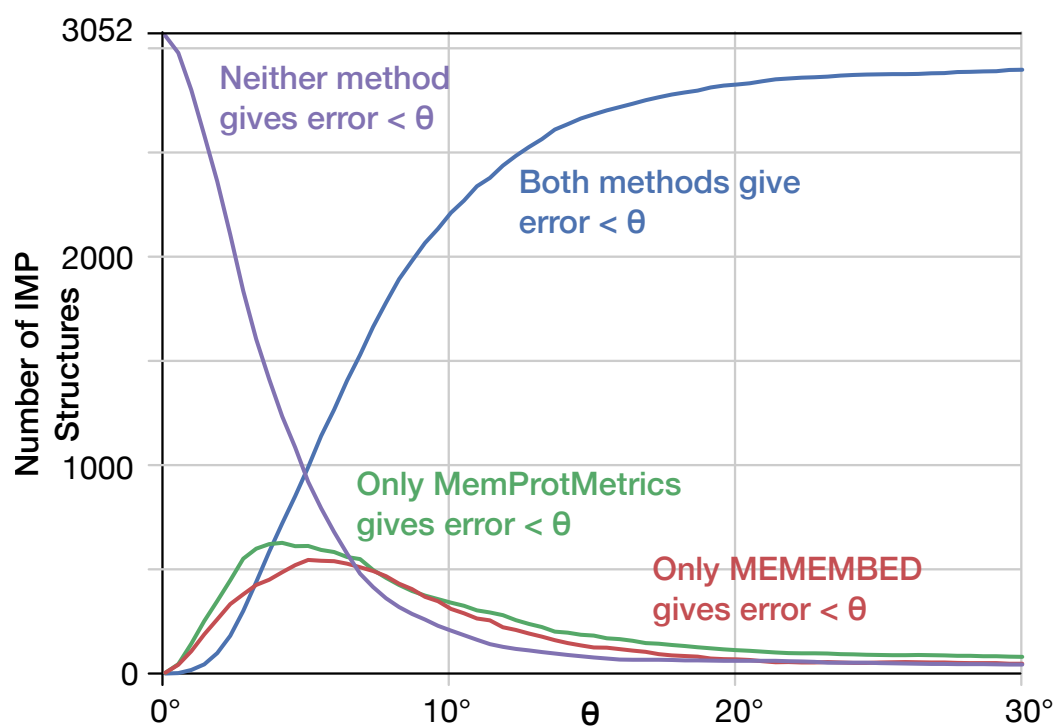


Figure 3.10: Cumulative error frequency comparison of MemProtMetrics and MEMEMBED

For angle error (θ) between 0° and 40° , counts are shown for structures where neither method, only one method or both methods output a result with angle error less than or equal to θ .

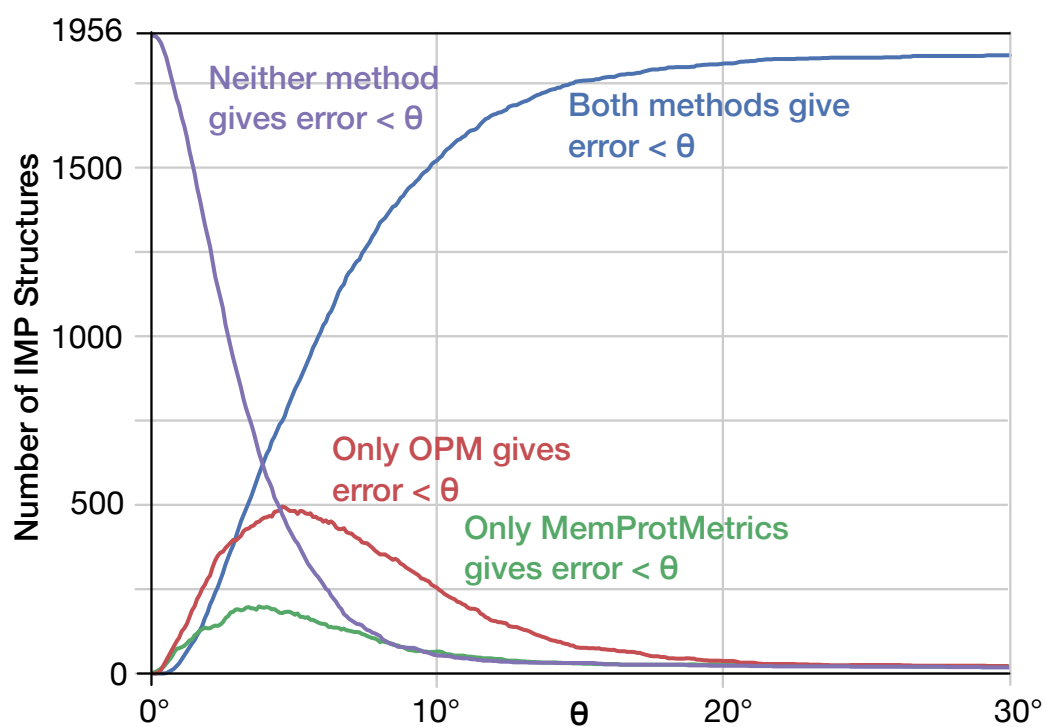


Figure 3.11: Cumulative error frequency comparison of MemProtMetrics and OPM

For angle error (θ) between 0° and 40°, counts are shown for structures where neither method, only one method or both methods output a result with angle error less than or equal to θ .

MEMEMBED alone, suggesting that MemProtMetrics is more accurate for a fractionally greater number of structures than MEMEMBED.

Comparison of OPM and MemProtMetrics shown in Figure 3.11 uses a reduced set of proteins for which OPM-oriented structures were available and directly comparable to MemProtMetrics-oriented structures. It can be seen that OPM is able to correctly orient considerably more structures to an accuracy of $\leq 10^\circ$ than MemProtMetrics. There are, however, several structures for which MemProtMetrics performs better, and both methods are able to align all but a few structures to $\leq 20^\circ$.

3.3.5 Cross-validation and fit bias

MemProtMetrics is currently parametrised using data from the entire MemProtMD database. This introduces potential for overfitting of MemProtMetrics to the particular set of IMPs for which structures are currently available. In order to perform a limited cross-validation of the results of MemProtMetrics, 152 IMP structures deposited in MemProtMD from September 2017 to January 2018 were aligned using parameters calculated from the July 2017 version of the MemProtMD database.

Results of this validation are summarised in Figure 3.12, showing tests performed on both training and testing sets compared to the MEMEMBED method. This analysis does not suggest that parametrisation of MemProtMetrics suffered from overfitting, with performance on the training set highly similar to performance on the testing set.

Accuracy of MemProtMetrics by protein family and fold

The uneven representation of folds within the MemProtMD database introduces a potential for bias in the MemProtMetrics model for families which are over-represented in the MemProtMD database. To test whether this had occurred, prediction error was compared between different SCOP folds.

In Figure 3.13, distribution of prediction errors for MEMEMBED and MemProtMetrics can be seen for the 12 best represented SCOP folds in MemProtMD. Error can be seen to vary slightly between SCOP folds, with single transmembrane helices and cytochrome components performing particularly well. MEMEMBED shows substantially increased prediction error for β -barrel proteins even when using the β -barrel-specific parameters. This is not seen with MemProtMetrics prediction errors.

The relationship between representation of a SCOP fold in the MemProtMD database and prediction error is shown in Figure 3.14. This shows some limited

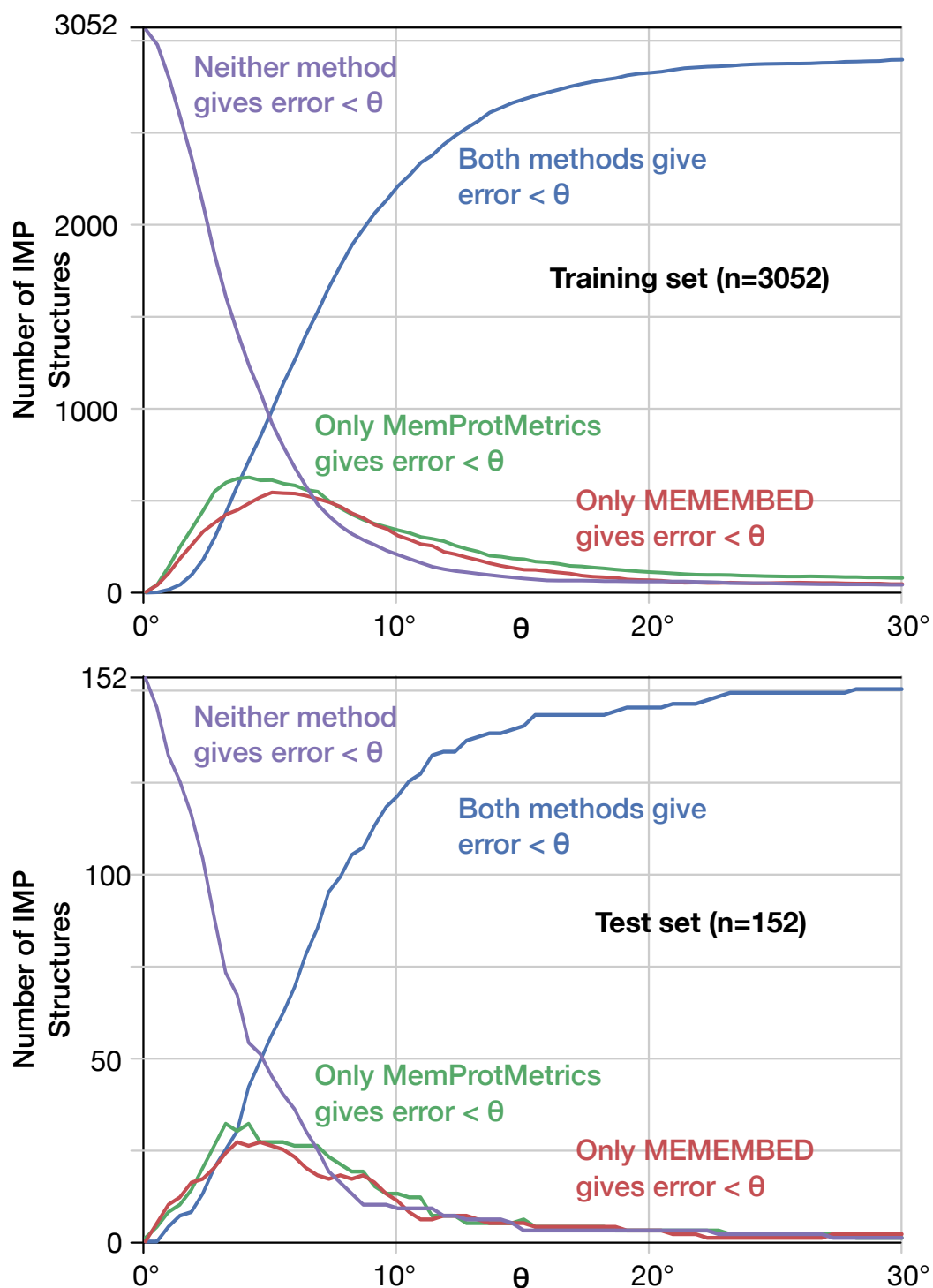


Figure 3.12: Cumulative error frequency for training and test sets
 Figure 3.10 is repeated above an identical analysis performed on the 152 IMPs added to MemProtMD between September 2017 and January 2018. For angle error (ϑ) between 0° and 40°, counts are shown for structures where neither method, only one method or both methods output a result with angle error less than or equal to ϑ .

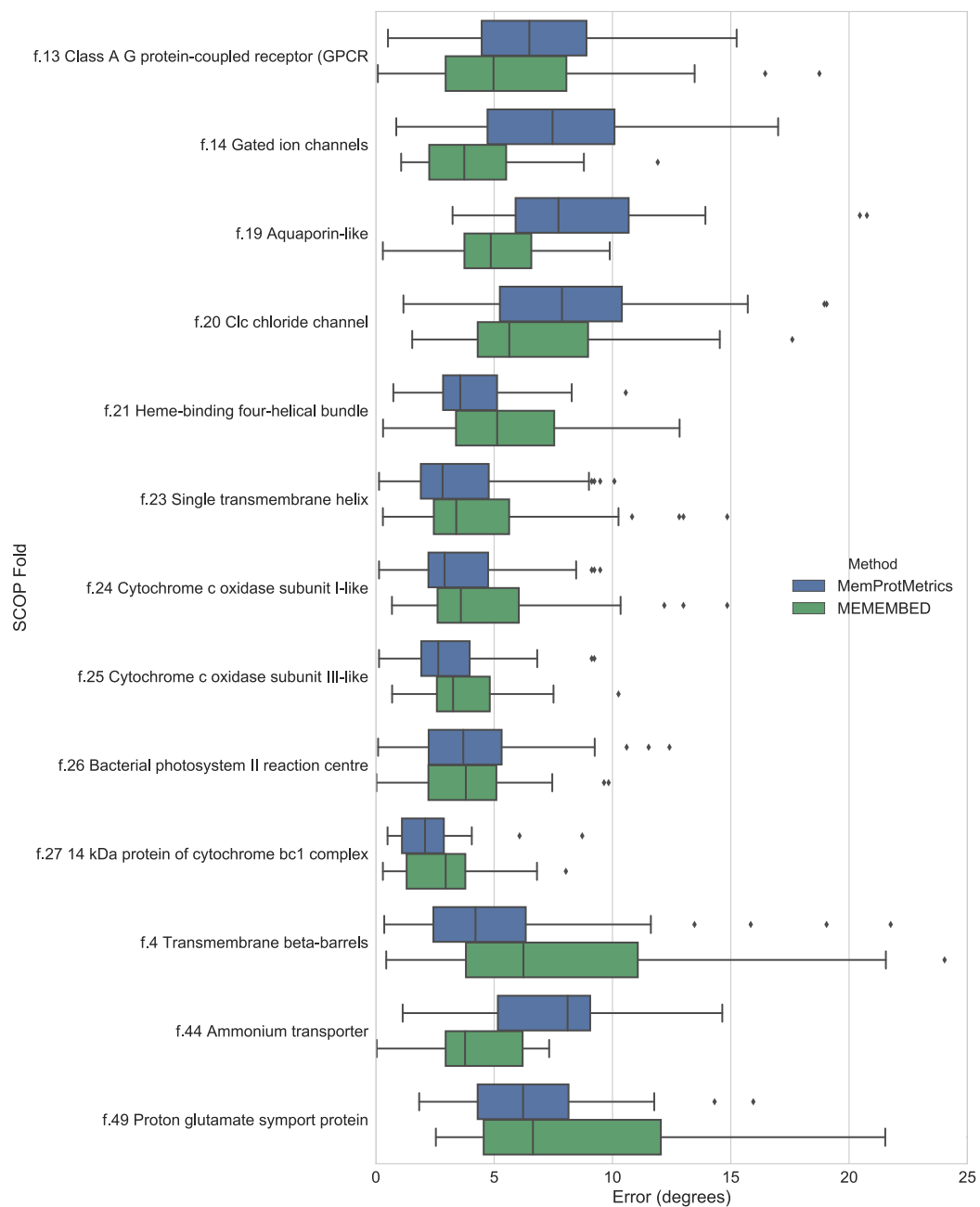


Figure 3.13: MemProtMetrics and MEMEMBED error by SCOP fold
 Box plots of prediction errors are shown for both MEMEMBED and MemProtMetrics for the 12 best represented SCOP folds in MemProtMD.

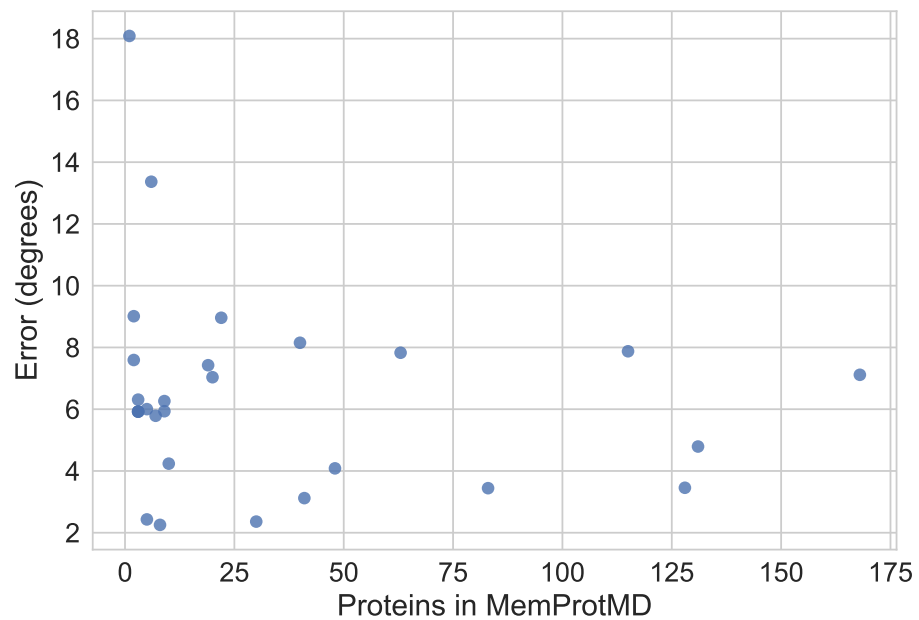


Figure 3.14: MemProtMetrics error by number of proteins

Mean MemProtMetrics prediction error is shown plotted against number of proteins in the MemProtMD database for each SCOP fold.

evidence of bias towards better represented proteins, with two of the poorest represented folds (f.3 Light-harvesting complex subunits and f.18 F1F0 ATP synthase subunit A) showing the greatest prediction error.

3.4 Fragment-based identification of lipid bilayer components in PDB depositions

3.4.1 Materials and Methods

In order to visualise common interactions between IMPs and lipid components of membranes, a system has been developed to identify, classify and align membrane lipids in PDB structures, including cholesterol, phospholipids and Lipopolysaccharide (LPS). Ligands in PDB files are first converted to accurate chemical structures and loaded using RDKit [232]. Attempts are then made to analyse the molecule as either a branched lipid or a lipid based on a steroid nucleus. If successful, model outputs are used to further classify the molecule into groups representing common membrane lipid components. Similar molecules can be aligned, and contacts with protein amino acids found frequently in deposited structures can then be overlaid.

Building accurate molecular structures from PDB depositions

Where data from structure determination experiments show a ligand bound to a protein structure, the likely atoms and coordinates of this ligand are deposited in the PDB with the protein structure. The PDB Ligand Expo [227] stores a database of around 25,000 standard ligands, which represents the majority of modelled ligands in PDB depositions. In some cases, particularly for large molecules such as Lipid A, a ligand in a PDB deposition may be composed of several fragments of standard ligands bonded together.

Ligands identified in PDB files are first converted to a list of atoms and bonds. The PDB Ligand Expo database is then used to populate atom attributes such as element and charge, and bond attributes such as multiplicity and aromaticity. Where a ligand is deposited as several bonded fragments, bonds between fragments cannot be queried and so are assumed to consist of a single σ bond. Atoms and bonds are then used to build a chemical structure using RDKit.

Identifying lipid bilayer components from molecular structures

Common membrane lipids are either based around a steroid nucleus (in the case of cholesterol, ergosterol and lanosterol), or consist of one or more hydrocarbon tails, connected to a head-group (in the case of phospholipids and glycolipids). Ligands are fitted to two models, optimised respectively for lipids based on a steroid nucleus and lipids based on connected fatty acids. Each model returns a measure of goodness of fit and several metrics capturing variation within the model for more precise classification. Many molecules will be inappropriate for a particular model and will produce a goodness of fit of zero.

Steroid nucleus-optimised model

The steroid nucleus-optimised model considers molecules as a set of accessory groups bonded to a central sterol group. A sterol group is first located, and numbered according to International Union of Pure and Applied Chemistry (IUPAC) recommendations for the nomenclature of sterols [233]. The presence or absence of double bonds between atoms of the sterol group is returned. For each atom of the sterol group, directly connected atom types are recorded, as well as the total number of atoms connected to the sterol group either directly or indirectly at that position. Goodness of fit is returned as either 1 (sterol group is present) or 0 (no sterol group present).

Membrane lipids fitting this model include cholesterol, ergosterol and lanosterol, which can be identified by a directly bonded oxygen atom at C-3, and at least four carbon atoms indirectly bonded to C-17. For further classification these may be differentiated by the presence of double bonds between C-5 and C-6 (cholesterol and ergosterol), C-7 and C-8 (ergosterol), and C-8 and C-9 (lanosterol). Examples of these lipid types can be seen in Figure 3.15.

Branched hydrocarbon-optimised model

The branched hydrocarbon-optimised model considers molecules as one of more hydrocarbon chains, joined to a connecting group which may in turn be joined to one or more head-groups (which may then be joined to further head-groups). Metrics include the number of each type of component present, and goodness of fit returned by this model quantifies the proportion of atoms which could be described within this formalism.

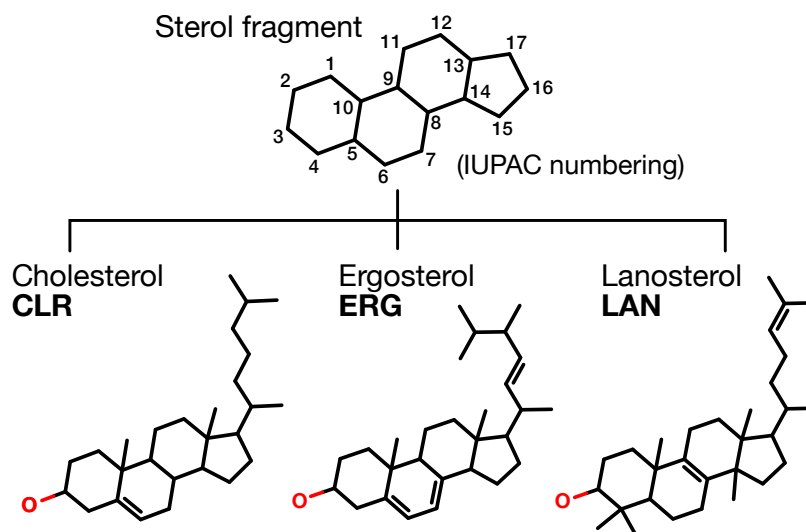


Figure 3.15: Classification of membrane ligands based on a steroid nucleus
 A sterol fragment is shown numbered according to IUPAC recommendations [233]. Below are shown the three most common sterols found in lipid bilayers; cholesterol, represented by the standard PDB ligand CLR, ergosterol represented by the standard PDB ligand ERG and lanosterol, represented by the standard PDB ligand LAN. All three molecules may be identified by an oxygen bonded to C-3 and tail bonded to C-17, and differentiated by the presence or absence of double bonds within the sterol nucleus.

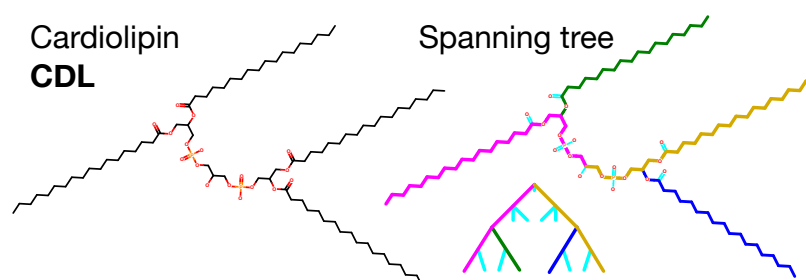


Figure 3.16: Minimum spanning tree construction for a cardiolipin molecule

The minimum spanning tree of a molecule is helpful when analysing molecules with a mostly tree-like topology. Shown here is, on the left, a cardiolipin molecule representing the standard PDB ligand CDL. On the right, the molecule is coloured according to the branches of a minimum spanning tree, and a schematic is shown depicting the branching order and root of the tree. The topology of cardiolipin forms a tree, and so there exists only one possible spanning tree.

To identify hydrocarbon chains, an undirected graph is constructed representing atoms as nodes and bonds as edges. This graph is divided into a set of linear branches according to a minimum spanning tree (see Figure 3.16), and any continuous sequence of bonded carbon atoms within a linear branch (which are not members of a ring structure) is considered a potential hydrocarbon chain. The connectivity of each potential hydrocarbon chain is then considered at each terminus, as are any side-chains attached to non-terminal atoms.

To be accepted, a hydrocarbon chain must be highly connected at one terminus, and connected to at most one non-carbon atom at the other terminus. It must also contain at least five carbon atoms on the main chain, and the main chain must contain at least five times as many atoms as the total number of atoms found by traversal of side-chains attached to non-terminal atoms. A special case exists for acceptance where an ester bond connects another hydrocarbon chain as a side-chain, as is seen in some LPS molecules. If accepted, a hydrocarbon chain is recorded in terms of the number of carbon atoms, ratio of main-chain atoms to side-chain atoms, number of each bond type, and types of atoms attached to termini.

Once hydrocarbon tails are found, the molecule is searched against a set of **connector fragments** which represent common backbones found in different types of membrane lipid. These include glycerol groups, found in phospholipids and glycolipids, sphingosine groups, found in ceramides and sphingolipids, and glucosamine groups, found in LPS molecules. Any connector fragment which is not bonded to a hydrocarbon chain is removed. Where any two remaining connector fragments share a third of atoms, they are deemed conflicting, the connector fragment bonded to the fewest hydrocarbon chains (or smallest fragment where fragments are equally connected) is removed. Examples of the three most common connector fragments can be seen in Figure 3.17.

Once connector fragments and hydrocarbon tails have been identified, the molecule is then queried against a set of **head-group fragments**. These include groups such as phosphate, ethanolamine, serine, commonly found in lipid head-groups. Fragments which share at least one-third of atoms with an existing connector fragment or hydrocarbon chain are removed, and conflicts between head-groups are resolved in a similar manner to connector fragments. Several examples of headgroup fragments can be seen in Figure 3.18.

A graph is constructed with hydrocarbon chains, connector fragments and head-group fragments as nodes and bonds between fragments as edges. Where multiple connected components exist within this graph, only the largest connected component

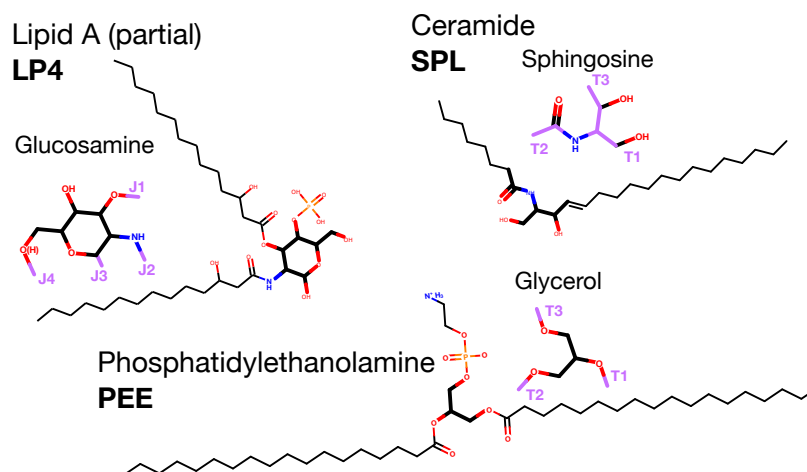


Figure 3.17: Examples of common connectors for branched fatty acid-type membrane lipids

The three most commonly occurring connectors are shown as molecular fragments next to example molecules selected from the standard PDB ligand library, with atoms of the fragment indicated by thicker bonds. Lipid A is shown connected by a glucosamine-like fragment, which may be attached at various points to up to four additional organic or inorganic groups (**T1-4**); here T1 and T2 connect to fatty acid chains, whilst T4 connects to a phosphate group. Ceramide is shown connected by a sphingosine fragment, connecting two hydrocarbon chains at T1 and T2. Phosphatidylethanolamine is shown connected by a glycerol fragment, with fatty acids connected to T1 and T2 and a phosphate group connected to T3.

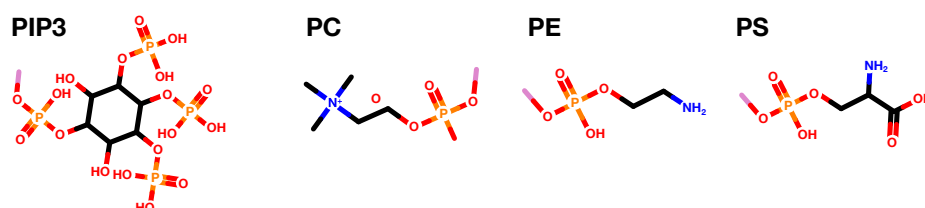


Figure 3.18: Examples of common head-groups for branched fatty acid-type membrane lipids

Four common phospholipid head-groups are shown; Phosphatidylinositol (3,4,5)-trisphosphate (PIP₃), Phosphatidylcholine (PC), Phosphatidylethanolamine (PE) and Phosphatidylserine (PS). Points at which a connector fragment is commonly bonded to the head-group are shown as a lilac coloured bond.

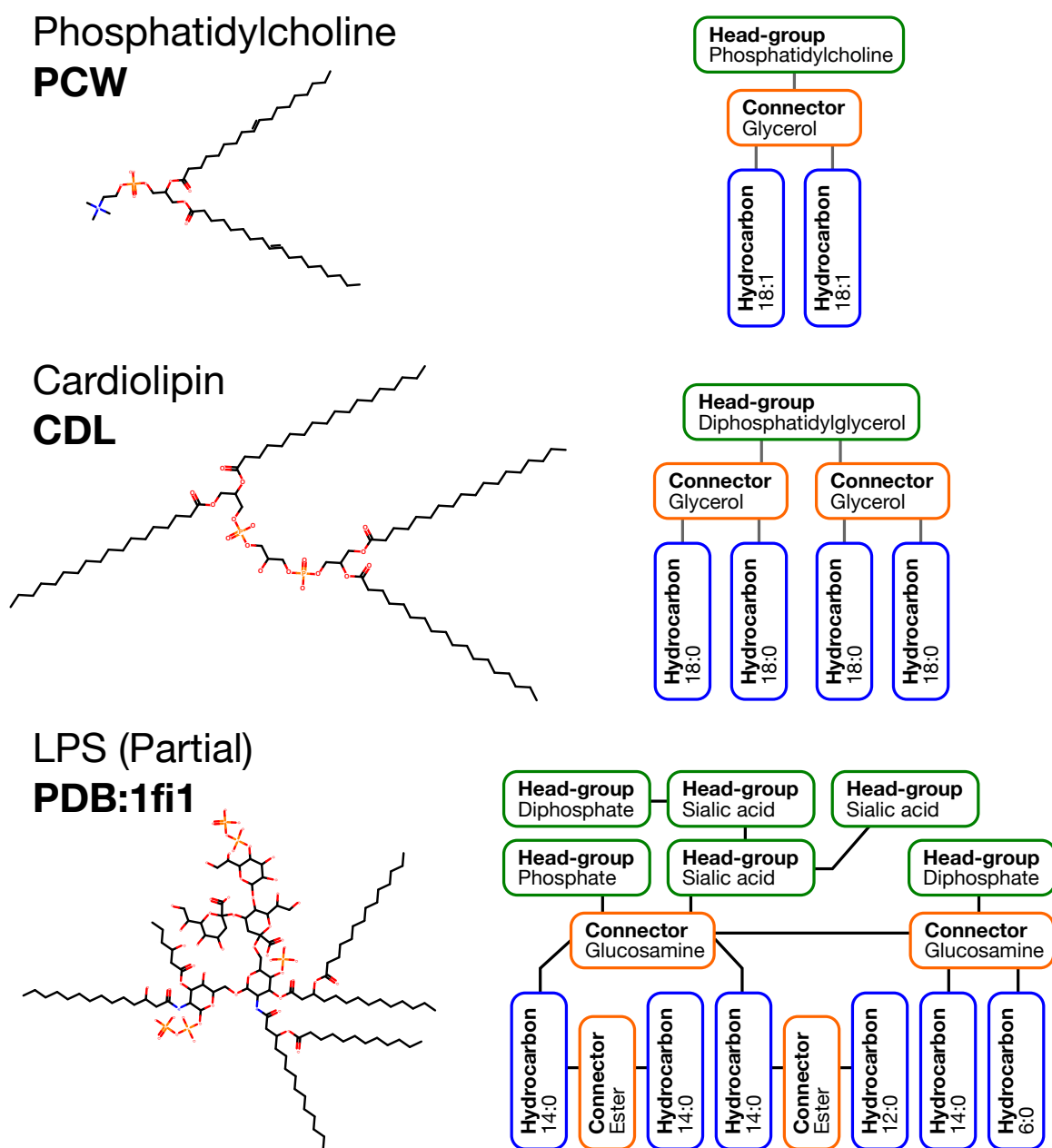


Figure 3.19: Branched fatty acid-optimised model applied to three lipid molecules

Three examples of lipid molecules are shown alongside the assembled graph of hydrocarbon chains, connector fragments and head-group fragments. PC, represented by the standard PDB ligand PCW, can be seen as two 18:1 hydrocarbon chains, bonded to a glycerol connector fragment which is in turn bonded to a PC headgroup fragment. Cardiolipin, represented by the standard PDB ligand CDL, is shown as two sets of two 18:0 hydrocarbon chains bound to a glycerol connector fragment, joined by a single Diphosphatidylglycerol (DPG) headgroup fragment bonded to both the glycerol fragment. A more complex example may be seen in the partial LPS structure found in PDB:1fi1, however this increased complexity is still captured entirely within the model.

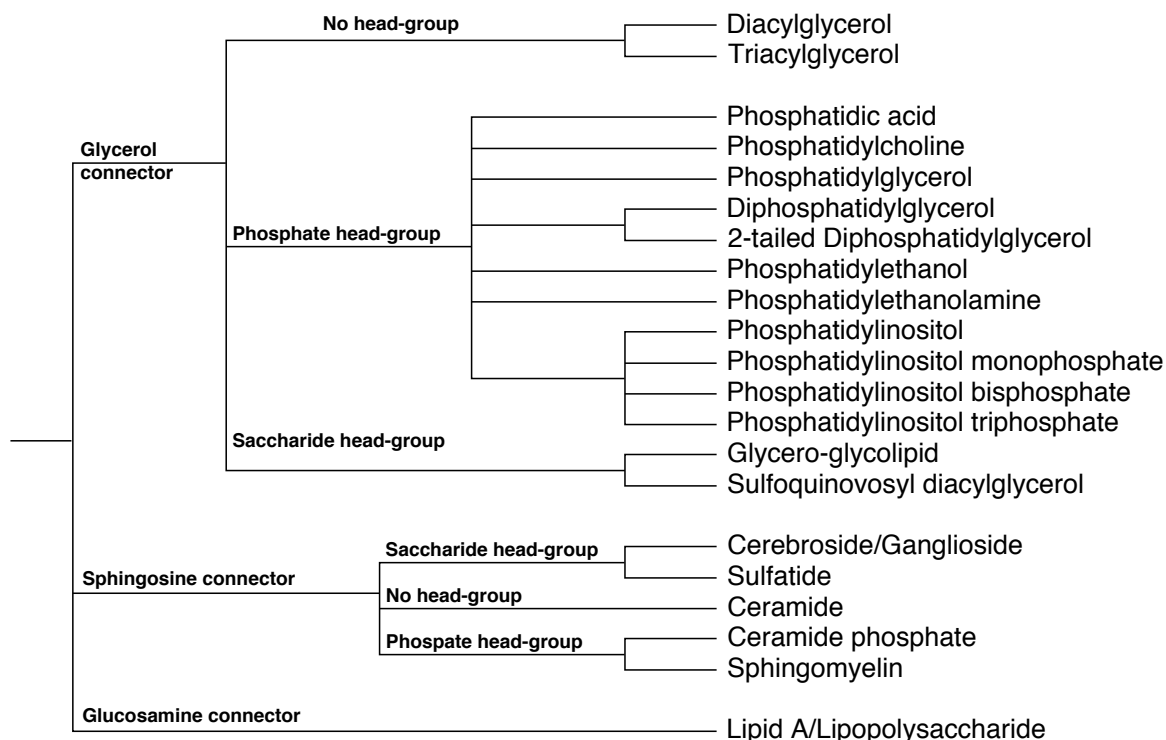


Figure 3.20: Classifications applied to molecules by the fatty acid-optimised model

Classifications applied by the branched fatty acid-optimised model are shown on a tree structure. Where the model is able to fit at least 80% of atoms in a molecule, it is classified based on the presence or absence and number of several different hydrocarbon tail, connector and head-group fragments.

is used. A goodness of fit score is returned, representing the proportion of atoms included within the fragments of this graph, along with the number of each type of fragment present. Several examples of this can be seen in Figure 3.19.

Where at least 80% of atoms in a molecule are accounted for by the fragments of this graph, the molecule is considered a likely branched hydrocarbon-type membrane lipid. Further classification can then be performed using the number of each type of element present. A comprehensive list of the classifications which may be applied to a molecule fitting this model can be seen in Figure 3.20.

Validation of lipid classification

In order to validate results of lipid identification, a large set of molecular structures were inspected visually. Initially, all 24728 standard PDB ligands were screened with extremely relaxed criteria to identify a superset unlikely to exclude any valid mem-

brane lipids. This included any molecule with a steroid core, and any molecule with at least two hydrocarbon chains containing at least four carbon atoms, resulting in a set of 653 molecules, containing mixture of detergents, drug and drug-like molecules, and membrane lipids. Of these, 350 were randomly selected for visual inspection. Amongst these, 87 molecules were identified by visual inspection as lipid components of cell or organelle membranes. Results of this visual classification were then used in the development of stricter parameters.

Quantification of false positive and false negative rates is complicated by the lack of any larger database of accurate membrane lipid structures. All lipid structures identified as membrane components by manual inspection are present in the automatically identified set, and all automatically identified lipids were subsequently identified by visual inspection as valid membrane lipid molecules. Some cross-validation has been possible by subsequent inspection of the remaining 303 molecules not manually inspected, which showed that no valid membrane lipid molecules omitted by automatic identification.

Annotation of lipid molecular structures with protein-lipid interactions based on PDB depositions

By combining data from multiple structures containing similar ligands, more generalised interactions between IMPs and membrane lipids can be characterised. A comprehensive list of membrane protein structures with ligands bound was obtained from MemProtMD (<http://memprotmd.bioch.ox.ac.uk>). All ligands were classified to identify likely membrane lipids, including non-standard ligands not included in the PDB Ligand Expo. All contacts within 0.5 nm between lipid and protein atoms were recorded, providing both atoms were assigned temperature (Debye-Waller) factors of less than 75 Å².

Protein residues were grouped according to their spatial distribution in MemProtMD simulations, shown in Table 3.1. Similarly classified membrane lipids were aligned using their maximum common substructure, computed by RDKit, ignoring bond aromaticity and multiplicity. The total abundance of amino acids was calculated within 0.5 nm for the entire molecule and used to calculate expected abundance of amino acids both individually and by group for each atom within the alignment. This was used to calculate the binary logarithm of the ratio of observed abundance to expected abundance ($\log_2(\frac{\text{observed}}{\text{expected}})$) for each pairing of amino acid type or amino acid group with an atom of the aligned structure. Where the observed abundance of

Group	Amino acids
Most hydrophobic	ILE, LEU, PHE
Intermediate hydrophobic	CYS, VAL
Mildly hydrophobic	ALA, MET
Polar hydrophobic	ASN, GLN, HIS
Small molecules	GLY, PRO, SER, THR
Aromatic	TRP, TYR
Positively charged	ARG, LYS
Negatively charged	ASP, GLU

Table 3.1: Protein amino acid groupings derived from MemProtMD

an amino acid type or group is substantially higher than the expected value, this is shown overlaid on a two-dimensional representation of the molecule.

3.4.2 Results

Identification of standard and non-standard membrane lipid ligands

Screening of the PDB Ligand Expo identified 197 classifiable membrane lipid types amongst the 24728 standard PDB ligands. Among IMP structures deposited in the PDB, 78 standard ligands and 34 non-standard ligands were found. Amongst 3052 IMP structures examined, 570 contained at least one modelled membrane lipid. Several examples of lipids bound to proteins in PDB structures are shown in Figure 3.21, including both standard ligands such as phosphatidylcholine, and non-standard ligands glycerophospholipid and LPS. It is not uncommon to see multiple membrane lipids modelled in a single PDB structure, with 176 structures containing five or more membrane lipid molecules and 47 structures containing at least 20.

The number of standard and non-standard membrane lipid types identified, and the number of structures for which at least one ligand is modelled is shown in Table 3.2. Rows are arranged similarly to the membrane lipid tree seen in Figure 3.20, and this tree is shown alongside the table. Phospholipids account for the vast majority of membrane lipids found in PDB structures, with phosphatidylcholine, diphosphatidylglycerol (cardiolipin) and phosphatidylethanolamine present in 149, 107 and 134 structures respectively. Outside phospholipids, 97 structures contain at least one diacylglycerol molecule, although many of these may represent partial modelling of larger lipids. Cholesterol is seen in 88 structures, and sulfoquinovosyl diacylglycerol in 35. Lipopolysaccharide molecules are seen in 18 different structures, each time as

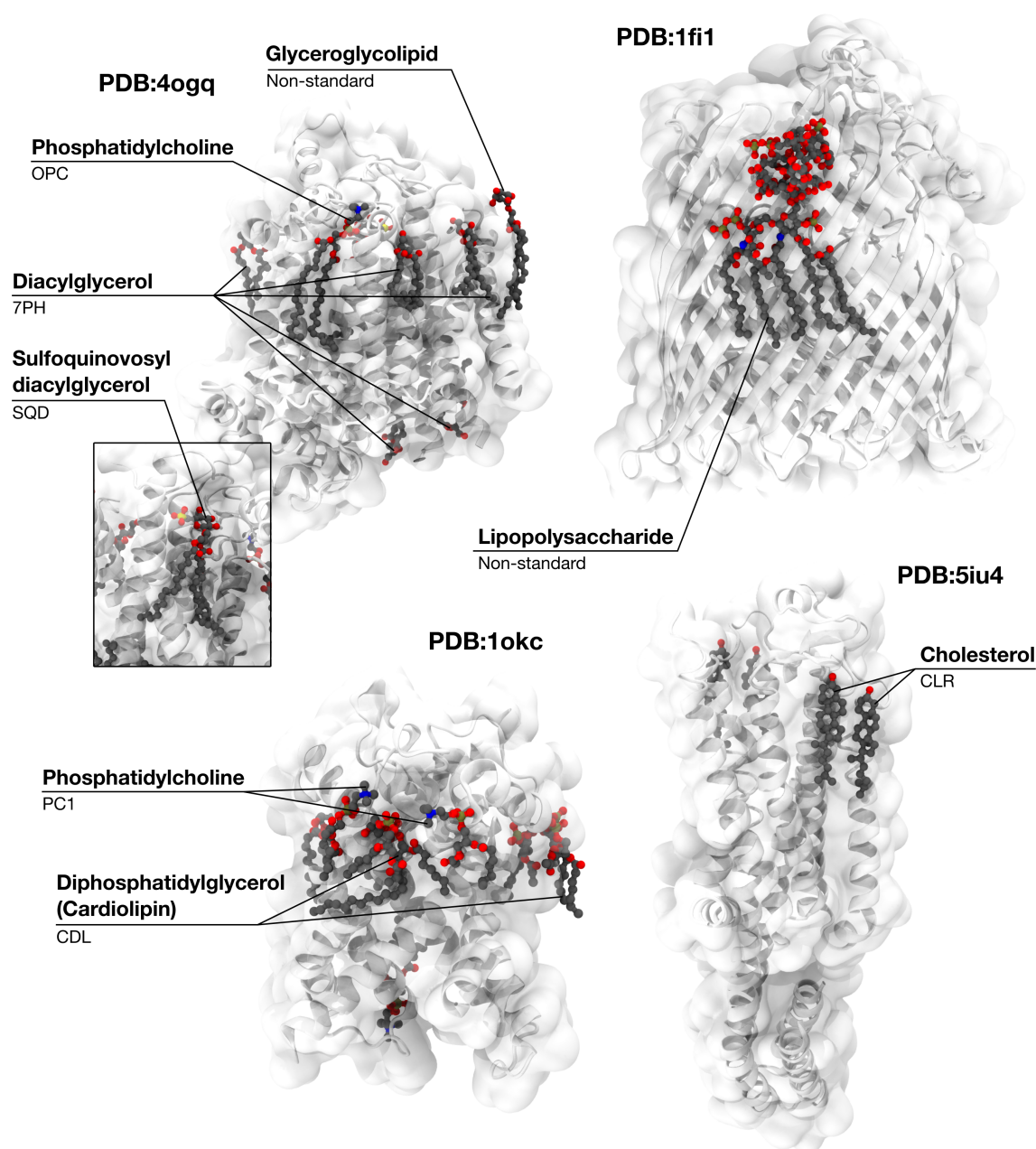


Figure 3.21: Examples of membrane lipids identified in PDB structures
Four IMP structures are shown in complex with a variety of identified lipid molecules. Where applicable, standard ligand IDs are indicated as well as the lipid's commonly used name.

		Membrane protein structures		
		Non-Standard ligand types		Standard ligand types
	Cholesterol	11	0	88
	Ergosterol	1	0	0
	Lanosterol	2	0	1
	Diacylglycerol	14	0	97
	Triacylglycerol	2	0	28
	Phosphatidic acid	20	0	28
	Phosphatidylcholine	28	0	149
	Phosphatidylglycerol	11	0	99
	Diphosphatidylglycerol	2	0	107
	2-tailed Diphosphatidylglycerol	5	0	87
	Phosphatidylethanol	1	0	0
	Phosphatidylethanolamine	18	0	134
	Phosphatidylinositol	4	0	1
	Phosphatidylinositol monophosphate	3	0	0
	Phosphatidylinositol biphosphate	4	0	8
	Phosphatidylinositol triphosphate	2	0	0
	Glycero-glycolipid	8	13	89
	Sulfoquinovosyl diacylglycerol	1	0	35
	Cerebroside/Ganglioside	4	0	0
	Sulfatide	6	0	3
	Ceramide	9	0	6
	Ceramide phosphate	6	0	0
	Sphingomyelin	3	0	0
	Lipid A/Lipopolysaccharide	3	18	18

Table 3.2: Lipid types found in PDB Ligand Expo and MemProtMD structures

For each type of ligand classified, the corresponding number of standard ligand types found in the PDB Ligand expo, the number of non-standard ligand types found bound to IMP structures, and the number of structures which contain at least one instance of that ligand type is shown. To the left is shown a tree representing various levels of classification as shown in Figure 3.20.

a different non-standard ligand. In many cases, membrane lipid molecules present in the PDB Ligand Expo are not found in any IMP structure, and are found only in soluble proteins which may be involved in lipid synthesis or transport.

Annotation of lipid chemical structures from structural data

Atoms of protein amino acids within 0.5 nm of membrane lipid atoms were identified and filtered to remove both protein and lipid atoms with temperature factors greater than 75 Å². Protein atoms were grouped according to the type of amino acid, using amino acid types determined by analysis of the MemProtMD database. Lipid atoms were aligned by considering the maximum common substructure between similarly classified lipid molecules. The results of annotations of a cholesterol and phosphatidylcholine structure are shown in Figure 3.22.

In the case of phosphatidylcholine, distributions of amino acid types match the likely position of lipid atoms within a lipid bilayer, with hydrophobic residues found near the ends of lipid acyl tails, and aromatic residues found near the lipid carbonyls. Around the head-group, the greatest diversity of amino acids is seen, with both positively and negatively charged residues, hydrophobic polar residues aspartic acid, glutamic acid and histidine, and residues with small side-chains such as glycine, proline, serine and threonine. It is interesting to note asymmetry between amino acids around atoms of acyl tails bonded at C1 and C2 of the glycerol group; the tail bonded at C1 is more commonly seen in contact with protein in PDB structures, although this is likely to be artificial since in many cases electron density for the glycerol group is highly ambiguous.

Also shown is an alignment of cholesterol molecules, in which distributions of amino acids are also indicative of the location of cholesterol in a lipid bilayer. Hydrophobic residues are seen mostly around the hydrocarbon tail, whilst a strong preference for aromatic residues can be seen around the cyclopentane ring, which typically sits at a similar point in a lipid bilayer to a phospholipid carbonyl. The hydroxyl group, carrying both positive and negative charge, can be seen occurring close to positively and negatively charged protein residues, as well as residues with small and polar side-chains.

3.5 Conclusions

MemProtMetrics has been used to assemble the most complete set of IMP structures, including several high-resolution structures omitted from the MemProtMD Database,

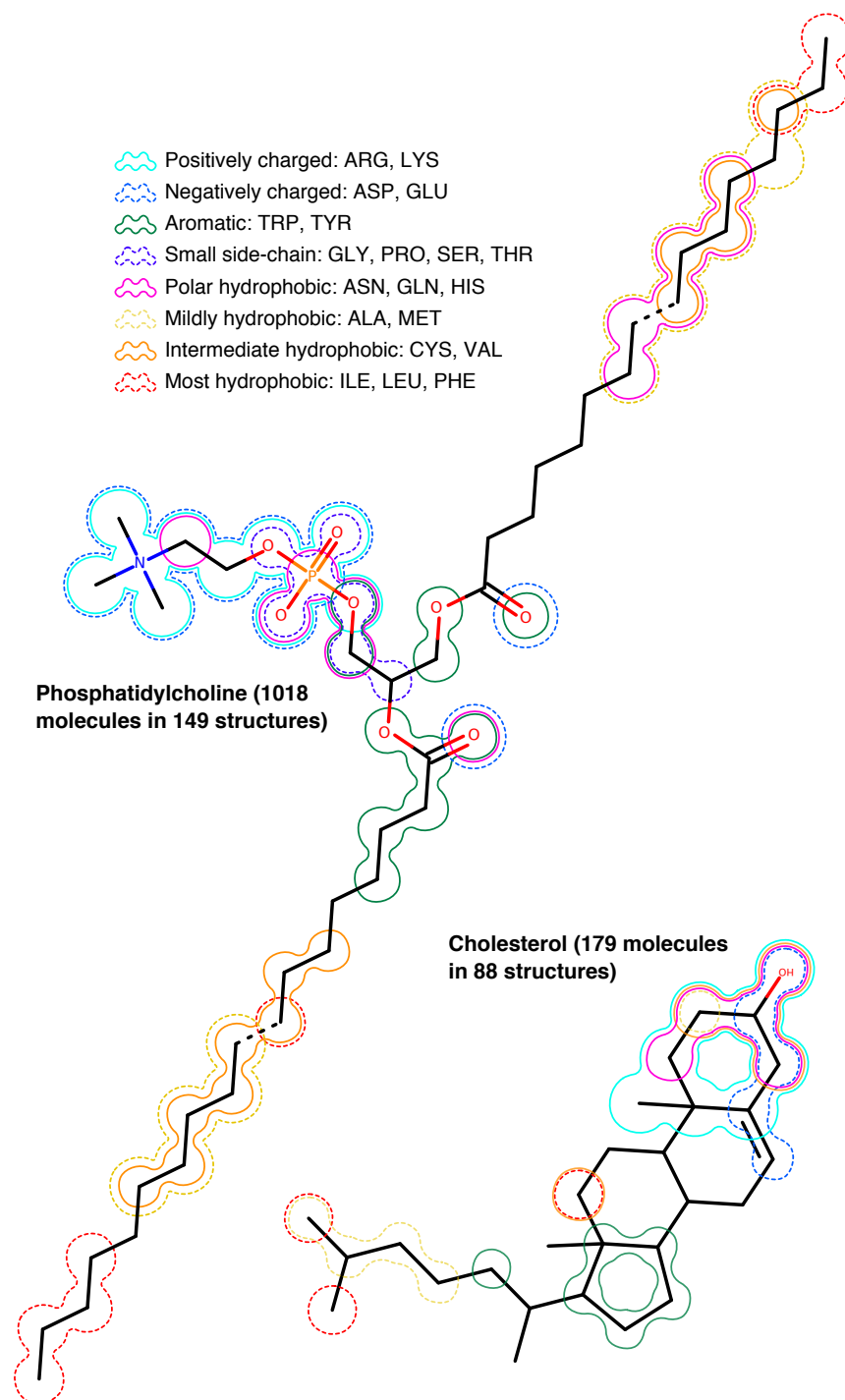


Figure 3.22: Structures of cholesterol and phosphatidylcholine lipids annotated using contacts found in PDB depositions

Alignments of all cholesterol and phosphatidylcholine ligands found in IMP structures are shown, overlaid with envelopes representing amino acid types found at least twice as frequently as expected within 0.5 nm of atoms in crystal structures. Envelopes are distinguished by colour and solid or dashed lines. Where both double and single bonds are found in aligned molecules, the bond is represented by a dotted line.

OPM and mpstruc. Bilayer surfaces output by MemProtMD metrics mostly provide an acceptable estimation of surfaces seen in simulations, and provide good initial orientations for MD simulation. Since membrane thickness is assumed to be constant, estimated bilayer surfaces for outer membrane proteins tend to be too far apart, although this does not have an effect on the tilt of the protein.

Whilst OPM achieved more accurate orientations for more structures than MemProtMetrics or MEMEMBED, structure orientation using both MemProtMetrics and MEMEMBED is in good agreement with tilts seen in simulations of IMPs with explicit lipid bilayers, with the majority of structures deviating by less than 10° between simulated and estimated orientations. MemProtMetrics and MEMEMBED may be seen as complementary, with MemProtMetrics producing more accurate results for β -barrel proteins and MEMEMBED producing more accurate results for α -helical proteins. This suggests substantial potential for improvement, either by improvements of methods individually, or by development of a hybrid method combining strengths of both approaches.

Although for the majority of proteins MemProtMetrics was able to accurately detect the inner surface of pores and cavities, errors in this step were the most frequent cause of misaligned or low-scoring IMP structures, and so improving the reliability of this step would substantially reduce numbers of misalignments, as well as increase objective function values for low-scoring IMP structures seen in Figure 3.6B. Modifications to the objective function which took hydrophobic mismatch into account would likely improve performance with single-pass α -helical proteins, although this would also require accurate prediction of membrane thickness.

Several approaches were considered to predict membrane thickness from maxima, minima, gradient and kurtosis of the distribution of $\rho_{mem} - \rho_{wat}$, as well as the distribution of surface polar and aromatic residues, none were able to predict membrane thickness seen in simulations better than $R_{pearson}^2 = 0.32$. Since many of the rules determining spatial distribution of amino acid residues differ substantially between evolutionary clades and cell compartments [110], it is likely that any successful approach would need to integrate some degree of awareness of these factors.

Some limited cross-validation was performed on MemProtMetrics, which did not suggest the model had been overfitted. This should be developed further using an approach such as k-fold cross-validation to perform model fitting on several subsets of the MemProtMD database, and perform tests on proteins not used for model fitting. There is also evidence that proteins following SCOP folds which are less well represented in the MemProtMD database suffer greater prediction error, so an

awareness of protein families and folds during parametrisation would reduce potential for bias towards protein families which are over-represented by PDB structures.

Classification of IMP structures remains problematic since IMP structures represent such a small proportion of PDB depositions; even with a low false-positive rate, the absolute number of false-positives may be comparable to the number of true-positives. It is likely that development of a more sophisticated objective function would substantially reduce the false-positive rate, since many false-positives are caused by disordered hydrophobic proteins, amyloid fibrils and steroidogenic enzymes.

Using a fragment-based approach, lipids bound to IMPs have been identified and classified, producing what is so far the most complete dataset of protein-lipid interactions in 570 PDB IMP structures, almost 20% of structures currently included in the MemProtMD Database. Using alignments of these molecules, phosphatidylcholine and cholesterol structures have been annotated according to protein-lipid interactions, revealing general properties of likely high-affinity binding sites.

This approach is limited, however, by the accuracy of ligand coordinates modelled in PDB depositions. Accurately fitting molecular structures to electron densities is difficult, particularly in the case of lipid hydrocarbon tails, and it is likely that inaccuracies in modelling will introduce noise and artefacts to this dataset. At present, a temperature factor threshold of $75 \text{ \AA}^2$ is applied, however it is likely data quality would be improved by excluding low-resolution structures. A preferable approach for structures determined by X-ray crystallography would be to cross-reference modelled coordinates with original electron densities.

With refinement and development, this method could be used to combine experimentally characterised protein-lipid interactions to construct pharmacophore-like models of the general properties of a high-affinity binding site for a number of different types of lipid. These models could be used to predict specific protein-lipid interactions, suggest mutagenesis approaches to alter the affinity of these sites, or support MD simulation studies of protein-lipid interactions.

Chapter 4

Simulations of IMP structures in mixed-lipid bilayers

4.1 Abstract

Biological membranes are composed of a vast number of lipid types, each with unique chemical properties. This composition may vary between leaflets, near certain proteins, and over time, playing vital roles in protein regulation and signalling. Interactions of Integral Membrane Proteins (IMPs) with specific lipid types, and their behaviour in bilayers of various compositions is therefore of considerable scientific interest.

Recent mass spectrometry experiments have shown ion mode dependence in the binding of Phosphatidylcholine (PC) and Phosphatidylglycerol (PG) lipids to the IMP OmpF, which was not seen in the IMP AmtB. A Molecular Dynamics (MD) approach was used to examine ion mode dependence of lipid binding. Several Coarse-Grained (CG) simulations were set up of OmpF and AmtB in a mixed PC/PG bilayer with both default forcefield parameters, and parameters modified to approximate a lower than physiological pH. PC and PG binding to OmpF in CG simulations was found to differ between these states, whilst binding to AmtB did not. This is consistent with data from mass spectrometry experiments, although the effect is less pronounced.

The Alchex (Alchemical Exchange) software tool has been developed with the aim of automating the process of altering the composition of a lipid bilayer in MD simulations. It is a robust method with a broad range of applications and can be downloaded from <https://github.com/tomnewport/alchex>.

TMEM16 proteins are a family of IMPs, of which many members are calcium-activated chloride ion channels or calcium-dependent lipid scramblases. By using a multi-scale MD simulation approach, a recently determined X-ray crystallographic

structure of TMEM16K has been computationally reconstituted in a mixed-lipid bilayer. Scramblase activity is observed in CG simulation, however conformational change of the protein prevents this in atomistic representation. Stable binding of calcium ions is found close to a groove involved in lipid translocation, suggesting a possible mechanism for the calcium activation of ion transport and lipid scramblase activity.

4.2 Introduction

Biological membranes contain many thousands of unique types of Integral Membrane Protein (IMP) and membrane-associated proteins, surrounded by many hundreds of distinct membrane lipid types. Membrane lipid composition is highly specific, and varies between cell compartments, between membrane leaflets and over time, with changes in lipid composition fulfilling vital regulatory and signalling roles [234, 235]. An understanding of the complex interplay between IMPs and membrane lipids is therefore vital to understanding biological processes taking place in this environment, and requires both experimental and computational approaches to characterise specific protein-lipid interactions.

Native mass spectrometry is one approach which can be used to detect and identify specific lipids bound to IMPs [60, 236]. In a native mass spectrometry experiment, proteins are first inserted into protective detergent micelles[237], amphipols [238] or nanodisks [239]. The sample is ionised using a "soft ionisation" method such as Electrospray Ionization (ESI) [240] or Matrix Assisted Laser Desorption Ionization (MALDI) [241] to minimise damage to the protein. ESI ionisation may be applied in positive ion mode, where the ESI nozzle is kept at a positive electrical potential or negative ion mode, where the nozzle is kept at a negative electrical potential. The sample is then accelerated using an electric field, and, using sample time-of-flight and response to a magnetic field, the weights and charges of individual ionised particles can be determined. If the weight of the protein monomer and possible ligands are known, the proportion of particles representing each oligomeric state and each combination of bound ligands can be determined. This method is therefore highly applicable to the study of interactions between IMPs and membrane lipids.

IMPs are considerably more stable than initially expected under the ionised gas phase conditions of a mass spectrometry experiment [62], however, these conditions are far from physiological, and so are likely to substantially alter protein-lipid dynamics. For IMPs in solution, interactions with lipids are mostly governed by polar

interactions with lipid head-groups [242, 243]. In the gas phase environment of a mass spectrometry experiment, however, polar and electrostatic interactions are likely to be altered substantially by ionisation and lack of water [244].

Biological membranes are typically divided into two leaflets composed of lipids in similar orientations, with a considerable energy barrier to translocation of a lipid molecule between leaflets [245], known as flip-flopping. Proteins which are able to reduce this energy barrier are known as scramblases, and play an important role in the regulation of bilayer asymmetry and lipid composition, particularly in the Endoplasmic Reticulum (ER), where newly synthesised lipids must be redistributed between leaflets [246]. Proteins which actively transport lipids between leaflets are known as flippases.

Flippase and scramblase proteins such as TMEM16F [95] also play a key role in exposure of Phosphatidylserine (PS) lipids in the extracellular leaflet during apoptosis. Exposure of PS lipids is a highly conserved anti-inflammatory and immunosuppressive signal [235], and failure or sabotage of PS signalling are linked to a broad range of infections and non-infectious diseases [247]. Mutations affecting Phosphatidylcholine (PC) exposure are often linked to neurological disease [248], such as Angelman syndrome, a heritable disease resulting from a deficiency of the flippase protein ATP10C [249] which causes severe learning difficulties, ataxia and epilepsy.

Many scramblases are members of the TMEM16 family, also known as anoctamins, which also play roles in ion transport and regulation of other membrane proteins [250]. Experimental studies using labelled lipid molecules have identified some degree of scramblase activity in all TMEM16 molecules, dependent on cytosolic calcium [41]. Considerable scramblase activity is seen in TMEM16C, TMEM16D, TMEM16F, TMEM16G, and TMEM16J. The publication of a high-resolution X-ray crystal structure of a fungal nhTMEM16 protein shown experimentally to exhibit scramblase activity [96] has shed considerable light on the mechanism by which TMEM16 facilitates lipid flip-flopping, and Molecular Dynamics (MD) simulations using this structure have identified a hydrophilic groove through which lipid head-groups can pass while lipid tails interact with the hydrophobic core of the membrane [97].

Many TMEM16 proteins act as calcium-activated chloride channels, in some cases in addition to their role as lipid scramblases. It has been suggested that, since in many cases scramblase activity and ion conduction are both calcium-dependent [251], this ion channel activity may occur through the lipid translocation groove, in tandem with lipid flip-flopping [98]. One such channel which does not appear to have scramblase

activity, however, is TMEM16A, for which a low-resolution crystal structure has been determined by Cryo-Electron Microscopy (Cryo-EM) [94].

Many recent advances in the understanding of protein-lipid interactions have come about in part with the use of MD simulations. MD can simulate proteins in either atomistic representation, or use a coarse-grained representation such as the MARTINI forcefield [119] in which groups of atoms are represented by a single bead, reducing complexity and facilitating simulations of systems larger in space or time than that which would be achievable by atomistic simulation. Computational studies of IMPs can use MD to reconstitute an IMP structure in a lipid bilayer *in silico*, and study interactions between the protein and a variety of lipid types.

A great strength of MD is the ability to simulate a variety of membrane types, composed of many different lipid types, and several methods exist to create new lipid bilayers with a variety of lipid compositions, including the Charm-GUI membrane builder [65] and Insane [205]. Equilibration of these bilayers is resource intensive, however, and so where an existing equilibrated bilayer is available a common approach is to use EXCHANGE_LIPIDS [252], which uses a fragment-based approach to exchange between several lipid types. This approach, however, is limited in application by the small number of lipid models which can be exchanged.

Here, two MD studies of IMP structures are presented, and a software tool designed to alter composition of lipid bilayers in MD simulations, known as Alchex (Alchemical Exchange), is described.

In a recent native mass spectrometry study, it was observed that by changing ion mode from positive to negative, binding of PC and Phosphatidylglycerol (PG) lipids to the IMP OmpF was substantially altered, with PC binding only in negative-ion mode and PG binding only in positive ion mode. Conversely, PG and PC were found bound to another IMP, AmtB, regardless of ion mode. It seems likely that this difference is due to the considerably larger number of acidic and basic residues on the surface of OmpF compared to the surface of AmtB. Presented here is a series of Coarse-Grained (CG) MD experiments which use a modified forcefield to simulate the effects of protonation on protein-lipid interactions. Analysis of these interactions suggests differences in ion mode dependence between AmtB and OmpF observed in the gas phase also exist to some small degree in solution.

To alter the composition of an equilibrated lipid bilayer in an MD simulation, the Alchex software tool [253] has been developed as an extension to the EXCHANGE_LIPIDS approach, and is described here. Using Alchex, simulated molecules can be replaced with any user-provided model, and leaflet compositions can be automatically altered

to match any desired symmetric or asymmetric composition, whether in planar or vesicle bilayers. The tool can be applied to both CG and atomistic simulations.

A crystal structure has been determined for TMEM16K, a TMEM16 protein linked to cerebellar ataxia [254]. Presented here are coarse-grained and atomistic simulations of this structure in symmetric and asymmetric mixed-lipid bilayers. Simulations show some degree of scramblase activity with a clear preference for flip-flops of PC lipids, consistent with the role of TMEM16K mutants in inflammatory neurological disease. Binding of calcium ions located near the lipid translocation groove is shown to be stable.

4.3 Materials and methods

4.3.1 Coarse-grained molecular dynamics simulations of AmtB and OmpF

To investigate the ion mode dependence of lipid binding to AmtB and OmpF, ten 1 μ s coarse-grained molecular dynamics simulations were performed of AmtB and OmpF in explicit bilayers composed of equal parts 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) and 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoglycerol (POPG). In five of these simulations, the forcefield was modified to approximate conditions at lower than physiological pH.

Protonation of a coarse-grained forcefield

Simulations were performed using the MARTINI coarse-grained forcefield [119, 118, 120], which reduces groups of three or four heavy (non-hydrogen) atoms to single particles. Default MARTINI forcefield parameters are calculated for physiological pH, and must be modified to simulate higher or lower pH values. To simulate the effects of a lower than physiological pH (pH 2) on a protein in a POPC/POPG bilayer, acid dissociation constants for amino acids (see Table 4.1) were used to predict changes in side-chain protonation at pH 2. Negative charges on side-chain beads of aspartate and glutamate were neutralised, and a positive charge was introduced to one histidine side-chain bead. Negative charges on the phosphate bead of POPC and POPG were also neutralised. These changes are illustrated in Figure 4.1. After forcefield protonation, net charge on AmtB was changed from 3 C to 69 C, and net charge on OmpF was changed from -36 to 90.

Table 4.1: pK_a of standard amino acids

Symbol	Name	pK_a	pK_b	pK_c	Sidechain ΔC (pH 7 $\rightarrow$ pH 2)
Ala	L-Alanine	2.33	9.71	-	
Arg	L-Arginine	2.03	9	12.1	
Asn	L-Asparagine	2.16	8.73	-	
Asp	L-Aspartate	1.95	9.66	3.71	+1
Cys	L-Cysteine	1.91	10.28	8.14	
Gln	L-Glutamine	2.18	9	-	
Glu	L-Glutamic Acid	2.16	9.58	4.15	+1
Gly	Glycine	2.34	9.58	-	
His	L-Histidine	1.7	9.09	6.04	+1
Ile	L-Isoleucine	2.26	9.6	-	
Leu	L-Leucine	2.32	9.58	-	
Lys	L-Lysine	2.15	9.16	10.67	
Met	L-Methionine	2.16	9.08	-	
Phe	L-Phenylalanine	2.18	9.09	-	
Pro	L-Proline	1.95	10.47	-	
Ser	L-Serine	2.13	9.05	-	
Thr	L-Threonine	2.2	8.96	-	
Trp	L-Tryptophan	2.38	9.34	-	
Tyr	L-Tyrosine	2.24	9.04	10.1	
Val	L-Valine	2.27	9.52	-	

Acid disassociation constants for amino acid carboxyl (pK_a), NH_2 - (pK_b) and sidechain (pK_c), and predicted change of side-chain between pH 7 and pH 2. Acid disassociation constants from [255].

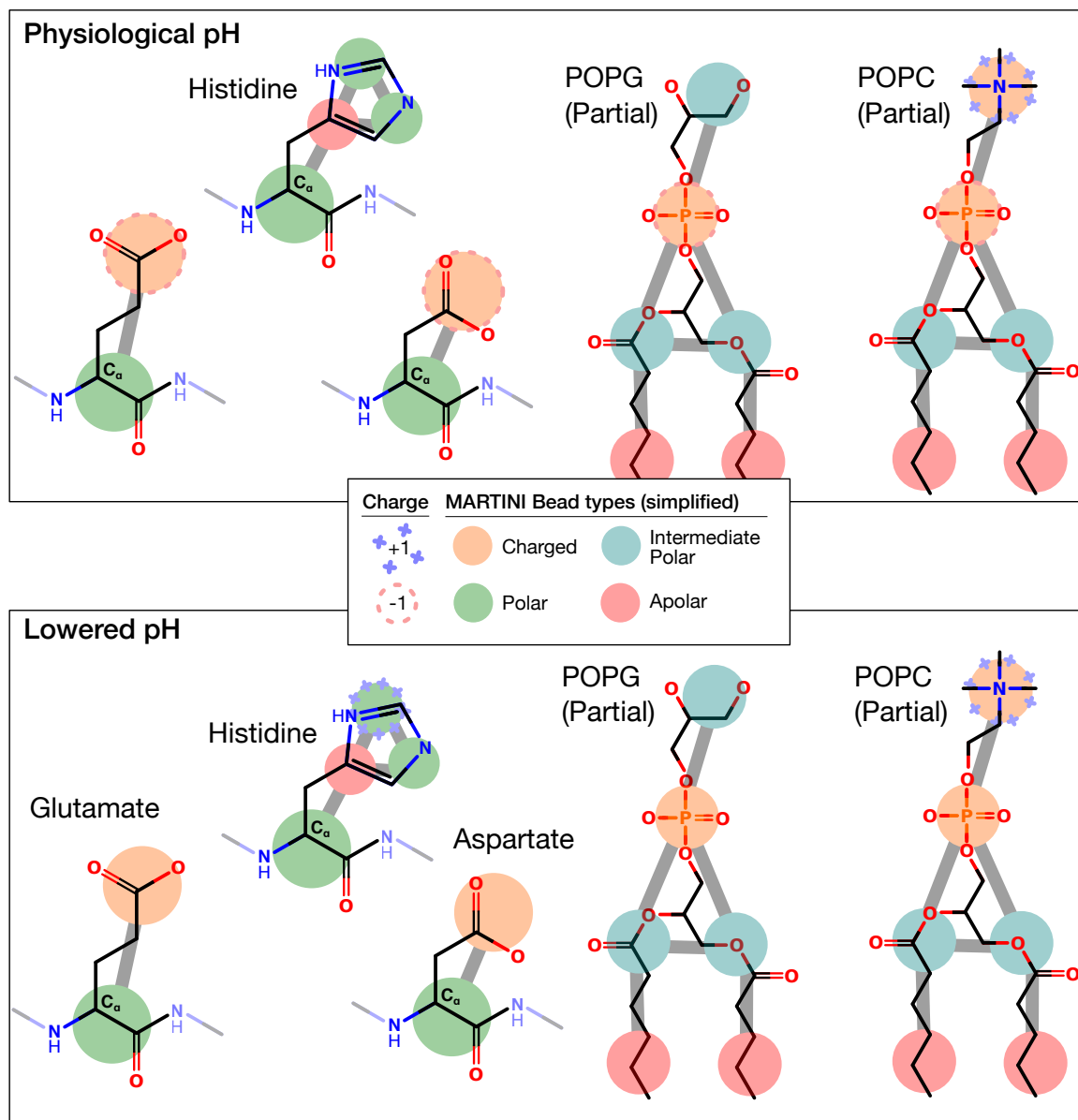


Figure 4.1: Protonation of the MARTINI forcefield

Simplified MARTINI beads and bonds are shown with full molecular structures overlaid for both physiological pH and lowered pH. To simulate effects of a lowered pH, negative charges on side-chain beads of glutamate and aspartate are neutralised and a positive charge is introduced to a histidine side-chain bead. The phosphate beads of POPC and POPG are also neutralised.

Simulation setup

Structures were obtained from the Protein DataBank (PDB) for AmtB and OmpF, (PDB:1u7g and PDB:2zfg respectively), and the MemProtMD pipeline [5] was used to set up ten 1 μ s Coarse-Grained Self-Assembly (CGSA) simulations for each protein with a 1:1 mix of POPC and POPG. For each protein, five of these simulations were run using the default MARTINI forcefield, whilst five were run using protonated parameters for histidine, aspartate, glutamate, POPG and POPC. Simulations were run using GROMACS 5.1 compiled with single precision [117] with a 20 fs timestep, with an elastic network applied to protein backbone beads. The V-rescale algorithm [256] was used for temperature coupling.

Simulation analysis

Simulations were analysed using MDAnalysis [206]. Analyses were performed on the final 800 ns of simulation time, and the first 200 ns were discarded to allow for self-assembly and equilibration of the lipid bilayer. Contacts between head-groups of POPG or POPC and protein side-chain beads within 0.6 nm were calculated for each protein residue for each frame, and a POPG preference was calculated for each residue as $\frac{N_{pg}}{N_{pg}+N_{pc}}$, where N_{pg} and N_{pc} are the number of frames in which a given residue makes contact with POPC and POPG respectively. System-wide preference for POPG binding was calculated by taking the mean of POPG preference for all residues contacting a lipid head-group for at least 30% of analysed frames. Density of POPG and POPC headgroups around the protein was visualised using PyMol [257]. Additional visualisation was performed using VMD [258].

4.3.2 Alchex: automated alteration of membrane composition

IMPs inhabit a broad variety of membrane types, each with a unique composition of lipid types. For the purposes of MD studies of IMPs, therefore is often useful to alter membrane composition by replacing individual lipid molecules to minimise disruption of an assembled bilayer. Alchex (ALChemical EXchange) is a software tool designed to automatically replace molecules in CG or atomistic MD simulations, either on an individual basis, or in bulk to match a specified composition. Target compositions may be applied to specific leaflets, within a three-dimensional bounding box, or using any valid MDAnalysis [206] selection string.

Alchex performs a number of replacements to achieve a requested composition, introducing a **replacement** molecule and manipulating its coordinates to match, as closely as possible, those of a **template** molecule. The template molecule is then removed.

Where a replacement molecule shares a substantial Maximum Common Subgraph (MCS) (based on atoms and bonds/constraints) with a template molecule, a topology-based replacement is performed. Atoms within the MCS are common to both molecules and so atoms of the replacement molecule are given identical coordinates. MCS calculation ignores atom and bond types, so replacement molecules may contain different atoms with a similar topology to that of the template molecule. To calculate coordinates of replacement molecule atoms not within the MCS, remaining atoms are split into connected components and attached to at least three closest atoms within the MCS. Atoms are then aligned to the template molecule by aligning attached MCS atoms to their counterparts by least-squares fit. Where atoms of the template molecule have no counterpart in the replacement molecule and a vacuum would otherwise be created, parts of the replacement molecule are stretched to fill this vacuum.

A special case exists for replacement of two template molecules with a single larger replacement molecule. Pairs of nearby template molecules are first identified and combined to create a single template with two connected components. MCSs are determined between each component of the template and the replacement molecule, and the shortest path connecting these components within the replacement molecule is identified. Atoms and bonds of the replacement molecule on this path are added to the combined template, connecting the two components. Further replacement may now proceed as previously described. Illustrations of this process, as well as examples of Alchex output can be seen in Figure 4.2.

Topology-based replacement is inappropriate in cases where molecules do not share a significant MCS. In these cases, molecules will be replaced by least-squares fit aligning a mapping between at least three atoms supplied by the user. This allows for small molecules such as cholesterol to be oriented correctly relative to the membrane.

New molecular structure and topology files are created incorporating newly created molecules and removing original template molecules. At this point, the molecular structure may contain clashes or inappropriate bond lengths or angles, and so a steepest-descent energy minimisation is performed using GROMACS, typically for up to 50,000 steps, or until all forces are less than 50 kJ/mol/nm, although user-supplied parameters may be used to run more or less stringent energy minimisations

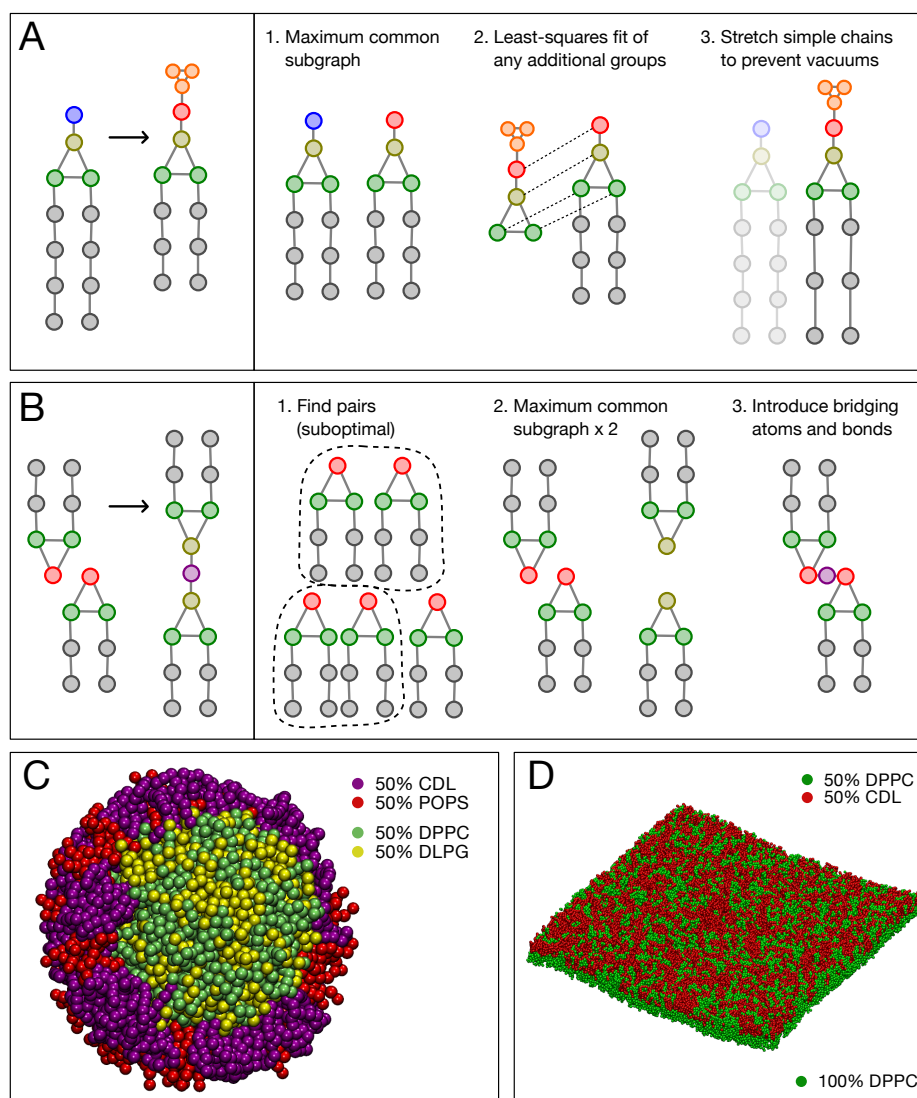


Figure 4.2: Illustration of the Alchex molecule replacement method

(A) a lipid molecule is replaced by another lipid with shorter acyl tails and a larger head-group. This is accomplished by 1. direct replacement of shared atoms by MCS, 2. adding accessory groups not within the MCS by least-squares fit to bonded atoms within the MCS, 3. stretching of simple chains to fill vacuums. (B) two lipid molecules are replaced by a single 4-tailed lipid (such as cardiolipin). This is accomplished by 1. finding nearby pairs 2. finding two instances of the MCS within the target structure for direct replacement of shared atoms 3. adding bridging atoms to connect two halves. (C) results of replacing molecules of a 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine (POPE) with a 50% cardiolipin/1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoserine (POPS) mixture in the outer leaflet and a 50 % dipalmitoylphosphatidylcholine (DPPC)/1,2-Dilauroyl-sn-glycero-3-phosphorylglycerol (DLPG) mixture in the inner leaflet. (D) results of replacing 50% of upper leaflet residues in a pure DPPC bilayer with cardiolipin.

where required. The resulting structure and topology files may now be used to run further simulations.

Testing Alchex lipid replacement

Alchex has been successfully used in several simulations, and has been confirmed by visual inspection not to introduce long-lived artefacts to the behaviour of the system. In addition, the impact of Alchex alteration of lipid tail length on bilayer properties has been tested using serial substitutions performed on a simple simulation of a lipid bilayer in water.

A lipid bilayer was set up using Insane [205], containing 338 CG DPPC molecules and 4541 CG water molecules. Steepest-descent energy minimisation was performed, and a 200 ns simulation was then run to allow the bilayer to equilibrate. The lipids 1,2-Dilauroyl-sn-glycero-3-phosphorylcholine (DLPG) and 1,2-Dilignoceroyl-sn-glycero-3-phosphorylcholine (DXPC) were chosen to perform substitutions, having, respectively, one fewer and two more tail beads than DPPC in CG representation.

After equilibration, all DPPC molecules were converted to DLPG using an Alchex automated substitution, and a further 200 ns simulation was performed. This was then repeated with substitution of resulting DLPG coordinates for DXPC, and finally substitution of DXPC for DPPC. Bilayer thickness was measured for each frame of each simulation, and can be seen plotted against simulation time in Figure 4.3. In each case, the final coordinates from the previous simulation were used as starting coordinates for Alchex substitution.

From this, it can be seen that, after automated substitution using Alchex, bilayer thickness is perturbed and takes around 20 - 50 ns to equilibrate. Larger perturbations result in greater time required for equilibration, with equilibration after substitution of DXPC for DLPG taking longer than substitution of DLPG for DPPC. Thickness of the initial and final DPPC bilayers is identical once equilibration is completed.

This suggests that alteration of lipid tail length using automated Alchex substitution is unlikely to introduce artefacts of membrane thickness persisting beyond 50 ns. Future development Alchex will be aided by similar tests looking at properties such as bilayer curvature, lipid order parameter and lipid diffusion, and introducing a variety of integral and peripheral membrane proteins to the system.

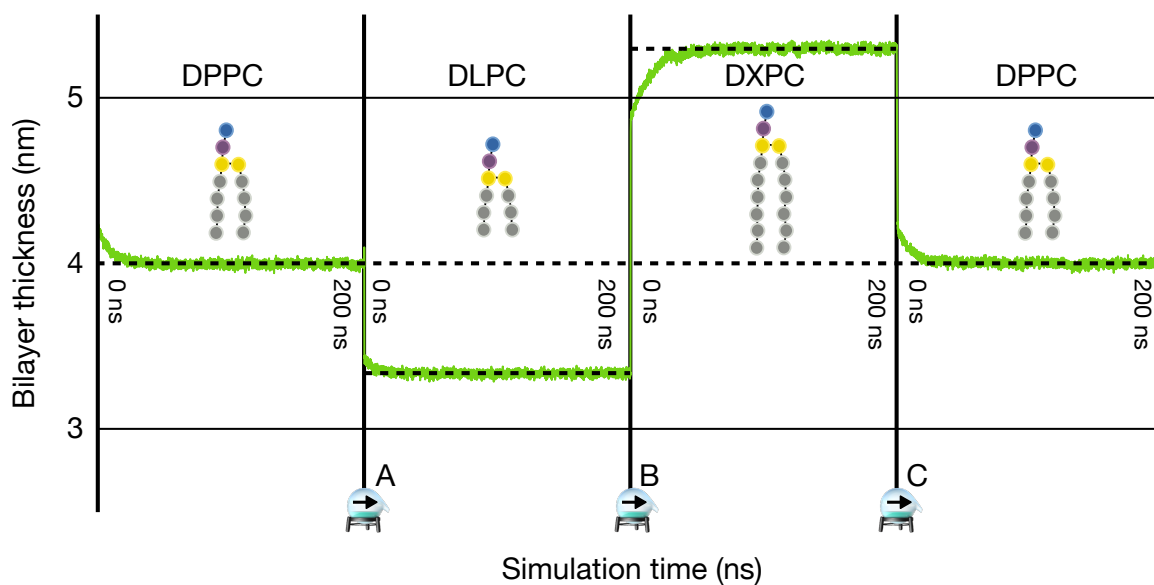


Figure 4.3: Impact of Alchex lipid replacement on bilayer thickness

Bilayer thickness, calculated for each simulation frame as the mean distance between lipid phosphate beads, is plotted over time for four separate simulations. Initially a pure DPPC bilayer is set up using Insane [205] and run for 200 ns. An automated Alchex substitution (A) is performed to replace all DPPC molecules with DLPG, and the simulation is run for a further 200 ns. This is repeated with substitution of DXPC for DLPG, and finally DPPC for DXPC, matching the initial bilayer composition. Idealised bulk thickness of a DLPG and DXPC bilayer is shown as an additional dashed line.

4.3.3 Multi-scale molecular dynamics simulations of a novel TMEM16K scramblase structure

A novel structure of TMEM16K (also known as Anoctamin-10) has been resolved by X-ray crystallography in Lipidic Cubic Phase (LCP) to 3.2 Å. To investigate whether ion channel or scramblase behaviour seen in homologous proteins is also seen in this structure, several CG and atomistic MD simulations were performed.

Bilayer insertion using CGSA

Two loop regions of four amino acids each were modelled in using MODELLER¹ [259]. The MemProtMD pipeline [5] was used to set up two 10 µs CGSA simulations of TMEM16K in a pure DPPC bilayer. Simulations were run using GROMACS 5.1 compiled with single precision [117] with a 20 fs timestep, using the V-rescale algorithm [256] for temperature coupling. The protein backbone was constrained using an elastic network.

Simulation of an ER membrane

TMEM16K is thought to reside in the ER membrane, and so Alchex was used to vary lipid composition to match that of an ER membrane. 17 frames were taken at randomly selected times from each 10 µs CGSA simulation, for a total of 34 frames, and membrane composition in each frame was then altered using Alchex. 17 frames were given an asymmetric ER membrane composition, with a 45:30:15:10 ratio of POPE, POPC, POPS and Phosphatidylinositol 4,5-bisphosphate (PIP₂) in the lower leaflet and an upper leaflet composed entirely of POPC. The remaining 12 frames were given a symmetric 45:30:15:10 ratio of POPE, POPC, POPS and PIP₂ in both leaflets. A 1 µs CG simulation was run from each frame, using the same simulation parameters and elastic network as used for CGSA simulations. ten simulations were extended to run for a further 2 µs, making for a total of 44 µs of mixed-lipid CG simulation time.

Once completed, mixed-lipid CG simulations were analysed to identify upper and lower leaflets, and lipids moving between leaflets, using the MemProtMD analysis application (see Chapter 2). To identify specificity of lipid interactions and flip-flop activity at each protein residue, the proportion of frames for which at least one side-chain bead was within 0.6 nm of the head-group beads of each lipid type (PIP₂,

¹Modelling of loop regions performed by Dr Phillip Stansfeld

POPC, POPE and POPS), as well as any bead of a lipid currently moving between leaflets was assessed.

Conversion to atomistic simulation

A single frame from a mixed-lipid CG simulation was selected and converted to atomistic representation in the GROMOS 53a6 forcefield [125] using CG2AT [122]. Seven calcium ions were resolved with the protein structure but had not been included in CG simulations, and six of these which were within 0.35 nm of a protein side-chain were added to the simulated atomistic structure. Five 300 ns atomistic simulations were performed using GROMACS 5.1 compiled with single precision [117] with a timestep of 2 fs, using the PME algorithm [260] for electrostatic interactions, the V-rescale [256] algorithm for temperature coupling and the Parrinello-Rahman algorithm [261] for pressure coupling. A total of 1.5 μ s of atomistic simulation was performed.

To visualise protein dynamics and identify stable equilibria, Root-mean-square deviation (RMSD) of protein Carbon- α atom positions of each chain was calculated between each frame and the starting structure. To quantify the dynamics of individual residues, DSSP [210] was used to assess protein secondary structure for each frame, and Root-mean-square fluctuation (RMSF) was calculated for each residue.

To assess the stability of calcium ions identified in the crystal structure, contacts between calcium ions and protein side-chain atoms within 0.35 nm were recorded for each frame of each simulation repeat using MDAnalysis [206].

Visualisation

Molecular visualisation was performed using VMD [258], with secondary structure calculated by STRIDE [262].

4.4 Results

4.4.1 Ion mode dependence of lipid binding to AmtB and OmpF

CG simulations were set up with both the default MARTINI forcefield and a protonated MARTINI forcefield to investigate the ion mode-dependence of lipid binding to AmtB and OmpF. Both AmtB and OmpF can be seen in assembled lipid bilayers in Figure 4.4. As seen in Figure 4.5A, whilst protonation of the entire system resulted no significant change in the POPG preference of AmtB averaged across all

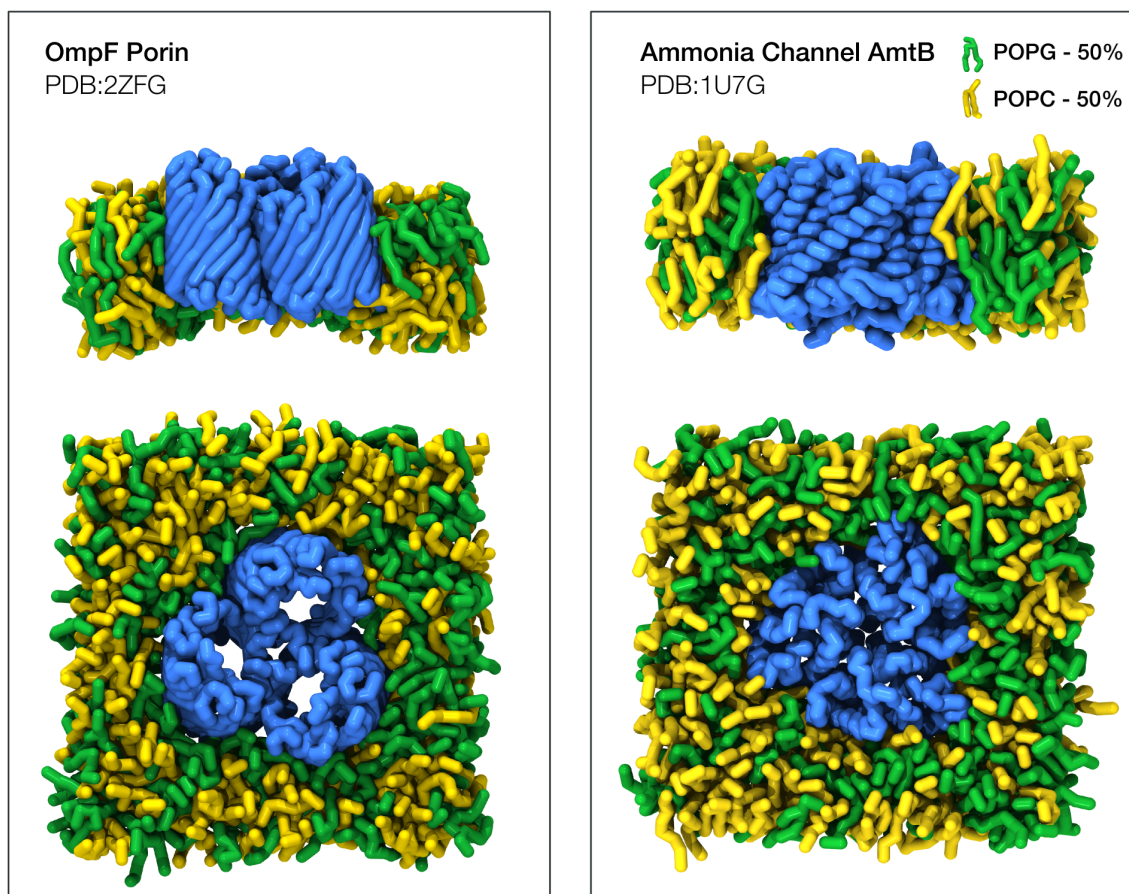


Figure 4.4: Coarse-grained representations of AmtB and OmpF in lipid bilayers

Top-down and side-on views are shown for frames of CG simulations of AmtB and OmpF. Molecules are displayed using bonds between CG beads. Protein is shown in blue whilst POPG and POPE are shown in green and yellow respectively. Water beads, and lipids between the protein and viewer are removed.

residues, a small but significant decrease can be seen in the POPG preference of OmpF, amounting to around a 2% difference. In Figure 4.5B, densities of POPE and POPG headgroups are shown over OmpF and AmtB structures for both deprotonated and protonated systems. It can be seen for both proteins that protonation results in minimal changes in lipid binding, although POPG binding between OmpF monomers does appear to be somewhat reduced.

4.4.2 Alchex: Automated exchange of simulated molecules

During in-house trials, Alchex was used by several researchers to replace a variety of molecules including lipids and enzyme substrates in both CG and atomistic simula-

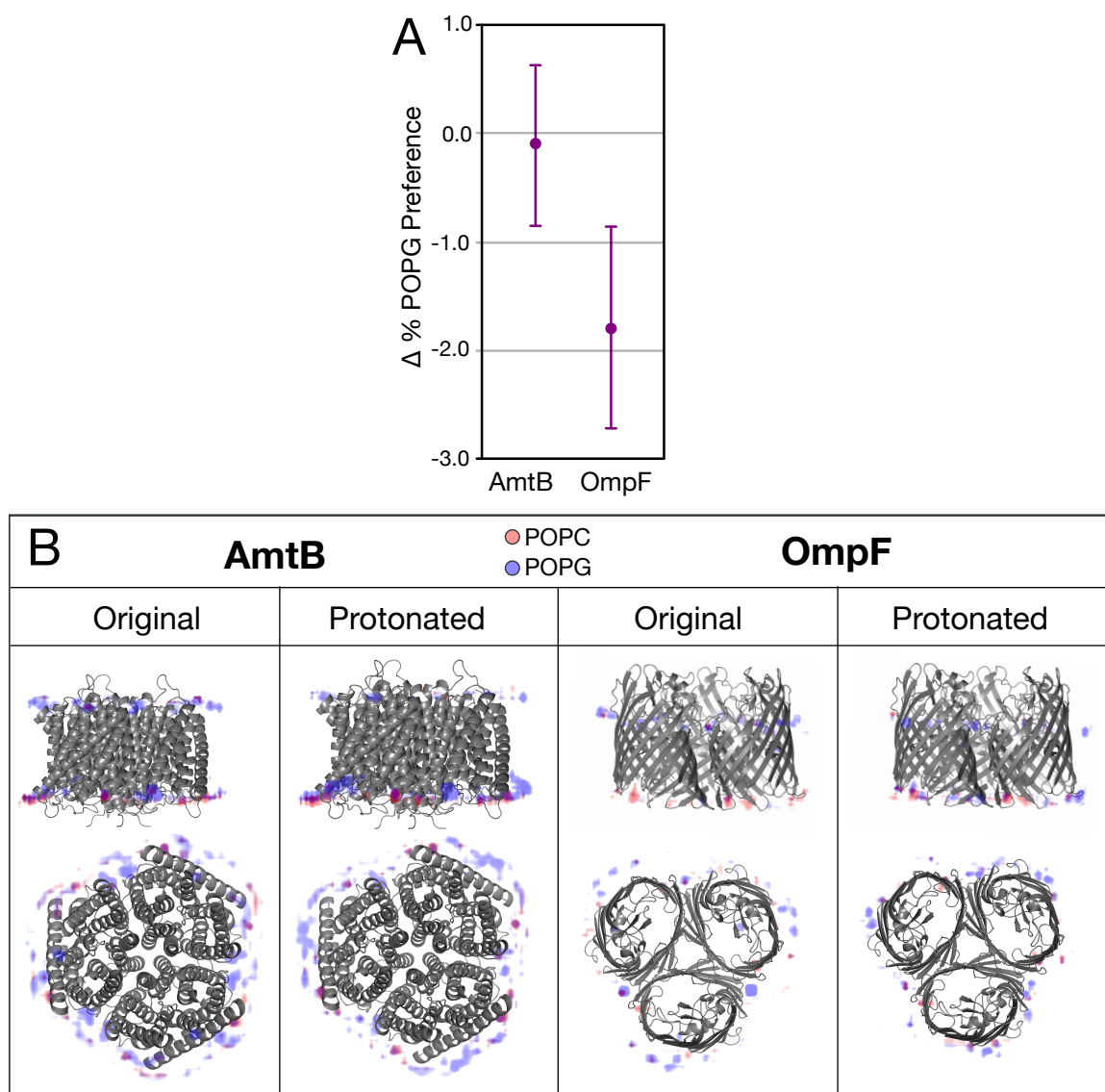


Figure 4.5: Effects of protonation on lipid binding to AmtB and OmpF
(A) change in mean preference for POPG with whole-system protonation compared with deprotonated state is shown for AmtB and OmpF. Error bars show 95% confidence intervals from a T-distribution. **(B)** Density of POPC (red) and POPG (blue) headgroups is shown over cartoon representations of AmtB and OmpF structures in both deprotonated and protonated states. Top-down and side-on views are shown.

tions. Resulting molecules were stable, although in some cases it was necessary to run 100,000 steps of steepest-descent energy minimisation. Following the success of this trial, a pre-alpha release of Alchex has been published [253] as a public git repository and can be downloaded from <https://github.com/tomnewport/alchex/>.

4.4.3 Multi-scale molecular dynamics simulations of a novel TMEM16K scramblase structure

To investigate potential scramblase or ion channel activity of a novel TMEM16K structure, a total of 34 μ s of CG simulation and 1.5 μ s of atomistic simulation was performed. Backbone restraints applied in CG simulations prevented large changes in protein conformation, however in atomistic simulations these restraints were not present, allowing changes in protein secondary structure and large-scale conformational changes.

RMSD plots shown in Figure 4.6 suggest that the conformation seen in the original LCP crystal structure is not far from a stable equilibrium in the presence of solvent and a lipid bilayer. It can be seen that each chain of each repeat rapidly undergoes conformational change amounting to a carbon- α RMSD of around 0.3 nm in the first 50 ns. Change then becomes more gradual, with some chains in some repeats appearing to reach a stable equilibrium, such as both chains of the second repeat, for which RMSD remains steady at 0.4 nm for the later 250 ns of simulation time, and in chain A of repeat 3, which remains at an RMSD of around 0.35 nm. In the majority of cases, however, RMSD indicates that the protein did not reach a stable equilibrium, and was sampling new conformational space throughout the simulation.

RMSF plots shown in Figure 4.7 quantify protein dynamics for each individual residue. RMSF for most residues is around 0.3 nm, however several regions can be identified for which RMSF is considerably higher, indicating greater flexibility. These typically occur in the extra-membrane domain of the protein (up to residue 170), and between transmembrane α -helices. This can be compared with protein secondary structure seen in simulations, also shown in Figure 4.7. Most elements of secondary structure are stable and remain intact for all simulation frames, with some less stable secondary structural elements corresponding with regions of elevated RMSF. These include the modelled region at the residue 350, which is flanked by unstable α -helices. Another modelled region around residue 180 also coincides with elevated RMSF and a less stable α -helix, although this is less pronounced.

To identify lipid specificity of sites on the protein surface, head-group contact frequencies were calculated for all CG simulations with asymmetric bilayers for PIP2,

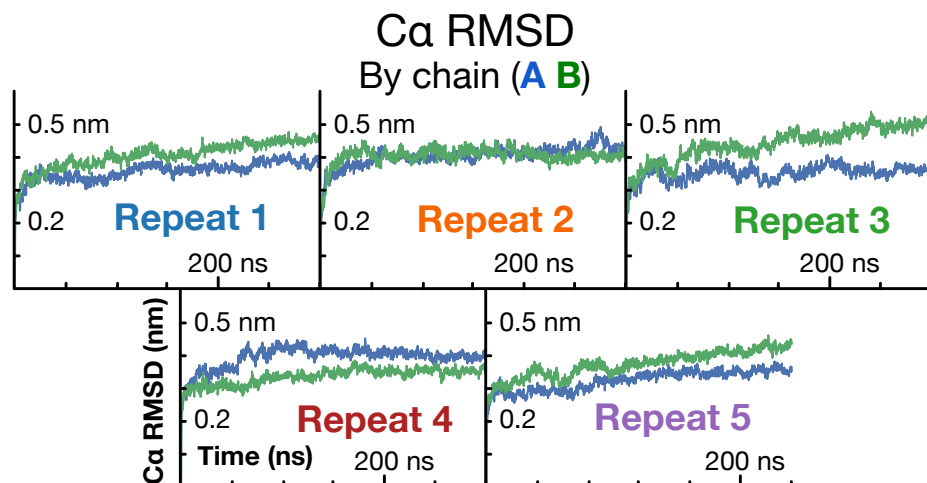


Figure 4.6: Carbon- α RMSD of a TMEM16K structure during atomistic simulations

RMSD is calculated separately for each chain using carbon- α atom positions for each frame of a 300 ns atomistic simulation of a TMEM16K structure.

POPS, POPC and POPE. These can be shown plotted along the sequence of TMEM16K in Figure 4.7, as well as colouring surface and cartoon representations of the TMEM16K structure in Figure 4.8. Changes in bilayer composition are negligible, with PIP2, POPE and POPS confined to the lower leaflet, and an upper leaflet composed entirely of POPC. Headgroup contacts shown over both sequence and structure show a tendency for PIP2 to cluster around the protein.

Scramblase activity is seen in CG simulations, with lipids flip-flopping between leaflets through a groove formed by two α -helices at residues 360-380 and residues 430-450. This can be seen in the flip-flop contacts shown over the protein sequence in Figure 4.7 and structure in Figure 4.8. Lipid head-group contacts near this groove show that this is one of few sites where PIP2 headgroups are depleted, with POPC and POPE accounting for the majority of lipid contacts. During CG simulations, whilst POPC, POPE and POPS lipids flip-flopped between leaflets, no flip-flop activity was seen for PIP2 lipids. This is shown in a bar chart in Figure 4.9. From this, it can also be seen that POPS lipids flip-flopped more frequently relative to the number of lipids in a given system, with POPC and POPE flip-flopping at similar rates.

No flip-flop events were seen in atomistic simulations, although lipids did enter the groove from both leaflets. Four lipids can be seen in different stages of flip-flopping in Figure 4.9. A more detailed view, also shown in this figure, shows the groove occluded by side-chain atoms of a glutamate residue and two asparagine residues, and it seems likely that this occlusion may be sufficient to prevent flip-flopping within the 300 ns

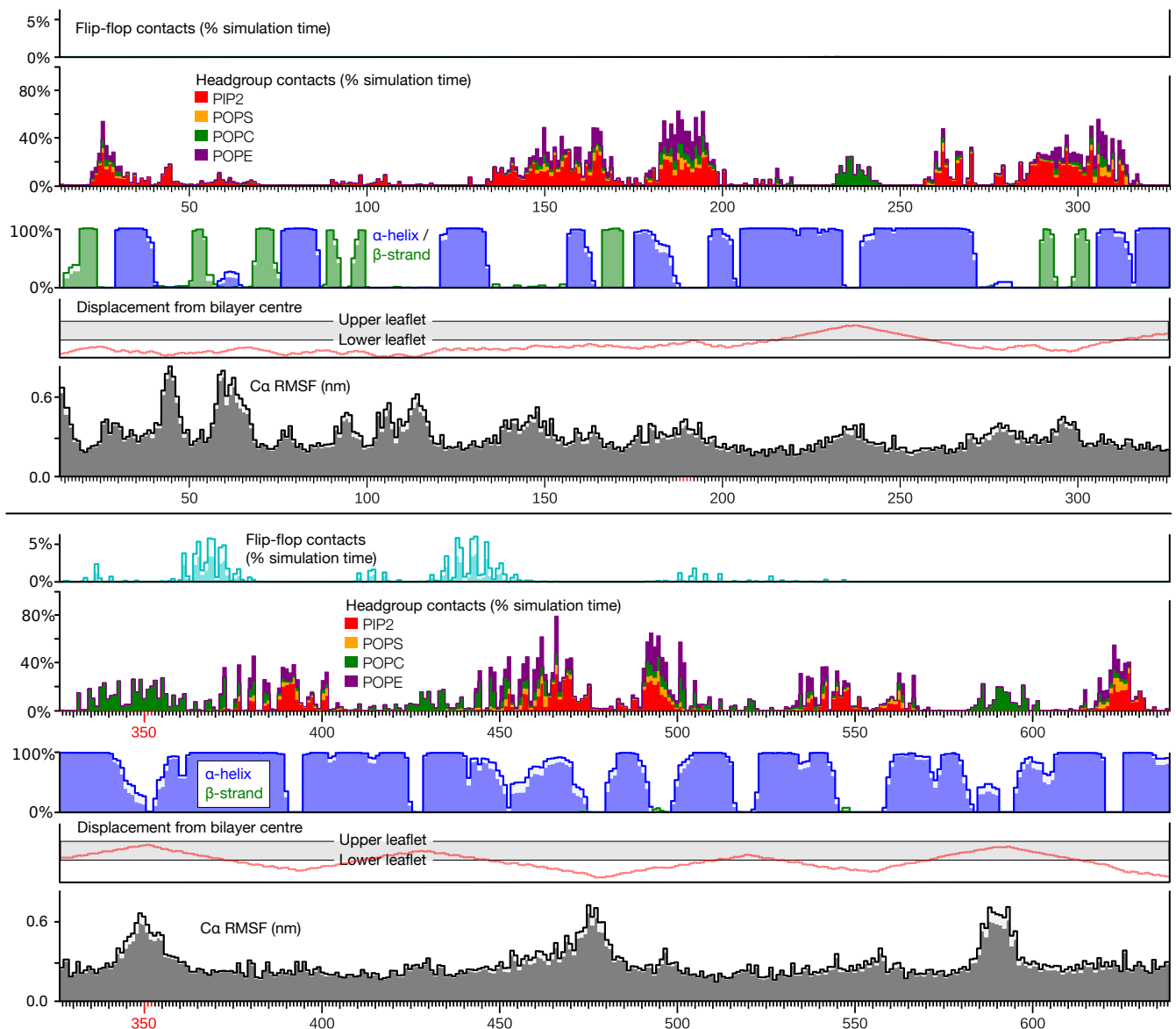


Figure 4.7: Lipid contacts, secondary structure displacement from bilayer centre and RMSF along the TMEM16K amino acid sequence

Simulation metrics are shown calculated for each residue along the sequence of TMEM16K. The sequence is split over two lines. From top to bottom: contacts with any part of any lipid in the process of moving between leaflets as a proportion of simulation time, plotted in cyan and overlaid with standard error for the mean shown in white; stacked plot showing contacts with headgroups of PIP2, POPS, POPC and POPE as a proportion of simulation time in red, orange, green and purple respectively; proportion of frames for which a residue is part of an α -helix or β -strand in blue and green respectively, with standard error for the mean overlaid in white; protein displacement from the bilayer centre shown in red, with approximate phosphate positions shown for each leaflet in black; and RMSF of the carbon- α atom of each residue plotted in grey with standard error for the mean overlaid in white. Hatch marks shown below some plots represent the TMEM16K protein structure. These are coloured red to indicate residues not present in the original structure.

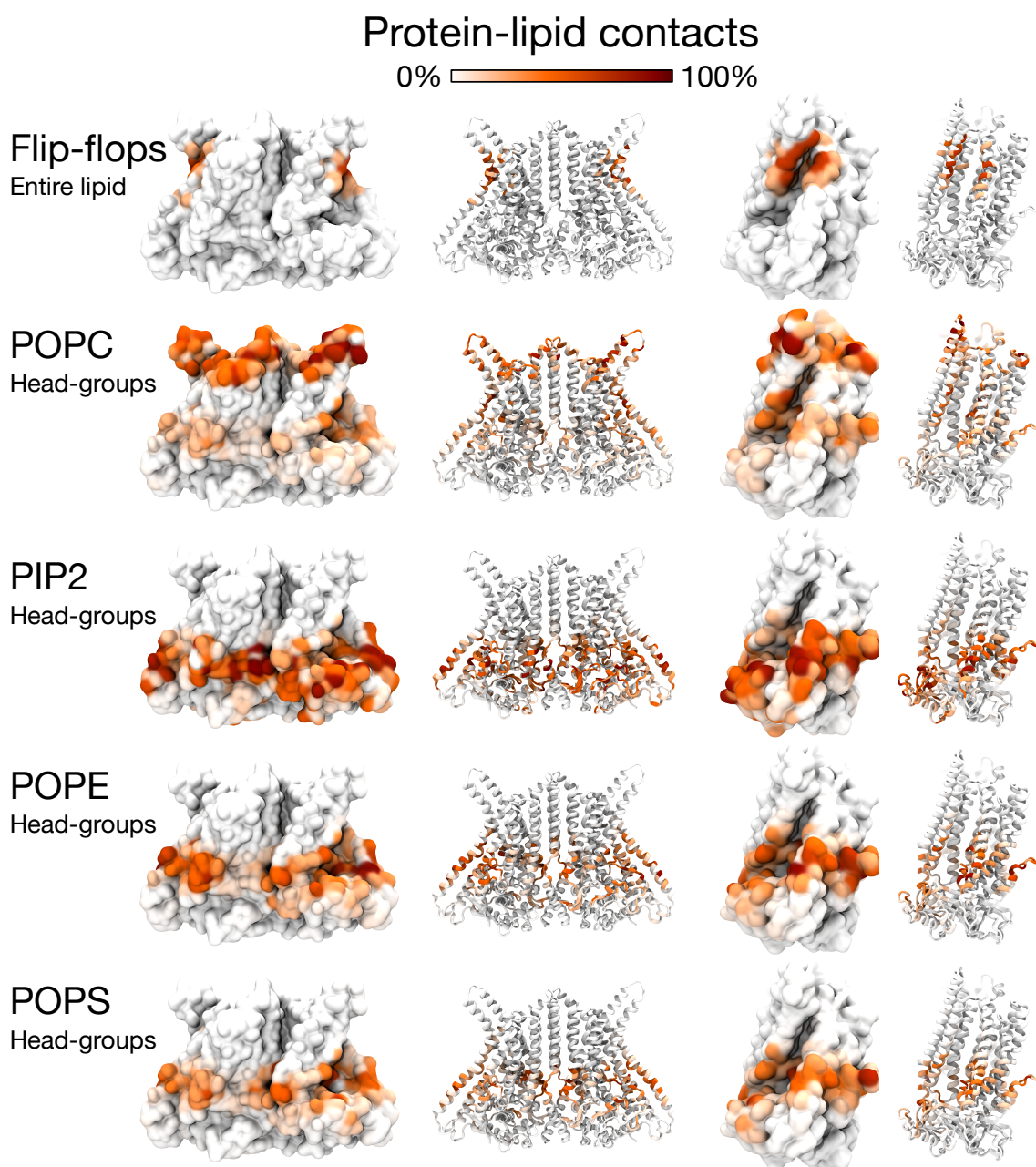


Figure 4.8: Surface and cartoon representations of TMEM16K coloured according to protein-lipid contacts

Protein surface and cartoon representations are shown from two different angles, coloured according to protein-lipid contacts seen in CG simulations. From top to bottom: contacts with any lipid in simulations with both symmetric and asymmetric bilayers and contacts with POPC, PIP2, POPE and POPS headgroups in simulations with asymmetric bilayers.

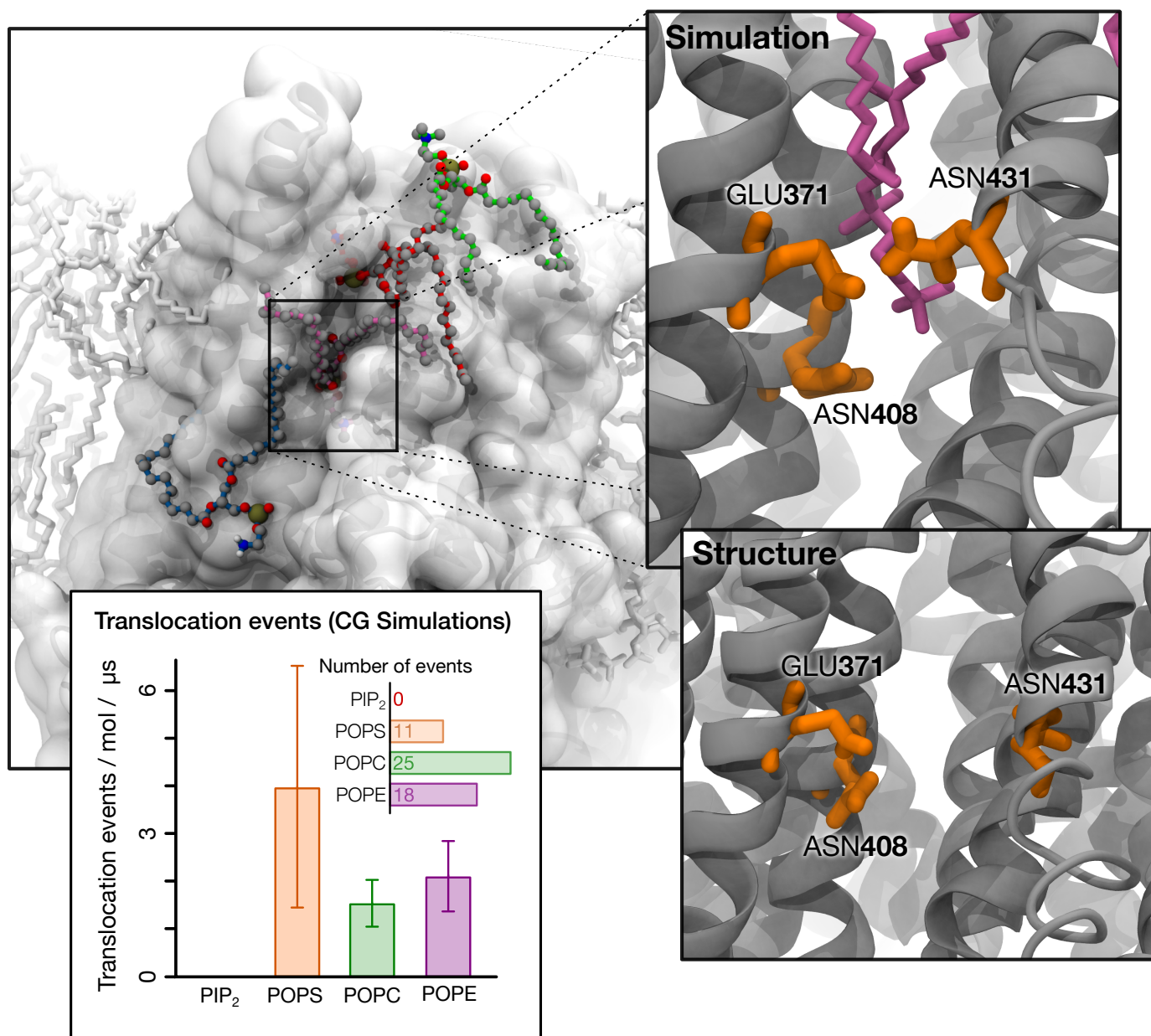


Figure 4.9: Flip-flopping of lipids between leaflets facilitated by a simulated TMEM16K structure

Four different lipids are shown at various stages of flip-flopping between leaflets through a groove formed by two alpha helices in a single frame of atomistic simulation. Detail (top right) shows a single lipid encountering residues 371, 408 and 431, which occlude the groove in most atomistic simulations, preventing lipid flip-flops from occurring. Below is shown the crystal structure, in which the three α -helices occupy different conformations, resulting in an unoccluded groove. Inset bottom right are shown flip-flopping frequencies for CG simulations based on the crystal structure. The number of events per molecule per microsecond is shown, with error bars representing standard error for the mean.

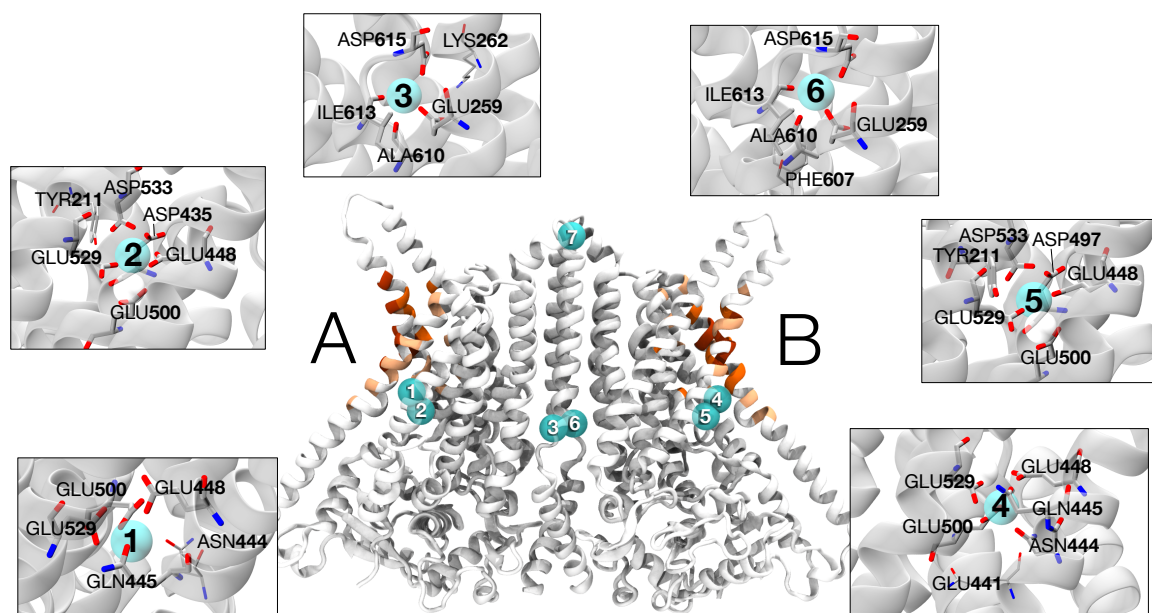


Figure 4.10: Calcium ions present in the TMEM16K crystal structure
Seven calcium ions present in the TMEM16K crystal structure are shown over a cartoon representation of the protein, coloured according to scramblase activity. Each ion is numbered, and, for the six ions appear bound to the protein, residues within 0.35 nm of calcium ion centres in either crystal structures or MD simulations are shown. Residues which contact calcium ions in MD structures alone are indicated by thinner bonds.

time-scale used for atomistic simulations. The α -helices to which these residues belong are in different conformations in the crystal structure used to determine backbone restraints used for CG simulations, and this may explain why flip-flopping is seen only in CG simulations.

Of seven calcium ions found in the TMEM16K crystal structure, six appeared to be buried within the crystal structure and made contact with several amino acid side-chains. During simulation, several calcium ions also made prolonged contacts with different amino acids. The positions of these calcium ions and amino acids can be seen in Figure 4.10. The calcium ion labelled 7 did not appear to be buried within the structure and so was not simulated, whilst ions labelled 3 and 6 exited their original sites almost immediately and did not form any further prolonged contacts.

Calcium ions labelled 1, 2, 4 and 5, however, remained in sites identified in the crystal structure for prolonged periods in most atomistic simulations. This can be seen in Figure 4.11, which shows contacts between calcium ions and amino acid side-chains for each frame of each repeat simulation. Great variation can be seen between

the five simulations, with ions remaining mostly in contact with the same set of amino acids in some simulations (such as ions 1 and 2 in the fifth repeat), or escaping entirely in others (such as ion 5 in the fifth repeat). Each binding site makes prolonged contact with a calcium ion from several amino acids in at least one simulation, with ion 4 being the most well-defined, forming prolonged contacts in all five repeats. It is interesting to note that whilst ions 1 and 4 and ions 2 and 5 occupy equivalent sites of their respective chains, their interactions with protein amino acids differs considerably.

4.5 Conclusions

4.5.1 Ion-mode dependence of lipid binding to AmtB and OmpF

By modification of the MARTINI forcefield, it has been possible to simulate mild protonation seen at a lower than physiological pH. The result of this appears to be a small but significant decrease in POPG binding to OmpF compared to POPC, but no significant change in lipid binding to AmtB. Whilst this is consistent with mass spectrometry experiments, the effect size is small, amounting to around a 2% decrease. This is perhaps unsurprising given that the AmtB assembly used here possesses 66 acidic residues whilst the OmpF used here possesses 126.

As mass spectrometry becomes increasingly used to examine binding of lipids to IMP, understanding of the dynamics of protein-lipid interactions under ionised gas phase conditions remains limited. The development of an MD methodology better simulating these conditions would be invaluable to better understand the results of mass spectrometry experiments. Any such approach would necessarily be complicated by the need to accurately model the effects of different ion modes and protective molecules such as amphipols, nanodisks and detergents used in mass spectrometry, about which very little is presently understood.

4.5.2 Alchex: automated alteration of membrane composition

The Alchex software tool provides a robust and scalable method for MD, automating the alteration of bilayer composition, as well as the introduction and manipulation of a broad range of ligand types, without specialist knowledge. The software interface still requires some improvement, as the version which appears in public pre-alpha release is somewhat unintuitive for new users.

Chain A

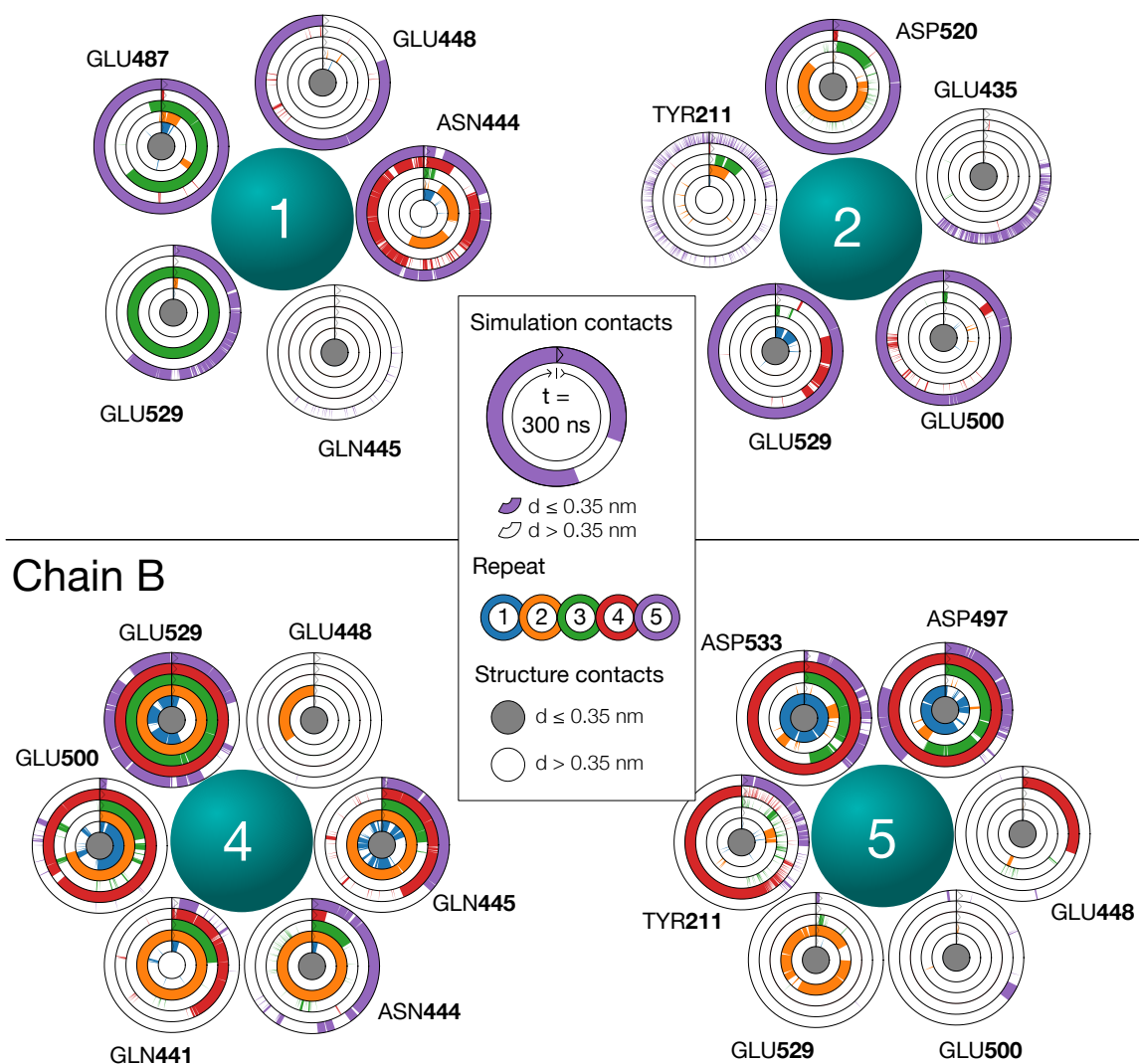


Figure 4.11: Contacts between bound calcium ions and amino acid side-chains

Contacts between calcium ions and amino acid side-chains are shown for five 300 ns atomistic simulation repeats for four calcium ions. Simulation repeats are arranged concentrically, with time running clockwise from the top. Sectors are coloured where a contact is made between a protein side-chain and an ion within 0.35 nm of the ion centre. The central circle is shaded grey if an amino acid side-chain is within 0.35 nm of the ion centre in the crystal structure.

4.5.3 Multi-scale molecular dynamics simulations of a novel TMEM16K scramblase structure

During atomistic simulations of TMEM16K, the protein remains in a conformation close to that seen during structure determination, with an RMSD of around 0.4-0.5 nm, as shown in Figure 4.6. RMSD does, however, suggest a larger conformational space not explored during the five 300 ns simulations. RMSF and secondary structure stability analysis shown in Figure 4.7 show that, for the majority of residues, the protein backbone and secondary structure are stable.

Lipid diffusion in atomistic simulations is too slow to provide adequate sampling of protein-lipid interactions, and so CG simulations are required to identify specific and non-specific lipid binding. Protein lipid interaction seen in CG simulations shown in Figure 4.7 and Figure 4.8 show that PIP2 clusters around the protein, but rarely enters lipid translocation groove. In the lower leaflet, POPC, POPE and POPS are all seen interacting with residues near the groove. Although only low levels of scramblase activity have been seen in experimental studies of TMEM16K [41], significant numbers of scramblase events were seen in CG simulations along the same lipid translocation groove seen in simulations of nhTMEM16 [97], as can be seen in Figure 4.8. Lipid flip-flops were not seen in atomistic simulations. This may be due to the shorter timescale used in atomistic simulations, with a total of 1.5 μ s as compared to 44 μ s of CG simulation, although it is also possible that constraints applied to the protein backbone based on the protein crystal structure in CG simulations introduced bias towards a less favourable conformation.

RMSD plots seen in Figure 4.6 support the idea that the protein undergoes a conformational change during atomistic simulation, and occlusion of the lipid scrambling groove due to conformational changes in protein structure, can be seen in Figure 4.9. The consequent lack of scramblase activity seen in atomistic simulations, suggests that *in vivo* lipid flip-flop rates would be low, or require some further conformational change of the protein. The lack of convergence seen in the majority of RMSD plots in Figure 4.6 is consistent with a wider conformational space not sampled in the 300 ns timescale.

The selectivity for POPS lipids of the TMEM16K scramblase mechanism seen in atomistic simulations, and shown in Figure 4.9, is consistent with the role of TMEM16F in translocation of PS lipids [95]. Given the role of TMEM16K mutants in cerebellar ataxia [254], a disease often linked to a failure of PS signalling and mutations of other proteins involved in lipid translocation [235, 249], this suggests a potential function of TMEM16K which merits further study.

Scramblase and chloride channel activity of other TMEM16 proteins is dependent on calcium ions, and four of the seven calcium ions identified in the crystallographic structure (shown in Figure 4.10) remained in well-defined sites for prolonged periods during simulations, as can be seen in Figure 4.11. These four ions are located close to the lipid translocation groove, and so it is entirely plausible that the binding and release of these ions results in movement of α -helices shown in Figure 4.9 to occlude the lipid translocation groove and thereby regulate scramblase activity. Longer atomistic MD simulations in the presence and absence of these ions are likely to shed considerable light on the regulation of TMEM16 scramblase activity.

Chapter 5

Conclusions

5.1 Results

The results of work presented in this thesis demonstrate the value of computational methods, and in particular, MD to the study of interactions between IMPs and the lipid constituents of biological membranes. Analysis and visualisation methods described in Chapter 2 provide new ways to analyse and visualise IMP structures, which go on to inform both the development of new IMP-specific methods detailed in Chapter 3, and studies of individual IMPs discussed in Chapter 4.

Consistent with my original aim of developing methods and resources to assist in the computational study of IMPs structure and function, the MemProtMD database and web server (Chapter 2) now provide a useful interactive resource which makes the results of MD simulations of IMPs in explicit lipid bilayers available to a wide audience through a constantly updated web database. In addition, ensemble analyses of the MemProtMD database have revealed general principles of membrane lipid interactions with amino-acid side-chains.

Using these findings, the MemProtMetrics tool described in Chapter 3 uses knowledge of amino-acid side-chain behaviour in lipid bilayers to identify IMPs in the PDB, and orient identified IMPs in a simplified lipid bilayer ready for CGSA simulation using the MemProtMD pipeline. MemProtMetrics' ability to predict orientation of IMPs seen in simulations is promising, although results of OPM [4] remain closest to simulations. It is likely, however, that further optimisations to the MemProtMetrics objective function may be able to improve accuracy substantially.

Described in Chapter 3 is a method to identify membrane lipids from chemical structure, which is then used to assemble what is to date the most complete set of structurally determined IMP-lipid interactions so far identified. Results of annotation of lipid molecules using this dataset are interesting, although represent only a tiny

fraction of insights which may be gained through further analysis of this dataset (see 5.2 *Future prospects* below).

The development of native mass spectrometry of IMPs and membrane lipids is a promising method for examining IMP-lipid interactions *in vitro*, however the effects of the ionised gas phase on these interactions is still poorly understood. The MD study of AmtB and OmpF described in Chapter 4 attempts to re-create conditions in the ionised gas phase by protonation of the MARTINI CG forcefield. Whilst similar trends are seen in these simulations to mass spectrometry experiments, the smaller effect size seen in simulations suggests this approximation leaves considerable scope for improvement.

Using forcefields such as MARTINI 2.2 [119], it is possible to simulate an increasing number of membrane lipids and therefore membrane compositions, bringing greater realism to biomolecular simulations of IMPs and revealing interactions between IMPs and a vast number of lipid species. The Alchex software tool allows simulated lipid compositions to be altered to include new lipid types in various symmetric and asymmetric compositions, automatically and requiring little technical expertise.

Multi-scale simulations of a novel TMEM16K structure in a realistic lipid bilayer give new insights into the mechanism and regulation of lipid scramblase activity, and shows the determined structure to be stable although conformationally highly flexible. The apparent selectivity of TMEM16K scramblase activity for PS is intriguing given the role of TMEM16K mutants in neurological disease linked to PS signalling, and merits further study.

5.2 Future prospects

Completed work described in this thesis suggested several extensions, some under active development by myself and others planned as future projects by others. Many of these feed into larger plans for the future of the MemProtMD pipeline, database and web server, shown in Figure 5.1 and discussed below.

5.2.1 Better use of PDB data

Represented by the leftmost panel of Figure 5.1, the MemProtMD pipeline could be improved by utilising more data from PDB depositions. Protein-lipid interactions seen in structures are not currently considered automatically, and could easily be cross-referenced against simulation results. By combining data from simulations and structures, conserved lipid binding motifs such as the Cholesterol Consensus Motif

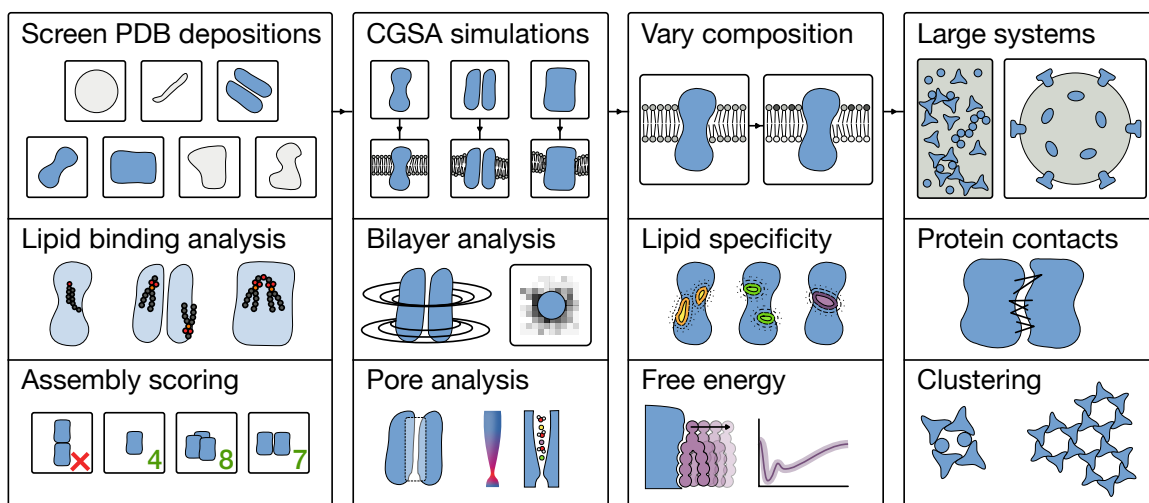


Figure 5.1: The extended MemProtMD pipeline

Several extensions to the MemProtMD pipeline are shown. From left to right - PDB depositions would be screened using MemProtMetrics to identify and orient membrane proteins, with membrane lipids automatically identified and assessed using original electron densities. Multiple biological assemblies would be built and scored using MemProtMetrics to identify likely subunit interfaces in the membrane. CGSA simulation analysis would be extended beyond current bilayer surface analysis to include annotations of pores or gating of channel proteins. Lipid composition could then be varied, and specific protein-lipid interactions identified, as well as CG or atomistic calculations of free energy of interactions. Large systems with multiple protein molecules could be set up using annular lipid positions from the MemProtMD database to view large-scale clustering, or identify likely contacts between different protein molecules or subunits.

(CCM) or Cholesterol Recognition/interaction Amino acid Consensus (CRAC) could be identified, either for all IMPs or individual IMP families.

By considering various biological assemblies suggested by structure authors or automated subunit interface identification using by PISA [263] using MemProtMetrics, anomalous assemblies in PDB depositions could be removed, and, where several valid oligomers or assemblies exist, each valid assembly could be simulated in a lipid bilayer.

5.2.2 Integration of specialised analyses

The second panel from the left in Figure 5.1 depicts analyses presently performed by the MemProtMD analysis application, as well as planned future developments. At present, the MemProtMD analysis application uses the same analysis methods for all IMPs. This is useful for comparisons between IMP families, however many interesting aspects of IMP structure and function specific to IMP subtypes such as receptors, channels or pores are not currently considered. A channel-specific annotation system is currently under development by other researchers which will integrate with the MemProtMD pipeline and web application to automatically characterise qualities of pores in proteins, including pore gating and selectivity. Once complete, visualisations of analysis results will be accessible through the MemProtMD web application.

5.2.3 Extension of MemProtMD to mixed lipid systems

Several sites on IMPs are highly selective for particular lipid types found in their native bilayers. The Alchex software tool, in addition to existing single-lipid species simulations of IMPs in the MemProtMD database makes setup of mixed lipid simulations including a vast array of physiologically relevant lipid types trivial. Preliminary analysis of simulations of membrane proteins using these mixed-lipid bilayers shown in Figure 5.2 suggests annular thickness of pure DPPC bilayers is typically around 0.2 nm greater than those of mixed species counterparts. This may be due to the stabilising effects of non-specific interactions with lipid head-groups closer to the protein's native bilayer.

Analysis of these simulations could reveal high-affinity binding sites for specific lipid types, and examine conservation of these sites across protein families, in a manner similar to the conserved single-species contact analysis performed by MemProtMD at present. Where likely lipid binding sites are found, the energy landscape of the interaction could be calculated using a method such as PLUMED [264], and the site

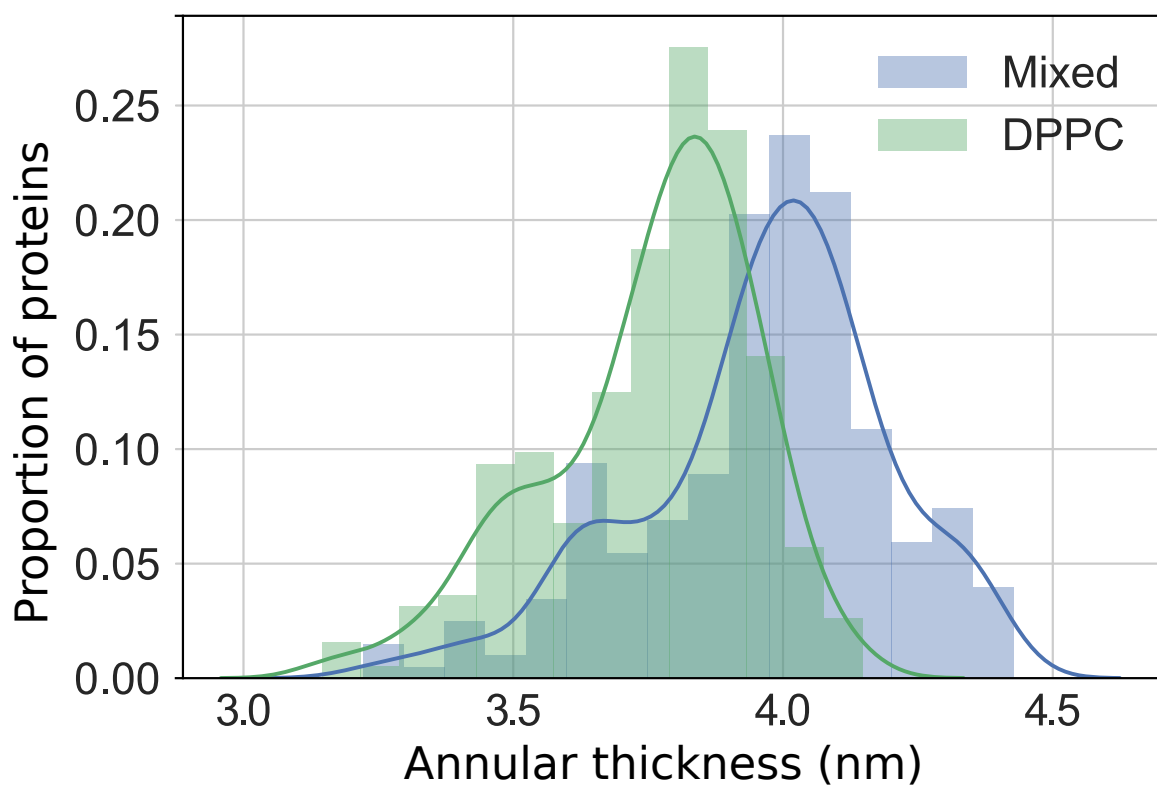


Figure 5.2: Annular thickness of simulated pure DPPC and mixed simulations

Distribution of annular membrane thickness is shown for both pure DPPC and mixed-lipid membranes for a set of 250 proteins for which both comparable simulations are available.

could be cross-referenced against lipid contacts seen in structures of related proteins. This addition to the MemProtMD pipeline is represented by the third panel from the left in Figure 5.1.

5.2.4 Automated setup of large systems

CG simulations of large systems including tens or even thousands of different IMPs are now computationally feasible, and raise the exciting prospect of bridging the gap between length scales seen in single-particle super-resolution microscopy and biomolecular simulation (see Chavent 2016 [63]). Typically, setup of these systems relies on insertion of protein structures into pre-equilibrated bilayers. Further equilibration is then required to allow the bilayer to adjust to newly inserted proteins, allowing lipid annular shells to form and the bilayer to thin or thicken according to localised variation in protein hydrophobic thickness.

The MemProtMD database lends itself to vastly accelerating the setup of such systems, containing simulations consisting of thousands of frames from which proteins may be taken complete with one or more equilibrated lipid annular shells. Positions of the outermost annular shells may then be used to align the protein to a pre-equilibrated bilayer structure, and clashes removed using the Alchembed method [265]. Time taken for equilibration is likely to be considerably reduced since proteins are transplanted with annular shells intact. This process is shown, along with two proof-of-concept systems, in Figure 5.3.

5.2.5 Extension of MemProtMD to lipoproteins and peripheral membrane proteins

The high-throughput approach of MemProtMD currently excludes peripheral or monotopic membrane proteins and lipoproteins, both of which interact frequently with lipid bilayers. Expansion of the MemProtMD pipeline to include peripheral membrane proteins and lipoproteins would reveal much about the interactions of these proteins with biological membranes and lipids.

5.2.6 Improving transparency and repeatability of MemProtMD

At present, the MemProtMD analysis application and web server are frequently backed up, and git is used for source control. Source control is not applied to the MemProtMD database, although regular incremental backups are taken. Whilst data is securely stored and well documented, the approach currently taken leaves much to be

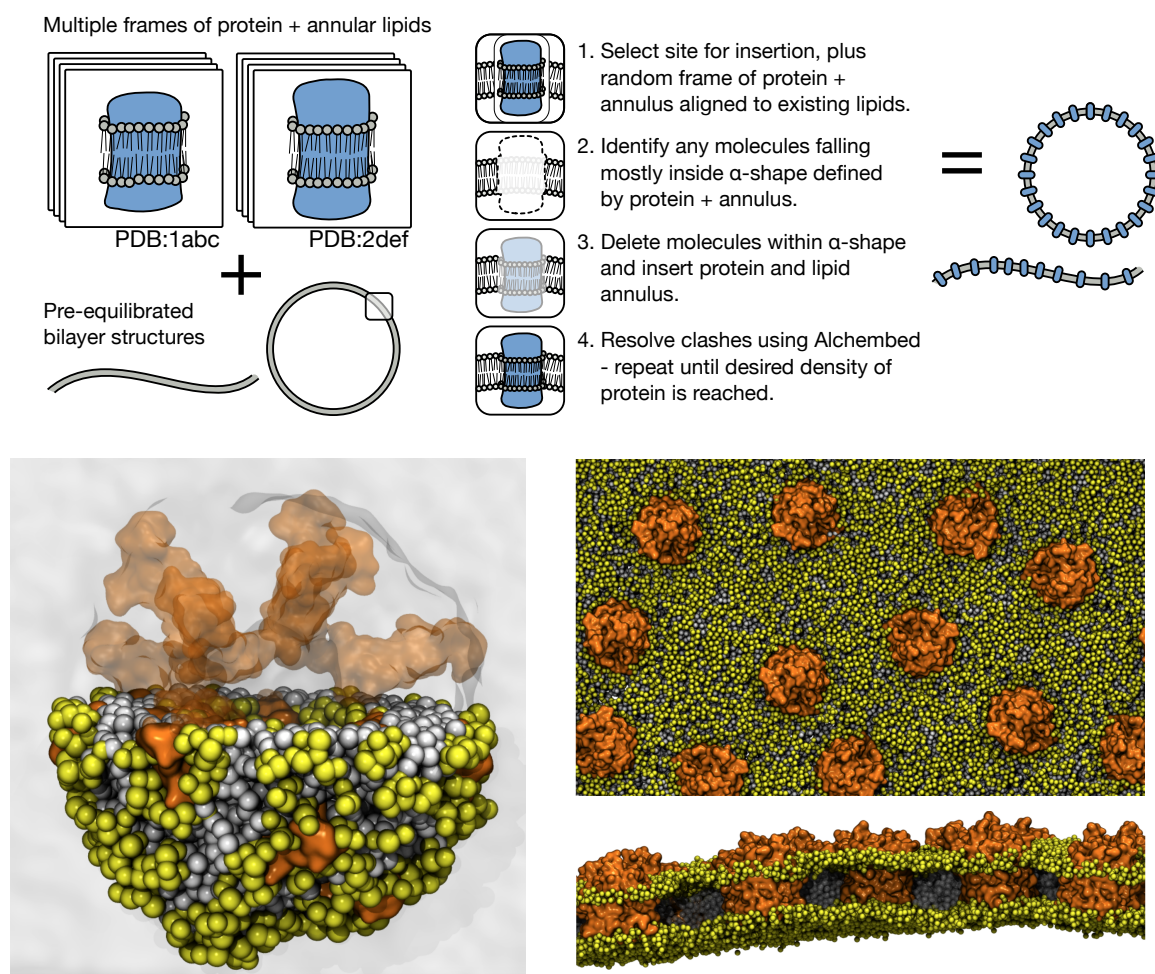


Figure 5.3: Automated setup of large systems using MemProtMD

A method under development for the setup of large membrane/protein systems using the MemProtMD Database is shown. 1 μ s simulations of proteins in explicit lipid bilayers are loaded from the MemProtMD Database, from which multiple frames may be extracted, each with differing (but equilibrated) annular lipids. A bilayer structure is selected from a library of pre-equilibrated planar and spherical lipid bilayers. Sites within the pre-equilibrated bilayer are selected randomly, and the protein is then aligned to the bilayer using positions of its annular lipids. Molecules of the pre-equilibrated structure within a α -shape defined by the protein and annular lipids are removed, and the two systems are combined, using Alchembed to resolve clashes. Shown below are two examples of systems awaiting clash removal with Alchembed: left - small vesicle with Influenza Matrix Protein 2 inserted (cutaway), right - section of a large undulating lipid bilayer with AmtB inserted shown from top-down and side-on view with lipid tails hidden.

desired in terms of transparency and repeatability, with software frequently updated and difficult to install on external hosts without considerable technical expertise. In future, this will be improved by using a Continuous Integration (CI) / Continuous Delivery (CD) approach to development and deployment of the MemProtMD source code and database. This is depicted in Figure 5.4.

Using this model, software updates are grouped into releases, each with clear version numbering. Creation of a release triggers archiving to an OpenAIRE [266] compliant service such as Zenodo (<https://www.zenodo.com>), making specific software releases citeable using a Digital Object Identifier (DOI). At the same time, a container is built, bundling all dependencies and source code into a complete system image which may be deployed on any machine running a recent Linux distribution using containerisation provided by Docker. This would allow frictionless deployment of specific versions of MemProtMD code by external users, either to reproduce MemProtMD data or analyse new structures or models of IMPs.

The MemProtMD database would also be subject to version control and archiving, although the size of simulation trajectories may complicate this, given current size restrictions of free-to-use OpenAIRE implementations. One approach would be to instead store a cryptographic hash of trajectory data which could be used to verify trajectories, or, alternatively, reduce the trajectory to 10-50 representative frames based on clustering of protein and lipid coordinates.

5.2.7 MD simulation of IMPs in the ionised gas phase

Simulations of system ionisation described in Chapter 4 did not completely reproduce effects seen in mass spectrometry experiments. It is therefore desirable to develop an MD methodology capable of accurately depicting the environment of a mass spectrometry experiment. This may include better parametrisation of ionised CG or atomistic forcefields, as well as simulation of more extreme ionisation and lipid micelle or bicelle structures.

5.2.8 Free energy landscape of TMEM16 flip-flops

Scramblase activity of TMEM16 proteins appears to play a key role in PS signalling, and so the specificity of this groove for various lipid head-groups is of considerable interest. One method to quantitatively examine this for a broad range of lipids and TMEM16 proteins would use of a tool such as PLUMED [264] to calculate the free energy landscape of pulling a lipid through the groove, either in CG or atomistic

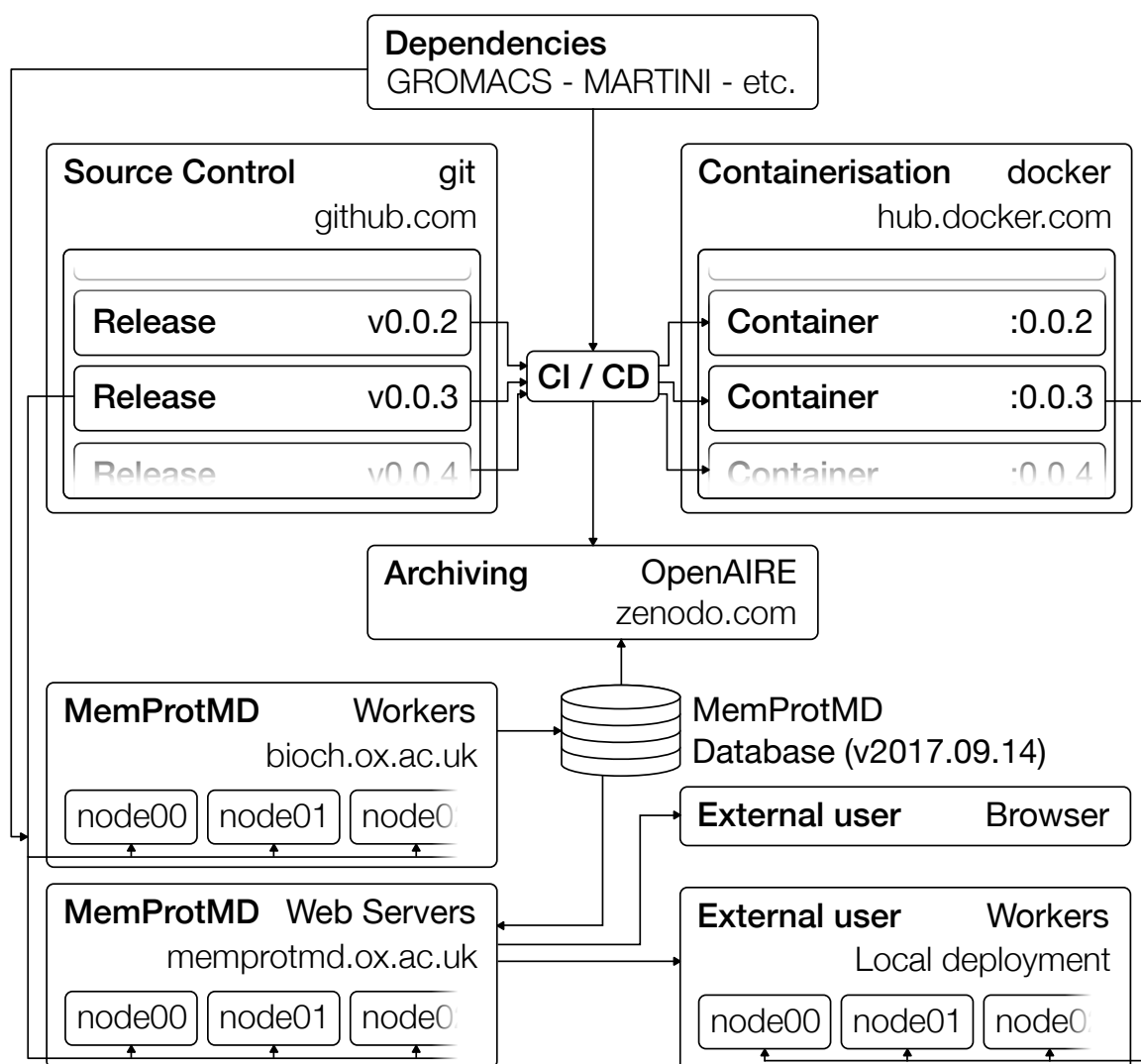


Figure 5.4: Repeatable, transparent research through continuous integration and archiving

An extension to the source control currently used by MemProtMD is shown, using CI, CD and archiving to improve repeatability and transparency of pipeline results. Updates to MemProtMD code through git version control are grouped into numbered releases. Releases trigger a set of CI and CD processes which archive software using an OpenAIRE-compliant host such as zenodo.com and build containers using Docker including all external dependencies, which can be deployed on any host. Updated software is then automatically deployed on in-house and external installations, providing both simulation management workers and web servers. Updates to the MemProtMD database would also be assigned version numbers and pushed to an OpenAIRE compliant archive.

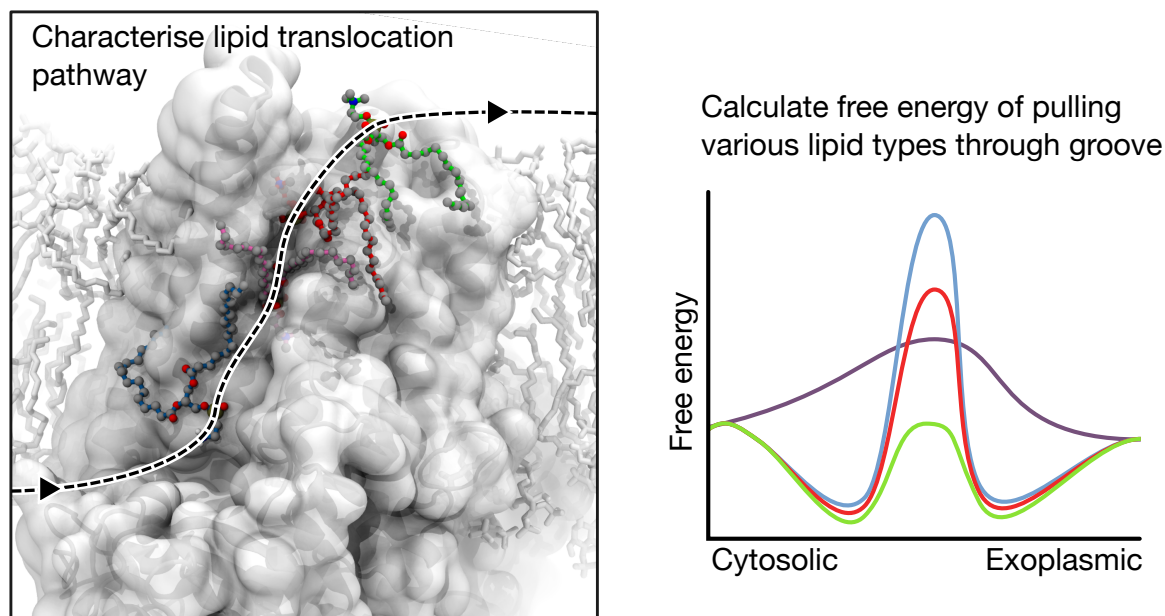


Figure 5.5: Determining the energy landscape of a TMEM16 lipid flip-flop

A method for calculation of the free energy landscape of TMEM16 lipid flip-flops is shown. Atomistic or CG simulations can be used to determine the exact pathway taken by lipid molecules flipped by a TMEM16 scramblase. Using an approach such as Plumed, multiple simulations could be set up to estimate the free energy of lipid headgroups at each point along the pathway, and this approach could be applied to many different lipid head-groups and TMEM16 structures or homology models. Right - the results of such free energy calculations are imagined for four different lipid head-groups following this pathway.

representation. This would first require the pathway followed by a lipid to be characterised, which would be possible using simulations described in Chapter 4.

As more TMEM16 structures become available, specificity of TMEM16 structures and homology models can be assessed for many different lipid types, and amino acid sidechains responsible for groove selectivity could be identified by energy peaks.

5.3 Closing remarks

Since the publication of the first IMP structure in 1985, enormous progress has been made in understanding the structure and function of IMPs, with often transformational results advancing science and technology. Much of this progress has come from the integration of many and varied methods applied to IMP, from structure determination and structure prediction to experimental biophysical assays, bioinformatics studies and biomolecular simulation. Vital to this integration is the transparent ex-

change of data and methods, and careful curation of datasets such as the PDB and Membrane Proteins of Known Structure (mpstruc). Tools and resources described here provide, and will continue to provide, valuable insights into the structure and function of IMPs, and play some small part in driving closer integration between the growing array of techniques currently available for the analysis of IMPs.

References

- [1] Linn Fagerberg, Kalle Jonasson, Gunnar von Heijne, Mathias Uhlén, and Lisa Berglund. Prediction of the human membrane proteome. *Proteomics*, 10(6):1141–1149, 2010.
- [2] John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. How many drug targets are there? *Nature reviews Drug discovery*, 5(12):993–996, 2006.
- [3] J Deisenhofer, O Epp, K Miki, R Huber, and H Michel. Structure of the protein subunits in the photosynthetic reaction centre of rhodospseudomonas viridis at 3Å resolution. *Nature*, 318(6047):618–624, 1985.
- [4] Mikhail A. Lomize, Irina D. Pogozheva, Hyeon Joo, Henry I. Mosberg, and Andrei L. Lomize. OPM database and PPM web server: Resources for positioning of proteins in membranes. *Nucleic Acids Research*, 40(D1):370–376, 2012.
- [5] Phillip J Stansfeld, Joseph E Goose, Martin Caffrey, Elisabeth P Carpenter, Joanne L Parker, Simon Newstead, and Mark SP Sansom. Memprotmd: automated insertion of membrane protein structures into explicit lipid membranes. *Structure*, 23(7):1350–1361, 2015.
- [6] George Hedger and Mark SP Sansom. Lipid interaction sites on channels, transporters and receptors: recent insights from molecular dynamics simulations. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1858(10):2390–2400, 2016.
- [7] Sebastian Kelm, Jiye Shi, and Charlotte M Deane. Medeller: homology-based coordinate generation for membrane proteins. *Bioinformatics*, 26(22):2833–2840, 2010.
- [8] Phillip J Stansfeld, Richard Hopkinson, Frances M Ashcroft, and Mark SP Sansom. Pip2-binding site in kir channels: definition by multiscale biomolecular simulations. *Biochemistry*, 48(46):10926–10933, 2009.

- [9] Jérôme Hénin, Reza Salari, Sruthi Murlidaran, and Grace Brannigan. A predicted binding site for cholesterol on the gaba a receptor. *Biophysical journal*, 106(9):1938–1949, 2014.
- [10] M Remmert, A Biegert, D Linke, AN Lupas, and J Söding. Evolution of outer membrane β -barrels from an ancestral $\beta\beta$ hairpin. *Molecular biology and evolution*, 27(6):1348–1358, 2010.
- [11] Timothy J. Stevens and Isaiah T. Arkin. Substitution rates in α -helical trans-membrane proteins. *Protein Science*, 10(12):2507–2517, 2001.
- [12] Eoin Fahy, Shankar Subramaniam, Robert C Murphy, Masahiro Nishijima, Christian RH Raetz, Takao Shimizu, Friedrich Spener, Gerrit van Meer, Michael JO Wakelam, and Edward A Dennis. Update of the lipid maps comprehensive classification system for lipids. *Journal of lipid research*, 50(Supplement):S9–S14, 2009.
- [13] Markus R Wenk. Lipidomics: new tools and applications. *Cell*, 143(6):888–895, 2010.
- [14] Gerrit Van Meer, Dennis R Voelker, and Gerald W Feigenson. Membrane lipids: where they are and how they behave. *Nature reviews. Molecular cell biology*, 9(2):112, 2008.
- [15] John D Scott and Tony Pawson. Cell signaling in space and time: where proteins come together and when theyre apart. *Science*, 326(5957):1220–1224, 2009.
- [16] LL Hsiao, RJ Howard, M Aikawa, and TF Taraschi. Modification of host cell membrane lipid composition by the intra-erythrocytic human malaria parasite plasmodium falciparum. *Biochemical Journal*, 274(1):121–132, 1991.
- [17] Karine Aupeix, Bénédicte Hugel, Thierry Martin, Pierre Bischoff, Helmut Lill, Jean-Louis Pasquali, and Jean-Marie Freyssinet. The significance of shed membrane particles during programmed cell death in vitro, and in vivo, in hiv-1 infection. *Journal of Clinical Investigation*, 99(7):1546, 1997.
- [18] Kathrin Marheineke, Sylvia Grünewald, William Christie, and Helmut Reiländer. Lipid composition of spodoptera frugiperda (sf9) and trichoplusia ni (tn) insect cells used for baculovirus infection. *FEBS letters*, 441(1):49–52, 1998.

- [19] M Garnier, JR Attali, P Valensi, E Delatour-Hanss, F Gaudey, and D Koutsouris. Erythrocyte deformability in diabetes and erythrocyte membrane lipid composition. *Metabolism*, 39(8):794–798, 1990.
- [20] Salvatore Pepe and Peter L McLennan. Cardiac membrane fatty acid composition modulates myocardial oxygen consumption and postischemic recovery of contractile function. *Circulation*, 105(19):2303–2308, 2002.
- [21] Patricia D Schley, David N Brindley, and Catherine J Field. (n-3) pufa alter raft lipid composition and decrease epidermal growth factor receptor levels in lipid rafts of human breast cancer cells. *The Journal of nutrition*, 137(3):548–553, 2007.
- [22] Pablo V Escribá. Membrane-lipid therapy: a new approach in molecular medicine. *Trends in molecular medicine*, 12(1):34–43, 2006.
- [23] AG Lee. Lipid–protein interactions in biological membranes: a structural perspective. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1612(1):1–40, 2003.
- [24] Mijo Simunovic and Gregory A Voth. Membrane tension controls the assembly of curvature-generating proteins. *Nature communications*, 6, 2015.
- [25] AG Lee. How lipids interact with an intrinsic membrane protein: the case of the calcium pump. *Biochimica et Biophysica Acta (BBA)-Reviews on Biomembranes*, 1376(3):381–390, 1998.
- [26] Heidi Koldsø and Mark SP Sansom. Local lipid reorganization by a transmembrane protein domain. *The journal of physical chemistry letters*, 3(23):3498–3502, 2012.
- [27] JM East, D Melville, and AG Lee. Exchange rates and numbers of annular lipids for the calcium and magnesium ion dependent adenosine triphosphatase. *Biochemistry*, 24(11):2615–2623, 1985.
- [28] Diomedes E Logothetis, Taihao Jin, Dmitry Lupyan, and Avia Rosenhouse-Dantsker. Phosphoinositide-mediated gating of inwardly rectifying k⁺ channels. *Pflügers Archiv-European Journal of Physiology*, 455(1):83–95, 2007.

- [29] Donald W Hilgemann, Siyi Feng, and Cem Nasuhoglu. The complex and intriguing lives of pip2 with ion channels and transporters. *Sci. STKE*, 2001(111):re19–re19, 2001.
- [30] Joanne Oates and Anthony Watts. Uncovering the intimate relationship between lipids, cholesterol and gpcr activation. *Current opinion in structural biology*, 21(6):802–807, 2011.
- [31] Natalia Levina, Sabine Töttemeyer, Neil R Stokes, Petra Louis, Michael A Jones, and Ian R Booth. Protection of escherichia coli cells against extreme turgor by activation of mscs and mscl mechanosensitive channels: identification of genes required for mscs activity. *The EMBO journal*, 18(7):1730–1737, 1999.
- [32] Catherine Berrier, Alexandre Pozza, Agnes de Lacroix de Lavalette, Solenne Chardonnet, Agnes Mesneau, Christine Jaxel, Marc Le Maire, and Alexandre Ghazi. The purified mechanosensitive channel trek-1 is directly sensitive to membrane tension. *Journal of Biological Chemistry*, 288(38):27307–27314, 2013.
- [33] Florian Lesage, Cécile Terrenoire, Georges Romey, and Michel Lazdunski. Human trek2, a 2p domain mechano-sensitive k⁺ channel with multiple regulations by polyunsaturated fatty acids, lysophospholipids, and gs, gi, and gq protein-coupled receptors. *Journal of Biological Chemistry*, 275(37):28398–28405, 2000.
- [34] Carl P Sparrow and CR Raetz. Purification and properties of the membrane-bound cdp-diglyceride synthetase from escherichia coli. *Journal of Biological Chemistry*, 260(22):12084–12091, 1985.
- [35] Louisa Wu, Barbara Niemeyer, Nansi Colley, Michael Socolich, and Charles S Zuker. Regulation of plc-mediated signalling in vivo by cdp-diacylglycerol synthase. *Nature*, 373(6511):216–222, 1995.
- [36] Jérôme Nigou and Gurdyal S Besra. Cytidine diphosphate-diacylglycerol synthesis in mycobacterium smegmatis. *Biochemical Journal*, 367(1):157–162, 2002.
- [37] Bruno Antonsson. Phosphatidylinositol synthase from mammalian tissues. *Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism*, 1348(1):179–186, 1997.

- [38] Isabel Mérida, Antonia Ávila-Flores, and Ernesto Merino. Diacylglycerol kinases: at the hub of cell signalling. *Biochemical Journal*, 409(1):1–18, 2008.
- [39] Jos AF Op den Kamp. Lipid asymmetry in membranes. *Annual review of biochemistry*, 48(1):47–71, 1979.
- [40] Karthik M Kodigepalli, Kiah Bowers, Arielle Sharp, and Meera Nanjundan. Roles and regulation of phospholipid scramblases. *FEBS letters*, 589(1):3–14, 2015.
- [41] Jun Suzuki, Toshihiro Fujii, Takeshi Imao, Kenji Ishihara, Hiroshi Kuba, and Shigekazu Nagata. Calcium-dependent phospholipid scramblase activity of tmem16 protein family members. *Journal of Biological Chemistry*, 288(19):13305–13316, 2013.
- [42] Alex J Laude and Ian A Prior. Plasma membrane microdomains: organization, function and trafficking. *Molecular membrane biology*, 21(3):193–205, 2004.
- [43] Ken Ritchie, Ryota Iino, Takahiro Fujiwara, Kotonno Murase, and Akihiro Kusumi. The fence and picket structure of the plasma membrane of live cells as revealed by single molecule techniques. *Molecular membrane biology*, 20(1):13–18, 2003.
- [44] Ubaldo E Martinez-Outschoorn, Federica Sotgia, and Michael P Lisanti. Caveolae and signalling in cancer. *Nature Reviews. Cancer*, 15(4):225, 2015.
- [45] Thomas Harder and Kai Simons. Caveolae, digs, and the dynamics of sphingolipidcholesterol microdomains. *Current opinion in cell biology*, 9(4):534–542, 1997.
- [46] Chen Song, Holm Holmsen, and Willy Nerdal. Existence of lipid microdomains in bilayer of dipalmitoyl phosphatidylcholine (dppc) and 1-stearoyl-2-docosahexenoyl phosphatidylserine (sdps) and their perturbation by chlorpromazine: a 13 c and 31 p solid-state nmr study. *Biophysical chemistry*, 120(3):178–187, 2006.
- [47] Teun Munnik. Phosphatidic acid: an emerging plant lipid second messenger. *Trends in plant science*, 6(5):227–233, 2001.

- [48] Drazen Raucher, Thomas Stauffer, Wen Chen, Kang Shen, Shuling Guo, John D York, Michael P Sheetz, and Tobias Meyer. Phosphatidylinositol 4, 5-bisphosphate functions as a second messenger that regulates cytoskeleton-plasma membrane adhesion. *Cell*, 100(2):221–228, 2000.
- [49] Jason G Kay and Sergio Grinstein. Phosphatidylserine-mediated cellular signaling. In *Lipid-mediated Protein Signaling*, pages 177–193. Springer, 2013.
- [50] Antoine-Emmanuel Saliba, Ivana Vonkova, and Anne-Claude Gavin. The systematic analysis of protein-lipid interactions comes of age. *Nature reviews. Molecular cell biology*, 16(12):753, 2015.
- [51] Martin Caffrey. A comprehensive review of the lipid cubic phase or in meso method for crystallizing membrane and soluble proteins and complexes. *Acta Crystallographica Section F: Structural Biology Communications*, 71(1):3–18, 2015.
- [52] Philip L Yeagle. Non-covalent binding of membrane lipids to membrane proteins. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1838(6):1548–1559, 2014.
- [53] Kutti R Vinothkumar. Membrane protein structures without crystals, by single particle electron cryomicroscopy. *Current opinion in structural biology*, 33:103–114, 2015.
- [54] Jean-Louis Rigaud, Gervaise Mosser, Jean-Jacques Lacapere, Anders Olofsson, Daniel Levy, and Jean-Luc Ranck. Bio-beads: an efficient strategy for two-dimensional crystallization of membrane proteins. *Journal of structural biology*, 118(3):226–235, 1997.
- [55] Tamir Gonen, Yifan Cheng, Piotr Sliz, Yoko Hiroaki, Yoshinori Fujiyoshi, Stephen C Harrison, and Thomas Walz. Lipid-protein interactions in double-layered two-dimensional aqp0 crystals. *Nature*, 438(7068):633, 2005.
- [56] Xiao Tao, Richard K Hite, and Roderick MacKinnon. Cryo-em structure of the open high conductance Ca^{2+} -activated K^{+} channel. *Nature*, 541(7635):46, 2017.
- [57] Charles R Sanders and Frank Sönnichsen. Solution nmr of membrane proteins: practice and challenges. *Magnetic Resonance in Chemistry*, 44(S1), 2006.

- [58] Daniel Huster. Solid-state nmr spectroscopy to study protein–lipid interactions. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1841(8):1146–1160, 2014.
- [59] Chih-Yun Hsia, Mark J Richards, and Susan Daniel. A review of traditional and emerging methods to characterize lipid–protein interactions in biological membranes. *Analytical Methods*, 7(17):7076–7094, 2015.
- [60] Nelson P Barrera, Natalie Di Bartolo, Paula J Booth, and Carol V Robinson. Micelles protect membrane complexes from solution to vacuum. *Science*, 321(5886):243–246, 2008.
- [61] Kallol Gupta, Joseph A. C. Donlan, Jonathan T. S. Hopper, Povilas Uzdevinys, Michael Landreh, Weston B. Struwe, David Drew, Andrew J. Baldwin, Phillip J. Stansfeld, and Carol V. Robinson. The role of interfacial lipids in stabilizing membrane protein oligomers. *Nature*, 541(7637):1–4, 2017.
- [62] Arthur Laganowsky, Eamonn Reading, Timothy M Allison, Martin B Ulmschneider, Matteo T Degiacomi, Andrew J Baldwin, and Carol V Robinson. Membrane proteins bind lipids selectively to modulate their structure and function. *Nature*, 510(7503):172, 2014.
- [63] Matthieu Chavent, Anna L Duncan, and Mark SP Sansom. Molecular dynamics simulations of membrane proteins and their interactions: from nanoscale to mesoscale. *Current opinion in structural biology*, 40:8–16, 2016.
- [64] S Doerr, MJ Harvey, Frank Noe, and G De Fabritiis. Htmd: High-throughput molecular dynamics for molecular discovery. *Journal of chemical theory and computation*, 12(4):1845–1852, 2016.
- [65] Sunhwan Jo, Joseph B Lim, Jeffery B Klauda, and Wonpil Im. Charmm-gui membrane builder for mixed bilayers and its application to yeast membranes. *Biophysical journal*, 97(1):50–58, 2009.
- [66] George Hedger, Mark SP Sansom, and Heidi Koldsø. The juxtamembrane regions of human receptor tyrosine kinases exhibit conserved interaction sites with anionic lipids. *Scientific reports*, 5, 2015.
- [67] David Filmore. It’s a gpcr world. *Modern drug discovery*, 7:24–28, 2004.

- [68] Michael Zocher, Cheng Zhang, Søren GF Rasmussen, Brian K Kobilka, and Daniel J Müller. Cholesterol increases kinetic, energetic, and mechanical stability of the human β 2-adrenergic receptor. *Proceedings of the National Academy of Sciences*, 109(50):E3463–E3472, 2012.
- [69] Sourav Ganguly, Andrew HA Clayton, and Amitabha Chattopadhyay. Organization of higher-order oligomers of the serotonin 1a receptor explored utilizing homo-fret in live cells. *Biophysical Journal*, 100(2):361–368, 2011.
- [70] Kaleeckal G Harikumar, Vishwajeet Puri, Raman Deep Singh, Kentaro Hanada, Richard E Pagano, and Laurence J Miller. Differential effects of modification of membrane cholesterol and sphingolipids on the conformation, function, and trafficking of the g protein-coupled cholecystokinin receptor. *Journal of Biological Chemistry*, 280(3):2176–2185, 2005.
- [71] Carlos J Baier, Jacques Fantini, and Francisco J Barrantes. Disclosure of cholesterol recognition motifs in transmembrane domains of the human nicotinic acetylcholine receptor. *Scientific reports*, 1:69, 2011.
- [72] Michael A Hanson, Vadim Cherezov, Mark T Griffith, Christopher B Roth, Veli-Pekka Jaakola, Ellen YT Chien, Jeffrey Velasquez, Peter Kuhn, and Raymond C Stevens. A specific cholesterol binding site is established by the 2.8 Å structure of the human β 2-adrenergic receptor. *Structure*, 16(6):897–905, 2008.
- [73] Ji Young Lee and Edward Lyman. Predictions for cholesterol interaction sites on the a2a adenosine receptor. *Journal of the American Chemical Society*, 134(40):16512–16515, 2012.
- [74] Wei Liu, Eugene Chun, Aaron A Thompson, Pavel Chubukov, Fei Xu, Vsevolod Katritch, Gye Won Han, Christopher B Roth, Laura H Heitman, Adriaan P IJzerman, et al. Structural basis for allosteric regulation of gpcrs by sodium ions. *Science*, 337(6091):232–236, 2012.
- [75] Durba Sengupta and Amitabha Chattopadhyay. Identification of cholesterol binding sites in the serotonin1a receptor. *The journal of physical chemistry B*, 116(43):12991–12996, 2012.

- [76] Jufang Shan, George Khelashvili, Sayan Mondal, Ernest L Mehler, and Harel Weinstein. Ligand-dependent conformations and dynamics of the serotonin 5-HT_{2A} receptor determine its activation and membrane-driven oligomerization properties. *PLoS computational biology*, 8(4):e1002473, 2012.
- [77] Stefan Gahbauer and Rainer A Böckmann. Membrane-mediated oligomerization of G protein coupled receptors and its implications for GPCR function. *Frontiers in physiology*, 7, 2016.
- [78] Sayan Mondal, Jennifer M Johnston, Hao Wang, George Khelashvili, Marta Filizola, and Harel Weinstein. Membrane driven spatial organization of GPCRs. *Scientific reports*, 3, 2013.
- [79] Xiaohui Cang, Linlin Yang, Jing Yang, Cheng Luo, Mingyue Zheng, Kunqian Yu, Huaiyu Yang, and Hualiang Jiang. Cholesterol- β 1AR interaction versus cholesterol- β 2AR interaction. *Proteins: Structure, Function, and Bioinformatics*, 82(5):760–770, 2014.
- [80] Rosie Dawaliby, Cataldo Trubbia, Cédric Delporte, Matthieu Masureel, Pierre Van Antwerpen, Brian K Kobilka, and Cédric Govaerts. Allosteric regulation of GPCR activity by phospholipids. *Nature chemical biology*, 12(1):35, 2016.
- [81] Chris Neale, Henry D Herce, Régis Pomès, and Angel E García. Can specific protein-lipid interactions stabilize an active state of the beta 2 adrenergic receptor? *Biophysical journal*, 109(8):1652–1662, 2015.
- [82] Jae Hyun Bae and Joseph Schlessinger. Asymmetric tyrosine kinase arrangements in activation or autophosphorylation of receptor tyrosine kinases. *Molecules and cells*, 29(5):443–448, 2010.
- [83] Ünal Coskun, Michał Grzybek, David Drechsel, and Kai Simons. Regulation of human EGFR by lipids. *Proceedings of the National Academy of Sciences*, 108(22):9044–9048, 2011.
- [84] Ioannis E Michailidis, Radda Rusinova, Anastasios Georgakopoulos, Yibang Chen, Ravi Iyengar, Nikolaos K Robakis, Diomedes E Logothetis, and Lia Baki. Phosphatidylinositol-4, 5-bisphosphate regulates epidermal growth factor receptor activation. *Pflügers Archiv-European Journal of Physiology*, 461(3):387–397, 2011.

- [85] George Hedger, David Shorthouse, Heidi Koldsø, and Mark SP Sansom. Free energy landscape of lipid interactions with regulatory binding sites on the trans-membrane domain of the egf receptor. *The Journal of Physical Chemistry B*, 120(33):8154–8163, 2016.
- [86] T Sooksawate and MA Simmonds. Effects of membrane cholesterol on the sensitivity of the gaba a receptor to gaba in acutely dissociated rat hippocampal neurones. *Neuropharmacology*, 40(2):178–184, 2001.
- [87] John E Baenziger, Jaimee A Domville, and JP Daniel Therien. The role of cholesterol in the activation of nicotinic acetylcholine receptors. *Current topics in membranes*, 80:95, 2017.
- [88] Grace Brannigan, Jérôme Hénin, Richard Law, Roderic Eckenhoff, and Michael L Klein. Embedded cholesterol in the nicotinic acetylcholine receptor. *Proceedings of the National Academy of Sciences*, 105(38):14418–14423, 2008.
- [89] Ryan E Hibbs and Eric Gouaux. Principles of activation and permeation in an anion-selective cys-loop receptor. *Nature*, 474(7349):54, 2011.
- [90] Coeli MB Lopes, Hailin Zhang, Tibor Rohacs, Taihao Jin, Jian Yang, and Diomedes E Logothetis. Alterations in conserved kir channel-pip 2 interactions underlie channelopathies. *Neuron*, 34(6):933–944, 2002.
- [91] Scott B Hansen, Xiao Tao, and Roderick MacKinnon. Structural basis of pip2 activation of the classical inward rectifier k+ channel kir2. 2. *Nature*, 477(7365):495, 2011.
- [92] Christos Pliotas, A Caroline E Dahl, Tim Rasmussen, Kozhinjampara R Mahendran, Terry K Smith, Phedra Marius, Joseph Gault, Thandiwe Banda, Akiko Rasmussen, Samantha Miller, et al. The role of lipids in mechanosensation. *Nature structural & molecular biology*, 22(12):991, 2015.
- [93] Prafulla Aryal, Viwan Jarerattanachai, Michael V Clausen, Marcus Schewe, Conor McClenaghan, Liam Argent, Linus J Conrad, Yin Y Dong, Ashley CW Pike, Elisabeth P Carpenter, et al. Bilayer-mediated structural transitions control mechanosensitivity of the trek-2 k2p channel. *Structure*, 25(5):708–718, 2017.

- [94] Cristina Paulino, Yvonne Neldner, Andy KM Lam, Valeria Kalienkova, Janine Denise Brunner, Stephan Schenck, and Raimund Dutzler. Structural basis for anion conduction in the calcium-activated chloride channel tmem16a. *eLife*, 6, 2017.
- [95] Toshihiro Fujii, Asuka Sakata, Satoshi Nishimura, Koji Eto, and Shigekazu Nagata. Tmem16f is required for phosphatidylserine exposure and microparticle release in activated mouse platelets. *Proceedings of the National Academy of Sciences*, 112(41):12800–12805, 2015.
- [96] Janine D Brunner, Novandy K Lim, and Stephan Schenck. X-ray structure of a calcium activated tmem16 lipid scramblase. *Biophysical Journal*, 108(2):343a, 2015.
- [97] Neville P Bethel and Michael Grabe. Atomistic insight into lipid translocation by a tmem16 scramblase. *Proceedings of the National Academy of Sciences*, page 201607574, 2016.
- [98] Jarred M Whitlock and H Criss Hartzell. A pore idea: the ion conduction pathway of tmem16/ano proteins is composed partly of lipid. *Pflügers Archiv-European Journal of Physiology*, 468(3):455–473, 2016.
- [99] Tomoki Naito, Hiroyuki Takatsu, Rie Miyano, Naoto Takada, Kazuhisa Nakayama, and Hye-Won Shin. Phospholipid flippase atp10a translocates phosphatidylcholine and is involved in plasma membrane dynamics. *Journal of Biological Chemistry*, 290(24):15004–15017, 2015.
- [100] Jonathan A Coleman, Michael CM Kwok, and Robert S Molday. Localization, purification, and functional reconstitution of the p4-atpase atp8a2, a phosphatidylserine flippase in photoreceptor disc membranes. *Journal of Biological Chemistry*, 284(47):32670–32679, 2009.
- [101] Xianjun Zhu, Richard T Libby, Wilhelmine N De Vries, Richard S Smith, Dana L Wright, Roderick T Bronson, Kevin L Seburn, and Simon WM John. Mutations in a p-type atpase gene cause axonal degeneration. *PLoS genetics*, 8(8):e1002853, 2012.
- [102] Jens P Andersen, Anna L Vestergaard, Stine A Mikkelsen, Louise S Mogensen, Madhavan Chalat, and Robert S Molday. P4-atpases as phospholipid flippases: structure, function, and enigmas. *Frontiers in physiology*, 7, 2016.

- [103] Bartholomew P Roland and Todd R Graham. Directed evolution of a sphingomyelin flippase reveals mechanism of substrate backbone discrimination by a p4-atpase. *Proceedings of the National Academy of Sciences*, page 201525730, 2016.
- [104] Aaron K. Chamberlain, Yohan Lee, Sanguk Kim, and James U. Bowie. Snorkeling preferences foster an amino acid composition bias in transmembrane helices. *Journal of Molecular Biology*, 339(2):471 – 479, 2004.
- [105] J. Antoinette Killian and Gunnar Von Heijne. How proteins adapt to a membrane-water interface. *Trends in Biochemical Sciences*, 25(9):429–434, 2000.
- [106] Jie Liang, Larisa Adamian, and Ronald Jackups. The membranewater interface region of membrane proteins: structural bias and the anti-snorkeling effect. *Trends in Biochemical Sciences*, 30(7):355 – 357, 2005.
- [107] M B Ulmschneider and Mark S. P. Sansom. Amino acid distributions in integral membrane protein structures. *Biochim Biophys Acta*, 1512(1):1–14, 2001.
- [108] Carlos Baeza-Delgado, Marc A. Marti-Renom, and Ismael Mingarro. Structure-based statistical analysis of transmembrane helices. *European Biophysics Journal*, 42(2-3):199–207, 2013.
- [109] Carolina Landolt-Marticorena, Karen a Williams, Charles M Deber, and Reinhard A.F. Reithmeier. Non-random distribution of amino acids in the transmembrane segments of human type I single span membrane proteins. *Journal of molecular biology*, 229(3):602–8, feb 1993.
- [110] Irina D. Pogozheva, Stephanie Tristram-Nagle, Henry I. Mosberg, and Andrei L. Lomize. Structural adaptations of proteins to different biological membranes. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1828(11):2592–2608, 2013.
- [111] Gunnar von HEIJNE and Ylva GAVEL. Topogenic signals in integral membrane proteins. *European Journal of Biochemistry*, 174(4):671–678, 1988.
- [112] Johan Nilsson, Bengt Persson, and Gunnar von Heijne. Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes. *Proteins*, 60(4):606–616, 2005.

- [113] JS Richardson and DC Richardson. Amino acid preferences for specific locations at the ends of alpha helices. *Science*, 240(4859):1648–1652, 1988.
- [114] Jack L.C.M. Van De Vossenberg, Sonja-Verena Albers, Chris Van Der Does, Arnold J. M. Driessen, and Wim Van Klompenburg. The positive inside rule is not determined by the polarity of the . *Molecular Microbiology*, 29(4):1125–1126, 1998.
- [115] Veit Goder, Tina Junne, and Martin Spiess. Sec61p contributes to signal sequence orientation according to the positive-inside rule. *Molecular biology of the cell*, 15(3):1470–1478, 2004.
- [116] Hao Dong, Mukesh Sharma, Huan Xiang Zhou, and Timothy A. Cross. Glycines: Role in α -helical membrane protein structures and a potential indicator of native conformation. *Biochemistry*, 51(24):4779–4789, 2012.
- [117] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, 2005.
- [118] Siewert J Marrink, H Jelger Risselada, Serge Yefimov, D Peter Tieleman, and Alex H De Vries. The martini force field: coarse grained model for biomolecular simulations. *The journal of physical chemistry B*, 111(27):7812–7824, 2007.
- [119] Djurre H. de Jong, Gurpreet Singh, W. F. Drew Bennett, Clement Arnarez, Tsjerk A. Wassenaar, Lars V. Schfer, Xavier Periole, D. Peter Tieleman, and Siewert J. Marrink. Improved parameters for the martini coarse-grained protein force field. *Journal of Chemical Theory and Computation*, 9(1):687–697, 2013. PMID: 26589065.
- [120] Luca Monticelli, Senthil K Kandasamy, Xavier Periole, Ronald G Larson, D Peter Tieleman, and Siewert-Jan Marrink. The martini coarse-grained force field: extension to proteins. *Journal of chemical theory and computation*, 4(5):819–834, 2008.
- [121] Clément Arnarez, Jaakko J Uusitalo, Marcelo F Masman, Helgi I Ingólfsson, Djurre H de Jong, Manuel N Melo, Xavier Periole, Alex H de Vries, and Siewert J Marrink. Dry martini, a coarse-grained force field for lipid membrane simulations with implicit solvent. *J. Chem. Theory Comput*, 11(1):260–275, 2015.

- [122] Phillip J Stansfeld and Mark SP Sansom. From coarse grained to atomistic: a serial multiscale approach to membrane protein simulations. *Journal of Chemical Theory and Computation*, 7(4):1157–1166, 2011.
- [123] Tsjerk A Wassenaar, Kristyna Pluhackova, Rainer A Bockmann, Siewert J Marrink, and D Peter Tieleman. Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. *Journal of chemical theory and computation*, 10(2):676–690, 2014.
- [124] Julija Zavadlav, Manuel N Melo, Ana V Cunha, Alex H De Vries, Siewert J Marrink, and Matej Praprotnik. Adaptive resolution simulation of martini solvents. *Journal of chemical theory and computation*, 10(6):2591–2598, 2014.
- [125] Chris Oostenbrink, Alessandra Villa, Alan E Mark, and Wilfred 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*, 25(13):1656–1676, 2004.
- [126] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.
- [127] Naama Hurwitz, Dina Schneidman-Duhovny, and Haim J Wolfson. Memdock: an α -helical membrane protein docking algorithm. *Bioinformatics*, 32(16):2444–2450, 2016.
- [128] Yana Gofman, Turkan Haliloglu, and Nir Ben-Tal. Monte carlo simulations of peptide–membrane interactions with the mcpep web server. *Nucleic acids research*, 40(W1):W358–W363, 2012.
- [129] James A. Cuff, Michele E. Clamp, Asim S. Siddiqui, Matt Finlay, and Geoffrey J. Barton. Jpred: a consensus secondary structure prediction server. *Bioinformatics (Oxford, England)*, 14(10):892–893, 1998.
- [130] David T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195–202, 1999.
- [131] Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

- [132] Gunnar Von Heijne. Membrane protein structure prediction: hydrophobicity analysis and the positive-inside rule. *Journal of molecular biology*, 225(2):487–494, 1992.
- [133] Gunnar von Heijne. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *The EMBO journal*, 5(11):3021, 1986.
- [134] DT Jones, WR Taylor, and JM Thornton. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 33(10):3038–3049, 1994.
- [135] Burkhard Rost, Piero Fariselli, and Rita Casadio. Topology prediction for helical transmembrane proteins at 86% accuracy—topology prediction at 86% accuracy. *Protein Science*, 5(8):1704–1718, 1996.
- [136] Erik LL Sonnhammer, Gunnar Von Heijne, Anders Krogh, et al. A hidden markov model for predicting transmembrane helices in protein sequences. In *Ismb*, volume 6, pages 175–182, 1998.
- [137] Gabor E Tusnady and Istvan Simon. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *Journal of molecular biology*, 283(2):489–506, 1998.
- [138] Håkan Viklund and Arne Elofsson. Best α -helical transmembrane protein topology predictions are achieved using hidden markov models and evolutionary information. *Protein Science*, 13(7):1908–1917, 2004.
- [139] David T Jones. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, 23(5):538–544, 2007.
- [140] Håkan Viklund, Andreas Bernsel, Marcin Skwark, and Arne Elofsson. Spoctopus: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics*, 24(24):2928–2929, 2008.
- [141] Timothy Nugent and David T Jones. Transmembrane protein topology prediction using support vector machines. *BMC bioinformatics*, 10(1):159, 2009.
- [142] László Dobson, István Reményi, and Gábor E Tusnady. Cctop: a consensus constrained topology prediction web server. *Nucleic acids research*, 43(W1):W408–W412, 2015.

- [143] Andreas Bernsel, Håkan Viklund, Aron Hennerdal, and Arne Elofsson. Topcons: consensus prediction of membrane protein topology. *Nucleic acids research*, 37(suppl_2):W465–W468, 2009.
- [144] Michael Remmert, Dirk Linke, Andrei N Lupas, and Johannes Söding. Hhomp-prediction and classification of outer membrane proteins. *Nucleic acids research*, 37(suppl_2):W446–W451, 2009.
- [145] Nitesh Kumar Singh, Aaron Goodman, Peter Walter, Volkhard Helms, and Sikander Hayat. Tmbhmm: a frequency profile based hmm for predicting the topology of transmembrane beta barrel proteins and the exposure status of transmembrane residues. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1814(5):664–670, 2011.
- [146] Sikander Hayat and Arne Elofsson. Boctopus: improved topology prediction of transmembrane β barrel proteins. *Bioinformatics*, 28(4):516–522, 2012.
- [147] Arlo Randall, Jianlin Cheng, Michael Sweredoski, and Pierre Baldi. Tmbpro: secondary structure, β -contact and tertiary structure prediction of transmembrane β -barrel proteins. *Bioinformatics*, 24(4):513–520, 2007.
- [148] Christoph Peters and Arne Elofsson. Why is the biological hydrophobicity scale more accurate than earlier experimental hydrophobicity scales? *Proteins: Structure, Function, and Bioinformatics*, 82(9):2190–2198, 2014.
- [149] Gábor E Tusnády, Zsuzsanna Dosztányi, and István Simon. Transmembrane proteins in the protein data bank: identification and classification. *Bioinformatics*, 20(17):2964–2972, 2004.
- [150] Timothy Nugent and David T Jones. Membrane protein orientation and refinement using a knowledge-based statistical potential. *BMC bioinformatics*, 14(1):276, jan 2013.
- [151] Torsten Schwede, Jurgen Kopp, Nicolas Guex, and Manuel C Peitsch. Swiss-model: an automated protein homology-modeling server. *Nucleic acids research*, 31(13):3381–3385, 2003.
- [152] Johannes Söding, Andreas Biegert, and Andrei N Lupas. The hhpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, 33(suppl_2):W244–W248, 2005.

- [153] Lawrence A Kelley, Stefans Mezulis, Christopher M Yates, Mark N Wass, and Michael JE Sternberg. The phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, 10(6):845–858, 2015.
- [154] Ch Surendhar Reddy, K Vijayasathathy, E Srinivas, G Madhavi Sastry, and G Narahari Sastry. Homology modeling of membrane proteins: a critical assessment. *Computational biology and chemistry*, 30(2):120–126, 2006.
- [155] DT Jones, WR Taylor, and JM Thornton. A mutation data matrix for transmembrane proteins. *FEBS letters*, 339(3):269–275, 1994.
- [156] Pauline C Ng, Jorja G Henikoff, and Steven Henikoff. Phat: a transmembrane-specific substitution matrix. *Bioinformatics*, 16(9):760–766, 2000.
- [157] Walter Pirovano, K Anton Feenstra, and Jaap Heringa. Praline: a strategy for improved multiple alignment of transmembrane proteins. *Bioinformatics*, 24(4):492–497, 2008.
- [158] Jamie R Hill, Sebastian Kelm, Jiye Shi, and Charlotte M Deane. Environment specific substitution tables improve membrane protein alignment. *Bioinformatics*, 27(13):i15–i23, 2011.
- [159] Jamie R Hill and Charlotte M Deane. Mp-t: improving membrane protein alignment for structure prediction. *Bioinformatics*, 29(1):54–61, 2012.
- [160] Sebastian Kelm, Jiye Shi, and Charlotte M Deane. imembrane: homology-based membrane-insertion of proteins. *Bioinformatics*, 25(8):1086–1088, 2009.
- [161] Jia-Ming Chang, Paolo Di Tommaso, Jean-François Taly, and Cedric Notredame. Accurate multiple sequence alignment of transmembrane proteins with psi-coffee. *BMC bioinformatics*, 13(4):S1, 2012.
- [162] Marcus Stamm, René Staritzbichler, Kamil Khafizov, and Lucy R Forrest. Alignmea membrane protein sequence alignment web server. *Nucleic acids research*, 42(W1):W246–W251, 2014.
- [163] Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.
- [164] Jean-Paul Ebejer, Jamie R Hill, Sebastian Kelm, Jiye Shi, and Charlotte M Deane. Memoir: template-based structure prediction for membrane proteins. *Nucleic acids research*, 41(W1):W379–W383, 2013.

- [165] Yoonjoo Choi and Charlotte M Deane. Fread revisited: accurate loop structure prediction using a database search algorithm. *Proteins: Structure, Function, and Bioinformatics*, 78(6):1431–1440, 2010.
- [166] Nicholas Hamilton and Thomas Huber. An introduction to protein contact prediction. *Bioinformatics: Structure, Function and Applications*, pages 87–104, 2008.
- [167] David T Jones, Tanya Singh, Tomasz Kosciolk, and Stuart Tetchner. Metapsicov: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7):999–1006, 2014.
- [168] Tomasz Kosciolk and David T Jones. Accurate contact predictions using co-variation techniques and machine learning. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):145–151, 2016.
- [169] Bohdan Monastyrskyy, Daniel D’Andrea, Krzysztof Fidelis, Anna Tramontano, and Andriy Kryshtafovych. New encouraging developments in contact prediction: Assessment of the casp11 results. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):131–144, 2016.
- [170] Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O’Donovan, Nicole Redaschi, and Lai Su L Yeh. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(DATABASE ISS.), 2005.
- [171] Protein Data Bank. Protein data bank. *Nature New Biol*, 233:223, 1971.
- [172] Sameer Velankar, Christoph Best, B Beuth, CH Boutselakis, N Cobley, AW Sousa Da Silva, Dimitris Dimitropoulos, Adel Golovin, Miriam Hirshberg, Melford John, et al. Pdbe: protein data bank in europe. *Nucleic acids research*, 38(suppl_1):D308–D317, 2009.
- [173] Akira R Kinjo, Hirofumi Suzuki, Reiko Yamashita, Yasuyo Ikegawa, Takahiro Kudou, Reiko Igarashi, Yumiko Kengaku, Hasumi Cho, Daron M Standley, Atsushi Nakagawa, et al. Protein data bank japan (pd bj): maintaining a structural data archive and resource description framework format. *Nucleic acids research*, 40(D1):D453–D460, 2012.

- [174] Peter W. Rose, Andreas Prli, Ali Altunkaya, Chunxiao Bi, Anthony R. Bradley, Cole H. Christie, Luigi Di Costanzo, Jose M. Duarte, Shuchismita Dutta, Zukang Feng, Rachel Kramer Green, David S. Goodsell, Brian Hudson, Tara Kalro, Robert Lowe, Ezra Peisach, Christopher Randle, Alexander S. Rose, Chenghua Shao, Yi-Ping Tao, Yana Valasatava, Maria Voigt, John D. Westbrook, Jesse Woo, Huangwang Yang, Jasmine Y. Young, Christine Zardecki, Helen M. Berman, and Stephen K. Burley. The rcsb protein data bank: integrative view of protein, gene and 3d structural information. *Nucleic Acids Research*, 45(D1):D271, 2017.
- [175] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.
- [176] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl_1):D138–D141, 2004.
- [177] Stephen H. White. mpstruc - Membrane Proteins of Known Structure. <http://blanco.biomol.uci.edu/mpstruc/>. Accessed: 2017-02-24.
- [178] Milton H Saier Jr, Can V Tran, and Ravi D Barabote. Tcdb: the transporter classification database for membrane transport protein analyses and information. *Nucleic acids research*, 34(suppl_1):D181–D186, 2006.
- [179] Florence Horn, Emmanuel Bettler, Laerte Oliveira, Fabien Campagne, Fred E Cohen, and Gerrit Vriend. Gpcrdb information system for g protein-coupled receptors. *Nucleic acids research*, 31(1):294–297, 2003.
- [180] Nicolas Le Novère and Jean-Pierre Changeux. Lgicdb: the ligand-gated ion channel database. *Nucleic acids research*, 29(1):294–295, 2001.
- [181] Dániel Kozma, István Simon, and Gábor E Tusnády. Pdbtm: Protein data bank of transmembrane proteins after 8 years. *Nucleic acids research*, 41(D1):D524–D529, 2012.

- [182] László Dobson, Tamás Langó, István Reményi, and Gábor E Tusnady. Expediting topology data gathering for the topdb database. *Nucleic acids research*, 43(D1):D283–D289, 2014.
- [183] Georgios N Tsaousis, Konstantinos D Tsirigos, Xanthi D Andrianou, Theodore D Liakopoulos, Pantelis G Bagos, and Stavros J Hamodrakas. Ex-topodb: a database of experimentally derived topological models of transmembrane proteins. *Bioinformatics*, 26(19):2490–2492, 2010.
- [184] Julia Varga, László Dobson, and Gábor E Tusnady. Topdom: database of conservatively located domains and motifs in proteins. *Bioinformatics*, 32(17):2725–2726, 2016.
- [185] Sunhwan Jo, Taehoon Kim, Vidyashankara G Iyer, and Wonpil Im. Charmm-gui: a web-based graphical user interface for charmm. *Journal of computational chemistry*, 29(11):1859–1865, 2008.
- [186] Bernard R Brooks, Robert E Bruccoleri, Barry D Olafson, David J States, S a Swaminathan, and Martin Karplus. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*, 4(2):187–217, 1983.
- [187] Michele Clamp, James Cuff, Stephen M Searle, and Geoffrey J Barton. The jalview java alignment editor. *Bioinformatics*, 20(3):426–427, 2004.
- [188] Daniel WA Buchan, Federico Minneci, Tim CO Nugent, Kevin Bryson, and David T Jones. Scalable web services for the psipred protein analysis workbench. *Nucleic acids research*, 41(W1):W349–W357, 2013.
- [189] Guy Yachdav, Tatyana Goldberg, Sebastian Wilzbach, David Dao, Iris Shih, Saket Choudhary, Steve Crouch, Max Franz, Alexander García, Leyla J García, et al. Cutting edge: anatomy of biojs, an open source community for the life sciences. *Elife*, 4:e07009, 2015.
- [190] Chris Marrin. WebGL specification. *Khronos WebGL Working Group*, 2011.
- [191] Marco Biasini. pv - WebGL protein viewer. <https://github.com/biasmv/pv>. Accessed: 2016-09-30.
- [192] Alexander S Rose and Peter W Hildebrand. Ngl viewer: a web application for molecular visualization. *Nucleic acids research*, 43(W1):W576–W579, 2015.

- [193] Stephen H White. The progress of membrane protein structure determination. *Protein Science*, 13(7):1948–1949, 2004.
- [194] Norbert Kučerka, John F Nagle, Jonathan N Sachs, Scott E Feller, Jeremy Pencer, Andrew Jackson, and John Katsaras. Lipid bilayer structure determined by the simultaneous analysis of neutron and x-ray scattering data. *Biophysical journal*, 95(5):2356–2367, 2008.
- [195] J. A. Poveda, A. M. Giudici, M. L. Renart, M. L. Molina, E. Montoya, A. Fernández-Carvajal, G. Fernández-Ballester, J. A. Encinar, and J. M. González-Ros. Lipid modulation of ion channels through specific binding sites. *Biochimica et Biophysica Acta - Biomembranes*, 1838(6):1560–1567, 2014.
- [196] Vadim Cherezov, Daniel M Rosenbaum, Michael A Hanson, Søren GF Rasmussen, Foon Sun Thian, Tong Sun Kobilka, Hee-Jung Choi, Peter Kuhn, William I Weis, Brian K Kobilka, et al. High-resolution crystal structure of an engineered human β 2-adrenergic g protein-coupled receptor. *science*, 318(5854):1258–1265, 2007.
- [197] JP Walsh and RM Bell. sn-1, 2-diacylglycerol kinase of escherichia coli. structural and kinetic analysis of the lipid cofactor dependence. *Journal of Biological Chemistry*, 261(32):15062–15069, 1986.
- [198] W Richard Burack and Rodney L Biltonen. Lipid bilayer heterogeneities and modulation of phospholipase a2 activity. *Chemistry and physics of lipids*, 73(1-2):209–222, 1994.
- [199] Seul Ki Min, Gwang Heum Yoon, Jung Hyun Joo, Sang Jun Sim, and Hwa Sung Shin. Mechanosensitive physiology of chlamydomonas reinhardtii under direct membrane distortion. *Scientific reports*, 4:4675, 2014.
- [200] F.-Xabier Contreras, Lisette Snchez-Magraner, Alicia Alonso, and Flix M. Goi. Transbilayer (flip-flop) lipid motion and lipid scrambling in membranes. *FEBS Letters*, 584(9):1779 – 1786, 2010. Frontiers in Membrane Biochemistry.
- [201] Victor Sojo, Christophe Dessimoz, Andrew Pomiankowski, and Nick Lane. Membrane proteins are dramatically less conserved than water-soluble proteins across the tree of life. *Molecular biology and evolution*, 33(11):2874–2884, 2016.

- [202] Nicolas J Tourasse and Wen-Hsiung Li. Selective constraints, amino acid composition, and the rate of protein evolution. *Molecular biology and evolution*, 17(4):656–664, 2000.
- [203] Amit Oberai, Nathan H Joh, Frank K Pettit, and James U Bowie. Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. *Proceedings of the National Academy of Sciences*, 106(42):17747–17750, 2009.
- [204] Phillip J Stansfeld, Elizabeth E Jefferys, and Mark SP Sansom. Multiscale simulations reveal conserved patterns of lipid interactions with aquaporins. *Structure*, 21(5):810–819, 2013.
- [205] Tsjerk A Wassenaar, Helgi I Ingolfsson, Rainer A Bockmann, D Peter Tieleman, and Siewert J Marrink. Computational lipidomics with insane: a versatile tool for generating custom membranes for molecular simulations. *Journal of chemical theory and computation*, 11(5):2144–2155, 2015.
- [206] Richard J. Gowers, Max Linke, Jonathan Barnoud, Tyler J. E. Reddy, Manuel N. Melo, Sean L. Seyler, Jan Domaski, David L. Dotson, Sbastien Buchoux, Ian M. Kenney, and Oliver Beckstein. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In Sebastian Benthall and Scott Rostrup, editors, *Proceedings of the 15th Python in Science Conference*, pages 98 – 105, 2016.
- [207] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, volume 836, pages 11—15, 2008.
- [208] Songrit Maneewongvatana and David M. Mount. Analysis of approximate nearest neighbor searching with clustered point sets. *CoRR*, cs.CG/9901013, 1999.
- [209] Travis E. Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20, 2007.
- [210] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

- [211] Robbie P Joosten, Tim AH Te Beek, Elmar Krieger, Maarten L Hekkelman, Rob WW Hooft, Reinhard Schneider, Chris Sander, and Gert Vriend. A series of pdb related databases for everyday needs. *Nucleic acids research*, 39(suppl_1):D411–D419, 2010.
- [212] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.
- [213] Oscar D Monera, Terrance J Sereda, Nian E Zhou, Cyril M Kay, and Robert S Hodges. Relationship of sidechain hydrophobicity and α -helical propensity on the stability of the single-stranded amphipathic α -helix. *Journal of peptide science*, 1(5):319–329, 1995.
- [214] Jimin Pei and Nick V Grishin. Promals: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, 23(7):802–808, 2007.
- [215] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D³ data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.
- [216] Arthur Lesk. *Introduction to Bioinformatics*. Oxford University Press, 2014.
- [217] Naomi K Fox, Steven E Brenner, and John-Marc Chandonia. Scope: Structural classification of proteinextended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309, 2013.
- [218] C Preston Moon and Karen G Fleming. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proceedings of the National Academy of Sciences*, 108(25):10174–10177, 2011.
- [219] Oliver S Smart, Julia M Goodfellow, and BA Wallace. The pore dimensions of gramicidin a. *Biophysical journal*, 65(6):2455–2460, 1993.
- [220] Sang Ho Park and Stanley J Opella. Tilt angle of a trans-membrane helix is determined by hydrophobic mismatch. *Journal of molecular biology*, 350(2):310–318, 2005.
- [221] Maddalena Venturoli, Berend Smit, and Maria Maddalena Sperotto. Simulation studies of protein-induced bilayer deformations, and lipid-induced protein tilting, on a mesoscopic model for lipid bilayers with embedded proteins. *Biophysical journal*, 88(3):1778–1798, 2005.

- [222] Lei Shi, Alessandro Cembran, Jiali Gao, and Gianluigi Veglia. Tilt and azimuthal angles of a transmembrane peptide: a comparison between molecular dynamics calculations and solid-state nmr data of sarcolipin in lipid membranes. *Biophysical journal*, 96(9):3648–3662, 2009.
- [223] Kaustubh R Mote, T Gopinath, and Gianluigi Veglia. Determination of structural topology of a membrane protein in lipid bilayers using polarization optimized experiments (poe) for static and mas solid state nmr spectroscopy. *Journal of biomolecular NMR*, 57(2):91–102, 2013.
- [224] Chaim A Schramm, Brett T Hannigan, Jason E Donald, Chen Keasar, Jeffrey G Saven, William F DeGrado, and Ilan Samish. Knowledge-based potential for positioning membrane-associated structures and assessing residue-specific energetic contributions. *Structure*, 20(5):924–935, 2012.
- [225] Frances Jane Sharom. Complex interplay between the p-glycoprotein multidrug efflux pump and the membrane: its role in modulating protein function. *Frontiers in oncology*, 4, 2014.
- [226] Yuan Gao, Erhu Cao, David Julius, and Yifan Cheng. Trpv1 structures in nanodiscs reveal mechanisms of ligand and lipid action. *Nature*, 534(7607):347, 2016.
- [227] Zukang Feng, Li Chen, Himabindu Maddula, Ozgur Akcan, Rose Oughtred, Helen M Berman, and John Westbrook. Ligand depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, 20(13):2153–2155, 2004.
- [228] Mihoko Saito, Naomi Takemura, and Tsuyoshi Shirai. Classification of ligand molecules in pdb with fast heuristic graph match algorithm complig. *Journal of molecular biology*, 424(5):379–390, 2012.
- [229] Brad Barber and H Huhdanpaa. Qhull. *The Geometry Center, University of Minnesota*, <http://www.geom.umn.edu/software/qhull/>, 1995.
- [230] David J Wales and Jonathan PK Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1997.
- [231] Michael JD Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in optimization and numerical analysis*, pages 51–67. Springer, 1994.

- [232] Greg Landrum. Rdkit: Open-source cheminformatics. <http://www.rdkit.org>. Accessed: 2016-09-21.
- [233] GP Moss. Nomenclature of steroids (recommendations 1989). *Pure and Applied Chemistry*, 61(10):1783–1822, 1989.
- [234] Wei Wu, Xiaoshan Shi, and Chenqi Xu. Regulation of t cell signalling by membrane lipids. *Nature Reviews Immunology*, 16(11):690–701, 2016.
- [235] Shigekazu Nagata, J Suzuki, K Segawa, and T Fujii. Exposure of phosphatidylserine on the cell surface. *Cell Death & Differentiation*, 23(6):952–961, 2016.
- [236] Min Zhou, Nina Morgner, Nelson P Barrera, Argyris Politis, Shoshanna C Isaacson, Dijana Matak-Vinković, Takeshi Murata, Ricardo A Bernal, Daniela Stock, and Carol V Robinson. Mass spectrometry of intact v-type atpases reveals bound lipids and the effects of nucleotide binding. *Science*, 334(6054):380–385, 2011.
- [237] Eamonn Reading, Ildir Liko, Timothy M Allison, Justin LP Benesch, Arthur Laganowsky, and Carol V Robinson. The role of the detergent micelle in preserving the structure of membrane proteins in the gas phase. *Angewandte Chemie International Edition*, 54(15):4577–4581, 2015.
- [238] Aneika C Leney, Lindsay M McMorran, Sheena E Radford, and Alison E Ashcroft. Amphipathic polymers enable the study of functional membrane proteins in the gas phase. *Analytical chemistry*, 84(22):9841–9847, 2012.
- [239] Jonathan TS Hopper, Yvonne Ting-Chun Yu, Dianfan Li, Alison Raymond, Mark Bostock, Ildir Liko, Victor Mikhailov, Arthur Laganowsky, Justin LP Benesch, Martin Caffrey, et al. Detergent-free mass spectrometry of membrane protein complexes. *Nature methods*, 10(12):1206–1208, 2013.
- [240] CRAIG M Whitehouse, RN Dreyer, M Yamashita, and JB Fenn. Electrospray ionization for mass-spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.
- [241] Koichi Tanaka, Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida, Tamio Yoshida, and T Matsuo. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid communications in mass spectrometry*, 2(8):151–153, 1988.

- [242] Francesc-Xabier Contreras, Andreas Max Ernst, Felix Wieland, and Britta Brügger. Specificity of intramembrane protein–lipid interactions. *Cold Spring Harbor perspectives in biology*, 3(6):a004705, 2011.
- [243] Mizuki Morita, AVSK Mohan Katta, Shandar Ahmad, Takaharu Mori, Yuji Sugita, and Kenji Mizuguchi. Lipid recognition propensities of amino acids in membrane proteins from atomic resolution data. *BMC biophysics*, 4(1):21, 2011.
- [244] Sheng Yin, Yongming Xie, and Joseph A Loo. Mass spectrometry of protein–ligand complexes: enhanced gas-phase stability of ribonuclease–nucleotide complexes. *Journal of the American Society for Mass Spectrometry*, 19(8):1199–1208, 2008.
- [245] Timothy C Anglin, Michael P Cooper, Hao Li, Katherine Chandler, and John C Conboy. Free energy and entropy of activation for phospholipid flip-flop in planar supported lipid bilayers. *The Journal of Physical Chemistry B*, 114(5):1903–1914, 2010.
- [246] David L Daleke. Regulation of transbilayer plasma membrane phospholipid asymmetry. *Journal of lipid research*, 44(2):233–242, 2003.
- [247] RB Birge, S Boeltz, S Kumar, J Carlson, J Wanderley, D Calianese, M Barcinski, RA Brekken, X Huang, JT Hutchins, et al. Phosphatidylserine is a global immunosuppressive signal in efferocytosis, infectious disease, and cancer. *Cell death and differentiation*, 23(6):962, 2016.
- [248] Hee-Yong Kim, Bill X Huang, and Arthur A Spector. Phosphatidylserine in the brain: metabolism and function. *Progress in lipid research*, 56:1–18, 2014.
- [249] Makiko Meguro, Akiko Kashiwagi, Kohzoh Mitsuya, Mitsuyoshi Nakao, Ikuko Kondo, Shinji Saitoh, and Mitsuo Oshimura. A novel maternally expressed gene, atp10c, encodes a putative aminophospholipid translocase associated with angelman syndrome. *Nature genetics*, 28(1):19, 2001.
- [250] Nicoletta Pedemonte and Luis JV Galiotta. Structure and function of tmem16 proteins (anoctamins). *Physiological reviews*, 94(2):419–459, 2014.
- [251] Björn Christian Schroeder, Tong Cheng, Yuh Nung Jan, and Lily Yeh Jan. Expression cloning of tmem16a as a calcium-activated chloride channel subunit. *Cell*, 134(6):1019–1029, 2008.

- [252] Heidi Koldsø, David Shorthouse, Jean Hélie, and Mark SP Sansom. Lipid clustering correlates with membrane curvature as revealed by molecular simulations of complex lipid bilayers. *PLoS computational biology*, 10(10):e1003911, 2014.
- [253] Tom Newport. tomnewport/alchex: Alchex pre-alpha release. <https://doi.org/10.5281/zenodo.895349>, 2017.
- [254] Sascha Vermeer, Alexander Hoischen, Rowdy PP Meijer, Christian Gilissen, Kornelia Neveling, Nienke Wieskamp, Arjan de Brouwer, Michel Koenig, Mathieu Anheim, Mirna Assoum, et al. Targeted next-generation sequencing of a 12.5 mb homozygous region reveals ano10 mutations in patients with autosomal-recessive cerebellar ataxia. *The American Journal of Human Genetics*, 87(6):813–819, 2010.
- [255] Thomas J. Bruno William M. Haynes, David R. Lide. Properties of amino acids. In *CRC Handbook of Chemistry and Physics, 95th Edition (100 Key Points)*, chapter 7, pages 7a,7b. CRC Press, 2014.
- [256] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1):014101, 2007.
- [257] Warren L DeLano. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter On Protein Crystallography*, 40:82–92, 2002.
- [258] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.
- [259] Benjamin Webb and Andrej Sali. Protein structure modeling with modeller. *Protein Structure Prediction*, pages 1–15, 2014.
- [260] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An n log (n) method for ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.
- [261] M. Parrinello and A Rahman. Crystal structure and pair potentials: A molecular-dynamics study. *Physical Review Letters*, 45(14):1196, 1980.
- [262] Matthias Heinig and Dmitrij Frishman. Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic acids research*, 32(suppl_2):W500–W502, 2004.

- [263] Evgeny Krissinel and Kim Henrick. Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, 372(3):774–797, 2007.
- [264] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, 2014.
- [265] Elizabeth Jefferys, Zara A Sands, Jiye Shi, Mark SP Sansom, and Philip W Fowler. Alchembed: a computational method for incorporating multiple proteins into complex lipid geometries. *Journal of chemical theory and computation*, 11(6):2743–2754, 2015.
- [266] Paolo Manghi, Natalia Manola, Wolfram Horstmann, and Dale Peters. An infrastructure for managing ec funded research output-the openaire project. *The Grey Journal (TGJ): An International Journal on Grey Literature*, 6(1), 2010.