Imperial College London
Department of Life Sciences

Integrating theory and experiments to study complex protein dynamics at the surface of lipid membranes

Máximo Sanz Hernández

Submitted for the degree of Doctor of Philosophy, September 2018

Abstract

The cellular membrane is a complex biophysical environment that occupies a central position in many fundamental biological processes. In association with the membrane there are a wide variety of proteins, macromolecular actors with crucial roles in driving those processes. Being a primary point of contact between the inside and outside of cells, membrane proteins are targeted by the majority of approved therapeutic drugs. New methods that can unveil the mechanisms by which these molecules function are therefore crucial in deepening our biomolecular understanding and can provide rational approaches for the design of novel therapeutic approaches.

Conformational dynamics are an important factor in driving the behaviour of membrane proteins. Receptors, channels and other membrane-embedded proteins function by populating distinct structural states and interconverting between them. However, the molecular detail of these heterogeneous processes is extremely challenging to elucidate with the current analytical tools, generally tailored for homogeneous protein states.

This thesis focuses on the methodological development of approaches to accurately characterise the dynamical behaviour of membrane proteins. The methods outlined combine the use of experiments and theoretical sampling to generate ensemble descriptions of membrane proteins. Specifically, the experimentally-restrained molecular dynamics approach is further developed for the incorporation of solid-state NMR observables.

The accuracy of this method is validated in the context of membrane proteins and subsequently applied to the study of three different biological systems: the cardiac regulatory protein phospholamban, the influenza proton channel M2 and the G-protein coupled receptor CXCR1. The generated ensembles reveal some important conformational transitions in these membrane proteins, and highlight the importance of combining theory and experiments in order to achieve an accurate, detailed description of biomolecular dynamics in lipid membranes.

Declaration

I declare that the work contained in this document is my own except when stated otherwise in the text by reference or acknowledgment. The thesis was composed by myself in its entirety and has not been submitted for any other degree or qualification.

Parts of the work here presented have led to the publication of two peer-reviewed articles:

- Sanz-Hernández, Vostrikov, Veglia & De Simone. Accurate determination of conformational transitions in oligomeric membrane proteins. *Scientific Reports* (2016).
- Sanz-Hernández & De Simone. The PROSECCO server for chemical shift predictions in ordered and disordered proteins. *Journal of Biomolecular NMR* (2017).

Figures from these two publications have been adapted or reproduced in this thesis (when specified in the figure description). The original material is licensed under a Creative Commons CC BY 4.0 license. A copy of the license can be found at

https://creativecommons.org/licenses/by/4.0/

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work

Acknowledgements

First of all I am hugely grateful to Alfonso for his constant support and guidance over the past six years. You have taught me so many things since I started taking my first steps in research, but most importantly always encouraged me to keep learning and dive fearlessly into new disciplines and challenges.

From the other side of the Atlantic, I thank Gianluigi for his insights, time and openness to critically discuss our results, hugely broadening my perspective on the power of experiments. Suhail for his endless patience and dedication to fixing (and teaching me how to fix myself) the innumerable computational errors I have encountered.

I would especially like to thank all the people who have come and gone over these years on the 6$^{\text{th}}$ floor of SEC. Particular thanks go to Jay, Justin and Flavia, who took the time to chaperone me when I went back into the lab, enduring too many basic questions from my long-forgotten Biochemistry lectures. Sarah and James for all their explanations, helping me through my head-scratching moments with MD and solid-state. Carlos for sharing my coding frustrations this past year and always helping out when Nvidia Santa came around. Yingqi for the many hours of NMR he set up. And many who were always willing to hear me complain about the lab, and so often gave me solutions to solve my problems: Mostafa, Nessim, Joe, Inma, Lee, Nick, as well as everybody who over all these years I have had the great chance of working with and sharing everyday joys and frustrations.

Huge thanks go to a very special group of friends who have accompanied me since I first arrived to Imperial as a very naïve boy, having since moved on with their lives (and still loving me even if I haven't).

A very warm thank you to my family, which I have enlarged since I started this endeavour. Nerina, Alessandro, Costanza and Margherita, thank you for welcoming me into your beautiful lives and being a newly found source of support.

I thank my brother Dédalo, the most ingenious person I know, for leading the way in our adventures and showing me the great worth of exploring the world. And Irene, for always knowing how to convey her peace and joy to me, in good and bad moments.

To my parents, Ángela and Máximo, I owe the priceless joy I feel every time I understand something new about the world. Thank you for showing me the power of questioning everything (including myself), for imparting the value of curiosity and enabling me with the tools to entertain it, but never fully satisfy it.

And finally, I would like to thank Dedè, a pillar that never let me collapse even when I was staggering, who knows when to push me further along the way and when to help me step back to gain perspective. Thank you for giving colour to my research and every single one of my days.

Contents

Abstract **3**

Declaration **5**

Acknowledgements **7**

List of tables **13**

List of figures **15**

Abbreviations **19**

1 Introduction **21**

1.1 Protein structure and dynamics in the context of cellular membranes 21

1.2 Experimental approaches for the study of membrane protein dynamics . . . 27

1.3 Theoretical methods in structural biology 30

1.4 Combining experiments and theory to characterise protein dynamics 35

1.5 Characterising membrane protein conformational landscapes 37

1.6 Summary and structure of the thesis 39

2 Development and testing of NMR restraints 41

2.1 Chemical shifts 41

2.1.1 Physical principles and challenges in their interpretation 41

2.1.2 An overview of chemical shift modelling in proteins 44

2.2 PROSECCO: PROtein SEquences and Chemical shift COrrelations 47

2.2.1 Curating a chemical shift database for a new predictor 47

2.2.2 $\text{PROSECCO}_{\text{IDP}}$: Improving the prediction of chemical shifts in disordered proteins 49

2.2.3 Predicting chemical shifts from sequence in secondary structure elements 53

2.2.4 $\text{PROSECCO}_{\text{FOLDED}}$: an accurate sequence-based method to predict chemical shifts in structured proteins 57

2.2.5 An outlook for improving protein CS models 64

2.3 Oriented solid-state NMR 65

2.3.1 Chemical shift anisotropy and dipolar coupling 65

2.3.2 Implementing ssNMR restraints in molecular simulations 70

3 Structural heterogeneity of phospholamban and its complex with SERCA 83

3.1 Phospholamban and the regulation of the SERCA pump 83

3.2 Conformational transitions in pentameric phospholamban 87

3.2.1 Methods: Restrained simulations setup and structural calculations . . 87

3.2.2 Experimentally-driven protein dynamics of the PLN pentamer 90

3.2.3 Phosphorylation leads to the dynamical activation of PLN 91

3.2.4 Calcium channeling properties in pentameric PLN 95

3.3 The dynamical complex between SERCA and phospholamban 98

3.3.1 Methods: Restrained simulations setup 99

3.3.2 Restrained simulations reproduce experimental results 105

3.3.3 Transient disorder-mediated interaction between SERCA and PLN . . 107

4 Dynamical studies of the influenza M2 tetramer 115

4.1 The role of M2 in influenza 115

4.2 Methods: Simulation setup and ensemble analysis 117

4.2.1 Restrained simulations setup 117

4.2.2 Structural and hydration analyses 118

4.3 Topological properties of the M2 tetramer 119

4.4 Pore profile of the influenza channel 121

4.5 Role of hydration in proton transport by M2 125

5 Structural dynamics of the G-protein coupled receptor CXCR1 129

5.1 CXCR1: an interleukin-8 receptor 129

5.2 Methods: DC-restrained simulations and analysis of CXCR1 ensembles . . . 131

5.2.1 Solid-state NMR restrained simulations 131

5.2.2 Network description of CXCR1 133

5.3 Experimentally derived dynamics of a GPCR 135

5.4 Allosteric communication network of CXCR1 139

6 Discussion **143**

6.1 Toward an accurate description of membrane protein conformational ensembles 143

6.1.1 Improving the interpretation of protein chemical shifts 143

6.1.2 Accurate topological dynamics of membrane proteins 145

6.2 Uncovering insights into membrane protein dynamics 148

6.2.1 Dynamical activation of PLN and its complex with SERCA 148

6.2.2 Conformational dynamics and hydration properties of the M2 influenza channel 151

6.2.3 Allosteric motions revealed in the IL-8 receptor CXCR1 153

6.3 Future perspectives 154

Bibliography **157**

List of Tables

3.1 Force constants employed in the simulations of pentameric phospholamban . 88

3.2 Experimental PRE-derived distances for the SERCA-PLN complex in its B-state 103

3.3 Experimental crosslinking-derived distances for the SERCA-PLN complex in its B-state 103

5.1 Residues used to define the transmembrane helical axes of CXCR1 136

List of Figures

1.1 Conformational landscapes of proteins 22

1.2 Functional dynamics of membrane proteins 25

1.3 Large-scale dynamics resolved by Cryo Electron Microscopy 28

1.4 Tetrameric KcsA potassium channel conductance mechanism revealed by Molecular Dynamics 32

1.5 Conformational equilibrium of phospholamban revealed by ssNMR-restrained replica-averaged MD simulations 38

2.1 Electronic shielding phenomena that influence chemical shifts 42

2.2 Method of kernel density estimation employed in $\text{PROSECCO}_{\text{IDP}}$ 50

2.3 Calibration of the weight normalising factor N_w 52

2.4 Benchmark of sequence-based prediction of CS in IDPs 53

2.5 Application of kernel-based CS prediction to secondary structure elements . 54

2.6 The effect of secondary structure boundaries on the CS 56

2.7 Neural network structure in $\text{PROSECCO}_{\text{FOLDED}}$ 58

2.8 Performance benchmark for $\text{PROSECCO}_{\text{FOLDED}}$ 59

2.9 Direct comparison between $\text{PROSECCO}_{\text{FOLDED}}$ and SPARTA+ for individual BMRB entries 60

2.10 Chemical shift prediction in the GA^{95}/GB^{95} system 62

2.11 Methyl sidechain predictions and pH-dependent corrections in PROSECCO . 63

2.12 Structural models for chemical shift anisotropy and dipolar coupling 67

2.13 CSA and DC as a footprint of helical orientation 69

2.14 Parallelisation of the ssNMR restrained MD algorithm 73

2.15 Internal averaging scheme 76

2.16 The D_{iff} metric for comparison of distributions 78

2.17 Illustration of the reference ensemble validation method 79

2.18 Validation of the restrained sampling approach in the phospholamban pentamer 81

3.1 Sequence and structure of monomeric phospholamban 84

3.2 Conformational heterogeneity of phospholamban 86

3.3 Agreement between PLN pentamer simulations and the experimental CSA and DC 89

3.4 Dynamical activation of PLN upon phosphorylation 93

3.5 Static fitting of CSA and DC data to phospholamban 94

3.6 Channeling properties of pentameric phospholamban 96

3.7 Internal transmembrane motions in phosphorylated phospholamban 97

3.8 Structure of the PLN-SERCA complex in the B-state 101

3.9 Q-factors for the ensemble of PLN in complex with SERCA 105

3.10 Distribution of transmembrane tilt angles of PLN in complex with SERCA . 106

3.11 Isotropic chemical shifts of SERCA-bound PLN 108

3.12 Dynamical cytoplasmic interaction between SERCA and PLN 109

3.13 Electrostatic interaction network between SERCA and PLN 111

3.14 Diversity in the electrostatic interactions of R9 and R13/R14 with SERCA . 112

4.1 Structure of the M2 influenza A proton channel in lipid bilayers 116

4.2 Agreement between experimental and back-calculated CSA and DC in M2 ensembles 120

4.3 Distribution of topological properties in the structural ensembles of M2 . . . 121

4.4 Pore radius profile of the M2 influenza channel 122

4.5 M2 channel fluctuations in relation to the topological motions of helices . . . 124

4.6 Hydration properties of the M2 proton channel 125

4.7 Structure of water molecules within the pore of M2 126

5.1 Structure of the CXCR1 receptor in lipid bilayers 130

5.2 Agreement between experimental and back-calculated DCs in CXCR1 ensembles 136

5.3 Helical reorientations between H5 and H6 of CXCR1 137

5.4 Generalised correlation matrix of C_α atoms in the CXCR1 ensemble 139

5.5 Allosteric network structure of CXCR1 141

5.6 Allosteric paths between distant loops of CXCR1 142

Abbreviations

ANN - Artificial neural network

ATP - Adenosine triphosphate

BMRB - Biological magnetic resonance data bank

CPU - Central processing unit

CS - Chemical shift(s)

CSA - Chemical shift anisotropy

CSI - Chemical shift index

DC - Dipolar coupling

DMPC - 1,2-dimyristoyl-sn-glycero-3-phosphocholine

DOPC - 1,2-dioleoyl-sn-glycero-3-phosphocholine

DOPE - 1,2-dioleoyl-sn-glycero-3-phosphoethanolamine

ECL - Extracellular loop

EM - Electron microscopy

GPCR - G-protein coupled receptor

GPU - Graphics processing unit

ICL - Intracellular loop

IDP - Intrinsically disordered protein

IDR - Intrinsically disordered region

IL - Interleukin

MAS - Magic angle spinning

MC - Monte Carlo

MD - Molecular dynamics

NMR - Nuclear magnetic resonance

PDB - Protein data bank

PKA - Protein kinase A

PISEMA - Polarization inversion spin-exchange at the magic-angle

PLN - Phospholamban

PRE - Paramagnetic relaxation enhancement

RA - Rotationally aligned

RDC - Residual dipolar coupling

RMSD - Root mean squared deviation

RMSF - Root mean square fluctuation

RNA - Ribonucleic acid

SERCA - Sarco/endoplasmic reticulum Ca^{2+}-ATPase

SLF - Separated local field

SR - Sarcoplasmic reticulum

ssNMR - Solid-state nuclear magnetic resonance

Chapter 1

Introduction

1.1 Protein structure and dynamics in the context of cellular membranes

The study of protein dynamics is a topic of great interest in biophysical research. Initial descriptions of the importance of protein motions arose following the establishment of the field of structural biology. The first protein three-dimensional structures were determined by X-ray crystallography: myoglobin by Kendrew and co-workers,[1] and haemoglobin by Perutz and co-workers.[2] In a crucial result in his studies, Perutz would furthermore describe that in the reaction of haemoglobin with oxygen a structural change occurs in the protein,[3] building evidence that protein structures are not static objects. During the following decades, a description of the process of protein folding was progressively built, starting from the initial observations that a protein's amino-acid sequence determines their tertiary structure.[4] This was followed up by energetic and kinetic descriptions of the process,[5,6] described as an energy funnel that guides a protein from an extended, unfolded state into its native structural conformation. Crucially, this energy funnel was described as a 'rugged' surface, where different possible local energy minima can be metastable either as intermediates of the folding process, or even co-existing as multiple conformations

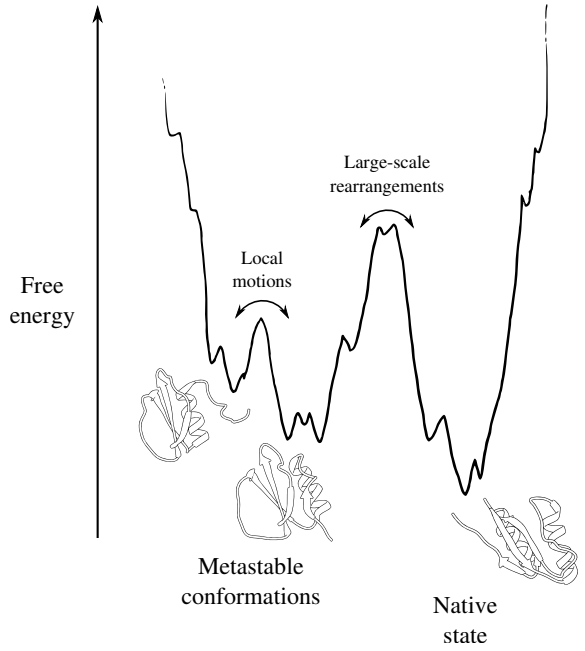


Figure 1.1: Conformational landscapes of proteins – Depiction of a simplified protein conformational landscape along a hypothetical one-dimensional reaction coordinate. The rugged nature of the surface means that proteins constantly fluctuate between different local conformations that are structurally similar. Large energy barriers are often associated with large-scale structural rearrangements that result in metastable conformations. Interconversion between states separated by such barriers normally occur in long timescales (microsecond to millisecond), whereas local motions are often in the nanosecond range.

of an ensemble of native states.[7–9]

The conceptual description of protein folding in terms of energy was therefore expanded to consider the complete conformational space that proteins can adopt, known as their conformational landscape, and studied both theoretically[10,11] and experimentally.[12–14] A simplified protein conformational landscape is depicted in Figure 1.1. The consensus arose that proteins do not behave as static objects, but rather make use of motions to function, populating different states or conformations along an often complex energy landscape.[15] It is a finely-tuned equilibrium and interchange between different structural states along the landscape that allows proteins to perform a multitude of fundamental biological functions. This principle of protein conformational plasticity has indeed been shown to be essential for key processes such as molecular recognition in protein complexes[16–18] or enzyme catalysis,[19–23] amongst many others.[24–30]

During those years the rise of nuclear magnetic resonance (NMR) as a tool for structural determination of proteins[31] also played an important role in the inclusion of dynamics into the structural biology field. With NMR scientists had the ability to probe the structures of proteins in their native solution conformations, where they were no longer trapped inside a crystal lattice and were free to move and interchange between different structural states. The field started to see the development of new NMR methods to gain insight into these motions. In particular, the NMR relaxation phenomenon in proteins started being probed as a tool to access the high resolution structures of different co-existing conformational states, as well as gaining information of their relative lifetimes and populations.[32] Given the essentiality of including dynamics in the picture when thinking of a protein's structure and function, many people in the field of structural biology progressively began to adopt an 'ensemble' approach to protein structure determination, whereby efforts are made to characterise not only the lowest energy conformational state, but obtain a complete picture of the full conformational landscape that a protein populates.[33] In recent times, the protein ensemble approach has become even more relevant in the study of intrinsically disordered proteins (IDPs), which lack a well-defined native folded conformation but rather exist as a collection of constantly interchanging unfolded states.[34]

Despite the great interest in characterising the dynamics of proteins, building an exhaustive description of the conformational landscape of proteins is no easy task. Most experimental techniques are well-suited to characterising only the low-energy, mostly populated conformation of a protein, traditionally referred to as its native state. The methods that can tackle dynamics often have limitations, such as only being able to access limited timescales of interconversion between structures, lacking in resolution or providing only average observables of a heterogeneous pool of conformations.[35] Theoretical models of protein dynamics have therefore been developed and exploited in combination with experimental information in order to gain a more detailed description of protein structural ensembles. In this context, molecular dynamics (MD) simulations have proven very successful in their ability to obtain valuable information about protein dynamics that is inaccessible to experiments. However, theoretical approaches also have shortcomings of

their own. Dynamical models are built from first-principle physical interactions, which are expressed as energy functions known as force fields, semi-empirically parametrised based on the available experimental evidence and considerations from quantum mechanical models.[11] Given the semi-empirical nature of this parametrisation, there are intrinsic errors in the force fields that can often yield inaccurate results,[36,37] despite the very encouraging recent improvements in protein force field accuracy.[38,39] Furthermore, the complete sampling of a protein conformational landscape is too exhaustive in most cases to be achieved by standard MD simulations. The long timescales at which large-scale molecular motions occur are currently inaccessible to MD simulations. Although current times are seeing an exceptionally fast growth in the processing power of computational resources, we are far from reaching the necessary simulation timescales to achieve a complete sampling of most protein energy landscapes.

These challenges that emerge from both experimental and theoretical approaches become even greater when studying proteins that are in association with biological membranes. In classical structural biology, the characterisation of membrane proteins has traditionally lagged behind that of their solution counterparts, primarily due to the difficulties in adequately reproducing their hydrophobic native environment *in vitro*.[40] Similarly, theoretical studies on membrane protein dynamics have also arisen long after those of soluble proteins.[41] It is clear however that despite this inherent difficulties, it is of great importance to study proteins at the membrane environment, where crucial protein-driven biochemical processes take place. The cellular membrane is primarily formed by a bilayer of lipid molecules that provides a barrier between the cell and its surroundings. In association with lipids there are carbohydrates and proteins, the latter of which are key players in regulating the interactions between the cell and the outside world. Membrane proteins receive and transduce chemical signals, transfer molecules across the bilayer, adhere to neighbouring cells and generate energy from metabolic processes.[42] In addition, being displayed at the surface of the cell makes membrane proteins a good target for medicinal drugs. Indeed, membrane proteins have become crucial in modern therapeutics, with approximately 60% of approved drugs targeting them.[43]

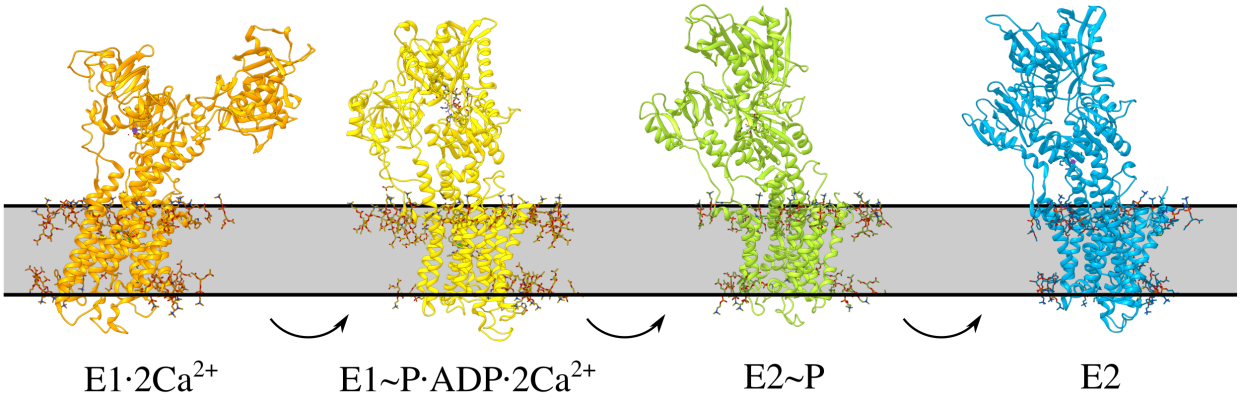


Figure 1.2: Functional dynamics of membrane proteins – The Sarco/endoplasmic Reticulum Ca^{2+} ATPase (SERCA) is responsible for heart muscle relaxation. This membrane protein actively transports calcium ions across the membrane of the sarcoplasmic reticulum, hydrolysing ATP in the process. To carry out its function it undergoes a series of conformational changes, cycling through different structural isoforms. Four of these isoforms, here depicted, have been crystallised in lipid membranes by Norimatsu and co-workers,[53] and show the large extent of conformational plasticity required for the protein to exert its function.

In carrying out these very important functions, membrane proteins need to make use of dynamics. Signals are often transduced by allosteric changes in protein structural conformations upon binding of a chemical, channels and transporters fluctuate between open and closed conformations, protein complexes associate and dissociate, all as part of very complex dynamical equilibria. Despite the challenges in the study of these dynamical mechanisms, there is substantial evidence for their importance in processes such as the regulation of membrane protein self-assembly,[44] protein-protein interactions,[45] signal transduction,[46–48] transport across the membrane[49,50] or cellular import and export.[51] An example of the importance of dynamics for membrane protein function is shown in Figure 1.2, which depicts the enzymatic cycle of the Sarco/endoplasmic Reticulum Ca^{2+} ATPase (SERCA). In cardiac myocites, SERCA is responsible for the active transport of calcium ions from the cell cytosol into the lumen of the sarcoplasmic reticulum, leading to cardiac muscle relaxation.[52] This process is governed by drastic conformational changes in the pump over each heartbeat cycle, which have been revealed by X-ray crystallography.[53] These conformational changes are also accompanied by transient interactions with two other membrane protein regulatory partners: sarcolipin and phospholamban (PLN), which alter the pump's affinity for calcium in an allosteric manner.[54]

Dynamics are not only crucial in the correct functioning of membrane proteins, but they are often also associated with the development of disease. In many cases, the misregulation of the natural dynamics of membrane proteins can lead to aberrant behaviours. In the case of SERCA, disruptions to the allosteric interactions between the pump and its regulatory partners can induce dilated cardiomyopathy.[55] Aberrant dynamics have also been described in ion channel voltage sensor domains, leading to neurological disorders such as erythromelalgia,[56,57] as well as in cardiovascular diseases related to G-protein coupled receptor (GPCR) misfunction.[58,59]

It is not only the loss of functional protein dynamics that can lead to the development of disease. Sometimes proteins deviate from their native conformational landscape and undergo non-functional and non-native structural transitions that lead to protein self-assembly. This is the case in amyloid diseases, where proteins form large oligomeric species that interact with the lipid membrane and disrupt it causing cellular death. Over time, these protein species deposit into insoluble fibrillar aggregates.[60] The most common amyloid diseases are neurodegenerative disorders, such as Alzheimer's or Parkinson's disease. One membrane protein involved in amyloid disease is the Prion protein. Prions are localised on the membrane of neurons via a glycosylphosphatidylinositol anchor. It contains a highly dynamical intrinsically disordered region, which upon misfolding can aggregate into toxic oligomeric species leading to the development of transmissible spongiform encephalopathies.[61]

Membrane protein dynamics have very wide implications, in the context of functional motions and their misbehaviour as well as regarding non-native pathological dynamical transitions. There is need for novel methods that help advance our understanding of these processes and provide the necessary level of detail associated with them. The accurate description of membrane protein conformational dynamics will have important consequences in modern therapeutics, as an initial step in the goal of developing rational drug design strategies that can re-establish the physiological dynamical behaviour of membrane proteins.

1.2 Experimental approaches for the study of membrane protein dynamics

The goal of accurately characterising the energy landscape of proteins is a challenging one, as it goes beyond the classical structural determination of the most energetically favourable native states. Transient protein conformations are generally difficult to detect and describe experimentally, so these studies often need the complementarity of theoretical approaches. Traditionally, experimental structural biology has focused on the determination of static protein structures, a task that is already difficult in itself. In the context of membrane proteins, the field has suffered from considerable difficulties in their experimental characterisation, such as the successful mimicking of their native hydrophobic environment. In fact, membrane protein structures are considerably under-represented in the Protein Data Bank (PDB).[62] At the time of writing, only 4,787 out of 140,591 (3%) entries are annotated as membrane proteins, in contrast with the estimated number of 26% in the human proteome.[63]

Despite this major gap, recent years have seen increasingly more success in overcoming the challenges of membrane protein structural determination. Since the first determination of membrane protein structures,[64,65] many important advances have been made leading to the discovery of important structures. In this context, X-ray crystallography,[66,67] cryo electron microscopy (cryo-EM)[68,69] and solid-state Nuclear magnetic resonance (ss-NMR)[70,71] are becoming well-established and successful techniques for the study of membrane proteins.

However, the general applicability of these techniques remains better tailored to those membrane proteins that feature compact and structured folds. A more challenging area is encountered in the case of dynamical membrane proteins, those structurally heterogeneous macromolecules that perform their function through finely tuned collective motions of their backbone and side-chain atoms. The accurate description of these membrane proteins of dynamical nature still remains elusive to most analytical techniques.

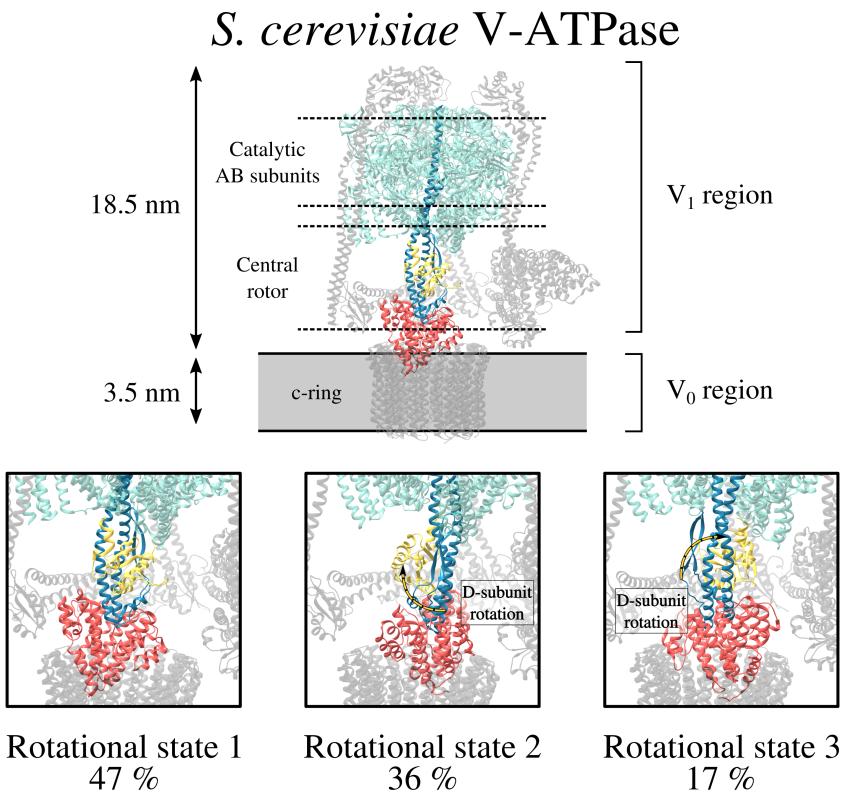


Figure 1.3: Large-scale dynamics resolved by Cryo Electron Microscopy – Zhao and co-workers resolved three different rotational states of the eukaryotic vacuolar H^+-ATPase by means of classification of electron microscopy particles in micrographs.[75] The structures of the individual states were calculated from maps at resolutions of 6.9 Å, 7.6 Å and 8.3 Å. The relative populations of each state can also be quantified (47%, 36% and 17% for states 1, 2 and 3 respectively). The application of these methods is only possible in systems that undergo considerably large conformational changes. In addition, each distinct state within the ensemble must be considerably populated for its structure to be resolved.

Classical experimental structural biology techniques have had some success in addressing the question of large-scale membrane protein dynamics. A common approach is to chemically trap a protein at different conformational states within a long and complex dynamical cycle. Each trapped structure can then be individually resolved, providing a description of the major conformations populated along the pathway. X-ray crystallography has successfully described enzymatic and transport cycles in this way.[53,72] More recently, the fast development of cryo-EM data processing methods has allowed for the description of similar large-scale dynamics, by means of clustering the imaged particles into different groups that are resolved separately to produce different structures.[73,74] Such an analysis was successfully applied by Zhao and co-workers to resolve different rotational conformations of the eukaryotic vacuolar H^+-ATPase (See Figure 1.3).[75]

However, the success of these approaches is limited to dynamical processes in very long timescales, with structurally well-defined and distinct low-energy states that can be trapped along a pathway, or are highly populated within the ensemble. The study of membrane proteins with more heterogeneous structures, which exert their function through order-disorder transitions or faster dynamics, remains intractable to classical experimental techniques, which cannot resolve the atomic detail of their concerted dynamics.

Alternatively, recently developed single-molecule fluorescence methods provide information about protein dynamics and oligomerisation at a much lower resolution. These methods probe protein-protein, interdomain and protein-lipid interactions, but they do so with a very coarse level of detail.[76–78] Historically, the experimental technique that has been most successful in probing protein structural dynamics at high resolution is NMR.

In recent years, ssNMR has emerged as a powerful technique to address the study of membrane protein structure and dynamics. Membrane protein structures are often solved by means of ssNMR, and combined with further ssNMR or solution NMR experiments to extract information about their dynamics. An example of membrane proteins studied by such methods are GPCRs. The dynamics of GPCRs have been studied by solution NMR[79] and ssNMR.[80] More recently, the structure of the GPCR CXCR1 in a lipid bilayer was solved by ssNMR.[81] The dynamics of many other proteins have been studied by ssNMR, including potassium[82] and proton channels,[82] β-barrel proteins[83,84] and viral fusion proteins.[85]

Although ssNMR provides more high-resolution dynamical information than any other experimental technique, this description of protein dynamics is often limited to probing the timescale of movement of different protein regions or the populations between different chemical states. Even if combining the information from different experimental techniques, the ultimate goal of a full structural characterisation of membrane proteins' conformational landscapes is not accessible to experiments only. The inclusion of theoretical approaches appears therefore crucial in reaching a detailed understanding of dynamical membrane proteins.

1.3 Theoretical methods in structural biology

In attempting to conquer the challenges to describe the protein dynamics that are inaccessible to experiments, theoretical methods derived from the realm of statistical mechanics have emerged as powerful tools. These methods rely on a quantitative energetic description of proteins and their surrounding environments, which is generally provided by semi-empirically constructed force fields, built after decades of research and optimisation by repeated iterations of adjustments and comparison with experimental evidence.[86]

Given an accurate description of the energetics is provided by such force fields, several methods have been proposed to achieve the ultimate goal of thoroughly sampling the conformational landscape of proteins. Monte Carlo (MC) based methods aim to reconstruct the ensemble of protein conformations by introducing random perturbations to the system. Each step, a configurational transition is proposed and its acceptance is based on its probability (derived from its calculated potential energy). With sufficient sampling, MC simulations can be used to reconstruct conformational ensembles.[87] There are examples of MC simulations of membrane protein systems,[88–90] however its applicability in that context has remained limited to date. A drastically different sampling method is provided by the energy landscape framework[91] and discrete path sampling.[92] These methods are based on building a potential energy surface from a set of different configurations of the protein. Local minima along this surface can be identified and their transition states sampled, building a complete description of the energy landscape. Although such methods have been proposed in the context of cellular membranes,[93] their applicability has been limited to the study of Amyloid beta oligomers within the membrane.[94]

Amongst all the theoretical approaches for the study of membrane protein dynamics, molecular dynamics (MD) simulation methods have undoubtedly been the most exploited to date. In MD simulations, the temporal evolution of a system's atomic coordinates is studied. At each step of the simulation, the forces acting on each atom are computed according to the potential described by the force field. Newton's equations of motion are then solved to calculate the displacement of atoms in the integration time step (typically

in the order of 1-2 fs), and the process is repeated. MD simulations therefore provide a trajectory that, if sampled for long enough, can be processed to obtain a temporal distribution of a protein's structural configurations, allowing the reconstruction of its conformational landscape.[95] In recent times, the fast advances in computational resources and their decrease in cost have made MD methods more and more accessible to researchers. In performing MD simulations, as the size of the biological system becomes larger, or the necessary sampling timescale becomes longer, the computational power required drastically increases. Under the constraint of current computing capabilities, MD has been successful in characterising many biological processes across a wide range of different timescales.

MD simulations have been employed to study a variety of questions about protein dynamics, from the problem of folding[39,96] to functional conformational transitions[97,98] and protein oligomerisation.[99] In association with lipid membranes, MD simulations have been instrumental in describing the atomic detail of biological processes such as ion conductance. In the case of potassium channels, MD simulations were able to accurately describe their selectivity filter and the *knock-on* mechanism by which they conduct ions,[100–102] depicted in Figure 1.4. Other simulation studies have been able to provide the atomic detail of voltage-sensing and gating of these channels.[103] MD simulations have also been successful in describing the activation mechanism of GPCRs[104] and the process of drug/ligand binding.[105]

Despite these successes of molecular simulations, the limitation of computational power is still a barrier for studying many interesting biological processes that occur on longer timescales. In order to overcome this limitation, different methods have been proposed to enhance the sampling performed by MD, in an attempt to capture rare events that would otherwise not be sampled along a standard MD trajectory. Some enhanced sampling methods take advantage of the increased thermal fluctuations at higher temperatures to overcome large free energy barriers. For instance, in Replica Exchange MD the system is simulated as multiple copies that evolve at different temperatures, periodically exchanging their configurations. The heating events accelerate the jumping of free energy barriers and therefore accelerate the sampling of the overall landscape.[106] In cases where

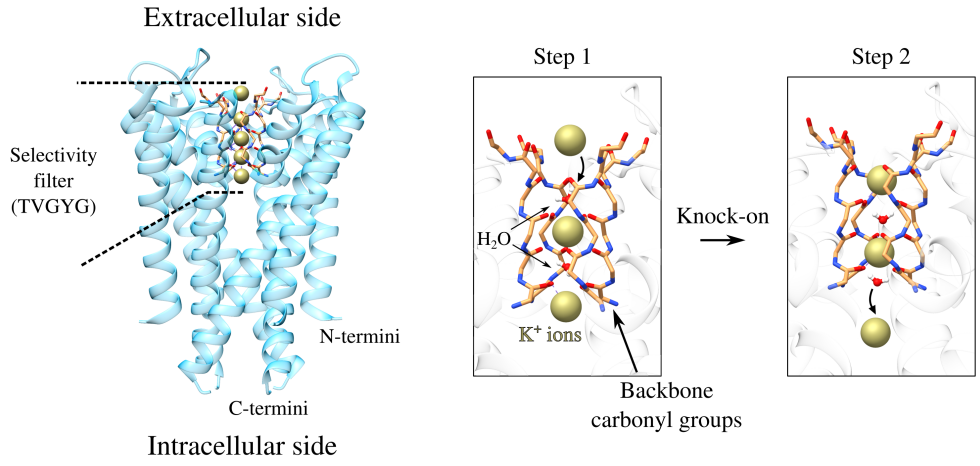


Figure 1.4: Tetrameric KcsA potassium channel conductance mechanism revealed by Molecular Dynamics – The atomic detail of the mechanism of K^+ ion conductance in transmembrane channels has been revealed by several studies employing MD simulations.[100–102] Ion permeation occurs through a Coulombic *knock-on* effect of water molecules and ions that alternate their positions in the ion coordination sites of the protein. This mechanism also ensures ion selectivity by the channels.

two different low-energy states are known (e.g. open and closed channel conformations), there are enhanced sampling methods to find the optimal free energy pathway that connects the two points. One of such methods is transition path sampling, which build an ensemble of intermediate configurations by sampling the possible paths between the two conformations according to their associated probabilities.[107] Finally, some enhanced sampling methods require the definition of one or multiple reaction coordinates, along which one can introduce a biasing force that accelerates the exploration of the energy landscape by increasing the population of high-energy regions. This principle was initially exploited by umbrella sampling methods,[108] and more recently by metadynamics,[109] an approach that has been very widely used to sample rare events in biomolecular simulations.

Enhanced sampling MD methods have been successfully applied to study longer dynamics in membrane proteins. For example, the mechanism of pore gating in pentameric ligand-gated ion channels has been studied in this way,[110] as well as the ion transport by the Na^+/K^+ pump.[111]

Another alternative to overcome the challenge of the computational cost of simulations is to decrease the level of detail in the models. This can be done by replacing the solvent

or the membrane system by implicit potentials,[112,113] which generally decreases the accuracy of the model.[114] Alternatively, atoms that are covalently bonded can be grouped into larger spheric particles and their interactions reparametrised. This approach is know as coarse graining.[115] Coarse grained simulations have established themselves as a very useful tool in the study of membrane proteins, with many recent improvements and constant development of lipid force fields to accurately mimic the membrane environment.[116] The most widely used coarse grained model in membrane protein MD is MARTINI.[117] Its lower level description of protein and lipids allows its employment in much more complex systems, ranging from large membranes of varied lipid compositions[118] to full viral capsids.[119,120] MARTINI models have been applied to explore protein-protein interactions of GPCRs,[121,122] protein-lipid interactions in the bacterial outer membrane[123] and the association of peripheral membrane proteins.[124] In more recent study, MARTINI models were combined with enhanced sampling by metadynamics in order to characterise the dimerisation mechanism of the epidermal growth factor receptor.[125]

The success and applicability of MD simulations to biological problems is unquestionable. However, the technique still suffers from one major drawback that comes from the imperfect nature of force fields. The energy potentials are semi-empirically constructed and subject to constant revision and improvement as more data becomes available to the modelling community. As more experimental and simulation studies have become available, many groups have reported on the comparison between the two in order to assess the accuracy of force fields. Many reviews highlight the recent improvements in force field accuracy and their ability to reproduce experimental data, often coming from NMR,[126,127] but also from other sources such as vibrational spectroscopy[128] and Φ-value analysis.[129] However, most studies that highlight the overall good agreement with experiments still warn against making quantitative predictions from MD.[130,131]

The main accuracy issue reported in the literature is force field discrepancy. Sampling a system under different state-of-the-art force fields will generally lead to different ensembles, with the impossibility to discern which one is closer to the native one.[132–134] A frequent source of discrepancy in force fields is the over/under-stabilisation of certain sec-

ondary structural elements over others.[135,136] Other studies have reported MD simulations that fail to reproduce correct protein folding pathways by overstabilising misfolded states that are inexistent in nature.[137] Other systematic studies describe that although MD generally describes the dynamics of the native states of small proteins with a high degree of accuracy, more problems are encountered in studying larger systems with more flexible regions.[138,139]

An especially challenging area of modelling for MD is that of intrinsic protein disorder. Historically, the empirical parametrisation of force fields has been heavily weighted on the known dynamical aspects of compact, globular proteins, resulting in energy functions that are biased toward secondary and tertiary structural features.[140,141] The most widely-used force fields are therefore generally unable to accurately describe intrinsic disorder or functional order-disorder transitions that often occur in native protein ensembles. Several studies point at incorrect water-protein interactions as a source of error in disordered proteins MD simulations.[142,143]

When simulating membrane proteins, the lipid molecules that form the hydrophobic bilayer must also be modelled. Many common lipid molecules have now been studied by MD, and again several different force fields exist. Most systematic validations of lipid force fields conclude that large-scale properties of the system (e.g. area per lipid, overall membrane scattering density profiles) are generally well reproduced.[144,145] However, atomic-level details are not generally well described in comparison with NMR dipolar couplings or deuterium quadrupolar splittings.[146,147] Finally, as was the case with proteins, one of the most significant concerns with lipid MD accuracy is force field discrepancy.[148,149] In the case of coarse grained models, many more large-scale properties of the membranes have been tested and are generally very well reproduced, under many different lipid compositions.[150–152]

Overall, theoretical approaches are highly useful to reach the atomic detail description of transient protein dynamics that experiments cannot access. However, their imperfection and limitations in accuracy mean that, whenever possible, theoretical studies will always

benefit from the inclusion of experiments to improve the description of any system of interest.

1.4 Combining experiments and theory to characterise protein dynamics

With the objective of overcoming the shortcomings of both the experimental and theoretical approaches, there is a new prominent view that the combination of experimental data and molecular simulation provides a powerful analytical approach to elucidate protein dynamics. By integrating experiments and theory, researchers aim to get the best of both worlds: an accurate characterisation from experiments and the high-level detailed description of dynamics from simulations. Furthermore, molecular simulations can be instrumental in the interpretation of certain experimental data which require a dynamical consideration of the system. This is especially true in the case of experimental data that are the result of time and ensemble averaged quantities over many different conformational states in the experimental sample.[153,154]

The problem of reconciling theory and observation in the realm of statistical mechanics was formalised by Jaynes by applying concepts from information theory.[155] Jaynes poses the problem of constructing the best possible probability distribution if only a number of experimental observations from that distribution are available. By calculating the informational entropy of the system (also known as Shannon's entropy[156]), he put forward the maximum entropy principle, which results in a posterior distribution that is compatible with the data but minimises the perturbation of the theoretical distribution, thus remaining as faithful as possible to the original model.

The maximum entropy principle has been applied in several forms to the study of protein structural ensembles.[157,158] A common approach starts by performing MD simulations of a system, which yields an initial ensemble of conformations. If sampled adequately, the

ensemble can be reconciled with experimental data *a posteriori*, by modifying the statistical weight of each conformation (and therefore the energy associated with it in the landscape). This is known as the reweighting approach,[159,160] and it has been successfully applied using data from a variety of sources, including NMR,[161,162] small-angle X-ray scattering,[163,164] infrared spectroscopy[165] and electron paramagnetic resonance.[166] Alternative reweighting approaches have also been proposed based on Bayesian methods[167,168] or the maximum parsimony principle.[169–171] Reweighting methods have the advantage that they can be applied very fast and repeatedly as new data becomes available, without the need to run repeated MD simulations that are computationally very expensive. However, they heavily rely on the simulation sampling all relevant states of the system. There may be certain conformations that contribute to the experimental observables but remain unexplored throughout the simulation, and thus cannot be reweighted into the final ensemble. In fact, the accuracy of these methods is dependent on the overlap between the prior and posterior distributions,[172] so they remain less applicable in cases where the original ensemble has to be considerably altered.

As an alternative to reweighting approaches, one can include explicit restraints to drive the sampling of a simulation toward the distribution that is in agreement with the experimental data. When dealing with time and ensemble-averaged data, the most common way to do this is to run a number of simultaneously evolving independent simulations (known as *replicas*) that only share information about their properties relating to the experimental data. The observable is back-calculated as an average of all the replicas and restrained toward the experimental value. This approach is known as replica-averaged restrained MD, and it has been formally shown to fulfil the maximum entropy principle.[173–175]

Replica-averaged restrained MD has been mostly exploited and validated in combination with NMR data.[33,176] A notable early application of the method using residual dipolar couplings (RDCs) was able to characterise the molecular recognition mechanism of ubiquitin.[177] Many efforts have been employed in characterising disordered and unfolded states of proteins in this way.[178–183] NMR-restrained simulations have since been used to study the conformational equilibria of a wide variety of soluble proteins,[184–187] folding interme-

diates,[188] partially misfolded structures on the pathway of protein aggregation,[189,190] allosteric motions in signalling proteins,[191–194] enzyme catalysis[195] and molecular recognition between proteins[196,197] and protein-nucleotide.[198,199]

The wide applicability of replica-averaged MD is an encouraging demonstration of the potential of integrating theory and experiments in building a comprehensible picture of protein structural dynamics. This picture is of course far from complete, as the approach suffers from similar limitations as traditional simulations, derived from their intensive computational nature, as well as from the usual experimental challenges in obtaining reliable data in complex systems. Nonetheless, these examples of success advocate that the integration between diverse sources of information is only beneficial to achieving the final objective of accurately characterising protein dynamics.

1.5 Characterising membrane protein conformational landscapes

Integrative approaches that combine experiments and theory to describe protein dynamics in detail have had plenty of success. However, their applicability to membrane protein systems is still at its infancy. Partly due to the experimental and theoretical challenges associated with the study of membrane proteins, examples of combining both sources of information are very scarce. Specifically in the context of replica-averaged MD, the primary source of experimental data comes from ssNMR, a setup in which protein can be studied in a close-to-native lipid environment. The first attempt at including ssNMR restraints into MD simulations made use of an implicit membrane system,[200] in the study of the orientation of a single-helix transmembrane domain of viral protein “u”.

A method for ssNMR-restrained MD in explicit bilayers was later developed and applied to study the conformational equilibrium of SERCA regulatory proteins sarcolipin and phospholamban.[201] The main structural transition described in that work is the order-

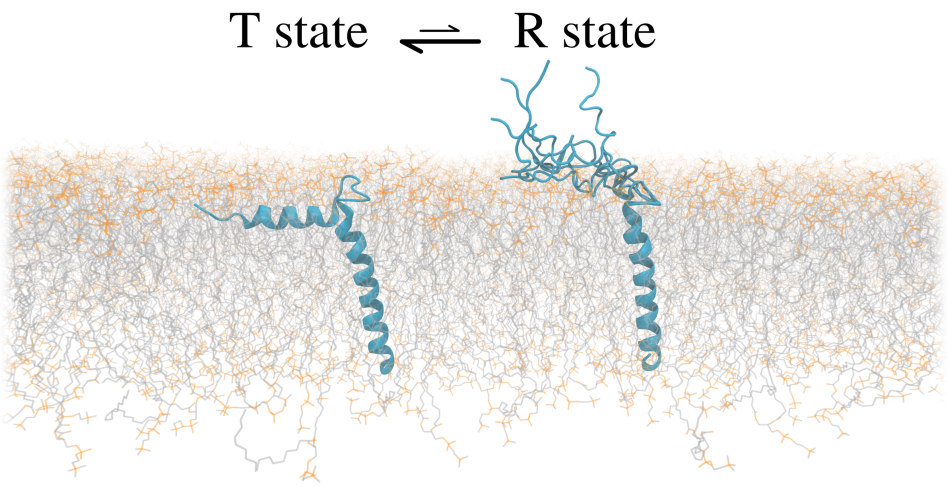


Figure 1.5: Conformational equilibrium of phospholamban revealed by ssNMR-restrained replica-averaged MD simulations – The dynamic equilibrium between the ground and excited states of PLN (T and R respectively) has been accurately described by the combination of experimental data from solid-state NMR and molecular simulations.[201] This study showcases the potential of including experimental information into molecular simulations, which are generally unable to accurately describe order-disorder transitions such as this one.

disorder equilibrium in the N-terminus of PLN. In its ground state (T-state), the N-terminal segment is adsorbed on the membrane as an amphipathic helix. In the excited R-state, the N-terminus detaches from the membrane and adopts a dynamical random coil conformation (See Figure 1.5). The excited state had already been observed spectroscopically[202] and studied by replica-averaged MD in micelles, with experimental RDCs from solution NMR.[203] The ssNMR-restrained simulations in bicelles are however carried out in an environment that closely mimics native conditions, and therefore provide a highly accurate description of the relative populations of the two states and their structural features.

Restrained MD simulations of membrane proteins have since been applied in only one other occasion, a study using ssNMR data regarding the dynamics of pVIII coat protein from fd phage.[204] The full potential of replica-averaged MD in characterising membrane protein dynamics is yet to be exploited, and the field of structural biology can greatly benefit from a systematic approach that can tackle these difficult questions with a high level of detail and accuracy.

1.6 Summary and structure of the thesis

The work I present in this thesis builds precisely on the framework of lipid membrane-based ssNMR-restrained MD.[201] This restraining approach is here developed further and applied to a variety of biological systems. The thesis therefore outlines methodological advances and practical applications of such methods, so I have structured it accordingly in an attempt to maintain a clear flow throughout.

Chapter 2 is a methodological chapter, where I describe two different sources of experimental NMR data and the progress made in building accurate models for them as well as their implementation into molecular simulations. Firstly, chemical shifts are introduced and the challenges in building accurate models for them are explained. I will present results from an accurate model of chemical shifts based exclusively in protein sequence, an encouraging first step in the ultimate goal of generating and applying accurate structural models for chemical shifts in the study of protein dynamics. Secondly, I will introduce two types of data associated with oriented solid-state NMR measurements, namely chemical shift anisotropy and dipolar coupling. Methods for the incorporation of these data into molecular dynamics simulations are outlined and I will present advancements in their capabilities as well as an extensive methodological validation of the approach.

The following three chapters focus on practical applications of the sampling methodology to characterise membrane protein dynamics. Chapter 3 describes the dynamics of phospholamban, a transmembrane protein with an important regulatory role in cardiac muscle relaxation. This chapter first describes the conformational transitions of phospholamban in its homo-pentameric state, and then moves onto the complex that phospholamban establishes with the cardiac pump SERCA.

Chapter 4 describes the application of the restrained MD sampling approach to another transmembrane protein, the homo-tetrameric proton channel M2 from influenza. Dynamical ensembles for the channel are generated and analysed, outlining important dynamical transitions that govern its mechanism of action.

The final practical application of the method, corresponding to chapter 5, focuses on the dynamics of CXCR1, a G-protein coupled receptor that has a crucial function in the inflammatory response and is linked to multiple forms of cancer. The experimentally-guided structural ensembles of this receptor provide some insights into the allosteric motions necessary for its function.

Finally, chapter 6 provides a discussion of all the results presented in the thesis, aiming to provide a wider perspective of their significance in the context of combining theoretical and experimental approaches to accurately describe membrane protein dynamics.

Chapter 2

Development and testing of NMR restraints

2.1 Chemical shifts

2.1.1 Physical principles and challenges in their interpretation

Chemical shifts (CS, often also noted as δ) are the most readily obtainable data by NMR both in solution and solid state. In proteins, CS are very precise probes of local structure and dynamics. However, their interpretation is highly challenging as they arise from a multitude of complex interactions and factors that are difficult to model. For any particular nuclear spin system in a protein, its CS is defined as its resonant frequency in an externally applied magnetic field (commonly noted as B_0). This frequency (measured in Hz) is directly proportional to the difference in energy between the two energy levels of the nuclear spin. As B_0 becomes larger (in more powerful NMR magnets), so does the resonant frequency of a nucleus. In order to compare CS values measured at different B_0 field strength, it is desirable to remove this dependency of the CS on the field. This can be done by quoting CS as a relative quantity in comparison to a measured standard (e.g. tetramethylsilane in ^{13}C and ^{1}H protein NMR, or liquid NH_3 in ^{15}N protein NMR), ex-

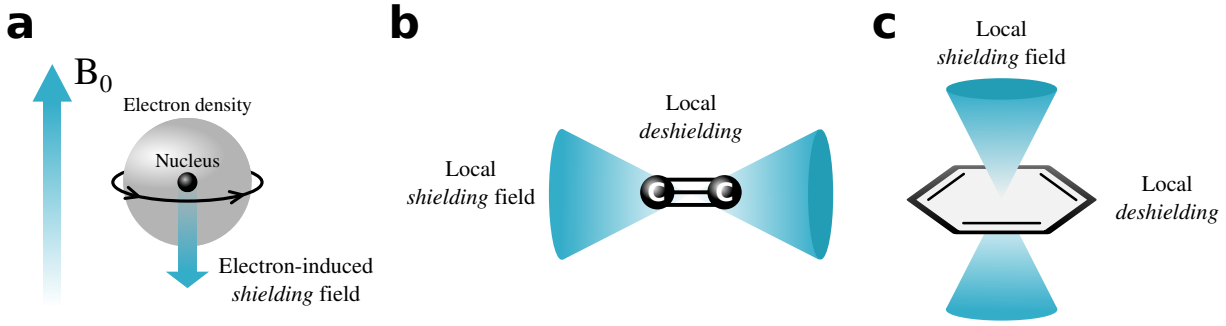


Figure 2.1: Electronic shielding phenomena that influence chemical shifts – Some of the interactions affecting the chemical shift are here depicted. (a) Nuclear shielding by the outer electron shell. The electronic environment around the nucleus generate an opposing magnetic field to the spectrometer field B_0, lowering the chemical shift of the nuclear species. (b,c) Local anisotropic shielding fields in different chemical groups. The geometry of chemical bonds limits the range of movement of electrons, creating non-uniform shielding fields depicted as cones. If a proton in close proximity finds itself in the space inside a cone, it will be shielded and experience a lower chemical shift. If by contrast it is in the space outside of the cone, it will experience a local deshielding and have a higher chemical shift.

pressed in parts per million (ppm).

The differences in resonant frequencies between different nuclei arise from magnetic interactions in the local environment of the nucleus in response to the external magnetic field. The primary factor determining the CS comes from the electronic shell surrounding the nucleus. The electrons around a nucleus have an electric charge, which (according to Lenz's law) responds to the external magnetic field (B_0) by generating an induced magnetic field in the opposite direction of B_0. This field is generally referred to as a *shielding* field, as it reduces the effective magnetic field experienced by the nucleus (Figure 2.1a). Since the shielding effect reduces the magnitude of the effective field experienced by the nucleus, the energy levels of the spins get closer to each other and the resonant frequency decreases. Therefore, shielding will decrease the observed CS of a nucleus.

Another factor affecting CS is the magnetic anisotropy of chemical groups. Electrons in chemical groups generate shielding fields according to the same physical principle as electrons in the atomic shell. However, bonded electrons cannot move freely, and therefore the shielding field they generate is not uniform, or *anisotropic*. These anisotropic shielding fields are generally represented as cones (see Figure 2.1b,c). If a nucleus is situated within a cone, it will be shielded and its CS decreases. In contrast, if it is located out-

side, deshielding takes place and the CS increases. For example, a nucleus located in orthogonal proximity to an aromatic ring will experience a strong shielding field from the π-orbitals in the ring, considerably decreasing its CS. This is often the case for methyl protons at the core of folded globular proteins, whose CS shifts to regions below zero ppm when near an aromatic ring. This phenomenon is commonly used as a footprint for detecting protein folding in proton NMR spectra.

Other molecular interactions can affect the chemical shift, such as hydrogen bonding. Protons involved in hydrogen bonds are generally shifted to considerably higher frequencies. Similarly, the presence of strong electronegative groups in the vicinity of a nucleus will drag electrons away causing deshielding, and increasing the CS. Temperature and solvent effects also influence the strength and prevalence of these interactions, and therefore have an important role in determining the CS.

Finally, the consideration of dynamics is again crucial. In NMR experiments, the observed CS is averaged over all chemical states visited during the recording of each scan. The protein dynamics in the millisecond timescale therefore influence the final value of the CS, whether it comes from local fluctuations in the electronic environment or more pronounced conformational transitions.

Given how all these factors can subtly influence the CS, it is evident how these data can serve as very sensitive probes of local structure and dynamics in proteins. In addition, CS can be measured in a highly accurate and reproducible manner by all NMR methods. However, after describing the complexity of these interactions, the challenge of accurately modelling CS from protein structure also becomes evident. There have been multiple attempts and some progress at building CS models that emanate from the physical phenomena of chemical shielding.[205–211] Such models require a full quantum-mechanical description of the protein, and calculating all the interactions that give rise to the CS is computationally very expensive. In addition, these models are still unable to capture the full complexity of the physical phenomena and the accuracy of the predicted CS is generally lower in comparison to empirical methods.[212] However, since these methods are built from

first principles, they should be applicable in any context. It is worth noting their value in modelling systems for which not much similarity is found in CS databases, such as protein complexes with nucleotides, carbohydrates or small molecules.[212] However, in the general case of protein CS modelling, empirical methods based on statistical tools have found more success and applicability. In the following subsection I will provide a more detailed overview of this subset of methods and evaluate their current capabilities.

2.1.2 An overview of chemical shift modelling in proteins

The dependence of chemical shifts on protein structure

Despite the difficulties in accurately modelling CS in complex structures such as proteins, correlations between CS and certain chemical or structural features have been described for many years. Each atom in an amino-acid has a *typical* CS based on the average observations across all publicly available NMR data, which is now compiled in the Biological Magnetic Resonance Data Bank (BMRB).[213] A crucial factor in accounting for deviations from these typical CS is secondary structure. This phenomenon has been described in a variety of atoms, where pronounced variations were observed in one direction for α-helical structures and in the opposite direction for β-sheet segments.[214,215] For example, in the case of $^{1}H_{\alpha}$, one of the most sensitive probes for secondary structure in NMR, a lower chemical shift is observed in α-helices, whereas the opposite is true for β-sheets.[216] In disordered, random coil conformations, the CS is found to be generally close to its typical value. The characteristic shift for an atom in a particular amino acid is therefore referred to as its random coil chemical shift. In fact, disordered proteins are characterised by a low degree of variability (or dispersion) in their NMR spectra.[217]

The realisation of the strong influence of secondary structure in the value of the CS led to the definition of metrics such as the *secondary chemical shift*, which is obtained by subtracting the random coil CS to the value of an observed CS.[218–220] An approach built on the premise of secondary CS is the chemical shift index (CSI),[221,222] still commonly

used today to study the secondary structure of proteins at a residue-specific level. A later method named TALOS[223,224] provided a new level in this description by correlating CS to the dihedral angles of the protein backbone (which are determinant of secondary structure elements).

Given the success of these approaches in extracting structural information from CS, there was great interest in solving the reverse problem: predicting CS starting from a protein's three dimensional structure. In this respect, the most successful models have applied empirical and statistical methods. These models process the protein structural information into inputs that are relevant for the prediction (e.g. residue identity, dihedral angles, geometric features near the nucleus) and optimize a set of statistical weights to fit a database of high-confidence CS. Different programs have been proposed which apply this idea, for example by using statistical hypersurfaces (SHIFTX[225]), artificial neural networks (PROSHIFT,[226] SPARTA[227]) or a combination of linear equations (Camshift[228]).

In characterising protein dynamics, the main scope of CS models is to provide structural restraints for their use in molecular simulations. In this context, it is worth a particular mention to the Camshift approach. Inspired by the geometrical phenomena that give rise to the CS, Camshift employs a set of weighted linear equations that account for torsion angles and the distances to nearby atoms. This simplified description provides an easily differentiable expression for the CS that can be readily used as an energy potential in molecular simulations. There are already many examples of its successful application in soluble proteins.[181–184] The accuracy of Camshift is however generally poorer than its more elaborate, updated counterparts SPARTA+[229] and SHIFTX2,[230] since the simpler linear regression terms are unable to capture the full complexity of the CS-structure dependence. The more complex models have not been applied in NMR-restrained MD simulations, due to the difficulty (or in some cases impossibility) in differentiating the CS terms, a factor that hampers their use as energy potentials.

There are however no restrictions to differentiating and building energy potentials from more complex non-linear expressions such as those arising from machine-learning methods

(e.g. parametrised neural networks). Using such approaches, I will show in the following sections that CS can be modelled accurately based exclusively on protein sequence information, in a step toward the ultimate goal of building highly reliable and applicable CS structural models. Given how commonly CS are measured in ssNMR experiments, the study of membrane protein dynamics could greatly benefit from accurate, differentiable CS potentials.

Modelling protein chemical shifts exclusively from primary sequence information

The impact of accurate CS models goes beyond their use in molecular simulations of protein dynamics. Having accurate descriptions of CS can be extremely useful in structural calculations and refinement based on CS data,[231–233] CS re-referencing,[234] or assisting in the process of NMR spectral assignment.[235–237] In many cases the structure of the protein is not available, and it is therefore not possible to apply the structure-based models previously outlined. Understanding the relationship between the information contained in the protein sequence and the variation of CS is an important step in building tools that are applicable when the structural information is missing.

The relationship between sequence and CS has been especially exploited in studies regarding protein disorder. Since IDPs lack a defined tertiary structure, NMR is the only applicable technique that can yield information about their transient interactions and conformational preferences, with examples both *in vitro*[238,239] and in the cellular environment.[240,241] In IDPs, their sequence is the only available source of information in modelling CS. Deviations from the random coil CS are generally accounted for by applying empirical corrections based on the neighbouring amino acids.[217,242–245]

When it comes to folded, globular proteins, the sequence-CS relationship has not been exploited as deeply. There are methods that rely on querying sequence databases to identify homology to proteins with known, deposited CS data.[246,247] However, these approaches rely on the presence of a closely related protein variant having been studied by NMR. The

relatively low number of proteins with published NMR data heavily limits the applicability of these methods. There is room for a generalised approach that accurately correlates CS directly to protein sequences, without requiring a three-dimensional structure or homology to previously studied systems.

Here I present a novel approach to the interpretation of protein CS exclusively from primary amino acid sequence, both in ordered and disordered proteins. The method was developed from the analysis of experimental CS applying statistical tools that improve the state of the art of the sequence-based IDP prediction and in the case of structured proteins achieves the accuracy of methods that require the three-dimensional structure of the proteins. In addition to providing a unified tool for CS prediction in ordered and disordered proteins, the development of the method has revealed key insights into the structural information contained in chemical shifts.[248] Following a long-standing tradition in modern science of naming computational methods with memorable acronyms, the model was baptised *protein sequences and chemical shift correlations* (PROSECCO). The success of PROSECCO in accurately describing CS exclusively from sequence is a promising step in the direction of developing more accurate CS models based on protein structure that can be reliably used in combination with molecular simulations.

2.2 PROSECCO: PROtein SEquences and Chemical shift COrrelations

2.2.1 Curating a chemical shift database for a new predictor

In gathering the necessary data to parametrise the predictor, the first step involved collecting the set of available chemical shift assignments from the BMRB.[213] The initial set was of 9,514 entries of protein assignment, which were later refined through several curating filters. We remove redundant sequences to avoid the over-representation of heavily studied protein families. Using the USEARCH algorithm,[249] we allowed a maximum se-

quence identity of 90%, filtering out entries that had close homologues already present in the dataset. This step lowered the number of entries to 5,140, but it ensures that PROSECCO will not be biased or reliant on protein homology information to model CS . The next data filtering step involved the removal of entries that were incorrectly referenced. This was done entry by entry, assessing the difference between the chemical shifts of each atom type and the overall distribution in the BMRB corresponding to that specific type of atom and amino acid in its associated secondary structure element. Entries that had an overall deviation in their CS distribution greater than a $\delta^{\text{tolerance}}$ were subsequently removed. The $\delta^{\text{tolerance}}$ values were set to 1.5, 0.5 and 3.5 ppm either side from the center of the BMRB CS distributions for ^{13}C, ^{1}H and ^{15}N respectively. This step removed a further 147 BMRB entries. Finally, within the remaining well-referenced entries, specific CS values were removed if deviating more than 3.5 standard deviations from their expected value (center of their corresponding distribution), to remove mistyped or incorrectly measured CS values. This final step discarded 0.51% values of the data.

The final dataset of folded proteins consisted of 4,993 BMRB entries, which included 3,953,878 experimental CS. Only 2,943 (59%) out of the final entries had an associated three-dimensional structure deposited in the protein data bank (PDB),[62] a sign of the importance of developing a predictor that does not rely on structure. In order to include secondary structure information, we assigned the secondary structure of each entry based on the Q3 notation (helix, strand and coil) using the δ2D method,[250] which employs the chemical shift information to accurately determine the secondary structure.

In order to benchmark the predictor, additional entries were curated. The final benchmark consisted in 77 BMRB entries that were all deposited after 2016, therefore ensuring that these data were not involved in the parametrization of any of the methods that were compared in this study.

In the case of disordered proteins, a different database was constructed. BMRB entries were cross-referenced with Disprot[251] or specific annotations in the literature regarding protein disorder. 44 non-redundant BMRB entries were found to be intrinsically disor-

dered or to contain an intrinsically disordered region (IDR) of at least 20 residues. A total of 20,154 chemical shift data were contained in this database. Given the smaller size of the IDP dataset, rather than dividing it into separate training and validation sets, a leave-one-out approach was used for validation in all the benchmarks presented, as done in other studies.[252]

2.2.2 PROSECCO$_{\mathbf{IDP}}$: Improving the prediction of chemical shifts in disordered proteins

In order to generate a sequence-based CS predictor for IDPs, a database known to contain only disordered, random-coil conformations in proteins was used. The method, PROSECCO$_{\mathrm{IDP}}$, is based on the generation of probability density functions for the CS of each atom of interest by using Gaussian kernel functions:

$$\hat{d}_i^A(\delta) = \frac{1}{n_i^A} \sum_{l=1}^{n_i^A} G_K(\delta - \delta_l) \tag{2.1}$$

Specifically, a probability density function $\hat{d}_i^A$ is generated for each atom of type A of an amino acid of type i. This is done by summing n_i^A Gaussian kernels , G_K, each centered at the CS value of the experimental observation in the entry (See Figure 2.2). The expression of each Gaussian kernel function is

$$G_K(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \tag{2.2}$$

where σ is the kernel bandwidth that determines the aperture of the kernel. This parameter was set according to Silverman's rule,[253] taking values of 0.05, 0.2 and 0.5 for ^{1}H, ^{13}C and ^{15}N respectively.

From these probability distributions an expectation value, δ_i^A is obtained. In cases with less than 400 observations, δ_i^A is chosen as CS value that corresponds to the maximum of

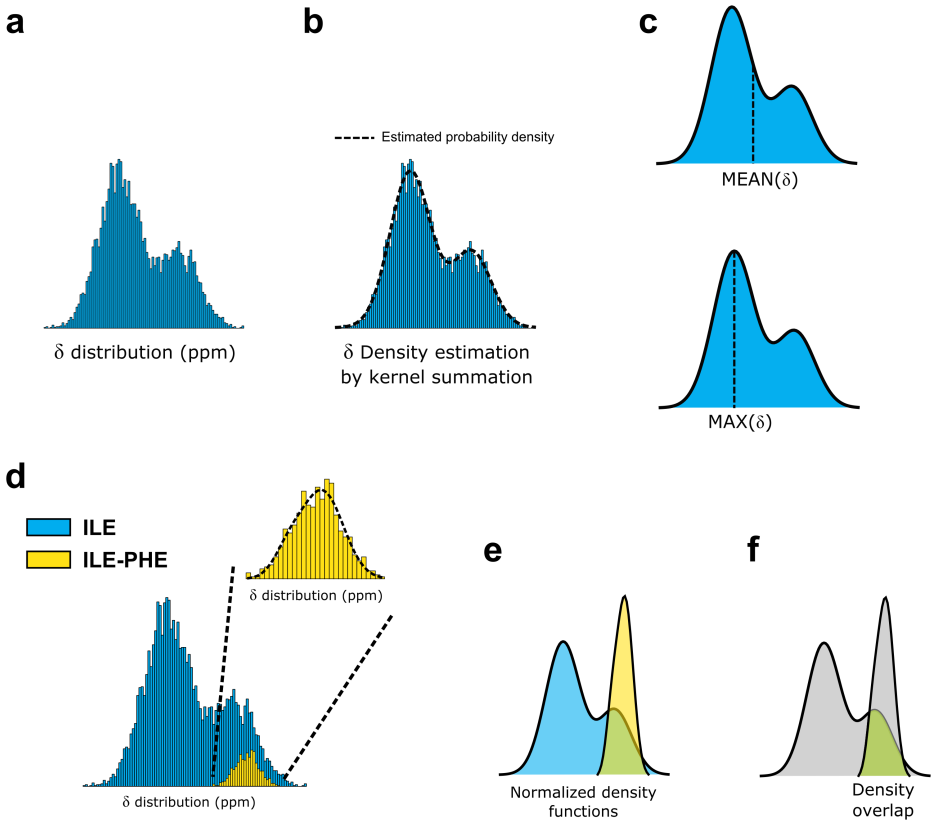


Figure 2.2: Method of kernel density estimation employed in PROSECCO$_{\text{IDP}}$ – (a) For a particular spin system, an illustrative distribution of experimental CS deposited in the BMRB is depicted. (b) The distribution can be transformed into a probability density function by summing a series of Gaussian kernels, centered at each experimental value available. The resulting sum provides the density estimation $\hat{d}_i^A(\delta)$. (c) Given a probability distribution, an expectation for the CS value is obtained by choosing the value associated with the mean of the density function or the CS corresponding to the maximum of the density function (if less than 400 observations are available). (d) Neighbour corrections for the value of the CS are applied based on the local protein sequence. For a specific atom in a particular amino acid (e.g. ILE, blue distribution), we identify the subset of CS associated with the specific residue pair present in the input sequence (e.g. ILE-PHE, in yellow). Another probability density function, $\hat{d}_{i,j}^A(\delta)$, is built from this secondary distribution (see Equation 2.5). (e) Both distributions are normalised and their expectation values used to calculate the neighbour correction term (Equation 2.6). (f) The overlap between the two distributions is used to compute the weight of the correction term (Equation 2.7). Adapted from Sanz-Hernández & De Simone, *J Bio NMR* (2017)[248] (licensed under CC BY 4.0).

the distribution:

$$\delta_i^A = \arg\max_{\delta}(\hat{d}_i^A(\delta)) \tag{2.3}$$

Otherwise, δ_i^A is evaluated as the mean CS value of the probability distribution (see Figure 2.2c):

$$\delta_i^A = \int_{-\infty}^{\infty} \delta\, \hat{d}_i^A(\delta)\, d\delta \tag{2.4}$$

This primary estimation of the CS is further refined using pair-wise correction terms based on the sequence of neighbour residues. If we consider i as the central residue, a correction can be applied based on the neighbour residue j. The optimal employment of nearest-neighbour terms involves a window of five residues, centred at the residue i of interest. Each correction term is evaluated via pair-wise density functions calculated using sub-datasets of the experimental database that reflect the specific pattern of amino acids i and j as in the input sequence (See Figure 2.2d,e,f).

$$\hat{d}_{i,j}^A(\delta) = \frac{1}{n_{i,j}^A} \sum_{l=1}^{n_{i,j}^A} G_K(\delta - \delta_l) \tag{2.5}$$

The expectation values of the pair-wise residue distributions, $\delta_{i,j}^A$, are similarly evaluated and used in the calculation of a neighbour correction term:

$$\Delta\delta_{i,j}^A = \delta_{i,j}^A - \delta_i^A \tag{2.6}$$

If a specific pair of residues has very limited occurrences in the dataset (set as ¡ 20 observations), the nearest-neighbour correction term $\Delta\delta_{i,j}^A$ is calculated as an average, by considering the effect that amino-acids of type j exert on all possible types of neighbour residues at the position i. For each neighbour correction term, a weight was constructed

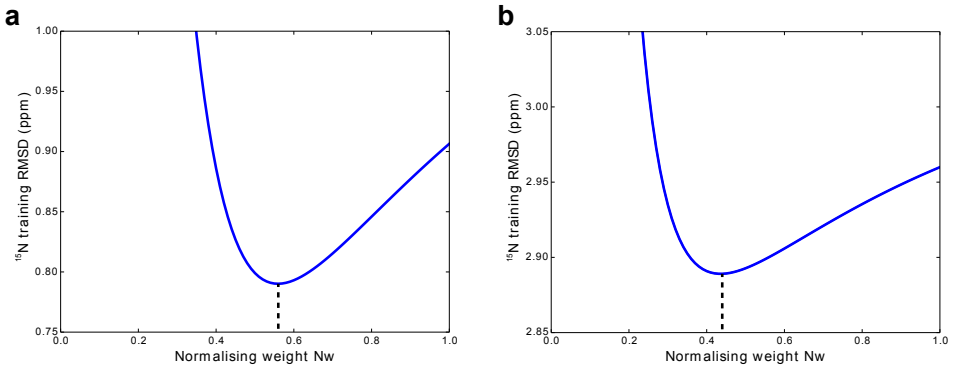


Figure 2.3: Calibration of the weight normalising factor N_w – A normalisation factor is introduced to adjust the overall correction between the primary CS estimation and the local sequence correction terms (Equation 2.8). The calibration was performed by minimising the RMSD between calculated and experimental CS values in the traning sets. In all cases a single minimum was found in the calibration curves. The examples shown correspond to the curves for the amide ^{15}N atom in PROSECCO$_{IDP}$ (a) and the folded counterpart (b). Adapted from Sanz-Hernández & De Simone, *J Bio NMR* (2017)[248] (licensed under CC BY 4.0).

using the negative overlap between the primary and the pair-wise density functions:

$$w = \left(1 - \int_{-\infty}^{\infty} \min\left(\hat{d}_i^A(\delta), \hat{d}_{i,j}^A(\delta)\right) d\delta\right) \tag{2.7}$$

A normalising factor for the weights, N_w, is introduced to complete the final expression. These factors were calibrated individually for each atom by minimizing the error between calculated and experimental CS (See Figure 2.3).

Given all these terms, a final CS value can be calculated by PROSECCO$_{IDP}$ for an atom A of a residue i in a local sequence *k-j-i-l-m* according to the full equation:

$$CS_i^A = \delta_i^A + \frac{1}{N_w}(w_{i,k-2}^A \Delta\delta_{i,k-2}^A + w_{i,j-1}^A \Delta\delta_{i,j-1}^A + w_{i,l+1}^A \Delta\delta_{i,l+1}^A + w_{i,m+2}^A \Delta\delta_{i,m+2}^A) \tag{2.8}$$

The accuracy of this method was compared against the two most readily used approaches for calculating CS in disordered proteins,[244,245] by benchmarking on all proteins in our dataset using a leave-one-out approach (Figure 2.4). The benchmark was performed by computing the root mean squared deviation (RMSD) between predicted and experimen-

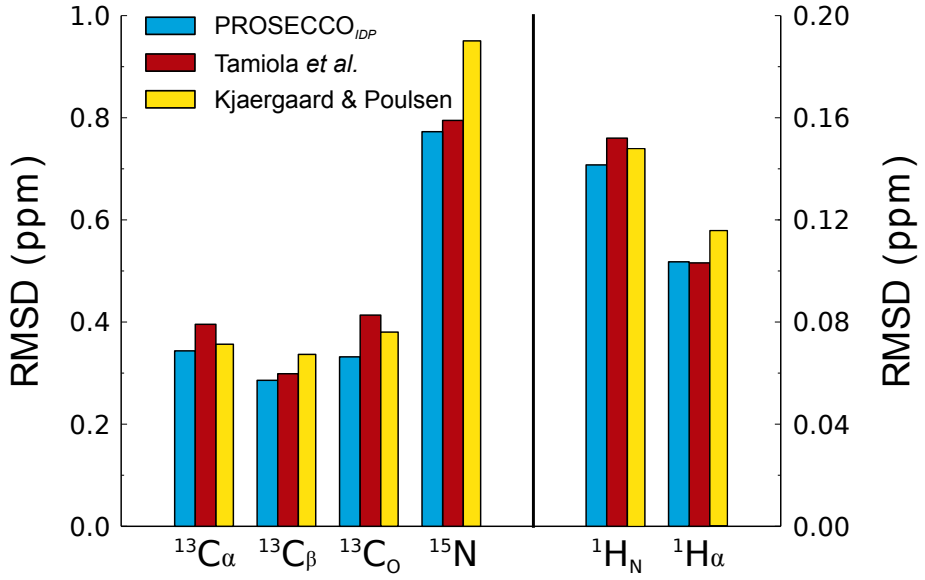


Figure 2.4: Benchmark of sequence-based prediction of CS in IDPs – The agreements between experimental and calculated CS are reported in terms of RMSD. The performance of PROSECCO$_{\text{IDP}}$ (cyan) is compared to the methods by Tamiola et al.[244] (red) and Kjaergaard and Poulsen[245] (yellow). A *leave-one-out* approach was employed in the case of PROSECCO$_{\text{IDP}}$, whereby the method is re-parametrised every time it evaluates a different BMRB entry, by leaving it out of the training dataset to avoid bias. Adapted from Sanz-Hernández & De Simone, *J Bio NMR* (2017)[248] (licensed under CC BY 4.0).

tal CS:

$$RMSD = \sqrt{\frac{1}{n}\sum_{l=1}^{n}\left(\delta_{l,exp} - \delta_{l,calc}\right)^2} \tag{2.9}$$

Overall, PROSECCO$_{\text{IDP}}$ improves the sequence-based models of CS in disordered proteins. The CS of backbone carbon atoms are reproduced with remarkable accuracy, an important achievement considering the high amount of information they contain to identify transient secondary structure formation in disordered proteins.[254,255]

2.2.3 Predicting chemical shifts from sequence in secondary structure elements

The Gaussian kernel-based approach of PROSECCO$_{\text{IDP}}$ was extended to individual secondary structure elements in folded proteins. The probability density functions and neigh-

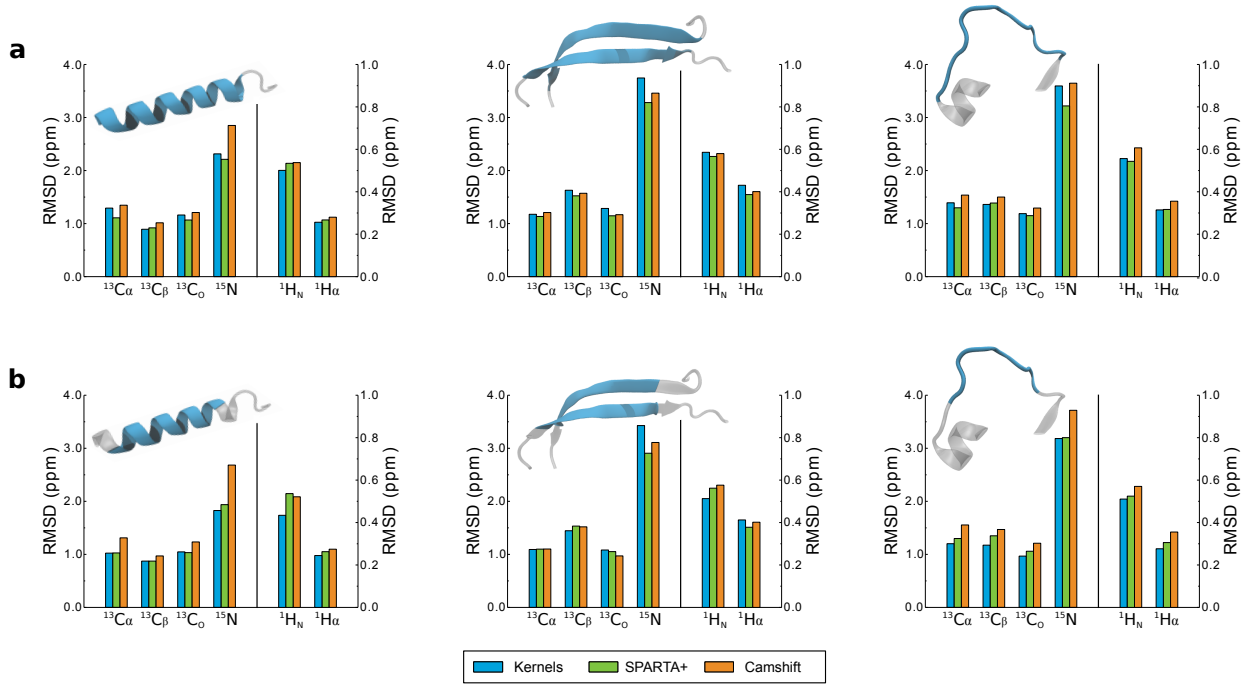


Figure 2.5: Application of kernel-based CS prediction to secondary structure elements – The Gaussian kernel-based predictor employed by $\text{PROSECCO}_{\text{IDP}}$ was separately parametrised for each type of Q3 elements present in proteins (helices, strands and loops). This benchmark was performed in a subset of 77 BMRB entries of structured proteins that were deposited from 2016 onwards, and discarded from the training set. The structured sequence-based kernels method (cyan) is compared with two structure-based methods: SPARTA+[229] (green) and Camshift[228] (orange). (a) The benchmark performed on the full protein sequences. (b) Benchmark performed by discarding two residues from each termini of every individual Q3 element. Adapted from Sanz-Hernández & De Simone, *J Bio NMR* (2017)[248] (licensed under CC BY 4.0).

bouring residue pair-wise corrections were established separately for three types of secondary structural elements, according to the Q3 classification that most predictors of secondary structures use,[256] namely helices (α-helices and 3_{10}-helices), β-strands (parallel and antiparallel) and coils. Initially, sequences were classified into the different Q3 categories according to the DSSP[257] analysis of their structures. The accuracy of this approach was directly compared to methods that predict CS from a protein three-dimensional tertiary structure. We chose to benchmark against SPARTA+,[229] a widely used neural network predictor, and Camshift,[228] which consists in a simplified linear combination model that can however be readily differentiated and used in molecular simulations (as described in Section 2.1.2).

The results show a similar level of accuracy in the kernel-based CS prediction from sequence to the structure-based methods (Figure 2.5a). A much higher accuracy is found

when excluding the boundary regions of Q3 elements. If removing two residues at each of the ends (N- and C-) of every Q3 regions, the kernel-based approach predicts CS with similar RMSD values to SPARTA+ and better than those associated with Camshift (Figure 2.5b). This finding points towards the dominance of protein sequence as the main factor in determining the CS value within internal segments of secondary structure elements.

Using the kernel-based method, boundary regions of secondary structure elements are predicted with 11.4% less accuracy than internal segments. These stark differences in RMSD suggest that Q3 transition segments require additional terms to accurately interpret their CS. In analysing the CS behaviour along secondary structure element boundaries, the concept of secondary CS is very helpful (this was briefly described in Section 2.1.2). The secondary CS is defined as the deviation from an atom's characteristic random coil CS:

$$\Delta\delta = \delta - \delta_{rc} \tag{2.10}$$

These deviations are generally caused by the presence of secondary structure, and have opposite signs in α-helices and β-strands. The distribution of these secondary chemical shifts along protein sequences in the PROSECCO database revealed some interesting patterns at the boundaries between different Q3 elements. Generally, the CS value progressively morphs from that of a Q3 segment into the characteristic CS of the following secondary structure element (Figure 2.6a). However, there were some clear anomalous transitions along these boundaries. Notably, the backbone amide ^{15}N CS follows a very unexpected pattern in segments that transition from helix to coil (Figure 2.6b), a previously unobserved phenomenon that may serve as a characteristic footprint of α-helical C-terminal capping. These secondary CS analyses served to establish simple Q3 boundary correction terms for the kernel-based predictor, which simply add the mean CS deviation for an atom at a specific position within the Q3 boundary segment.

$$CS_i^{A,corr} = CS_i^A + \Delta CS_i^{A,Q3} \tag{2.11}$$

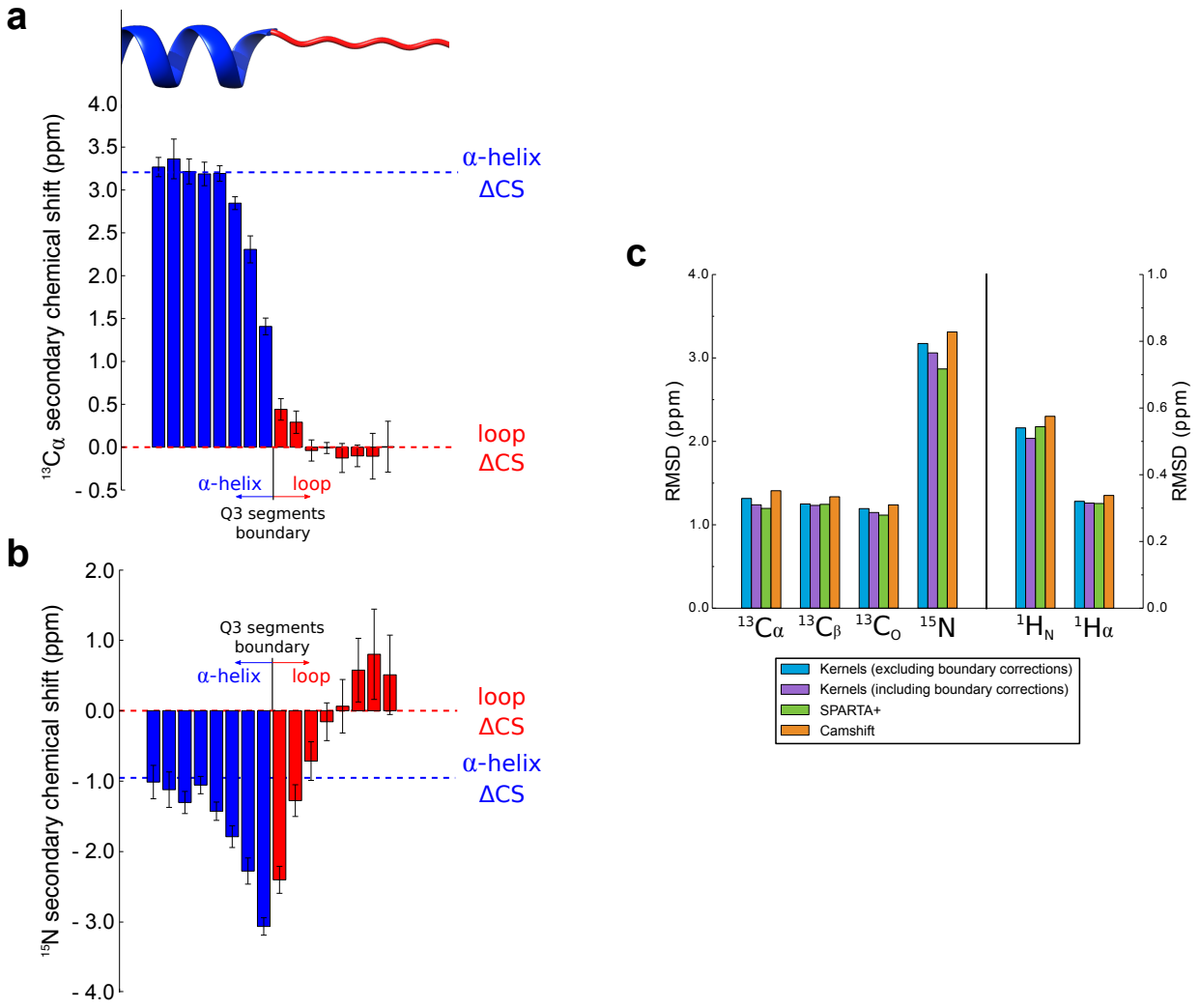


Figure 2.6: The effect of secondary structure boundaries on the CS – (a,b) Secondary CS at the boundary between two different Q3 elements, in this case an N-terminal α-helix followed by a loop. The height of each bar represents the average secondary CS as a function of the distance from the boundary between the two elements, and error bars correspond to the standard deviation. In the case of $^{13}C_\alpha$ (a), the secondary CS gradually morphs from its positive α-helical value of 3.2 ppm to its random coil value (0.0 ppm). An anomalous trend is found in the case of backbone amide ^{15}N secondary CS (b). Starting from the negative value characteristic of α-helices (-0.95 ppm), the secondary CS becomes considerably more extreme as we approach the Q3 boundary, reaching a maximum value of -3.0 ppm at the C-terminal capping residue of the helix, after which the secondary CS fades back to the random coil value. (c) Based on the behaviour of the secondary CS, position-specific Q3 boundary corrections can be derived and included in the kernel-based model (Equation 2.11). The performance of the predictor is reported for the structured benchmark dataset, without Q3 boundary corrections (cyan) and upon their inclusion (purple), and compared to SPARTA+[229] (green) and Camshift[228] (orange). Adapted from Sanz-Hernández & De Simone, *J Bio NMR* (2017)[248] (licensed under CC BY 4.0).

Applying these correcting terms to the kernel-based model improved the prediction RMSDs by 7.7% in residues within the boundary regions and 3.6% overall (Figure 2.6c).

2.2.4 PROSECCO$_{\text{FOLDED}}$: an accurate sequence-based method to predict chemical shifts in structured proteins

The level of accuracy reached by the kernel-based CS model suggests that protein sequence and secondary structure information are the dominant factors that need to be taken into account to accurately model CS. Given the recent development of highly accurate algorithms to predict secondary structure from protein primary sequences, it seems possible to generate models that bypass the experimentally-determined tertiary structure by coupling CS prediction to the prediction of Q3 segments in proteins. Building on the kernel-based method described above, and using the sequence-based predictor of secondary structure PSIPRED,[256] the CS RMSDs increase 5.5% as a result of the errors in Q3 prediction. Although this lowers the overall performance of the method, and results in noticeably worse RMSDs than SPARTA+, the predictions are still more accurate than those of the structure-based model Camshift.

In order to reduce the uncertainty associated with the prediction of Q3 elements by PSIPRED, I built a more complex model, PROSECCO$_{\text{FOLDED}}$, for the sequence-based prediction of CS using artificial neural networks (ANNs). ANNs are machine-learning tools that are remarkably powerful in modelling non-linear, complex dependencies between variables. They have been successfully employed in a variety of contexts across all scientific disciplines.[258,259]

The structure of the ANN used in PROSECCO$_{\text{FOLDED}}$ is depicted in Figure 2.7. This type of ANN is known as a feedforward single-layer regressor. It consists in a directed mathematical graph which transforms information from inputs (i.e. sequence and predicted Q3) into an output (the CS value) by means of a series of operations, which are explained as follows. The input layer of the network contains 143 nodes, which encode

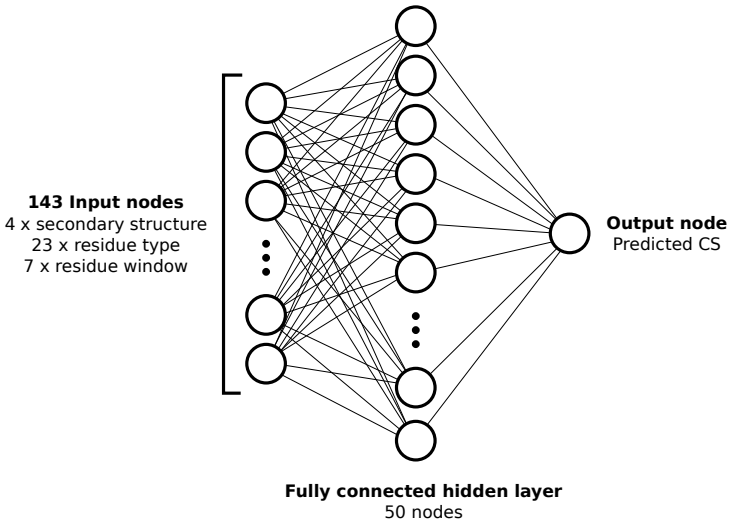


Figure 2.7: Neural network structure in PROSECCO$_{\text{FOLDED}}$ – The ANN employed in PROSECCO$_{\text{FOLDED}}$ is a fully connected directed mathematical graph with a single hidden layer. Sequence and secondary structure information is encoded into an input layer, and this information is transformed through the hidden nodes resulting in a single output value, which corresponds to the predicted CS.

the information for each residue in the sequence. The Q3 predictions are evaluated using a rolling window of seven residues. For each residue in the window, four nodes are used to describe the secondary structure (one node per Q3 region and an extra node to account for N- and C-terminal segments), amounting to 28 nodes. The remaining 115 nodes account for the primary sequence in a window of five residues (23 nodes per residue: 20 standard amino-acids and specific nodes for oxidized cysteine, *cis*-proline and protonated histidine). Each one of these nodes contains the score from the BLOSUM-62 similarity matrix.[260]

The input layer is connected to a single so-called *hidden* layer of nodes, where each node is the result of a weighted sum of all input nodes (with weights that must be calibrated to fit the data). In our case, the optimal amount of hidden nodes was 50. The values in each hidden node are then transformed by means of an activation function, the hyperbolic tangent function in this case. This steps allows the network to model non-linear relationships between the data and the output. Finally, the output layer receives another weighted sum of all the activated hidden nodes. In the PROSECCO$_{\text{FOLDED}}$ case, the output layer contains a single node: the predicted CS value.

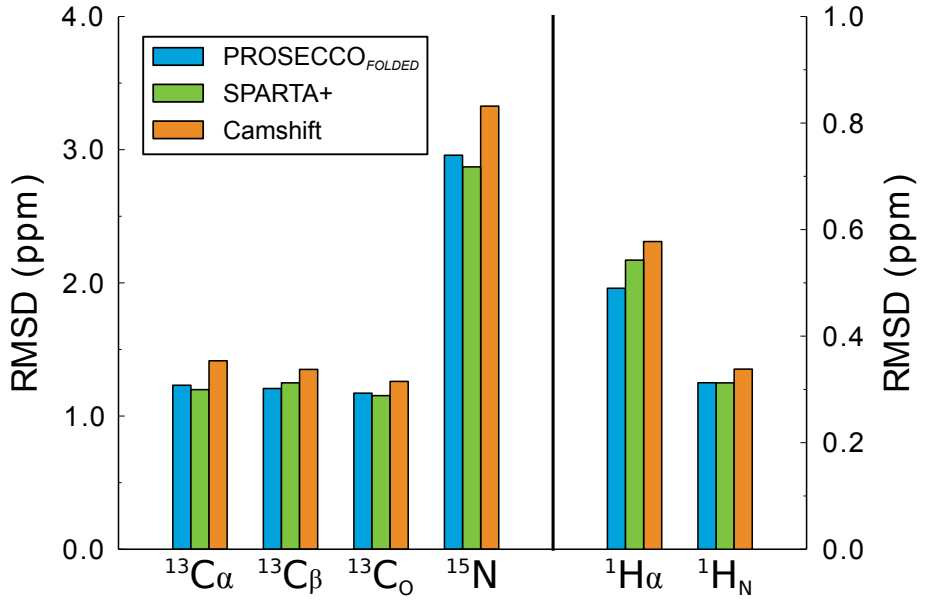


Figure 2.8: Performance benchmark for PROSECCO$_{\text{FOLDED}}$ – The final version of PROSECCO$_{\text{FOLDED}}$ consists of an artificial neural network that receives information from the protein sequence and its PSIPRED[256] predicted secondary structure. The RMSDs reported for the benchmark dataset reveal that PROSECCO$_{\text{FOLDED}}$ (cyan) has an overall performance similar to the structure-based method SPARTA+[229] (green) and improves the quality of the predictions from Camshift[228] (orange). Adapted from Sanz-Hernández & De Simone, *J Bio NMR* (2017)[248] (licensed under CC BY 4.0).

Once it is constructed, the ANN must be trained by calibrating the weights in the model to fit the experimental data in the database. This calibration generally involves minimizing the mean squared error between the outputs and the experimental CS data with algorithms such as gradient descent. In parametrising the PROSECCO$_{\text{FOLDED}}$ network, the Adam algorithm[261] (a stochastic variation of gradient descent) was employed.

The fully parametrised ANNs result in a model, PROSECCO$_{\text{FOLDED}}$, that uses exclusively protein sequence information to predict all possible CS in a protein. Its accuracy was benchmarked against the same structure-based models, showing that the added complexity of the method is able to improve the predictions considerably. PROSECCO$_{\text{FOLDED}}$ reaches the level of accuracy of complex knowledge-based approaches such as SPARTA+ and improves the description of CS provided by Camshift, the most widely used approach in CS-driven molecular simulations (Figure 2.8).

A more detailed comparison between the benchmark of PROSECCO$_{\text{FOLDED}}$ and SPARTA+ is given in Figure 2.9. By comparing the performance entry-by-entry, it is evident that

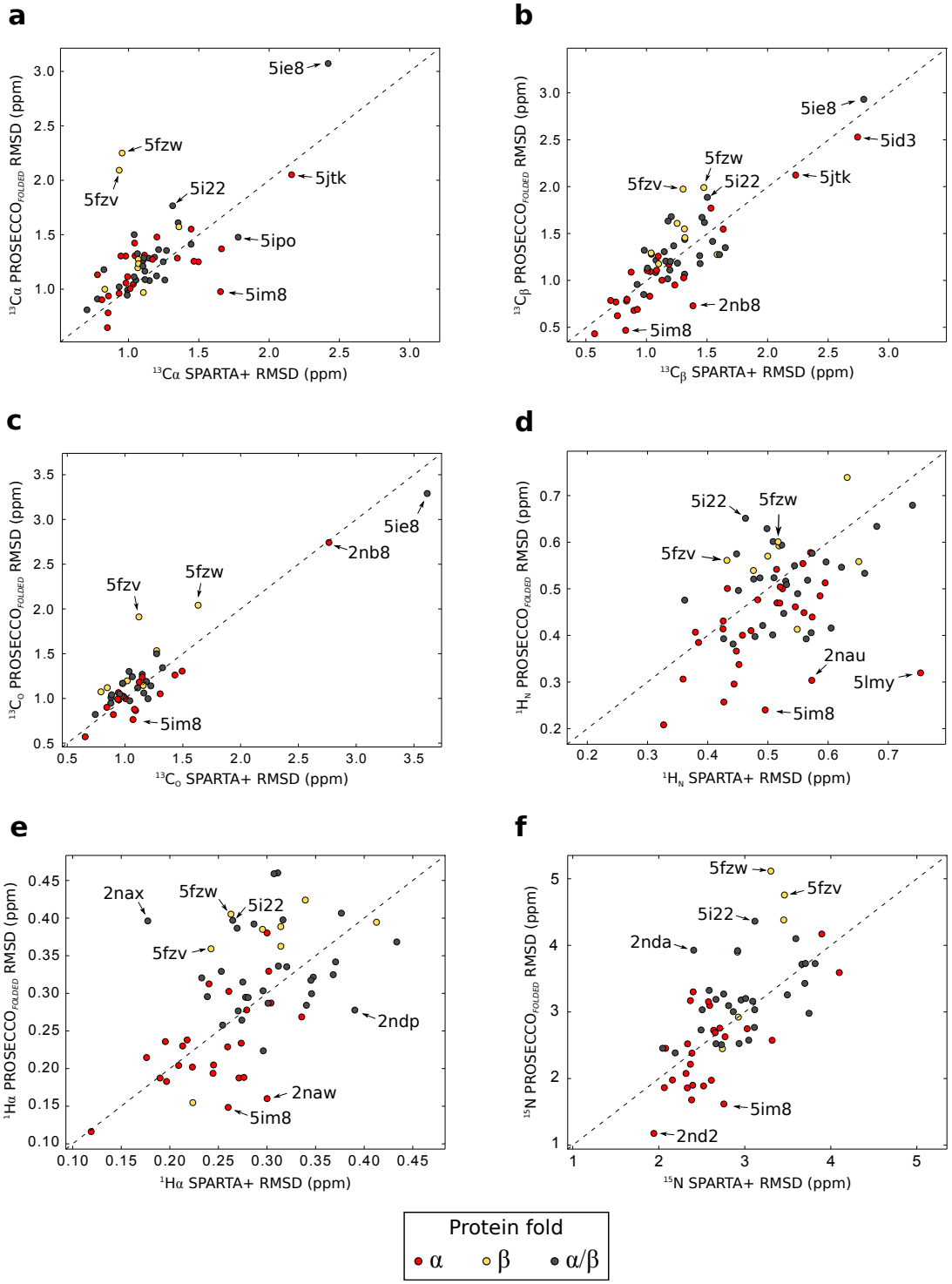


Figure 2.9: Direct comparison between PROSECCO$_{\text{FOLDED}}$ and SPARTA+ for individual BMRB entries – For each entry evaluated in the benchmark, the RMSD for PROSECCO$_{\text{FOLDED}}$ is plotted on the y-axis, and the SPARTA+[229] RMSD on the x-axis. Entries below the dotted line are therefore better predicted by PROSECO$_{\text{FOLDED}}$, and the opposite is true for entries below the line. Each dot is coloured to represent their type of protein fold. SPARTA+ generally produces better predictions for β-rich structures, with PROSECCO$_{\text{FOLDED}}$ often performing better in α-helical structures. Only entries with at least 10 CS observations are plotted. Adapted from Sanz-Hernández & De Simone, *J Bio NMR* (2017)[248] (licensed under CC BY 4.0).

$\text{PROSECCO}_{\text{FOLDED}}$ performs better in structures that are rich in α-helices, often beating SPARTA+ in the benchmark. However, the sequence-based approach is less accurate in β-rich entries. This decrease in performance may arise from the tertiary contacts that intrinsically occur in β-sheets when strands come together, between residues that are far away in the sequence.

The high level of CS accuracy achieved by only considering sequence information is a very encouraging step towards building better CS models. However it should be noted that the accuracy of $\text{PROSECCO}_{\text{FOLDED}}$ is reliant on the quality of the Q3 prediction by PSIPRED. Although the algorithm is generally very accurate, there may be cases where a bad Q3 prediction will have a *knock-on* effect onto the CS prediction. This is here illustrated with the extreme case of the GA95/GB95 proteins, which are artificially designed proteins that have 95% sequence identity but clearly distinct folds.[262] Since the proteins differ only in three residues, the accurate prediction of their secondary structure is an extremely challenging case. PSIPRED generates broadly similar Q3 predictions for both, resulting in errors that propagate to the CS prediction by $\text{PROSECCO}_{\text{FOLDED}}$, which performs considerably worse than SPARTA+ (Figure 2.10). However, if the correct Q3 assignments are provided, $\text{PROSECCO}_{\text{FOLDED}}$ is able to match the accuracy of SPARTA+, showing that in these cases the limiting factor of the model comes from the Q3 prediction.

The applicability of PROSECCO was extended to model CS in amino-acid sidechains in both folded and disordered proteins. Given the large datasets curated in this study, the model is able to generate accurate CS tables not only for backbone atoms, which are the ones generally modelled by most predictors. A particularly interesting group of sidechain CS is that of methyl groups, since they are often used in the context of Transverse relaxation optimized spectroscopy methods[264] as tools to study structure and dynamics of very large proteins that would otherwise not be feasible to study by NMR. The prediction of methyl CS in IDP is considerably more accurate than in folded proteins, as expected due to their lower CS dispersion (Figure 2.11a). $\text{PROSECCO}_{\text{FOLDED}}$ was compared against the structure-based method CH3SHIFT[263] (Figure 2.11b). Remarkably, the sequence-based method is able to provide better predictions for ^{13}C atoms, despite the common

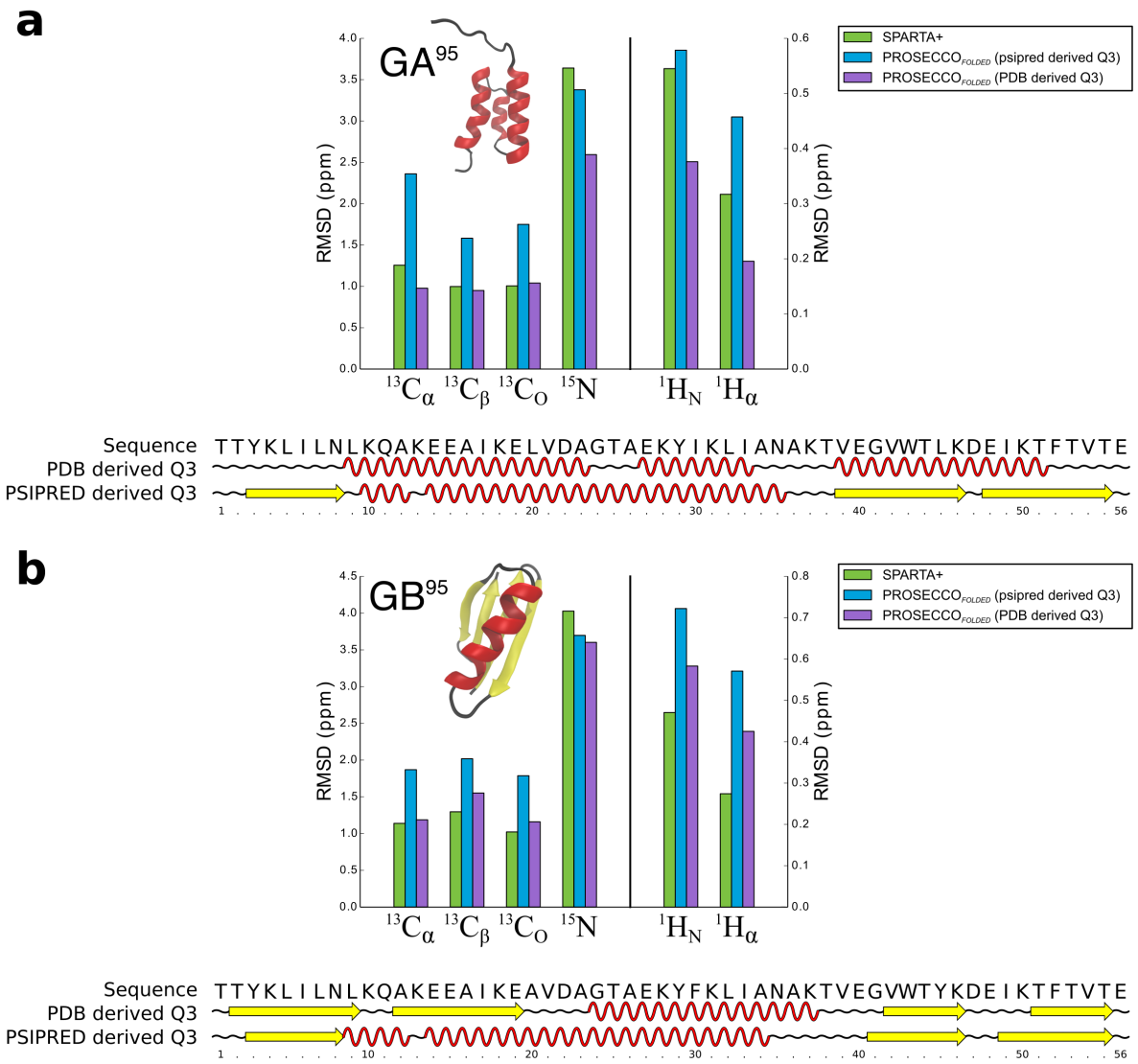


Figure 2.10: Chemical shift prediction in the GA^{95}/GB^{95} system – These two proteins share 95% sequence identity but fold into completely different structures.[262] Their secondary structure predictions by PSIPRED[256] are practically identical due to their high sequence similarity, which results in overall inaccurate CS predictions by $PROSECCO_{FOLDED}$. By contrast, if the model is provided with Q3 information derived directly from the PDB structures, the accuracy of the predictions become much better, comparable to SPARTA+,[229] or even providing an improvement in the case of the helical GA^{95}. Adapted from Sanz-Hernández & De Simone, *J Bio NMR* (2017)[248] (licensed under CC BY 4.0).

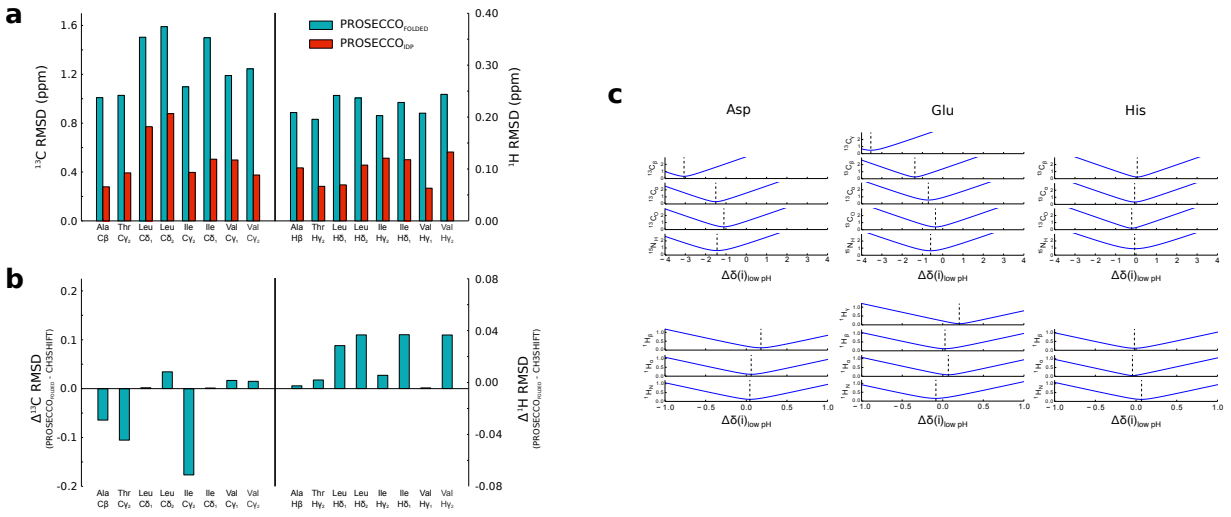


Figure 2.11: Methyl sidechain predictions and pH-dependent corrections in PROSECCO – The final CS models were extended to predict sidechain atoms CS in both folded and disordered proteins, as well as to consider pH effects. (a) RMSD for predicted ^{13}C and ^{1}H CS of methyl groups in folded and intrinsically disordered proteins. (b) Differences in prediction RMSDs for methyl ^{13}C and ^{1}H atoms between PROSECCO$_{\mathrm{FOLDED}}$ and the structure-based method CH3SHIFT.[263] (c) Calibration of pH correction factors in disordered proteins. The pH value in the BMRB entries employed for the calibration was centered at 6.4. A separate set of low pH entries centered at pH 2.8 was used to derive pH-correction factors $\Delta\delta_{i,pH}$ for protonable amino acids Asp, Glu and His. Adapted from Sanz-Hernández & De Simone, *J Bio NMR* (2017)[248] (licensed under CC BY 4.0).

tertiary effects (such as ring current effects) that often influence the CS of methyl groups, as they tend to be buried in the hydrophobic core of the protein. The prediction of CS in ^{1}H atoms was however modelled more accurately by CH3SHIFT, by taking the tertiary structure into consideration.

Finally, given the large dataset used to calibrate the method, PROSECCO can provide different predictions depending on the oxidation state of cysteine (which strongly affects the CS of its atoms, particularly the $^{13}C_{\beta}$), protonation state of histidine and cis-/trans-conformation of prolines. In addition, a further correction term was introduced to account for low pH experimental conditions. These terms act by adding an *ad hoc* value, calibrated at pH 2.8, to atoms in protonable residues (Figure 2.11c).

2.2.5 An outlook for improving protein CS models

The results outlined in the previous section provide very encouraging prospects in the advancement toward highly accurate models of CS in proteins. I have presented an approach, PROSECCO, to accurately interpret CS in both disordered and folded proteins employing only their primary amino acid sequences. In the context of IDPs, PROSECCO is able to provide better predictions than currently available similar methods that use protein sequence information. In the case of structured proteins, PROSECCO provides a method that is applicable for any protein sequence, without relying on sequence alignments and homology as the other available sequence-based methods. Furthermore, the method generates CS predictions that are as accurate as elaborate structure-based models such as SPARTA+ and better than Camshift, the only CS model used to bias molecular simulations of protein dynamics. More importantly, PROSECCO can be applied in cases where the experimental structure is not yet available, as is often the case when performing initial structural studies, providing a very useful tool for the initial interpretation of these ubiquitous NMR data.

In addition to generating accurate CS models based on protein sequence, some considerations leading to the parametrisation of PROSECCO have revealed important dependencies of the CS on protein secondary structure. In particular, it is here shown that backbone CS are mostly dependent on their secondary structure rather than tertiary structure, with internal segments of Q3 elements being accurately predicted from their sequence information (perhaps with some exceptions in the case of β-sheets). Moreover, there is a strong influence from the secondary structure elements that flank an atom. In particular, the behaviour of the CS at the boundaries between secondary structure elements shows that there are clear CS deviations around secondary structure breaks. These deviations may propagate up to three residues on each direction away from the break in secondary structure, suggesting that larger residue windows should be employed to better describe the CS in protein structures (considering up to seven residues, in contrast to the three used in SPARTA+).

In the broader context of this thesis, the application of PROSECCO to experimental models of protein dynamics will require the extension of the method to include structural information. The work that has been outlined here paves the way for the development of such a method that improves the currently available CS models. The curated database can be readily used to generate a model that builds on the structure of $\text{PROSECCO}_{\text{FOLDED}}$, a neural network that incorporates tertiary structural information in a quantitative manner (e.g. dihedral angles and distances). This model can provide a mathematical expression that may be differentiated and employed in molecular simulations, improving the accuracy of CS-driven studies of protein dynamics. The development of such a method can serve as a very valuable tool in both soluble and membrane environments, given the great availability of NMR chemical shifts in all these protein systems.

The published PROSECCO method[248] is available to the community through a web server at http://desimone.bio.ic.ac.uk/prosecco/

2.3 Oriented solid-state NMR

2.3.1 Chemical shift anisotropy and dipolar coupling

Solid-state NMR (ssNMR) is one of the most popular structural techniques in the study of membrane proteins. It allows the high resolution reconstruction of the structure and dynamics of proteins embedded in lipid bilayers, an environment very close to their native. Despite its current popularity and applicability, the development of ssNMR has been historically hampered by strong technical limitations that were very challenging to overcome. A key difference between solution NMR and ssNMR is that in solution protein molecules are free to move and constantly reorient with respect to the external magnetic field B_0 (a phenomenon known as *molecular tumbling*). As a result, anisotropic nuclear spin interactions always average to their isotropic value in solution. In particular, the very strong dipolar and quadripolar interactions average to zero, as they have no isotropic

component. In solids, proteins are constrained without possibility of tumbling (whether it is in powdered crystals, lipid membranes, or other solid materials). This means that the anisotropic nuclear spin interactions are not averaged, and when recording an NMR spectrum one would observe the superposition of all signals corresponding to every possible orientation of the protein in the sample. This naturally leads to very broad spectral lines, which are impossible to interpret in complex systems.

There are two mainstream approaches to overcoming this problem and obtaining high resolution ssNMR spectra. The first involves artificially spinning the sample about an axis that is situated at an angle of approximately 54.7 degrees (known as the *magic angle*) with respect to B_0. The strongest anisotropic spin interactions (the dipolar and the first-order quadripolar) are orientationally dependent with respect to B_0 according to the following term

$$3\cos^2\theta - 1 \tag{2.12}$$

The magic angle is that at which this term becomes zero. By spinning at a very high speed about this angle, those anisotropic interactions are therefore averaged to zero and the broad lines disappear, yielding high resolution signals that correspond to the isotropic nuclear spin interactions (e.g. isotropic chemical shifts as described in Section 2.1). Nonetheless, anisotropic interactions contain very valuable information, and methods have been developed to re-introduce some of them through recoupling radio-frequency pulses.[265]

The second approach to obtaining high resolution ssNMR spectra is to produce samples that are precisely aligned, with the membrane normal axis pointing in the same direction as B_0, so that all proteins find themselves in the same orientation. In that case, the anisotropic interactions can be directly measured as narrow signals, since we no longer have an aggregated signal of all possible orientations of the system. There are two different ways to obtain alignment of the proteins in the sample. The first is to mechanically align lipid bilayers by drying them onto layers of a flat substrate (normally glass), and

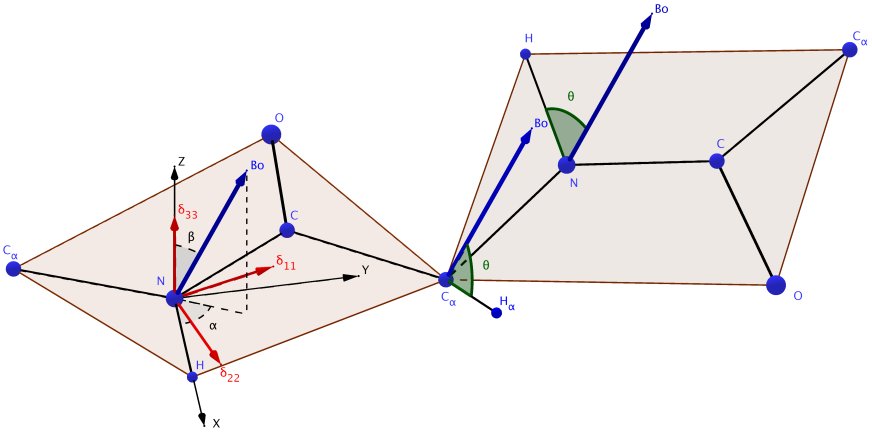


Figure 2.12: Structural models for chemical shift anisotropy and dipolar coupling – CSA and DC can be calculated based on the atomic coordinates of a protein's atoms.[268] The amide ^{15}N CSA is modelled as a tensor, δ, centered on the atom of interest, whose orientation with respect to the external magnetic field, B_0, determines the calculated value. α and β are Euler angles used in the calculation to transform to the principal axis frame. The DC between two atoms is only dependent on the angle formed between their bond and B_0. The DC between $^{1}H_{\alpha}$-$^{13}C_{\alpha}$ and amide ^{1}H-^{15}N are depicted here.

slowly re-hydrating them afterwards.[266] The second approach makes use of lipid bicelles instead of bilayers. Tuning the ratio of long- and short-chain lipids allows the creation of bicelles that will spontaneously align in the presence of a strong magnetic field.[267] Although oriented ssNMR approaches involve more difficulties in terms of sample preparation, they allow the direct detection of two crucial orientational NMR parameters, namely dipolar couplings (DC) and chemical shift anisotropy (CSA).

Dipolar couplings arise from the direct interaction through space of two nuclear spins. The strength of the interaction is dependent on the inverse cube of the distance between the spins. Its orientation dependence comes simply from the angle of the vector that joins the nuclei with respect to the external magnetic field. Figure 2.12 shows a depiction of DCs between amide ^{1}H-^{15}N and $^{1}H_{\alpha}$-$^{13}C_{\alpha}$. Precisely, dipolar couplings can be calculated according to the equation

$$DC = \frac{1}{2}\zeta_{DC}(3\cos^2\theta - 1) \tag{2.13}$$

where θ is the angle between the vector that joins both nuclei and B_0. ζ_{DC} is a DC con-

stant specific to the chemical bond of interest, with the physical expression (in Hz) as follows

$$\zeta_{DC} = -\frac{\mu_0}{8\pi^2}\frac{\gamma_A\gamma_B\hbar}{r_{AB}^3} \tag{2.14}$$

where μ_0 is the magnetic permeability of free space, $\hbar$ is the reduced Planck constant, γ_A and γ_B are the gyromagnetic ratios of nuclei A and B respectively and r_{AB} is the distance between the nuclei. In the case of amide ^{1}H-^{15}N DCs, the ζ_{DC} constant is set to 10.52 kHz, corresponding to an N-H distance of 1.05 Å. In the case of $^1H_\alpha$-$^{13}C_\alpha$, ζ_{DC} is set to -22.68 kHz, corresponding to a H_α-C_α distance of 1.10 Å.

CSA is a result of the anisotropic electronic environment that influences the chemical shift (as described in Section 2.1.1). The shielding provided by the local nuclear environment is dependent on the orientation with respect to B_0 . The orientation-dependence of the CSA is mathematically expressed as a 3×3 tensor, δ. A depiction of such a tensor for the amide ^{15}N atom is shown in Figure 2.12. The value of the ^{15}N can then be calculated according to the following expression

$$\delta_{15_N} = \delta_{11} \times \sin^2(\alpha - 17) \times \sin^2\beta + \delta_{22} \times \cos^2\beta + \delta_{33} \times \cos^2(\alpha - 17) \times \sin^2\beta \tag{2.15}$$

where δ_{11}, δ_{22} and δ_{22} are the magnitudes of the three principal tensor components, respectively set to 64.0, 76.0, 216.9 ppm for non-glycine residues and 46.5, 66.3, 211.6 ppm in glycine. α and β are the Euler angles (expressed in degrees) to transform from the laboratory frame to the principal axis frame.

The most common experiments to recover the CSA and DC from aligned samples are ^{1}H-^{15}N double-resonance experiments. Specifically, separated local field (SLF) experiments[269] give two-dimensional spectra that correlate the CSA to the DC. The most popular variant of these experiments in oriented samples is the polarization-inversion spin-exchange at the

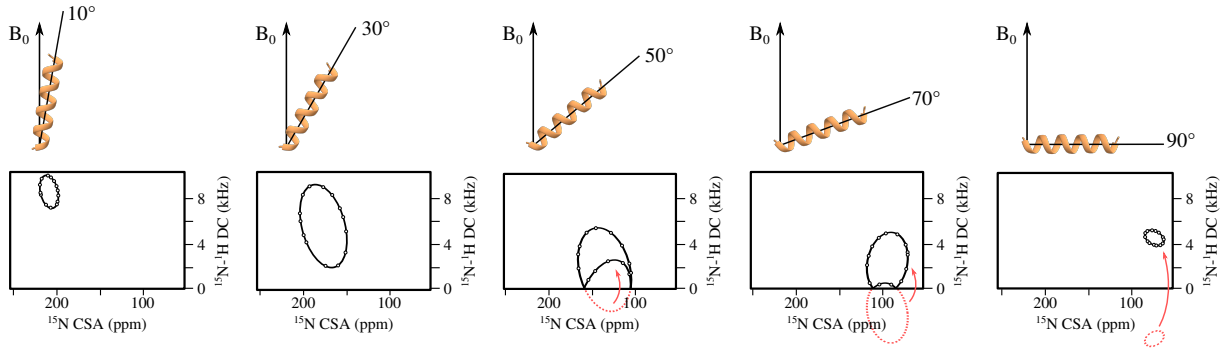


Figure 2.13: CSA and DC as a footprint of helical orientation – α-helices give rise to characteristic circular patterns in PISEMA oriented ssNMR spectra, known as PISA wheels. The location and shape of the PISA wheel within the spectrum reports on the orientation of the helix within the membrane. As helical tilt angles approach 90 degrees, the DCs become negative. However, standard PISEMA experiments are unable to discriminate the sign of DCs, and generally only the positive section of the spectrum is shown. There are specific pulse sequences that can recover the correct sign of the DC in ambiguous cases.[271]

magic angle (PISEMA).[270] Peaks in these spectra correspond to the CSA and DC of backbone amide groups, and their location in the spectrum is dependent on the orientation of the peptide plane with respect to the membrane normal axis. These data therefore carry very large amounts of information about the protein topology and orientation within the lipid membrane.

In α-helices, the repeating pattern of the helix results in a periodic arrangement of the amide N-H vector orientations with respect to B_0. This results in PISEMA peaks that are arranged into circles, or so-called *PISA wheels*.[272] In ideal helices, different orientations in the membrane will result in specific positions of the PISA wheel within the spectrum (see Figure 2.13). Several structural refinement methods based on oriented ssNMR experiments have exploited this footprint of helical orientation, by fitting the CSA and DC to find the optimal crossing and rotational angles of helices in transmembrane protein structures.[273,274] However, as useful as CSA and DC may be for structural determination, these static fitting methods neglect the protein dynamics that often contribute to the CSA and DC spectral signals, since these are averaged over the millisecond timescale. In that context, a dynamical interpretation of such models is desirable in order to decipher the protein conformational flexibility encoded by oriented ssNMR data.

2.3.2 Implementing ssNMR restraints in molecular simulations

Classical molecular dynamics simulations

In biomolecular molecular dynamics (MD) simulations, the temporal evolution of molecules is sampled by considering the three-dimensional positions of atoms and the interactions between them. These interactions are described by means of semi-empirical force fields, which are functions that describe the potential energy of the system as a sum of different contributions. The different contributions are separated into bonded (between atoms that share covalent bonds) and non-bonded.

$$V_{FF} = V_{bonded} + V_{non-bonded} \tag{2.16}$$

Bonded contributions include those coming from bond stretching and angle bending, which involve two and three bonded atoms respectively. Additionally, there are four-body interactions in the form of torsions (or dihedral) angles between sequential atoms, and improper dihedrals between non-consecutively bonded atoms.

$$V_{bonded} = V_{bonds} + V_{angles} + V_{torsions} + V_{impropers} \tag{2.17}$$

Non-bonded contributions are further divided into Van der Waals and electrostatic interactions. Van der Waals forces are described by a Lennard-Jones potential which accounts for the strong repulsion of atoms that are very close in space as well as the weaker attraction forces arising from the induced dipole effect. The electrostatic forces are computed by assigning partial charges to each particle in the system, using a Coulombic potential to calculate the interaction force.

$$V_{nonbonded} = V_{VdW} + V_{electrostatic} \tag{2.18}$$

Once we have a description of the potential force provided by the force field, the forces that each atom experiences is calculated as the negative derivative of that potential.

$$F_i = -\frac{dV}{dr_i} \tag{2.19}$$

where r_i denotes the Cartesian coordinates of each atom i. Once the forces are known, the trajectory of atoms can be computed by solving Newton's equations of motion. More specifically, Newton's second law states that the acceleration experienced by an object is dependent on the force applied to it and its mass:

$$F = ma \tag{2.20}$$

The above equation can be expressed in terms of the Cartesian positions of each atom, r_i:

$$\frac{d^2 r_i}{dt^2} = \frac{F_i}{m_i} \tag{2.21}$$

In molecular simulations, the spatial motions of all particles are coupled. Solving the equations of motion therefore becomes a many-body problem which cannot be solved analytically. Numerical methods are therefore employed to approximate the equations. In particular, finite difference methods based on the Verlet integration algorithm are employed. The Verlet algorithm uses a discrete timestep of integration, δt, and the atomic positions at two different timepoints, the current and the preceding. The solution to Newton's equation is based on a Taylor series, expanding the atomic positions at times $t + \delta t$ and $t - \delta t$:

$$r_i(t + \delta t) = r_i(t) + r_i'(t)\delta t + \frac{\delta t^2}{2!} r_i''(t) + \frac{\delta t^3}{3!} r_i'''(t) + O(\delta t^4) \tag{2.22}$$

$$r_i(t-\delta t) = r_i(t) - r_i'(t)\delta t + \frac{\delta t^2}{2!}r_i''(t) - \frac{\delta t^3}{3!}r_i'''(t) + O(\delta t^4) \tag{2.23}$$

Adding the above two equations and substituting the second derivative of r_i as in Equation 2.21 gives a solution to the equation:

$$r_i(t+\delta t) = 2r_i(t) - r_i(t-\delta t) + \frac{\delta t^2}{m_i}F_i(t) + O(\delta t^4) \tag{2.24}$$

which uses the positions from the current timestep, $r_i(t)$ and the previous one, $r_i(t-\delta t)$ as well as the forces computed from the force field, $F_i(t)$.

It is important to note that the first order derivative terms of $r_i(t)$ cancel out, meaning that atomic velocities cannot be calculated from this expression. Modifications to the Verlet algorithm, such as the leap-frog or the Velocity Verlet integrators have been developed to allow the direct calculation of velocities.

Algorithmic improvements in the implementation of restraints

The inclusion of CSA and DC into molecular simulations was first described in the context of SERCA regulatory proteins sarcolipin and phospholamban (PLN).[201] This thesis builds on that work, by using the same algorithmic approach to guide MD simulations of membrane proteins. However, several improvements were made to the code, allowing for considerable performance improvements.

The original algorithm is based on the GROMACS package,[275] but could only be ran in parallel across multiple central processing units (CPUs) by using the particle decomposition method, which distributes the calculation of particle interactions across all computing cores based on their position at the start of the simulation. During the simulation the particles randomly diffuse across the simulation box, and each node needs to consider interactions with all nearby atoms, which will be split across all other CPUs. Therefore, all CPUs must communicate amongst themselves to compute the force interactions. An

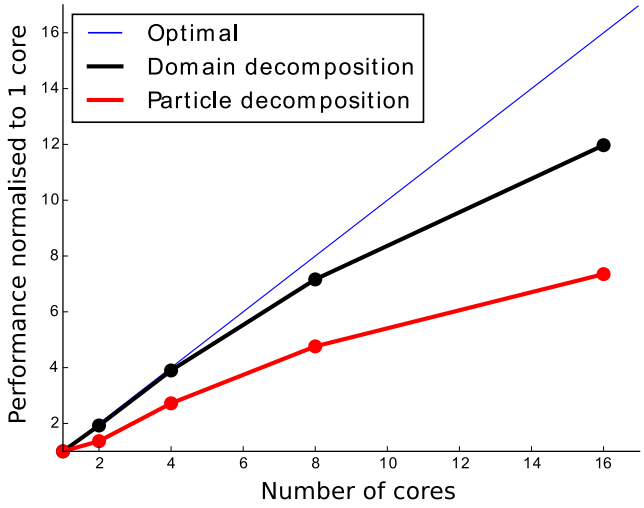


Figure 2.14: Parallelisation of the ssNMR restrained MD algorithm – The original implementation of the ssNMR restraints was further developed to render it compatible with the domain decomposition method, which provides a noticeable increase in performance, particularly as more CPUs are used in a simulation. This test was carried out on a cluster of AMD Opteron 6180 SE processors, simulating 3 parallel replicas of a solvated membrane protein system, the phospholamban pentameric structure 2KYV,[276] containing 78,262 atoms.

alternative method can be found in domain decomposition, where the simulation box is split into smaller cells that are assigned to different CPUs based on their spatial location. The atoms are periodically re-assigned to different CPUs according to the cell they find themselves in during the simulation, ensuring that a CPU must only communicate with the ones assigned to its neighbouring cells. This algorithm scales much better as the number of CPUs employed in the simulation becomes larger. I implemented a version of the ssNMR-restrained MD algorithm compatible with domain decomposition, allowing its more efficient scaling (Figure 2.14).

Secondly, the algorithm was further developed to make it compatible with the Verlet method to compute neighbour lists[277] (the list of particles below a certain distance cutoff). Crucially, this improvement allows the incorporation of ssNMR restraints in MD simulations that employ graphical processing units (GPUs) rather than CPUs. The use of GPUs in MD has provided a qualitative jump in performance over recent years. Together, these algorithmic improvements allowed the efficient application of ssNMR-guided simulations to a variety of large membrane systems, which would have otherwise been impossible to

tackle in a timely fashion.

Application to oligomeric systems

The ssNMR-restrained MD approach requires the back-calculation of CSA and DC at every step of the simulation from the protein's atomic coordinates. This calculation is based on the structural models illustrated in Figure 2.12. At every step, CSA and DC are back-calculated and compared to their experimental values. The restraints are then introduced through the addition of an external, experimentally-driven potential, V^{Exp}, on top of the force field:

$$V_{total} = V_{FF} + V_{Exp} \tag{2.25}$$

This potential takes the form of a harmonic function that exerts a force in the direction that minimises the disagreement.

$$V_{exp} = k \sum_{i} \left(Obs_i^{Calc} - Obs_i^{Exp} \right)^2 \tag{2.26}$$

where i runs over all available experimental data. The external harmonic potential has a flat bottom, whereby no force is applied if the disagreement is within the experimental error, δ^{Exp}.

$$Obs_i^{Calc} - Obs_i^{Exp} = \begin{cases} Obs_i^{Calc} - Obs_i^{Exp} & \text{if } Obs_i^{Calc} > Obs_i^{Upp} \\ 0 & \text{if } Obs_i^{Low} < Obs_i^{Calc} < Obs_i^{Upp} \\ Obs_i^{Calc} - Obs_i^{Exp} & \text{if } Obs_i^{Calc} < Obs_i^{Low} \end{cases} \tag{2.27}$$

where

$$Obs_i^{Upp} = Obs_i^{Exp} + \delta_i^{Exp} \tag{2.28}$$

$$Obs_i^{Low} = Obs_i^{Exp} - \delta_i^{Exp} \tag{2.29}$$

The values of the experimental errors can be explicitly included in the algorithm. In the case of undefined errors in the dataset, default values of 5 ppm and 0.5 kHz were use for CSA and DC respectively.

As described in Section 1.4, a necessary consideration when employing NMR restraints in MD simulations is that the experimental data is a result of a time and ensemble average over all the copies in the NMR sample. This can be accounted for by running a number of simulations in parallel (usually between 4 and 16) , known as replicas, which are used to back-calculate and restrain the average NMR observables rather than their individual instantaneous values.

$$Obs_i^{Calc} = \frac{1}{M} \sum_{m}^{M} Obs_{i,m}^{Calc} \tag{2.30}$$

where M is the number of replicas.

In homo-oligomeric proteins, the NMR data is further averaged between the individual protomers that make up the quaternary assembly. These protomers are in direct contact with each other and therefore their dynamics are not independent of each other, as is the case with different replicas that evolve separately. Nonetheless, the averaging that occurs in the test tube must be reproduced by the simulation, so we need a scheme that averages the data among individual protomers in a structure as well as different quaternary assem-

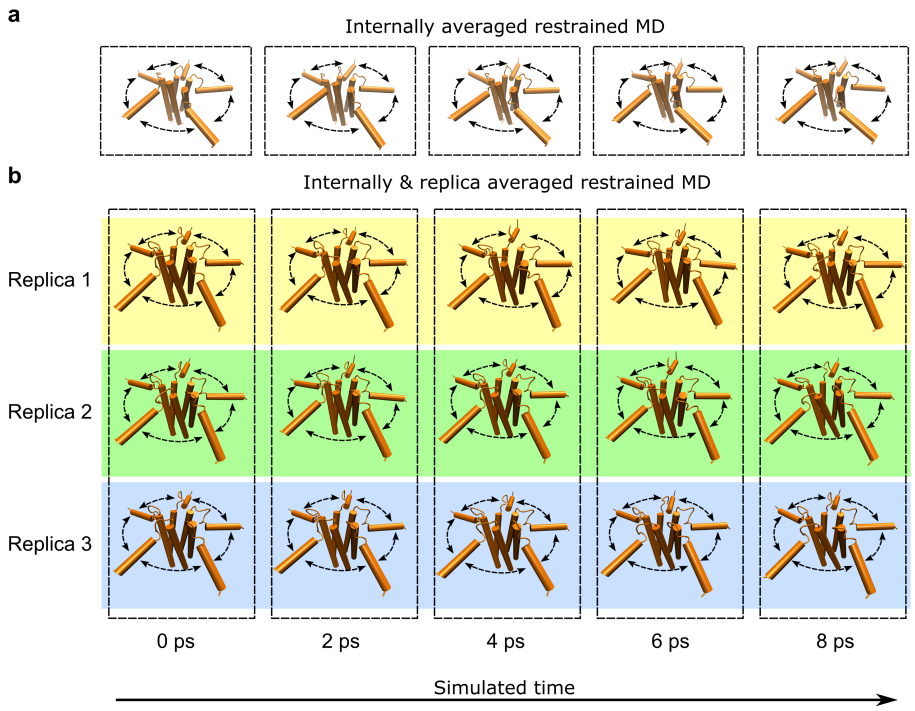


Figure 2.15: Internal averaging scheme – In ssNMR-driven simulations of homo-oligomeric proteins, two levels of averaging must be considered. (a) The first is the averaging that occurs between the different protomers within each quaternary assembly, referred to as internal averaging. (b) Secondly, internal averaging must be combined with the averaging across multiple copies of the system that evolve independently (replica averaging). Adapted from Sanz-Hernández *et al.*, *Sci Rep* (2016)[278] (licensed under CC BY 4.0).

blies that evolve independently.

$$Obs_i^{Calc} = \frac{1}{M}\frac{1}{P}\sum_{m}^{M}\sum_{p}^{P} Obs_{i,m,p}^{Calc} \tag{2.31}$$

where P corresponds to the number of protomers.

In this thesis I present simulations of two homo-oligomeric systems: the PLN pentamer and the tetrameric M2 channel from influenza. The oligomeric averaging scheme was implemented and illustrated in the case of PLN (Figure 2.15). The averaging between the protomers is here designated as *internal averaging*, whereas the classical averaging between different simulations is referred to as *replica averaging*.

Validating the sampling scheme

Before applying the ssNMR-restrained MD approach to experimental systems, the sampling scheme must be calibrated and validated. In order to do so, I employed the *reference ensemble* method, which has been previously described in the literature[176,185] and was extended here. The calibration was performed on the PLN pentameric structure embedded in DOPC/DOPE membranes (simulation details are described in Section 3.2.1).

The method involves the generation of a so-called reference ensemble of conformations, created by running an unrestrained simulation under an arbitrary force field A. This reference ensemble is then treated as an experimental collection of structures that we wish to reproduce. The reference ensemble is used to back-calculate the values of the ssNMR data that would be observed in an experiment, which we refer to as a *synthetic dataset*. The system is then simulated without restraints under a different force field B, generating an ensemble that highly differs from the reference one due to the discrepancies between the different force fields. Given that we have the atomic coordinates of both ensembles, we can directly compute structural properties of the system (e.g. interatomic distances, angles or secondary structure content) and compare their distribution in the two ensembles. If these distributions diverge, one would hope that the inclusion of *synthetic* restraints would drive the convergence toward the reference ensemble when sampling with force field B. In order to measure the similarity of any two distributions of structural properties, we used a histogram-based metric here referred to as D_{iff} (Figure 2.16). If A and B are two normalized distributions, then

$$D_{iff} = \sum_i \mid P_i^A - P_i^B \mid \tag{2.32}$$

where i runs over equally spaced bins. The value of D_{iff} ranges from zero (in identical distributions) to two (in absolutely non-overlapping distributions). In this study, all comparisons were performed over 20 bins.

The specific reference ensemble approach adopted in the case of PLN is illustrated in Fig-

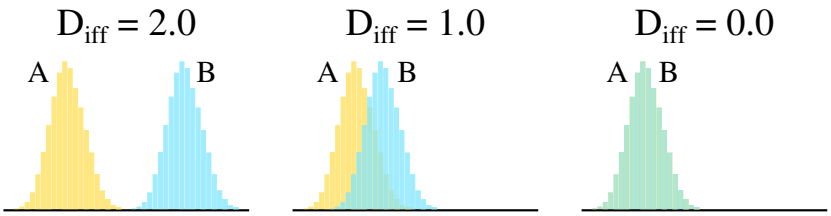


Figure 2.16: The D_{iff} metric for comparison of distributions – When comparing two distributions of structural properties such as tilt angles or pairwise interatomic distances, the histogram-based D_{iff} metric (Equation 2.32) provides a value for the similarity of the distributions. If the two distributions are completely different and non-overlapping, their associated D_{iff} value is 2. As the similarity between the distributions increases, the D_{iff} decreases until it reaches 0 for a pair of identical distributions.

ure 2.17. All simulations started from the PLN pentamer structure 2KYV[276] equilibrated and solvated in a DOPC/DOPE membrane. The reference ensemble was generated by conducting a 100 ns simulation with the AMBER99SB-ILDN force field.[279] From this simulation I extracted 10,000 equally spaced structures, which were used to back-calculate the average CSA and DC values for each residue (the synthetic data). These structures were also used to compute the distribution of structural properties that will be described later (e.g. helical tilt angles).

The same system was then simulated under the CHARMM36 force field.[126] In the absence of ssNMR restraints, CHARMM36 samples a conformational space with different characteristics to the one sampled by AMBER99SB-ILDN. This is evidence by high D_{iff} values when comparing their distributions of structural properties. By introducing the synthetic restraints, the sampling by CHARMM36 should approximate that of AMBER99SB-ILDN, and the similarities between the distributions should improve. This improvement can be quantified by computing the restrained D_{iff} value and comparing it to the free simulation:

$$\Delta D_{iff} = D_{iff}^{Restr} - D_{iff}^{Unr} \tag{2.33}$$

The resulting ΔD_{iff} will therefore have a negative value in cases where the restraints improve the representation of the reference ensemble. If the restraints worsen the structural description of the ensemble, ΔD_{iff} will be positive.

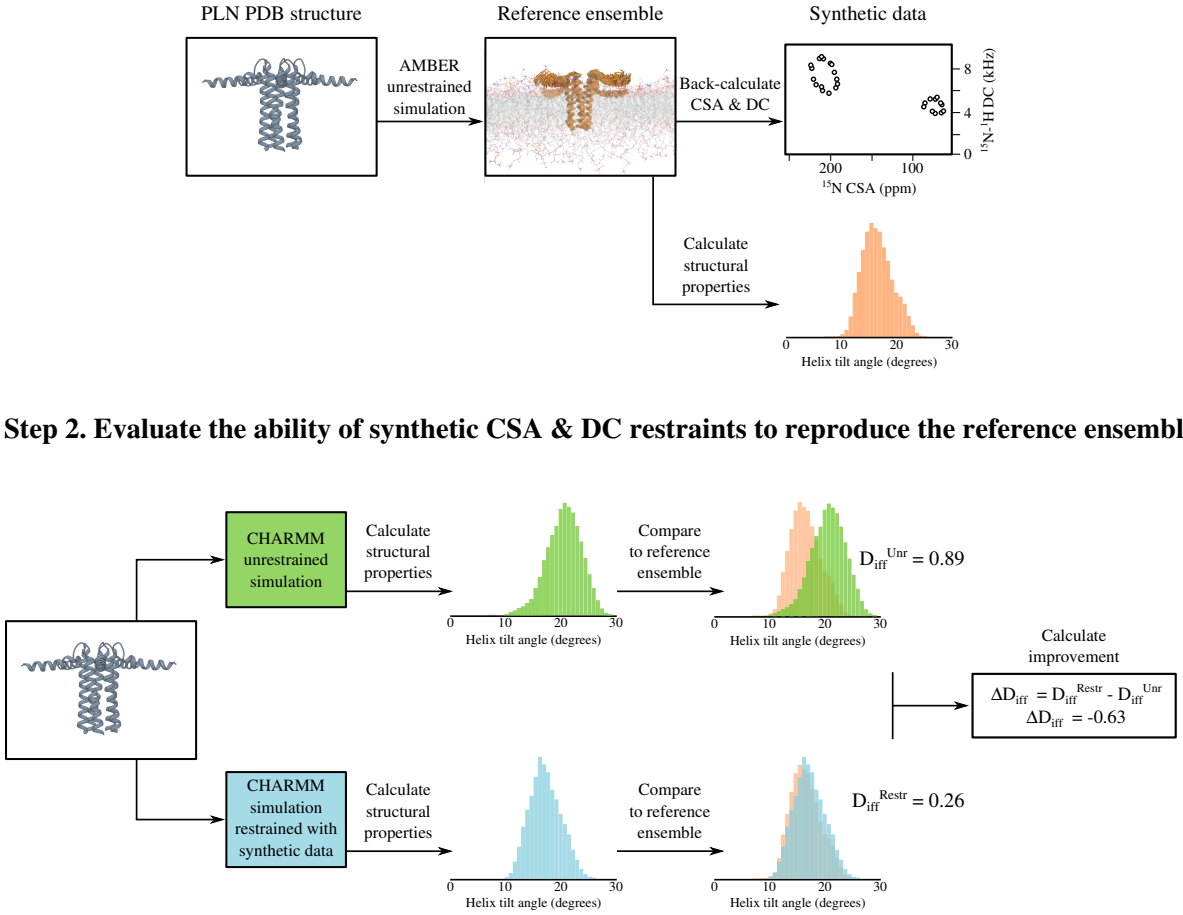


Figure 2.17: Illustration of the reference ensemble validation method – The goal of the reference ensemble test is to validate the ability of a restraining method to reproduce not only the experimental data but also the structural properties of a target ensemble. It proceeds by performing a long unrestrained simulation with a particular force field (in our case AMBER-based), and a large ensemble of structures is extracted from the simulation: the *reference* ensemble (step 1). From the reference ensemble we can back-calculate a *synthetic* set of CSA and DC, that will be later used as restraints. In addition, since we have the atomic coordinates of all structures, we can calculate the distribution of any structural property of interest (e.g. helical tilt angles). In step 2, a different force field is chosen to run a new simulation (in our case, CHARMM-based). The CHARMM simulation is first run unrestrained, and structures are extracted from it generating an *unrestrained* ensemble. We can then calculate the distribution of helical tilt angles in this unrestrained ensemble and compare it to the reference distribution, obtaining an unrestrained D_{iff} value, D_{iff}^{Unr}. A new CHARMM simulation is then run, but now incorporating the synthetic CSA and DC as restraints. This simulation generates a *restrained* ensemble, from which a distribution of helical tilt angles can be computed and compared to the reference, obtaining a restrained D_{iff} value, D_{iff}^{Restr}. By calculating a ΔD_{iff}, we can evaluate whether the restraints succeeded at improving the description of the structural property under investigation. If ΔD_{iff} is positive, the unrestrained value was lower than the restrained one and therefore the restraints have worsened the structural description of the system. Negative ΔD_{iff} values are on the other hand indicative of the success of the restraining approach in improving the description of the structural properties in the protein.

In addition to verifying the ability of the restraints to accurately describe structural ensembles, this method can be used to calibrate the sampling scheme by testing different numbers of replicas, the inclusion of internal averaging and the value of the restraining force constants.

In the specific case of PLN, the comparison of the ensembles focused on the structural and topological parameters depicted in Figure 2.18a,b. These are the tilt angles of transmembrane (τ) and cytoplasmic (θ) helices, the global tilt angle of the pentameric assembly (γ) and the distribution of intramolecular distances between every pair of C_α atoms (d). This last parameter has been previously described in the literature as a sensitive reporter of atomic structure and overall fluctuations in protein ensembles.[176]

Using these reporting structural properties, twelve different ensemble-averaging schemes were tested for the incorporation of synthetic ssNMR restraints. These correspond to six setups with a different number of replicas (1, 2, 3, 4, 8 and 16), each setup with and without internal averaging. The ability of each setup to reproduce the reference ensemble was evaluated by calculating their ΔD_{iff} values, summarised in Figure 2.18d. In addition, an extra simulation was run using AMBER99SB-ILDN, the same force field employed in the generation of the reference ensemble. This provided a new ensemble that is highly similar to the reference, at the level of statistical sampling error. The ΔD_{iff} for that ensemble therefore corresponds to the desirable amount of improvement that can be obtained by the ssNMR-restrained sampling approach.

Overall, the results from the validation test show that the restraining approach is very successful at reproducing all the structural properties of the reference ensemble. The inclusion of internal averaging improves in all cases the quality of the ensembles without resulting in over-fitting of the data. The optimal sampling scheme requires the combination of replica and internal averaging, with up to 16 replicas to obtain maximum accuracy. This setup reaches the highest desirable level of accuracy, corresponding to the statistical error associated with sampling under the same force field.

Under this sampling, a detailed description of the improvement in reproducing the cor-

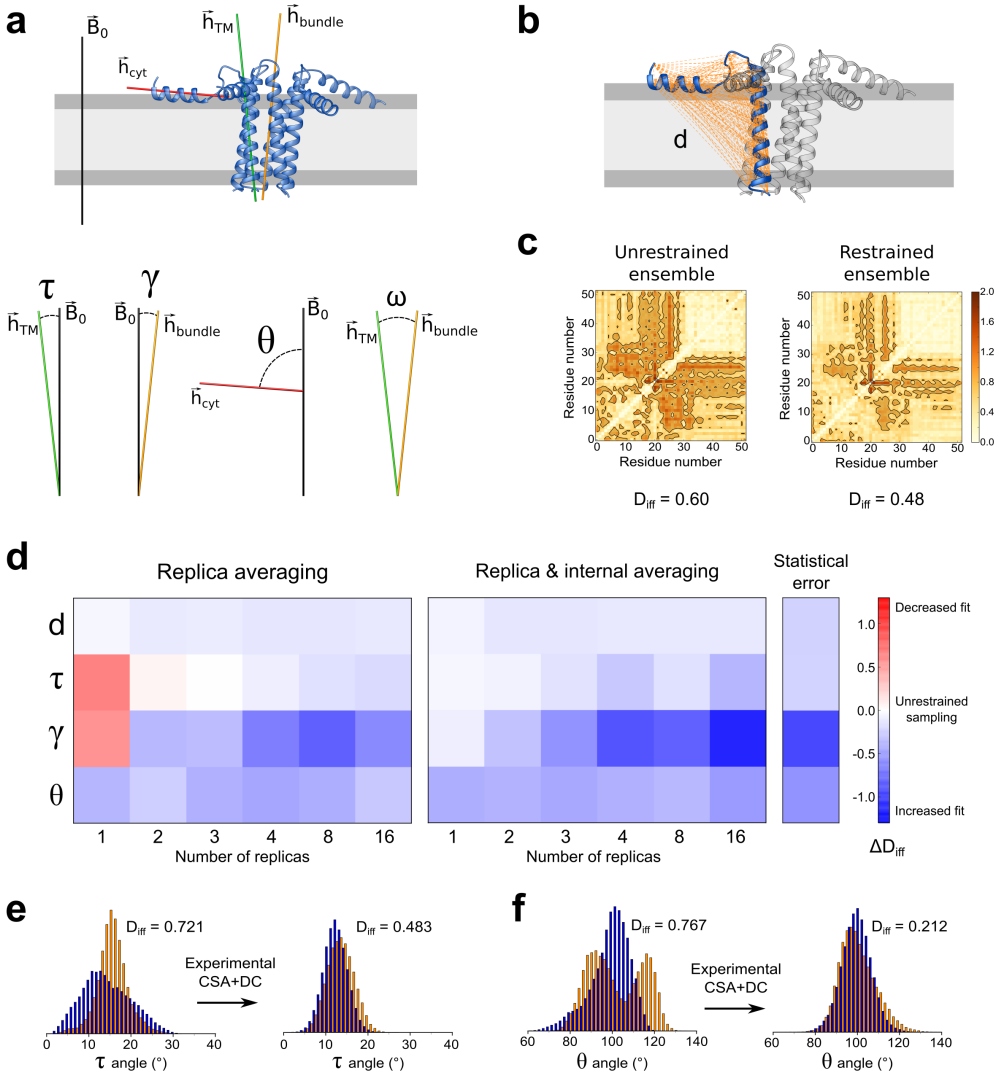


Figure 2.18: Validation of the restrained sampling approach in the phospholamban pentamer – (a) Tilt angles formed between different helical axes in PLN and B_0. (b) Distributions of all possible pairwise intramolecular distances between C_α atoms in PLN. (c) Intramolecular C_α distances D_{iff} matrices for the unrestrained ensemble (left) and a synthetic restrained ensemble (right). Each point in the matrix corresponds to a distribution of pairwise distances between two residues, showing their associated D_{iff} with the reference. In the case of the unrestrained ensemble, the D_{iff} values are higher, so the distributions are more dissimilar. By introducing the restraints, the D_{iff} values decrease overall, indicating a better fit to the reference distributions. (d) Results of the reference ensemble test for 12 different restrained sampling setups. The four structural properties defined in (a,b) are used to evaluate the success of the restraints. The ΔD_{iff} value reflects the ability of the restraints to improve the description of the system (Equation 2.33 and Figure 2.17). Positive ΔD_{iff} values (red) indicate a worse description of the reference structural distributions, with positive values (blue) reporting an improvement. The rightmost column corresponds to an unrestrained simulation under the same force field as the reference ensemble, which illustrates the statistical error limit of the sampling. (e, f) Samplings were produced under two different force fields, which lead to divergent distributions of τ and θ angles if unrestrained. However, when incorporating the experimental restraints, both angles converge toward much more similar distributions. Adapted from Sanz-Hernández *et al.*, *Sci Rep* (2016)[278] (licensed under CC BY 4.0).

rect residue-by-residue interatomic distances is provided in Figure 2.18c, a depiction of the D_{iff}^2 matrix previously described in the literature.[176] Each point in this matrix represents the D_{iff} value associated with the distribution of pairwise atomic distances between the C_α atoms of two residues. A D_{iff} value of 0.7 is used as a threshold of significantly different distributions, and higher values are highlighted. In this context, the inclusion of ssNMR restraints shows a much better representation of the interatomic distances within the transmembrane and cytoplasmic helices, as well as a drastic improvement in the representation of interdomain distances.

Finally, the last test to validate the sampling approach involves the application of the actual experimental CSA and DC under the two different force fields CHARMM36 and AMBER99SB-ILDN. If both force fields are run unrestrained, the distribution of their topological parameters is highly divergent (Figure 2.18e,f). Upon introduction of the experimental ssNMR data, both force fields converge toward a single, highly similar distribution, driven by the experimental information.

Taken together, these results indicate that the restrained MD approach is able to accurately describe experimental structural ensembles of membrane proteins to the level of statistical sampling errors. This is a crucial step toward the goal of accurately characterising the conformational motions and transitions that take place in membrane protein systems with an atomic level of detail.

Chapter 3

Structural heterogeneity of phospholamban and its complex with SERCA

3.1 Phospholamban and the regulation of the SERCA pump

The functioning of the heart in healthy organisms relies on the activity of cardiac muscle cells (myocites), which periodically contract during systole and relax during diastole, thus resulting in a correct cardiac cycle. Calcium ions act as an essential second messenger in this process, causing contraction of the muscle by activating the myofilaments.[280] Muscle relaxation is triggered by a reduction of the Ca^{2+} concentration in the cytoplasm, which is relocated into the sarcoplasmic reticulum (SR) or the extracellular space. Sequestration of Ca^{2+} into the SR is achieved by the transmembrane pump Sarco(endo)plasmic reticulum Ca^{2+}-ATPase (SERCA). In human myocites, SERCA accounts for the removal of over 70% of cytosolic Ca^{2+}, and is therefore the main actor during diastole. The Ca^{2+} stored in the SR is subsequently released to achieve contraction during the next heartbeat. As a

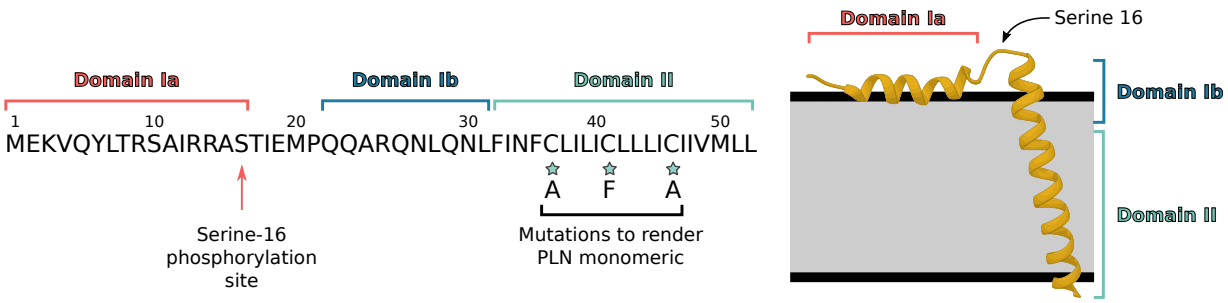


Figure 3.1: Sequence and structure of monomeric phospholamban – Phospholamban exists as a 52-residue protein, with two α-helices in its ground state. The N-terminal regulatory domain Ia features an amphipathic sequence that adopts a helical conformation, adsorbed on the cytosolic side of the SR membrane. The transmembrane segment is divided into a more hydrophilic domain Ib and the hydrophobic domain II. PLN can be phosphorylated at position 16 (serine) by protein kinase A. In order to shift the oligomeric equilibrium of PLN completely to monomeric, the three cysteine residues in domain II are often mutated to alanine, phenylalanine and alanine respectively.[282] The resulting construct is normally referred to as PLN^{AFA}.

result, the activity of SERCA strongly influences cardiac contractility by controlling the amount of Ca^{2+} that is available for release from the SR.[55]

A crucial regulator of SERCA is phospholamban (PLN). PLN is a 52-residue transmembrane protein (see Figure 3.1) that is primarily found in the SR membrane of cardiac myocites. When bound to SERCA, PLN allosterically reduces the pump's affinity for Ca^{2+}, leading to a reduced activity of SERCA. PLN is phosphorylated at serine-16 by cAMP-dependent protein kinase A (PKA) upon the stimulation of β-adrenergic receptors. Phosphorylation relieves the inhibition of SERCA, promoting cardiac relaxation and increasing contractility.[281]

PLN is found in an equilibrium between two different oligomeric species: monomeric and pentameric. When studied in the absence of SERCA, this equilibrium is heavily tilted toward the pentameric isoform, and indeed only the pentameric species is observed *in vitro*.[283] There are three cysteine residues in the transmembrane domain II that are essential in the formation of the pentamer. Although they do not form disulphide bonds, they form important packing contacts in the quaternary assembly. In order to obtain exclusively monomeric PLN, these three cysteines can be mutated to alanine, phenylalanine and alanine respectively (see Figure 3.1), generating a construct that is referred to in the literature as PLN^{AFA}. In its monomeric state, the function of PLN as a regulator

of SERCA through direct protein-protein interaction is extremely well documented. However, the biological role of the pentameric state of PLN is still a matter of debate. One hypothesis is that PLN pentamers act as a reservoir of monomers that can quickly de-oligomerise to form a complex to inhibit SERCA when needed. This is supported by biophysical evidence, such as observations that the presence of SERCA alters the oligomeric equilibrium of PLN.[284–286]

Another hypothesis is that PLN may act as an ion channel in its pentameric form. Initial biophysical and electrochemical studies suggested PLN formed pores that were permeable to Ca^{2+}.[287,288] Later studies pointed at the possibility of PLN forming Cl^- selective channels[287,288] and more recently, monovalent cations.[289,290] Other experimental,[291] structural[276,292] and theoretical[293] studies have argued against the role of PLN as an ion channel, but the argument is still open.

The dynamical equilibrium of PLN goes beyond its oligomeric state. When free in its monomeric form, PLN populates two structural states. Its ground state (T-state), maintains the domain Ia in its folded, helical structure. In the excited R-state, this regulatory domain unfolds and detaches from the membrane, an effect that is enhanced by phosphorylation (see Figure 3.2).[203] When bound to SERCA, these conformational dynamics of PLN play a crucial role in the allosteric modulation of the pump's activity. Inhibition of SERCA is mediated by the transmembrane domain II of PLN. When bound in its ground T-state or excited R-state, PLN only interacts with SERCA through its transmembrane section, and inhibits the activity of the pump (See Figure 3.2). However, the regulatory domain Ia may interact with the cytosolic domains of SERCA, forming a B-state in which the inhibition is relieved. Phosphorylation of PLN increases the population of the B-state, leading to increased activity of SERCA.[54]

These complex dynamical equilibria of PLN have been primarily studied spectroscopically, discerning between distinct structural states and qualitatively describing their populations. A detailed structural understanding of these processes is still missing. In this chapter I will present work that, exploiting the combination of experiments and molecular

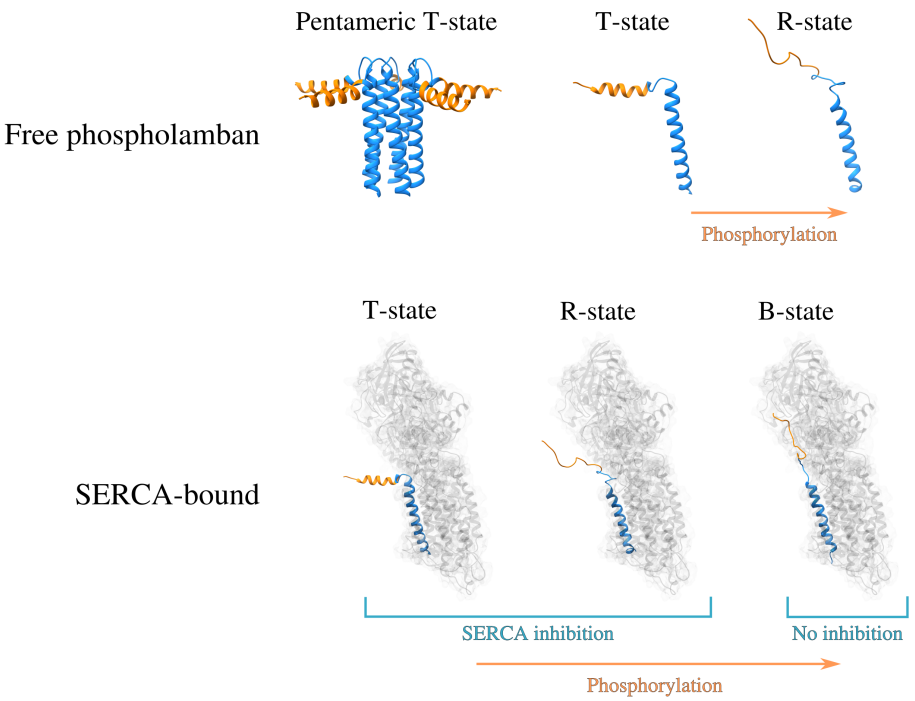


Figure 3.2: Conformational heterogeneity of phospholamban – Free PLN exists at an equilibrium between two different oligomeric states: a pentameric assembly and a monomeric form. The monomer exhibits a degree of conformational heterogeneity in its N-terminal cytoplasmic domain Ia, which exchanges between a helical conformation in its ground (T) state and an excited (R) state where it becomes partially disordered and detaches from the surface of the membrane. Phosphorylation at position 16 enhances the population of the excited state. When bound to SERCA, PLN also populates a variety of structural states. In its ground T-state, PLN acts as an inhibitor of SERCA. However, its domain Ia still populates excited states. In its R-state, the domain Ia partially loses its secondary structure, but does not interact with SERCA and PLN still acts as an inhibitor. Domain Ia may come into contact with the cytoplasmic domains of SERCA, populating the excited B-state, which relieves the inhibition of SERCA through an allosteric mechanism. Phosphorylation of PLN shifts the equilibrium toward the B-state, and thus increases the activity of the SERCA pump.

simulation, advances the understanding of the dynamics of PLN in its pentameric state and in association with SERCA.

3.2 Conformational transitions in pentameric phospholamban

3.2.1 Methods: Restrained simulations setup and structural calculations

Setting up the restrained MD simulations

All the simulations carried out in this study were set up to mimic the experimental conditions. The starting structures were obtained from the PDB (codes 2KYV for the pentameric non-phosphorylated PLN[276] and 2M3B for the pentamer phosphorylated at serine-16 - pS16[292]). The protein structures were embedded in phospholipid bilayers according to the 4:1 DOPC:DOPE ratio employed in the experiments. Tip3p waters were used to solvate the system. The generalised AMBER lipid force field[294] was used to parametrise lipids in the AMBER simulations, while the CHARMM36 lipid parameters[295] were used in combination with the CHARMM36 protein force field.

The non-phosphorylated PLN was solvated in a membrane of 215 DOPC and 54 DOPE molecules, with 10,731 water molecules, resulting in a total of 73,284 atoms. The pS16-PLN was embedded in a membrane of 291 DOPC and 73 DOPE lipids, solvated with 12,716 waters and totalling to 92,193 atoms. Both systems were simulated under two different protein force fields: AMBER99SB-ILDN[279] and CHARMM36.[126]

The systems were equilibrated during 100 ns of MD at 300 K keeping the position of heavy atoms fixed by a force constant of 1,000 kJ/(mol $\times$ nm^2), followed by 50 ns of unrestrained simulation. All simulations were performed in the NPT ensemble, coupling pressure with the semi-isotropic Berendsen method[296] (coupling constant of 1 ps, reference pressure of

No. of replicas	Replica averaging		Replica & internal averaging	
	k_{CSA}(J/(mol×ppm^2))	k_{DC}(J/(mol×kHz^2))	k_{CSA}(J/(mol×ppm^2))	k_{DC}(J/(mol×kHz^2))
1	4.5	150	12.5	250
2	12.0	200	12.5	250
3	12.5	350	12.5	350
4	12.5	375	12.5	375
8	20.0	800	20.0	800
16	25.0	1000	25.0	1000

Table 3.1: Force constants employed in the simulations of pentameric phospholamban.

1 bar). The V-rescale method[297] was employed to couple temperature to 300 K, with a constant of 0.1 ps. All simulations were carried out under periodic boundary conditions. The integration timestep was set to 2 fs and the LINCS algorithm was used for constraints.[298] Electrostatic interactions were accounted for using the Particle mesh Ewald method.[299]

Restrained simulations were performed by linearly increasing the restraining force constant over the initial 20 ns of each simulation, from zero to its final value. The values of the force constants under the different simulation setups are summarised in Table 3.1.

The agreement between the experimental and back-calculated ssNMR data can be expressed quantitatively as a Q-factor:

$$Q = \sqrt{\frac{\sum_i \left(Obs_i^{Exp} - Obs_i^{Calc}\right)^2}{\sum_i \left(Obs_i^{Exp}\right)^2}} \tag{3.1}$$

In restrained simulations, the Q-factors converge during the force equilibration phase of 20 ns and remain stable for the rest of the simulation, whereas in unrestrained cases they are much more fluctuating (Figure 3.3a). In the final averaged ensembles, the overall agreement with the experimental CSA and DC also improves through the incorporation of the restraints (Figure 3.3b).

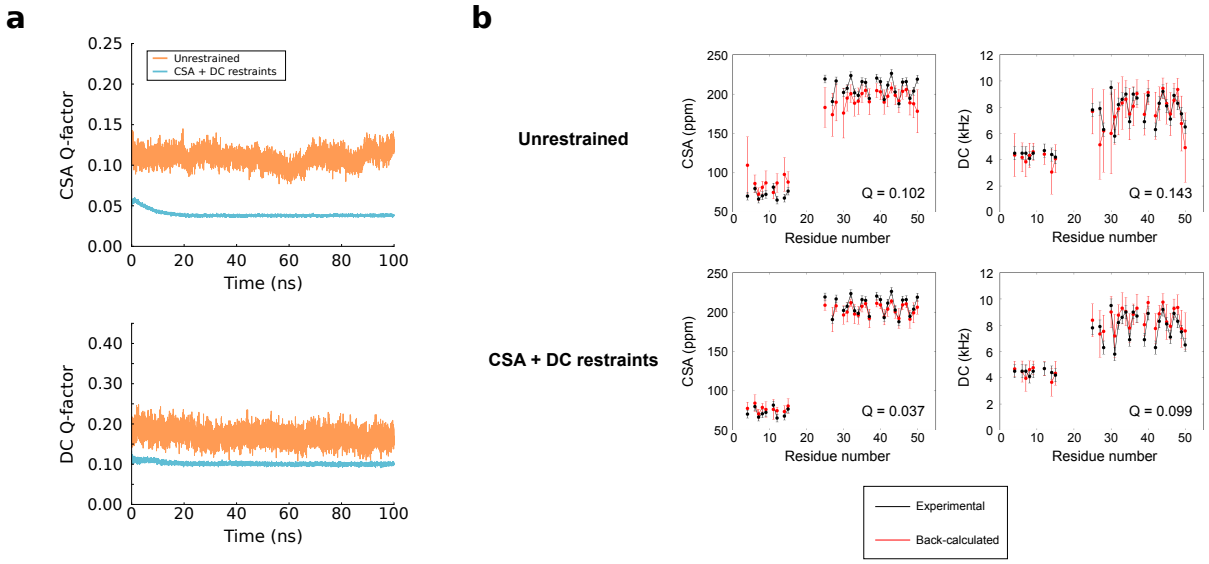


Figure 3.3: Agreement between PLN pentamer simulations and the experimental CSA and DC – (a) Q-factor progression during the simulations of PLN in unrestrained (orange) and restrained (cyan) cases. (b) Overall agreement between the back-calculated and experimental CSA and DC in the unrestrained ensemble (top) and ssNMR restrained ensemble (bottom). Adapted from Sanz-Hernández *et al.*, *Sci Rep* (2016)[278] (licensed under CC BY 4.0).

Computing helical tilt angles

Tilt angles τ and θ (Figure 2.18a) were calculated between the membrane normal (B_0) and the axis of the helices (h_{TM} and h_{cyt}). Each of these helical axis was calculated by a linear least-squares fitting of the backbone atoms in the helix (residues 25 to 51 in the transmembrane helix and 3 to 14 in the cytoplasmic helix). The global tilt angle of the transmembrane helical bundle, γ, was computed between the membrane normal, B_0 and the normalised sum of all the individual transmembrane helical axes.

S^2 order parameters

In order to calculate the NMR order parameters, S^2, of amide N-H bonds (shown in Figure 3.4c), all structures in the ensemble were superposed, using the C_α atoms of the transmembrane segments of PLN for the least-square fitting algorithm. The S^2 values were then calculated according to the approach by Best and co-workers.[300] An order parameter was also computed for the global axis of the helical bundle, h_{bundle} (Figure 3.4d). The same method was applied, except the h_{bundle} vector was considered instead of the N-H

vector.

Correlation of transmembrane helical motions

In order to assess the correlation in the motions of transmembrane helix orientations (as done in Figure 3.4e), we calculated the average h_{TM} vector (see Figure 2.18a) throughout the ensemble for each of the five protomers in PLN. For each structure in the ensemble, the average h_{TM}^{avg} of every helix was subtracted to its corresponding instantaneous h_{TM}. This generates five vectors of orientational motion, mot_{TM} (one per helix), for every structure in the ensemble. Subsequently, for each structure in the ensemble I computed the angle formed between the two mot_{TM} vectors of every possible pair of helices. If the orientational motions of two helices generally occur in the same direction (they change their tilt in a correlated way), their mot_{TM} vectors will populate an angle close to zero, whereas if they are uncorrelated they are free to move independently, populating a wider range of angles.

Pore radius calculations

Pore radius were calculated by using the program MOLE 2.0.[301] In order to compute the Z-dependency of the pore radius, all structures in the ensemble were aligned using the backbone atoms of the transmembrane segments for the least-squares fitting.

3.2.2 Experimentally-driven protein dynamics of the PLN pentamer

The optimal setup found by the reference ensemble test uses 16 replicas in combination with internal averaging. This sampling scheme was subsequently employed to study the conformational dynamics of PLN in lipid bilayers using CSA and DC data experimentally measured in mechanically aligned DOPC:DOPE bilayers.[276] Simulations were carried out

under two different force fields, AMBER99SB-ILDN and CHARMM36. When these force fields are employed in unrestrained MD simulations, they result in divergent ensembles with considerably different structural and dynamical properties. However, the inclusion of experimental restraints drives the sampling toward unique distributions of structural and topological properties (Figure2.18e,f). The convergence of these structural proteins suggests that the ssNMR-restrained approach is able to overcome the intrinsic biases of individual force fields, therefore providing an accurate, experimentally-driven description of the structure and dynamics.

The restrained MD simulations were extended for 100 ns for each replica (total sampling time of 1.6 µs per force field). In the resulting ensembles, the PLN pentamer preserves its characteristic pinwheel topology, where the domain II (composed of the C-terminal transmembrane helices) is inserted in the membrane at an average crossing (τ) angle of 12.8° ± 3.17. The domain Ia, corresponding to the cytoplasmic region, retains its helical structure and remains adsorbed on the lipid bilayer, although it shows more topological variability than the domain II, with θ angles of 98.8° ± 8.58.

3.2.3 Phosphorylation leads to the dynamical activation of PLN

Phosphorylation of PLN by cAMP-dependent PKA plays a crucial role in the regulation of diastole, by relieving the inhibition of SERCA. In the monomeric state of PLN, the addition of the phosphate group on the serine residue at position 16 enhances the population of the excited R-state.[203] However, the effects of phosphorylation on the dynamics of the PLN pentamer have not been described in detail yet.

The ssNMR-restrained MD approach was here applied to the phosphorylated PLN pentamer, using CSA and DC measured in mechanically aligned bilayers.[292] Using the sampling approach described in Section 3.2.1 (and performing the simulations under the two different force fields), I obtained experimentally-driven ensembles with highly convergent structural properties and great agreement with the experimental data. These ensembles revealed a conformational activation of the pentamer upon phosphorylation, featuring an

increase in the dynamics of domains I and II. Particularly, in the transmembrane segment (domain II), the helices assume a much wider range of τ angles (Figure 3.4a).

The dynamics of the regulatory domain Ia undergo an activation on the nanosecond timescale, as evidenced by the back-calculation of the backbone root-mean square fluctuations (RMSF) and the backbone amide S^2 order parameters (Figure 3.4b,c - Section 3.2.1 for details on their calculation). The S^2 order parameters are in excellent agreement with ^{15}N NMR relaxation measurements that were not used to restrain the simulations,[292] which report a decrease in the order parameters of domain Ia (particularly at the phosphorylation site) but no effect on the transmembrane region. In contrast, when running unrestrained MD simulations these orthogonal experimental data are not well described, adding to the demonstration of the accuracy of this ssNMR-restrained MD approach and its ability to overcome force field biases and imperfections in studying membrane protein dynamics. The cytosolic helices also differ in their topological properties. The θ angle adopts higher values in pS16-PLN in comparison to the non-phosphorylated pentamer (98.8 and 104.0 for the non-phosphorylated and pS16-PLN, respectively), again showing convergence between the AMBER and CHARMM ensembles.

Despite the unaffected nanosecond dynamics of the transmembrane domain, its dynamical activation in terms of topological orientations is evident from the wider range of tilts adopted. Similarly to the backbone amide S^2, an order parameter can be used to describe the fluctuations of the transmembrane helical barrel in pentameric PLN. These fluctuations can be described as a cone of motion for the axis of the barrel (Figure 3.4d). This analysis reveals a drastic activation of PLN orientational dynamics which occurs at a longer timescale, extending to the millisecond, as probed by CSA and DC and therefore encoded in the experimental restraints.

The enhanced topological variability within the membrane of pS16-PLN does not disrupt the quaternary interfaces that pack the pentameric helical bundle. This is evidenced by the strong correlation in the orientational motions between the transmembrane segments of the individual protomers (Figure 3.4e - Section 3.2.1 for details on their calculation).

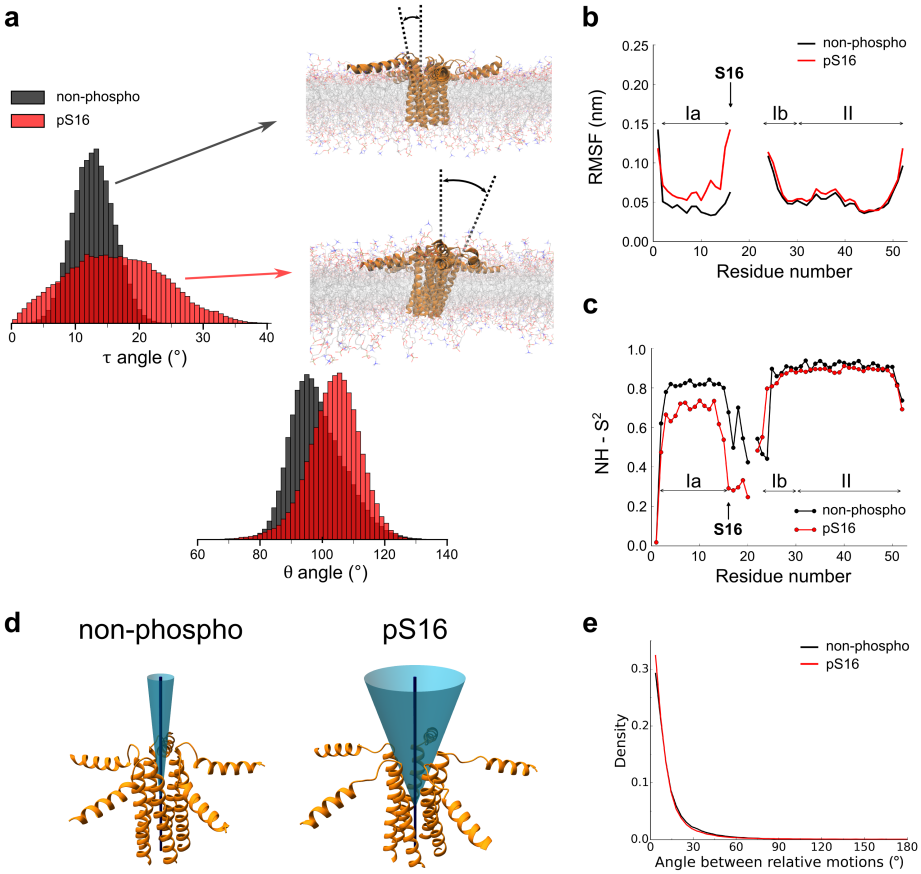


Figure 3.4: Dynamical activation of PLN upon phosphorylation – (a) Distributions of topological helical angles (τ and θ) of PLN in its non-phosphorylated (black) and pS16 (red) pentameric states. (b) Root mean square fluctuations (RMSF) of the C_α atoms in the two ensembles. In order to remove artifacts from inter-domain motions, the calculations were performed separately for domains I and II. (c) Back-calculated S^2 order parameters of amide N-H bonds in both ensembles. (d) Conical representation of the S^2 order parameter applied to the helical bundle axis (see methods in Section 3.2.1). The pentameric bundle axis populates a wider range of angles in the pS16 form, leading to a larger amplitude of orientational motion. (e) Correlation of relative helical motions between protomers in PLN (see methods in Section 3.2.1). The relative orientations of protomers are equally highly correlated in both ensembles, demonstrating that despite the large topological motions in pS16 PLN, the protomers re-orientate together without disrupting the transmembrane quaternary interface of the pentamer. Adapted from Sanz-Hernández *et al.*, *Sci Rep* (2016)[278] (licensed under CC BY 4.0).

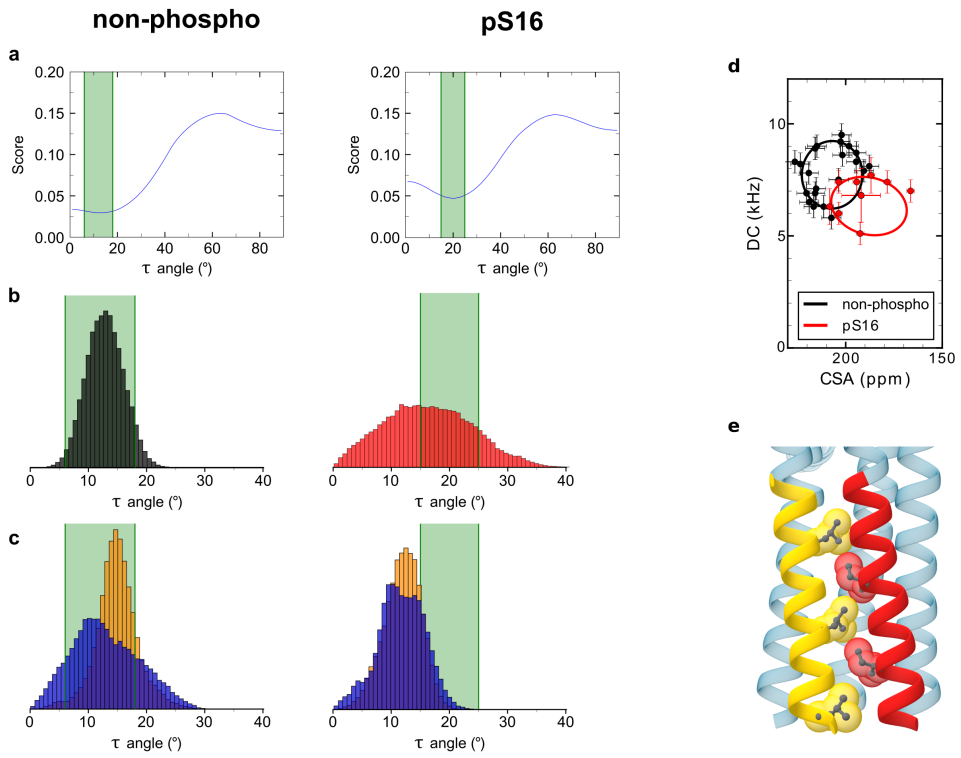


Figure 3.5: Static fitting of CSA and DC data to phospholamban – (a) Score of a static-based orientational fitting of the PISEMA data to PLN, where lower scores represent better fittings.[274] The area with the optimal τ angles according to the model is highlighted in green. (b) Distributions of τ angles extracted from the experimentally restrained ensembles, with the static best-fit area superposed in green. (c) Distributions of τ angles extracted from the unrestrained ensembles. (d) PISEMA spectral representation of the CSA and DC data for the transmembrane segments of pentameric PLN. (e) Depiction of the quaternary transmembrane leucine-isoleucine zipper that holds the pentameric assembly in PLN. Adapted from Sanz-Hernández *et al.*, *Sci Rep* (2016)[278] (licensed under CC BY 4.0).

I computed the average helical tilts and evaluated for each structure in the ensemble the tilt displacement of each helix from its average. Most helices move in a correlated manner, by tilting in the same direction and therefore forming an angle close to zero between their relative motions. This correlation emerges from a tight packing of the quaternary assembly, which is absolutely preserved upon phosphorylation despite the wide range of orientations assumed by the pentamer.

Overall, the ssNMR-restrained MD approach is able to provide a detailed dynamical description of the effects of phosphorylation in PLN. In contrast, the application of the classical PISEMA static fitting models of CSA and DC (See Figure 2.13) is unable to model the data adequately. In the case of non-phosphorylated PLN, the angles observed in our

ensemble, centered at 12.8°, are in very good agreement with the angle of 12° that would be estimated by fitting a static model[274] to the CSA and DC data (Figure 3.5a,b). When applying this model to the pS16-PLN CSA and DC, the best fit corresponds to a τ angle of 20°, which is incompatible with the correct packing of the PLN domain II helices. In fact, the quaternary interface is primarily stabilised by a leucine-isoleucine zipper (Figure 3.5e), which would be disrupted if the helices adopted such a high tilt. Given the inaccuracies of unrestrained MD in reproducing the experimental CSA and DC, and the inability of static models to maintain the structural topology of PLN, the dynamical considerations provided by the experimental ensembles here presented are the only possible way to account for differences in the CSA and DC data, which indeed show a shift toward their isotropic values upon phosphorylation (Figure 3.5d). This result particularly highlights the need to develop new integrative methodologies that take dynamics into account when interpreting complex ensemble-averaged experimental data.

3.2.4 Calcium channeling properties in pentameric PLN

In addition to serving as a storage mechanism as an interacting partner of SERCA, the pentameric state of PLN has been proposed to act as an ion channel, with different studies pointing at its ability to conduct calcium, chlorine and monovalent cations. The ssNMR-based structural characterisations of pentameric PLN both in its non-phosphorylated[276] and pS16 forms,[292] highlighted the difficulty for PLN to permeate ions given the tight packaging of its hydrophobic core at the interface between the five protomers. Given the dynamical considerations contained in the ssNMR-restrained MD approach, the experimental ensembles here generated can provide an accurate description of whether ion passage through the PLN pentamer is possible.

In our dynamical ensemble of non-phosphorylated PLN, the average radius of the pore formed by domain II is of 2.11 Å (Figure 3.6). The principal limiting points for ion passage through PLN are three hydrophobic bottlenecks formed by leucine residues at positions 37, 44 and 51 (with radii of 1.78, 1.54 and 1.81 Å respectively). Upon phosphory-

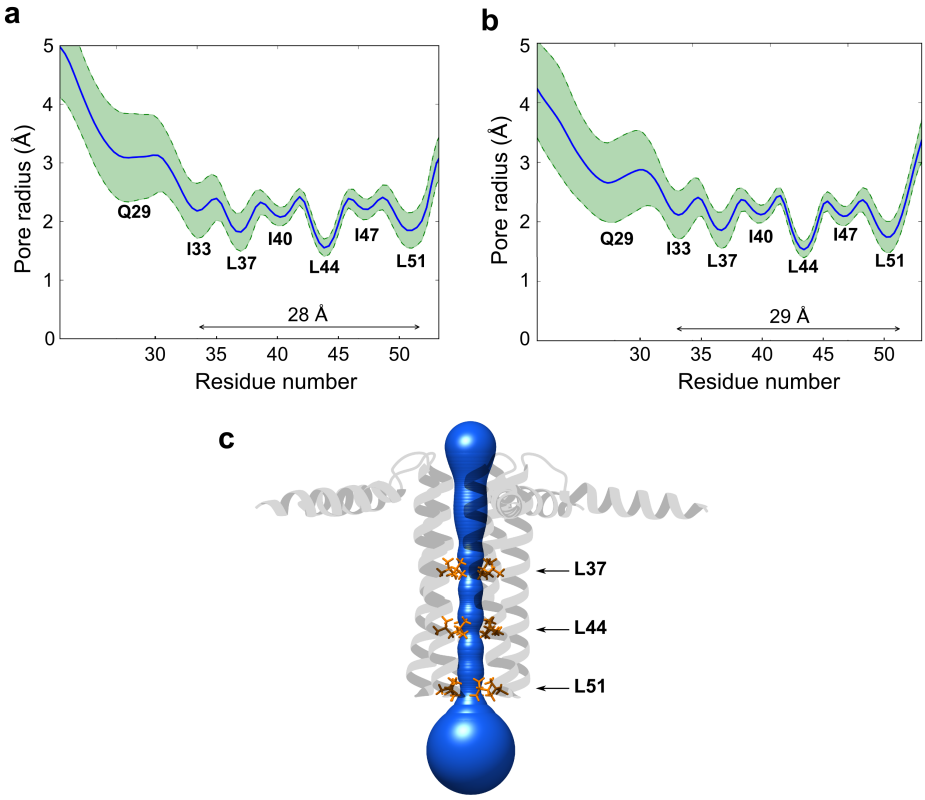


Figure 3.6: Channeling properties of pentameric phospholamban – (a) Channel profile of non-phosphorylated PLN. The blue line shows the average pore radius, with green dashed lines representing the standard deviations. The positions of the inward facing residues that constrict the pore width are labelled. (b) Channel profile of the pS16 PLN form. (c) Average pore radius in non-phosphorylated PLN overlayed onto the protein structure. The three primary bottlenecks for ion passage are shown in orange. Adapted from Sanz-Hernández *et al.*, *Sci Rep* (2016)[278] (licensed under CC BY 4.0).

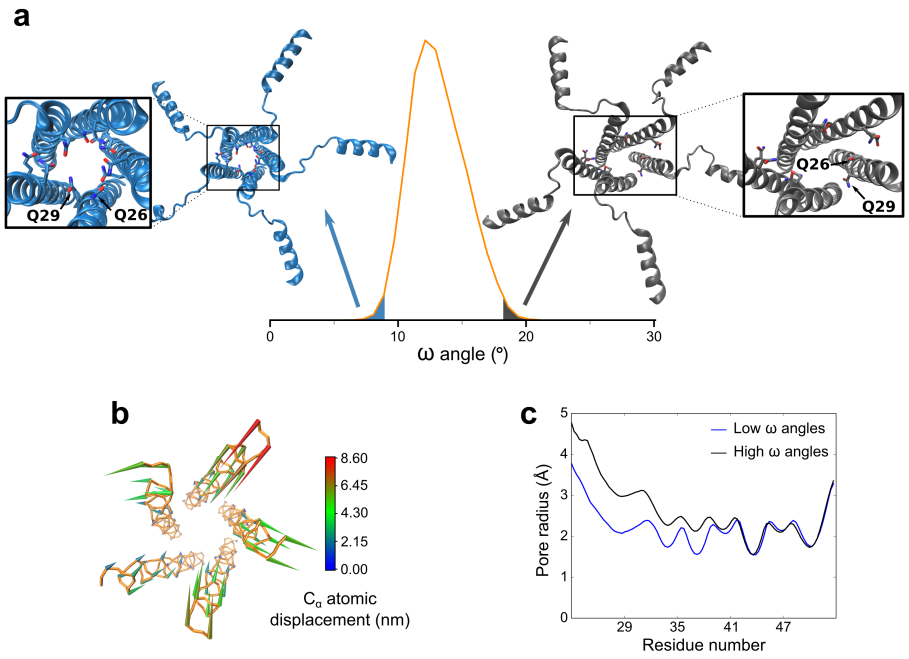


Figure 3.7: Internal transmembrane motions in phosphorylated phospholamban – (a) Distribution of the ω angle in pS16 PLN. Extreme low values are associated with a highly symmetrical disposition of the five helices, which adopt a circular shape in which Q26 and Q29 sidechains (shown as sticks) form quaternary contacts. High values are associated with a more distorted elliptical arrangement where the Q26/Q29 contacts are disrupted. (b) Porcupine plot depicting the backbone dynamics of the transmembrane helices along the first eigenvector of the covariance matrix of the ensemble. Cones illustrate the atomic displacement of C_α atoms. (c) Pore radius profile calculated for the structures at extreme values (top and bottom one-percentiles) of the distribution in (a). Adapted from Sanz-Hernández *et al.*, *Sci Rep* (2016)[278] (licensed under CC BY 4.0).

lation of serine-16, there is no evidence of any significant alteration to the profile of the pore, which retains a similar average radius of 2.09 Å and preserves all three bottlenecks.

With a dynamical description of PLN in hand, I searched for possible gating motions that could result in transient conformations relieving the impediments for ion passage. Principal component analysis of the trajectories of transmembrane helices was performed to identify the presence of any such fluctuations. The analysis is able to describe a collective motion in the domain Ib and an N-terminal segment of domain II, concerning residues 25 to 40. This portion of the helices experience some degree of twisting relative to each other (see Figure 3.7), in a motion that is observed both in non-phosphorylated and pS16 PLN. This collective motion can also be evidenced from evaluating the structures with extreme values of the ω angle (for its definition see Figure 2.18a), and appears to be linked to the

packing of the glutamine sidechains at positions 26 and 29 (Figure 3.7a). At low ω values, the interactions between these sidechains can form stably and this leads to an extension of the narrow hydrophobic pore, which develops an additional bottleneck at position 34. At high ω values, the glutamine sidechain interactions are disrupted and the region spanning residues 25 to 35 loosen their quaternary packing, relieving the first of the three hydrophobic bottlenecks. However, the pore along the C-terminal segment of domain II remains intact, and particularly the major bottlenecks at positions 44 and 51 are unaltered (Figure 3.7c).

Overall, the ensembles generated for pentameric PLN suggest that, despite the strong dynamical activation of domain II upon phosphorylation, the tight packaging of the quaternary assembly is fully retained, with the pentameric pore remaining too narrow and hydrophobic for favourable ion passage. Furthermore, these ensembles do not feature any large-scale fluctuations leading to transiently open conformations that may favour the passage of ions.

3.3 The dynamical complex between SERCA and phospholamban

Our collaborators in the Veglia group at the University of Minnesota have measured ssNMR data and performed a range of other experiments to characterise the SERCA-PLN complex. Given the success in applying the ssNMR-restrained approach to the study of the PLN pentamer, I set out to apply the same methodology to study the transient regulatory interaction between SERCA and PLN. In particular, this section will describe a sampling methodology focusing in the characterisation the B-state of the SERCA-PLN complex (see Figure 3.2). I will first describe the methodology employed to build the SERCA-PLN complex from different sources of experimental data, and the sampling strategy used to incorporate ssNMR and other experimental information. Later I will describe the ensembles that resulted from the sampling and the insights into the structural dynam-

ics of the regulatory complex.

3.3.1 Methods: Restrained simulations setup

Building the structure of the SERCA-PLN complex

A crystal structure of the SERCA-PLN complex was determined in 2013 by Akin and coworkers.[302] With an overall resolution of 2.8 Å, it shows that PLN locks SERCA into the E2 apo state of its enzymatic cycle, with no bound calcium nor ATP. Despite the overall good quality of the structure, the interface between SERCA and PLN is not very well defined in the structure, which only has clear electron density for the backbone of domain II in PLN. The absence of a clear definition for sidechains of PLN and the presence of density for only half of PLN may have led to a limited model of the SERCA-PLN interface in this structure. Despite these problems, this structure was chosen as a starting template to build the complex, as it is the only available experimental structural study of it. The conformation of SERCA was directly taken from the crystal structure (PDB code 4Y3U). Missing loops were introduced using the MODELLER software[303] based on other available crystal structures of SERCA alone.

The interface between SERCA and PLN was built *in silico*, based on CSA, DC and cross-linking data measured by collaborators in the Veglia group, in Minneapolis. The experimental construct used in their experiments was the monomeric species PLN^{AFA}, which was therefore used for the model. The transmembrane section of PLN^{AFA} was docked onto SERCA *in vacuo*, maintaining backbone positional restraints on the pump and restraining the dihedral angles of PLN^{AFA} to retain the helical structure. The docking was guided by the cross-linking data, by performing short MD runs (500 ps) where a harmonic upper wall potential is applied to restrain the distances between cross-linked residues to below 5 Å. 100 such runs were performed starting from different orientations of PLN^{AFA}, and the resulting docked complexes were clustered according to the backbone RMSD of the proteins. The center of the most highly populated cluster was picked as the most rep-

resentative structure, and used to carry on with the modelling. The interface between SERCA and PLN^{AFA} in the modelled complex was similar to the one found in the crystal structure. There are some modifications in the sidechain stacking between SERCA and PLN^{AFA}, with PLN^{AFA} shifted higher on the Z-axis in our model, toward the cytosolic side of SERCA.

The membrane model employed in this simulation was a DOPC:DOPE 1:4 mixture (307 DOPC molecules and 76 DOPE molecules), mimicking the experimental setup. An important consideration is the orientation of the transmembrane segment of SERCA and PLN^{AFA} with respect to the lipid bilayer. A recent study by Norimatsu and co-workers produced crystal structures of SERCA in lipid bilayers in 4 different states along its enzymatic cycle[53] (Figure 1.2). Since some lipids are visible in these structures, the orientation of SERCA in the membrane can be modelled from them. The transmembrane segment of SERCA in the PLN^{AFA}-docked model was aligned to the structure that is most representative of the E2 enzymatic state (PDB code 2AGV), and the membrane model was built following its orientation. The resulting orientation of PLN^{AFA} was in good agreement with the tilt angle expected from statically fitting the transmembrane CSA and DC, resulting in a crossing degree of 32°.

The resulting membrane model of SERCA and the transmembrane segment of PLN^{AFA} was equilibrated thoroughly using a coarse-grained description of the system, provided by the MARTINI model.[117] The position of all protein particles was restrained, allowing the lipid bilayer and solvent to equilibrate for 500 ns. The resulting system was re-converted into a full atomic description using the Backward MARTINI approach.[304]

Finally, the dynamical N-terminus of PLN was modelled as a random coil, completely detached from SERCA and the lipid bilayer. Eight different randomised conformations of the disordered terminus were generated to serve as different starting structures for the independent replicas. The system was fully solvated in 61,423 Tip3p waters, resulting in a total of 252,806 atoms in a tetragonal box of dimensions $12.3 \times 12.3 \times 16.8$ nm^3. All 8 structures were further equilibrated for 20 ns in the NPT ensemble with positional re-

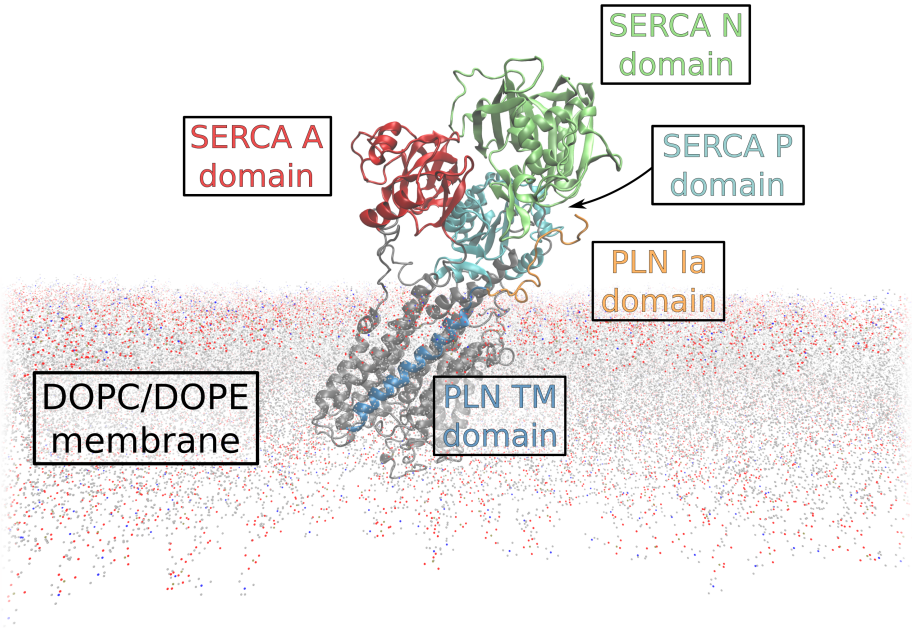


Figure 3.8: Structure of the PLN-SERCA complex in the B-state – The complex between PLN^{AFA} and SERCA was built using the crystal structure 4Y3U as a template. Distances from crosslinking and PRE experiments were incorporated to generate a model of the complex in the B state, where the cytoplasmic domain of PLN adopts a disordered conformation bound to the ATPase.

straints on their heavy atoms. Positional restraints were then released for another 20 ns of further equilibration at a temperature of 300 K, where experimental CSA and DC restraints were gradually incorporated for the transmembrane segment of PLN, with respective force constants of 50 J/(mol×ppm^2) and 800 J/(mol×kHz^2). This resulted in the 8 initial conformations employed in the replica-averaged ssNMR-restrained simulations. A snapshot of the resulting PLN-SERCA complex is showin in Figure 3.8.

Experimentally restrained simulations of the PLN-SERCA complex

The resulting PLN-SERCA complex structures feature a disordered domain Ia of PLN bound to the SERCA cytoplasmic domains, corresponding to the B-state of the complex (see Figure 3.2). In order to characterise the dynamical nature of this complex, I applied the restrained MD simulation approach, using a combination of data measured by the Veglia group in Minnesota from different experimental sources. Specifically, there are available experimental CSA and DC for the PLN transmembrane segment, crosslink-

ing information between the PLN and SERCA transmembrane regions and paramagnetic relaxation enhancement (PRE) measurements between the N-terminal region of PLN and the SERCA cytoplasmic domains.

The CSA and DC were incorporated into the simulation as described in Section 2.3.2, averaging them over 8 parallel replicas. The crosslinking and PRE experiments provide information that can be translated into interatomic distances. In crosslinking experiments, cysteine probes in specific parts of the protein form disulphide bridges with cysteines in close proximity of their interacting partner. This generally occurs only if the two cysteine residues are in close proximity, and the resulting distances are therefore treated as short-ranged (<6 Å) with fluctuations in the order of a few Angstrom. In PRE measurements, a paramagnetic label is introduced at a specific position of a protein. This label quenches the NMR signal of nearby nuclear spins, causing the disappearance of peaks from residues in close proximity. The signal-depletion effect of PREs is however observed at a much higher distances and over a wider range of distances (commonly beyond 15 Å). In the PLN-SERCA complex, PRE measurements were available from two experimental sources. Firstly, PLN was isotopically labelled with ^{13}C and ^{15}N, and paramagnetic probes introduced in SERCA bound to exposed cysteine residues. Secondly, ^{13}C isotopic methyl labels were crossed-linked to exposed cysteines in SERCA, and the paramagnetic probe was introduced in the PLN domain Ia.

Taken together, all these sources of distance information resulted in 21 PRE-derived and 7 crosslinking distance restraints for the ensemble calculations (shown in Tables 3.2 and 3.3). Distance restraints were introduced as flat-bottomed potentials, as previously described for ssNMR restraints, but were not replica-averaged in order to avoid overfitting issues often reported when employing distance measurements in MD simulations.[305] In this study, crosslinking distances were set to 4.6 Å ± 3.0 Å, and PRE-derived distances at 15 Å ± 15 Å.

The interaction between SERCA and PLN in the B-state is mediated by the fully disordered domain Ia of PLN. Given the transient and heterogeneous nature of this interac-

Paramagnetic Relaxation Enhancement	
PLN Residue	SERCA Residue
A0	C364
A0	C471
A0	C498
A0	C561
A0	C636
A0	C674
K3	C674
K3	K397
V4	C674
Q5	C674
Y6	C674
L7	C364
L7	C471
L7	C498
L7	C561
L7	C636
L7	C674
T8	C674
S10	C674
A11	C674
I12	C674

Table 3.2: Experimental PRE-derived distances for the SERCA-PLN complex in its B-state, measured by the Veglia group in Minneapolis.

Crosslinking	
PLN Residue	SERCA Residue
Q23	K328
N27	L321
N30	L321
I48	V89
V49	V89
L51	V89
L52	V89

Table 3.3: Experimental crosslinking-derived distances for the SERCA-PLN complex in its B-state, measured by the Veglia group in Minneapolis.

tion, there is a very large conformational sampling space for the simulations to explore. With the current computing power capabilities, a standard method of sequential restrained MD may prove intractable for such a large system. In order to overcome this hurdle, an approach based on annealing cycles was implemented, whereby the PLN N-terminus periodically binds and detaches from SERCA, incorporating the experimental distances to drive each binding event.

In more detail, at the beginning of each cycle the domain Ia of PLN is fully detached from SERCA by introducing a lower-wall harmonic potential that pushes PLN residues away from several SERCA residues present in the cytoplasmic domains of the pump. Specifically, the lower wall potential pushes PLN residues 1 to 10 at least 50 Å away from I140 (SERCA P domain), T430 (SERCA N domain) and C674 (SERCA A domain). Residues 11 to 14 are pushed at least 25 Å away from those residues and in addition the interfacial SERCA residues T742, A327 and L119. The detachment occurs gradually, by linearly increasing the force of the lower-wall potential to its maximum after 500 ps. During this time, the temperature is also linearly increased to 370 K to enhance the conformational sampling space of the disordered domain Ia. CSA and DC force constants are linearly decreased to half their value to avoid instabilities during this high-energy phase of the cycle. This detachment step yields a fully detached PLN domain Ia with no SERCA contacts and enough surrounding free space to explore its unbound disordered conformational space.

The detachment step is followed by another 500 ps of sampling at 370 K, where the detached PLN domain Ia is allowed to fluctuate. Subsequently, the lower wall potential is linearly removed over 1 ns of sampling. The PRE-derived distance restraints are gradually incorporated over this time, and the force constants of CSA and DC restraints restored to their full value. The temperature is also linearly decreased to 300 K. After the binding event, the PLN N-terminus adopts a SERCA-bound conformation. This step leads onto the sampling phase of the cycle, where the restraint forces and the temperature are kept constant for 2 ns. The structures sampled during these 2 ns per cycle are the ones included in the final experimentally-driven SERCA-PLN B-state conformational ensemble.

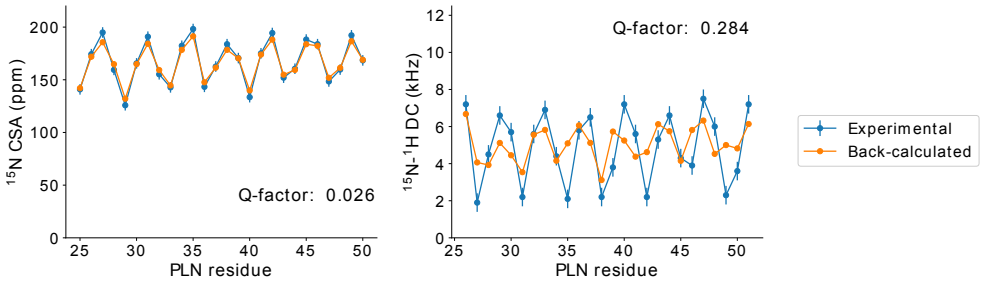


Figure 3.9: Q-factors for the ensemble of PLN in complex with SERCA – Agreement between the experimental and back-calculated CSA and DC of the transmembrane segment of PLN when bound to SERCA.

The described annealing strategy results in cycles of 4 ns in length, with a effective sampling time of 2 ns per cycle. To generate the final ensemble a total of 33 cycles were completed for each of the 8 parallel replicas, resulting in a total simulation time of 1,056 ns. This translates into 528 ns of sampling of the B-state, from which 10,000 equally spaced structures were extracted to form the final conformational ensemble.

3.3.2 Restrained simulations reproduce experimental results

The generated B-state conformational ensemble of the PLN-SERCA complex incorporates experimental information from a variety of sources: CSA and DC from oriented ssNMR measurements, and distances from PRE and crosslinking experiments. Overall, the agreement between the experimental and back-calculated ssNMR observables is very good (see Figure 3.9), with an excellent Q-factor of 0.026 in the case of CSA and 0.284 for the DCs. The agreement in the latter observable shows room for improvement, and the simulation could perhaps tolerate a higher force constant for the DC restraints.

The good agreement between experimental and simulated ssNMR variables is indicative of an accurate reproduction of the transmembrane topology of SERCA-bound PLN. In the previously characterised structure of free monomeric PLN, the tilt angle of the transmembrane helix is centered at 24°. When bound to SERCA, the tilt of the helix increases considerably, with its distribution centered at an average value of 31.3° ± 6.08 (see Figure 3.10). This distribution appears to be unimodal and largely symmetrical, sug-

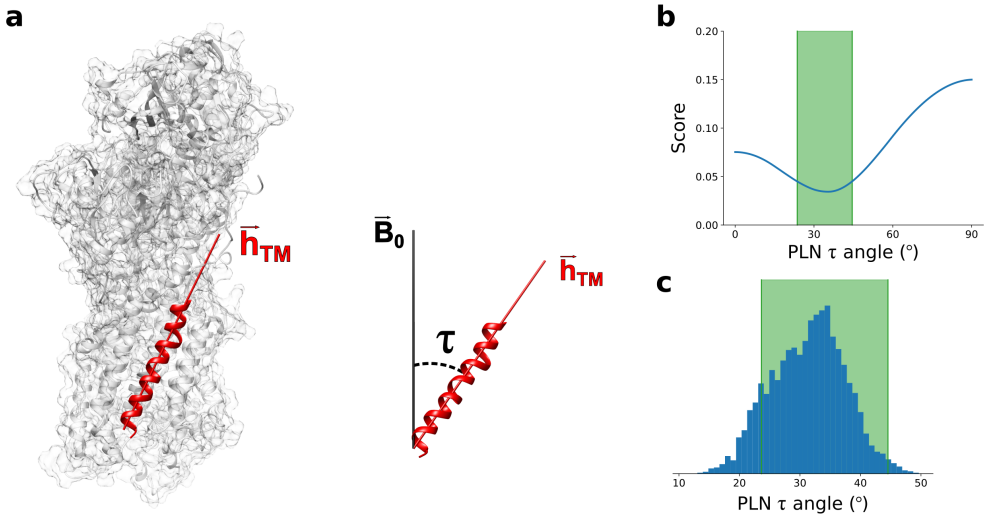


Figure 3.10: Distribution of transmembrane tilt angles of PLN in complex with SERCA – (a) Depiction of the PLN τ angle when bound to the SERCA pump. (b) Static fitting model of the PISEMA data applied to the SERCA-bound CSA and DC, with lower scores representing a better fit. The best-fit area is shaded in green. (c) Distribution of τ angles in the SERCA-bound PLN ensemble, with the static best-fit area overlayed in green.

gesting that PLN fluctuates about a single tilt value when bound to SERCA. The angles are in good agreement with the best-fit values described by classic static fitting models of the PISEMA data, as was the case for the non-phosphorylated PLN pentamer (see Section 3.2.2 and Figure 3.5). However, the distribution of transmembrane tilt angles is considerably wider when PLN is bound to SERCA than in the non-phosphorylated pentameric assembly, with the standard deviation approximately doubling, from 3.17° in the pentameric case to 6.08° in the SERCA-bound complex. This reflects on an enhanced level of orientational fluctuations in PLN when bound to the enzymatic pump.

In addition to the data used to restrain the MD simulations, an additional source of experimental information is available for the complex, in the form of isotropic chemical shifts for the domain Ia of PLN. As described in Section 2.1, chemical shifts are very sensitive probes of an atom's local environment and its secondary structure, so they can be accurately employed as orthogonal validation data for the present ensemble of the PLN-SERCA complex. Specifically, the chemical shifts of domain Ia should highly differ between the disordered B-state conformation and the ground PLN T-state, where domain Ia adopts a helical fold. The predictor SPARTA+[229] was used to back-calculate the aver-

age isotropic chemical shifts of domain Ia in the ensemble, and compared to their experimental values (Figure 3.11). The results show excellent agreement between experimental and back-calculated chemical shifts, with RMSDs of 1.01, 1.05 and 0.97 ppm for C_α, C_β and C_O respectively, which are below the average error generally incurred by the predictor (as shown in Section 2.2.4 and Figure 2.8). By contrast, when SPARTA+ is employed to calculate the chemical shifts corresponding to the helical domain Ia in its T-state, the divergence with the experimental values increases for all atoms, resulting in RMSDs of 2.39, 1.47 and 1.27 ppm for C_α, C_β and C_O respectively, which are all above the average error of the predictor. This result highlights the accurate reproduction of the disordered nature of domain Ia of PLN, which populates a highly heterogeneous collection of states.

3.3.3 Transient disorder-mediated interaction between SERCA and PLN

The regulation of the SERCA pump by PLN is multi-step and complex. It involves a variety of different structural states, which differ in the activity levels of SERCA and lead to a fine tuning of the balance of calcium concentration in cardiac myocites.[54] PLN normally acts as an inhibitor of SERCA when bound to it, but in its B-state, PLN is able to revert the inhibition and restore the ATPase activity. The structural ensemble for the PLN-SERCA B-state reveals that the binding between SERCA and PLN is mediated by a highly heterogeneous interaction at the cytoplasmic side of the membrane, where the disordered domain Ia of PLN establishes transient contacts with many different residues along SERCA's exposed surface.

Overall, the domain Ia of PLN establishes contacts over a very wide interaction surface of SERCA, primarily clustering over domains N and P (Figure 3.12). The disordered N-terminus of PLN extends away from the lipid membrane and samples a large variety of sites in the cytoplasmic region of SERCA, with some contacts being more prevalent than others. In particular, there is a primary surface cleft of interaction formed by flexible loops in SERCA, located at positions 383 to 388 (in domain N), 551 to 559 (in domain

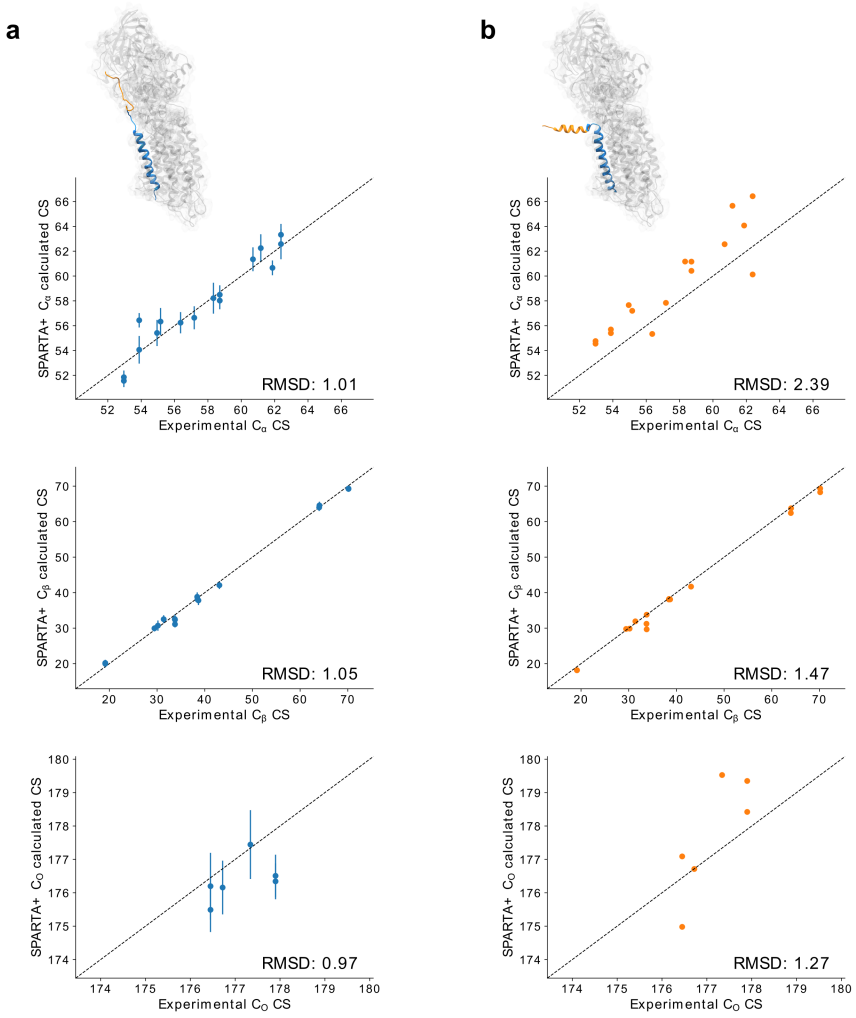


Figure 3.11: Isotropic chemical shifts of SERCA-bound PLN – (a) Comparison between experimental and back-calculated isotropic chemical shifts (using SPARTA+[229]) of C_α, C_β and C_O atoms in the B-state restrained ensemble of the PLN-SERCA complex. (b) The equivalent comparison using a structure of the SERCA-PLN complex in the T-state, with domain Ia of PLN folded into a helical conformation.

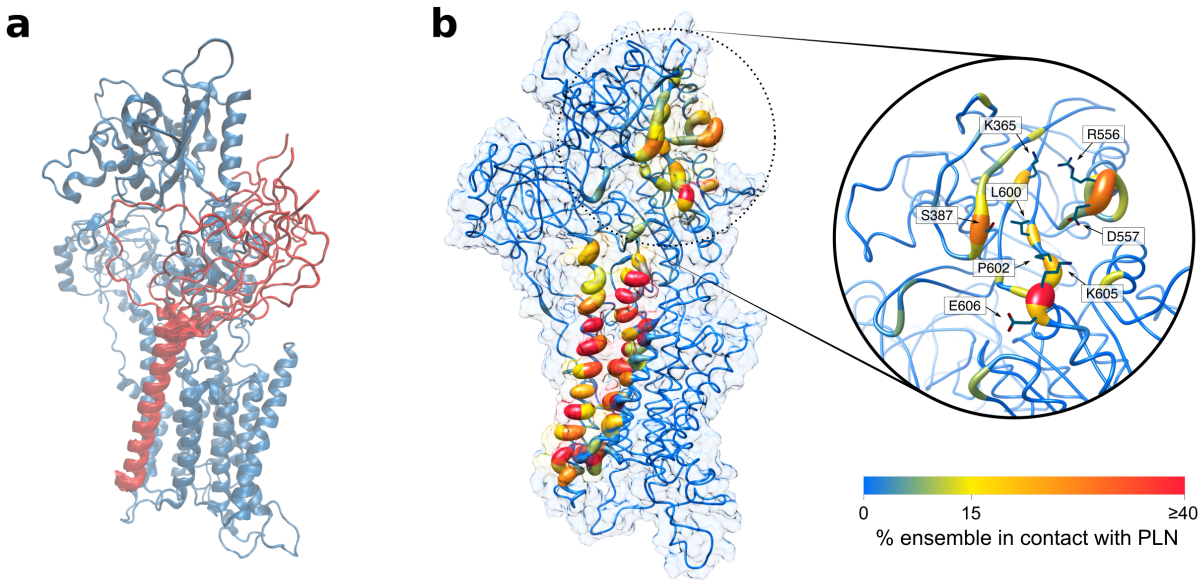


Figure 3.12: Dynamical cytoplasmic interaction between SERCA and PLN – (a) Overlay of distinct PLN structures populated during the simulation. Domain Ia populates a wide variety of disordered conformations, interacting with multiple regions of the SERCA cytoplasmic domains. (b) Interface of the PLN-SERCA complex mapped onto the SERCA structure. Residues are coloured according to the percentage of structures in the ensemble that exhibit at least one heavy-atom contact (< 5 Å) between the two proteins. The residues with the most prevalent contacts in the SERCA cytoplasmic domains are labelled.

P), 600 to 606 and 358 to 365 (these two latter loops link domains P and N). Figure 3.12b depicts the main SERCA residues involved in atomic contacts (<5 Å) with PLN, and the prevalence of those contacts across the ensemble. Notably, the most prevalent cytoplasmic contacts are present in 40% or less of ensemble structures, highlighting the transient nature of the interactions sampled by PLN.

Interestingly, 5 out of the 8 most contacted residues in SERCA are charged (K605, R556, D557, E606 and K365), a suggestion that the regulatory interaction is governed by electrostatic forces between the two membrane proteins. In order to evaluate this possibility in detail, the Coulomb energy contributions between every pair of residues in the complex was evaluated (see Figure 3.13). This analysis reveals a large charged surface of interaction on SERCA, characterised by different electrostatic hotspots available for PLN to interact with. In domain Ia of PLN, all charged residues establish some electrostatic interactions with these patches. More specifically, the positive sidechains of PLN exhibit a large degree of promiscuity in establishing interactions with different negative patches of

SERCA.

The R13/R14 pair exhibit strong interactions with two negative patches: one formed by residues E606/D737/D738 and another by residues E392/E394. R9 interacts with up to four negatively charged sites: the two previous one as well as two other formed by D399 and D557 respectively. Both these positions in PLN are crucially involved in mutations associated with familial lethal cardiomiopathies. The mutation R9C results in a loss of function, with PLN heavily reducing its ability to inhibit SERCA.[306] On the other hand, R14del completely supresses SERCA by locking PLN into a super-inhibitory conformation that cannot be relieved.[307] Interestingly, the interaction between the R13/R14 pair and E606 is the most prominent electrostatic contact in the ensemble. A mutation that disrupts this contact would heavily affect the shape and complementarity of the interaction surface, having potential knock-on effects in shifting the fine equilibrium between T-, R- and B-states of the SERCA-PLN complex (see Figure 3.2), and thus altering the overall activity of SERCA.

From the SERCA side, the negative E606 is the residue with the highest involvement in electrostatic interactions with PLN. Interestingly, E606 is flanked by positively-charged residue K605. This lysine residue is one of the SERCA positions that features the highest prevalence of atomic contacts throughout the ensemble (see Figure 3.12b), establishing regular contacts with R9 and R13/R14 in PLN. Taken together, these results point toward a mechanism of electrostatic counter-balancing by SERCA, whereby strong electrostatic interactions may be disrupted by nearby opposing charges in order to enhance the conformational variability of the protein complex.

Both R9 and the R13/R14 pair exhibit heterogeneity in their electrostatic interactions across the SERCA surface (see Figure 3.14). R13/R14 are primarily in contact with the largest negative patch on the surface, formed by the sidechains of SERCA residues E606, D737 and D738. A second patch is also accessible to R13/R14, formed by residues E392 and E394, which is contacted considerably less frequently (Figure 3.14b). R9 shows even more heterogeneity, as it is able to reach further onto the cytoplasmic side of SERCA,

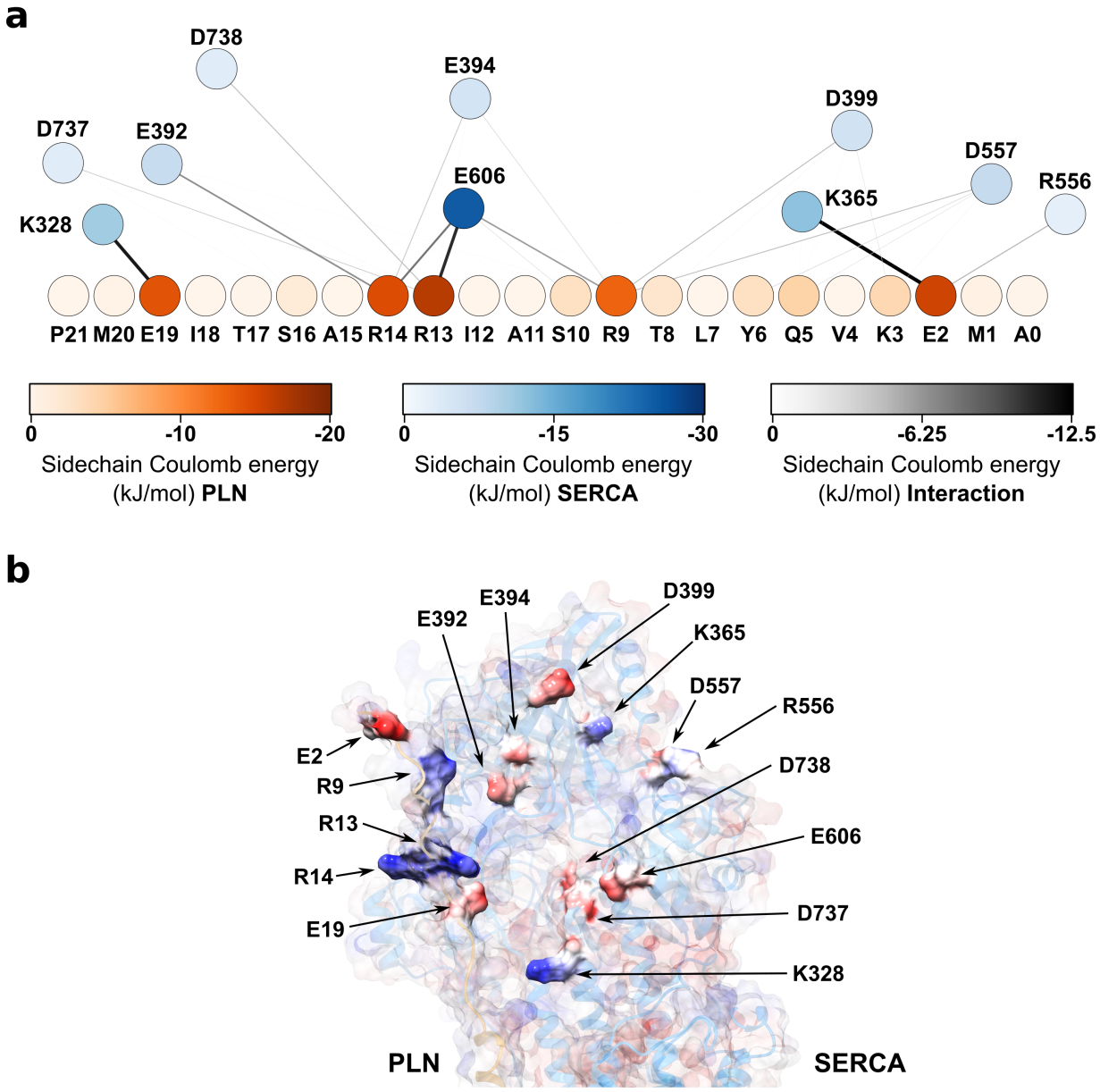


Figure 3.13: Electrostatic interaction network between SERCA and PLN – (a) Network of electrostatic contacts between the cytoplasmic domains of PLN (orange) and SERCA (blue). The intensity of the node colours corresponds to the residue's average sidechain Coulomb energy computed in the ensemble. Edges (in grayscale) are drawn between any two residues in electrostatic contact, with the color intensity corresponding to the interaction's average Coulomb energy. (b) Map of the electrostatic surface of the complex, with labels showing the exposed hotspots in PLN and SERCA.

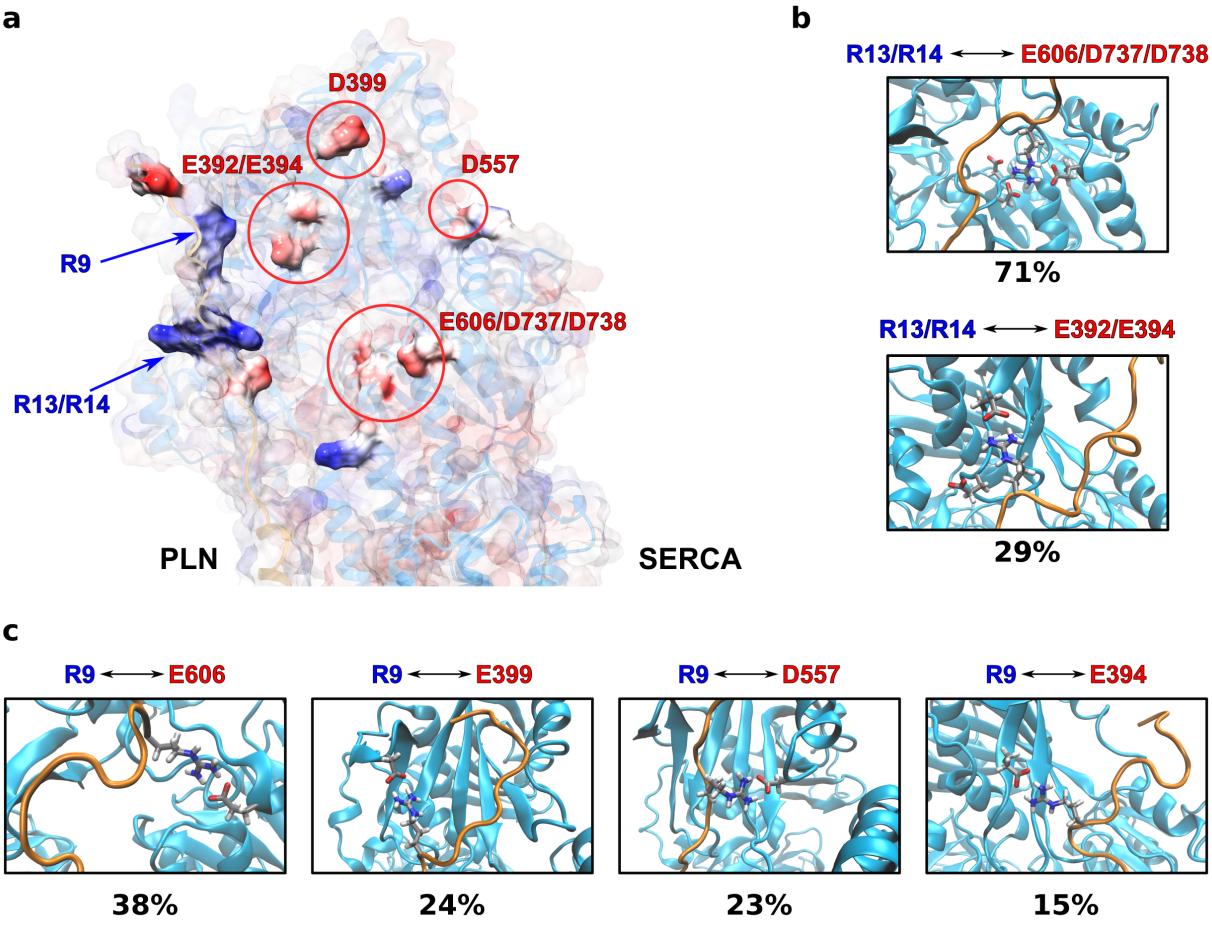


Figure 3.14: Diversity in the electrostatic interactions of R9 and R13/R14 with SERCA – (a) Representation of the three arginine residues in the domain Ia of PLN and the four negatively charged patches exposed in the cytoplasmic side of SERCA. (b) Electrostatic interactions of the R13/R14 pair are split throughout the ensemble between two negative clusters, with representative structures shown for each. Percentages for each structure correspond to the fraction of Coulomb sidechain energy contributed by each interaction. (c) Representative structures of the four electrostatic contacts established by R9.

where two additional negative patches rest (one formed by residue D399 and the other by residue D557). R9 still interacts primarily with the E606 patch, but has a more even split of electrostatic interactions over the four accessible negatively charged sites (Figure 3.14c).

Overall, the experimentally guided structural ensemble here described for the B-state of the SERCA-PLN complex reveals the very transient and heterogeneous nature of the cytoplasmic interaction between both proteins. It is known that by binding SERCA in its B-state, PLN crucially restores the activity of SERCA. The transient interactions here described must therefore act as a trigger for allosteric transitions that restore the pump's ability to transport Ca^{2+} ions. These results point toward the importance of electrostatic forces in driving the cytoplasmic interplay between the two proteins, and pinpoint the atomic detail of certain contacts that are disrupted in pathological variants of PLN.

Chapter 4

Dynamical studies of the influenza M2 tetramer

4.1 The role of M2 in influenza

Influenza viruses are negative-sense RNA viruses whose virion particles are enclosed by lipid bilayers. The most virulent form of the virus is classed as influenza A, and has been involved in a number of human and avian pandemics.[308] The viral genome is composed of 8 RNA segments encoding 11 genes. Among these, there are three integral membrane proteins: hemagglutinin (HA), neuraminidase (NA) and the matrix 2 protein (M2). HA is responsible for the virus binding to the host cell, whereas NA plays a crucial role in viral release.[309]

The third integral membrane protein, M2, is a 97-residue tetrameric protein that acts primarily as a proton channel. It has an essential role in the viral infectious cycle, by acidifying the virion's interior once it has been contained in an endosome. Proton transport by M2 is triggered by the lower pH inside the endosome, and the acidification of the viral interior leads to membrane fusion with the endosome and the release of viral RNA, leading to the infection of the host cell.[310] The activity of M2 has been therapeutically targeted

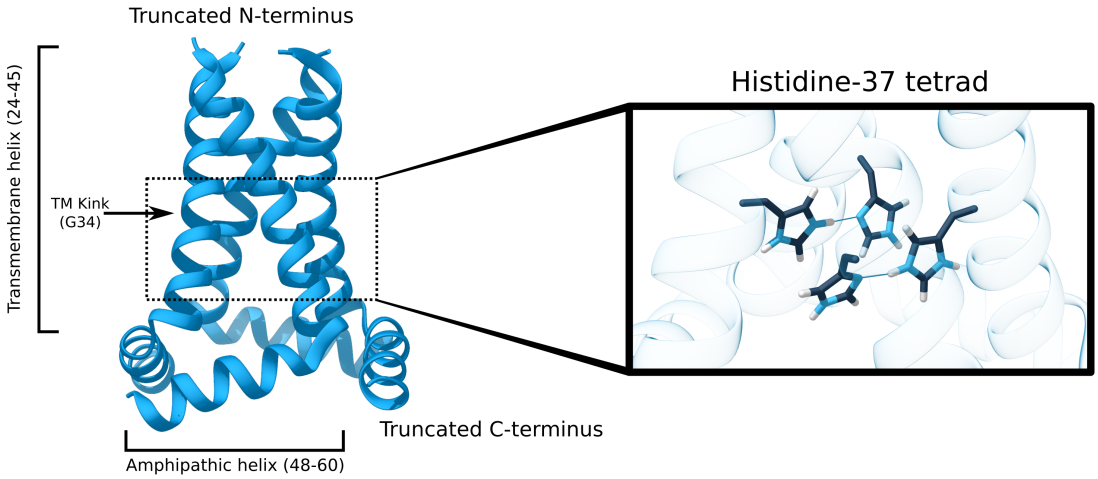


Figure 4.1: Structure of the M2 influenza A proton channel in lipid bilayers – This structure of M2 (PDB accession code: 2L0J) was determined by solid-state NMR in aligned DOPC/DOPE lipid bilayer preparations at pH 7.5,[318] employing chemical shift anisotropy and dipolar couplings as structural restraints. Topologically, the tetrameric structure features a transmembrane helix, with a substantial kink at position G34, and an amphipathic helix adsorbed on the membrane. The histidine tetrad at position 37, responsible for gating and proton transport, was found to be asymmetrically protonated, with two out of four possible sites occupied.

by the anti-viral compound amantadine,[311] but the success of this drug has been very limited due to the fast development of resistant variants of the virus.[312]

Given the importance of M2 in the infectious cycle, many structural efforts have been directed at understanding the molecular basis of proton transport. M2 has been characterised structurally in different instances, by X-ray crystallography,[313,314] solution NMR[315] and solid-state NMR.[316–319] Most of these structures focus on the transmembrane segment of M2 (residues 25 to 46). One of the solid-state NMR structures, determined in aligned lipid bilayer samples,[318] features a longer sequence (22-62), and reveals the presence of another set of shorter helices of amphipathic nature, which sit on the surface of the membrane on the virion interior (See Figure 4.1). These shorter helices have been shown to be important in inducing membrane curvature during the viral budding process, providing an independent membrane scission mechanism that does not rely on the host cell's endosomal sorting complex.[320]

Within the transmembrane domain, a highly conserved histidine residue, at position 37, is essential for pH-dependent channel gating and selective proton conductance. An additionally conserved tryptophan residue is located at position 41, forming a characteristic

motif of the channel commonly referred to as HxxxW.[321] The H37 tetrad has experimentally been shown to be specifically responsible for proton transport,[322] but the mechanism of transport has been subject of debate. A so-called *shutter* mechanism postulates that at low pH all the histidine residues become protonated, leading to a pore opening driven by repulsion through which protons diffuse.[323,324] Alternatively, a *shuttle* mechanism has been proposed, whereby the imidazole ring reacts with nearby protons and releases them on the other side of the pore.[319,325]

The dynamical process of proton transport by M2 has also been heavily addressed computationally, in studies of molecular dynamics[324,326–329] and methods incorporating quantum considerations to simulate proton reactive transfer.[330–332] The availability of oriented ssNMR data allows a further investigation in this direction, through the application of CSA and DC-restrained simulations to this system.

The aim of studying the dynamics of M2 with the ssNMR-restrained MD is to achieve an accurate characterisation of the orientational dynamics in M2. The topological fluctuations in the helices within the membrane are likely to play an important role in governing the channel's mechanism of conductance. In this chapter I present the results from ssNMR-restrained simulations in the closed conformation of M2, as determined in DOPC/DOPE lipid bilayers.[318]

4.2 Methods: Simulation setup and ensemble analysis

4.2.1 Restrained simulations setup

Simulations of the M2 tetramer were started from an ssNMR structure spanning residues 22 to 62[318] (PDB code 2L0J). The structures were embedded in a DOPC:DOPE bilayer mimicking the 4:1 ratio employed in the experiments (175 DOPC and 48 DOPE molecules). The protein-lipid system was solvated with 9,998 Tip3p water molecules and 10 Cl^- ions,

resulting in a total of 63,128 atoms. The simulation box was equilibrated for 100 ns with positional restraints on the protein heavy atoms, coupling temperature at 300 K with the V-rescale method[297] and pressure with the Berendsen algorithm.[296] The system was simulated under two different force fields, namely AMBER99SB-ILDN[279] and CHARMM36.[126] The Particle mesh Ewald method[299] was used to describe electrostatic interactions, and LINCS was used in defining the constraints.[298] The integration timestep was 2 fs in all simulations.

After the equilibration phase, the structure was further simulated for 75 ns under each of the two force fields, and a set of 8 structures was extracted to serve as a starting point for the replica-averaged simulations. Unrestrained samplings were carried out for an additional 100 ns. The restrained simulations were averaged among 8 parallel replicas as well as internally among the four protomers that compose the tetrameric assembly. CSA and DC restraints were incorporated by linearly increasing their force constants over a force equilibration phase of 20 ns per replica, reaching final equilibrium values of 12.5 J/(mol×ppm^2) and 625 J/(mol×kHz^2) for CSA and DC respectively. The restrained samplings were extended for 80 ns per replica. Overall, the total sampling time resulted in 1.8 µs.

4.2.2 Structural and hydration analyses

An ensemble of 10,000 equally spaced structures was extracted from each of the two simulations under different force fields. Topological properties of these ensembles were described in the form of helical tilt angles, computed as described in Section 3.2.1. The agreement between back-calculated and experimental CSA and DC is expressed in terms of Q-factors, also computed as described in Section 3.2.1. Calculations of the internal pore radii were performed using the software HOLE.[333]

Hydration analyses of the ensembles were performed by building solvent density maps, according to previously described methods.[334,335] Briefly, the method proceeds by partitioning the space around the protein into a solvation shell (up to 6 Å from the protein

surface) and an outer shell corresponding to the bulk density (6 to 8 Å away from the surface). A grid is then build with a 0.5 Å spacing and the solvent density evaluated at every point based on the oxygen atom positions across the ensemble. All M2 structures were superposed onto the transmembrane helical segment (residues 24 to 45) to remove protein rotation and translation effects. A C_α-RMSD cutoff of 2 Å from the original ssNMR structure was employed, analysing only structures below this similarity threshold in order to eliminate artefacts coming from the protein backbone flexibility.

4.3 Topological properties of the M2 tetramer

Using CSA and DC measured in aligned samples of lipid bilayers as restraints, the structural ensembles of the M2 tetrameric channel were sampled using two different force fields, CHARMM and AMBER-based. Due to the oligomeric nature of the protein, the restraints were applied using an replica and internally-averaged approach, as successfully calibrated and validated for pentameric phospholamban (see Section 2.3.2). The restrained samplings were able to produce structural ensembles with very high agreements with the experimental data, as shown in Figure 4.2. Q-factors are improved with respect to the unrestrained samples, reaching values of 0.058 (0.060) and 0.165 (0.177) for CSA and DC respectively (values in brackets correspond to the AMBER ensemble).

Aside from improving the values of the ssNMR observables, the effect of the experimental restraints is also observed in the description of topological properties of the ensemble. When evaluating the helical crossing angles in the structure, the unrestrained ensembles of AMBER and CHARMM simulations feature divergent distributions, which derive from the different characteristics of the two force fields. As was shown in the case of phospholamban, the introduction of the restraints is able to overcome the individual biases of force fields and provide a more accurate, experimentally-driven description of these orientational properties, as evidenced by the convergence toward single distributions of topological angles (Figure 4.3).

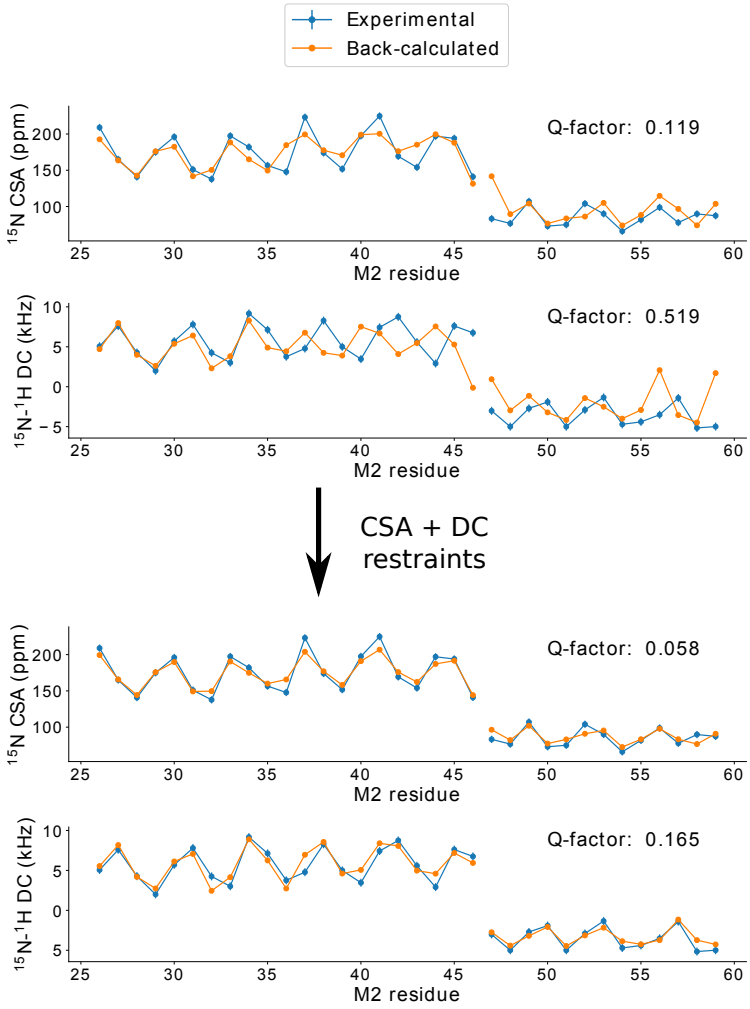


Figure 4.2: Agreement between experimental and back-calculated CSA and DC in M2 ensembles – The comparison between experimental and backcalculated CSA and DC is illustrated for the CHARMM-based ensembles. The unrestrained ensemble (top) features some important deviations in comparison to the experiments, particularly in the dipolar couplings, where the C-terminal segment of the transmembrane helix appears out of phase. Inclusion of the experimental restraints provides an overall better agreement for both experiment observables, illustrated by the decrease in Q-factors.

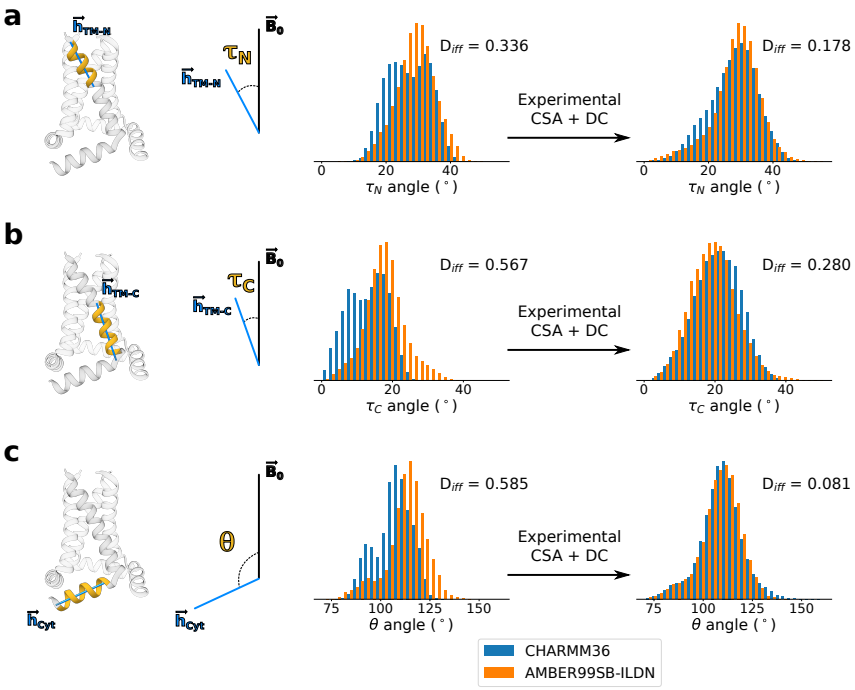


Figure 4.3: Distribution of topological properties in the structural ensembles of M2 – Distributions of (a) τ_N, (b) τ_C and (c) θ angles in CHARMM36 (blue) and AMBER99SB-ILDN (orange) ensembles of the M2 proton channel. Unrestrained ensembles (left) result in dissimilar, less overlaping distributions between the two ensembles. Inclusion of the experimental retraints drive the ensemble toward unique, highly overlapping distributions.

In the resulting ensembles, M2 retains its overall topology, with the four transmembrane helices inserted into the bilayer, exhibiting a kink induced by the glycine residue at position 34. The amphipathic helix rests stably adsorbed on the membrane. Due to its kinked nature, the transmembrane helix was divided into two segments when analysing its topological properties. The N-terminal region (residues 24 to 33) tilts at a mean value of 27.5° ± 7.49, whereas the C-terminal segment (residues 35 to 45) crosses at 20.6° ± 6.46. The C-terminal amphipathic helix, which exhibits the highest degree of convergence amongst the two ensembles, sits at an average angle of 108° ± 11.0.

4.4 Pore profile of the influenza channel

M2 channels protons across the membrane through the interfacial pore formed by its four transmembrane helices. The width profile of the channel determines the ability of protons to diffuse through or react with pore lining residues, more specifically with the histidine

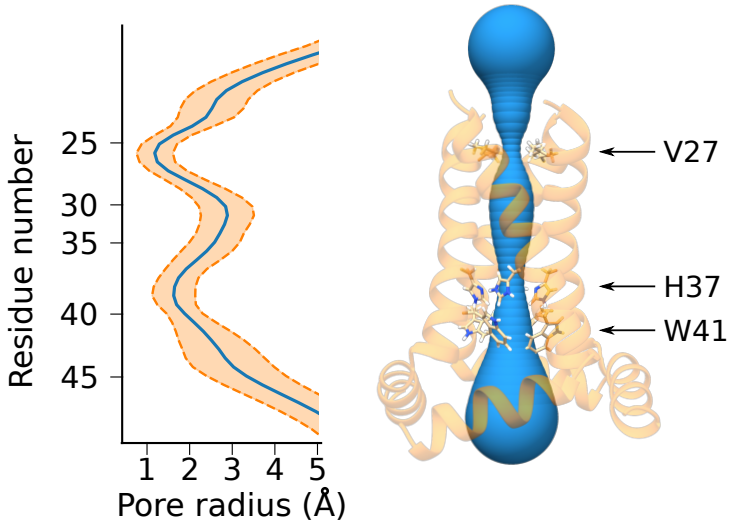


Figure 4.4: Pore radius profile of the M2 influenza channel – The proton channel is characterised by the presence of two major bottlenecks. The first, provided by the sidechains of V27, provides a tight entry point at the N-terminal end of the channel. A second bottleneck is provided by H37, the residues responsible for channel activity. This second gate is followed by the sidechains of W41, which provide a looser constriction point that undergoes wider fluctuations.

tetrad at position 37, essential for channel activity and selectivity. The sampled structural ensembles of M2 provide a dynamical description of the channel, which allows for an analysis of the average properties of the pore as well as the identification of collective protein backbone motions and helical reorientations that may result into an alteration of the channel profile.

The pore of M2 features two major constriction sites (Figure 4.4). The first, located at position 27, is formed by the inward-facing valine sidechains, and creates an entry barrier for water and ions on the N-terminal side of the channel, having an average radius of 1.18 Å $\pm$ 0.42. This constriction is preserved throughout the ensemble, showing small fluctuations in its radius. A second bottleneck is found at the histidine tetrad, which acts as an exit barrier at the opposite side of the pore. This bottleneck appears looser than the one at the entry side, featuring an average radius of 1.62 Å $\pm$ 0.51. A secondary exit gate is provided by the bulky tryptophan sidechains at position 41. Although centered at a loose value of 2.63 Å in radius, this secondary gate features much larger fluctuations across the ensemble. The tryptophan sidechains exhibit inward packing and unpacking motions, resulting in a closing and opening of the pore respectively. These wide fluctua-

tions in the conformations of W41 result in a much wider range of radii at this secondary gate, which exhibits a standard deviation of 0.93 Å across the ensemble.

Between these two entry and exit points of the channel lies a segment where the pore radius becomes wider. This section is lined up by an inward-facing serine sidechain at position 31 and the kink-inducing glycine at position 34. The absence of large interfacial sidechains results in the wider profile of this section, whose average radius across the ensemble reaches the value of 2.89 Å. The wider segment extends for a total of 17 Å, the distance between the two gates, thus allowing for the accommodation of several water molecules inside.

The transmembrane pore of M2 does not remain fixed throughout the ensemble, but its radius is rather subject to fluctuations which occur as a result of large-scale concerted motions of the helical backbones. These fluctuations were found to be dependent on the protein's topological flexibility, and were analysed by evaluating the profile of the channel in relation to orientational properties of the transmembrane helices (Figure 4.5).

Specifically, the ω angle, which reports on the tilt of transmembrane segments relative to the global tetrameric axis (see Figure 4.5a), exhibits a bimodal distribution with a major and a minor conformation. The major conformation corresponds to a topology similar to that observed in the original ssNMR structure, which retains the average radius profile of the interfacial pore. The minor conformation, characterised by lower ω angles, features a more compact packing of the transmembrane helices, with all the pore-lining sidechains pointing inward and leading to a highly closed configuration. The profile of the pore is considerably affected by the collective fluctutations about this angle, exhibiting a global decrease in radius. Particulary, the length of the proton exit bottleneck is extended due to the coming together of W41 sidechains at the quaternary interface. Similarly, the tighter helical packing in the region spanning residues 27 to 37 drastically decreases the volume of the cavity formed between the two bottlenecks.

Global fluctuations such as this one have important consequences in altering the overall profile of the channel, and therefore affecting the accessibility of the inner regions of the

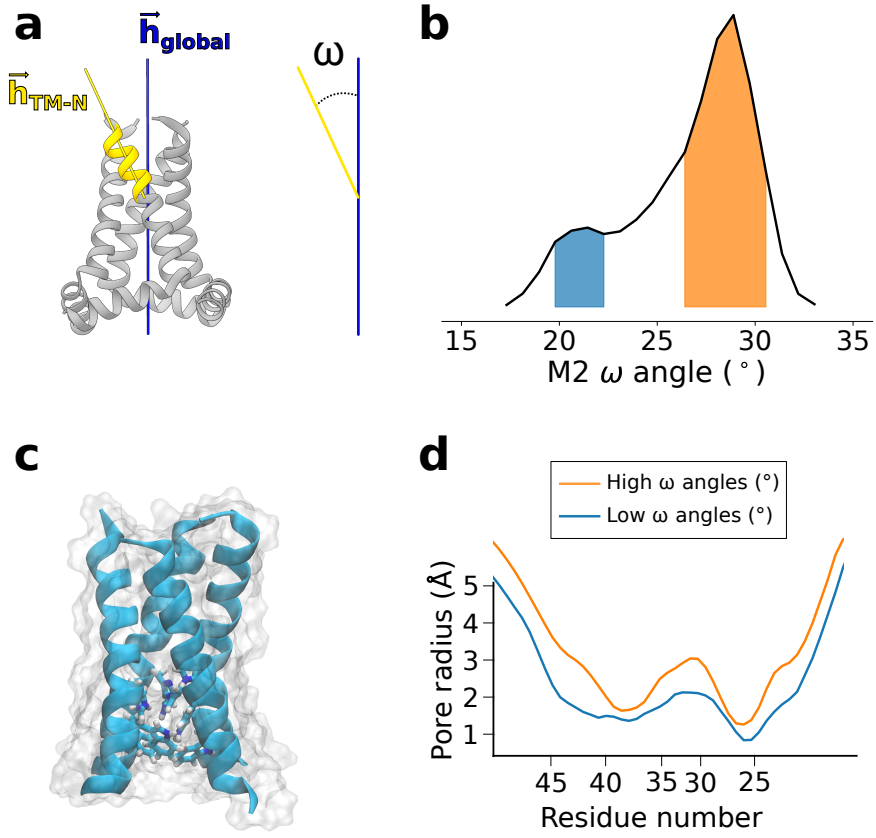


Figure 4.5: M2 channel fluctuations in relation to the topological motions of helices – (a) Definition of the ω angle in M2, calculated between the N-terminal segment of the transmembrane helix (residues 24 to 33) and the global quaternary axis of the channel's transmembrane segment. (b) Distribution of the ω angle, featuring a bimodal shape with two distinct conformations. (c) Representative structure of the minor conformation, where the channel assumes a highly closed conformation as a result of the tight packing of sidechains at the quaternary interface. (d) Pore radius profile comparison between the two conformations described in (b).

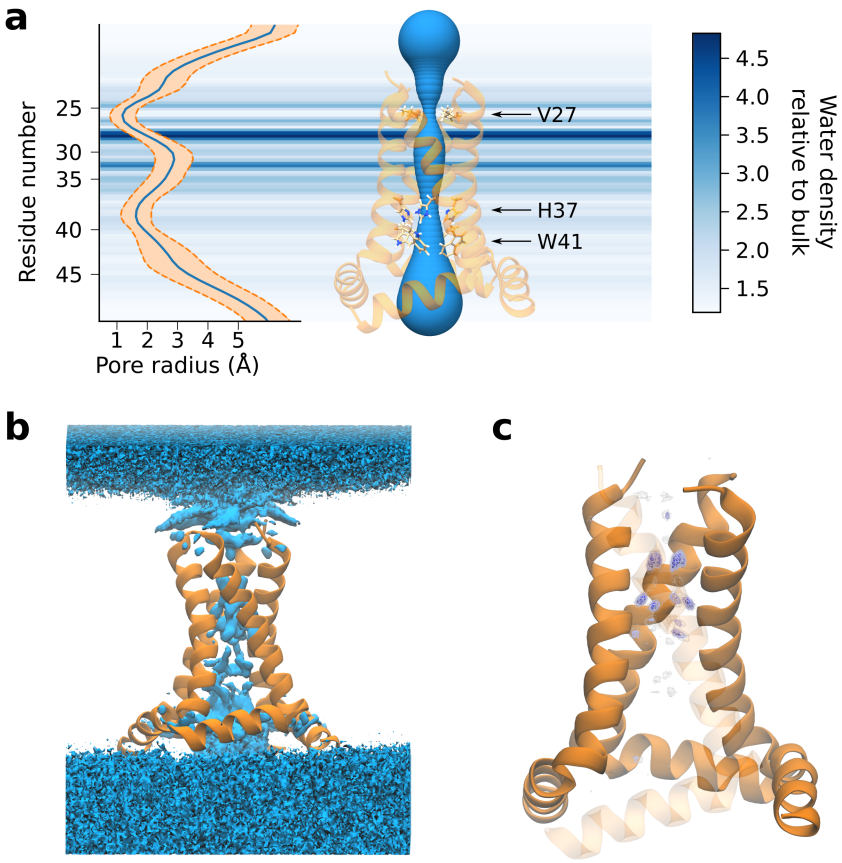


Figure 4.6: Hydration properties of the M2 proton channel – (a) Average water density computed along cross-sections of the tetrameric pore, overlayed onto the pore radius profile. (b) Illustration of the bulk water density within the simulation box. (c) Clusters of structured waters, with high density relative to the bulk. Three dimensional contours represent densities of 4σ(transparent wireframe), 6σ(grey wireframe) and 8σ(blue solid surface)

pore to water and protons. These conformational transitions, characterised by collective reorientations of transmembrane helices, are likely to be a driving factor in the mechanism of proton channeling through the M2 pore.

4.5 Role of hydration in proton transport by M2

Water dynamics plays a fundamental role in enabling the transfer of ions through membrane protein channels.[336] Particularly, the structure induced in water by its confinement within protein channels has been shown to influence the transfer of protons across the membrane.[337] In the structural ensembles of M2, the presence of such structured waters was analysed by computing a grid-based water density function relative to the position of

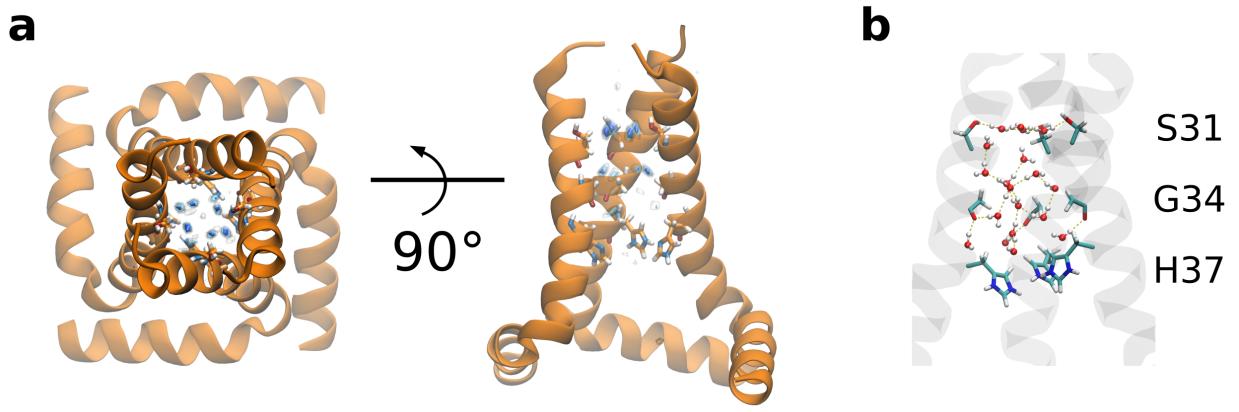


Figure 4.7: Structure of water molecules within the pore of M2 – (a) Highly ordered clusters of dense water within the inner cavity of the M2 channel pore. Contours are drawn at 5σ, 7σ and 10σ. One protomer was removed from the side view for clarity. (b) Structured water molecules are stabilised by a heavy network of hydrogen bonds, including contributions from the sidechain of S31, the carbonyl of G34 and the imidazole ring of H37.

the proton channel (see Section 4.2.2 for details).

Focusing on the water dynamics within the interior of the pore, there is a clear spike in water density inside the cavity delimited by the two bottlenecks at positions 27 and 37. In this confined space, water exhibits an average density close to 5 times higher than that of the bulk solvent (denoted as 5σ), with some structured clusters that reach 10σ (Figure 4.6). The entry and exit of water molecules into this cavity during the simulation is considerably slow, being heavily limited by the gates illustrated in Figure 4.4.

Water molecules are not only confined inside the pore's internal cavity, but they are also stabilised by polar interactions with inward exposed chemical groups of the protein. At the center of the cavity, the polar sidechain of serine at position 31 provides hydrogen bonding opportunities for water. Similarly, the backbone carbonyl group in kink-inducing glycine-34 is not involved in secondary structure hydrogen bonding, and is therefore able to interact with nearby waters. Finally, the imidazole rings of the pore-delimiting histidine tetrad at position 37 also exhibit a high degree of hydrogen bonding throughout the ensemble. In fact, the positions of the highest density water clusters correlate well with the locations of these chemical groups (Figure 4.7). A very dense network of hydrogen bonds is observed within the cavity throughout the simulation, resulting in a high degree of water structuring and the observed increase in water density.

In this inactive conformation of M2, the high degree of water ordering in the vicinity of H37 may have implications in the rate of proton transport, likely acting as an impediment to diffusion and presenting an additional energetic barrier that prevents the movement of molecules through the pore.

Chapter 5

Structural dynamics of the G-protein coupled receptor CXCR1

5.1 CXCR1: an interleukin-8 receptor

G-protein coupled receptors (GPCRs) are a major class of membrane proteins. They are crucial in modern medicine, being targeted by over 25% of all available therapeutical drugs.[43] They have a characteristic, highly conserved topological fold, composed of seven transmembrane helices. The function of GPCRs is to relay signals to the inside of the cell upon binding of molecules on the extracellular space. They do so by activating guanine nucleotide-binding proteins (G-proteins), which leads to the activation of a downstream signalling cascade that eventually results in cellular changes.

GPCRs are involved in a wide variety of cellular functions, ranging from sensory tasks to the regulation of homeostasis.[338] One particular class of GPCRs are chemokine receptors, involved in the regulation of the immune system response and inflammation. Chemokines are small proteins that may be produced and secreted under certain conditions to trigger the migration of cells to a specific site. Many chemokines, such as Interleukin-8 (IL-8, also known as CXCL-8) are associated with the inflammatory response, and act by recruiting leukocytes to the site of inflammation.[339]

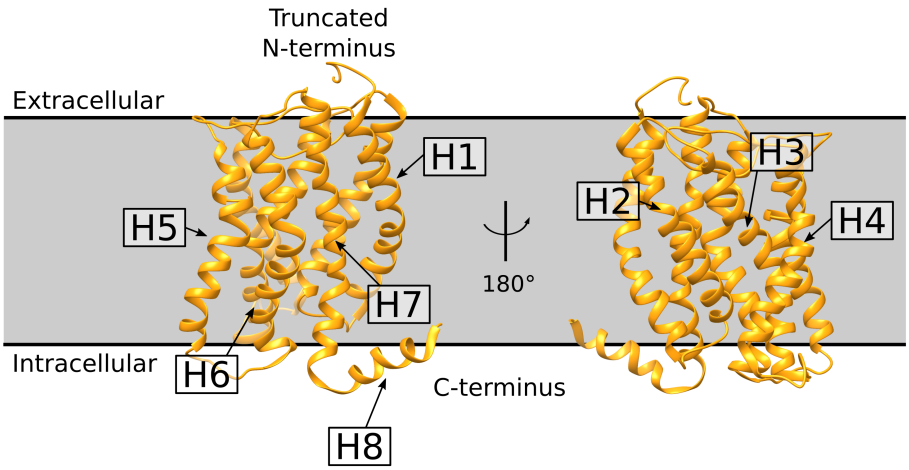


Figure 5.1: Structure of the CXCR1 receptor in lipid bilayers – The structure of CXCR1 was determined by MAS ssNMR, combining torsion angle restraints from isotropic chemical shifts with orientational restraints from ^{15}N-^{1}H and $^{13}C_{\alpha}$-$^{1}H_{\alpha}$ dipolar couplings. The structure features the characteristic seven transmembrane helix fold of GPCRs, with perceptible kinks in the helices.

More specifically, IL-8 primarily acts by activating the GPCRs CXCR1 and CXCR2, triggering signal cascades that may result in different cellular responses.[340] IL-8 commonly binds to CXCR1 receptors expressed at the surface of neutrophils, inducing a conformational change in the receptor which exposes intracellular epitopes for G-protein binding. Activated G-proteins can target effectors of chemotaxis such as the phosphatidylinositol-3 kinase, which subsequently increase the phosphorylation and activity of Protein Kinase B (also known as Akt).[341] Signalling pathways such as this one are crucial in the correct modulation of the inflammatory response, but they are also found to be upregulated in many common forms of cancer.[340]

Despite the many advances in the biological characterisation of these signalling pathways and their implications in disease, the detail of the molecular mechanisms governing the process is still scarce. The structure of CXCR1 was recently elucidated by solid-state NMR (see Figure 5.1), employing dipolar couplings as a primary structural restraint.[81] The availability of such data enables the study of the conformational dynamics of CXCR1 by means of experimentally-restrained molecular dynamics. In this chapter I present conformational ensembles of CXCR1 in its IL-8 free state, which shed some light on dynamical structural transitions with potential implications for signal transduction, as well as

allowing for the evaluation of the membrane protein's allosteric network at an atomic detail.

5.2 Methods: DC-restrained simulations and analysis of CXCR1 ensembles

5.2.1 Solid-state NMR restrained simulations

Experimental background

The CXCR1 structure elucidated by Park and co-workers[81] employed dipolar couplings (DCs) as structural restraints. These DCs were not obtained from oriented ssNMR measurementes such as the PISEMA-based experiments described in the previous chapters of this thesis. In the case of CXCR1, magic angle spinning (MAS) separated local field (SLF) experiments were performed, where the heteronuclear DCs are recoupled during the pulse sequence and correlated to the isotropic chemical shifts of the nuclei.[342] In this manner, the authors obtained DCs for amide ^{15}N-^{1}H as well as $^{13}C_{\alpha}$-$^{1}H_{\alpha}$ nuclear pairs. Since they are performed under MAS conditions, SLF type experiments do not require the mechanical or magnetic alignment of samples that is characteristic of oriented ssNMR approaches. However, in order to recover the DCs in SLF experiments an additional condition must be met in the experimental setup. The protein in the MAS rotor must be undergoing fast rotational diffusion about the bilayer normal, that is the protein must be globally rotating at a faster speed than the experimental NMR timescale of the experiment. This often requires careful tuning of the experimental temperature, which normally needs to be above the gel to liquid crystalline phase transition of the membrane to observe the rapid rotational diffusion. Samples that satisfy this condition are often referred to as *rotationally aligned* (RA).[343] The motionally averaged DCs obtained by this MAS approach may be used as structural restraints in the same way as those obtained by oriented ssNMR approaches.

CXCR1 equilibration and restrained MD sampling

The initial CXCR1 structure (PDB accession code: 2LNL) was energy minimised and embedded into a bilayer composed of 399 DMPC lipids, which reproduces the experimental conditions. The system was solvated and equilibrated using the MARTINI force field[117] for 500 ns employing positional restraints on the protein particles. The system was subsequently translated back into a full atomic description by the Backward MARTINI method.[304] Tip3p waters were used to solvate the full atomic model, adding 32,392 molecules to the system. 13 chlorine ions were added to fully neutralize the charge of the system. In total, the final simulation system contained 149,122 atoms in a dodecahedron box of principal dimensions $12.6\times10.9\times10.8$ nm^3. During all simulations, temperature was coupled at 300 K using the V-rescale method[297] (0.1 ps constant). Pressure was coupled at 1 bar every 1 ps using the semi-isotropic Berendsen algorithm.[296] The Particle mesh Ewald method[299] was employed to calculate electrostatic interactions and LINCS[298] was used for bond constraints.

The system was further equilibrated using the AMBER99SB-ILDN[279] force field, maintaining positional restraints of protein heavy atoms for 20 ns. The structure was further equilibrated by removing the positional restraints and including distance restraints to maintain the hydrogen-bonding pattern of the transmembrane helices, which were linearly relieved during 30 ns of sampling, in order to minimise possible aberrations caused by the force field. The resulting systems were further simulated for 40 ns, and 8 equally spaced structures from the last 10 ns of sampling were extracted as starting conformations for the ssNMR restrained MD simulations.

The DC-restrained samplings were replica-averaged across 8 parallel simulations. Restraints were gradually introduced by linearly increasing the force constants over 20 ns of sampling, to final values of 800 J/(mol$\times$kHz^2) for both ^{15}N-^{1}H and ^{13}C$_\alpha$-^{1}H$_\alpha$ DCs. Restrained simulations were run for 80 ns per replica after equilibration, generating a restrained structural ensemble of CXCR1, composed of 10,000 equally spaced structures. Unrestrained simulations were also ran. Specifically, 4 independent simulations of 50 ns

each were performed and an unrestrained ensemble was generated for comparison to the experimental sampling. The samplings above described resulted in a total simulation time of 1 µs.

An analysis of the experimental agreement and topological properties of the system was carried out for CXCR1. Q-factors and helical axes were computed following the same approach described in analysing the phospholamban ensembles (see Section 3.2.1).

5.2.2 Network description of CXCR1

In order to study the allosteric communication network of CXCR1, mathematical graph-based methods were employed.[344] Generally speaking, a protein structure may be modelled as a mathematical graph where each node corresponds to a residue, and edges are drawn between residues that are in close contact within the structure. Having a dynamical ensemble, one may compute the correlations in the relative motions of residues and identify allosteric paths that connect separate regions of the protein.

Specifically for CXCR1, a network of 296 nodes was constructed (the number of residues in the protein structure). Pairwise contacts between residues were defined using a cut-off of 5 Å between any two heavy atoms. Connections (formally known as *edges*) were drawn between two nodes in the network only if a contact was present in at least 75% of the structures in the ensemble. The graph-based model was built using the Python model NetworkX.[345]

Once the network is constructed, the edges are weighted according to the motional correlations between the two residues involved. Traditionally, these correlations can be extracted by computing a Pearson correlation coefficient from the Cartesian covariance matrix of atomic positions.[346] However, this metric often disregards correlations based on their spatial orientations (e.g. perfectly correlated orthogonal motions appear non-correlated), a limitation that can be overcome by applying information theory based metrics.[347] In particular, the correlation between the motions of two atoms i and j can be derived from

their mutual information:

$$MI_{i,j} = \int \int p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} dx_i dx_j \tag{5.1}$$

where $p(x_i)$ and $p(x_j)$ are the marginal probability distributions of the atomic positions of i and j respectively, and $p(x_i, x_j)$ is their joint probability distribution. The value of $MI_{i,j}$ ranges from 0 (for non-correlated motions) to $+\infty$ (for fully correlated motions). To aid in interpreting the value, its range can be transformed into [0,1] by means of a Pearson-like coefficient:

$$C_{i,j} = \sqrt{1 - e^{-(2/d)MI_{i,j}}} \tag{5.2}$$

where d corresponds to the dimensionality of the data ($d = 3$ when evaluating Cartesian coordinates of molecular structures). The computation of these generalised correlation coefficients was performed using the *g_correlation* tool built for GROMACS.[347]

Finally, this coefficient is transformed into a *distance* which is directly associated which each edge in the graph:

$$d_{i,j} = \log \mid C_{i,j} \mid \tag{5.3}$$

In this way, residues whose motions are highly correlated are separated by a *short* distance (low value of $d_{i,j}$), whereas uncorrelated residues are separated by longer distances.

The resulting graph representation of CXCR1 allows the study of communication networks in the structure. For example, by studying the *shortest path* between two distant residues in the structure one can find the most likely allosteric path connecting them, by following the short-ranged correlated motions of residues in close contact. An overall measure of a residue's involvement in the allosteric network can be obtained by its *centrality*. Different measures of centrality exists, according to different considerations of importance

within the network. One such definition is *closeness* centrality, which for every node measures the sum of the length of the shortest paths to every other node in the network. The closeness centrality of node x is defined as:

$$C(x) = \frac{N}{\sum_y d_{y,x}} \tag{5.4}$$

where N is the number of nodes in the graph and y runs through all nodes in the graph different to x. According to this metric, the more central residues in the network are those with the highest allosteric proximity to the other residues in the structure.

5.3 Experimentally derived dynamics of a GPCR

The structural ensemble of CXCR1 were calculated using ^{15}N-^{1}H as well as $^{13}C_{\alpha}$-$^{1}H_{\alpha}$ DCs measured in lipid bilayers by MAS ssNMR. The replica-averaged ssNMR-restrained MD approach was employed with 8 parallel replicas. Unrestrained MD simulations were run as a control.

Unrestrained simulations are unable to reproduce the available DCs, resulting in very high deviations between the experimental and back-calculated values (Figure 5.2). Upon inclusion of the restraints, the agreement with the data drastically improves, yielding Q-factors of 0.175 and 0.255 for $^{13}C_{\alpha}$-$^{1}H_{\alpha}$ and ^{15}N-^{1}H DCs respectively.

It has become evident in previous chapters that DCs carry plenty of information regarding the topological orientation of α-helices. The high quality agreement between the ensemble and the experiment is therefore suggestive of a highly accurate representation of the topological dynamics of CXCR1. Interestingly, the ensemble features a very large degree of topological variability in the orientations of transmembrane helices. Every transmembrane helix in the receptor exhibits some degree of kinking in the ensemble (see Table 5.1), ranging from small bends such as that of helix-1 (13.2°) to larger helical reorientations (>45° in helix-3 and helix-6).

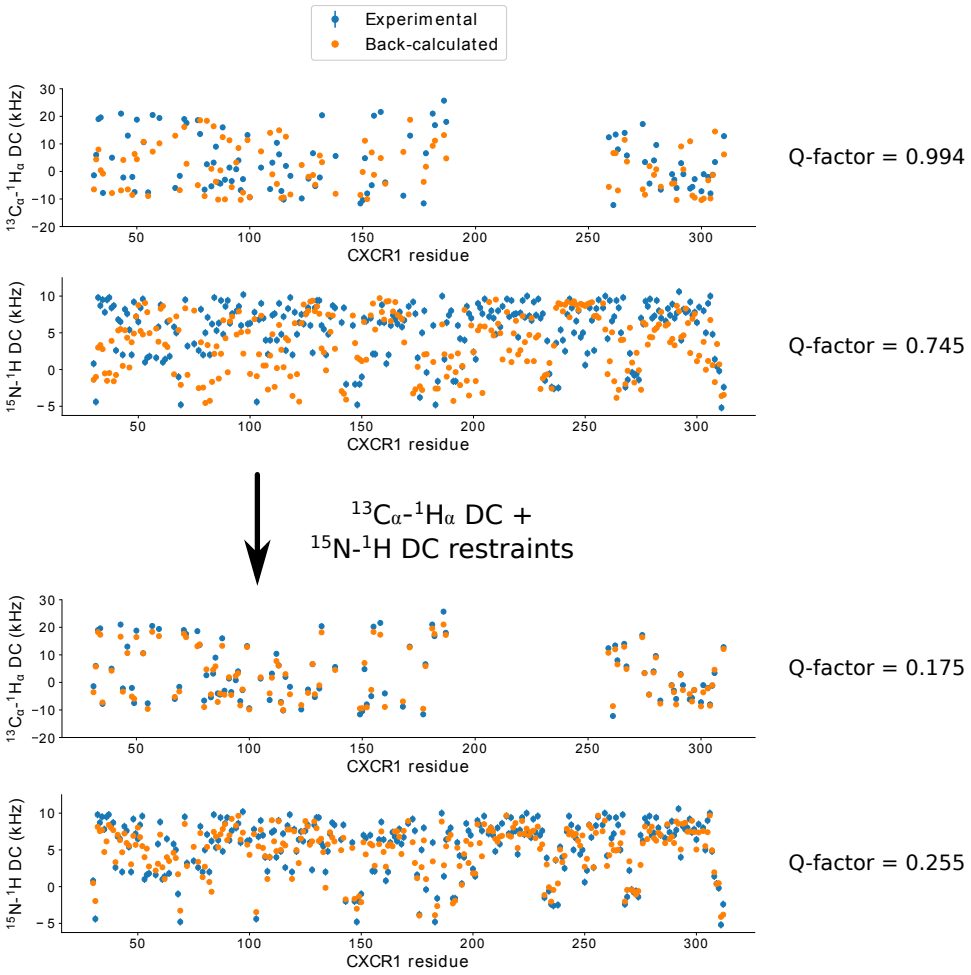


Figure 5.2: Agreement between experimental and back-calculated DCs in CXCR1 ensembles – Unrestrained simulations are not able to reproduce the experimental values of the DCs. Inclusion of the experimental restraints drastically improves the agreement, reaching considerably lower Q-factors.

Helix	Top residues	Bottom residues	Kink position	Average kink angle (°)
H1	37 to 55	57 to 67	G56	13.2
H2	90 to 101	74 to 87	F88	39.7
H3	110 to 123	126 to 138	G124	49.4
H4	163 to 173	150 to 161	G162	34.8
H5	200 to 209	211 to 226	G210	25.0
H6	258 to 268	238 to 256	P257	46.5
H7	275 to 293	295 to 308	G294	35.3

Table 5.1: Residues used to define the transmembrane helical axes of CXCR1. All seven transmembrane helices exhibit some degree of kinking during the simulations. The longest segments of unkinked helices were selected to compute the axes of their top parts (closer to the extracellular side) and the bottom segments (closer to the intracellular interface). The positions of the kinking residues and the average kink angle across the ensemble are outlined.

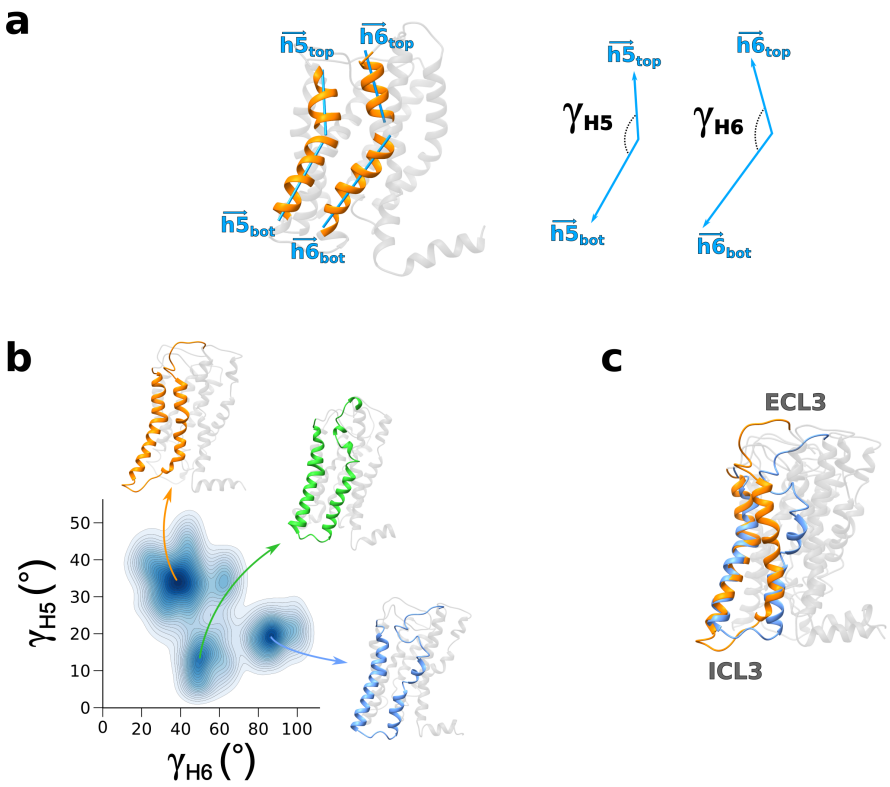


Figure 5.3: Helical reorientations between H5 and H6 of CXCR1 – (a) Kink angles (γ) of helix-5 and helix-6 of CXCR1. The range of residues used to define the helical segments are outlined in Table 5.1. (b) Distribution of γ_{H5} and γ_{H6} angle distributions, illustrating the codependence of both angles. Darker colours indicate a higher population in the ensemble. (c) Overlay of the representative structures of the two most populated conformations along the landscape in (b). The intracellular loop 3 (ICL3) and extracellular loop 3 (ECL3) show a considerable degree of variation among the structures as a result of the helical reorientations and the partial unfolding of helix-6.

Helical kinks have been identified and studied in both soluble and membrane proteins.[348,349] Most studies point toward their importance in dictating the tertiary helix packing of proteins, as well as creating regions of helical flexibility that can serve as hinges in transducing long-range signals along the protein structure.[350] More specifically in GPCRs, helical kinks have been shown to provide a source of structural diversity among these proteins, which have such a highly conserved structural fold.[351] The introduction of kinks at different positions of helices in highly similar proteins may lead them to exhibit distinct topological reorientations that can account for their unique functional characteristics.

The structural ensemble of CXCR1 not only features kinks in all the transmembrane helices, but also reveals topological reorientations of the protein where these kinks act as hinges, or bending points for helices. An example of such topological motions is illus-

trated in Figure 5.3. Helix-5 and helix-6 both exhibit a large degree of topological variability in their kink angles, and being in contact with each other, their orientational motions exhibit a high degree of codependence (Figure 5.3b). According to the reorientations of helices 5 and 6, the ensemble can be separated into three conformations. The most populated helical orientations correlate to the conformation where the helices are fully folded, with helix-5 kinked at 34° and helix-6 kinked at 39°. A very lowly populated state is observed where helix-5 shifts to a very low tilt (13°) and a partial unfolding at the top segment of helix-6 starts to occur (from position 263 onwards), resulting in an observed kink of 49°. The third conformation corresponds to a more populated metastable configuration of the helices, where the top of helix-6 kinks considerably more and retains its partial unfolding.

Interestingly, these dynamical rearrangements propagate within the structure and considerably alter the conformation of the loops that connect the different helices. In GPCRs, these extracellular and intracellular loops are indeed responsible for establishing specific interactions with the ligand and the G-protein respectively.[58,352,353] In the case of CXCR1, the rearrangements of helix-5 and helix-6, coupled with the partial unfolding at the top of helix-6, crucially alter the shape of extracellular loop 3 (ECL3), with the loop translocating to the side and leaving a wider cleft at the extracellular side of the receptor. The structure of the intracellular loop 3 (ICL3) that connects both helices also sees some variation, driven by the reduced C-terminal extension of helix-5 in the metastable configuration.

This result highlights the importance of these extremely flexible hinges in dictating the shape of exposed protein surfaces. Together with the overall topological variability and codependent orientational dynamics in the ensemble, the analysis points toward a potential role of helical kinks in relaying allosteric signals upon ligand binding to GPCRs, potentially acting as hinges to propagate ligand-induced structural rearrangements in the extracellular side to the intracellular surface.

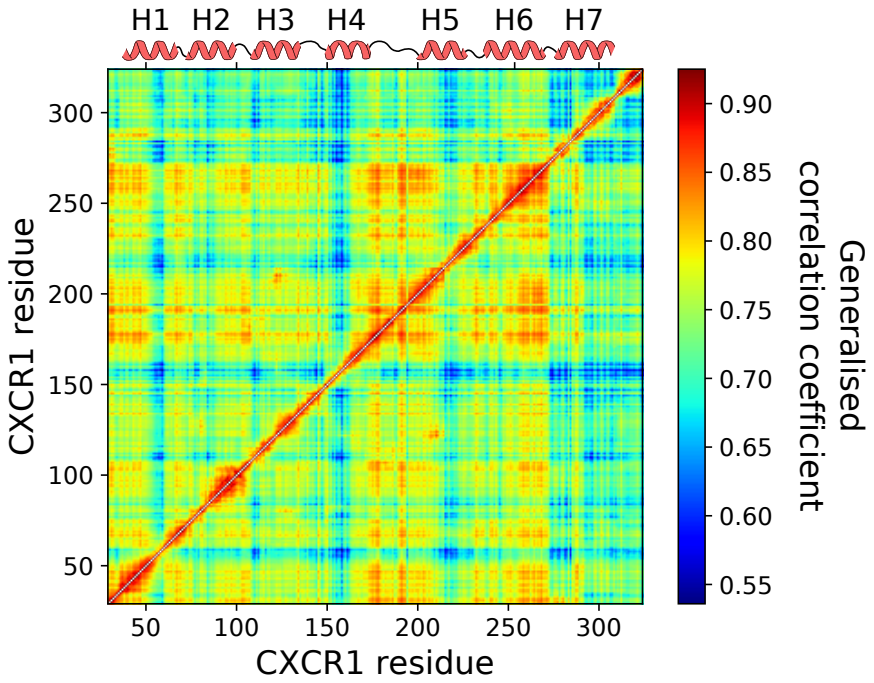


Figure 5.4: Generalised correlation matrix of C_α atoms in the CXCR1 ensemble – Generalised Pearson-like correlation coefficients (Equation 5.2) were computed between every pair of C_α atoms in CXCR1. Regions with a high value of the coefficient exhibit a high degree of correlation in their motions, whereas the opposite is true for lower values

.

5.4 Allosteric communication network of CXCR1

Having an extensive structural ensemble of CXCR1, there are computational analysis methods to reveal the allosteric communication pathways in the GPCR. Allosteric motions are often revealed through long-range correlations between protein residues. Having molecular trajectories, the codependence between motions can be established by computing the covariance matrix of atomic coordinates,[346] or more elaborately, a generalised correlation metric based on the mutual information between residues (Equations 5.1 and 5.2). The generalised correlation of motions was computed for the C_α atoms of CXCR1, and is depicted in Figure 5.4.

Overall, the matrix reveals periodic sections of short-ranged correlations, mostly aligning to the transmembrane helices of the receptor, as would be expected. A considerably higher degree of long-range correlations is found at helix-6, particularly at its C-terminal top segment (between positions 263 and 275) This region is indeed highly dynamical and,

as shown in Figure 5.3, it features a high degree of topological variability and transient unfolded conformations. The global correlation matrix reflects that these motions have long-range effects, triggering rearrangements at distant regions of the receptor, including a number of segments ranging from H1 to H5 in the sequence.

A closer look at the allosteric mechanisms of proteins can be achieved by studying its communication network. The phenomenon of allostery is commonly described as the movements through a network of contacts along which structural changes are propagated, resulting in correlated long-range effects in protein conformational transitions.[354] The challenge of modelling networks across a multitude of disciplines has commonly been approached by mathematics, applying elements from graph theory. Mathematical graphs are objects composed of *nodes* (also known as vertices) and *edges* that join nodes together. There is a history of modelling protein structures as graphs, for example by using nodes to represent residues and drawing edges between residues in close spatial proximity.[355]

It is possible to incorporate the dynamical information from protein conformational ensembles into a graph-based description, and infer from it some properties of the communication network between residues. In the case of CXCR1, a graph of 296 nodes (one per residue) was constructed, and a connected graph was built from the interatomic contacts in the ensemble (see Section 5.2.2 for detailed methods). The generalised correlation metrics (depicted in Figure 5.4) are transformed into a distance value (Equation 5.3) in order to weigh each edge in the network. Finally, the closeness centrality (Equation 5.4) was computed for each residue as a measure of their importance within the global communication network of the protein. The resulting graph is depicted in Figure 5.5a.

Overall, the network features a highly connected structure, with two main distal sides (the intra- and extracellular faces of the receptor) linked by a dense central network. The network properties can be projected onto the structure by mapping the centrality of each residue (Figure 5.5b). Interestingly, the most central residues are found at the interface between helices 3 and 5, where both helical kinks encounter each other. In fact, the kinking residues G124 and G210 are among the three most central residues in the whole net-

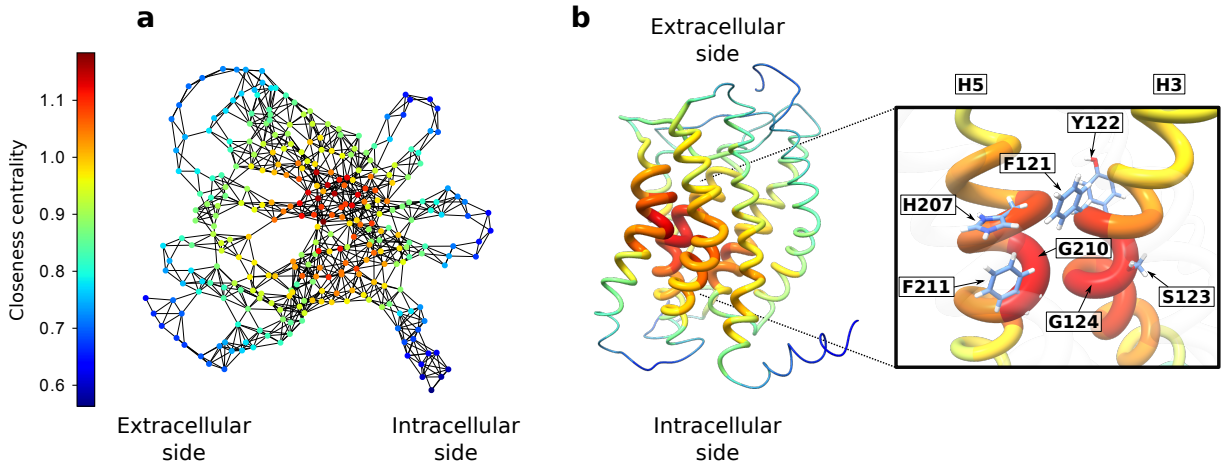


Figure 5.5: Allosteric network structure of CXCR1 – (a) The communication network of CXCR1 was modelled as a mathematical graph of 296 nodes, one per each residue. Connections between nodes are based on interatomic contacts and weighed according to the generalised correlations in their motions (see Section 5.2.2 for details). Nodes are coloured according to the value of their closeness centrality. (b) Closeness centrality values mapped onto the structure of CXCR1. The most central region is shown in detail, with the sidechains of central aminoacids depicted as sticks.

work. These kinks are surrounded by an aromatic-rich region that is also central in the network.

The key position of these two helical kinks within the network is further evidence of their importance in relaying allosteric signals between the extra- and the intracellular sides of the receptor. Given their high centrality, these residues are involved in the optimal communication pathways between different segments of the protein. In the graph-based model, the shortest paths between two nodes can be readily computed by minimising the distance required to travel from one to the other. These are indicative of the optimal communication path between those two residues.

An example of several shortest paths that travel through the kinked H3-H5 interface is shown in Figure 5.6. Specifically, the calculated paths join residues between two distant surface-exposed loops, ICL3 (on the intracellular face) and ELC2 (extracellular). The optimal communication pathways between these loops are relayed primarily through helices 3 and 5, finding their common bottleneck at around the kink region. The communication between loops at the opposite interfaces of the receptor is crucial for GPCR activity and the specificity of their interactions. Helical kinks are here shown to play a central role in

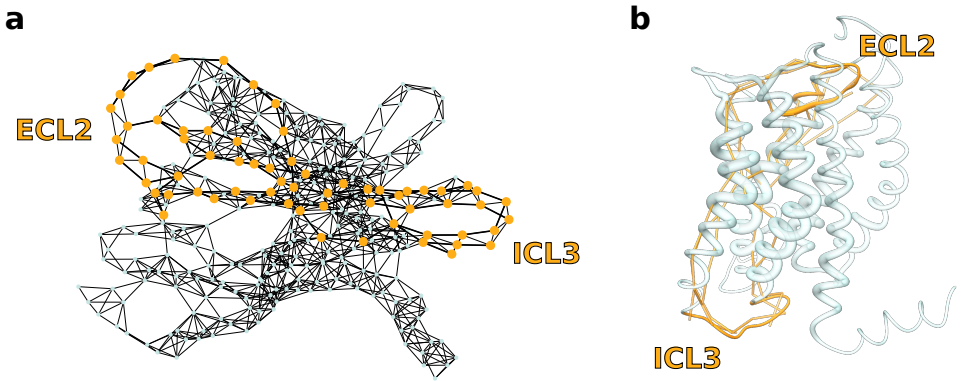


Figure 5.6: Allosteric paths between distant loops of CXCR1 – (a) Shortest paths between residues in intracellular loop 3 (ICL3) and extracellular loop 2 (ECL2) of CXCR1. Exposed loops are crucial in the function of GPCRs, by interacting with ligands on the extracellular site and with effector G-proteins on their intracellular face. (b) Structural representation of the shortest paths outlined in (a), mapped onto the three dimensional structure of CXCR1.

the allosteric paths joining these segments, underlining the importance of orientational dynamics in governing GPCR function.

Chapter 6

Discussion

The results presented in this thesis combine methodological advancements and practical applications of computational sampling methodologies that integrate experiments and theory to characterise the conformational dynamics of membrane proteins. The approaches here presented have focused primarily on experiments of nuclear magnetic resonance, but the principles of ensemble calculations are applicable to any source of molecular information. The overall scope of this research is to highlight the ability of experimental information to improve the accuracy of state-of-the-art theoretical models, as well as the necessity to employ such theoretical methods to achieve a full detailed understanding of the mechanisms that govern membrane protein dynamics.

6.1 Toward an accurate description of membrane protein conformational ensembles

6.1.1 Improving the interpretation of protein chemical shifts

Nuclear magnetic resonance (NMR) measurements in proteins fundamentally rely on one crucial observable: the chemical shift (CS). They are unique for each atom in a protein

structure and are affected heavily by subtle alteration of their chemical environment, therefore being very highly sensitive reporters of local structure and dynamics in proteins. Given the great deal of structural information they carry and their ubiquity within the field of NMR, the absence of highly accurate structural models considerably hinders their interpretation.

The methods illustrated in Section 2.1, with the implementation of PROSECCO, aim to advance the understanding of the dependence of CS on protein sequence and local secondary structure. In the case of intrinsically disordered proteins, given their absence of a well-defined tertiary structure, they are generally correlated to the protein's primary sequence. The kernel-based statistical approach presented in $\text{PROSECCO}_{\text{IDP}}$ improved the accuracy of the available sequence-based predictors.[244,245] Sequence-based models of CS in disordered proteins are indeed a useful tool, as deviations from the disordered CS values generally indicates the transient formation of secondary structure elements, often involved in mechanisms of molecular recognition characteristic of disordered segments.[34]

In folded proteins the interpretation of CS becomes much more difficult, due to the degree of complexity added by the intramolecular interactions of its tertiary structure. The correlations between protein structure and CS has been explored by a variety of statistical approaches,[225–228] which despite providing useful predictions, still have shortcomings in terms of accuracy. The new approach presented in this work, $\text{PROSECCO}_{\text{FOLDED}}$, shows indeed that by only considering primary sequence information a similar level of accuracy can be achieved to the widely employed currently available models of CS.[229] In addition, $\text{PROSECCO}_{\text{FOLDED}}$ is able to consistently improve the predictions provided by Camshift,[228] currently the only method employed in CS-guided dynamical refinements of proteins. The principles upon which $\text{PROSECCO}_{\text{FOLDED}}$ is constructed therefore serve as a starting step to build more elaborate structure-based models of CS, allowing for better interpretation of these observables.

The analyses leading to the development of PROSECCO have in fact already revealed several structural factors that heavily influence protein chemical shifts. With regards to

structured segments, it appears that the CS of backbone atoms at internal regions of secondary structure elements are primarily reliant on the protein sequence, and mostly independent from tertiary and quaternary structural features. The behaviour at the boundaries of secondary structure elements is however considerably different, with secondary CS of different atoms experiencing clearly distinct patterns at the interface between Q3 elements. The results from PROSECCO highlight the importance of carefully studying and quantifying these effects, which should be accounted for in parametrising more accurate structure-based models of CS.

6.1.2 Accurate topological dynamics of membrane proteins

Membrane proteins are exposed to a unique physicochemical environment that is provided by the lipid bilayer. The membrane is however not a static environment, but rather a highly fluid one where lipids and proteins undergo a wide variety of motions. In proteins, these range from atomic fluctuations to large-scale conformational transitions.[356] In integral membrane proteins, the larger scale fluctuations often involve reorientations of the protein backbone with respect to the membrane axis, generally resulting in collective motions that can be crucial for protein function.[201]

In Section 2.3 I outlined a powerful methodology to study such dynamics, by combining the information encoded in oriented solid-state NMR (ssNMR) measurements with theoretical models of protein dynamics in the form of molecular dynamics (MD) simulations. The sampling methodology, based on the widely described ensemble-restrained MD approach,[33,184,189] was extended and applied to homo-oligomeric protein systems, by including two layers of averaging (one among simulations evolving in parallel and another between the identical protomers conforming the quaternary assembly).

Using the structure of non-phosphorylated phospholamban pentamer in a DOPC/DOPE bilayer as a model oligomeric system, an exhaustive calibration and validation of the sampling approach has been conducted by means of the *reference ensemble* test.[176,185] The calibration revealed the importance of averaging the restraints at all possible levels (at

the replica level as well as internally among the phospholamban protomers) in order to achieve an optimal structural characterisation. More importantly, the results from the test highlight the divergence in the structural ensembles generated by two different force fields commonly employed in biomolecular simulations (AMBER99SB-ILDN[279] and CHARMM36[126]). The divergence between empirical force fields is a considerable limitation in the application of theoretical sampling methods. Every force field has an intrinsic bias in their parametrisation, which can lead to artefacts in the resulting structural ensembles. It is impossible to know *a priori* which force field is most suitable to study a particular biomolecular system, and it is likely that every possible option carries a high level of error when compared to the experimental structural ensemble.[132,133] The reference ensemble validation of pentameric phospholamban shows that CSA and DC restraints are able to overcome force field biases in the description of topological dynamics (of orientational helical angles) and general structural propensities (reported by intramolecular distances). In fact, the results from the optimal restrained sampling setup reaches the statistical error limit in sampling the ensemble, providing encouraging evidence that the inclusion of experimental information can help overcome the intrinsic inaccuracies of empirical force fields in membrane protein systems.

Further evidence of the structural accuracy provided by the restraints was obtained when applying the experimental CSA and DC restraints to phospholamban. In unrestrained simulations, the two force fields employed produce ensembles that are noticeably different in their topological properties. This divergence is corrected when applying the experimental CSA and DC as restraints, always resulting in an increased similarity between the two force fields. This convergence in the structural distributions is evidence that the restraints are driving the sampling toward a single ensemble that is closer to the experimental one. In some cases, such as in phosphorylated phospholamban, the resulting experimental ensemble is considerably different to both unrestrained samplings, a result that highlights the importance of including experimental information to achieve an accurate description of biomolecular dynamics. The restraint-driven convergence of topological ensemble properties is also described in the influenza channel M2, demonstrating the general applicabil-

ity of these methods to the study of different biological problems.

The accuracy of the experimentally-driven ensembles can be further validated by comparing the results against orthogonal experimental data that were not included as restraints. The ability to reproduce independent data is generally treated as evidence of ensemble accuracy.[154] In the case of the phospholamban pentamer there is an excellent agreement between the back-calculated amide S^2 order parameters and relaxation NMR measurements,[292] an agreement that is drastically improved by the inclusion of the restraints. This supports the hypothesis that experimental restraints do not simply influence the description of a few highly correlated structural variables, but rather they provide an overall improvement in defining the structural properties of the ensemble.

Finally, it is important to note that despite the high level of information carried by experimental CSA and DC, their detailed dynamical interpretation would be impossible without the inclusion of theoretical methods such as MD. Traditionally, these data have been fitted with static models to obtain a single optimal value for the helical crossing angle.[273,274] However, these models neglect the dynamical information encoded in the data, and can yield wildly inaccurate results in situations of high topological heterogeneity. Indeed, the application of this model to the CSA and DC of phosphorylated phospholamban results in a crossing angle that is incompatible with the protein's transmembrane fold. In this case, only the incorporation of MD calculations can yield an accurate description of the topological heterogeneity of the pentamer. The inclusion of dynamics is important even in cases where the ensemble distributions correspond to the optimal topological angle, since the MD sampling approach provides detailed information on the extent of the fluctuations about that central angle.

Overall, the ssNMR-restrained MD samplings provide considerable evidence on the ability of these experiments to correct the biases present in empirical force field models. The results from the generated ensembles also highlight the need to include the dynamical considerations provided by MD in order to achieve a complete, detailed description of the conformational heterogeneity in membrane protein structural ensembles.

6.2 Uncovering insights into membrane protein dynamics

6.2.1 Dynamical activation of PLN and its complex with SERCA

The role of phospholamban (PLN) as a modulator of the calcium pump SERCA is known to be governed by a complex dynamical equilibrium.[54] The application of the ssNMR-driven MD sampling methodology to PLN revealed some interesting insights into the behaviour of this crucial regulatory protein. The exhaustive structural ensembles generated for PLN in its pentameric state reveal a drastic conformational transition upon the naturally-occurring phosphorylation at position 16 (pS16).

Specifically, the pentamer is activated by increasing its topological dynamics in its transmembrane helical bundle, with helices adopting a wide range of orientations with respect to the membrane. This dynamical activation is manifested by the isotropic scaling of the CSA and DC, and occurs at the longer timescales sampled by PISEMA experiments, in the order of milliseconds. The faster (nanosecond) internal dynamics of the transmembrane segment of PLN are unaffected by phosphorylation, as evidenced by the amide S^2 order parameters of the ensemble. The cytoplasmic domain Ia, on the other hand, displays increased dynamics in the nanosecond timescale. At the pentameric state, the motions of transmembrane and cytoplasmic segments of PLN are therefore coupled under different timescales. This evidence could only be obtained through the application of detailed, experimental dynamical models, and had been elusive in previous structural studies of the PLN pentamer, both experimental and theoretical.[276,292,357,358] The importance of such dynamical coupling between the transmembrane and cytoplasmic regions of PLN has been shown to be important in the monomeric state, with important consequences in the regulation of SERCA.[359] In the pentameric case, the described dynamical activation may play a role in the alteration of the equilibrium between monomeric and pentameric species. Controlling the relative populations of each oligomeric species is a crucial step in ensuring that the right concentration of free monomer is available to bind with SERCA,

and therefore guaranteeing that the pump is not over- or under-regulated.[276,360]

The structural ensembles of the pentamer also allow for the evaluation of a previous hypothesis regarding PLN: its ability to conduct calcium ions. Despite initial electrochemical studies in favour of this hypothesis,[288] other structural, biophysical and theoretical studies have shown contradicting evidence.[276,291,292,361] However, a complete evaluation of the energetics involved in ion conductance requires a full dynamical description, since the passage of ions could in principle be achieved through PLN populating transient conformations with a more open quaternary pore, resulting in a lowered energy barrier for the entry of Ca^{2+}. In analysing the ensembles there were no such conformations that would allow favourable ion passage through the pore. There are some distortions of the pore achieved by a collective twisting of the transmembrane helices, which however retains the major hydrophobic bottlenecks that impede ion passage. Even upon phosphorylation, the characteristics of the pore are highly conserved, despite the large-scale dynamical activation observed. Overall, the PLN structural ensembles here described dismiss the possibility of favourable Ca^{2+} channelling through the quaternary pore of the pentameric assembly.

In addition to studying the PLN pentamer, the experimentally restrained simulations were applied to the challenging PLN-SERCA complex, a highly heterogeneous system that exists in a variety of very slow interconverting states.[54] Using information from CSA, DC, PRE and crosslinking experiments, a full atomic-level structural complex of PLN-SERCA was built and sampled in its B-state, where domain Ia of PLN adopts a fully disordered conformation that interacts with the cytoplasmic domains of SERCA. Independent measurements of isotropic chemical shifts are in excellent agreement with the resulting structural ensemble of the complex, confirming the disordered nature of domain Ia.

The main feature of the PLN-SERCA ensemble is a great degree of heterogeneity in the cytoplasmic interaction surface of the proteins. Domain Ia of PLN retains an unstructured conformation, establishing fuzzy contacts across two of the cytoplasmic domains of SERCA (N and P). The absence of a well-defined binding mode at the cytoplasmic in-

terface is significant in the biological context of SERCA regulation. Transient disorder-mediated complexes have been shown to provide functional advantages in situations where a controlled and reversible regulation of protein activity is required.[362] In contrast to a highly structured strongly interacting complex, the transient contacts of phospholamban can be readily modulated in response to external changes during the sequential phases of the cardiac cycle, leading to a fast and direct impact in the allosteric regulation of SERCA. The disordered, heterogeneous nature of domain Ia may also have implications for its regulation through phosphorylation. Indeed, it has been shown that protein disorder plays a crucial role in the ability of residues to reversibly undergo this type of post-translational modifications.[363]

A crucial driving factor in the cytoplasmic interaction of SERCA and PLN are electrostatic forces. The exposed surface of interaction in SERCA features several patches of strong charge, both positive and negative. Domain Ia of PLN shows prominent interactions with those patches, primarily through positively charged arginine sidechains (R9 and the R13/R14 pair). Interestingly, these positions are both affected by pathological mutations, in the form of R9C and R14del,[306,307] suggesting that pathology may arise from a disruption of the SERCA-PLN complex leading to the misregulation of cardiac contractility. Electrostatic interactions between the proteins still exhibit a heterogeneous nature, with multiple distant sites (up to four in the case of R9) being contacted by PLN throughout the ensemble. This promiscuity in the interactions of PLN may be further promoted by the presence in SERCA of destabilising positive charges in close proximity to the exposed negative patches of interaction. An electrostatic basis of interaction between SERCA and PLN also fits well with recent findings that oligonucleotides can disrupt the PLN-SERCA complex, and fully restore SERCA activity.[364] Indeed, the presence of charge in the nucleotides was found to be essential for their activity.[365] Having an accurate atomic level description of the process, as provided by the ensembles here described, can be an important step toward the rational design of such therapeutic strategies.

6.2.2 Conformational dynamics and hydration properties of the M2 influenza channel

The dynamics of the M2 proton channel from influenza have been the subject of many studies, due to their essential role in the development and spreading of the disease. A complete molecular understanding of the mechanism of proton transport by M2 must bring together elements from experimental and modelling studies, due to the complexity of the process and high-resolution detail required to describe the molecular interactions involved. In this work, a dataset of CSA and DC measured in aligned phospholipid bilayers was employed to obtained experimentally-guided structural ensembles of M2.

The analysis of the M2 channel profile revealed an overall narrow pore, with a central cavity that is primarily delimited by two gates formed by the sidechains of residues V27 and H37. The C-terminal bottleneck is formed by the highly conserved histidine residue known to be the main responsible for pH sensing and proton conductance,[366] whereas the N-terminal gate is subject to a mutation, V27A, that confers resistance to adamantane drugs.[367] Replacing the pore-facing valine sidechains with smaller alanine is likely to disrupt the entry gate of the channel, enabling a higher degree of solvent diffusion into the pore's inner cavity.

The ensembles generated in this work revealed some topological motions of the transmembrane helices, which were correlated to considerable changes in the radius of the pore. The rearrangements of helical orientations leads to a low populated subset of structures that feature an exceptionally tightly packed quaternary interface, drastically reducing the accessibility of the channel to the solvent and imposing additional barriers to proton diffusion. More drastic topological motions have indeed been suggested to occur (with the opposite effect) as a result of lowering the pH, which would render the channel active by relieving the primary bottlenecks of the pore.[324]

The analyses of M2 also provided insights into the structure of water molecules within the transmembrane pore. The computation of water densities revealed the presence of several

highly structured clusters of water in the inner cavity delimited by V27 and H37. This increased density of water molecules correlates well with observations of electron density from water molecules at those positions in several crystal structures of the channel.[314,368] This ordering of water within the M2 pore has been described by other simulation studies, which point toward an ice-like transition once water enters the pore.[327,332,369]

The ordered water clusters confined in the pore were found to be stabilised by a dense network of hydrogen bonds, facilitated by interactions with protein chemical groups. Specifically, the sidechain of S31 was found to establish prevalent hydrogen bonds associated with the most water-dense clusters. The importance of this residue within the channel has been previously described in relation to its binding to adamantane drugs, as it establishes key interactions that are essential for drug binding,[70,370] with water dynamics playing a crucial role in mediating these interactions.[371] These interactions are indeed disrupted in the most widespread drug-resistant form of the virus, S31N.[368]

In addition to S31, the ordering of water molecules inside the cavity is facilitated by the carbonyl group of G34, the residue that induces the topological kink in the transmembrane helices. The interplay between this residue and water has indeed been evaluated by infrared spectroscopy.[372] These experimental studies demonstrate an overall rigidity in the structure of water at high pH when the channel is inactive, which is relieved when lowering the pH. Hydrogen bonds between G34 and water have also been shown to be highly prevalent in simulation studies.[332]

The M2 channel is an excellent example of a membrane protein system where collective dynamics play a crucial role in its function. The inclusion of experimental observables into molecular models provides an increased level of accuracy in describing the conformational fluctuations of the channel, which are heavily influenced by transmembrane helical reorientations. In this context, the information provided by aligned ssNMR experiments under different conditions would be highly valuable in elucidating the details of the proton conductance mechanism by M2.

6.2.3 Allosteric motions revealed in the IL-8 receptor CXCR1

A very prominent class of membrane proteins is formed by G-protein coupled receptors (GPCRs). These molecules are involved in a wide range of biological functions, and their misbehaviour often leads to disease. They have been the subject of much structural and biochemical study, and are currently the major class of proteins targeted therapeutically.[43] Recent developments in solid-state NMR allowed the measurement of dipolar couplings (DCs) for the GPCR CXCR1 in rotationally aligned (RA) samples.[81] These DCs were here employed to generate an experimentally-driven structural ensemble of the chemokine receptor CXCR1.

In its ligand-free form, the CXCR1 ensemble is characterised by considerable conformational fluctuations, particularly regarding the topological orientations of transmembrane helices. All helices in the structure exhibit some degree of kinking about residues that serve as flexible hinges for helical reorientation. The observed topological motions of helices are found to be mostly codependent with one another, resulting in collective motions along the hydrophobic interfaces of the transmembrane helices. An example of such a collective motion involving helices 5 and 6 was outlined, highlighting its consequences in the conformational rearrangements of intra- and extracellular loops, which are the surface exposed regions of the GPCR that interact with the G-protein and the ligand respectively.[352]

In proteins, these global structural rearrangements are generally facilitated via an allosteric communication network, an evolved series of interactions that connect distant parts of the protein.[354] Using the information from the DC-driven structural ensembles, the allosteric network of CXCR1 was modelled employing mathematical methods from the field of graph theory. These analyses revealed that helical kinks (particularly those of helices 3 and 5) are crucial points in the communication network, establishing long-range connections between distant regions of the receptor. These kinks in fact hold a central position in the optimal communication pathways between intra- and extracellular loops (as illustrated with intracellular loop 3 and extracellular loop 2). The central role of kinks in

connecting the two faces of the receptor points to an important functional role of these hinges in allosterically relaying the molecular signal after ligand binding. In fact, topological reorientations of helices have been observed when comparing experimental structures of active and inactive GPCRs.[373] Other studies point at the conservation of kink-inducing proline residues in GPCR families,[374] as well as the more general conservation and evolution of structural kinks as a means of functional diversification.[351] Mutations of residues associated with helical kinks have also been shown to have an effect in the activity of GPCRs.[375]

The experimentally-driven ensembles of CXCR1 provide an atomic level description of the allosteric mechanisms of GPCRs that result from their topological flexibility. Given the high degree of orientational information that DCs carry, their inclusion into theoretical models was essential to achieve an accurate description of the process. Applying these sampling methodologies to other GPCRs could allow the identification of their unique structural and topological properties in relation to their specific biological roles.

6.3 Future perspectives

The work outlined in this thesis resulted in some encouraging results and marks more than one avenue forward to expand the present research into the study of membrane protein dynamics.

With regards to accurately modelling chemical shifts in folded proteins, I have shown that a relatively accurate model can be built by only considering protein sequence. In fact, this model improves the accuracy of Camshift, the only method currently available to perform dynamical refinements in proteins. The inclusion of structural information into the $\text{PROSECCO}_{\text{FOLDED}}$ framework could therefore greatly enhance the accuracy of the CS models available for MD-based ensemble calculations. A common limitation to the employment of experimental terms as restraints in theoretical calculations is that the structural model of the observable must be differentiable with respect to the atomic Cartesian

coordinates. Incorporating structural terms such as interatomic angles or distances into the $\text{PROSECCO}_{\text{FOLDED}}$ neural network would yield an accurate, differentiable model that can be employed in such refinements.

The application of orientational ssNMR data (CSA and DCs) in MD samplings of membrane protein ensembles has here been greatly advanced. The biological systems here studied are considerably greater in size and complexity than any of the previously addressed.[201] The advances in oriented ssNMR[272] as well as the development of new methodologies to recover CSA and DC from magic-angle-spinning ssNMR samples[376] is an encouraging step toward the further applicability of these data in determining structural ensembles of important membrane protein systems.

In a wider context, the results from this investigation advocate for the integration of experimental data into theoretical models of protein dynamics. The two have been shown to highly benefit from each other: experimental information can improve the accuracy in the description provided by theoretical models, whereas theoretical models can provide a high level of detail into protein dynamics, and are often necessary for the correct interpretation of the data. In coming years, modelling tools such as molecular dynamics will only become more accessible as a result of the increase in overall computational power. Finally, the insights gained from integrating theory and experiments can be fed back into the models *a posteriori*, assisting in the re-parametrisation of empirical force fields.

Going beyond the scope of biomolecular dynamics, we are experiencing a drastic increase in the transfer of knowledge between different disciplines of science. This new era of scientific research is an exciting place for experiments and theory not only to coexist, but to complement each other. It is by integrating all of our available tools and information that we can be strongest in facing the complex challenge of understanding nature.

Bibliography

[1] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181(4610):662–666, 1958.

[2] M. F. Perutz, M. G. Rossmann, Ann F. Cullis, Hilary Muirhead, Georg Will, and A. C. T. North. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature*, 185(4711):416–422, feb 1960.

[3] M. F. Perutz. X-ray Analysis of Hemoglobin. *Science*, 140(3569):863–869, may 1963.

[4] C. J. Epstein, R. F. Goldberger, and C. B. Anfinsen. The Genetic Control of Tertiary Protein Structure: Studies With Model Systems. *Cold Spring Harbor Symposia on Quantitative Biology*, 28:439–449, jan 1963.

[5] C B Anfinsen. Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, 181(4096):223–30, jul 1973.

[6] P E Leopold, M Montal, and J N Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences of the United States of America*, 89(18):8721–5, sep 1992.

[7] P. Wolynes, J. Onuchic, and D Thirumalai. Navigating the folding routes. *Science*, 267(5204):1619–1620, mar 1995.

[8] K a Dill and H S Chan. From Levinthal to pathways to funnels. *Nature structural biology*, 4(1):10–19, 1997.

[9] Christopher M. Dobson, Andrej Sali, and Martin Karplus. Protein folding: A perspective from theory and experiment. *Angewandte Chemie - International Edition*, 37(7):868–893, 1998.

[10] Michael Levitt and Shneior Lifson. Refinement of protein conformations using a macromolecular energy minimization procedure. *Journal of Molecular Biology*, 46(2):269–279, dec 1969.

[11] J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, jun 1977.

[12] Joël Janin and Shoshanna J. Wodak. Structural domains in proteins and their role in the dynamics of protein function. *Progress in Biophysics and Molecular Biology*, 42:21–78, jan 1983.

[13] H. R. Faber and B. W. Matthews. A mutant T4 lysozyme displays five different crystal conformations. *Nature*, 348(6298):263–266, nov 1990.

[14] Mark Gerstein, Arthur M. Lesk, and Cyrus Chothia. Structural Mechanisms for Domain Movements in Proteins. *Biochemistry*, 33(22):6739–6749, jun 1994.

[15] Hans Frauenfelder, Stephen G Sligar, and Peter G Wolynes. The Energy Landscapes and Motions of Proteins. *Science*, 254:1598–1603, 1991.

[16] Alexander Koglin, Frank Löhr, Frank Bernhard, Vladimir V. Rogov, Dominique P. Frueh, Eric R. Strieter, Mohammad R. Mofid, Peter Güntert, Gerhard Wagner, Christopher T. Walsh, Mohamed A. Marahiel, and Volker Dötsch. Structural basis for the selectivity of the external thioesterase of the surfactin synthetase. *Nature*, 454(7206):907–911, 2008.

[17] J. P. Junker, F. Ziegler, and M. Rief. Ligand-Dependent Equilibrium Fluctuations of Single Calmodulin Molecules. *Science*, 323(5914):633–637, jan 2009.

[18] Alexander Koglin. Conformational Switches Modulate Protein Interactions in Peptide Antibiotic Synthetases. *Science*, 312(5771):273–276, apr 2006.

[19] James S. Fraser, Michael W. Clarkson, Sheena C. Degnan, Renske Erion, Dorothee Kern, and Tom Alber. Hidden alternative structures of proline isomerase essential for catalysis. *Nature*, 462(7273):669–673, 2009.

[20] Elan Z Eisenmesser, Oscar Millet, Wladimir Labeikovsky, Dmitry M Korzhnev, Magnus Wolf-Watz, Daryl a Bosco, Jack J Skalicky, Lewis E Kay, and Dorothee Kern. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(7064):117–121, 2005.

[21] David D. Boehr, Dan McElheny, H. Jane Dyson, and Peter E. Wright. The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. *Science*, 313(5793):1638–1642, 2006.

[22] Magnus Wolf-Watz, Vu Thai, Katherine Henzler-Wildman, Georgia Hadjipavlou, Elan Z. Eisenmesser, and Dorothee Kern. Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nature Structural and Molecular Biology*, 11(10):945–949, 2004.

[23] Katherine A. Henzler-Wildman, Vu Thai, Ming Lei, Maria Ott, Magnus Wolf-Watz, Tim Fenn, Ed Pozharski, Mark A. Wilson, Gregory A. Petsko, Martin Karplus, Christian G. Hübner, and Dorothee Kern. Intrinsic motions along an enzymatic reaction trajectory. *Nature*, 450(7171):838–844, 2007.

[24] G Wagner. Characterization of the distribution of internal motions in the basic pancreatic trypsin inhibitor using a large number of internal NMR probes. *Quarterly reviews of biophysics*, 16(1):1–57, feb 1983.

[25] J. Ma and M. Karplus. The allosteric mechanism of the chaperonin GroEL: A dynamic analysis. *Proceedings of the National Academy of Sciences*, 95(15):8502–8507, 1998.

[26] M Karplus and J Kuriyan. Molecular dynamics and protein function. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 102(19):6679–6685, 2005.

[27] Gabriel E. Zentner and Steven Henikoff. Regulation of nucleosome dynamics by histone modifications. *Nature Structural and Molecular Biology*, 20(3):259–266, 2013.

[28] Qi Zhang, Andrew C. Stelzer, Charles K. Fisher, and Hashim M. Al-Hashimi. Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature*, 450(7173):1263–1267, 2007.

[29] Charalampos G Kalodimos, Nikolaos Biris, Alexandre M J J Bonvin, Marc M Levandoski, Marc Guennuegues, Rolf Boelens, and Robert Kaptein. Structure and Flexibility Adaptation in Nonspeci c and Speci c Protein-DNA Complexes. *Science*, 305(July):386–389, 2004.

[30] David D. Boehr, Ruth Nussinov, and Peter E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology*, 5(11):789–796, 2009.

[31] Michael P. Williamson, Timothy F. Havel, and Kurt Wüthrich. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. *Journal of Molecular Biology*, 182(2):295–315, 1985.

[32] Lewis E Kay. Protein dynamics from NMR. *Nature*, 5(Suppl S):513–517, 1998.

[33] Kresten Lindorff-Larsen, Robert B Best, Mark a Depristo, Christopher M Dobson, and Michele Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–132, 2005.

[34] H. Jane Dyson and Peter E. Wright. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, 6(3):197–208, 2005.

[35] M. Vendruscolo and C. M. Dobson. Towards complete descriptions of the free-energy landscapes of proteins. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363(1827):433–452, feb 2005.

[36] Stefano Piana, Kresten Lindorff-Larsen, and David E. Shaw. How robust are protein folding simulations with respect to force field parameterization? *Biophysical Journal*, 100(9):L47–L49, 2011.

[37] Kresten Lindorff-Larsen, Paul Maragakis, Stefano Piana, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Systematic validation of protein force fields against experimental data. *PLoS ONE*, 7(2):1–6, 2012.

[38] David E Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Yibing Shan, and Willy Wriggers. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science*, 330(October):341–347, 2010.

[39] H. S. Chung, S. Piana-Agostinetti, D. E. Shaw, and W. A. Eaton. Structural origin of slow diffusion in protein folding. *Science*, 349(6255):1504–1510, 2015.

[40] Stephen H. White. Biophysical dissection of membrane proteins. *Nature*, 459(7245):344–346, 2009.

[41] Matthieu Chavent, Anna L. Duncan, and Mark S.P. Sansom. Molecular dynamics simulations of membrane proteins and their interactions: From nanoscale to mesoscale. *Current Opinion in Structural Biology*, 40(Md):8–16, 2016.

[42] Markus Sällman Almén, Karl J V Nordström, Robert Fredriksson, and Helgi B. Schiöth. Mapping the human membrane proteome: A majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biology*, 7:50, 2009.

[43] John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. How many drug targets are there? *Nature reviews. Drug discovery*, 5(12):993–6, 2006.

[44] Mijo Simunovic, Anand Srivastava, and Gregory A Voth. Linear aggregation of proteins on the membrane as a prelude to membrane remodeling. *Proceedings of the National Academy of Sciences of the United States of America*, 110(51):20396–20401, 2013.

[45] Colin Kleanthous, Patrice Rassam, and Christoph G. Baumann. Protein-protein interactions and the spatiotemporal dynamics of bacterial outer membrane proteins. *Current Opinion in Structural Biology*, 35:109–115, 2015.

[46] Ka Young Chung, Søren G. F. Rasmussen, Tong Liu, Sheng Li, Brian T. DeVree, Pil Seok Chae, Diane Calinski, Brian K. Kobilka, Virgil L. Woods, and Roger K. Sunahara. Conformational changes in the G protein Gs induced by the $\beta 2$ adrenergic receptor. *Nature*, 477(7366):611–615, 2011.

[47] Andrew C. Kruse, Jianxin Hu, Albert C. Pan, Daniel H. Arlow, Daniel M. Rosenbaum, Erica Rosemond, Hillary F. Green, Tong Liu, Pil Seok Chae, Ron O. Dror, David E. Shaw, William I. Weis, Jürgen Wess, and Brian K. Kobilka. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature*, 482(7386):552–556, 2012.

[48] Joel R Meyerson, Janesh Kumar, Sagar Chittori, Prashant Rao, Jason Pierson, Alberto Bartesaghi, Mark L Mayer, and Sriram Subramaniam. Structural mechanism of glutamate receptor activation and desensitization. *Nature*, 514(7522):328–334, 2014.

[49] Tatsuro Shimamura, Simone Weyand, Oliver Beckstein, Nicholas G Rutherford, Jonathan M Hadden, David Sharples, Mark S P Sansom, So Iwata, Peter J F Henderson, and Alexander D Cameron. Molecular Basis of Alternating Access Membrane Transport by the Sodium-Hydantoin Transporter Mhp1. *Science*, 328(5977):470–473, apr 2010.

[50] Zoltan Bozoky, Mickael Krzeminski, Ranjith Muhandiram, James R Birtley, Ateeq Al-zahrani, Philip J Thomas, Raymond A Frizzell, Robert C Ford, and Julie D Forman-Kay. Regulatory R region of the CFTR chloride channel is a dynamic integrator of phospho-dependent intra- and intermolecular interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 110(47):E4427–E4436, 2013.

[51] Mihaela Anitei and Bernard Hoflack. Bridging membrane and cytoskeleton dynamics in the secretory and endocytic pathways. *Nature Cell Biology*, 14(1):11–19, 2012.

[52] D H MacLennan, C J Brandl, B Korczak, and N M Green. Amino-acid sequence of a Ca2+ + Mg2+-dependent ATPase from rabbit muscle sarcoplasmic reticulum, deduced from its complementary DNA sequence. *Nature*, 316(6030):696–700, 1985.

[53] Yoshiyuki Norimatsu, Kazuya Hasegawa, Nobutaka Shimizu, and Chikashi Toyoshima. Protein-phospholipid interplay revealed with crystals of a calcium pump. *Nature*, 545(7653):193–198, 2017.

[54] Martin Gustavsson, Raffaello Verardi, Daniel G Mullen, Kaustubh R Mote, Nathaniel J Traaseth, T Gopinath, and Gianluigi Veglia. Allosteric regulation of SERCA by phosphorylation-mediated conformational shift of phospholamban. *Proceedings of the National Academy of Sciences of the United States of America*, 110(43):17338–43, oct 2013.

[55] Marisa Brini and Ernesto Carafoli. Calcium Pumps in Health and Disease. *Physiological Reviews*, 89(4):1341–1378, oct 2009.

[56] William A. Catterall. Ion Channel Voltage Sensors: Structure, Function, and Pathophysiology. *Neuron*, 67(6):915–928, sep 2010.

[57] Sulayman D. Dib-Hajj, Theodore R. Cummins, Joel A. Black, and Stephen G. Waxman. From genes to pain: Nav1.7 and human pain disorders. *Trends in Neurosciences*, 30(11):555–563, nov 2007.

[58] Naomi R. Latorraca, A. J. Venkatakrishnan, and Ron O. Dror. GPCR dynamics: Structures in motion. *Chemical Reviews*, 117(1):139–155, 2017.

[59] David S Feldman, Cynthia A Carnes, William T Abraham, and Michael R Bristow. Mechanisms of Disease: β-adrenergic receptorsalterations in signal transduction and pharmacogenomics in heart failure. *Nature Clinical Practice Cardiovascular Medicine*, 2(9):475–483, sep 2005.

[60] Fabrizio Chiti and Christopher M Dobson. Protein misfolding, functional amyloid, and human disease. *Annual review of biochemistry*, 75:333–66, 2006.

[61] Gabor G. Kovacs and Herbert Budka. Prion Diseases: From Protein to Cell Pathology. *The American Journal of Pathology*, 172(3):555–565, mar 2008.

[62] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic acids research*, 28(1):235–42, jan 2000.

[63] Linn Fagerberg, Kalle Jonasson, Gunnar Von Heijne, Mathias Uhlén, and Lisa Berglund. Prediction of the human membrane proteome. *Proteomics*, 10(6):1141–1149, 2010.

[64] R. Henderson and P. N. T. Unwin. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature*, 257(5521):28–32, sep 1975.

[65] J. Deisenhofer, O. Epp, K. Miki, R. Huber, and H. Michel. Structure of the protein subunits in the photosynthetic reaction centre of Rhodopseudomonas viridis at 3Å resolution. *Nature*, 318(6047):618–624, dec 1985.

[66] S. Faham, A. Watanabe, G. M. Besserer, D. Cascio, A. Specht, B. A. Hirayama, E. M. Wright, and J. Abramson. The Crystal Structure of a Sodium Galactose Transporter Reveals Mechanistic Insights into Na+/Sugar Symport. *Science*, 321(5890):810–814, aug 2008.

[67] Simone Weyand, Tatsuro Shimamura, Shunsuke Yajima, Shun'ichi Suzuki, Osman Mirza, Kuakarun Krusong, Elisabeth P Carpenter, Nicholas G Rutherford, Jonathan M Hadden, John O'Reilly, Pikyee Ma, Massoud Saidijam, Simon G Patching, Ryan J Hope, Halina T Norbertczak, Peter C J Roach, So Iwata, Peter J F Henderson, and Alexander D Cameron. Structure and molecular mechanism of a nucleobase-cation-symport-1 family transporter. *Science (New York, N.Y.)*, 322(5902):709–13, oct 2008.

[68] Jianping Wu, Zhen Yan, Zhangqiang Li, Chuangye Yan, Shan Lu, Mengqiu Dong, and Nieng Yan. Structure of the voltage-gated calcium channel Cav1.1 complex. *Science (New York, N.Y.)*, 350(6267):aad2395, dec 2015.

[69] Zhen Yan, Xiaochen Bai, Chuangye Yan, Jianping Wu, Zhangqiang Li, Tian Xie, Wei Peng, Changcheng Yin, Xueming Li, Sjors H W Scheres, Yigong Shi, and Nieng Yan. Structure of the rabbit ryanodine receptor RyR1 at near-atomic resolution. *Nature*, 517(7532):50–55, jan 2015.

[70] Sarah D Cady, Klaus Schmidt-Rohr, Jun Wang, Cinque S Soto, William F Degrado, and Mei Hong. Structure of the amantadine binding site of influenza M2 proton channels in lipid bilayers. *Nature*, 463(7281):689–692, 2010.

[71] Antoine Loquet, Nikolaos G Sgourakis, Rashmi Gupta, Karin Giller, Dietmar Riedel, Christian Goosmann, Christian Griesinger, Michael Kolbe, David Baker, Stefan Becker, and Adam Lange. Atomic model of the type III secretion system needle. *Nature*, 486(7402):276–9, may 2012.

[72] Harini Krishnamurthy and Eric Gouaux. X-ray structures of LeuT in substrate-free outward-open and apo inward-open states. *Nature*, 481(7382):469–474, jan 2012.

[73] S. Asano, Y. Fukuda, F. Beck, A. Aufderheide, F. Forster, R. Danev, and W. Baumeister. A molecular census of 26S proteasomes in intact neurons. *Science*, 347(6220):439–442, jan 2015.

[74] Kazuyoshi Murata and Matthias Wolf. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1862(2):324–334, feb 2018.

[75] Jianhua Zhao, Samir Benlekbir, and John L. Rubinstein. Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase. *Nature*, 521(7551):241–245, 2015.

[76] Xiaohua Wang, Xiaojuan Li, Xin Deng, Doan-Trung Luu, Christophe Maurel, and Jinxing Lin. Single-molecule fluorescence imaging to quantify membrane protein dynamics and oligomerization in living plant cells. *Nature Protocols*, 10(12):2054–2063, nov 2015.

[77] Christopher King, Valerica Raicu, and Kalina Hristova. Understanding the FRET Signatures of Interacting Membrane Proteins. *Journal of Biological Chemistry*, 292(13):5291–5310, mar 2017.

[78] Valeriya M Trusova, Galyna P Gorbenko, Todor Deligeorgiev, and Nikolai Gadjev. Probing protein-lipid interactions by FRET between membrane fluorophores. *Methods and Applications in Fluorescence*, 4(3):034014, sep 2016.

[79] Michael P. Bokoch, Yaozhong Zou, Søren G. F. Rasmussen, Corey W. Liu, Rie Nygaard, Daniel M. Rosenbaum, Juan José Fung, Hee-Jung Choi, Foon Sun Thian, Tong Sun Kobilka, Joseph D. Puglisi, William I. Weis, Leonardo Pardo, R. Scott Prosser, Luciano Mueller, and Brian K. Kobilka. Ligand-specific regulation of the extracellular surface of a G-protein-coupled receptor. *Nature*, 463(7277):108–112, jan 2010.

[80] Daryl B. Good, Shenlin Wang, Meaghan E. Ward, Jochem Struppe, Leonid S. Brown, Józef R. Lewandowski, and Vladimir Ladizhansky. Conformational Dynamics of a Seven Transmembrane Helical Protein Anabaena Sensory Rhodopsin Probed by Solid-State NMR. *Journal of the American Chemical Society*, 136(7):2833–2842, feb 2014.

[81] Sang Ho Park, Bibhuti B. Das, Fabio Casagrande, Ye Tian, Henry J. Nothnagel, Mignon Chu, Hans Kiefer, Klaus Maier, Anna A. De Angelis, Francesca M. Marassi, and Stanley J. Opella. Structure of the chemokine receptor CXCR1 in phospholipid bilayers. *Nature*, 491(7426):779–783, 2012.

[82] Adam Lange, Karin Giller, Sönke Hornig, Marie-France Martin-Eauclaire, Olaf Pongs, Stefan Becker, and Marc Baldus. Toxin-induced conformational changes in a potassium channel revealed by solid-state NMR. *Nature*, 440(7086):959–962, apr 2006.

[83] Matthias Hiller, Ludwig Krabben, Kutti R Vinothkumar, Federica Castellani, Barth-Jan van Rossum, Werner Kühlbrandt, and Hartmut Oschkinat. Solid-state magic-angle spinning NMR of outer-membrane protein G from Escherichia coli. *Chembiochem : a European journal of chemical biology*, 6(9):1679–84, sep 2005.

[84] P. M. Hwang, W.-Y. Choy, E. I. Lo, L. Chen, J. D. Forman-Kay, C. R. H. Raetz, G. G. Prive, R. E. Bishop, and L. E. Kay. Solution structure and dynamics of the outer membrane enzyme PagP by NMR. *Proceedings of the National Academy of Sciences*, 99(21):13560–13565, oct 2002.

[85] Yan Sun and David P Weliky. 13C-13C correlation spectroscopy of membrane-associated influenza virus fusion peptide strongly supports a helix-turn-helix motif and two turn conformations. *Journal of the American Chemical Society*, 131(37):13228–9, sep 2009.

[86] Pedro E. M. Lopes, Olgun Guvench, and Alexander D. MacKerell. Current Status of Protein Force Fields for Molecular Dynamics Simulations. In *Molecular Modeling of Proteins*, pages 47–71. Springer, 2015.

[87] Kristen A. Fichthorn and W. H. Weinberg. Theoretical foundations of dynamical Monte Carlo simulations. *The Journal of Chemical Physics*, 95(2):1090–1096, jul 1991.

[88] Richard W. Pastor. Molecular dynamics and Monte Carlo simulations of lipid bilayers. *Current Opinion in Structural Biology*, 4(4):486–492, jan 1994.

[89] Sanguk Kim, Aaron K. Chamberlain, and James U. Bowie. A Simple Method for Modeling Transmembrane Helix Oligomers. *Journal of Molecular Biology*, 329(4):831–840, jun 2003.

[90] Amit Kessel, Dalit Shental-Bechor, Turkan Haliloglu, and Nir Ben-Tal. Interactions of hydrophobic peptides with lipid bilayers: Monte Carlo simulations with M2delta. *Biophysical journal*, 85(6):3431–44, dec 2003.

[91] David J. Wales. Energy landscapes: calculating pathways and rates. *International Reviews in Physical Chemistry*, 25(1-2):237–282, jan 2006.

[92] David J. Wales. Discrete path sampling. *Molecular Physics*, 100(20):3285–3305, oct 2002.

[93] Zoe Cournia, Toby W. Allen, Ioan Andricioaei, Bruno Antonny, Daniel Baum, Grace Brannigan, Nicolae-Viorel Buchete, Jason T. Deckman, Lucie Delemotte, Coral del Val, Ran Friedman, Paraskevi Gkeka, Hans-Christian Hege, Jérôme Hénin, Marina A. Kasimova, Antonios Kolocouris, Michael L. Klein, Syma Khalid, M. Joanne Lemieux, Norbert Lindow, Mahua Roy, Jana Selent, Mounir Tarek, Florentina Tofoleanu, Stefano Vanni, Sinisa Urban, David J. Wales, Jeremy C. Smith, and Ana-Nicoleta Bondar. Membrane Protein Structure, Function, and Dynamics: a Perspective from Experiments and Theory. *The Journal of Membrane Biology*, 248(4):611–640, aug 2015.

[94] Birgit Strodel, Jason W. L. Lee, Christopher S. Whittleston, and David J. Wales. Transmembrane Structures for Alzheimer's Aβ 142 Oligomers. *Journal of the American Chemical Society*, 132(38):13300–13312, sep 2010.

[95] Ron O. Dror, Robert M. Dirks, J.P. Grossman, Huafeng Xu, and David E. Shaw. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annual Review of Biophysics*, 41(1):429–452, 2012.

[96] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science (New York, N.Y.)*, 334(6055):517–20, oct 2011.

[97] J. D. Faraldo-Gomez and B. Roux. On the importance of a funneled energy landscape for the assembly and regulation of multidomain Src tyrosine kinases. *Proceedings of the National Academy of Sciences*, 104(34):13643–13648, aug 2007.

[98] R. O. Dror, D. H. Arlow, D. W. Borhani, M. O. Jensen, S. Piana, and D. E. Shaw. Identification of two distinct inactive conformations of the 2-adrenergic receptor reconciles structural and biochemical observations. *Proceedings of the National Academy of Sciences*, 106(12):4689–4694, mar 2009.

[99] A. Barducci, M. Bonomi, M. K. Prakash, and M. Parrinello. Free-energy landscape of protein oligomerization from atomistic simulations. *Proceedings of the National Academy of Sciences*, 110(49):E4708–E4713, dec 2013.

[100] Sergei Yu. Noskov, Simon Bernèche, and Benoît Roux. Control of ion selectivity in potassium channels by electrostatic and dynamic properties of carbonyl ligands. *Nature*, 431(7010):830–834, oct 2004.

[101] M. O. Jensen, D. W. Borhani, K. Lindorff-Larsen, P. Maragakis, V. Jogini, M. P. Eastwood, R. O. Dror, and D. E. Shaw. Principles of conduction and hydrophobic gating in K+ channels. *Proceedings of the National Academy of Sciences*, 107(13):5833–5838, mar 2010.

[102] D. A. Kopfer, C. Song, T. Gruene, G. M. Sheldrick, U. Zachariae, and B. L. de Groot. Ion permeation in K+ channels occurs by direct Coulomb knock-on. *Science*, 346(6207):352–355, oct 2014.

[103] Lucie Delemotte, Mounir Tarek, Michael L Klein, Cristiano Amaral, and Werner Treptow. Intermediate states of the Kv1.2 voltage sensor from atomistic molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(15):6109–14, apr 2011.

[104] R. O. Dror, D. H. Arlow, P. Maragakis, T. J. Mildorf, A. C. Pan, H. Xu, D. W. Borhani, and D. E. Shaw. Activation mechanism of the 2-adrenergic receptor. *Proceedings of the National Academy of Sciences*, 108(46):18684–18689, nov 2011.

[105] Ron O. Dror, Albert C. Pan, Daniel H. Arlow, David W. Borhani, Paul Maragakis, Yibing Shan, Huafeng Xu, and David E. Shaw. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, 108(32):13118–13123, aug 2011.

[106] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151, nov 1999.

[107] Christoph Dellago, Peter G. Bolhuis, Félix S. Csajka, and David Chandler. Transition path sampling and the calculation of rate constants. *The Journal of Chemical Physics*, 108(5):1964–1977, feb 1998.

[108] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, feb 1977.

[109] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12562–12566, 2002.

[110] F. Zhu and G. Hummer. Pore opening and closing of a pentameric ligand-gated ion channel. *Proceedings of the National Academy of Sciences*, 107(46):19814–19819, nov 2010.

[111] Haibo Yu, Ian M Ratheal, Pablo Artigas, and Benoît Roux. Protonation of key acidic residues is critical for the K+-selectivity of the Na/K pump. *Nature Structural & Molecular Biology*, 18(10):1159–1163, oct 2011.

[112] Alexey Onufriev. Continuum Electrostatics Solvent Modeling with the Generalized Born Model. In *Modeling Solvent Environments*, pages 127–165. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2010.

[113] Seiichiro Tanizaki and Michael Feig. A generalized Born formalism for heterogeneous dielectric environments: application to the implicit modeling of biological membranes. *The Journal of chemical physics*, 122(12):124706, mar 2005.

[114] Ramu Anandakrishnan, Aleksander Drozdetski, Ross C. Walker, and Alexey V. Onufriev. Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations. *Biophysical Journal*, 108(5):1153–1164, mar 2015.

[115] Siewert J. Marrink, Alex H. de Vries, and Alan E. Mark. Coarse Grained Model for Semiquantitative Lipid Simulations. *The Journal of Physical Chemistry B*, 108(2):750–760, jan 2004.

[116] 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, oct 2016.

[117] 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, jul 2007.

[118] D. Jefferies and S. Khalid. Molecular Simulations of Complex Membrane Models. In *Modeling of Microscale Transport in Biological Processes*, pages 1–18. Elsevier, 2017.

[119] John M.A. Grime and Gregory A. Voth. Early Stages of the HIV-1 Capsid Protein Lattice Formation. *Biophysical Journal*, 103(8):1774–1783, oct 2012.

[120] Tyler Reddy, David Shorthouse, Daniel L. Parton, Elizabeth Jefferys, Philip W. Fowler, Matthieu Chavent, Marc Baaden, and Mark S.P. Sansom. Nothing to Sneeze At: A Dynamic and Integrative Computational Model of an Influenza A Virion. *Structure*, 23(3):584–597, mar 2015.

[121] Xavier Periole, Adam M. Knepp, Thomas P. Sakmar, Siewert J. Marrink, and Thomas Huber. Structural Determinants of the Supramolecular Organization of G Protein-Coupled Receptors in Bilayers. *Journal of the American Chemical Society*, 134(26):10959–10965, jul 2012.

[122] Heidi Koldsø and Mark S. P. Sansom. Organization and Dynamics of Receptor Proteins in a Plasma Membrane. *Journal of the American Chemical Society*, 137(46):14694–14704, nov 2015.

[123] Thomas J. Piggot, Daniel A. Holdbrook, and Syma Khalid. Conformational dynamics and membrane interactions of the E. coli outer membrane protein FecA: A molecular dynamics simulation study. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1828(2):284–293, feb 2013.

[124] Fiona B Naughton, Antreas C Kalli, and Mark S P Sansom. Association of Peripheral Membrane Proteins with Membranes: Free Energy of Binding of GRP1 PH Domain with Phosphatidylinositol Phosphate-Containing Model Bilayers. *The journal of physical chemistry letters*, 7(7):1219–24, apr 2016.

[125] Mickaël Lelimousin, Vittorio Limongelli, and Mark S. P. Sansom. Conformational Changes in the Epidermal Growth Factor Receptor: Role of the Transmembrane Domain Investigated by Coarse-Grained MetaDynamics Free Energy Calculations. *Journal of the American Chemical Society*, 138(33):10611–10622, aug 2016.

[126] Jing Huang and Alexander D. MacKerell. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25):2135–2145, sep 2013.

[127] Paul Robustelli, Kate A. Stafford, and Arthur G. Palmer III. Interpreting protein structural dynamics from NMR chemical shifts. *J Am Chem Soc*, 134(14):6365–6374, 2012.

[128] Kyle A. Beauchamp, Yu-Shan Lin, Rhiju Das, and Vijay S. Pande. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *Journal of Chemical Theory and Computation*, 8(4):1409–1414, apr 2012.

[129] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences*, 109(44):17845–17850, oct 2012.

[130] M. B. Prigozhin and M. Gruebele. Microsecond folding experiments and simulations: a match is made. *Physical Chemistry Chemical Physics*, 15(10):3372, 2013.

[131] Filip Persson and Bertil Halle. Transient Access to the Protein Interior: Simulation versus NMR. *Journal of the American Chemical Society*, 135(23):8735–8748, jun 2013.

[132] Oliver F. Lange, David van der Spoel, and Bert L. de Groot. Scrutinizing Molecular Mechanics Force Fields on the Submicrosecond Timescale with NMR Data. *Biophysical Journal*, 99(2):647–655, jul 2010.

[133] Pedro E. M. Lopes, Olgun Guvench, and Alexander D. MacKerell. Current Status of Protein Force Fields for Molecular Dynamics Simulations. In *Molecular Modeling of Proteins*, pages 47–71. Springer, 2015.

[134] Evan S. O'Brien, A. Joshua Wand, and Kim A. Sharp. On the ability of molecular dynamics force fields to recapitulate NMR derived protein side chain order parameters. *Protein Science*, 25(6):1156–1160, jun 2016.

[135] Kalliopi K. Patapati and Nicholas M. Glykos. Three Force Fields' Views of the 310 Helix. *Biophysical Journal*, 101(7):1766–1771, oct 2011.

[136] Kresten Lindorff-Larsen, Paul Maragakis, Stefano Piana, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Systematic Validation of Protein Force Fields against Experimental Data. *PLoS ONE*, 7(2):e32131, feb 2012.

[137] Peter L. Freddolino, Sanghyun Park, Benoît Roux, and Klaus Schulten. Force Field Bias in Protein Folding Simulations. *Biophysical Journal*, 96(9):3772–3780, may 2009.

[138] Stefano Piana, John L Klepeis, and David E Shaw. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Current Opinion in Structural Biology*, 24:98–105, feb 2014.

[139] Alpan Raval, Stefano Piana, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, pages n/a–n/a, 2012.

[140] Wei Wang, Wei Ye, Cheng Jiang, Ray Luo, and Hai-Feng Chen. New force field on modeling intrinsically disordered proteins. *Chemical biology & drug design*, 84(3):253–69, sep 2014.

[141] Dong Song, Ray Luo, and Hai-Feng Chen. The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *Journal of Chemical Information and Modeling*, 57(5):1166–1178, may 2017.

[142] Robert B. Best, Wenwei Zheng, and Jeetain Mittal. Balanced ProteinWater Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation*, 10(11):5113–5124, nov 2014.

[143] Stefano Piana, Alexander G. Donchev, Paul Robustelli, and David E. Shaw. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *The Journal of Physical Chemistry B*, 119(16):5113–5123, apr 2015.

[144] Alexander P. Lyubartsev and Alexander L. Rabinovich. Force Field Development for Lipid Membrane Simulations. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1858(10):2483–2497, oct 2016.

[145] Ryan W Benz, Francisco Castro-Román, Douglas J Tobias, and Stephen H White. Experimental validation of molecular dynamics simulations of lipid bilayers: a new approach. *Biophysical journal*, 88(2):805–17, feb 2005.

[146] Louic S. Vermeer, Bert L. de Groot, Valérie Réat, Alain Milon, and Jerzy Czaplicki. Acyl chain order parameter profiles in phospholipid bilayers: computation from molecular dynamics simulations and comparison with 2H NMR experiments. *European Biophysics Journal*, 36(8):919–931, nov 2007.

[147] O.H. Samuli Ollila and Georg Pabst. Atomistic resolution structure and dynamics of lipid bilayers in simulations and experiments. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1858(10):2512–2528, oct 2016.

[148] Thomas J. Piggot, Ángel Piñeiro, and Syma Khalid. Molecular Dynamics Simulations of Phosphatidylcholine Membranes: A Comparative Force Field Study. *Journal of Chemical Theory and Computation*, 8(11):4593–4609, nov 2012.

[149] Xiaohong Zhuang, Judah R. Makover, Wonpil Im, and Jeffery B. Klauda. A systematic molecular dynamics simulation study of temperature dependent bilayer structural properties. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1838(10):2520–2529, oct 2014.

[150] Siewert J. Marrink and D. Peter Tieleman. Perspective on the Martini model. *Chemical Society Reviews*, 42(16):6801, 2013.

[151] Philip W. Fowler, Jean Hélie, Anna Duncan, Matthieu Chavent, Heidi Koldsø, and Mark S. P. Sansom. Membrane stiffness is modified by integral membrane proteins. *Soft Matter*, 12(37):7792–7803, 2016.

[152] 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, oct 2016.

[153] Henry van den Bedem and James S Fraser. Integrative, dynamic structural biology at atomic resolutionit's about time. *Nature Methods*, 12(4):307–318, apr 2015.

[154] Massimiliano Bonomi, Gabriella T. Heller, Carlo Camilloni, and Michele Vendruscolo. Principles of protein structural ensemble determination. *Current Opinion in Structural Biology*, 42:106–116, feb 2017.

[155] E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, may 1957.

[156] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, jul 1948.

[157] Wouter Boomsma, Jesper Ferkinghoff-Borg, and Kresten Lindorff-Larsen. Combining Experiments and Simulations Using the Maximum Entropy Principle. *PLoS Computational Biology*, 10(2):e1003406, feb 2014.

[158] Enrico Ravera, Luca Sgheri, Giacomo Parigi, and Claudio Luchinat. A critical assessment of methods to recover information from averaged data. *Physical Chemistry Chemical Physics*, 18(8):5686–5701, 2016.

[159] Simon Olsson, Jes Frellsen, Wouter Boomsma, Kanti V. Mardia, and Thomas Hamelryck. Inference of Structure Ensembles of Flexible Biomolecules from Sparse, Averaged Data. *PLoS ONE*, 8(11):e79439, nov 2013.

[160] Hoi Tik Alvin Leung, Olivier Bignucolo, Regula Aregger, Sonja A. Dames, Adam Mazur, Simon Bernèche, and Stephan Grzesiek. A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content. *Journal of Chemical Theory and Computation*, 12(1):383–394, jan 2016.

[161] W Y Choy and J D Forman-Kay. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *Journal of molecular biology*, 308(5):1011–32, may 2001.

[162] David H. Brookes and Teresa Head-Gordon. Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins. *Journal of the American Chemical Society*, 138(13):4530–4538, apr 2016.

[163] Bartosz Różycki, Young C. Kim, and Gerhard Hummer. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure*, 19(1):109–116, jan 2011.

[164] L D Antonov, S Olsson, W Boomsma, and T Hamelryck. Bayesian inference of protein ensembles from SAXS data. *Physical chemistry chemical physics : PCCP*, 18(8):5832–8, feb 2016.

[165] Anurag Sethi, Divina Anunciado, Jianhui Tian, Dung M. Vu, and S. Gnanakaran. Deducing conformational variability of intrinsically disordered proteins from infrared spectroscopy with Bayesian statistics. *Chemical Physics*, 422:143–155, aug 2013.

[166] Thomas H Edwards and Stefan Stoll. A Bayesian approach to quantifying uncertainty from experimental noise in DEER spectroscopy. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 270:87–97, 2016.

[167] Kyle A Beauchamp, Vijay S Pande, and Rhiju Das. Bayesian energy landscape tilting: towards concordant models of molecular ensembles. *Biophysical journal*, 106(6):1381–90, mar 2014.

[168] Gerhard Hummer and Jürgen Köfinger. Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of Chemical Physics*, 143(24):243150, dec 2015.

[169] Ivano Bertini, Andrea Giachetti, Claudio Luchinat, Giacomo Parigi, Maxim V. Petoukhov, Roberta Pierattelli, Enrico Ravera, and Dmitri I. Svergun. Conformational Space of Flexible Biological Macromolecules from Average Data. *Journal of the American Chemical Society*, 132(38):13553–13558, sep 2010.

[170] Konstantin Berlin, Carlos A Castañeda, Dina Schneidman-Duhovny, Andrej Sali, Alfredo Nava-Tudela, and David Fushman. Recovering a representative conformational ensemble from underdetermined macromolecular structural data. *Journal of the American Chemical Society*, 135(44):16595–609, nov 2013.

[171] Gabrielle Nodet, Loïc Salmon, Valéry Ozenne, Sebastian Meier, Malene Ringkjøbing Jensen, and Martin Blackledge. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *Journal of the American Chemical Society*, 131(49):17908–18, dec 2009.

[172] M. Ceriotti, G. A. R. Brain, O. Riordan, and D. E. Manolopoulos. The inefficiency of re-weighted sampling and the curse of system size in high-order path integration. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 468(2137):2–17, jan 2012.

[173] Jed W. Pitera and John D. Chodera. On the use of experimental observations to bias simulated ensembles. *Journal of Chemical Theory and Computation*, 8(10):3445–3451, 2012.

[174] Andrea Cavalli, Carlo Camilloni, and Michele Vendruscolo. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *The Journal of Chemical Physics*, 138(9):094112, mar 2013.

[175] Benoît Roux and Jonathan Weare. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *Journal of Chemical Physics*, 138(8):1–8, 2013.

[176] Alfonso De Simone, Barbara Richter, Xavier Salvatella, and Michele Vendruscolo. Toward an accurate determination of free energy landscapes in solution states of proteins. *Journal of the American Chemical Society*, 131(11):3810–1, mar 2009.

[177] Oliver F Lange, Nils-Alexander Lakomek, Christophe Farès, Gunnar F Schröder, Korvin F A Walter, Stefan Becker, Jens Meiler, Helmut Grubmüller, Christian Griesinger, and Bert L de Groot. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science (New York, N.Y.)*, 320(5882):1471–5, jun 2008.

[178] Kresten Lindorff-Larsen, Sigridur Kristjansdottir, Kaare Teilum, Wolfgang Fieber, Christopher M. Dobson, Flemming M. Poulsen, and Michele Vendruscolo. Determination of an Ensemble of Structures Representing the Denatured State of the Bovine Acyl-Coenzyme A Binding Protein. *Journal of the American Chemical Society*, 126(10):3291–3299, mar 2004.

[179] Matthew M Dedmon, Kresten Lindorff-Larsen, John Christodoulou, Michele Vendruscolo, and Christopher M Dobson. Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *Journal of the American Chemical Society*, 127(2):476–7, jan 2005.

[180] Jane R. Allison, Peter Varnai, Christopher M. Dobson, and Michele Vendruscolo. Determination of the Free Energy Landscape of α-Synuclein Using Spin Label Nuclear Magnetic Resonance Measurements. *Journal of the American Chemical Society*, 131(51):18314–18326, dec 2009.

[181] Carlo Camilloni, Andrea Cavalli, and Michele Vendruscolo. Replica-Averaged Metadynamics. *Journal of Chemical Theory and Computation*, 9(12):5610–5617, dec 2013.

[182] Carlo Camilloni and Michele Vendruscolo. Statistical mechanics of the denatured state of a protein using replica-averaged metadynamics. *Journal of the American Chemical Society*, 136(25):8982–91, 2014.

[183] Daniele Granata, Fahimeh Baftizadeh, Johnny Habchi, Celine Galvagnion, Alfonso De Simone, Carlo Camilloni, Alessandro Laio, and Michele Vendruscolo. The inverted free energy landscape of an intrinsically disordered peptide by simulations and experiments. *Scientific Reports*, 5(1):15449, dec 2015.

[184] Carlo Camilloni, Paul Robustelli, Alfonso De Simone, Andrea Cavalli, and Michele Vendruscolo. Characterization of the Conformational Equilibrium between the Two Major Substates of RNase A Using NMR Chemical Shifts. *Journal of the American Chemical Society*, 134(9):3968–3971, mar 2012.

[185] Alfonso De Simone, Rinaldo W Montalvao, and Michele Vendruscolo. Determination of Conformational Equilibria in Proteins Using Residual Dipolar Couplings. *Journal of chemical theory and computation*, 7(12):4189–4195, dec 2011.

[186] Alfonso De Simone, Rinaldo W. Montalvao, Christopher M. Dobson, and Michele Vendruscolo. Characterization of the Interdomain Motions in Hen Lysozyme Using Residual Dipolar Couplings as Replica-Averaged Structural Restraints in Molecular Dynamics Simulations. *Biochemistry*, 52(37):6480–6486, sep 2013.

[187] Rinaldo W. Montalvao, Alfonso De Simone, and Michele Vendruscolo. Determination of structural fluctuations of proteins from structure-based calculations of residual dipolar couplings. *Journal of Biomolecular NMR*, 53(4):281–292, aug 2012.

[188] Daniela Bonetti, Carlo Camilloni, Lorenzo Visconti, Sonia Longhi, Maurizio Brunori, Michele Vendruscolo, and Stefano Gianni. Identification and Structural Characterization of an Intermediate in the Folding of the Measles Virus X Domain. *Journal of Biological Chemistry*, 291(20):10886–10892, may 2016.

[189] Alfonso De Simone, Anne Dhulesia, Gemma Soldi, Michele Vendruscolo, Shang-Te Danny Hsu, Fabrizio Chiti, and Christopher M Dobson. Experimental free energy surfaces reveal the mechanisms of maintenance of protein solubility. *Proceedings of the National Academy of Sciences of the United States of America*, 108(52):21057–62, dec 2011.

[190] Philipp Neudecker, Paul Robustelli, Andrea Cavalli, Patrick Walsh, Patrik Lundström, Arash Zarrine-Afsar, Simon Sharpe, Michele Vendruscolo, and Lewis E. Kay. Structure of an Intermediate State in Protein Folding and Aggregation. *Science (New York, N.Y.)*, 336(6079):362–366, 2012.

[191] Hoi Tik Alvin Leung, Predrag Kukic, Carlo Camilloni, Francesco Bemporad, Alfonso De Simone, Francesco A. Aprile, Janet R. Kumita, and Michele Vendruscolo. NMR characterization of the conformational fluctuations of the human lymphocyte function-associated antigen-1 I-domain. *Protein Science*, 23(11):1596–1606, nov 2014.

[192] Predrag Kukic, Carlo Camilloni, Andrea Cavalli, and Michele Vendruscolo. Determination of the individual roles of the linker residues in the interdomain motions of calmodulin using NMR chemical shifts. *Journal of molecular biology*, 426(8):1826–38, apr 2014.

[193] Predrag Kukic, Hoi Tik Alvin Leung, Francesco Bemporad, Francesco A. Aprile, Janet R. Kumita, Alfonso De Simone, Carlo Camilloni, and Michele Vendruscolo. Structure and Dynamics of the Integrin LFA-1 I-Domain in the Inactive State Underlie its Inside-Out/Outside-In Signaling and Allosteric Mechanisms. *Structure*, 23(4):745–753, apr 2015.

[194] Predrag Kukic, Patrik Lundström, Carlo Camilloni, Johan Evenäs, Mikael Akke, and Michele Vendruscolo. Structural Insights into the Calcium-Mediated Allosteric Transition in the C-Terminal Domain of Calmodulin from Nuclear Magnetic Resonance Measurements. *Biochemistry*, 55(1):19–28, jan 2016.

[195] Alfonso De Simone, Francesco a Aprile, Anne Dhulesia, Christopher M Dobson, and Michele Vendruscolo. Structure of a low-population intermediate state in the release of an enzyme product. *eLife*, 4:1–13, 2015.

[196] James M. Krieger, Giuliana Fusco, Marc Lewitzky, Philip C. Simister, Jan Marchant, Carlo Camilloni, Stephan M. Feller, and Alfonso De Simone. Conformational Recognition of an Intrinsically Disordered Protein. *Biophysical Journal*, 106(8):1771–1779, apr 2014.

[197] Domenico Sanfelice, Alfonso De Simone, Andrea Cavalli, Serena Faggiano, Michele Vendruscolo, and Annalisa Pastore. Characterization of the Conformational Fluctuations in the Josephin Domain of Ataxin-3. *Biophysical Journal*, 107(12):2932–2940, dec 2014.

[198] Arvind Kannan, Carlo Camilloni, Aleksandr B Sahakyan, Andrea Cavalli, and Michele Vendruscolo. A conformational ensemble derived using NMR methyl chemical shifts reveals a mechanical clamping transition that gates the binding of the HU protein to DNA. *Journal of the American Chemical Society*, 136(6):2204–7, feb 2014.

[199] Aditi N. Borkar, Michael F. Bardaro, Carlo Camilloni, Francesco A. Aprile, Gabriele Varani, and Michele Vendruscolo. Structure of a low-population binding intermediate in protein-RNA recognition. *Proceedings of the National Academy of Sciences*, 113(26):7171–7176, jun 2016.

[200] Sunhwan Jo and Wonpil Im. Transmembrane helix orientation and dynamics: Insights from ensemble dynamics with solid-state NMR observables. *Biophysical Journal*, 100(12):2913–2921, 2011.

[201] Alfonso De Simone, Kaustubh R Mote, and Gianluigi Veglia. Structural dynamics and conformational equilibria of SERCA regulatory proteins in membranes by solid-state NMR restrained simulations. *Biophysical journal*, 106(12):2566–76, jun 2014.

[202] Martin Gustavsson, Nathaniel J. Traaseth, and Gianluigi Veglia. Probing ground and excited states of phospholamban in model and native lipid membranes by magic angle spinning NMR spectroscopy. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1818(2):146–153, feb 2012.

[203] Alfonso De Simone, Martin Gustavsson, Rinaldo W. Montalvao, Lei Shi, Gianluigi Veglia, and Michele Vendruscolo. Structures of the excited states of phospholamban and shifts in their populations upon phosphorylation. *Biochemistry*, 52:6684–6694, 2013.

[204] Xi Cheng, Sunhwan Jo, Yifei Qi, FrancescaM. Marassi, and Wonpil Im. Solid-State NMR-Restrained Ensemble Dynamics of a Membrane Protein in Explicit Membranes. *Biophysical Journal*, 108(8):1954–1962, 2015.

[205] Jeffrey C. Hoch, Christopher M. Dobson, and Martin Karplus. Fluctuations and averaging of proton chemical shifts in the bovine pancreatic trypsin inhibitor. *Biochemistry*, 21(6):1118–1125, mar 1982.

[206] Adrian M. Lee and Ryan P. A. Bettens. First Principles NMR Calculations by Fragmentation. *The Journal of Physical Chemistry A*, 111(23):5111–5115, jun 2007.

[207] J. A. Vila, J. M. Aramini, P. Rossi, A. Kuzin, M. Su, J. Seetharaman, R. Xiao, L. Tong, G. T. Montelione, and H. A. Scheraga. Quantum chemical 13C chemical shift calculations for protein NMR structure determination, refinement, and validation. *Proceedings of the National Academy of Sciences*, 105(38):14389–14394, sep 2008.

[208] J. A. Vila, Y. A. Arnautova, O. A. Martin, and H. A. Scheraga. Quantum-mechanics-derived 13C chemical shift server (CheShift) for protein structure validation. *Proceedings of the National Academy of Sciences*, 106(40):16972–16977, oct 2009.

[209] Frans A A Mulder and Michael Filatov. NMR chemical shift data and ab initio shielding calculations: emerging tools for protein structure determination. *Chemical Society reviews*, 39(2):578–90, feb 2010.

[210] Tong Zhu, John Z. H. Zhang, and Xiao He. Quantum Calculation of Protein NMR Chemical Shifts Based on the Automated Fragmentation Method. In *Advance in Structural Bioinformatics*, pages 49–70. Springer, 2015.

[211] Jodi Kraus, Rupal Gupta, Jenna Yehl, Manman Lu, David A. Case, Angela M. Gronenborn, Mikael Akke, and Tatyana Polenova. Chemical Shifts of the Carbohydrate Binding Domain of Galectin-3 from Magic Angle Spinning NMR and Hybrid Quantum Mechanics/Molecular Mechanics Calculations. *The Journal of Physical Chemistry B*, 122(11):2931–2939, mar 2018.

[212] Andrea Frank, Heiko M. Möller, and Thomas E. Exner. Toward the Quantum Chemical Calculation of NMR Chemical Shifts of Proteins. 2. Level of Theory, Basis Set, and Solvents Model Dependence. *Journal of Chemical Theory and Computation*, 8(4):1480–1492, apr 2012.

[213] Eldon L. Ulrich, Hideo Akutsu, Jurgen F. Doreleijers, Yoko Harano, Yannis E. Ioannidis, Jundong Lin, Miron Livny, Steve Mading, Dimitri Maziuk, Zachary Miller, Eiichi Nakatani, Christopher F. Schulte, David E. Tolmie, R. Kent Wenger, Hongyang Yao, and John L. Markley. BioMagResBank. *Nucleic Acids Research*, 36(Database issue):402–408, 2008.

[214] John L. Markley, Donella H. Meadows, and Oleg Jardetzky. Nuclear magnetic resonance studies of helix-coil transitions in polyamino acids. *Journal of Molecular Biology*, 27(1):25–40, jul 1967.

[215] D S Wishart, B D Sykes, and F M Richards. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *Journal of molecular biology*, 222(2):311–33, nov 1991.

[216] N.J Clayden and R.J.P Williams. Peptide group shifts. *Journal of Magnetic Resonance (1969)*, 49(3):383–396, oct 1982.

[217] Stephan Schwarzinger, Gerard J. A. Kroon, Ted R. Foss, John Chung, Peter E. Wright, and H. Jane Dyson. Sequence-Dependent Correction of Random Coil NMR Chemical Shifts. *Journal of the American Chemical Society*, 123(13):2970–2978, apr 2001.

[218] D C Dalgarno, B A Levine, and R J Williams. Structural information from NMR secondary chemical shifts of peptide alpha C-H protons in proteins. *Bioscience reports*, 3(5):443–52, may 1983.

[219] D S Wishart and D A Case. Use of chemical shifts in macromolecular structure determination. *Methods in enzymology*, 338:3–34, 2001.

[220] Steven P Mielke and V V Krishnan. Characterization of protein secondary structure from NMR chemical shifts. *Progress in nuclear magnetic resonance spectroscopy*, 54(3-4):141–165, apr 2009.

[221] D. S. Wishart, B. D. Sykes, and F. M. Richards. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 31(6):1647–1651, feb 1992.

[222] DavidS. Wishart and BrianD. Sykes. The 13C Chemical-Shift Index: A simple method for the identification of protein secondary structure using 13C chemical-shift data. *Journal of Biomolecular NMR*, 4(2), mar 1994.

[223] Gabriel Cornilescu, Frank Delaglio, and Ad Bax. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *Journal of Biomolecular NMR*, 13(3):289–302, 1999.

[224] Yang Shen, Frank Delaglio, Gabriel Cornilescu, and Ad Bax. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of biomolecular NMR*, 44(4):213–23, aug 2009.

[225] Stephen Neal, Alex M. Nip, Haiyan Zhang, and David S. Wishart. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *Journal of Biomolecular NMR*, 26(3):215–240, 2003.

[226] Jens Meiler. PROSHIFT: Protein chemical shift prediction using artificial neural networks. *Journal of Biomolecular NMR*, 26(1):25–37, 2003.

[227] Yang Shen and Ad Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *Journal of Biomolecular NMR*, 38(4):289–302, 2007.

[228] Kai J Kohlhoff, Paul Robustelli, Andrea Cavalli, Xavier Salvatella, and Michele Vendruscolo. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *Journal of the American Chemical Society*, 131:13894–13895, 2009.

[229] Yang Shen and Ad Bax. SPARTA+: A modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *Journal of Biomolecular NMR*, 48(1):13–22, 2010.

[230] Beomsoo Han, Yifeng Liu, Simon W. Ginzinger, and David S. Wishart. SHIFTX2: Significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR*, 50(1):43–57, 2011.

[231] David D. Laws, Angel C. de Dios, and Eric Oldfield. NMR chemical shifts and structure refinement in proteins. *Journal of Biomolecular NMR*, 3(5):607–612, sep 1993.

[232] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker, and A. Bax. Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences*, 105(12):4685–4690, 2008.

[233] Yang Shen, Robert Vernon, David Baker, and A. Bax. De novo protein structure generation from incomplete chemical shift assignments. *Journal of Biomolecular NMR*, 43(2):63–78, 2009.

[234] Haiyan Zhang, Stephen Neal, and David S. Wishart. RefDB: A database of uniformly referenced protein chemical shifts. *Journal of Biomolecular NMR*, 25(3):173–195, 2003.

[235] W Gronwald, L Willard, T Jellard, R F Boyko, K Rajarathnam, D S Wishart, F D Sönnichsen, and B D Sykes. CAMRA: chemical shift based computer aided protein NMR assignments. *Journal of biomolecular NMR*, 12(3):395–405, oct 1998.

[236] J L Pons and M A Delsuc. RESCUE: an artificial neural network tool for the NMR spectral assignment of proteins. *Journal of biomolecular NMR*, 15(1):15–26, sep 1999.

[237] Bowei Wang, Yunjun Wang, and David S Wishart. A probabilistic approach for validating protein NMR chemical shift assignments. *Journal of biomolecular NMR*, 47(2):85–99, jun 2010.

[238] Malene Ringkjøbing Jensen, Loïc Salmon, Gabrielle Nodet, and Martin Blackledge. Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *Journal of the American Chemical Society*, 132(4):1270–1272, 2010.

[239] Elliott J. Stollar, Hong Lin, Alan R. Davidson, and Julie D. Forman-Kay. Differential Dynamic Engagement within 24 SH3 Domain: Peptide Complexes Revealed by Co-Linear Chemical Shift Perturbation Analysis. *PLoS ONE*, 7(12), 2012.

[240] Isabella C Felli, Leonardo Gonnelli, and Roberta Pierattelli. In-cell 13C NMR spectroscopy for the study of intrinsically disordered proteins. *Nature Protocols*, 9(9):2005–2016, 2014.

[241] Christopher A. Waudby, Carlo Camilloni, Anthony W P Fitzpatrick, Lisa D. Cabrita, Christopher M. Dobson, Michele Vendruscolo, and John Christodoulou. In-Cell NMR Characterization of the Secondary Structure Populations of a Disordered Conformation of α-Synuclein within E. coli Cells. *PLoS ONE*, 8(8), 2013.

[242] D S Wishart, C G Bigam, A Holm, R S Hodges, and B D Sykes. 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *Journal of biomolecular NMR*, 5(1):67–81, jan 1995.

[243] Yunjun Wang and Oleg Jardetzky. Investigation of the neighboring residue effects on protein chemical shifts. *Journal of the American Chemical Society*, 124(47):14075–14084, 2002.

[244] Kamil Tamiola, Burçin Acar, and Frans A. A. Mulder. Sequence-Specific Random Coil Chemical Shifts of Intrinsically Disordered Proteins. *Journal of the American Chemical Society*, 132:18000–18003, 2010.

[245] Magnus Kjaergaard and Flemming M. Poulsen. Sequence correction of random coil chemical shifts: Correlation between neighbor correction factors and changes in the Ramachandran distribution. *Journal of Biomolecular NMR*, 50(2):157–165, 2011.

[246] D S Wishart, M S Watson, R F Boyko, and B D Sykes. Automated 1H and 13C chemical shift prediction using the BioMagResBank. *Journal of biomolecular NMR*, 10(4):329–36, 1997.

[247] Barbara C.M. Potts and Walter J. Chazin. Chemical shift homology in proteins. *Journal of Biomolecular NMR*, 11(1):45–57, 1998.

[248] Máximo Sanz-Hernández and Alfonso De Simone. The PROSECCO server for chemical shift predictions in ordered and disordered proteins. *Journal of Biomolecular NMR*, 69(3):147–156, 2017.

[249] Robert C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.

[250] Carlo Camilloni, Alfonso De Simone, Wim F Vranken, and Michele Vendruscolo. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry*, 51(11):2224–31, mar 2012.

[251] Damiano Piovesan, Francesco Tabaro, Ivan Mičetić, Marco Necci, Federica Quaglia, Christopher J. Oldfield, Maria Cristina Aspromonte, Norman E. Davey, Radoslav Davidović, Zsuzsanna Dosztányi, Arne Elofsson, Alessandra Gasparini, András Hatos, Andrey V. Kajava, Lajos Kalmar, Emanuela Leonardi, Tamas Lazar, Sandra Macedo-Ribeiro, Mauricio Macossay-Castillo, Attila Meszaros, Giovanni Minervini, Nikoletta Murvai, Jordi Pujols, Daniel B. Roche, Edoardo Salladini, Eva Schad, Antoine Schramm, Beata Szabo, Agnes Tantos, Fiorella Tonello, Konstantinos D. Tsirigos, Nevena Veljković, Salvador Ventura, Wim Vranken, Per Warholm, Vladimir N. Uversky, A. Keith Dunker, Sonia Longhi, Peter Tompa, and Silvio C.E. Tosatto. DisProt 7.0: A major update of the database of disordered proteins. *Nucleic Acids Research*, 45(D1):D219–D227, 2017.

[252] Alfonso De Simone, Andrea Cavalli, Shang-Te Danny Hsu, Wim Vranken, and Michele Vendruscolo. Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. *Journal of the American Chemical Society*, 131(45):16332–3, nov 2009.

[253] Bernard. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, 1986.

[254] Alexander S. Maltsev, Jinfa Ying, and Ad Bax. Impact of N-terminal acetylation of α-synuclein on its random coil and lipid binding properties. *Biochemistry*, 51(25):5004–5013, 2012.

[255] Carlo Camilloni, Daniel Schaal, Kristian Schweimer, Stephan Schwarzinger, and Alfonso De Simone. Energy landscape of the prion protein helix 1 probed by metadynamics and NMR. *Biophysical journal*, 102(1):158–67, jan 2012.

[256] David T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202, 1999.

[257] W Kabsch and C Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[258] Kornel Papik, Bela Molnar, Rainer Schaefer, Zalan Dombovari, Zsolt Tulassay, and Janos Feher. Application of neural networks in medicine a review. *Med Sci Monit*, 4(3):538–546, 1998.

[259] Jure Zupan and Johann Gasteiger. Neural Networks in Chemistry. *Angewandte Chemie - International Edition*, 32:503–527, 1999.

[260] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[261] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pages 1–15, 2015.

[262] Yang Shen, Philip N. Bryan, Yanan He, John Orban, David Baker, and Ad Bax. De novo structure generation using chemical shifts for proteins with high-sequence identity but different folds. *Protein Science*, 19(2):349–356, 2010.

[263] Aleksandr B. Sahakyan, Wim F. Vranken, Andrea Cavalli, and Michele Vendruscolo. Structure-based prediction of methyl chemical shifts in proteins. *Journal of Biomolecular NMR*, 50(4):331–346, 2011.

[264] Jason E. Ollerenshaw, Vitali Tugarinov, and Lewis E. Kay. Methyl TROSY: explanation and experimental verification. *Magnetic Resonance in Chemistry*, 41(10):843–852, 2003.

[265] Gaël De Paëpe. Dipolar Recoupling in Magic Angle Spinning Solid-State Nuclear Magnetic Resonance. *Annual Review of Physical Chemistry*, 63(1):661–684, may 2012.

[266] Nabanita Das, Dylan T Murray, and Timothy A Cross. Lipid bilayer preparations of membrane proteins for oriented and magic-angle spinning solid-state NMR samples. *Nature protocols*, 8(11):2256–70, nov 2013.

[267] Charles R. Sanders, Brian J. Hare, Kathleen P. Howard, and James H. Prestegard. Magnetically-oriented phospholipid micelles as a tool for the study of membrane-associated molecules. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 26:421–444, jan 1994.

[268] Malcolm H. Levitt. *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. Wiley, 2nd edition, 2008.

[269] R. K. Hester, J. L. Ackerman, B. L. Neff, and J. S. Waugh. Separated Local Field Spectra in NMR: Determination of Structure of Solids. *Physical Review Letters*, 36(18):1081–1083, may 1976.

[270] A Ramamoorthy, Yufeng Wei, and Dong-Kuk Lee. PISEMA Solid-State NMR Spectroscopy. *Annual Reports on NMR Spectroscopy*, 52:1–52, 2004.

[271] T Gopinath and Gianluigi Veglia. Sensitivity enhancement in static solid-state NMR experiments via single- and multiple-quantum dipolar coherences. *Journal of the American Chemical Society*, 131(16):5754–6, apr 2009.

[272] Sara K. Hansen, Kresten Bertelsen, Berit Paaske, Niels Chr. Nielsen, and Thomas Vosegaard. Solid-state NMR methods for oriented membrane proteins. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 88-89:48–85, aug 2015.

[273] Francesca M Marassi and Stanley J Opella. A Solid-State NMR Index of Helical Membrane Protein Structure and Topology. *Journal of Magnetic Resonance*, 144(1):150–155, may 2000.

[274] Jarrod J. Buffy, Nathaniel J. Traaseth, Alessandro Mascioni, Peter L. Gor'kov, Eduard Y. Chekmenev, William W. Brey, and Gianluigi Veglia. Two-dimensional solid-state NMR reveals two topologies of sarcolipin in oriented lipid bilayers. *Biochemistry*, 45(36):10939–10946, 2006.

[275] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4 : Algorithms for Highly Efficient , Load-Balanced , and Scalable Molecular Simulation. *Journal of chemical theory and computation*, 4:435–447, 2008.

[276] Raffaello Verardi, Lei Shi, Nathaniel J Traaseth, Naomi Walsh, and Gianluigi Veglia. Structural topology of phospholamban pentamer in lipid bilayers by a hybrid solution and solid-state NMR method. *Proceedings of the National Academy of Sciences of the United States of America*, 108(22):9101–6, may 2011.

[277] Szilárd Páll and Berk Hess. A flexible algorithm for calculating pair interactions on SIMD architectures. *Computer Physics Communications*, 184(12):2641–2650, dec 2013.

[278] Máximo Sanz-Hernández, Vitaly V Vostrikov, Gianluigi Veglia, and Alfonso De Simone. Accurate Determination of Conformational Transitions in Oligomeric Membrane Proteins. *Scientific Reports*, 6:23063, mar 2016.

[279] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, 78(8):1950–1958, 2010.

[280] Donald M. Bers. Cardiac excitationcontraction coupling. *Nature*, 415(6868):198–205, jan 2002.

[281] David H MacLennan and Evangelia G Kranias. Phospholamban: a crucial regulator of cardiac contractility. *Nature reviews. Molecular cell biology*, 4(7):566–577, 2003.

[282] J Fujii, K Maruyama, M Tada, and D H MacLennan. Expression and site-specific mutagenesis of phospholamban. Studies of residues involved in phosphorylation and pentamer formation. *The Journal of biological chemistry*, 264(22):12950–5, aug 1989.

[283] D D Thomas, L G Reddy, C B Karim, M Li, R Cornea, J M Autry, L R Jones, and J Stamm. Direct spectroscopic detection of molecular dynamics and interactions of the calcium pump and phospholamban. *Annals of the New York Academy of Sciences*, 853:186–94, sep 1998.

[284] L G Reddy, L R Jones, and D D Thomas. Depolymerization of phospholamban in the presence of calcium pump: a fluorescence energy transfer study. *Biochemistry*, 38(13):3954–62, mar 1999.

[285] Seth L Robia, Kenneth S Campbell, Eileen M Kelly, Zhanjia Hou, Deborah L Winters, and David D Thomas. Förster transfer recovery reveals that phospholamban exchanges slowly from pentamers but rapidly from the SERCA regulatory complex. *Circulation research*, 101(11):1123–9, nov 2007.

[286] Eileen M Kelly, Zhanjia Hou, Julie Bossuyt, Donald M Bers, and Seth L Robia. Phospholamban oligomerization, quaternary structure, and sarco(endo)plasmic reticulum calcium ATPase binding measured by fluorescence resonance energy transfer in living cells. *The Journal of biological chemistry*, 283(18):12202–11, may 2008.

[287] R. J. Kovacs, M. T. Nelson, H. K B Simmerman, and L. R. Jones. Phospholamban forms Ca2+ selective channels in lipid bilayers. *Journal of Biological Chemistry*, 263(34):18364–18368, 1988.

[288] Roozbeh Aschar-Sobbi, Teresa L. Emmett, Gary J. Kargacin, and Margaret E. Kargacin. Phospholamban phosphorylation increases the passive calcium leak from cardiac sarcoplasmic reticulum. *Pflugers Archiv European Journal of Physiology*, 464(3):295–305, 2012.

[289] Serena Smeazzetto, Indra Schröder, Gerhard Thiel, and Maria Rosa Moncelli. Phospholamban generates cation selective ion channels. *Physical Chemistry Chemical Physics*, 13(28):12935, 2011.

[290] Serena Smeazzetto, Francesco Tadini-Buoninsegni, Gerhard Thiel, and Maria Rosa Moncelli. Selectivity of the phospholamban ion channel investigated by single channel measurements. *Journal of Electroanalytical Chemistry*, 812:244–248, mar 2018.

[291] Lucia Becucci, Alessandro Cembran, Christine B Karim, David D Thomas, Rolando Guidelli, Jiali Gao, and Gianluigi Veglia. On the function of pentameric phospholamban: ion channel or storage form? *Biophysical journal*, 96(10):L60–2, 2009.

[292] Vitaly V. Vostrikov, Kaustubh R. Mote, Raffaello Verardi, and Gianluigi Veglia. Structural dynamics and topology of phosphorylated phospholamban homopentamer reveal its role in the regulation of calcium transport. *Structure*, 21(12):2119–2130, 2013.

[293] Christopher Maffeo and Aleksei Aksimentiev. Structure, Dynamics, and Ion Conductance of the Phospholamban Pentamer. *Biophysical Journal*, 96(12):4853–4865, jun 2009.

[294] Callum J. Dickson, Lula Rosso, Robin M. Betz, Ross C. Walker, and Ian R. Gould. GAFFlipid: a General Amber Force Field for the accurate molecular dynamics simulation of phospholipid. *Soft Matter*, 8(37):9617, 2012.

[295] Jeffery B. Klauda, Richard M. Venable, J. Alfredo Freites, Joseph W. O'Connor, Douglas J. Tobias, Carlos Mondragon-Ramirez, Igor Vorobyov, Alexander D. MacKerell, and Richard W. Pastor. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *The Journal of Physical Chemistry B*, 114(23):7830–7843, jun 2010.

[296] H J C Berendsen, J P M Postma, W F van Gunsteren, a DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.

[297] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *Journal of Chemical Physics*, 126(1):014101–7, 2007.

[298] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[299] 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, 1993.

[300] Robert B. Best, Jane Clarke, and Martin Karplus. What contributions to protein side-chain dynamics are probed by NMR experiments? A molecular dynamics simulation analysis. *Journal of Molecular Biology*, 349(1):185–203, 2005.

[301] David Sehnal, Radka Svobodová Vaeková, Karel Berka, Lukáš Pravda, Veronika Navrátilová, Pavel Banáš, Crina Maria Ionescu, Michal Otyepka, and Jaroslav Koča. MOLE 2.0: Advanced approach for analysis of biomacromolecular channels. *Journal of Cheminformatics*, 5(8):1–13, 2013.

[302] B. L. Akin, T. D. Hurley, Z. Chen, and L. R. Jones. The Structural Basis for Phospholamban Inhibition of the Calcium Pump in Sarcoplasmic Reticulum. *Journal of Biological Chemistry*, 288(42):30181–30191, 2013.

[303] Narayanan Eswar, Ben Webb, Marc A Marti-Renom, M S Madhusudhan, David Eramian, Min-Yi Shen, Ursula Pieper, and Andrej Sali. Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics*, Chapter 5:Unit–5.6, oct 2006.

[304] Tsjerk A. Wassenaar, Kristyna Pluhackova, Rainer A. Böckmann, 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, feb 2014.

[305] Barbara Richter, Joerg Gsponer, Péter Várnai, Xavier Salvatella, and Michele Vendruscolo. The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *Journal of biomolecular NMR*, 37(2):117–35, feb 2007.

[306] K. N. Ha, L. R. Masterson, Z. Hou, R. Verardi, N. Walsh, G. Veglia, and S. L. Robia. Lethal Arg9Cys phospholamban mutation hinders Ca2+-ATPase regulation and phosphorylation by protein kinase A. *Proceedings of the National Academy of Sciences*, 108(7):2735–2740, feb 2011.

[307] K. Haghighi, F. Kolokathis, A. O. Gramolini, J. R. Waggoner, L. Pater, R. A. Lynch, G.-C. Fan, D. Tsiapras, R. R. Parekh, G. W. Dorn, D. H. MacLennan, D. Th. Kremastinos, and E. G. Kranias. A mutation in the human phospholamban gene, deleting arginine 14, results in lethal, hereditary cardiomyopathy. *Proceedings of the National Academy of Sciences*, 103(5):1388–1393, jan 2006.

[308] Andrew Rambaut, Oliver G. Pybus, Martha I. Nelson, Cecile Viboud, Jeffery K. Taubenberger, and Edward C. Holmes. The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453(7195):615–619, may 2008.

[309] Tasleem Samji. Influenza A: understanding the viral life cycle. *The Yale journal of biology and medicine*, 82(4):153–9, dec 2009.

[310] Lawrence H Pinto and Robert A Lamb. The M2 proton channels of influenza A and B viruses. *The Journal of biological chemistry*, 281(14):8997–9000, apr 2006.

[311] C Wang, K Takeuchi, L H Pinto, and R A Lamb. Ion channel activity of influenza A virus M2 protein: characterization of the amantadine block. *Journal of virology*, 67(9):5585–94, sep 1993.

[312] A J Hay, A J Wolstenholme, J J Skehel, and M H Smith. The molecular basis of the specific anti-influenza action of amantadine. *The EMBO journal*, 4(11):3021–4, nov 1985.

[313] Amanda L Stouffer, Rudresh Acharya, David Salom, Anna S Levine, Luigi Di Costanzo, Cinque S Soto, Valentina Tereshko, Vikas Nanda, Steven Stayrook, and William F DeGrado. Structural basis for the function and inhibition of an influenza virus proton channel. *Nature*, 451(7178):596–9, jan 2008.

[314] Rudresh Acharya, Vincenzo Carnevale, Giacomo Fiorin, Benjamin G Levine, Alexei L Polishchuk, Victoria Balannik, Ilan Samish, Robert A Lamb, Lawrence H Pinto, William F DeGrado, and Michael L Klein. Structure and mechanism of proton transport through the transmembrane tetrameric M2 protein bundle of the influenza A virus. *Proceedings of the National Academy of Sciences of the United States of America*, 107(34):15075–80, 2010.

[315] Jason R Schnell and James J Chou. Structure and mechanism of the M2 proton channel of influenza A virus. *Nature*, 451(7178):591–5, 2008.

[316] Junfeng Wang, Sanguk Kim, Frank Kovacs, and Timothy A Cross. Structure of the transmembrane region of the M2 protein H + channel. *Protein Science*, 10(1992):2241–2250, 2001.

[317] Sarah D. Cady, Klaus Schmidt-Rohr, Jun Wang, Cinque S. Soto, William F. DeGrado, and Mei Hong. Structure of the amantadine binding site of influenza M2 proton channels in lipid bilayers. *Nature*, 463(7281):689–692, feb 2010.

[318] Mukesh Sharma, Myunggi Yi, Hao Dong, Huajun Qin, Emily Peterson, David D Busath, Huan-xiang Zhou, and Timothy A Cross. Insight into the Mechanism of the Influenza A Proton Channel from a Structure in a Lipid Bilayer. *Science*, 5(October):509–513, 2010.

[319] F. Hu, W. Luo, and M. Hong. Mechanisms of Proton Conduction and Gating in Influenza M2 Proton Channels from Solid-State NMR. *Science*, 330(6003):505–508, oct 2010.

[320] Jeremy S. Rossman, Xianghong Jing, George P. Leser, and Robert A. Lamb. Influenza Virus M2 Protein Mediates ESCRT-Independent Membrane Scission. *Cell*, 142(6):902–913, sep 2010.

[321] Rafal M Pielak and James J Chou. Influenza M2 proton channels. *Biochimica et biophysica acta*, 1808(2):522–9, feb 2011.

[322] Chunlong Ma, Alexei L Polishchuk, Yuki Ohigashi, Amanda L Stouffer, Arne Schön, Emma Magavern, Xianghong Jing, James D Lear, Ernesto Freire, Robert A Lamb, William F DeGrado, and Lawrence H Pinto. Identification of the functional core of the influenza A virus A/M2 proton-selective ion channel. *Proceedings of the National Academy of Sciences of the United States of America*, 106(30):12283–8, jul 2009.

[323] A Okada, T Miura, and H Takeuchi. Protonation of histidine and histidine-tryptophan interaction in the activation of the M2 ion channel from influenza a virus. *Biochemistry*, 40(20):6053–60, may 2001.

[324] Ekta Khurana, Matteo Dal Peraro, Russell DeVane, Satyavani Vemparala, William F DeGrado, and Michael L Klein. Molecular dynamics calculations suggest a conduction mechanism for the M2 proton channel from influenza A virus. *Proceedings of the National Academy of Sciences of the United States of America*, 106(4):1069–74, 2009.

[325] Mei Hong, Keith J Fritzsching, and Jonathan K Williams. Hydrogen-bonding partner of the proton-conducting histidine in the influenza M2 proton channel revealed from 1H chemical shifts. *Journal of the American Chemical Society*, 134(36):14753–5, sep 2012.

[326] Alexander M. Smondyrev and Gregory A. Voth. Molecular Dynamics Simulation of Proton Transport through the Influenza A Virus M2 Channel. *Biophysical Journal*, 83(4):1987–1996, oct 2002.

[327] Myunggi Yi, Timothy A. Cross, and Huan-Xiang Zhou. A Secondary Gate As a Mechanism for Inhibition of the M2 Proton Channel by Amantadine. *The Journal of Physical Chemistry B*, 112(27):7977–7979, jul 2008.

[328] Myunggi Yi, Timothy A Cross, and Huan-xiang Zhou. Conformational heterogeneity of the M2 proton channel and a structural model for channel activation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(32):13311–13316, 2009.

[329] Hanning Chen, Yujie Wu, and Gregory a Voth. Proton transport behavior through the influenza A M2 channel: insights from molecular simulation. *Biophysical journal*, 93(10):3470–9, nov 2007.

[330] Vincenzo Carnevale, Giacomo Fiorin, Benjamin G. Levine, William F. DeGrado, and Michael L. Klein. Multiple Proton Confinement in the M2 Channel from the Influenza A Virus. *The Journal of Physical Chemistry C*, 114(48):20856–20863, dec 2010.

[331] Arindam Bankura, Michael L Klein, and Vincenzo Carnevale. Proton affinity of the histidine-tryptophan cluster motif from the influenza A virus from ab initio molecular dynamics. *Chemical physics*, 422:156–164, aug 2013.

[332] Ruibin Liang, Hui Li, Jessica M J Swanson, and Gregory a Voth. Multiscale simulation reveals a multifaceted mechanism of proton permeation through the influenza A M2 proton channel. *Proceedings of the National Academy of Sciences*, 111(26):9396–9401, 2014.

[333] Oliver S. Smart, Joseph G. Neduvelil, Xiaonan Wang, B.A. Wallace, and Mark S.P. Sansom. HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *Journal of Molecular Graphics*, 14(6):354–360, dec 1996.

[334] V Lounnas, B M Pettitt, and G N Phillips. A global model of the protein-solvent interface. *Biophysical journal*, 66(3 Pt 1):601–14, mar 1994.

[335] Alfonso De Simone, Guy G Dodson, Chandra S Verma, Adriana Zagari, and Franca Fraternali. Prion and water: tight and dynamical hydration sites have a key role in structural stability. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7535–40, may 2005.

[336] M S Sansom, I H Shrivastava, K M Ranatunga, and G R Smith. Simulations of ion channels–watching ions and water move. *Trends in biochemical sciences*, 25(8):368–74, aug 2000.

[337] Jayendran C Rasaiah, Shekhar Garde, and Gerhard Hummer. Water in nonpolar confinement: from nanotubes to proteins and beyond. *Annual review of physical chemistry*, 59:713–40, 2008.

[338] Daniel M Rosenbaum, Søren G F Rasmussen, and Brian K Kobilka. The structure and function of G-protein-coupled receptors. *Nature*, 459(7245):356–63, 2009.

[339] Jason W Griffith, Caroline L Sokol, and Andrew D Luster. Chemokines and chemokine receptors: positioning cells for host defense and immunity. *Annual review of immunology*, 32:659–702, 2014.

[340] D. J.J. Waugh and C. Wilson. The Interleukin-8 Pathway in Cancer. *Clinical Cancer Research*, 14(21):6735–6741, nov 2008.

[341] C Knall, G S Worthen, and G L Johnson. Interleukin 8-stimulated phosphatidylinositol-3-kinase activity regulates the migration of human neutrophils independent of extracellular signal-regulated kinase and p38 mitogen-activated protein kinases. *Proceedings of the National Academy of Sciences of the United States of America*, 94(7):3052–7, apr 1997.

[342] Bibhuti B Das, Henry J Nothnagel, George J Lu, Woo Sung Son, Ye Tian, Francesca M Marassi, and Stanley J Opella. Structure determination of a membrane protein in proteoliposomes. *Journal of the American Chemical Society*, 134(4):2047–56, 2012.

[343] Bibhuti B Das, Sang Ho Park, and Stanley J Opella. Membrane protein structure from rotational diffusion. *Biochimica Et Biophysica Acta-Biomembranes*, 1848(1):1–17, 2014.

[344] S. Bowerman and J. Wereszczynski. Detecting Allosteric Networks Using Molecular Dynamics Simulation. *Methods in Enzymology*, 578:429–447, 2016.

[345] Aric A Hagberg, Daniel A Schult, and Pieter J Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11–15, Pasadena, CA USA, 2008.

[346] P H Hünenberger, A E Mark, and W F van Gunsteren. Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *Journal of molecular biology*, 252(4):492–503, sep 1995.

[347] Oliver F. Lange and Helmut Grubmüller. Generalized correlation for biomolecular dynamics. *Proteins: Structure, Function, and Bioinformatics*, 62(4):1053–1061, dec 2005.

[348] Gunnar von Heijne. Proline kinks in transmembrane α-helices. *Journal of Molecular Biology*, 218(3):499–503, apr 1991.

[349] Henry R. Wilman, Jiye Shi, and Charlotte M. Deane. Helix kinks are equally prevalent in soluble and membrane proteins. *Proteins: Structure, Function, and Bioinformatics*, 82(9):1960–1970, sep 2014.

[350] Frank S. Cordes, Joanne N. Bright, and Mark S.P. Sansom. Proline-induced Distortions of Transmembrane Helices. *Journal of Molecular Biology*, 323(5):951–960, nov 2002.

[351] S. Yohannan, S. Faham, D. Yang, J. P. Whitelegge, and J. U. Bowie. The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proceedings of the National Academy of Sciences*, 101(4):959–963, 2004.

[352] M Wheatley, D Wootten, MT Conner, J Simms, R Kendrick, RT Logan, DR Poyner, and J Barwell. Lifting the lid on GPCRs: the role of extracellular loops. *British Journal of Pharmacology*, 165(6):1688–1703, mar 2012.

[353] M.C. Peeters, G.J.P. van Westen, Q. Li, and A.P. IJzerman. Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation. *Trends in Pharmacological Sciences*, 32(1):35–42, jan 2011.

[354] Jingjing Guo and Huan Xiang Zhou. Protein Allostery and Conformational Dynamics. *Chemical Reviews*, 116(11):6503–6515, 2016.

[355] Ali Rana Atilgan, Pelin Akan, and Canan Baysal. Small-world communication of residues and significance for protein dynamics. *Biophysical journal*, 86(1 Pt 1):85–91, jan 2004.

[356] Lukas K. Tamm, editor. *Protein-Lipid Interactions*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG, aug 2005.

[357] Shobha Potluri, Anthony K. Yan, James J. Chou, Bruce R. Donald, and Chris Bailey-Kellogg. Structure Determination of Symmetric Homo-Oligomers by a Complete Search of Symmetry Configuration Space, Using NMR Restraints and van der Waals Packing. *Proteins*, 65(1):203–219, 2006.

[358] Peng Lian, Dong Qing Wei, Jing Fang Wang, and Kuo Chen Chou. An allosteric mechanism inferred from molecular dynamics simulations on phospholamban pentamer in lipid membranes. *PLoS ONE*, 6(4):2–7, 2011.

[359] Kim N. Ha, Martin Gustavsson, and Gianluigi Veglia. Tuning the structural coupling between the transmembrane and cytoplasmic domains of phospholamban to control sarcoplasmic reticulum Ca 2+-ATPase (SERCA) function. *Journal of Muscle Research and Cell Motility*, 33(6):485–492, 2012.

[360] Tanja Wittmann, Martin J. Lohse, and Joachim P. Schmitt. Phospholamban pentamers attenuate PKA-dependent phosphorylation of monomers. *Journal of Molecular and Cellular Cardiology*, 80:90–97, 2015.

[361] Lucia Becucci, Maria Luisa Foresti, Adrian Schwan, and Rolando Guidelli. Can proton pumping by SERCA enhance the regulatory role of phospholamban and sarcolipin? *Biochimica et Biophysica Acta - Biomembranes*, 1828(11):2682–2690, 2013.

[362] Peter Tompa and Monika Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in proteinprotein interactions. *Trends in Biochemical Sciences*, 33(1):2–8, jan 2008.

[363] Lilia M Iakoucheva, Predrag Radivojac, Celeste J Brown, Timothy R O'Connor, Jason G Sikes, Zoran Obradovic, and A Keith Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucleic acids research*, 32(3):1037–49, 2004.

[364] Kailey J. Soller, Raffaello Verardi, Meng Jing, Neha Abrol, Jing Yang, Naomi Walsh, Vitaly V. Vostrikov, Seth L. Robia, Michael T. Bowser, and Gianluigi Veglia. Rheostatic Regulation of the SERCA/Phospholamban Membrane Protein Complex Using Non-Coding RNA and Single-Stranded DNA oligonucleotides. *Scientific Reports*, 5(August):13000, 2015.

[365] Kailey J. Soller, Jing Yang, Gianluigi Veglia, and Michael T. Bowser. Reversal of Phospholamban Inhibition of the Sarco(endo)plasmic Reticulum Ca 2+ -ATPase (SERCA) Using Short, Protein-interacting RNAs and Oligonucleotide Analogs. *Journal of Biological Chemistry*, 291(41):21510–21518, oct 2016.

[366] J. Hu, R. Fu, K. Nishimura, L. Zhang, H.-X. Zhou, D. D. Busath, V. Vijayvergiya, and T. A. Cross. Histidines, heart of the hydrogen ion channel from influenza A virus: Toward an understanding of conductance and proton selectivity. *Proceedings of the National Academy of Sciences*, 103(18):6865–6870, may 2006.

[367] Guoying Dong, Chao Peng, Jing Luo, Chengmin Wang, Le Han, Bin Wu, Guangju Ji, and Hongxuan He. Adamantane-Resistant Influenza A Viruses in the World (19022013): Frequency and Distribution of M2 Gene Mutations. *PLOS ONE*, 10(3):e0119115, mar 2015.

[368] Jessica L Thomaston and William F DeGrado. Crystal structure of the drug-resistant S31N influenza M2 proton channel. *Protein science*, 25(8):1551–4, 2016.

[369] L R Forrest, A Kukol, I T Arkin, D P Tieleman, and M S Sansom. Exploring models of the influenza A M2 channel: MD simulations in a phospholipid bilayer. *Biophysical journal*, 78(1):55–69, jan 2000.

[370] Sarah D. Cady, Tatiana V. Mishanina, and Mei Hong. Structure of Amantadine-Bound M2 Transmembrane Peptide of Influenza A in Lipid Bilayers from Magic-Angle-Spinning Solid-State NMR: The Role of Ser31 in Amantadine Binding. *Journal of Molecular Biology*, 385(4):1127–1141, jan 2009.

[371] Ayanjeet Ghosh, Jun Wang, Yurii S. Moroz, Ivan V. Korendovych, Martin Zanni, William F. DeGrado, Feng Gai, and Robin M. Hochstrasser. 2D IR spectroscopy reveals the role of water in the binding of channel-blocking drugs to the influenza M2 channel. *The Journal of Chemical Physics*, 140(23):235105, jun 2014.

[372] Ayanjeet Ghosh, Jade Qiu, William F DeGrado, and Robin M Hochstrasser. Tidal surge in the M2 proton channel, sensed by 2D IR spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 108(15):6115–20, apr 2011.

[373] W. Liu, E. Chun, A. A. Thompson, P. Chubukov, F. Xu, V. Katritch, G. W. Han, C. B. Roth, L. H. Heitman, A. P. IJzerman, V. Cherezov, and R. C. Stevens. Structural Basis for Allosteric Regulation of GPCRs by Sodium Ions. *Science*, 337(6091):232–236, jul 2012.

[374] Tara Mirzadegan, Gil Benkö, Sławomir Filipek, and Krzysztof Palczewski. Sequence analyses of G-protein-coupled receptors: similarities to rhodopsin. *Biochemistry*, 42(10):2759–67, mar 2003.

[375] Lei Shi, George Liapakis, Rui Xu, Frank Guarnieri, Juan A Ballesteros, and Jonathan A Javitch. Beta2 adrenergic receptor activation. Modulation of the proline kink in transmembrane 6 by a rotamer toggle switch. *The Journal of biological chemistry*, 277(43):40989–96, oct 2002.

[376] Stanley J Opella. Structure Determination of Membrane Proteins in Their Native Phospholipid Bilayer Environment by Rotationally Aligned Solid-State NMR Spectroscopy. *Accounts of Chemical Research*, 46(9):2145–2153, sep 2013.