THE UNIVERSITY *of* EDINBURGH

A World of Probabilities: A Molecular Dynamics and Markov State Modelling Approach for Rational Design of Allosteric Modulators

Adele Hardie

A thesis presented for the degree of Doctor of Philosophy

THE UNIVERSITY
of EDINBURGH

School of Chemistry

University of Edinburgh

United Kingdom

2024

Declaration of Sole Authorship

I declare that the thesis *"A World of Probabilities: A Molecular Dynamics and Markov State Modelling Approach for Rational Design of Allosteric Modulators"* is an original report of my research. I confirm that:

- This thesis was written solely by me and has not been submitted for any previous degree.
- Except where outlined in the text, the work presented is entirely my own.
- Where I have used published work of others, this is clearly referenced and credit given.
- Any contributions from colleagues in a collaboration are explicitly referenced in the text.

Part of this thesis (Chapter 2) has been published in: Hardie, A.; Cossins, B. P.; Lovera, S.; Michel, *J. Commun. Chem.* **2023**, *6*, 125.

Year: 2024

Title: A World of Probabilities: A Molecular Dynamics and Markov State Modelling Approach for Rational Design of Allosteric Modulators

Author: Adele Hardie

Abstract

Even with current scientific and technological advances, drug discovery is a lengthy and expensive process. With a large number of pharmaceuticals already on the market and increasingly stricter regulations, it is difficult to design compounds that either are a significant improvement on the existing drugs or aimed at a novel target. In the light of this, allosteric modulators are a source of novelty in the field of drug discovery. Allosteric sites, i.e. sites that are distinct from the active site, tend to have high variety and low conservation even between proteins of the same family. Designing allosteric, rather than orthosteric, modulators allows for improved drug profiles, new ways of drugging already targeted proteins, and even revisiting targets previously deemed undruggable.

Aided by progress in structural biology and computing power available, computer aided drug design methods are heavily utilized in the study of allosteric modulation. There are multiple allosteric pocket detection and residue network analysis tools available to the computational chemist, however the effect a ligand binding to an allosteric site might have on the protein conformational ensemble remains difficult to quantify. Approaches using machine learning and Markov modelling have been in development, however they require the use of molecular dynamics (MD) simulations that are currently too time consuming for practical applications.

This thesis contains the development and application of a joint steered MD (sMD) and Markov State Modelling (MSM) approach, to reduce the computational time required to sample relevant conformational space of the protein. In this workflow, sMD simulations are used to bias the protein system from functionally "active"

to “inactive” conformations, and vice versa. From the sMD trajectories, a range of protein conformations is sampled, including unstable intermediate conformations not routinely accessible via standard MD methods. Each of these conformations serves as a new starting point for a swarm of unbiased MD simulations, allowing this methodology to leverage the increasingly available parallel computing infrastructures. These “seeded” MD simulations are combined to build MSMs, which describe the protein conformational ensemble. The MSMs are modelled in parallel, so that the probability values of states can be directly comparable across MSMs. The state probabilities of a protein system with no potential allosteric modulators are used as a baseline, and ligands are characterized based on the changes they induce. If the presence of a ligand decreases the probability of a state defined as “active”, the ligand is therefore an allosteric modulator. On the other hand, if the ligand increases the probability of this state, it is an activator.

The main body of this thesis consists of application of the above methodology to three protein systems: Protein Tyrosine Phosphatase 1B (PTP1B), Exchange Protein directly Activated by CAMP 1 (EPAC1), and Polycystic Kidney Disease 2 (PKD). Each system highlights a different class of drug target and activation mechanism. Additionally, each chapter emphasizes various considerations and caveats of applying sMD/MSMs to allosteric modulator assessment. Firstly, the workflow is validated for the first time on known inhibitors of PTP1B. The inhibitors target two distinct allosteric sites, and the trends in the experimentally measured inhibition are captured by the MSM modelled state probabilities. Additionally, the importance of comprehensively describing the protein conformational changes during sMD is discussed. The different effects of the ligands on PTP1B activity are also related to the different protein-ligand interactions observed in molecular dynamics simulations.

Secondly, the approach is applied to EPAC1, this time modelling activation by cAMP and partial activation by compound I942. Furthermore, while the function of PTP1B was defined by small loop motions, the activation of EPACs involves a large domain rearrangement and a two-step mechanism. A three-state conforma-

tional ensemble model is discussed for EPAC1, capturing activation by cAMP and partial activation by I942. As the description of protein dynamics in three states is more complex, data-driven method metastable state partitioning is less reliable. The state assignment was done manually, based on knowledge of EPACs activation, highlighting the non-triviality of biologically relevant state assignment. To investigate the differences between cAMP and I942, the latter is modelled with a variety of restraints that mimic protein-ligand interactions observed with cAMP. This allows to make MD-guided suggestions to further I942 lead development into a full activator.

Finally, the above sMD/MSM methodology is applied to PKD2, illustrating a more complex scenario where less data is available. Multiple considerations were taken when modelling PKD2, such as simulating a membrane and truncating the protein. As no small molecule modulators of PKD2 are known, the goal of this chapter is to investigate the regulatory effect of PI(4,5)P_2 membrane lipid on PKD2. As a control, the activation by a gain-of-function mutant was modelled first, followed by inhibition by PI(4,5)P_2. This example gives insight into future scalability of the sMD/MSM workflow presented in this thesis, and considerations for application to real-life drug design projects.

Lay Summary

While the wealth of knowledge and modern technologies available to scientists is constantly growing, it is harder than ever to get new medicines on the market. Any novel pharmaceutical has to either work on new, usually difficult, targets, or be a significant improvement on already existing treatment options. Additionally, the safety regulations for drugs are becoming more and more strict (with good reason). All of this together makes drug discovery very long and expensive, resulting in the huge multi-billion dollar pharmaceutical industry.

One of the most common methods of developing pharmaceuticals is designing (small) molecules, that bind to (large) proteins and change their behaviour in the human body, in such a way that prevents or alleviates disease. The question of where on the protein the molecule should bind is not always obvious. A straightforward option is the part of the protein that carries out some sort of function (the active site), which would result in blockage of the site and therefore preventing the protein from carrying out that function. However, that may not be an option for a multitude of reasons, and a different part of the protein may be targeted, hoping that it would perturb the structure of the protein enough to still get the desired effect. This is known as allosteric modulation, and a site that is not the active site is known as an allosteric site. Since our hypothetical molecule is no longer directly affecting the active part of the protein, its effects on protein function are harder to predict. Some molecules binding to allosteric sites may completely change the protein function, and some may have no effect at all.

A big part of the resource and time cost in the development of new medicines

comes from the need to make and test large numbers of molecules, before a safe and effective drug is designed. The inclusion of computer-aided drug design (CADD) aims to make this process cheaper and faster by creating computational models of the molecules and making predictions on their properties and how they may interact with the protein of interest. Due to advances in structural biology, atomic-resolution structures of target proteins are often available and provide insight into how a drug candidate may affect the target before it is ever tested in the lab. The advances in computational technologies make more and more complex calculations routinely available and allow more accurate models of the protein-drug system to be built.

This thesis covers a new computational approach to modelling the effects of molecules binding to allosteric sites. We employ two methodologies: molecular dynamics (MD) simulations, and Markov State Models (MSMs). MD simulations predict how atoms move over time, and so allow to have a dynamic movie instead of a static picture of our target system. However, many of the changes that are of interest occur more slowly than we can routinely simulate using the standard approaches. Therefore, we include an enhanced sampling method called steered MD simulations (sMD), where we purposefully push the system from one structural conformation to another (imagine a box being opened or closed). The simulations are inherently biased, and therefore are not a great source of data for further model building. Therefore we sample a range of conformations from these simulations (think of a box being barely open, half-way open, and almost but not fully open) and use those snapshots as starting points for more simulations, this time unbiased. Then the data from those simulations is used to build a Markov State Model, which uses the information about the movement of the system to give the probabilities of the system to exist in certain conformations. In our box analogy, if the box shuts closed every time is is partially open, then it has the highest probability to be in a closed state. On the other hand, if it pops open when it is partially open, a more stable conformation for the box is the open state. Now imagine placing an object inside the box - does this change whether the box is more likely to be open or closed,

or maybe has no effect at all? The object is our drug candidate molecule, and the approach outlined in this thesis works by comparing the probabilities of the protein being in a state that is active when the molecule is absent and present. Depending on the desired effect on the target, the goal may be to design molecules that increase (activators) or decrease (inhibitors) this probability.

This approach is validated on three different protein targets. First is Protein Tyrosine Phosphatase 1B, a protein involved in insulin signaling and a potential target for obesity and type II diabetes, outlined in Chapter 2. This includes the initial development of the methodology and proof of concept on a few known inhibitors. Secondly, the methodology was applied to Exchange Protein Activated by cAMP (cyclic Adenosine Monophosphate) 1 (EPAC1) in Chapter 3, which includes validation that the effects of activators are also captured. Lastly Chapter 4 outlines the modelling of Polycystic Kidney Disease 2 (PKD2). The approach is validated by modelling a mutant version of the protein that results in a more functionally active form, and then the effects of a membrane lipid Phosphotidyl Inositol Bisphosphate (PIP_2) are investigated. These examples showcase the success of the methodology in capturing effects of allosteric modulators, as well as illustrates the scope and limitations in applying it to drug design endeavors.

Acknowledgements

Firstly I would like to express my gratitude to my supervisor at the University of Edinburgh, Prof Julien Michel. Your mentorship and guidance throughout these years provided incredible support and made my PhD experience very enjoyable. Our weekly meetings always left me so motivated, optimistic and inspired. This opportunity has allowed me to grow so much as a scientist, and I hope we get to work together more in the future. I would also like to thank Dr Silvia Lovera, my supervisor at UCB. I am very grateful that we had the chance to have this unexpected collaboration and for all the support even over long distance. Hopefully our paths will cross again in the small world of computational chemistry!

I would also really like to thank everyone in lab 234. I was very lucky to have worked in such a wonderful environment, particularly the lunch time chats have kept me going through failed simulations and long days writing. Notably Anna, Cecilia, and Shivani, I miss you all already and look forward to lots of visits in Edinburgh.

Last, but definitely not least, I want to thank my husband Joe, who has been my heart and my rock all these years. None of this would have been possible without your constant support and amazing coffee. I am so grateful to you and our little April for being there for me for this chapter in our lives.

Contents

List of Figures

List of Tables

List of Abbreviations

3D 3-Dimensional

ADPKD Autosomal Dominant Polycystic Kidney Disease

AMBER Assisted Model Building and Energy Refinement

aMD Accelerated Molecular Dynamics

AML Acute Myeloid Leukemia

AMMo Allostery in Markov Models

ATP Adenosine Triphosphate

BB Benzbromarone

CADD Computer-Aided Drug Design

cAMP cyclic Adenosine Monophosphate

CDC25HD CDC25 Homology Domain

CHARMM Chemistry at Harvard Macromolecular Mechanics

CK Chapman-Kolmogorov

cNBD cyclic Nucleotide Binding Domain

COVID Coronavirus Disease

CR Catalytic Region

cryo EM Cryogenic Electron Microscopy

CV Collective Variable

DEP Disheveled Egl-10 Plecstrin

DFT Density Functional Theory

EC_{50} Half Maximal Effective Concentration

ED-EYA2 Eyes Absent Homolog 2

EPAC Exchange Protein directly Activated by cAMP

FBDD Fragment Based Drug Disvoery

FDA U.S. Food and Drug Administration

FEP Free Energy Perturbation

FF Force Field

GABA γ-Aminobutyric Acid

GAFF2 General AMBER Force Field 2

GaMD Gaussian Accelerated Molecular Dynamics

GDP Guanine Disphosphate

GEF Guanine Exchange Factor

GoF Gain of Function

GPU Graphical Processing Unit

GTP Guanine Triphosphate

HAD Haloacid Dehalogenase

HPC High Performance Computing

HTS High Throughput Screening

IC_{50} Half Maximal Inhibitory Concentration

IPF Idiopathic Pulmonary Fibrosis

IQR Inter-Quartile Range

ITS Implied Timescales

KIR Inward Rectifying K^+

KNF Koshland-Nemety-Filmer

KO Knock Out

LBDD Ligand Based Drug Discovery

LEaP Link Edit and Parameterize

MD Molecular Dynamics

MetaD Metadynamics

ML Machine Learning

MSM Markov State Model

MWC Monod-Wyman-Changeux

NMR Nuclear Magnetic Resonance

NPT Number Pressure Temperature

NVE Number Volume Energy

NVT Number Volume Temperature

OPLS Optimized Potentials for Liquid Simulations

OPM Orientations of Proteins in Membranes

PBC Periodic Boundary Conditions

PCCA Perron Cluster-Cluster Analysis

PCN Protein Contact Network

PDB Protein Data Bank

PDK1 Phosphoinositide Dependent Kinase 1

PI(3,4,5)P_2 Phosphatidyl Inositol-3,4,5-Triphosphate

PI(3,5)P_2 Phosphatidyl Inositol-4,5-Bisphosphate

PI(4,5)P_2 Phosphatidyl Inositol Bisphosphate

PKD2 Polycystic Kidney Disease 2

PME Particle Mesh Ewald

POPC 1-Palmitoyl-2-Oleoyl-sn-Glycero-3-Phosphocholine

PPD Post-Partum Depression

PSP Protein Serine/Threonine Phosphatase

PTM Post Translational Modification

PTP Protein Tyrosine Phosphatase

PTP1B Protein Tyrosine Phosphatase 1B

P-Tyr Phospho-Tyrosine

PyEMMA Python Emma's Markov Model Algorithms

QM Quantum Mechanics

RA Ras Association

RAP1 Ras-related Protein 1

REM Ras-Exchange Motif

RMSD Root Mean Square Deviation

RR Regulatory Region

SARS-CoV-2 Severe Acute Respiratory Syndrome Related Coronavirus 2

SBDD Structure Based Drug Discovery

sMD Steered Molecular Dynamics

SMILES Simplified Molecular-Input Line-Entries

TCPTP T-Cell Protein Tyrosine Phosphatase

tICA Time Independent Component Analysis

TIP3P Transferable Intermolecular Potential with 3 Points

TMEM16F Transmembrane Protein 16 F

TRP Transient Receptor Potential

TRPA Transient Receptor Potential Ankyrin

TRPC Transient Receptor Potential Canonical

TRPM Transient Receptor Potential Melastatin

TRPML Transient Receptor Potential Mucolipin

TRPN Transient Receptor Potential NO-Mechanopotential

TRPP Transient Receptor Potential Polycystin

TRPV Transient Receptor Potential Vanilloid

USD United States Dollars

VV Velocity Verlet

WT Wild Type

Chapter 1

Introduction

1.1 Computer-Aided Drug Design

1.1.1 Contemporary Drug Design

The modern drug design process has come a long way since the serendipitous and natural remedy based approach of the 19th century and before[1–3]. The concept of biomacromolecules as specific drug targets, advances in molecular structure elucidation, pharmacokinetic studies, characterization of "drug-likeness"[4], cell and animal models, and the rise of computational chemistry all have shaped the drug discovery landscape[2, 3]. A common approach has become to develop small molecule drugs - compounds that have a low molecular weight and are usually orally bioavailable[5].

The general process timeline is split into three stages: early discovery, pre-clinial studies, and phase 1-3 clinical trials (Figure 1.1)[6, 7]. The early discovery phase consists of target identification, high throughput screening (HTS) to find hits, developing those hits into leads, and lastly optimizing the leads to maximise their desired properties (such as affinity for the target or good oral bioavailability) while minimising adverse properties (e.g. toxicity). After testing in animal models during the pre-clinal stage, the drug candidates are trialed on humans in the clinic[6, 8]. In this pipeline, computational chemistry has a prominent and still growing role in the early stage drug discovery, leveraging the access to 3D target structures, growing

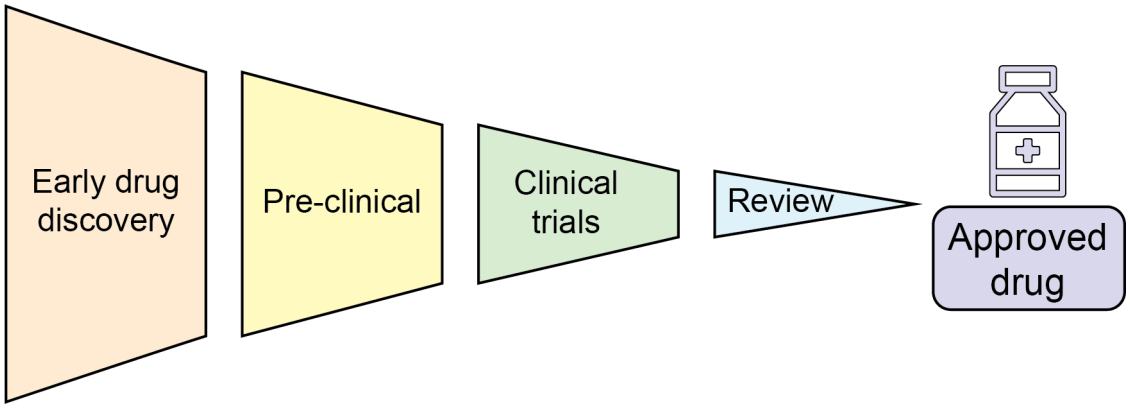

Figure 1.1: The general drug discovery timeline. The process begins with a large number of drug candidates, decreasing at each step.

accessible chemical space, and High Performance Computing (HPC)[9].

1.1.2 Examples of *in silico* Approaches

The goal of computational chemists supporting early stage drug discovery is to model and predict the properties of drug candidates, so that the usually more time-intensive and costly experimental testing is focused on a smaller number of molecules. There are two main categories of computational modelling in small molecule drug design: ligand- (LBDD) and structure-based drug discovery (SBDD)[10, 11].

Ligand-Based Drug Discovery

Ligand-based drug discovery strategies are employed when speed over accuracy is desired or when the structure of the target is not known[11]. In these cases, only data obtained from the *ligand* is considered, assuming that similar molecules will have similar biological activity. Simplified Molecular-Input Line-Entries (SMILES)[12, 13], molecular fingerprints[14], or sets of physical property descriptors[15]. These representations can then be employed to virtual screening of a large database for compounds similar to a reference molecule (e.g. via the Tanimoto similarity used for fingerprints)[16, 17]. A reference ligand can also be stripped down to its relevant structural features - the pharmacophore - to search for molecules that match those rather than ones that are simply similar to the reference[18].

Structure-Based Drug Discovery

The advances in X-Ray crystallography, cryogenic electron microscopy (cryo EM) and computational homology modelling has made atomic resolution 3D structures of drug targets more commonly accessible. Target structural information allows drug discovery to leverage the protein-ligand interactions into consideration, which allows to better predict how the drug candidate might act in the body. One of the most commonly used SBDD methods is molecular docking, which models the binding pose of a small molecule to a protein target, and the associated binding affinity in the

terms of a docking score[9, 10, 19]. While this is slower than LBDD methods such as Tanimoto similarity searching, recent advances in accelerating docking are making it feasible to routinely apply to screening large libraries[20–22]. Another key method is alchemical free energy calculations, which use molecular dynamics (MD) to compute the relative binding affinity between two compounds to some protein target. The improvement of force fields, MD engines, free energy computation methods and HPC now allow for fast and accurate (< 1 kcal mol^{-1}) results[23–25].

Fragment-Based Drug Discovery

Another drug design strategy that employs both ligand- and structure-based approaches is fragment-based drug design (FBDD). The possible chemical space of small drug-like compounds is estimated to be as large as 10^{20}-10^{24} molecules[26], meaning that extensively sampling such a large space remains a challenge, even with the improvements on HTS. FBDD is an alternative strategy, involving screening compounds that have lower molecular weight (< 300 g mol^{-1}). As they are made up of fewer atoms, the number of fragments that need to be sampled to cover a representative chemical space is much smaller, making it easier to find hits. Afterwards, the hit fragments can be further elaborated into larger, drug-like molecules via fragment growing, merging, or linking[27, 28].

1.1.3 Common Challenges in Drug Design

Even with the modern technological advances, getting a drug on the market is a lengthy (12-15 years) and expensive (>1B USD) process[6]. The attrition rate for clinical trials (i.e. for compounds that have already passed the rigorous preclinical testing) is still 90%[29]. The general decline in successful drug projects per billion USD is called 'Eroom's Law', the reverse of Moore's Law which describes the exponential improvement of computing power with time[7].

There is already a large variety of available effective drugs on the market. Any new drugs would have to be significantly improved or target a niche, difficult to treat

diseases. Only about 10% of the human genome is estimated to code for 'druggable' proteins (ones that could be targeted with small molecules)[30], making it difficult to even choose a drug target to start[31]. Additionally, the safety regulations have only become stricter[32]. Even the rapidly growing accessible chemical space poses a challenge. With more compounds to screen, larger portions of the process have to rely on automation and quick HTS methods, which can allow potentially good drug candidates to be filtered out[7, 9, 33].

1.2 Allosteric Modulation

Allosteric modulation describes the way two distant protein sites are connected via a dynamic network of residues, allowing changes in one site to have an effect on the other[34–37]. The first study of allostery was carried out by Bohr in the early 20^{th} century, analysing the effect of oxygen on hemoglobin[34, 38]. Since then, it has been researched extensively, and it is expected that most, if not all, proteins have some potential for allosteric modulation[39–41].

1.2.1 Models of Allosteric Modulation

Originally, two main mechanisms of allosteric modulation were proposed: the concerted Monod-Wyman-Changeux (MWC) model[42], and the Koshland-Nemety-Filmer (KNF) model of sequential induced fit[36, 43]. The MWC model suggests that the protein exists in two distinct states, tense and relaxed, and the binding of an allosteric modulator shifts the equilibrium between those states (Figure 1.2a)[36, 39, 42]. Contrarily, the KNF model claims that the protein shifts between the two major states sequentially upon ligand binding - the modulator induces a change in the protein structure, perturbing the distant active site (Figure 1.2b)[43, 44]. Both of these agree on the existence of two (or more) metastable states in the protein, and that the transition between them happens upon ligand binding[36, 37, 44].

The current dominant explanation of allosteric modulation is a more general

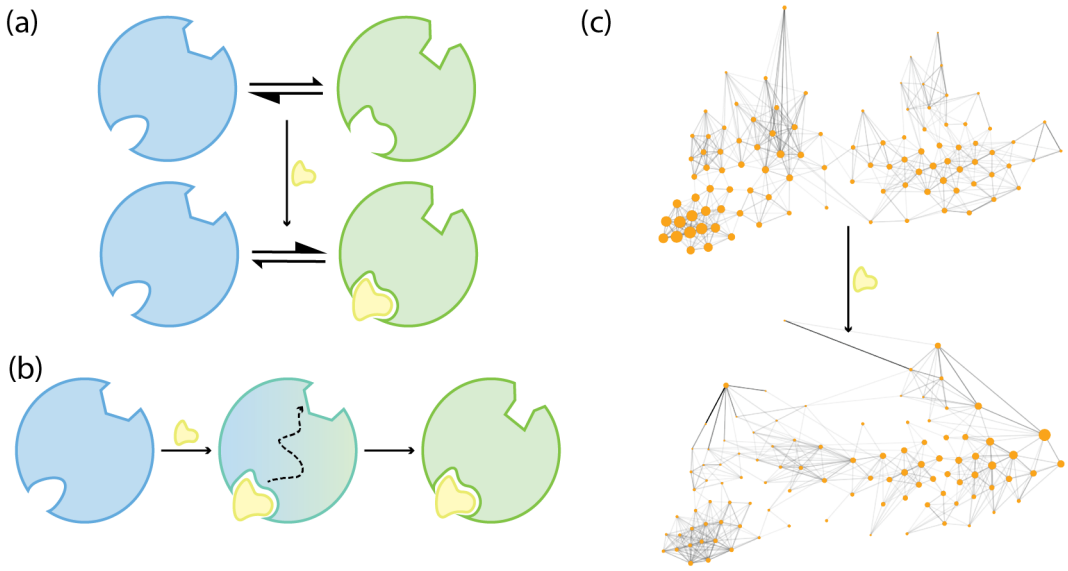

Figure 1.2: The different models to explain allosteric modulation. (a) The MWC model implies an equilibrium between two states (blue and green), which shifts upon binding of an allosteric modulator (yellow). (b) The KNF model proposes a sequential change from one state (blue) to another (green), upon binding of an allosteric modulator (yellow). (c) The ensemble model of allostery view protein dynamics as an ensemble of states. Each protein state is represented as an orange circle, where a larger size indicates a larger probability. The lines represent transitions between states, with darker lines indicating a more likely transition. When an allosteric modulator (yellow) binds to the protein, the states themselves remain the same, but their energies and the transition probabilities change.

model of a change in the population of the protein conformational ensemble. It is common to visualize and think of proteins in a static way, rendering and inspecting individual 3D structures. A simple two-state explanation of allostery then seems natural, as these most stable states would be the ones available for resolution. However, nuclear magnetic resonance (NMR) and simulation studies reveal that proteins are dynamic, displaying constant changes in conformation, even if they are subtle. Therefore, protein dynamics should be viewed as ensembles of states, which the protein visits at different frequency based on the stability of each state[45]. By changing the energetics of the ensemble, an allosteric modulator shifts the population distribution, making certain states more common (Figure 1.2c). Notably, all conformations are possible with or without the ligand - what changes is the probability of reaching them[39, 44, 46].

1.2.2 Allostery in Drug Design

The use of allosteric modulators provides new opportunities in pharmaceuticals that would not be available with focusing on orthosteric sites only. Because of the diversity of allosteric sites, allosteric drugs have the potential for higher selectivity and lower off-target effects[47, 48]. For instance, orthosteric phosphoinositide-dependent kinase 1 (PDK1) inhibitors have issues with selectivity due to competing with adenosine triphosphate (ATP), a common signalling molecule involved in many biological processes. On the other hand, allosteric inhibitors target different sites and show high selectivity[49, 50]. Additionally, some protein targets have been deemed undruggable due to the nature of their orthosteric site, e.g. the charge or shape[41, 51]. Some proteins lack a defined orthosteric site altogether, such as those involved in protein-protein interactions[52]. Bypassing the difficult ligand pockets and focusing on other binding sites on the protein may be the only feasible approach to target such proteins with small molecules.

While they still make up a very small portion of approved drugs on the market[53], allosteric modulation is still an active area of pharmaceutical research. Just

a few days prior to the writing of this chapter, a new thyroid hormone receptor β (-β) allosteric activator, resmetirom, was approved by the U.S. Food and Drug Administration (FDA)[54, 55]. Similarly, zuranolone, a γ-aminobutyric acid ($GABA_A$) allosteric activator, was approved for the treatment of post-partum depression (PPD) last year[56, 57]. Other allosteric drugs, such as enasidenib[58] (ClinicalTrials.gov ID: NCT02677922), for the treatment of acute myeloid leukemia (AML), and RLY-1971[59] (ClinicalTrials.gov ID: NCT04252339), for metastatic solid tumors, are also currently in clinical trials.

However, allosteric drug discovery is not without its challenges. One of the main difficulties lies in the discovery of the allosteric pocket itself. They are less conserved, often more shallow and less defined than orthosteric sites. Allosteric sites can also be cryptic, opening up only in the presence of the ligand, which may not be apparent in available 3D structures[41, 48]. A variety of methods exist to identify cryptic pockets, such as MD simulations[60, 61] and high throughput X-Ray crystallography[62, 63], but the possibility of any pocket, other than the orthosteric one, being an allosteric site makes it difficult to focus design efforts early on[48]. It is also not immediately clear whether an allosteric ligand will have an agonistic or antagonistic effect. Similar compounds binding to the same site can have vastly different regulatory effects on the protein, which is known as molecular switching[64–66]. Nonetheless, allosteric modulation has already made a significant impact on modern drug discovery that only continues to grow as more allosteric drugs reach the market.

1.2.3 Computationl Methods to Study Allostery

With the increase in computing power routinely available, modern computational simulation and analysis methods have the capability to provide new insights into the ensembles of protein dynamics. Combined with advances in structural biology, this makes *in silico* approaches valuable in the rational development of allosteric modulators[67].

Pocket Detection

As already mentioned in section 1.2.2, a significant issue with designing novel allosteric modulators is the initial discovery of an allosteric site to target. For sites present in available 3D structures, methods using protein figerprints[68] or physicochemical descriptors[69] have been trained based on information on known allosteric sites. Similarly, graph representations of proteins have also been used to detect distant sites linked to the orthosteric site[70]. While these methods are relatively quick to run, using just a few static structures risks excluding cryptic pockets that emerge only in the presence of a ligand. MD simulations provide a dynamic view of proteins and allow to probe the potential allosteric sites. Available methods include mixed solvent MD, where probe molecules can induce the opening of allosteric sites[71, 72], and enhanced sampling methods, where dynamics are accelerated to uncover hidden pockets on a faster timescale[60, 73, 74].

Residue Network Analysis

The identification and analysis of the residues that connect the allosteric site to the active site is useful in understanding the underlying allosteric mechanism. A common approach is to represent the protein as a graph, where each residue is denoted as a node, connected by edges determined based on geometric or energetic properties. For example, in protein contact networks (PCNs), the distance matrix between C_α atoms $\mathbf{d} = \{d_{ij}\}$ is used to determine residue-residue interactions (the smaller the distance, the stronger the interaction)[75]. The atom coordinates can come from an experimental structure on an average ensemble from an MD simulations. As with identifying allosteric pockets, this static view at a single protein conformation (or a simulation at only one end of the allosteric pathway) can cause loss of information on the allosteric mechanism. The answer to that is to look at how the residue contacts change between the two end states of an allosteric mechanism[76]. Residues that are distant to the orthosteric site, but still show a large difference in behaviour, can be inferred to be part of the allosteric network[76, 77].

Allosteric Modulator Assessment

A key step in rational design of allosteric modulators is to assess the effects of small molecules on protein activity. As with orthosteric sites, strategies such as docking[9, 10, 19] or free energy perturbation (FEP) methods[23–25] can be employed to assess any ligand binding affinity. Additionally, methods for modelling and ranking protein interactions with allosteric ligands specifically are available[78]. However, when it comes to allosteric modulation, it is not always clear whether a ligand binding at an allosteric site will necessarily have a modulatory effect on the distant orthosteric site. The currently emerging approaches to quantify the effects of allosteric ligands on protein conformational ensembles employ machine learning[79] or Markov modelling[80, 81] to compare unliganded and ligand-bound protein dynamics. The difficulty of these dynamics-based methods lies in the duration of the simulation required to capture how the ligands perturb the conformational ensemble of the proteins, making it difficult to routinely apply in drug design strategies.

1.3 Molecular Dynamics Simulations

While quantum mechanics (QM) simulation methods, such as density-functional theory (DFT), can predict particle behaviour from first principles[82], it is not currently practically feasible to apply to such large systems as a protein in solution. Instead, molecular dynamics simulations use Newton's equations of motion to model atom movement over time. For biomacromolecules, such as proteins, they illustrate dynamic system properties with atomistic resolution[83, 84]. The *in silico* nature of MD experiments also makes various perturbations of the system possible, for instance alchemical tranformations of ligands to calculate the difference in their energy, or applying an external force to part of the protein to observe how it will affect its dynamics[24, 85].

1.3.1 Integrators

The starting point for a simulation is a set of coordinates for all atoms of the system of interest. For proteins, Protein Data Bank (PDB) entries of experimentally resolved structures are often used. The aim of a simulation is to model how these coordinates will change with time. This is achieved by applying Newton's second law, which states that the force $\mathbf{F}$ acting on an object is the rate of change of momentum p, which can be expressed as the mass of the object (m) times acceleration (a):

$$\mathbf{F} = \frac{d\mathbf{p}}{dt} = m\mathbf{a} \tag{1.1}$$

And acceleration is the second derivative of the position $\mathbf{r}$ of an object:

$$\mathbf{a}(t) = \frac{d^2\mathbf{r}}{dt^2} \tag{1.2}$$

Integrating Equation 1.2 above yields the evolution of coordinates over time. In the cases of molecular dynamics, the simulation is divided into small timesteps and the finite difference method is applied. The usual timestep in MD simulation is 2 femtoseconds (or 10^{-15} seconds), to account for vibration motions in the system involving hydrogen atoms (the lightest atoms), such as the C-H bond stretching[86]. The algorithms for MD simulations operate under the assumption that the integral of Equation 1.2 can be expanded into a Taylor series:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 + ... \tag{1.3}$$

where $\mathbf{r}$ is the position of an atom, $\mathbf{v}$ is its velocity, and $\mathbf{a}$ is the acceleration. t is current time, and δt is the timestep of the simulation.

A common integration algorithm used for MD simulation is the Verlet algorithm. In addition to Equation 1.3 above, it also makes use of a similar expression for the position of the previous step:

$$\mathbf{r}(t - \delta t) = \mathbf{r}(t) - \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 \tag{1.4}$$

Adding equations 1.3 and 1.4 yields:

$$\mathbf{r}(t + \delta t) = 2\mathbf{r}(t) + \mathbf{r}(t - \delta t) + \mathbf{a}(t)\delta t^2 \tag{1.5}$$

Using Equation 1.5, only the acceleration, as well as the current and previous atom positions are required. This makes the Verlet algorithm quite simple and effiecient. However, due to the lack of a velocity term, the velocity of the current position at time t cannot be computed, only at previous step $t - \delta t$[87]. The Velocity Verlet (VV) algorithm computes the velocities at the same time as the positions. A commonly used similar algorithm is the leapfrog integrator, where the velocities are computed at half-steps instead, such that the computation of positions and velocities alternates[88].

1.3.2 Force Fields

The above computations of the evolution of atom coordinates with time rely on using acceleration $\mathbf{a}$. As per Equation 1.1, it relates to the forces acting upon the atoms in the system. The force can be expressed as a negative derivative of the potential energy of the system with respect to the change in position:

$$\mathbf{F} = -\frac{\delta U}{\delta \mathbf{r}} \tag{1.6}$$

In molecular dynamics, the potential energy of the system is usually defined by the following functional form:

$$
\begin{aligned}
U(R) = & \frac{1}{2}\sum_{bonds} K_b(b-b_0)^2 + \frac{1}{2}\sum_{\substack{bond\\angles}} K_\Theta(\Theta-\Theta_0)^2 \\
& + \frac{1}{2}\sum_{torsional} K_\phi[1+\cos(n\phi-\delta)] \\
& + \sum_{\substack{non-bonded\\pairs}} (\frac{A}{r^{12}} - \frac{B}{r^6} + \frac{q_1 q_2}{Dr})
\end{aligned}
\tag{1.7}
$$

In Equation 1.7 above, U is the energy, R is the atom coordinates used to calculate b (bond length), Θ (bond angles), ϕ (torsional angles) and r (distance between atoms). K for each term denotes bond force constant, bond angle force constant, and torsional angle force constant, respectively. b_0, Θ_0, and ϕ_0 represent the ideal values for these terms. The non-bonded interactions, represented in the last term, include the Lennard-Jones 6-12 potential and the electrostatic interactions, where A, B, q_1, and q_2 depend on the atoms involved in the pair[89].

Equation 1.7 involves constants such as ideal bond length and angle values, or the force constants. These values are obtained from fitting to quantum mechanics and experimental results, and are accumulated into large sets of parameters, called force fields (FFs). Even though the lower level of theory makes force field methods less accurate compared to QM, they have been significantly improved over the years, and are still crucial in simulating large systems[84].

Empirical Force Fields

The most common current force fields are AMBER[90], CHARMM[91] and OPLS[92]. Their development started at a similar time in the 80s, and began by representing non-polar hydrogen atoms as united with the carbon atoms they were bonded to. As this was shown to be insufficient to capture relevant system dynamics, they switched to fully explicit representations in the 90s. A significant effort has been made to provide a large number of parameters, with constant revisions and additions[93]. The downside of this approach is that specifying a (relatively simple) functional form is

limiting, as is the requirement to transfer parameters from a small set of molecules to more general scenarios[94, 95].

Machine Learning Force Fields

With the emerging boom of machine learning (ML) in computational chemistry, ML force fields are proposed as a bridge between the cheap but inaccurate MD simulations, and the high level QM methods. They involve using an ML model, such as a neural network, to model the energy of a system as a function of atom coordinates. The functional form does not have to be specified, allowing for much more flexibility. There is no need to predefine system properties such as bonds, as all behaviour can trained from the data used[96]. This requires a large amount of data, as the ML force field would not be able to represent conformational space not included in the training set[97]. However, they are already being applied to MD simulations and could potentially replace traditional force fields altogether[98].

1.3.3 Periodic Boundary Conditions and Long Range Interactions

The MD system has a finite size, and therefore some boundary where it "ends". As the simulated atoms move in space, they may move out of the defined boundary of the system. Similarly, adjacency to the boundary may cause changes in inter-particle interactions. To account for this, periodic boundary conditions (PBC) are commonly applied[99]. The system box is considered as the unit cell, and is repeated periodically, to imitate an infinitely large system.

While the treatment of bonded terms in Equation 1.7 is clearly defined by the bonds in the system, non-bonded interactions are less straightforward. There are $O(N^2)$ non-bonded atom pairs in a system, and evaluating each one can significantly increase simulation time for large systems. Additionally, this may not even be necessary, as these interactions decay with distance and so contributions of interactions between atoms above a certain distance can be negligible. As the Lennard-Jones

potential (first non-bonded term in Equation 1.7) decays quickly, it is set to 0 above a certain cutoff distance. A switching function can also be introduced to smooth the change in potential to 0. Similarly, the short-range Coulomb interactions are also computed in the direct space. The long-range electrostatic interactions are less trivial to compute, as they decay slowly. There are a few methods to compute these long-range interactions, with the most common being Particle Mesh Ewald (PME). It uses Fast Fourier Transforms to handle the summation in reciprocal space, thus accelerating the calculation significantly[100].

1.3.4 Thermostats and Barostats

Integrating the equations of motion as outlined in section 1.3.1 yields the micro-canonical (NVE) ensemble. In the NVE ensemble, the number of particles, system volume and energy are all conserved. The chemical potential μ, temperature T, and pressure P are not conserved and may change significantly. This does not represent most experimental conditions, and therefore thermostats and barostats are employed to keep the system temperature and pressure at required value. These represent relevant experimental conditions, such as 1 atm and 300 K for simulation of proteins. Two commonly used ensembles are the canonical ensemble NVT and the isothermal-isobaric ensemble NPT. In both ensembles the temperature is allowed to fluctuate around an average value, via the use of a thermostat algorithm. Depending on the algorithm, this may involve rescaling of the system velocities (e.g. Berendsen thermostat) or the modification of the equation of motion itself e.g. Langevin dynamics)[101].

Similarly in the NPT ensemble, a barostat is used to keep the pressure fluctuating around a set value. This is usually achieved by changing the unit cell vectors, and therefore manipulating the volume. The Berendsen barostat, for instance, couples the system to an external cell of constant pressure, and scales the simulation system volume to match the pressures[102].

1.3.5 Enhanced Sampling Methods

While MD simulations provide valuable insight into protein dynamics, many relevant biological processes occur on a timescale that is too long for routine MD simulation. To combat this, a variety of enhanced sampling approaches that accelerate system dynamics have been developed, some of which are outlined below.

Metadynamics

Rather than using atom coordinates to describe changes in a system, it may be more useful to use a collective variable (CV), which is a lower dimensional function of atom coordinates, such as a distance between atoms. Metadynamics (MetaD) introduces a history-dependent external force that is a function of the CV(s) used. As the system starts in an energy minimum, Gaussian boost potentials are deposited within it, growing the bias until the system escapes that minimum and enters another one. Then the process starts again, until the simulation turns to a random walk along a flat energy surface (Figure 1.3a). This approach does not require knowledge of the energy landscape before the simulation and allows exploration of the wider conformational space. However, that also makes it less suitable for simulating a well-defined pathway for a conformational change[103].

Steered Molecular Dynamics

Steered molecular dynamics (sMD) simulations involve applying a harmonic restraint to the system, based on one or more CVs:

$$V(\mathbf{s}, t) = \frac{1}{2}\kappa(t)(\mathbf{s} - \mathbf{s}_0(t))^2 \tag{1.8}$$

where $\kappa(t)$ is the time-dependent force constant, $\vec{s}$ is the actual CV value, and $\vec{s}_0$ is the expected CV value. The expected CV values are defined for each step of the simulation, defining the path the system is steered to take. While the CV values are close to the expected values, the bias potential is low, and when the CVs

deviate from the defined path, the bias increases (Figure 1.3b). As sMD requires the definition of the expected values at every step of the simulation, prior knowledge of the pathway of conformational change is required[104, 105].

Accelerated Molecular Dynamics

While the previous methods use CVs to define the conformational space the protein will explore, it is useful to simulate rare events or sample higher energy conformations without predefined expectations. Accelerated MD (aMD) does this by applying a bias boost potential that modifies the true potential when the system energy is below a certain value. This evens out the energy surface and prevents the system from being trapped in energy wells[106]. An example is shown in Figure 1.3c. The original aMD approach makes it difficult to recover the unbiased free energy ensemble. Instead, an approach using a Gaussian boost potential (GaMD) reduces noise in simulation and allows to recover the energy landscape more accurately. This approach is even less directed than MetaD simulations, and while the lack of predefined CVs may be useful when not a lot is known about the system dynamics, it also means that there is no way to ensure relevant conformational space is sampled[107].

1.4 Markov State Modelling

Given the protein ensemble view of allostery that has emerged, Markov State Models (MSMs), which aim to model the protein conformational ensemble, have been increasing in their popularity to study allostery[37]. MSMs provide a way to model protein dynamics described as a few deciding conformational changes. Simple observations of occurrence of certain conformations during a few simulations may not provide the whole picture of the conformational ensemble. For instance, let us consider ten simulations of a protein, all starting from some functionally "active" conformation. If this conformation is very energetically unfavourable, the protein will adopt the more stable "inactive" conformation in all ten simulations. Looking only at the start and end points of these simulations and the conformations the

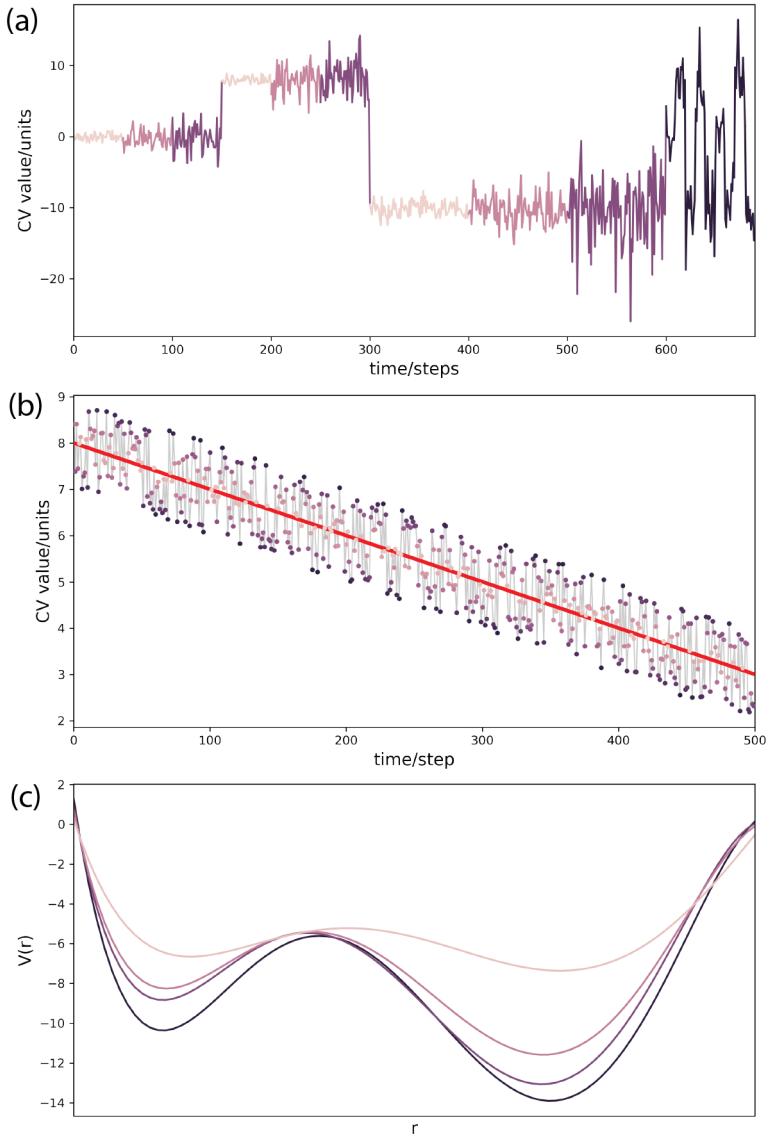

Figure 1.3: Examples of enhanced sampling methods. (a) During metaD simulations, the systems starts at a free energy basin at some CV value (0 in this example). As larger and larger biases are introduced, the system explores more and more of the basin, until it escapes and enters the basin closest in energy (CV value -10 units here). This continues until the simulation samples the whole energy surface randomly. Darker line colour here represents larger bias. (b) During sMD, the simulation pathway is described by defining the expected CV values for every step (red line). As the system deviates from the expected values during simulation (dots), a harmonic bias is added to steer the system towards the expected value. The higher the deviation, the higher the bias (dark dots), while if the system is already close to the expected value, the bias is lower (lighter dots). (c) An example free energy surface of a system, with different boost potentials during aMD. The darkest line indicates the original free energy. Lighter lines indicate increased acceleration. This raises the free energy minima, making it easier to explore a larger conformational space.

protein adopted, it might seem that the conformational ensemble has a 50/50 distribution of "active" and "inactive" conformations. However, knowing the *transitions* between these conformations reveals that the "active" conformation has very low probability, and protein would predominantly adopt the "inactive" conformation. In an infinitely (or at least reasonably) long simulation, the counts of states visited would reveal the relevant protein dynamics, but the computational cost of these is often prohibitive for routine application. By computing MSMs, the observed transitions between states from shorter simulations can be used to uncover the overall equilibrium dynamics of the conformational space sampled.

The theory behind MSMs is described in sections 1.4.1 and 1.4.2 below, as outlined by Prinz *et al*[108].

1.4.1 Continuous Dynamics

A Markov State Model is an approximation of a Markov chain, which in turn models a stochastic process where the next state transition probability only depends on the current state (a memoryless, or Markovian, process). Figure 1.4a shows a a simple example of a Markov chain involving two states. The transition probability density between two states is described as:

$$p(\mathbf{x}, \mathbf{y}; \tau)d\mathbf{y} = \mathbb{P}[\mathbf{x}(t+\tau) \in \mathbf{y} + d\mathbf{y} \mid \mathbf{x}(t) = \mathbf{x}] \tag{1.9}$$

where states $\mathbf{x}$ and $\mathbf{y}$ (e.g. conformations) belong to some state space Ω, τ is the timestep after time t, and $d\mathbf{y}$ is an infinitesimally small volume of space around state $\mathbf{y}$. In Ω, all states are connected, i.e. for an infinitely long walk along the chain ($t \to \infty$), each state will be visited an infinite number of times. At a given point in time t, the probability density $p_t(\mathbf{x})$ denotes the state distribution in the state space. After a timestep τ, the probability density changes to $p_{t+\tau}(\mathbf{x})$ according to a propagator $\mathcal{Q}(\tau)$.

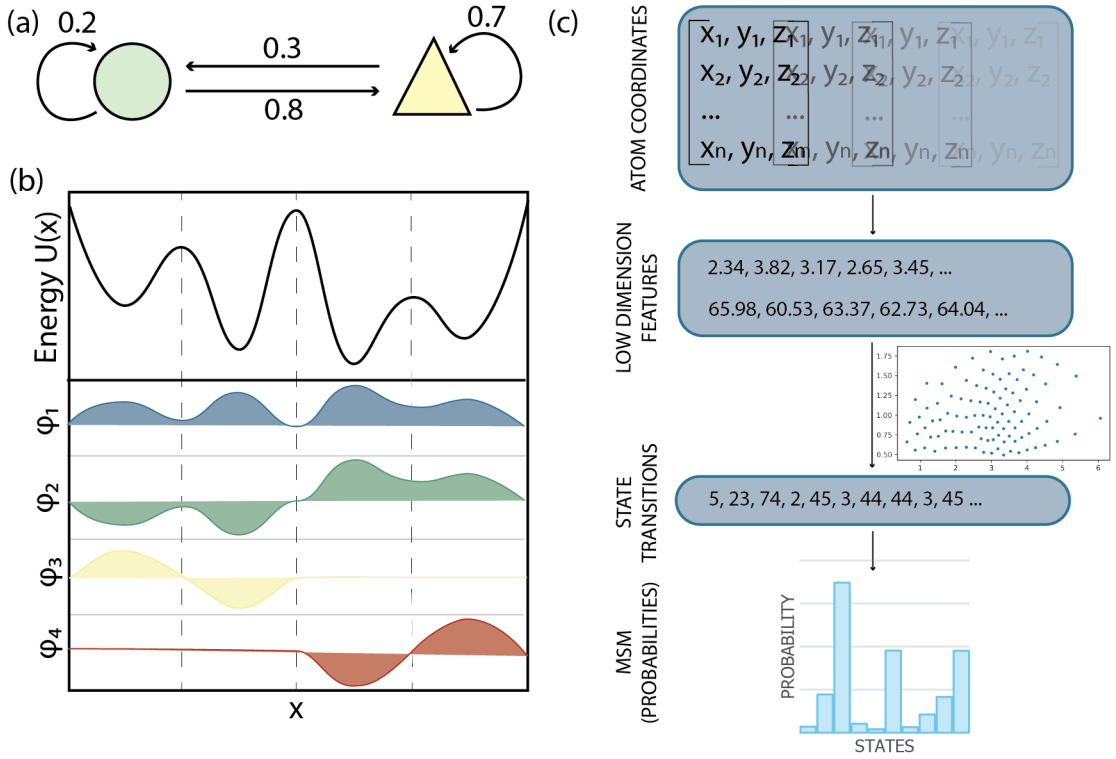

Figure 1.4: (a) A simple two state Markov chain, with the associated state transition probabilities. (b) The energy landscape of a model process with four main stable states. The 4 eigenfunctions are illustrated. (c) The processing an MD trajectory undergoes for MSM building. Initially the atom coordinates are reduced to a few select features, which are then clustered into states. Each frame of the MD trajectory is then assigned to a state.

$$p_{t+\tau}(\mathbf{y}) = \mathcal{Q}(\tau) \circ p_t(\mathbf{y}) = \int_{x\in\Omega} d\mathbf{x}\, p(\mathbf{x}, \mathbf{y}; \tau) p_t(\mathbf{x}) \tag{1.10}$$

i.e. all the state transitions are considered and a new probability density is established. Over an infinite time, the probability density must relax to the stationary density $\mu(\mathbf{x})$, which is an invariant property of the system. For molecular systems, that would be the distribution of the conformational ensemble at equilibrium. Re-weighting the probability densities by μ to yield functions $u_t(\mathbf{x})$ such that $p_t(\mathbf{x}) = \mu(\mathbf{x})u_t(\mathbf{x})$ and substituting it into Equation 1.10 gives the transfer operator, which will be considered for the rest of the section:

$$u_{t+\tau}(\mathbf{y}) = \mathcal{T}(\tau) \circ u_t(\mathbf{y}) = \frac{1}{\mu(\mathbf{y})} \int_{x\in\Omega} d\mathbf{x}\, p(\mathbf{x}, \mathbf{y}; \tau)\mu(\mathbf{x})u_t(\mathbf{x}) \tag{1.11}$$

The transfer operator has eigenfunctions with corresponding eigenvalues:

$$\mathcal{T}(\tau) \circ \psi_i(\mathbf{x}) = \lambda_i \psi_i(\mathbf{x}) \tag{1.12}$$

The first eigenfunction, ψ_1, has an associated eigenvalue of $\lambda_1 = 1$ and corresponds to the stationary process, or the overall system dynamics. The following eigenfunctions correspond to individual dynamic processes of the system (Figure 1.4b), with the corresponding eigenvalues all having a value between -1 and 1. Considering the first m slow processes (and distinguishing the rest as the combination of remaining faster processes) gives the following:

$$\begin{aligned} u_{t+k\tau} &= \mathcal{T}_{slow}(k\tau) \circ u_t(\mathbf{x}) + \mathcal{T}_{fast}(k\tau) \circ u_t(\mathbf{x}) \\ &= \sum_{i=1}^{m} \lambda_i^k \langle u_t, \phi_i \rangle \psi_i(\mathbf{x}) + \mathcal{T}_{fast}(k\tau) \circ u_t(\mathbf{x}) \\ &= \sum_{i=1}^{m} \lambda_i^k \langle u_t, \psi_i \rangle_\mu \psi_i(\mathbf{x}) + \mathcal{T}_{fast}(k\tau) \circ u_t(\mathbf{x}) \end{aligned} \tag{1.13}$$

where ϕ_i is the i^{th} eigenfunction on $\mathcal{Q}$, and k is the time index referring to the k-fold application of the transformer. As k increases, the slow processes decay

faster, and as $k \to \infty$, the only remaining slow process is ψ_1, i.e. the overall system dynamics. The eigenvalues $\lambda_{i=2,\dots,m}$ correlate to the timescale of their associated process[108]:

$$t_i(k\tau) = t_i = -\frac{\tau}{\ln|\lambda_i|} \tag{1.14}$$

1.4.2 Discrete Dynamics

The transformer operator outlined above applies to continuous dynamics of a system and describes the Markov chain of the process. However, a feasible description of system dynamics (e.g. an MD simulation trajectory) is inherently discretized by its timestep. Therefore what can be obtained from simulation data is only a Markov model, $\mathbf{T}$. The discrete state space is denoted as S and contains n number of sets, where points $\mathbf{x}$ each belong to a set. This means the transfer operator $\mathcal{T}$ becomes an $n \times n$ transition matrix $\mathbf{T}$, denoting transition probability from state i to state j after some time τ:

$$\begin{aligned}
\mathbf{T}_{i,j}(\tau) &= \mathbb{P}[\mathbf{x}(t+\tau) \in S_j \mid \mathbf{x}(t) \in S_i] \\
&= \frac{\mathbb{P}[\mathbf{x}(t+\tau) \in S_j \ \cap \ \mathbf{x}(t) \in S_i]}{\mathbb{P}[\mathbf{x}(t) \in S_i]} \\
&= \frac{\int_{\mathbf{x}\in S_i} d\mathbf{x}\mu_i(\mathbf{x})p(\mathbf{x}, S_j; \tau)}{\int_{\mathbf{x}\in S_i} d\mathbf{x}\mu_i(\mathbf{x})}
\end{aligned} \tag{1.15}$$

The decomposition of this transition matrix gives eigenvectors, which are an approximation of eigenfuctions outlined in Equation 1.12.The integration in Equation 1.15 run over sets $S_{i,\dots,n}$ and therefore for each integral, only information on dynamics local to the current set is required[108]. This property will be leveraged later in section 1.5.

It is important to note that the use of a Markov model ($\mathbf{T}(\tau)$) is only an approximate description of system dynamics and carries a systematic discretization

error. Instead of continuous transitions in the state space, the jump between some locations within sets is itself not Markovian. The projection onto the discrete space that transforms a continuous function $f(\mathbf{x})$ into a discrete function $\hat{f}(\mathbf{x})$ is denoted as:

$$\hat{f}(\mathbf{x}) = Qf(\mathbf{x}) = \sum_{i=1}^{n} a_i \chi_i(\mathbf{x}) \tag{1.16}$$

where $\chi_i(\mathbf{x})$ is the membership function of $\mathbf{x}$ belonging to set i (a step function in case of a crisp partitioning of space), and a_i are coefficients derived from the probability of each state. The approximation error is then defined as:

$$\delta_f = \|f(\mathbf{x}) - \hat{f}(\mathbf{x})\|_{\mu,2} \tag{1.17}$$

where $\|\cdot\|_{\mu,2}$ is the stationary density μ weighted Euclidean norm. Applying the projection Q to the initial density to project it onto discrete space gives $Qp_0(\mathbf{x})$, and substituting that into Equation 1.17 gives the difference ϵ between the discretized and the original functions after time $k\tau$:

$$\epsilon(k) = \|Q[\mathcal{T}(\tau)]^k Qp_0(\mathbf{x}) - Q[\mathcal{T}(\tau)Q]^k Qp_0(\mathbf{x})\|_{\mu,2} \tag{1.18}$$

where $Q[\mathcal{T}(\tau)]^k Qp_0(\mathbf{x})$ is the change in the initial density by true dynamics after k steps. Assuming the maximum possible error removes the dependence on the initial probability:

$$E(k) := \|Q[\mathcal{T}(\tau)]^k - Q[\mathcal{T}(\tau)Q]^k Q\|_{\mu,2} \tag{1.19}$$

The coarser the discretization, the larger the difference between the true continuous dynamics and the modelled discrete dynamics. However, the scale of acceptable discretization also depends on the processes modelled, i.e. if the discretized space still represents the relevant slow dynamics well, the loss of information on the very fast processes can be considered insignificant[108].

1.4.3 Practical Considerations

The above sections described the mathematical basis and some relevant properties of Markov state models. Here we outline the practicalities of building MSMs on MD simulation data.

MD trajectories usually contain all atom coordinates with some timestep between each frame. Using each unique set of atom coordinates as a definition of state sets $S_{i,\ldots,n}$ is not feasible, as such high dimensionality would mean that each state does not get visited more than once during an MD simulation. Some function of atom coordinates, such as protein backbone torsional angles, root-mean-square-deviation (RMSD), etc. are usually employed to reduce dimensionality (Figure 1.4C). Choosing relevant features to represent the relevant protein dynamics in low-dimensional space is non-trivial and can require trial and error. Automatic feature selection tools are available[109], but these may require prohibitively large amounts of simulation data.

Further reduction can be obtained by using time-independent component analysis (tICA), which applies a linear transformation on these values, combining them into collective coordinates[110]. This is followed by defining the discrete space sets, where center clustering methods are often used. The common clustering methods require prior selection of the number of clusters, which should be chosen to partition the feature space sufficiently finely[111]. Each frame of the available MD trajectory is then assigned to a state (Figure 1.4C)[112].

Using this transformed MD trajectory data, MSMs can be built. For this step, selection of an appropriate lag time τ is essential. It has to be long enough to make the Markov model as Markovian as possible, but short enough to still contain multiple state transitions within the limitation of MD simulation duration. Given Equation 1.14, the implied timescales (eigenvalues) of a process should be independent of the lag time. Therefore, a way of determining an appropriate τ value is to plot the implied timescales of the slowest few processes against a range of lag time values and identify when they start to plateau[113]. The MSM is then constructed

by counting state transitions after the lag time τ, and computing the transition probabilities. The stationary distribution of the states (i.e. the probability of each state) is the first eigenvector of this transition matrix[112]. The models can also be validating by using the Chapman-Kolmogorov (CK) test. The transition matrix ($\mathbf{T}$) predicted by a model after k lag times τ should be the same as the prediction by an independent model with a lag time $k\tau$:

$$[\mathbf{T}(\tau)]^k \approx \mathbf{T}(k\tau) \tag{1.20}$$

While the state probabilities are already obtained at this step, the number of states used to build an MSM that describes protein dynamics well is too large to easily make conclusions on the relevant dynamics of the system modelled. As such, further coarse-graining of the states is often required. The main approach is Perron Cluster-Cluster Analysis (PCCA), which uses the signs of the MSM eigenvectors to identify states that readily interconvert among themselves, but not others, therefore clustering them into metastable macrostates[110, 114]. Relating these states to major protein conformations allows application of MSMs to study protein folding, ligand binding, and other major conformational changes[110].

1.5 Joint sMD/MSM Workflow

In this work we present a joint sMD/MSM approach for the rational design of allosteric modulators. As outlined in section 1.2.3, while there are many allosteric pocket prediction and ligand affinity computation methods, modelling the quantitative effect of an allosteric modulator on the protein conformational ensemble remains difficult. Therefore, we propose the use of enhanced sampling methods to sample a larger conformational space, followed by unbiased MD simulations to gather unbiased trajectory data to build Markov State Models that capture the effects of allosteric ligands with significantly reduced overall sampling time. The methodology is described in more detail in this section, while its validation on a

variety of drug targets is outlined in chapters 2-4 of this thesis. To facilitate the application of this worklfow, the AMMo (Allostery in Markov Models) tool has been made available on GitHub.

1.5.1 Enhanced Sampling

The first step in the AMMo workflow is the steered MD simulation (Figure 1.5a). The change in the protein dynamic ensemble that takes place upon binding of allosteric modulator occurs on timescales not routinely accessible via equilibrium MD simulations. Here we employ sMD simulations to steer the system from functionally "inactive" to "active" conformations, and *vice versa*. The use of sMD requires strict definition of these end state conformations in terms of CVs used. This allows to move the system along a specific allosteric modulation pathway, probing the subset of conformational space that is relevant to effects of the ligands of interest. The exact steering protocol can include some trial and error, as a force that achieves the desired conformational change on a reasonable timescale, without biasing the system too harshly, needs to be identified. This requires close inspection of the sMD simulations until a suitable protocol has been devised.

The CV space must include descriptions of the orthosteric site that capture protein activity (e.g. the catalytic residues being positioned in the correct conformation for an enzyme to carry out a reaction). This can be a variety of variables, such as atom distances, bond angles, or RMSD. Additionally, in order to avoid hysteresis of the allosteric network, it can also be included in the CV set for the sMD simulation. This helps avoid the possibility that the allosteric network residues will remain in the sMD starting conformation, therefore biasing the following equilibrium MD simulations towards that initial conformation. This does require the knowledge of the allosteric network of the protein of interest, which is not always available, and in such cases allosteric network prediction methods outlined in section 1.2.3 can be of use. It is also worth mentioning that while the sMD used here require reliable structures of the target protein in active and inactive conformations, this information is

often not easily obtainable. Undirected enhanced sampling, such as aMD (section 1.3.5), can also be used to hasten state transitions[115].

1.5.2 Seeded MD Simulations

Once the protein transitions between functionally active and inactive conformations have been modelled with sMD, the resulting trajectories are used to obtain a range of representative protein conformations along this pathway. They are then used as "seeds" for a swarm of seeded MD simulations, where each conformation serves as a new starting point (shown in Figure 1.5a). This approach has a few benefits over a smaller number of standard long equilibrium MD simulations. Firstly, the seeded MD simulations can be run in parallel, making use of HPC capabilities and obtaining a larger total sampling time using less real time. Secondly, the simulations themselves can provide useful information while remaining quite short by leveraging the MSM requirement of local equilibrium only, i.e. only the close state transitions will need to be sampled (Equation 1.15). Multiple shorter simulations can be easily combined to build a single model describing all the relevant protein dynamics. Lastly, while the snapshots sampled from the start and end of the sMD simulation will closely resemble the energetically stable active and inactive conformations, the ones closer to the middle are expected to describe unstable, intermediate conformations. Therefore it is expected to see more rapid transitions during seeded MD simulations starting from those trajectories, which provides more useful data for the MSM.

1.5.3 MSM Building

The goal of the sMD/MSM workflow is to compare modelled state probabilities with and without the ligand, to identify the effect it has on the protein conformational ensemble. Because of this, particular care has to be taken during MSM building. All of seeded MD simulation data is reduced to relevant features, reducing the trajectory dimensionality from all atom coordinates to low dimension time series.

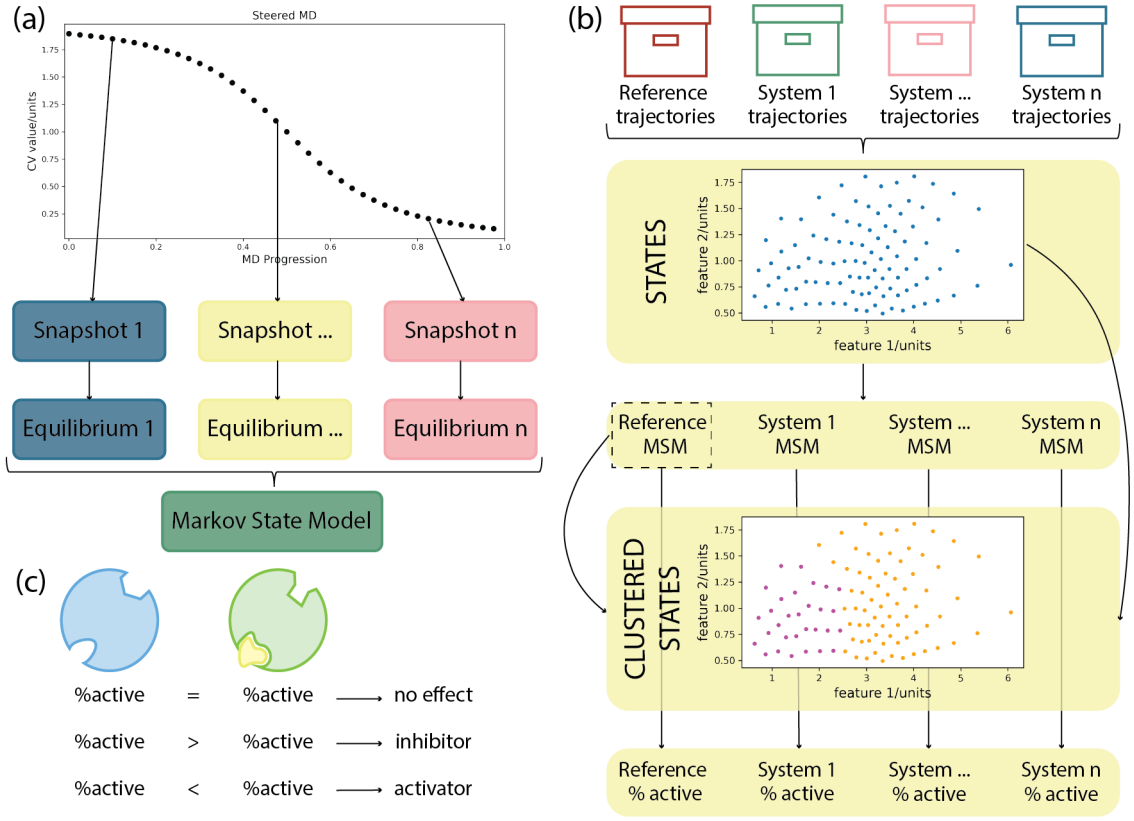

Figure 1.5: The sMD/MSM workflow to model effects of allosteric modulators on the protein conformational ensemble. (a) The initial step is to run steered MD simulations, exploring a wide conformational space. The trajectory is used to sample n number of snapshots, which serve as starting points for further equilibrium MD simulations. The data from these seeded MD simulations is combined into an MSM. (b) The MSM building protocol. The data from all systems that are being investigated is pooled together and clustered into states, in order to keep state assingment consistent across systems. These states are used to build individual MSMs that describe the dynamics of each system and give the stationary probabilities of each state. The states are then clustered furthered into metastable states, either manually or using PCCA of one of the MSMs (reference in this case). Based on knowledge of the target protein, the active state is identified, and its probability is computed. (c) Comparison of the active state probabilities. The unliganded protein is set as the reference point. If the active state probability is the same whether the ligand is bound or not, the ligand has on effect. If the probablity is lower or higher with the ligand present, this indicates that it is an allosteric inhibitor or activator, respectively.

All of the trajectories are pooled together and then clustered via *k-means* clustering (Figure 1.5b). This gives a set of microstates that are consistent for all MSMs built later. Each frame of the individual trajectories is then assigned to a state, further discretizing the state space and resulting in trajectories being represented as transitions between states. The MSMs are built for each system, to describe individual system dynamics and give the stationary probability values of each state. Since the states are defined by the exact same feature values for all MSMs considered, the probability values can be directly compared.

In order to better relate the MSM results to protein function, the microstates are further clustered into metastable states (Figure 1.5b). This can be done manually, by leveraging information available on the protein target, or by PCCA. Then, depending on the feature values for each state, the active state (or another state of interest) can be identified. The probability of the overall metastable state is simply the sum of the stationary probabilities of the microstates that belong to it. Additionally, in order to assess whether the sampling by MD simulation was sufficient to describe system dynamics, bootstrapping by resampling can be performed. It involves randomly resampling n number of trajectories for a system (i.e. some trajectories will be excluded, and some will be sampled more than once), and rebuilding the MSM with that data, whilst keeping the microstate and metastable state assignment consistent throughout. This gives an active (or any other) state probability distribution. If the sampling is sufficient, the distribution should be narrow, as changes in a few trajectories should not change the state probabilities much. A wide distribution can be an indication that more or longer seeded MD trajectories are required.

Finally, the macrostate probabilities can be compared to evaluate allosteric effects of the ligands of interest. Here the system with no potential allosteric modulator is used as a reference, to represent the unaffected protein conformation ensemble. If the state probabilities do not change significantly when the ligand is included, it does not have an effect on the protein dynamics and therefore is not an allosteric modulator. If the active state probability decreases (or the inactive state probability

increases), the MSMs indicate that the ligand is an allosteric inhibitor. On the other hand, if the active state probability increase, the ligand is an activator instead. The degree in probability change can also be used to rank the effectiveness of ligands as allosteric modulators.

Chapter 2

Computational Assessment of the Activity Determinants of Small Molecule Inhibitors of PTP1B

2.1 Introduction

The AMMo workflow (outlined in chapter 1.5) was first developed and tested on the Protein Tyrosine Phosphatase 1B (PTP1B). The charged and highly conserved nature of the PTP1B active site has made it challenging to develop orally bioavailable ligands that act as competitive inhibitors. Therefore allosteric inhibition of PTP1B enzymatic activity is an attractive drug design strategy[51]. The significant effort to modulate PTP1B allosterically has led to a large amount of NMR[116–118] and crystallographic[116, 119, 120] data on the allosteric network of this target, making it a suitable first target to validate the sMD/MSM approach.

2.1.1 Protein Phosphatases

Post-translational modifications (PTMs) are changes made to proteins after they have been synthesized. A chemical group may be added to one or more amino acids, or the protein itself broken up into smaller parts (proteolysis). By allowing

structural components not found in the natural 20 amino acids, PTMs are crucial to protein function, influencing activity, folding, localization, protein-protein interactions, and more. Some common PTMs are illustrated in Figure 2.1a-c. Many modifications have been discovered by chance, such as coincidental deletion of the amino acid carrying the modification. The modern improvements in protein expression and purification have allowed for the comparison of amino acid sequences of proteins and their final structures (such as discrepancy in mass as identified by mass spectrometry)[121, 122].

One of the most common PTMs is phosphorylation, which involves the addition of a phosphate group. It is modulated by the opposing activities of kinases, which add the phosphate group, and phosphatases, which remove it. The phosphoryl group drastically changes the properties of the amino acid and its local protein environment, adding a negative charge and increasing polarity, which creates a site for protein-protein interactions[121, 123]. Disregulation of phosphorylation is involved in diseases such as cancer[124, 125] and Alzheimer disease[126].

Human phosphatases can be classified into 3 main families, based on the dephosphorylation mechanism: PTPs (protein tyrosine phosphatases), PSPs (protein serine/threonine phosphatases), and HADs (haloacid dehalogenases). The PTP family proteins (which includes PTP1B) usually contain a CX_5R amino acid sequence in their active site. The Cys acts as the nucleophile, while the Arg coordinates the phosphate group. The PSP phosphatases employ divalent metal cations to coordinate the substrate, while HADs include a DxDx(V/T) motif, where an Asp acts as a nucleophile, with a magnesium cation cofactor[121, 123]. The different mechanisms are illustrated in Figure 2.1d-f.

2.1.2 The Structure and Reaction of PTP1B

Protein tyrosine phosphatase 1B, or PTP1B, is a member of the PTP phosphatase superfamily, and selectively catalyses the dephosphorylation of tyrosine residues. Over-expression of PTP1B is implicated in diseases such as cancer, type II diabetes,

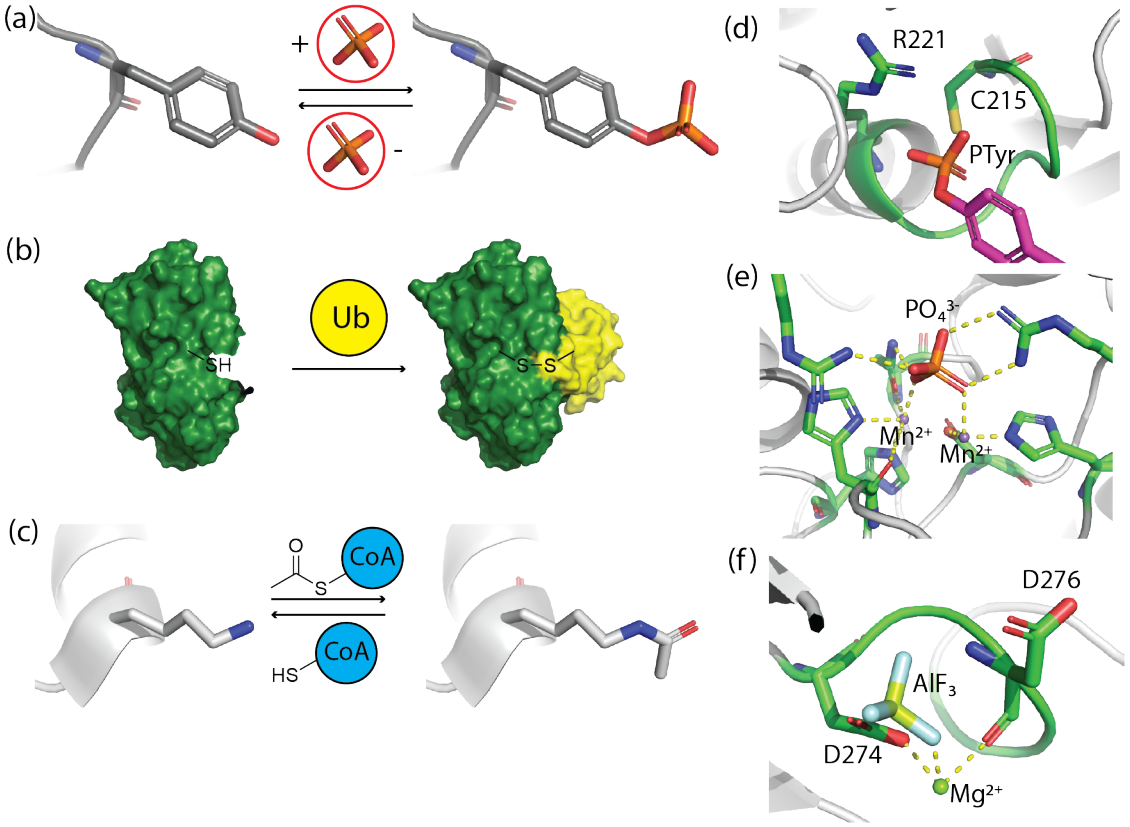

Figure 2.1: Illustration of some post-translational modifications and the 3 main phosphatase families. (a) Tyrosine phosphorylation (structure from PDB ID 1EEO). (b) Ubiquitination, with ubiquitin protein shown in yellow (structure from PDB ID 1FXT). (c) Acetylation of lysine, by coenzyme A shown in blue (structure from PDB ID 5ZS7). (d) The CX_5R motif of protein tyrosine phosphatase 1B (PTP1B), which belongs to the PTP family (structures from PDB IDs 1EEO and 1SUG). The catalytic residues are shown in green, and the P-Tyr residue is shown in magenta. (e) The two manganese ions coordinating a phosphate ion in protein phosphatase 1 (PP1), belonging to the PSP phosphatase family (structure from PDB ID 4MOV). Coordinating protein residues are shown in green. (f) The magnesium ion and the DXDX(V/T) motif of eyes absent homolog 2 (ED-EYA2) (structure from PDB ID 3HB0). The aluminium fluoride ion is shown as a surrogate to illustrate what the phosphate coordination would be.

and obesity[51]. As the structure of the full construct (residues 1-435) of the protein has not yet been resolved, the work described in this chapter focuses on the N-terminal catalytic domain (residues 1-300)[127].

The CX_5R motif common in PTP phosphatases makes up the P-loop of the active site, and includes catalytic residues C215 and R221. These residues are responsible for the binding of the phospho-tyrosine (P-Tyr) substrate. The other key component in the active site is the WPD loop, which contains the catalytic D181. In the *apo* PTP1B, the WPD loop mostly adopts the open (inactive) conformation, positioning D181 away from the active site. Upon substrate binding, the loop closes, moving D181 closer to the substrate phosphate ion and switching to the active conformation (Figure 2.2a)[118]. The aspartic acid acts as a proton donor in the first step of the mechanism, and as a base catalyst in the second. The P-Tyr substrate is coordinated by R221 and C215 on the P loop. Nucleophile C215 attacks the phosphate, releasing the dephosphorylated tyrosine residue. In the second step, a water molecule is used to recover the catalytic D181, and release the phosphate ion. The mechanism is depicted in Figure 2.2b[128]. Additionally, the WPD loop transitions between open and closed conformations on a μs timescale, slower than is usual for flexible protein loops[116].

In addition to the WPD and P loops in the active site, another key structural element of PTP1B is the α7 helix (residues 282-298). When the protein is in the active conformation, i.e. the WPD loop is closed, α7 docks between helices α3 and α6 (shown in Figure 2.2c). These interactions play a key role in the activity of PTP1B and stabilize the active conformation, as removing the helix decreases PTP1B activity by 40%. In the inactive conformation, i.e. when the WPD loop is open, the α7 helix is disordered and positioned away from the protein. The folding/unfolding of α7 also follows fast timescales, further suggesting that it is significant to protein activity and allosteric modulator[116]. This stabilizing role of an α7 helix is also seen in other protein phosphatases, such as the T-cell protein tyrosine phosphatase (TCPTP)[129].

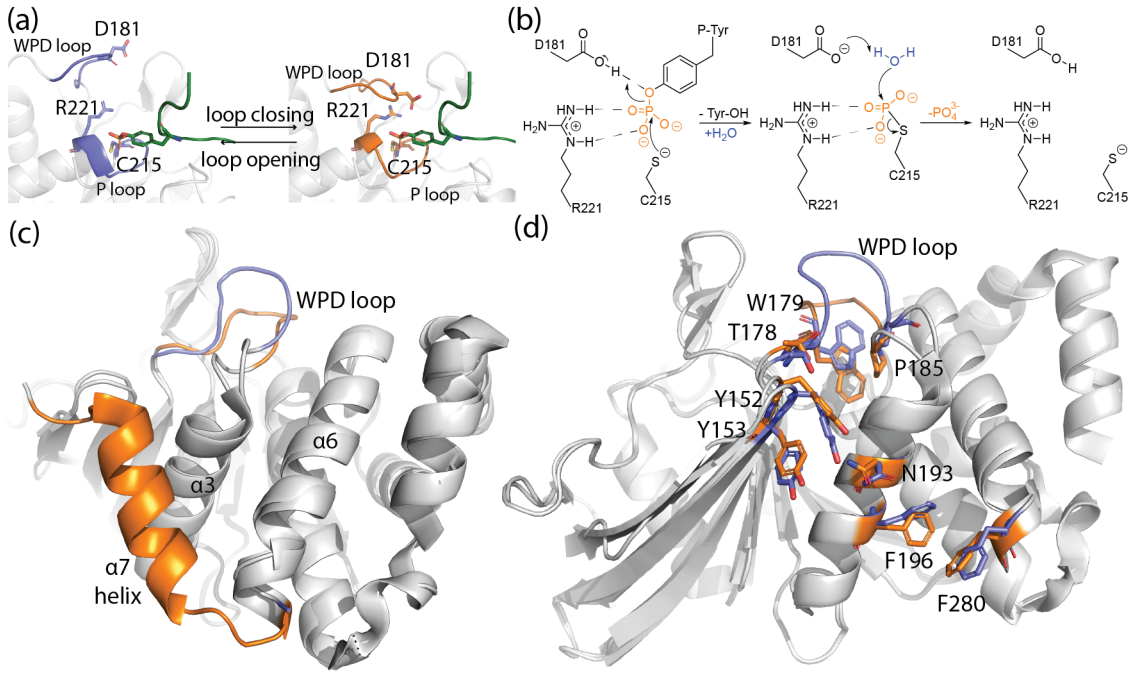

Figure 2.2: Key structural elements of PTP1B. (a) The open (blue, left) and closed (orange, right) conformations of the WPD loop. A model phosphotyrosine substrate peptide is shown in green. When the loop closes, D181 (WPD loop) is positioned close to the substrate, coordinated by residues C215 and R221 of the P loop. (b) The mechanism of dephosphorylation by PTP1B, adapted from Brandão *et al.*[128]. (c) The docked $\alpha 7$ helix in when the WPD loop is closed (orange, PDB ID 1SUG), and the disordered (and therefore unresolved) $\alpha 7$ when the loop is open (blue, PDB ID 2HNP). (d) The key allosteric residues connecting the $\alpha 7$ helix site to the active site.

Due to the challenges in targeting the PTP1B active site (more in section 2.1.3 below), the allosteric network of PTP1B has been the subject of extensive study. The α7 helix connects to the active site through residues F280, F196, N193, Y153, Y152, and T178, as seen in mutation studies, NMR, and multitemperature X-Ray crystallography[116–119, 130]. For instance, F280 and F196 show π-stacking of the side chain phenyl rings across the gap between helices α3 and α6 in the active conformation of PTP1B. The α7 helix pushes the F280 sidechain into the pocket when it is folded and docked along the protein. Similarly, Y152 shows both the "up" and "down" rotamers of the phenol sidechain in the inactive conformation, but in the active conformation only the "down" rotamer is stabilized[119]. Additionally, P185 has been identified as crucial to the WPD loop closure due to its interactions with W179[116]. These residues are illustrated in Figure 2.2d.

2.1.3 Small Molecule Inhibitors of PTP1B

Only a few PTP1B inhibitors targeting the active site have reached the clinical trial stage, primarily for the treatment of type II diabetes[127]. However, most have now been discontinued due to poor selectivity and bioavailability, which is influenced by the nature of the PTP1B active site. The catalytic site of PTP phosphatases is very highly conserved, and therefore P-Tyr mimetic compounds often exhibit off-target effects. Ertiprotafib, a monocarboxylic acid mimetic of P-Tyr, was removed from clinical trials owing to its interaction with other proteins, such as IκB kinase β (IKK-β)[131]. In addition to being highly conserved, the R221 in the active site of PTP1B carries a positive charge, making it beneficial for inhibitors to carry a negative charge in turn. However, charge or high polarity are associated with poor permeability and bioavailability[127]. Nonetheless, efforts to target the active site are continuing, with a promising PTP1B/PTPN2 inhibitor ABBV-CLS-484 in clinical trials[132].

As the nature of the PTP1B active site raises additional challenges for developing orthosteric inhibitors, leveraging allosteric modulation is a viable alternative

strategy. The first set of allosteric inhibitors was reported as early as 2004 (Figure 2.3a-b). It also aided in deciphering the importance of the $\alpha 7$ helix, as these inhibitors bind in the pocket between helices $\alpha 3$ and $\alpha 6$, blocking the docking of $\alpha 7$ there[130]. Due to the benzbromarone (BB) core of these molecules, the site has been referred to as the BB site. High throughput frgament screening has also led to identification of a another distinct allosteric site, the 197 site, so called because of K197 extending into it. A number of fragments binding at the 197 site, as well as a covalent inhibitor (bound to a K197C mutant), with EC_{50} of 7.8 $\pm$ 1.1 μM, have been reported. The covalent inhibitor and a represenative fragment are shown in Figure 2.3a-b[119]. Another covalent inhibitor, selectively tethering to C121 of yet another distinct allosteric site is shown in Figure[133]. PTP1B contains a variety of allosteric sites, which show better druggability and have less conserved residues than its catalytic site, making allosteric modulation a viable strategy to finally successfully put a PTP1B inhibitor on the drug market[119, 127].

In this chapter, three experimentally characterised inhibitors (**1**-**3**) and a fragment binder (**4**) with unknown functional effect (Figure 2.3a-b) were used to validate the joint sMD/MSM methodology as outlined in chapter 1.5. Our protocol successfully identified **1** as a potent allosteric inhibitor[130]. **2**, a fragment obtained by deconstructing inhibitor **1** that only shows very weak activity experimentally was classified as inactive by our protocol. **3**, a covalently bound fragment that weakly inhibit PTP1B shows activity intermediate between **1** and **2** in our protocol[119]. Fragment binder **4** is predicted functionally inactive by our approach. Through comparative analysis of the computed protein conformational ensembles we identify specific protein conformational states that could be used as blueprints for virtual screens of novel PTP1B allosteric modulators. Our efforts illustrate how our joint sMD/MSM protocol could be used to prioritise fragment hits for hit-to-lead chemistry efforts, and to plan virtual screening campaigns. Additionally, we compare the inclusion of the allosteric network (Figure 2.3c) in the sMD collective variable set to only steering the active site, and illustrate the significance of relevant comforma-

tional space sampling.

2.2 Methods

2.2.1 System Preparation

All systems with the WPD loop closed used protein coordinates from PDB ID 1SUG. All open loop protein conformations were from PDB ID 2HNP, with W179 rotated to match the rotamer in 1SUG using Flare v5[134]. Both protein conformations included residues 1-282, truncating the α7 helix. Peptide substrate was taken from PDB ID 1EEO. Missing PTP1B residues and all peptide ACE/NME caps were added using Flare.

In all cases, E97 was modelled as GLH and H214 as HID, due to predicted pK_a values of 8.59 and 3.71 respectivelly by propka3[135], using PDB ID 2HNP. Additionally D181 was modelled as ASH, and C215 as CYM, to match the proton-donor role of D181 and the coordination of substrate by C215. All system preparation was done through BioSimSpace[136], except in the case of tethered ligand **3**. The ff14SB force field was used for protein residues, with additional phosphate parameters from Case *et al.*[137]. The ligands were parameterised using GAFF2 and the AM1-BCC charge method. For all systems, the same PDB IDs were used for the open and closed conformations of PTP1B, 2HNP and 1SUG respectively. Where present, the same PDB ID was used for the peptide substrate, 1EEO. Ligands were inserted into the system by aligning the PDB entry of PTP1B containing the ligand to the relevant protein conformation, and copying ligand coordinates. This approach was taken to keep the starting coordinates of the protein consitent throughout the various simulations, only varying the ligand. System source PDB IDs and ligand charges are outlined in Table 2.1. In all cases, TIP3P water was used to explicitly solvate the system as a cuboid box, with 10 Å distance. Na^+ ions were added to neutralize the system, and Na^+ and Cl^- ions were added to achieve 150 mM NaCl concentration. All systems were minimized and equilibrated using GROMACS version 2020.2[138].

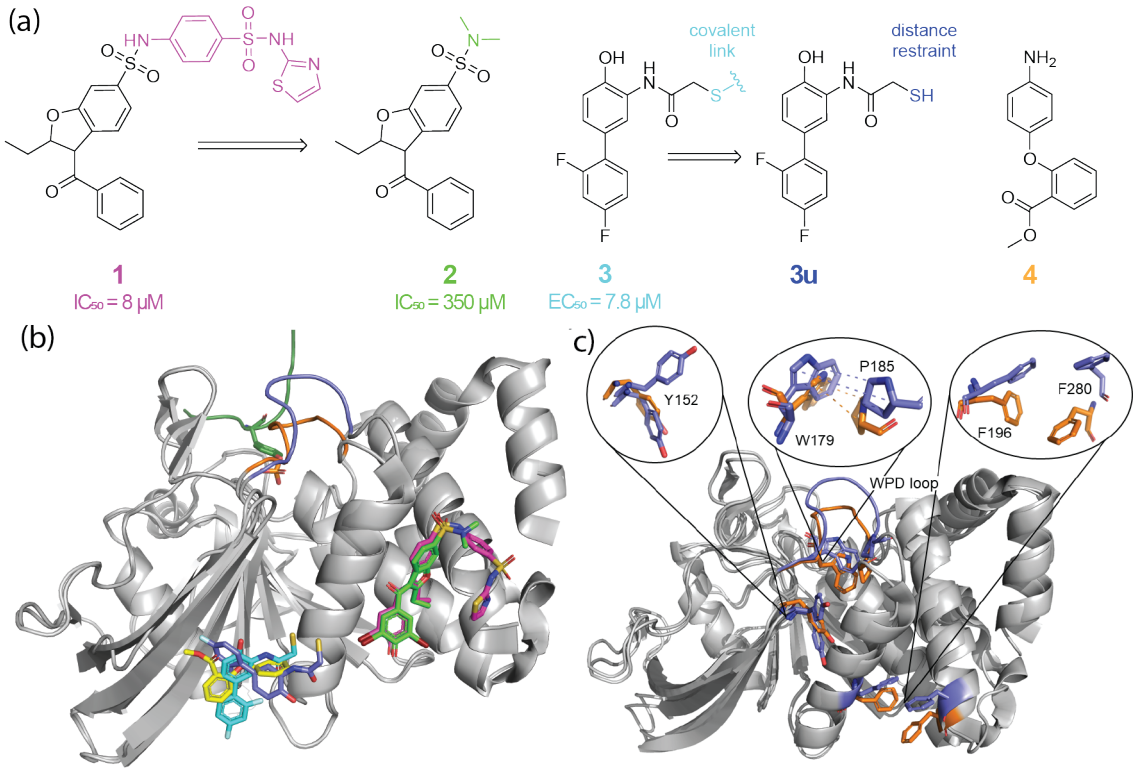

Figure 2.3: The compounds modelled in this chapter, and the allosteric network residues selected. (a) The structures of BB site binders **1** and **2**[130], 197 site covalent inhibitor **3**[119] (as well an untethered **3u**), and a fragment binder at 197 site of unknown effect, **4**. (b) The binding poses of compounds shown in (a): **1** in magenta, **2** in green, **3** in cyan, and **4** in yellow. (c) The representative allosteric network residues and their different conformations in active (orange) and inactive (blue) PTP1B conformations.

Minimisation was carried out over 7500 steepest descent steps. Systems were heated from 0 K to 300 K over 100 ps. Equilibration was performed in the NPT ensemble for an additional 250 ps.

2.2.2 Tethered Ligand Parameter Setup

To prepare the parameters for covalently linked **3**, the ligand and C197 (atoms SG, CA, CB, HB1-3) residues were obtained from PDB ID 6B95. CA in C197 was changed to a hydrogen, and the CYS was renamed to CYX. The PDB file was converted to mol2 format using antechamber[139] and the AM1-BCC charge method, with a neutral charge. The atom types in the mol2 file were set as follows: SB was set to S, CB to CT and HB1-3 to ha. The force field modification file was generated using parmchk2 and the parameter file was generated using tLEaP. The information corresponding to the Cys197 residue was removed from the parameter file, also modifying connectivity and atom number entries.

2.2.3 Long Equilibrium MD simulations

To confirm the agreement between the computational PTP1B model and experimental data, 1 μs unbiased MD simulations were run with Amber20[139] via BioSimSpace, at a temperature of 300 K and 1 atm pressure in the NPT ensemble. The collective variables (Table 2.2) were computed using cpptraj.

2.2.4 Steered MD

Steered molecular dynamics were run with Amber20 and PLUMED v2.6.1[140] via BioSimSpace. Two steering protocols were evaluated: steering the WPD loop only, and additionally including the allosteric network (Figure 2.3c). The collective variables used are outlined in Table 2.2. In all cases the first 4 ps were used to apply the force, maintaining the CVs at original values. Open to closed loop conformation sMD was carried out over 150 ns with a 3500 kJ mol^{-1} force constant, while closed to open sMD was carried out over 100 ns with a 2500 kJ mol^{-1} force constant.

Table 2.1: The PDB IDs and ligand information for all systems modelled in this chapter. The protein structures for closed and open WPD loop conformations in all cases were taken from 1SUG and 2HNP respectively. The PDB ID column for peptide and ligands indicates to source of the peptide/ligand structure and binding pose.

system	**protein conformation**	**protein PDB**	**peptide PDB**	**ligand PDB**	**ligand PDB**
Apo	open	2HNP	None	None	None
	closed	1SUG	None	None	None
Reference	open	2HNP	1EEO	None	None
	closed	1SUG	1EEO	None	None
compound 1	open	2HNP	1EEO	1T4J	-1
	closed	1SUG	1EEO	1T4J	-1
compound 2	open	2HNP	1EEO	1T48	-1
	closed	1SUG	1EEO	1T48	-1
compound 3	open	2HNP	1EEO	6B95	0
	closed	1SUG	1EEO	6B95	0
compound 4	open	2HNP	1EEO	5QDL	0
	closed	1SUG	1EEO	5QDL	0

Table 2.2: The steering collective variable parameters for steering only the WPD loop, and steering the allosteric network alongisde the WPD loop. Note that the residues in the CV definition column are offset by +1 to account for the ACE cap added. The angles are in radians and distances in nanometers, in line with PLUMED units.

CV	**definition**	**active to inactive target value**	**inactive to active target value**
WPD loop RMSD	residues 179-185 all heavy atoms RMSD to target conformation	0.0	0.0
Y152 χ1 angle	dihedral angle residue 153 atoms N, Cα, Cβ, and $C\gamma$	-1.047	1.047
P185 stacking to W179	absolute difference between the P186(Cδ)-W180(Cϵ2) and P186(Cα)-W180(Cδ1) distances	0.3	0.0
F196 stacking to F280	Cγ distance between residues 197 and 281	0.7	0.45

The target values for the allosteric residues were taken from 1 μs equilibrium MD simulations (Figure 2.6). All simulations were run at 300 K and 1 atm.

2.2.5 Seeded MD

100 snapshots were extracted from each sMD trajectory (200 total per model), equally sampling the WPD loop RMSD range, using cpptraj v4.25.6 (AmberTools20). The systems were resolvated and re-equilibrated as outlined above and 50 ns equilibrium MD simulations were carried out with Amber20 via BioSimSpace, saving snapshots every 10 ps (5000 frames per simulation). Total sampling time of trajectories used for a single MSM was 10 μs.

2.2.6 Ligand Restraints

In the cases when ligands were restrained for sMD or seeded MD simulations, flat bottomed distance restraints were used. The exact parameters are given in Table 2.3.

2.2.7 Markov State Modelling

The seeded MD trajectories were featurised using cpptraj[141]. The features used were WPD loop (residues 178-184) backbone RMSD to PTP1B with WPD loop closed (PDB ID 1SUG), and P loop (residues 214-219) RMSD. All MSM model buiding was done using PyEMMA version 2.5.7[142]. All system data was pooled together, and k-means clustering (100 cluster centers) was used to define microstates. Implied timescales (ITS) were calculated using a range of lag times between 1 and 3000 steps (10 ps to 30 ns) (Figure 2.4), to explore the appropriate MSM lag time. Based on the ITS, MSMs were generated with a lag time of 2000 steps (20 ns) in all cases. Perron Cluster-Cluster Analysis (PCCA) of the reference system was performed to define two macrostates. The metastable state with lower RMSD values corresponds to the active state, as lower RMSD values correspond to higher similarity to the crystal structure of PTP1B with the loop closed. The clusters not

Ligand	**Restraint atoms**	**Restraint bounds** /$\AA$	**Force constants** /**kcal mol**$^{-1}$
2r	2(N01) and E277(Cδ)	2.5, 3.0, 4.0, 4.5	0, 300
2r	2(O19) and N194(Cγ)	2.5, 3.0, 4.0, 4.5	0, 300
3u	3(S19) and C198(S)	2.5, 3.0, 4.0, 4.5	0, 50
4	4(N18) and C198(S)	2.5, 3.0, 4.0, 4.5	0, 50

Table 2.3: Flat bottom restraint parameters. The restraint bounds indicate the r_1-r_4 distance definitions, while the force constants k_2 and k_3 are used to compute the restraint energy at different points of the flat bottomed potential. When the restrained distance R is at the values between r_2 and r_3, the restraint energy is 0. When it is between r_1 and r_2, a parabolic restraint $k_2(R-r_2)^2$ is applied. Similarly, when R is between r_3 and r_4, a parabolic restraint $k_3(R-r_3)^2$ is applied. When the restrained distance is less than r_1 or greater than r_4, the restraint increases linearly with the slope of the adjacent parabola restraint[139].

sampled by the reference system were assigned to the inactive state. When clusters were not sampled by a particular system, they were assigned 0% stationary probability manually. Chapman-Kolmogorov (CK) tests for each MSM are available in Figure 2.5.

Bootstrapping by resampling was carried out for 100 iterations for each system. 200 random trajectories were selected, and the MSM was built using the same 100 cluster centers, and the same active state assignment as above. The stationary probabilities of clusters belonging to the active state were summed to give a single active state probability each time.

2.2.8 Conformational Analysis

To recreate a statistically weighted ensemble of each system, 10,000 frames were sampled out of all trajectory data for the system, using the MSM stationary probabilities as weights. These trajectories were used to compute the residue behaviour shown in the results below. To obtain the reference active and inactive conformation ensembles, the same was applied to the reference system. However, instead of the stationary probabilities, metastable distributions for the active and inactive states were used.

2.3 Results

2.3.1 Validation of the sMD/MSM Protocol on Substrate Simulations

In order to confirm the expected dynamics of the PTP1B model, 1 μs MD simulations with the peptide substrate were carried out for both the active (WPD loop closed) and inactive (WPD loop open) conformations. The values of CVs later used for steering were computed and are shown in Figure 2.6. The WPD loop remains in either open or closed conformation during the whole simulation. Additionally, in

line with experiment[119], Y152 does show only one, "down", rotamer during the simulation with the loop closed, but both "up" and "down" when the loop is open. P185 maintains its *pi*-stacking to W179, as expressed by the absolute difference between the P185(Cδ)-W179(Cϵ) and P185(Cα)-W179(Cδ1) distances. If the two distances are similar in value (i.e. low absolute difference), they are parallel and the two residues face each other. If one is larger than the other (larger difference), P185 has shifted sideways, breaking the stacking (Figure 2.3c). Finally, the F196-F280 stacking is also maintained during the simulation with the WPD loop closed, and is absent when the loop is open. The mean distribution values observed in these simulations were used as targets for later sMD simulations, and are outlined in Table 2.2.

Systems including *apo* PTP1B, PTP1B with peptide substrate (reference), and PTP1B with substrate and each of the compounds **1**-**4** (Figure 2.3), were put through the sMD/MSM workflow as follows. Steered MD simulations were performed, steering the WPD loop and the allosteric network residues, outlined in Figure 2.3c and Table 2.2. An example of sMD results is shown in Figure 2.7. From each trajectory, 100 snapshots evenly sampling the observed WPD loop conformations were saved and used as starting points for follow-up 50 ns seeded MD simulations (200 trajectories, 10 μs total sampling time per model). Effects of prolonging the seeded MD simulations to 100 ns are shown in Figure 2.8. Each trajectory was reduced to two features: WPD loop (residues 178-184) backbone root mean square distance (RMSD) to closed conformation, and the P loop (residues 214-218) backbone RMSD to closed conformation (Figure 2.9a). An example of the featurized data and the resulting clusters is shown in Figure 2.9b.

MSMs were built for all of the systems with a lag time of 20 ns. From state transition probabilities, the equilibrium probabilities of each state were computed. The states were clustered into two macrostates using PCCA, referred to as "active" and "inactive" based on the RMSD of the loops[108, 114] (Figure 2.9b). The WPD loop RMSD cutoff values observed for the metastable states correspond to the RMSD

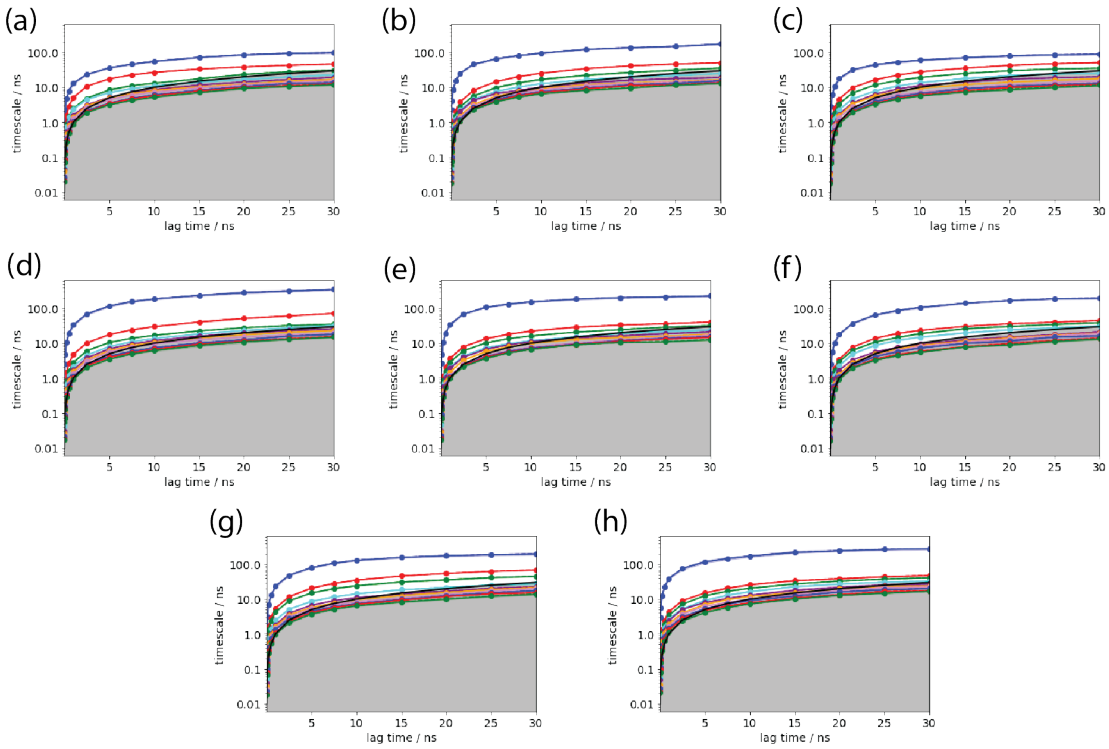

Figure 2.4: Implied timescales (ITS) of each MSM: **(a)** apo **(b)** reference **(c)** 1 **(d)** 2 **(e)** 2r **(f)** 3 **(g)** 3u **(h)** 4.

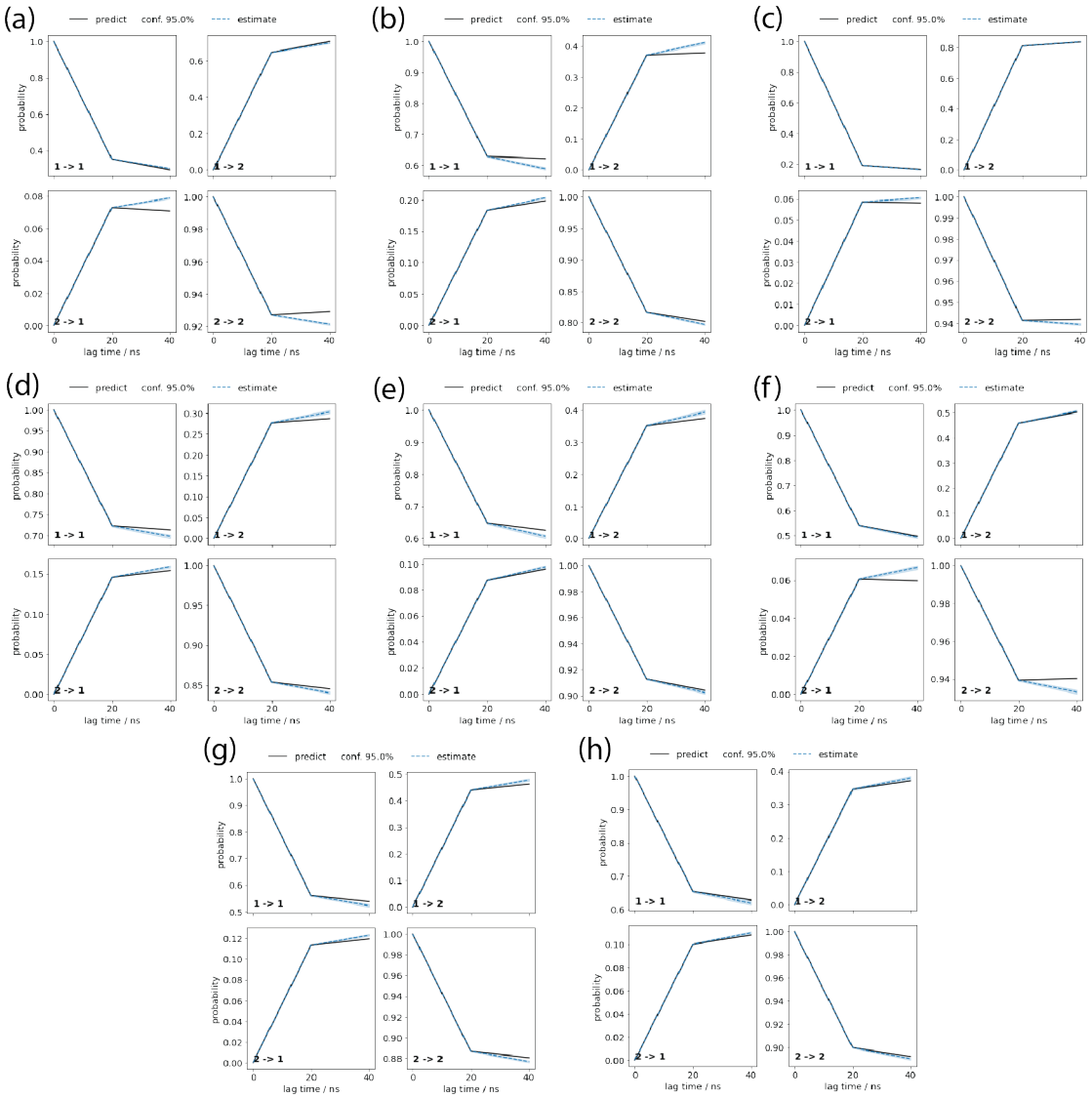

Figure 2.5: Chapman-Kolmogorov test of each MSM: **(a)** apo **(b)** reference **(c)** 1 **(d)** 2 **(e)** 2r **(f)** 3 **(g)** 3u **(h)** 4

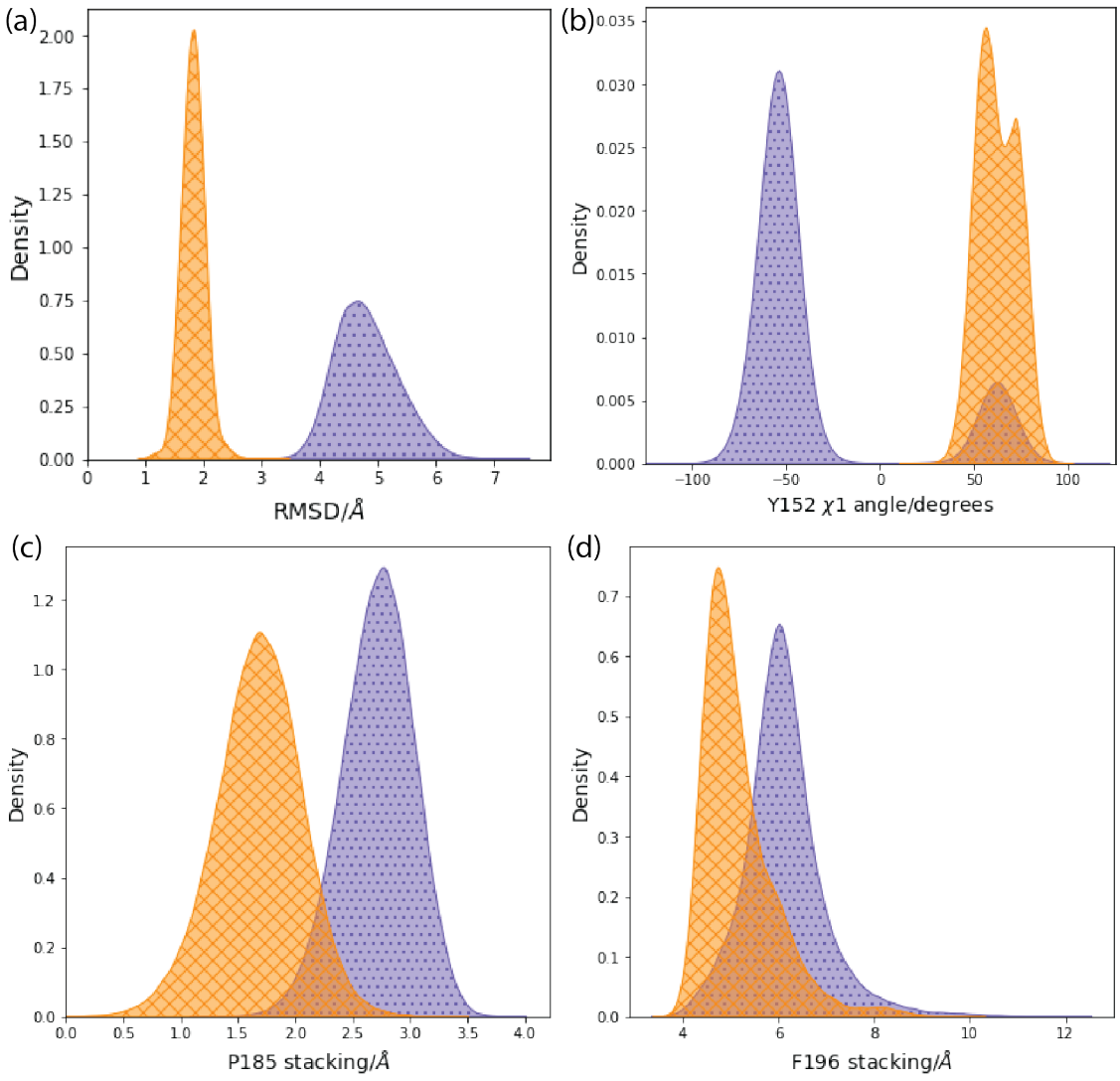

Figure 2.6: The collective variables used for sMD of PTP1B, during 1 μs equilibrium MD simulations of the reference PTP1B system (with the peptide substrate) when the WPD loop was open (blue, dots) and closed (orange, crosses). **a** WPD loop heavy atom RMSD to PTP1B with the closed loop conformation. **b** Y152 *chi*1 angle. **c** P185 stacking to W179 distance, which is defined as the absolute difference between the P185(Cδ)-W179(Cϵ) and P185(Cα)-W179(Cδ1) distances. **d** F196 stacking to F280 stacking, which is defined as the F196(Cγ)-F280(Cγ) distance.

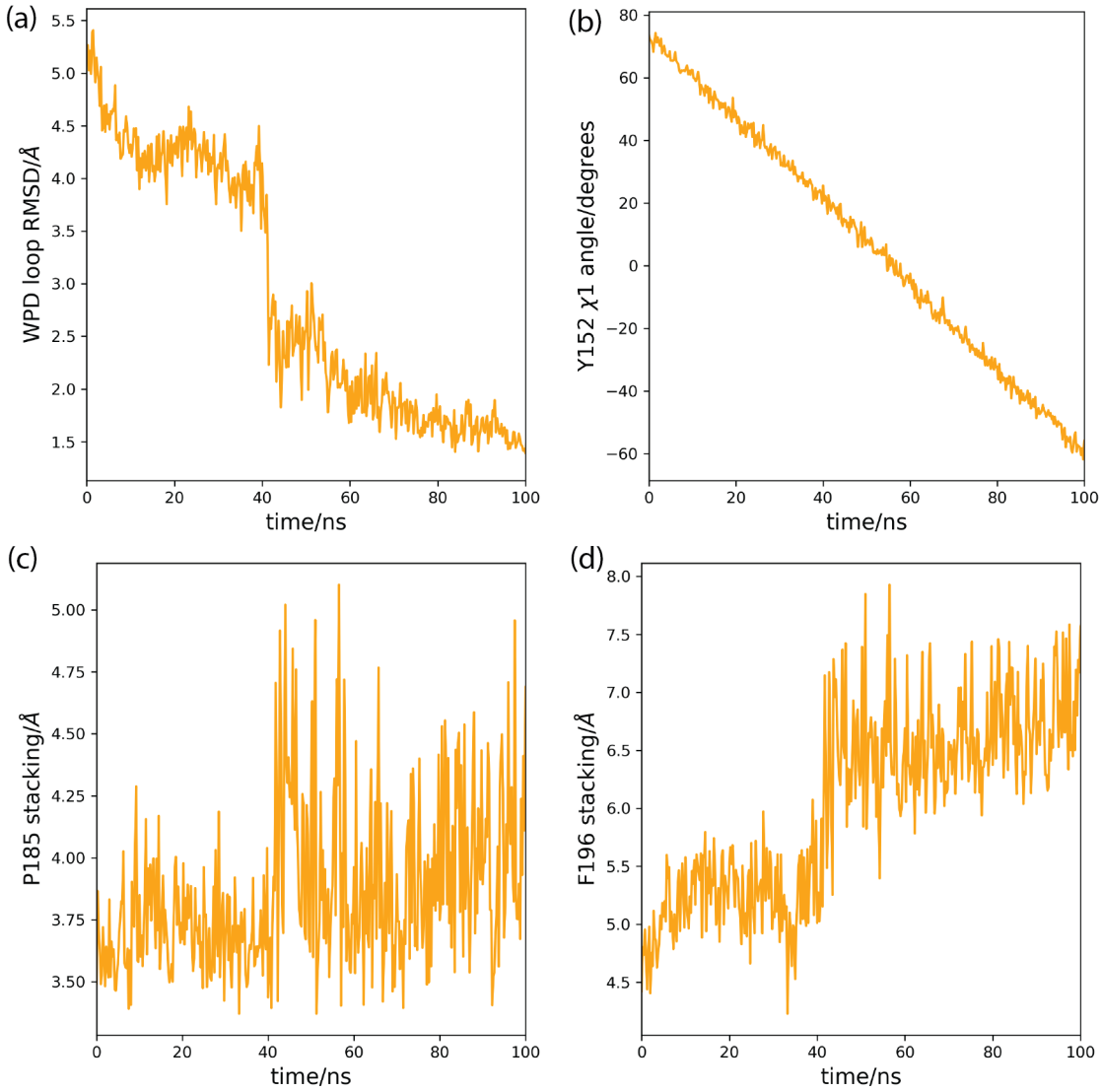

Figure 2.7: Example results of steering PTP1B with the peptide substrate (reference system) from closed (active) to open (inactive) conformation. (a) WPD loop RMSD (b) Y152 χ1 angle (c) P185 stacking to W179 (d) F196 stacking to F280.

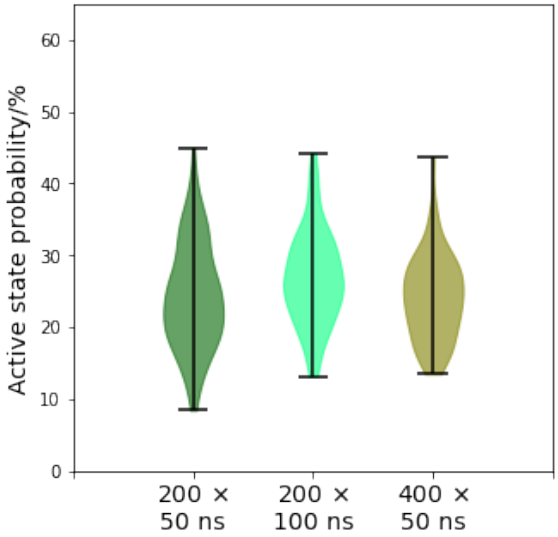

Figure 2.8: Bootstrapped active state probabilities of PTP1B with compound 2r when seeded MD duration was increased from 50 ns to 100 ns (20 μs total sampling time), and the number of seeded MD trajectories was increased from 200 to 400 (20 μs total sampling time), compared to original sampling.

value distribution during equilibrium MD simulations of PTP1B with WPD loop closed and open (Figure 2.6). Active state probabilities obtained from using PCCA assignments based on other system MSMs give very similar results and are shown in Appendix 2.10. The procedure was repeated a hundred times, using the initial micro- and macrostate definitions, to generate probability distributions for observing active states for each system (Figure 2.9c).

The major conformation for *apo* PTP1B is the inactive conformation, in agreement with experimental results that suggest a low fraction of active states (2.5%)[118] ("apo" in Figure 2.9c). Upon substrate binding there is a significant increase in active conformation probability, in agreement with experimental data[118] ("reference" in Figure 2.9c). However while NMR measurements suggest the active conformation dominates PTP1B's conformational ensemble when the enzyme is bound to a substrate (87% population[118]), the MSM indicates the active state is only formed 25% of the time. Experimental data suggests that activation of PTP1B by closure of the WPD loop is coupled with a disorder-to-order transition of helix $\alpha 7$. Owing to the difficulties in reliably simulating such large-scale conformational changes the PTP1B model used in the current study is a truncated variant that lacks helix $\alpha 7$. Experimental evidence shows that a mutant PTP1B-$\Delta 7$ lacking helix $\alpha 7$ is about 40% less active than wild-type[116]. Thus the incomplete activation of PTP1B in presence of a model peptide substrate is fully consistent with experimental observations. The goal of the present protocol is to classify ligands as allosteric effectors by comparison of relative shifts in active state populations, for which trends (relative to the reference system) are sufficient.

2.3.2 Compound 1 is Modelled as an Inhibitor, While the Deconstructed Analog 2 Shows no Inhibition

The active state probability distribution for compound **1** is significantly shifted towards lower values than that observed for the reference system (Figure 2.9c, "**1**"), strongly suggesting that compound **1** behaves as an allosteric inhibitor. This be-

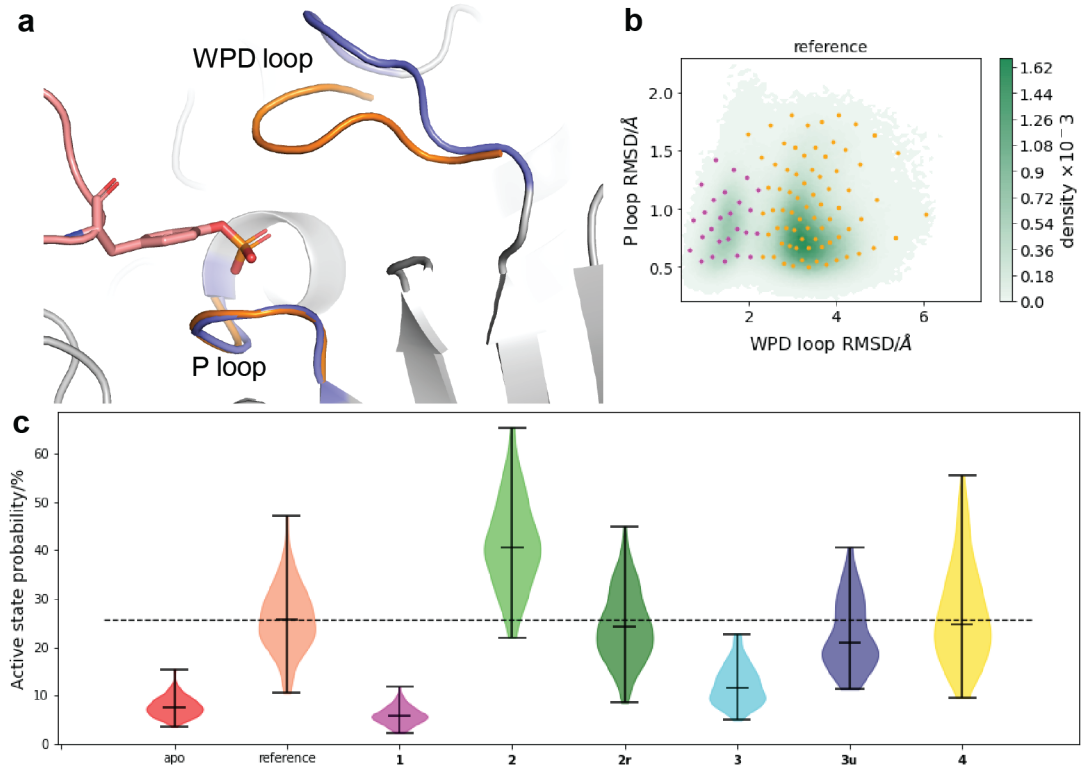

Figure 2.9: Markov State Model features and results. **a** The features used to reduce data dimensionality: backbone RMSD to closed WPD loop conformation, and D181(Cγ)-C215(S) distance. **b** An example of the reference system data, with the microstate clusters overlayed. Each cluster is assigned to a metastable macrostate via PCCA (magenta - active, orange - inactive). Since all of the data was clustered together for consistency, some of the microstates for a given system are not populated. **c** Violin plots of active state probability distributions for each system, after 100 iterations of bootstrapping by resampling. The middle horizontal bar of each violin plot indicates the median active state probability, while the upper and lower bars indicate the maximum and minimum values. The dashed line marks the median active state probability of the reference system. The x axis ticks indicate the PTP1B system composition: *apo* PTP1B (apo), PTP1B with a subtrate peptide (reference), and PTP1B with compounds **1**-**4**, in addition to the substrate peptide (**1**-**4**). **2r** stands for restrained compound **2**, while **3u** stands for untethered compound **3**, i.e. the covalent S-S bond is replaced with a distance restraint.

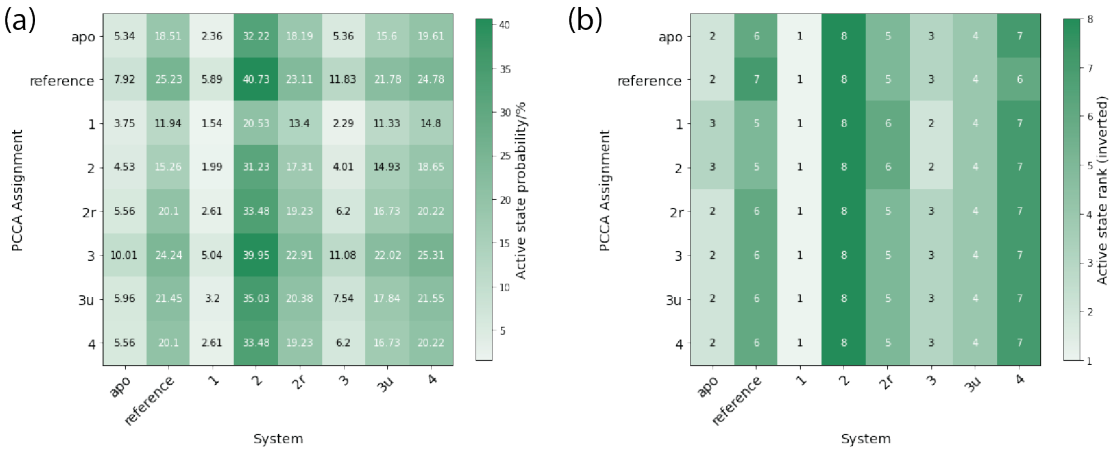

Figure 2.10: Active state probabilities **(a)** and inverse active state ranking (1 - lowest, 7 - highest) **(b)** when using PCCA assignments from each MSM to assign the active state.

haviour is consistent with an IC_{50} value of **1** ca. 8 μM[130] reported for compound **1**.

Compound **2** is a smaller analogue obtained by truncation of the aryl-sulfonamide moiety of **1** (Figure 2.3a). **2** is reported in literature as a very weak inhibitor (IC_{50} ca. 350 μM)[130]. The initial results (Figure 2.9c "**2**") did not show a decrease in active conformation probability, but rather a broad up-shifted distribution. Inspection of the MD trajectories used to build the MSM showed that **2** was only weakly bound and had a tendency to escape its binding site on a timescale of several nanoseconds, casting doubts on the reliability of the results obtained by the protocol. A second MSM was built, this time restraining intermolecular distance between **2** and N193 and E276 with weak flat-bottom biasing potentials (see Figure 2.11b and Methods). These distance restraints were selected to enforce **2** to adopt a binding pose consistent with that observed with **1** throughout the MD simulations. The resulting active state probability distribution (Figure 2.9c, "**2r**") was very similar to the reference system, suggesting a lack of functional effect. Extending simulation time or number of simulations reduces model error, but does not suggest inhibitory effect for compound **2r** (Figure 2.8).

The lack of inhibition by compound **2**, even when restrained to the binding pocket, may relate to its reduced interactions with F280, which has been suggested to be part of the allosteric network of PTP1B[117, 118]. Compound **1** wraps around the side chain and π stacks via its thiazole moiety, forcing F280 to adopt primarily an "up" rotamer (Figure 2.11a and Figure 2.11d, magenta χ 1 angle ca. -60 deg.). The "up" rotamer of F280 is observed in the inactive sub-ensemble of PTP1B "reference" more than it is in the active (Figure 2.11c). Compound **2** lacks a arylsulfonamide-thiazole moiety to wrap around F280, and consequently F280 adopts multiple rotameric states during the simulations (Figure 2.11b and Figure 2.11d green). The most populated "down" rotamer of F280 observed during simulations of **2r** is similar to the major rotamer observed in the active sub-ensemble of PTP1B "reference" simulations (Figure 2.11d green χ 1 angle ca. -180 deg. and Figure 2.11c, orange).

Such differences in behaviour in F280 dynamics are not apparent in crystal structures of **1** and **2** (PDB IDs 1T4J and 1T48) where F280 adopts a "down" rotamer exclusively (Figure 2.11d, dashed lines).

2.3.3 Covalent Tethering of Compound 3 Contributes to Allosteric Effect

Large-scale automated crystallography screening of fragments carried out by Keedy *et al.* has resulted in a tethered fragment **3** at a site distinct from that occupied by compounds **1**-**2**. The fragment is covalently linked to a K197C mutant and shows 60% maximum inhibition[119]. A ligand binding at the K197 site may interact with residues part of the allosteric network, such as Y152 or N193[117, 118]. Therefore, the joint sMD/MSM protocol was applied to compound **3**. The model produced a down-shift in active conformation probability distribution with respect to the reference system (Figure 2.9c "**3**"), suggesting an inhibitory effect intermediate between **1** and **2**.

In order to further assess the sensitivity of the sMD/MSM workflow to the effect of fragments, a model for untethered **3**, **3u** (with the K197C PTP1B mutant) was built. The covalent linkage was replaced by a flat-bottomed non-directional distance restraint to K197C (see Methods). The sMD/MSM produced a broad active state probability distribution with a median only slighly shifted down with respect to the reference system (Figure 2.9c "**3u**"). Comparison of the computed MSM ensembles for **3** and **3u** shows that **3** mainly adopts a "upright" binding pose owing to the covalent tether (Figure 2.12a) that resembles the crystallographic pose observed for this fragment. This pose enables the fragment phenol moeity to engage in hydrogen bonding interactions with K150, a suggested allosteric residue[118]. By contrast untethered fragment **3u** is more mobile and adopts predominantly a "sideways" pose (Figure 2.12b). This causes the phenol group to interact with E200, which has not been flagged as a residue of interest to the allosteric network [117–119]. The "upright" pose can still be detected albeit less frequently. These observations

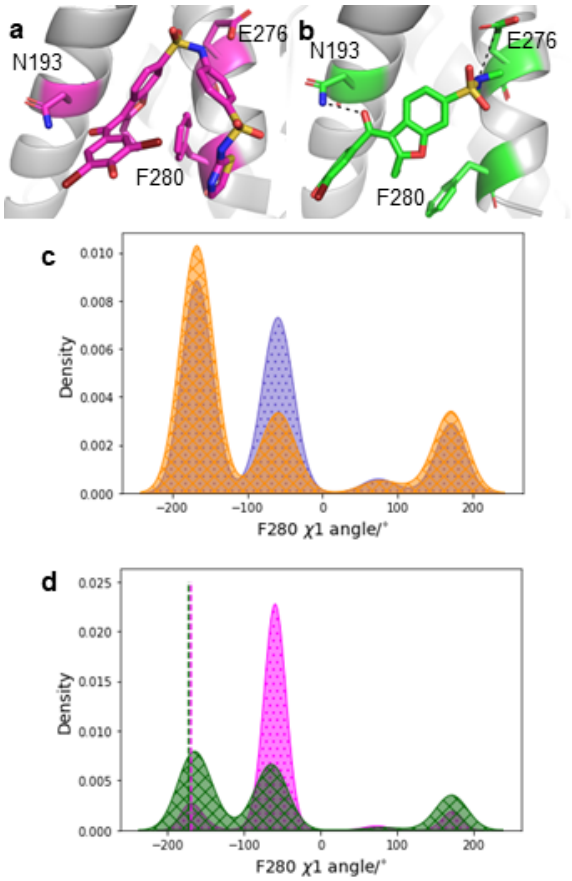

Figure 2.11: Protein and ligand conformations for PTP1B with compounds **1** and **2r**. **a** Compound **1** and key residues. **b** **2** and key residues. Distance restraints indicated by black dashed lines. **c** F280 χ1 dihedral when PTP1B is active (orange, crosses) and inactive (blue, dots). **d** F280 χ1 dihedral for PTP1B with **1**(magenta, dots) and **2r**(dark green, crosses). X-Ray values for structures with compounds **1** (PDB ID: 1T4J) and **2** (PDB ID: 1T48) are shown as (overlapping) dashed lines.

suggest that stabilisation of the "upright" pose could be a plausible design strategy to elaborate fragment **3** into a non-covalently bound allosteric inhibitor of wild-type PTP1B.

Finally, the protocol was tested on a fragment of unknown allosteric effect. Fragment binder **4**[119] was processed using a similar distance restraint scheme as for **3u**. The resulting active conformation probability distribution for **4** is broad and does not suggest allosteric inhibition when compared with the reference system (Figure 2.9c "**4**"). The major binding pose of **4** also corresponds to a "sideway" binding mode that engage in hydrogen bonds with E200, (Figure 2.12 c) on the α3 helix and adjacent to the binding site of **1** and **2**, but further away from the allosteric residues pictured previously. No minor "upright" pose was detected in the conformational ensemble. Overall these results suggest that fragment **4** does not show potential for allosteric inhibition of PTP1B without further elaboration to enforce adoption of a different binding pose.

2.3.4 Comparison of Steering Protocols Indicates the Importance of the Allosteric Network in Steered MD

The initial steering CV set included only the WPD loop, as it determines the activity of PTP1B. This sMD protocol was carried out on the reference, and compounds **1** and **2** systems, continuing with seeded MD and MSM building as outlined in section 2.3.1. In this case, compound **1** did not show significant inhibition in the initial model, and even showed an upward shift in active state probability in a replicate model (Figure 2.13. On the other hand, including the selected allosteric network residues in the steering CV set led to reproducible modelling of inhibitory effect by compound **1** (Figure 2.13b).

As the steering is carried out during 100-150 ns, the timescale is too fast for the allosteric network residues to adjust conformation in response to the change in the WPD loop conformation. Figure 2.13c and d illustrates the difference in the residue dynamics when they are left out of the steering and when they are

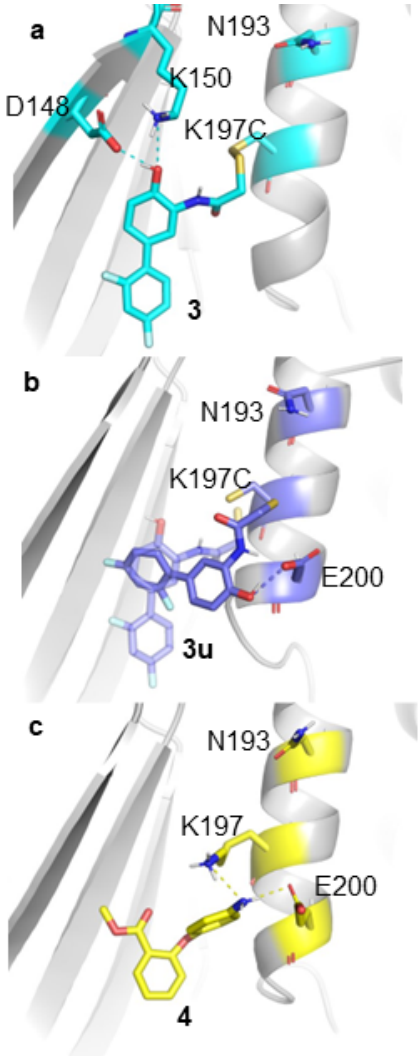

Figure 2.12: The major ligand conformations during seeded MD of **3** (cyan), **3u** (dark blue) and **4** (yellow). **a** Covalently linked **3** maintains its crystal binding pose, and forms hydrogen bonds with D148 and K150. **b** Replacing the covalent link with a distance restraint changes the binding mode, and interactions are mainly formed with E200 instead of K150. **c** Fragment **4** binds similarly to **3u**.

included. In the active WPD loop conformation, Y152 shows the "down" rotamer only, while both "up" and "down" are shown in the inactive conformation[119]. When steering from inactive to active PTP1B conformations, it is expected for Y152 to maintain or switch to the "down" rotamer. However, when the angle is not steered exclusively, the "down" roatmer is explored only briefly at the very end of the simulation. Additionally, when steering from inactive to active, the F196 to F280 stacking does not reform without exclusive steering. Interestingly, the P185 to W179 stacking does form and break in response to the WPD loop conformational change, which can be explained by the direct adjacency of P185 to the WPD loop.

2.4 Discussion

The results reported here demonstrate that the joint sMD/MSM protocol can be used to discriminate allosteric inhibitors from non-functional binders. They provide an inverse view of how this workflow could be applied in a computer-aided drug design (CADD) project. The most potent allosteric PTP1B inhibitor reported in the literature (**1**) was analysed and subsequently deconstructed into a less potent variant **2**[130]. The MSM model for **1** suggest potent inhibition in agreement with literature data. Reliable analysis of compound **2** requires the use of restraints to prevent spontaneous unbinding during MD simulations. The judicious use of distance restraints provides information on what interactions are important in the activity of compounds **1** and suggests which vectors could be grown or changed to achieve the desired functional results. Similarly, compound **3** is deconstructed into **3u** by replacing a covalent link with an *in silico* distance restraint, causing a decrease in inhibition. These different strategies to enforce proximity with PTP1B have a significant effect on the conformation of the ligand, and the interactions that are formed with the protein. Compound **4** behaves similarly to **3u**, demonstrating how the protocol may be used to profile compounds with unknown allosteric potential. Further developing **3u** or **4** to behave more like covalently linked **3** (such as moving the compound **4** acetyl group around the benzene ring) could lead to increased effi-

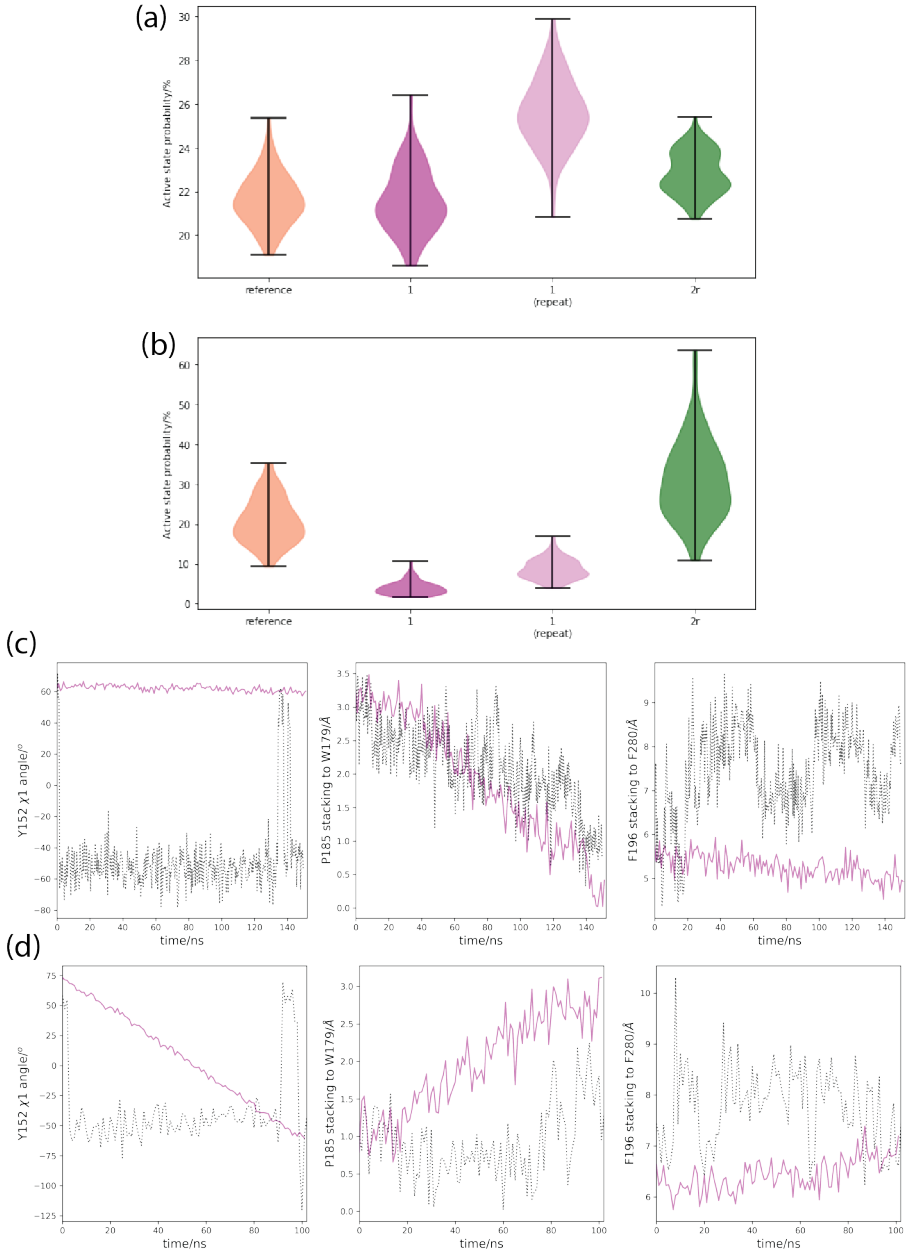

Figure 2.13: The effects of different CV sets on the active state probabilities modelled by MSMs and the allosteric network residues of PTP1B. (a) the active state probabilities of the reference, 2 compound **1** repeats, and compound **2** systems when only the WPD loop was steered during sMD. (b) The active state probabilities of the same systems when the allosteric network residues shown in Figure 2.3c were steered alongside the WPD loop. (c) The allosteric network residues during sMD when they were included in the CV set (magenta, solid line) and excluded (black, dotted line), in steering the WPD loop from open to closed. (d) The allosteric network residues during sMD when they were included in the CV set (magenta, solid line) and excluded (black, dotted line), in steering the WPD loop from closed to open.

cacy as allosteric inhibitors. Through modelling active state probabilities via MSMs, these binding pose changes can be related to protein activity.

As the modelled change in active state probability can be related to local changes in ligand binding site conformations, the seeded MD trajectories can be mined to select protein conformations associated with functional states. In turn, the resulting conformations can be used for further virtual screening, to find ligands that could induce the same binding site rearrangements. For example, the increased activity of compound **1** over compound **2** was related to differences in the preferred conformations of F280 during the MD simulations. This insight was not apparent from available X-ray crystallographic data since in existing crystal structures for **1** and **2** this residue is modelled in the "down" conformation (PDB IDs: 1T4J and 1T48 respectively)[130]. The simulations carried out here revealed an alternative "up" rotamer, which is predominantly adopted in inactive states of PTP1B. Therefore targeting the F280 "up" rotamer offers a potential focus for further drug discovery campaigns.

A key feature of the present approach is the use of steered MD simulations. Previous studies applying MSMs to study allosteric modulation have been successful in using unbiased MD simulation data[143] However, since the WPD loop of PTP1B changes conformation on multi-μs timescales[116], the simulation length required to observe a number of transitions that is statistically significant is unpractical for routine applications. sMD allows access to intermediate conformations with simulations on nanoseconds timescales, and the following short seeded MD simulations leverage parallel computing, reducing answer time even further. Sampling of relevant conformations, both of the ligand and the protein, is key to modelling activity probabilities consistent with experimental data. Steering only the active site residues does not ensure the allosteric network will adjust to new conformational states as the WPD loop moves, which causes inconsistent simulation results in disagreement with reported literature values (Figure 2.13). Future work may focus on extending the enhanced sampling methodologies used to seed the MSMs to decrease the amount

of experimental data required to ensure that the relevant protein conformational states have been sampled.

The use of Markov State Modelling enables to decrease the time-to-answer by modelling long-time scale dynamics as a set of shorter timescale simulations that may be run concurrently. However, obtaining equilibrium distributions require the use of dimensionality reduction. The MD data is reduced to chosen features, which in turn are clustered into discrete microstates. Those discrete states in this case were assigned to final "active" and "inactive" PTP1B states using Perron Cluster-Cluster Analysis (PCCA), which uses the eigenvectors of the transition matrix that makes up the MSM to find metastable states[144]. Tools that make this procedure simpler and data-driven, such as VAMPnets[145] are in development. The procedure is more complex when comparing multiple MSMs, rather than focusing on a single model. It is preferable that both the microstates and the coarser active/inactive assignments are based on the same feature values for the models to be more easily comparable. Additionally, to determine the active and inactive state partition, the assignments from PCCA of the reference system were used throughout to keep them consistent. However, seven total MSMs were built in this case, and any of those assignments could be used. Figure 2.10 shows the effects of using different active state definitions on the active state probability and ranking. The results remain qualitatively consistent but in general the automated selection of a suitable macrostate definition is non-trivial. Therefore future work focusing on automating the MSM construction and analysis steps is desirable to facilitate deployment of the technology at scale.

The work in this chapter relies on simulation of binders to assess their potential allosteric effects when bound to different pockets on the surface of a protein. Future developments of the protocol could be sought to allow characterisation of the allosteric potential of cryptic binding sites discovered by molecular dynamics simulations, enabling protein druggability assessments prior efforts to identify binders have been initiated[60, 73].

Overall the results presented in this chapter suggest that it is viable to routinely

compare numerous Markov State Models to assess the effects of ligand binding or point mutations on protein function. Extension of the present sMD/MSM methodology to other drug target classes is warranted to validate the generality of the approach for supporting allosteric drug design workflows, and is outlined in chapters 3 and 4.

Chapter 3

Elucidation of the Mechanism of Enzyme EPAC1 Partial Activation by the Small Molecule Agonist I942

3.1 Introduction

The previous chapter outlined an application of AMMo on modelling inhibitors, as well as a protein whose activity is defined by localised loop rearrangement. An important step in validating the sMD/MSM approach is to confirm that it can also capture the effects of activators, and expand to other drug target classes. In this chapter, we look at exchange proteins activated by cAMP (EPACs). The activation mechanism here is defined by large scale domain rearrangements, and we model activation by cAMP and a non-cAMP-like compound I942.

3.1.1 Idiopathic Pulmonary Fibrosis

Idiopathic pulmonary fibrosis (IPF) is a disease of the lungs, where chronic inflammation causes large amounts of scar tissue to accumulate on the alveoli and damage

the lung architecture. Patients diagnosed with IPF suffer from chronic coughing and shortness of breath, as the lung capacity and subsequently blood oxygen concentration decrease[146]. The most common risk factors are age and environmental circumstances, such as smoking or exposure to metal, plant, and animal dust. Currently available antifibrotic treatments, barring lung transplants, can only slow down the progression of IPF, not reverse the damage already present[147], while patients have a life expectancy of only around 3-4 years without any treatment[148]. As such, early diagnosis and treatment are crucial to preventing fatal outcomes. Additionally, the spread of the COVID-19 pandemic worldwide has drawn attention to the link between severe cases of SARS-CoV-2 and IPF[149].

3.1.2 Exchange Proteins Activated by cAMP

EPACs (Exchange Proteins Activated by cAMP) are guanine nucleotide exchange factors (GEFs) for RAP1[150]. RAP1 is a small GTPase, involved in cell adhesion[151], proliferation and migration[152]. There are two isoforms of EPACs: EPAC1 and EPAC2. They have similar structures, but are found in different types of tissues and are responsible for different biological functions[153]. For instance, EPAC2-knockout (KO) mice showed no Ca^{2+} leak induced arrhythmias that are present in wild type (WT) and EPAC1-KO mice, implicating EPAC2 but not EPAC1 in cardiac function[154]. In particular, EPAC1 activation has been linked to dose-dependent decrease in fibroblast proliferation[155], relaxation of airway smooth muscle cells[156], and regulation of lung epithelial cell adhesion and migration[157]. Therefore, targeting activation of EPAC1 could provide additional treatment to reduce further lung damage[147].

EPAC1 is comprised of a regulatory region (RR), containing the allosteric cAMP binding site, and a catalytic region (CR), containing the Rap binding (active) site[153]. The CR is comprised of the RAS-exchange motif (REM), the RAS association (RA), and CDC25 homology domain (CDC25HD). When RAP1 binds to EPAC1, the nucleotide binding site on RAP1 becomes deformed, decreasing both

GDP and GTP affinity. The significantly higher GTP concentration in the cell favours GTP binding over GDP[158]. The RR of EPAC1 contains Disheveled Egl-10 Plecstrin (DEP) and a single cyclic nucleotide binding domain (cNBD) (Figure 3.1A). In the absence of cAMP, EPACs exist in an auto-inhibited state, where the RR is blocking the Rap binding site on the catalytic region. The hinge between the two regions adopts a helical conformation, while the adjacent phosphate binding cassette (PBC) sterically blocks the hinge from moving and revealing the Rap binding site[159]. The coupling of the conformational changes between the PBC and hinge is highlighted in the L273W mutant. Replacement of the L273 residue on the PBC with a bulky tryptophan mutation prevents the hinge adopting the active conformation, and no GEF activity was observed in $EPAC1_{L273W}$, even in the presence of 500 μM cAMP[160]. The interface between the RR and the CR is also stabilized via a mixture of hydrogen bonding and ionic interactions between residues of the two domains (Figure 3.2) (ionic latch, or IL), and the helical conformation of the hinge prevents opening of EPAC1[153, 161]. When the phosphate-sugar group of cAMP binds to the PBC, it shifts it from the "out" to the "in" conformation (Figure 3.1a). This allows the hinge helix to unfold at its C-terminus end, moving the RR by approximately 45 Å and enabling RAP binding to the exposed active site (Figure 3.1a). Additionally, the active conformation of EPAC1 is further stabilized by the interactions between the adenosine group of cAMP and K353 of the REM domain on the catalytic region, as well as E315 of the C-terminus of the cNBD (Figure 3.1c). The terminal β-sheet strands of the cNBD and the first helix of the REM domain are known as the "lid", as they close off the previously solvent-exposed cAMP binding site upon activation[159].

3.1.3 Activators of EPACs

In the initial efforts to drug EPAC1, a number of cAMP analogue agonists were developed[163–168], such as 8-CPT[169] (Figure 3.1b). However, these have been associated with cardiac hypertrophy[170] and off-target effects[171]. This is poten-

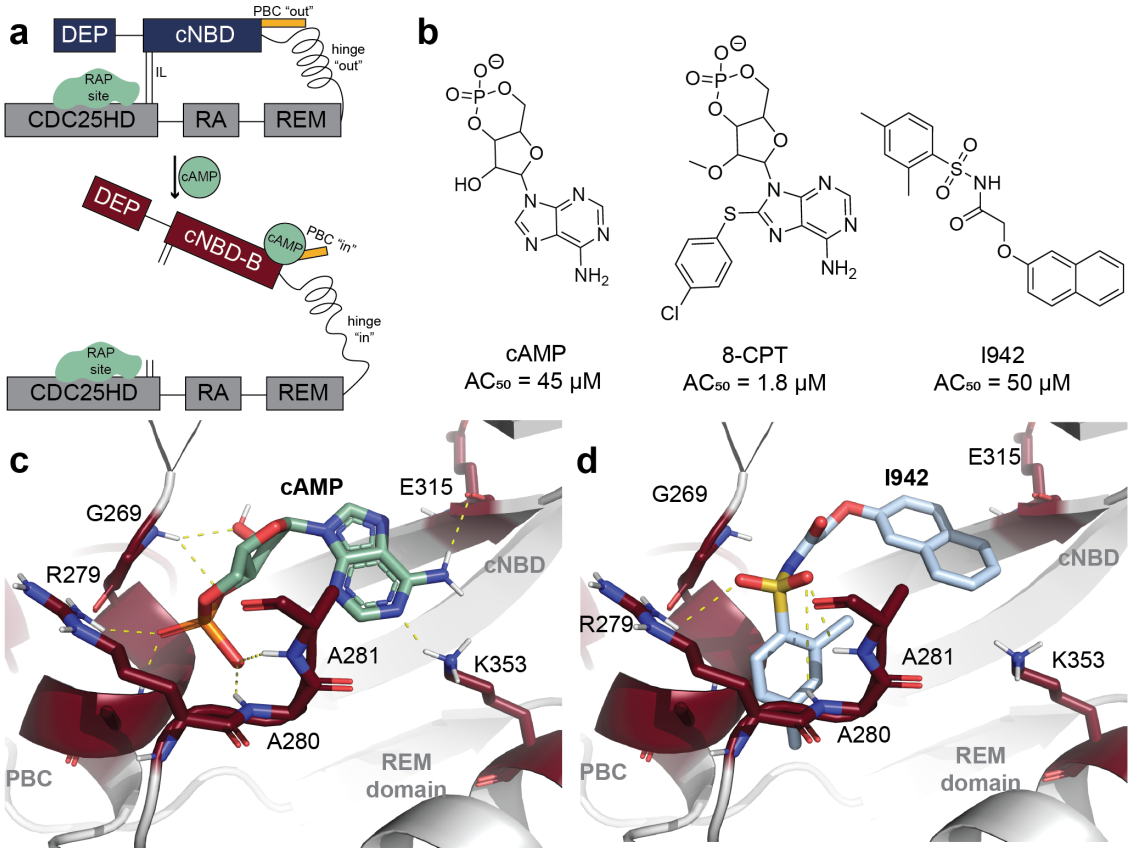

Figure 3.1: Structure of active and inactive EPAC1, and activators. (a) *apo* EPAC1 exists in an auto-inhibited state, with the RR region (blue) blocking the RAP binding site (green) on the CR (gray). Upon binding of cAMP, the RR moves away (red), exposing the RAP binding site. (b) Activators of EPAC1: cAMP, a cAMP analogue 8-CPT, and a non cAMP analogue partial activator I942. (c) The modelled binding pose of cAMP (light green) to EPAC1 in the active conformation (red), based on structure of cAMP analogue Sp-cAMP with EPAC2 (PDB ID 3CF6). Key interactions are shown in yellow dashes. (d) The modelled binding pose of I942 (light blue) to EPAC1 in the active conformation (red). The binding pose was modelled based on findings by Shao *et al.* on EPAC1-I942 interactions[162]. Key interactions are shown in yellow dashes.

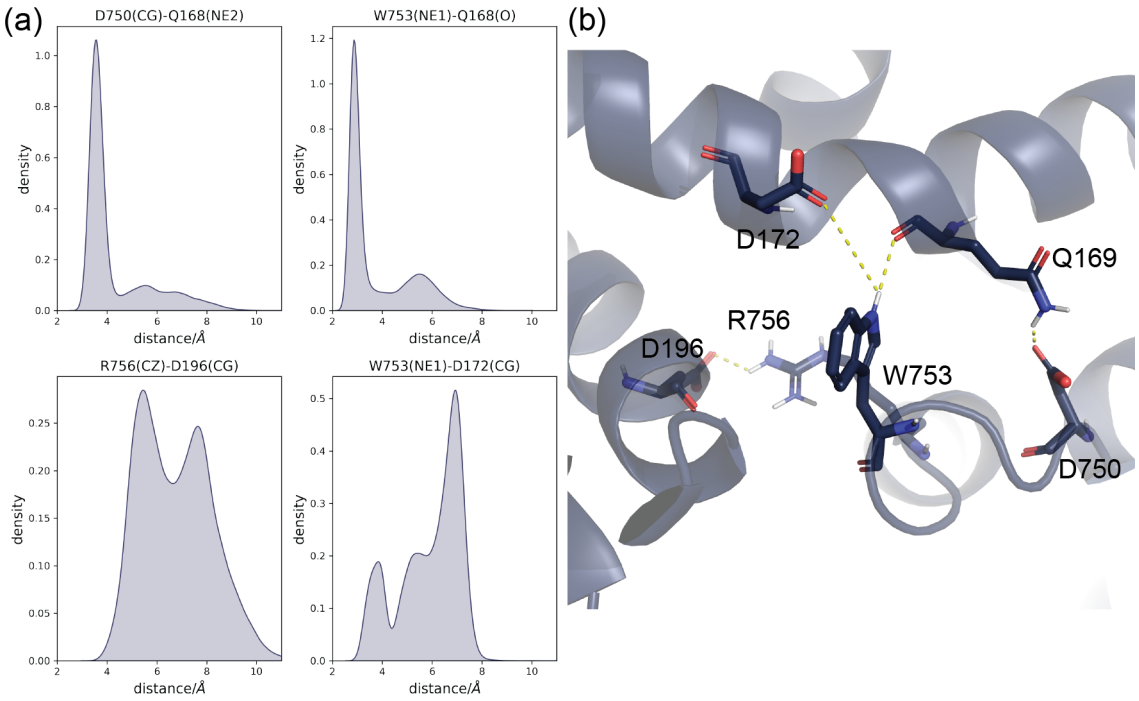

Figure 3.2: Ionic latch of EPAC1. (a) Distances between involved residues during 500 ns equilibrium MD simulation of the *apo* EPAC1 in inactive conformation. (b) Residues involved in the ionic latch.

tially due to non-selective activation of both EPAC isoforms, as EPAC2 is implicated in regulation of cardiomyocytes[154]. These side effects have shifted the focus of activating EPAC1 to non-cAMP-like small molecules[172], leading to the development of I942 (Figure 3.1b and d), a selective partial EPAC1 agonist. It was identified in a HTS of lead-like small molecules, where compounds were tested for competitive inhibition of a fluorescent cAMP analogue 8-NBD-cAMP. I942 was shown to bind to both EPAC isoforms, but elicit partial activation in EPAC1 only. Due to binding to both EPAC1 and EPAC2, I942 is also a competitive inhibitor of cAMP-induced GEF activity of both proteins. As I942 can only induce partial activation of EPAC1 (and inhibit EPAC2), it may have an net inhibitory effect of EPAC GEF activity *in vivo*[173]. However, I942 is much more drug-like compared to close cAMP analogues, and provides a more suitable scaffold for further lead optimisation towards a full selective activator. NMR studies suggest that while I942 binds to the same site as cAMP, the partial activation of EPAC1 comes from the stabilization of an intermediate state, rather than the active conformation of EPAC1. In this intermediate state, the PBC has adopted an "in", or active-like, conformation, but the hinge in the cNBD has not yet shifted and remains in the "out", or inactive-like" conformation[162].

While the NMR work done on EPAC1-I942 has provided evidence for the intermediate state stabilization to explain the partial activation observed, the data on I942 is still sparse and limited by the experimental techniques employed. Due to poor stability of full length EPAC1, only the more stable isolated EPAC1-cNBD domain was used for the original HTS screen that identified I942. When the full length protein was used in the activation assay, I942 affinity was found to be higher than seen previously with just cNBD[173]. Additionally, the work by Shao *et al.* also used only EPAC1-cNBD (residues 149-318), which excludes the significance of ligand interactions with the CR region (such as K353 shown in Figure 3.1c and d). On that account, molecular dynamics (MD) simulations are employed here to fill the gaps in available experimental data and provide atomistic resolution insight into the

dynamics of the whole protein. In Chapter 2, we have combined enhanced sampling molecular dynamics with Markov State Modelling (MSM), to model changes in the protein conformational ensemble in the presence of allosteric inhibitors. However, this work only covered small loop motions, rather than large domain rearrangements that characterise EPAC activity[174].

In this chapter, we use steered molecular dynamics (sMD) to push the conformation of EPAC1 from inactive to active conformation (and *vice versa*) via the intermediate state, followed by an ensemble of further, unbiased MD simulations, starting from various points along the sMD coordinate. This unbiased MD data is used to build MSMs, which capture a three state ("inactive", "intermediate", "active") partitioning of the EPAC1 conformational ensemble. As this methodology has only been applied to inhibitors previously, it is validated by first comparing the *apo* WT EPAC1 to the EPAC1-cAMP complex, as well as $EPAC1_{L273W}$-cAMP. Our modelling correctly captures the activation of WT EPAC1 by cAMP, and also the prevention of the activation by the L273W mutation. Following this, we investigate the effects of I942 on EPAC1 by modelling I942 with native interactions only, as well as with additional artificial protein-ligand distance restraints to mimic hydrogen bonding observed with cAMP but not I942. Comparisons of the computed conformational ensembles allowed us to isolate and determine the effects of key cAMP-EPAC1 interactions. Finally, we propose structural modifications that could turn I942 into a full EPAC1 agonist.

3.2 Methods

3.2.1 Protein Modelling

The homology models of EPAC1 in the active and inactive conformations were prepared by collaborators at Heriot-Watt University as follows, using the SWISS-MODEL webserver[175]. The FASTA sequence was obtained from the UniProt database (entry O95398). X-Ray diffracted structures of EPAC2 were used as tem-

plates when constructing the protein models: and EPAC2-cAMP analogue complex (PDB ID: 4MGK) and *apo* EPAC2 (PDB ID: 2BYV) for active and inactive conformations respectively. The active conformation template lacked the DEP domain, which therefore was copied from the inactive conformation using PyMol.

All protein structures were capped with ACE and NME caps for the N- and C-termini respectively using software Flare[134]. In all cases histidines 439 and 646 (model residues 392 and 599) were modelled as protonated at the δ position based on propka3[176] predicted pK_a values of 2.34 and 1.80 respectively.

3.2.2 Ligand Modelling

Cyclic AMP was manually docked to the active conformation of EPAC1 by aligning the EPAC2 complex with the cAMP analogue (Sp-cAMP) in the PDB entry 3CF6, and editing it in Flare[134]. cAMP was further manually docked to the inactive conformation of EPAC1 by aligning the cNBD regions of the protein using PyMol and copying the ligand coordinates.

Compound I942 was manually docked in-situ by editing the previously obtained cAMP coordinates in Flare. The ligand pose was in agreement with NMR measurements from Shao *et al.* that indicated formation of hydrogen bonds between I942 and EPAC2 residues R279, A280 and A281[162].

3.2.3 System Preparation

All system setup was carried out via BioSimSpace[136]. The proteins were parameterized using the AMBER ff14SB forcefield, and GAFF2 with the AM1-BCC charge method was used for the ligands. cAMP was modelled with a charge of -1, and I942 was modelled as neutral. All systems were solvated in TIP3P water with a 15 Å shell and a 150 mM NaCl concentration, adding ions as needed to neutralize the system. In all cases, minimization was carried out for 7500 steps, heating to 300 K for 500 ps, and further equilibration for 2 ns, all using GROMACS 2020.2[138]. Particle Mesh Ewald (PME) was used, with a direct space cutoff of 12 Å.

3.2.4 Equilibrium Molecular Dynamics

Simulations of *apo* EPAC1 in active and inactive conformations were carried out for 1 μs each, using pmemd.cuda from AMBER22[139]. Trajectories were written out every 10 ps, and simulations were run at 300 K temperature and 1 atm pressure, using Langevin dynamics with γ=2 ps^{-1}. Features later used for MSM building (section 3.2.7) were computed using cpptraj v4.25.6 (AmberTools22)[141].

3.2.5 Steered Molecular Dynamics

Steered molecular dynamics were performed using pmemd.cuda from AMBER22[139] with PLUMED v2.6.1[177, 178], using BioSimSpace for input file preparation. Simulations were run at a temperature of 300 K and 1 atm pressure, using Langevin dynamics with γ=2 ps^{-1}. Steering was carried out in both directions in two steps, i.e. inactive-intermediate-active and active-intermediate-inactive (Figure 3.1B). The steering CVs were (also see Table 3.1):

- Regulatory region backbone RMSD to the final target conformation (active or inactive)
- Hinge backbone RMSD to step target conformation (active, intermediate, or inactive)
- PBC backbone RMSD to step target conformation (active, intermediate, or inactive)

The reference for hinge and PBC in all cases included the cNBD up to and including the hinge region (residues 122-263 of the protein model), in order to remove the noise from the overall domain translation and only maintain the internal rearrangements. The reference for the intermediate state was prepared using PyMol by replacing the PBC (residues 270-274, model residues 223-227) of the inactive conformation with the one from the active conformation. No additional energy minimisation was required. The force constant applied to all CVs was 3500 kJ mol^{-1},

and the simulation duration was 60 ns for each step (120 ns total). All parameters are shown in Table 3.1.

To maintain relevant binding poses and prevent ligand dissociation during the large cNBD movement, both cAMP and I942 were restrained using flat bottomed restraints during all sMD simulations. During preliminary runs, dissociation of both cAMP and I942 was observed. The ligand restraints are shown in Table 3.2.

3.2.6 Seeded Molecular Dynamics

The two steering trajectories were combined into a single trajectory and 100 snapshots were extracted from each steering direction, equally sampling the CVs used for steering, using cpptraj. These conformations were used as starting points (or "seeds") for a further 50 ns of equilibrium MD simulations with pmemd.cuda (AMBER22). Trajectories were written out every 10 ps, and simulations were run at 300 K temperature and 1 atm pressure, using Langevin dynamics with γ=2 ps^{-1}. In the case of the I942 restrained system only, some ligand restraints were also maintained during the seeded MD simulations (Table 3.2, Figure 3.3).

3.2.7 Markov State Modelling

Each seeded MD trajectory was reduced to the following features: the RR-hinge-CR angle, the hinge RMSD to the inactive conformation and the PBC RMSD to the inactive conformation, using cpptraj. Residue masks are outlined in Table 3.3.

All Markov State modelling here was performed using pyemma version 2.5.7[179]. The data from *apo* EPAC1, EPAC1-cAMP, EPAC1 L273W, EPAC1-I942 was pooled together and clustered into 300 microstates using k-means clustering. Each frame of the seeded MD trajectories was then assigned to one of these 300 states, and used to build Markov State Models for each system, with a lag time of 25 ns. The microstates were assigned to 3 metastable states: active, inactive, and intermediate. The state with the lowest domain angle, hinge and PBC RMSD values would be considered the inactive state, while the state with the highest values of these features

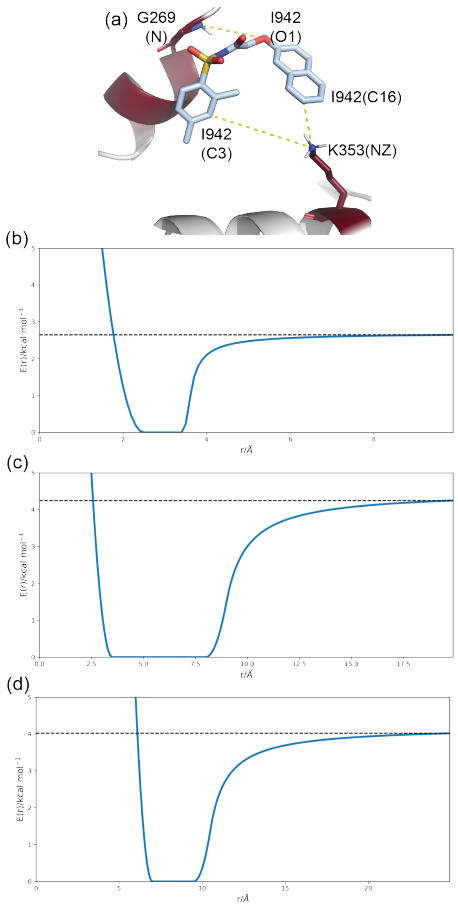

Figure 3.3: Restraints between EPAC1 and I942, mimicking cAMP interactions with the PBC and lid regions. (a) Atoms used for restraints. (b) Restraint energy as a function of distance between I942(O1) and G269(N). The maximum restraint value is shown as a black dashed line. (c) Restraint energy as a function of distance between I942(C16) and K353(NZ). The maximum restraint value is shown as a black dashed line. (d) Restraint energy as a function of distance between I942(C3) and K353(NZ). The maximum restraint value is shown as a black dashed line. This restraint was added to ensure the correct placement of K353, as the distance restraints lack the directionality of hydrogen bonds. All energies were computed as per the AMBER22 manual[139].

CV	**Residues**	**Reference**	**Target value /Å**	**Force /kcal mol^{-1}**
Inactive to Active				
Regulatory Region	48-339	active active	intial/2, 0	3500, 3500
Hinge	297-310	intermediate active	0, 0	3500, 3500
PBC	270-274	intermediate active	0, 0	3500, 3500
Active to Inactive				
Regulatory Region	48-339	inactive inactive	intial/2, 0	3500, 3500
Hinge	297-310	intermediate inactive	0, 0	3500, 3500
PBC	270-274	intermediate inactive	0, 0	3500, 3500

Table 3.1: The collective variable definitions for sMD simulations, as well as the references, target values, and forces at each of the 2 steps of the steering. For the RR RMSD, the target value at step 1 was half of the starting RMSD value. In all cases, backbone atoms were used for steering.

Atoms	**r1**	**r2**	**r3**	**r4**	**rk2**	**rk3**	**ialtd**
cAMP							
:233@H :cAMP@O1P	1.3	1.8	3.0	3.50	0.0	150.0	0
cAMP @C4 @N9 @C1' @O4'	46.0	47.0	48.0	49.0	150.0	150.0	0
I942							
:I942@O5' :232@NH1	2.5	3.0	4.0	4.5	0.0	150.0	0
:I942@O :233@N	2.5	3.0	4.0	4.5	0.0	150.0	0
:I942@O :234@N	2.5	3.0	4.0	4.5	0.0	150.0	0
I942 restrained							
:I942@O5' :232@NH1	2.5	3.0	4.0	4.5	0.0	150.0	0
:I942@O :233@N	2.5	3.0	4.0	4.5	0.0	150.0	0
:I942@O :234@N	2.5	3.0	4.0	4.5	0.0	150.0	0
:I942@O1 :222@N	2.0	2.5	3.4	3.6	5.0	22.5	1
:I942@C16 :306@NZ	2.0	3.5	8.0	9.0	5.0	1.5	1
:I942@C3 :306@NZ	6.0	7.0	9.5	10.5	5.0	1.4	1

Table 3.2: The flat bottomed restraint parameters used to restrain all ligands. Atoms are indicated using the AMBER atom masks. 2 atoms mean a distance restraint, and 4 atoms mean a torsional angle. r1-4 are the flat bottom well defining points (in Å for distances and degrees for dihedrals), and rk2-3 are the restraint energies (in kcal mol^{-1} $Å^{-2}$ for distances and kcal mol^{-1} deg^{-1} for dihedrals. The ialtd parameter indicates whether energy penalties plateu (ialtd=1) or increase indefinitely (ialtd=0). Restraints with atoms indicated in bold have also been applied in some seeded MD simulations. Note that atom masks are for protein models used in this chapter, which have an offset of -47 from the full protein.

Feature	**type**	**mask**	**reference (RMSD only)**	**alignment mask (RMSD only)**
domain angle	angle	:121-125 (@CA,C,N,O) :250-263 (@CA,C,N,O) :701-704 (@CA,C,N,O)	-	-
hinge	RMSD	:250-263 (@CA,C,N,O)	inactive	:122-263 !(@/H)
PBC	RMSD	:223-227 (@CA,C,N,O)	inactive	:122-263 !(@/H)

Table 3.3: AMBER selection masks for the features used to reduce data dimensionality when building MSMs. Note that atom masks are for protein models used in this work, which have an offset of -47 from the full protein.

would be the active. The intermediate state is characterised by a low hinge RMSD value ("inactive"-like) and a high PBC RMSD value ("active"-like), as outlined in Shao *et al.*[162]. Originally, PCCA as used in Chapter 2 was applied, however it did not yield a set of states with well-defined state clusters that satisfied the conditions above (Figure 3.4). Therefore, the macrostate assignment was done manually. The metastable state centres were assigned based on feature values during equilibrium and seeded MD simulations (see section 3.3.1), and each microstate was assigned to the closest centre. The total macrostate probabilities were computed by summing over the probabilities of the microstates that belong to the macrostate.

In order to asses the quality of the models built, the state probabilities of each system were bootstrapped by resampling. For each system, the seeded MD trajectory pool was resampled with replacement to select 200 trajectories. With this new resampled pool of trajectories, a new MSM was built, but using the same micro- and macro-state assignments. The probabilities of inactive, intermediate, and active states were computed as above. This was repeated for 100 iterations, yielding a distribution of probabilities for each system. The mean values were used to report metastable populations, and the standard deviation of each distribution was used to report statistical uncertainties. Models where the conformational space is well sampled will show a smaller error, as excluding a few trajectories should not significantly change the final probability values.

A Note on Errors

In Chapter 2, the above method of bootstrapping was also applied to evaluate the robustness of the MSM modelled probabilities of states. The bootstrapped probabilities were assumed to come from a natural distribution, and so the mean and standard deviation were reported, with the standard deviation providing information on the spread of the data. However, with the more complex dynamics of the systems reported in Chapters 3 and 4, some of the bootstrapped probability distributions, particularly those close to 0% and 100%, are skewed and using the standard

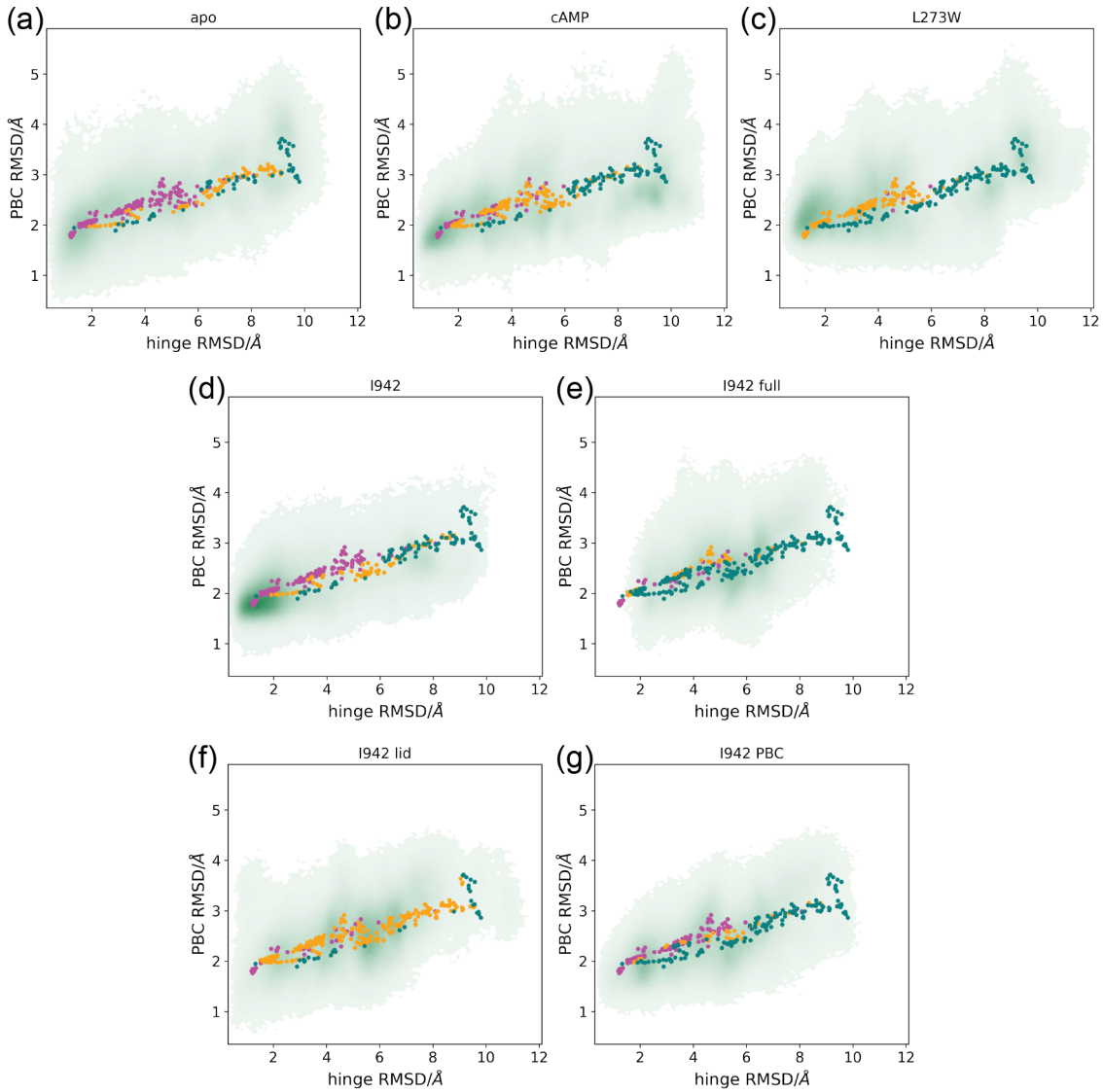

Figure 3.4: The results of PCCA on each of the EPAC1 MSMs used in this chapter: (a) *apo* (b) EPAC1-cAMP (c) $EPAC1_{L273W}$-cAMP (d) EPAC1-I942 (e) EPAC1-I942 restrained to both PBC and lid (f) EPAC1-I942 restrained to lid only (g) EPAC1-I942 restrained to PBC only

Feature	**inter-domain angle/°**	**hinge RMSD/Å**	**PBC RMSD/Å**
Inactive	30	1.0	1.0
Intermediate	60	1.0	2.5
Active	100	8.0	2.5

Table 3.4: The domain angle, hinge RMSD and PBC RMSD values defining the centers to partition microstates into inactive, intermediate, and active metastable states.

deviation as the error would take those values outside of the possible 0-100% probability range. In order to still report the spread of the boostrapped probabilities and provide insight into the quality of the models analysed in these Chapters, the inter-quartile range (IQR) between quartiles 1 and 3 is reported instead of the standard deviation, and the median is used instead of the mean for the final probability value.

Conformational ensembles were generated by drawing 10,000 frames from the seeded MD pool. The probability of any frame to be sampled was based on the MSM-computed equilibrium probability of the microstate that frame was assigned to. The distances and dihedrals discussed in the results below were computed from these ensembles using cpptraj. In the case of $EPAC1_{L273W}$-EPAC1 simulations, the active state ensemble was also recreated similarly, using only the active state probabilities to sample, to confirm the cAMP hydrogen bonds to K353 in the active state of the L273W mutant. This was necessary as the active state is barely sampled in the full equilibrium ensemble.

3.3 Results

3.3.1 Using Equilibrium and Steered MD Simulations to Define the Metastable States of EPAC1

In order to analyse the behaviour of relevant EPAC1 regions, equilibrium MD simulations in both active and inactive conformations were carried out for a duration of 1 μs each. The features later used to build Markov State Models were computed: RR-hinge-CR inter-domain angle, hinge RMSD to the inactive conformation, the PBC RMSD to the inactive conformation (Figure 3.5a, Figure 3.6). The domain angle for the inactive conformation remained below 25°, and in the region of 125-150° for the active conformation. The hinge RMSD values were similarly well separated and stable, however PBC RMSD showed some overlap between the active and inactive conformations. This is due to the fact that the change in PBC conformation is very

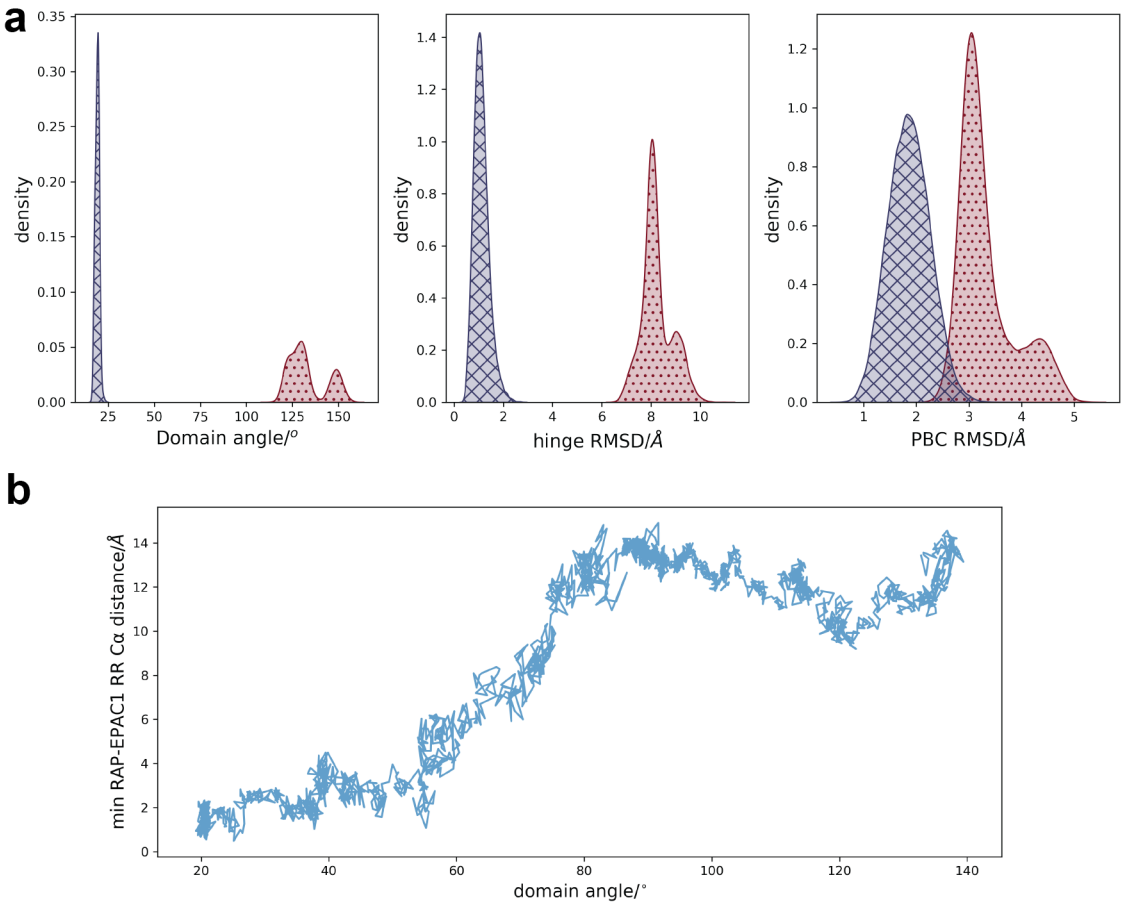

Figure 3.5: Simulations used to define the inactive, intermediate, and active states of EPAC1. (a) The domain angle, hinge RMSD to the inactive conformation, and PBC RMSD to the inactive conformation during 1 μs equilibrium MD simulations of EPAC1 in inactive (blue, crosses) and active (red, dots) conformations. (b) The minimum C(α)-C(α) distances between aligned RAP and EPAC1 regulatory region for each frame of a sMD trajectory, plotted against the associated domain angle values.

small between active and inactive EPAC1.

These results were sufficient to describe the active and inactive states of EPAC1, by using feature values observed during the equilibrium MD simulations. As the intermediate conformation is defined as a mix of hinge "out" (or inactive-like) and PBC "in" (or active-like) conformations, the definition of the hinge and PBC coordinates was also straightforward. However, the appropriate domain angle values to be used for the definition of the intermediate state were not evident from the equilibrium MD simulations, as EPAC1 remained in its active or inactive starting conformation on 1 μs timescale. To choose a suitable domain angle value for the intermediate state, a simulation steering EPAC1 from inactive to active conformation (outlined below) was aligned to an EPAC2 structure with RAP bound (PDB ID 3CF6). Only the catalytic regions were used for alignment, to capture the translation of the regulatory region. Pairwise C(α)-C(α) distances between all EPAC1 RR and all RAP residues were computed for the sMD simulation, taking the smallest one as the EPAC1(RR)-RAP distance. This distance was plotted against the domain angle, to make a direct relation between the inter-domain angle and the occlusion of the RAP binding site by the cNBD (Figure 3.5b). While the domain angle increased from 20^o to 60^o, the minimum Cα-Cα distance remained low, but jumped up as the domain angle increased from 60^o to 95^o (Supplementary Figure 2B). Therefore, the inter-domain angle of 60^o was chosen to define the centre of the intermediate state, with values of 30^o and 100^o chosen for the centres of the inactive and active states respectively.

3.3.2 Markov State Modelling Captures Activation by cAMP

The sMD/MSM workflow shown in Figure 3.6a was applied to *apo* EPAC1 and the EPAC1-cAMP complex. Steered MD simulations were performed, steering EPAC1 from active to inactive, and inactive to active conformations, via the intermediate state. The collective variables (CVs) used to define the conformational change were regulatory region (residues 48-339) backbone root-mean-square-deviation (RMSD),

hinge (residues 297-310) backbone RMSD, and PBC (residues 270-274) backbone RMSD, all to the target state conformation of each step (Table 3.1). To capture the internal rearrangements in the hinge and PBC, only the cNBD (residues 169-310) of EPAC1 was used as a reference for these RMSD calculations, in accordance with work by Shao *et al.*[162]. The intermediate state was defined by the hinge and PBC adopting inactive- and active-like conformations respectively, while the regulatory region RMSD was set at the halfway RMSD value to active or inactive conformation, depending on the final target state. The two steps were combined to yield an "active-intermediate-inactive" trajectory, and an "inactive-intermediate-active trajectory". Example sMD results are shown in Figure 3.7.

From each of sMD trajectories, 100 snapshots were evenly sampled, giving a range of protein conformations (Figure 3.6a, 200 snapshots total for each system). They were used as "seeds" for further 50 ns equilibrium MD simulations, which in turn were reduced to 3 features: hinge backbone RMSD to the inactive conformation and PBC backbone RMSD to the inactive conformation, as for the steering, but instead of using the RR RMSD, a cNBD-hinge-CR domain angle was used (Figure 3.6b). All featurized data used in this work was pooled together and clustered into 300 microstates via k-means clustering. The larger number of states was required for a finer space partitioning in 3 dimensions. These same states were used to build MSMs of each system with a 25 ns lag time, based on implied timescales (ITS) shown in Figure 3.8.

From the MSMs, the equilibrium probability of each microstate is computed. The density plot of each state is shown in Figure 3.9, plotted as a function of the hinge and PBC RMSD values. The microstates were further partitioned into inactive, intermediate and active metastable states, by defining metastable state centers in the 3D feature space, and assigning each cluster to the closest one. To calculate the probability of each metastable state, the equilibrium probabilities of the microstates comprising the metastable state were summed. To further evaluate the quality of the model, bootstrapping by resampling of the seeded MD trajectory

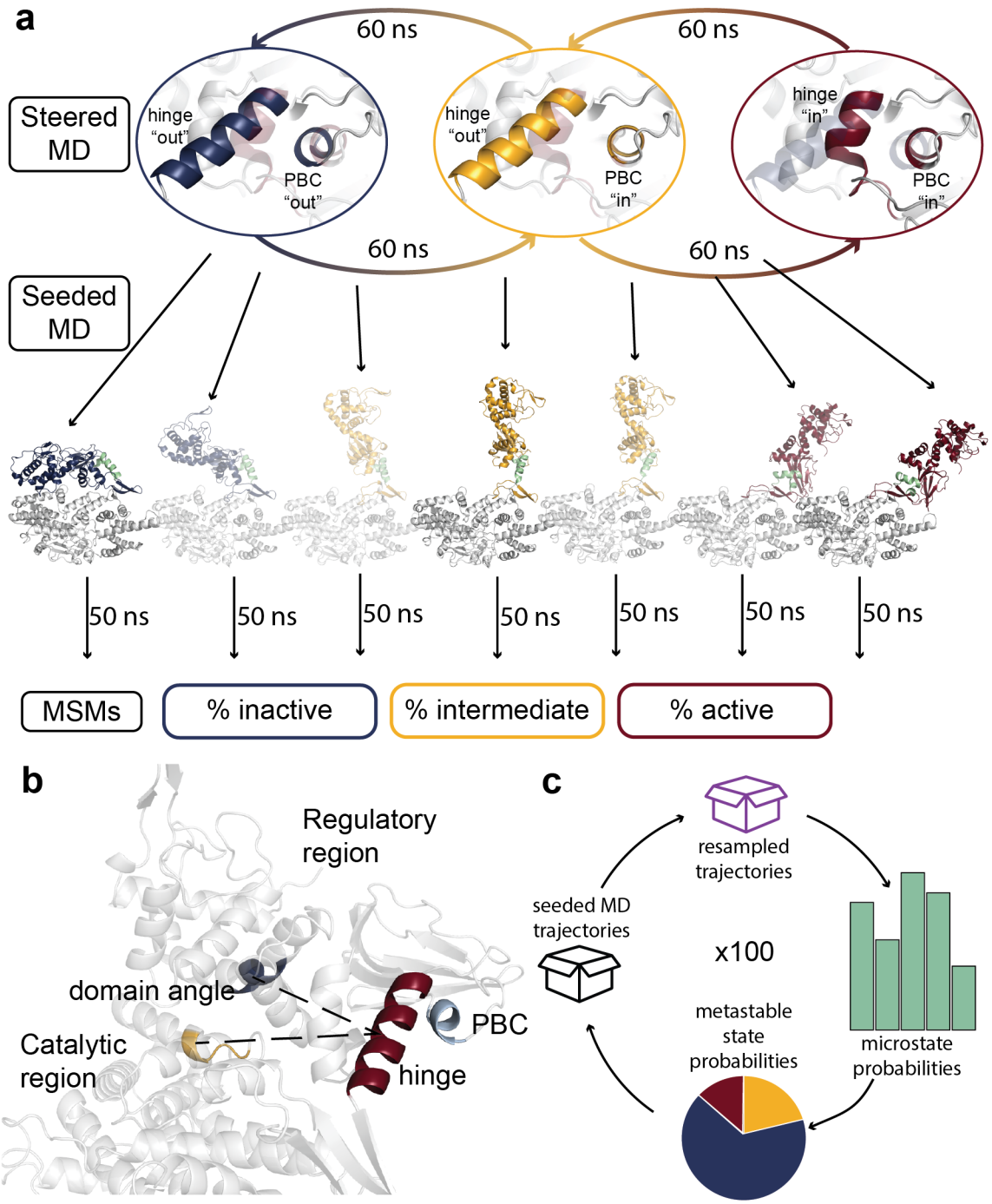

Figure 3.6: The sMD/MSM workflow applied to EPAC1. (a) sMD was used to drive EPAC1 from active to inactive state and *vice versa*. In the examples shown in this figure, the RR region is highlighted (hinge and PBC shown in green), while the CR is coloured in grey. Further unbiased 50 ns MD simulations (seeded MD) were run using each of the conformation snapshots as starting coordinates, which were used to build the Markov State Models. (b) The features used to reduce MD data dimensionality: cNBD-hinge-CR domain angle (shown in dashed black lines), hinge (red) RMSD and PBC (light green) RMSD, both to inactive conformation. (c) The bootstrapping cycle. At each iteration the seeded MD pool for each system is randomly resampled with replacement to select 200 trajectories.

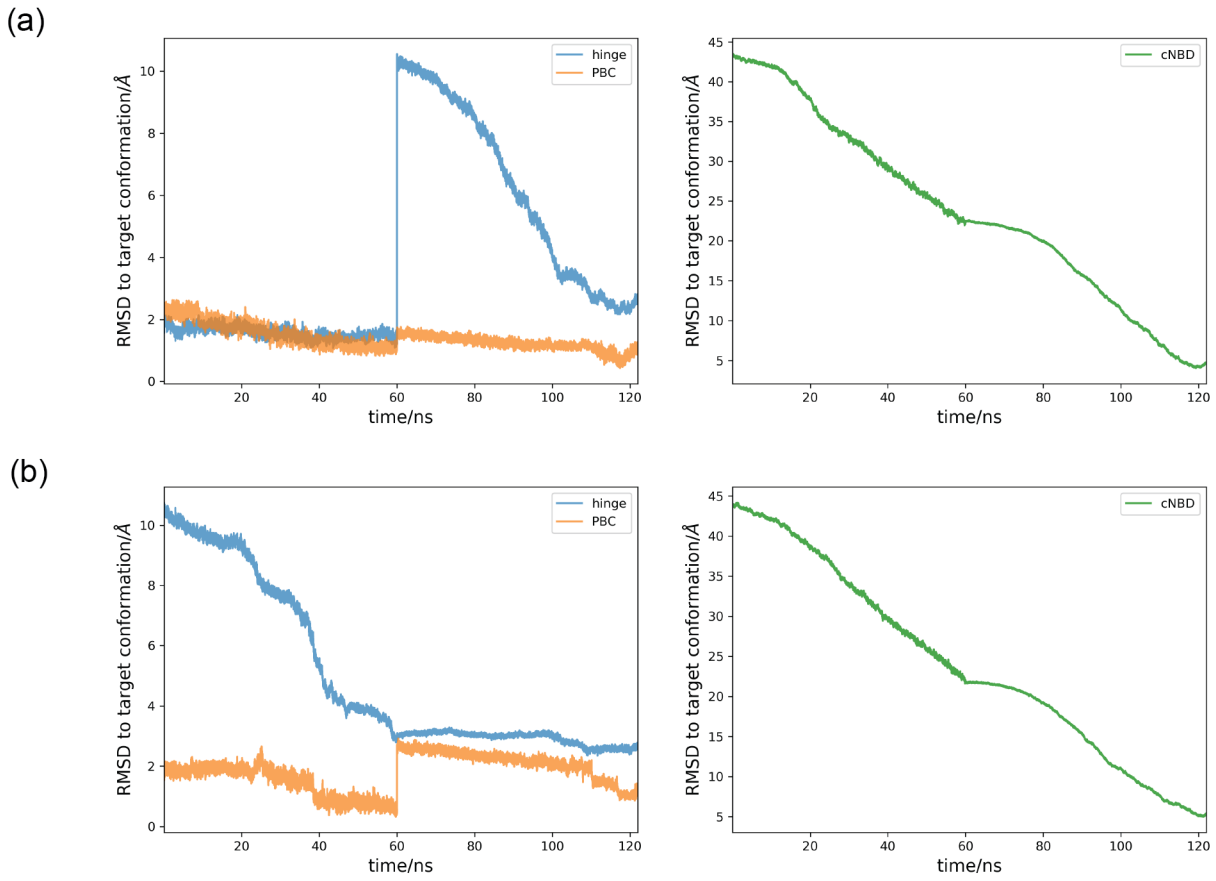

Figure 3.7: Example results of *apo* EPAC1 sMD simulation: (a) inactive to active, and (b) active to inactive. The shift in hinge and PBC RMSD corresponds to changes in RMSD reference conformation.

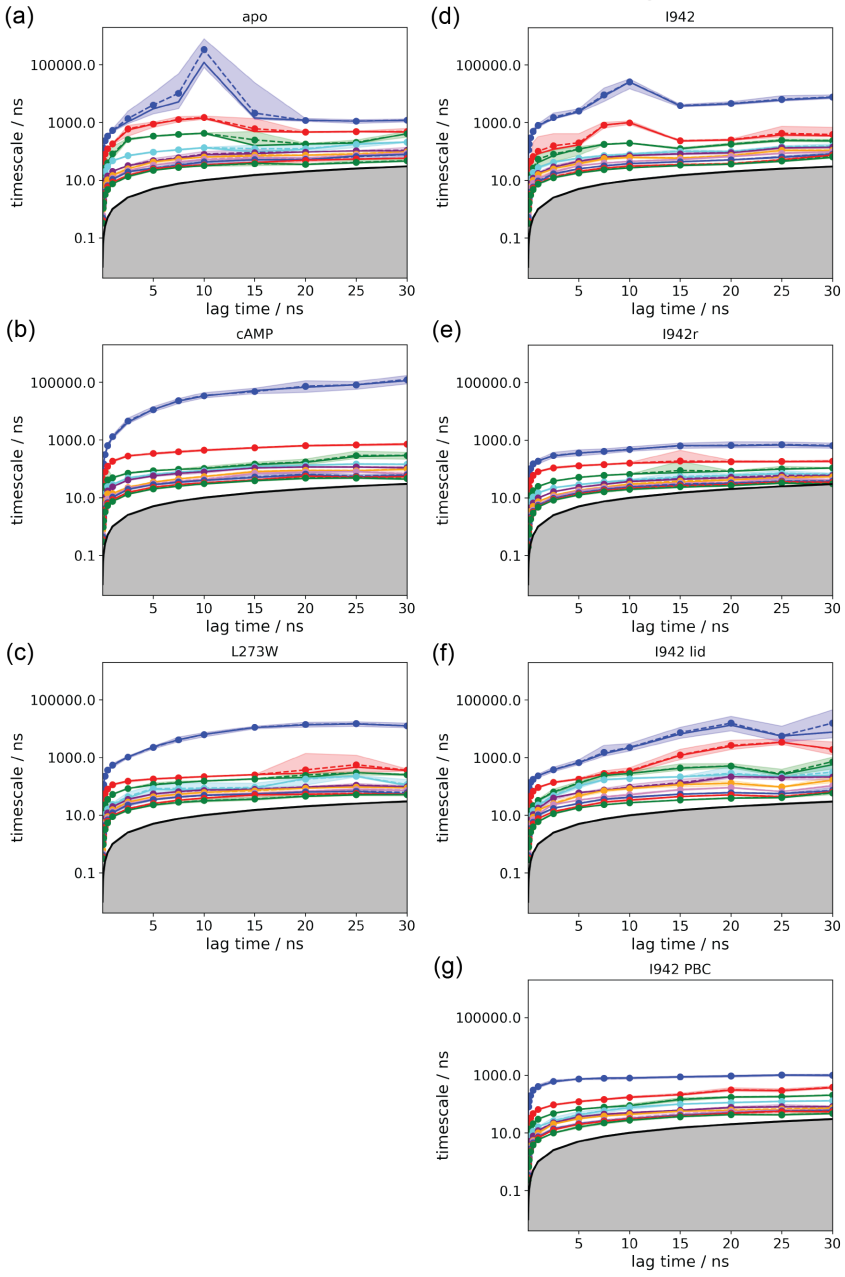

Figure 3.8: ITS plots of the EPAC1 MSMs used in this chapter: (a) *apo* (b) EPAC1-cAMP (c) $EPAC1_{L273W}$-cAMP (d) EPAC1-I942 (e) EPAC1-I942 restrained to both PBC and lid (f) EPAC1-I942 restrained to lid only (g) EPAC1-I942 restrained to PBC only

pool was performed for each system, keeping the micro- and metastable states the same as in the original MSM building (Figure 3.6c). The resulting state probabilities are shown in Figure 3.9.

In agreement with experimental results, *apo* EPAC1 is modelled as primarily adopting the inactive conformation (61%, IQR 55-67%), some intermediate conformation (38%, IQR 32-44%), and no active conformation (0%, IQR 0-0%) (Figure 3.10). These values are reflected in the stationary probability distribution shown in Figure 3.9, where the majority of the density is seen in the low hinge and PBC RMSD values, and none in the high value region. In the presence of cAMP, the active state probability increases to 38% (IQR 12-66%), demonstrating that the sMD/MSM approach applied here can capture activation by cAMP. Additionally, the cAMP stationary probability density in Figure 3.9 shows an increase in probability density in the high regions of hinge and PBC RMSD values, corresponding to the active state.

To further analyse the conformational changes induced by cAMP, 10,000 frames were sampled from the seeded MD data pool of *apo* EPAC1, as well as EPAC1-cAMP, according to the MSM state probabilities. The re-weighted ensemble confirms that cAMP maintains hydrogen bond interactions with the phosphate binding cassette (PBC) throughout the seeded MD simulations, as shown in Figure 3.11a. The distance between these two ends of the PBC, shown in Figure 3.11c as distance *d*, can be used to monitor the "in"/"out" conformation of the PBC. *Apo* EPAC1 shows large fluctuation between the two conformations, with the distance ranging between 8 Å and 11 Å (Figure 3.11b). This agrees with the high population of the intermediate state seen in *apo* EPAC1 (Figures 3.9 and 3.10). When cAMP is present, the distance distribution is much narrower with a median value of 8.5 Å, as the PBC is pulled into the "in" conformation (Figure 3.11b and c).

While a small change, this shift opens up the space between the hinge and the PBC, moving L273 out and up, and allowing F300 of the hinge to take its place. The hinge "melts" at the C-terminus, swinging the lid region towards the cAMP

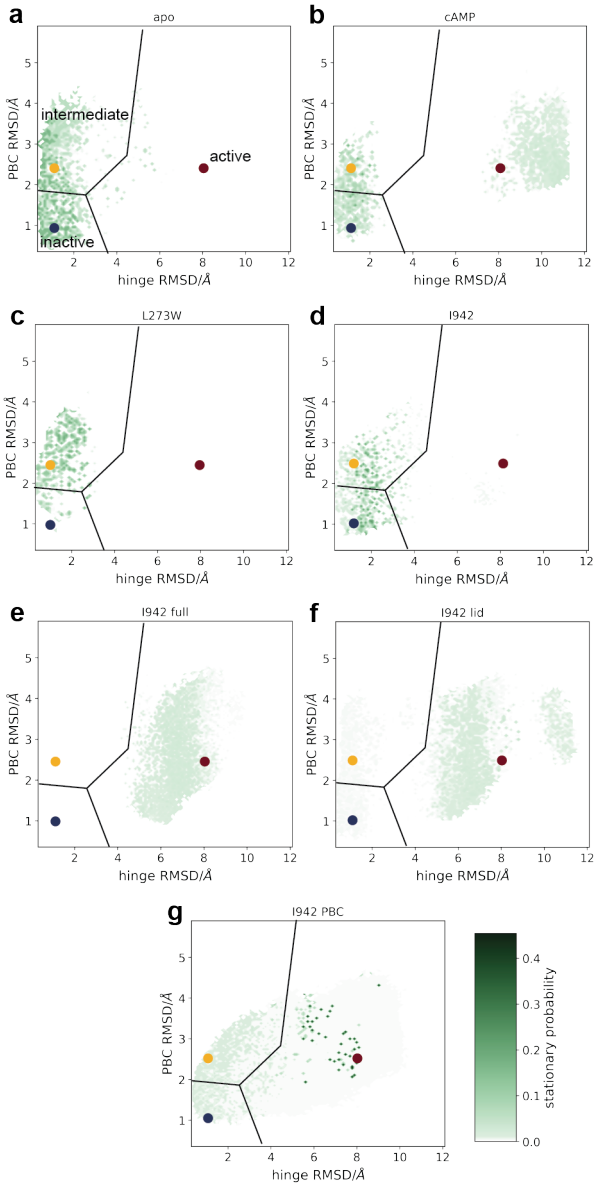

Figure 3.9: The equilibrium distribution density map in the hinge and PBC RMSD to inactive conformation space for each system: (a) *apo* EPAC1, (b) EPAC1-cAMP, (c) EPAC1$_{\mathrm{L273W}}$-cAMP, (d) EPAC1-I942, (e) EPAC1-I942 with I942 restrained to the lid and PBC regions, (f) EPAC1-I942 with I942 restrained to the lid region, and (g) EPAC1-I942 with I942 restrained to the PBC. On top, the macrostate centers and the resulting space partitioning are depicted: inactive (blue), intermediate (yellow) and active (red).

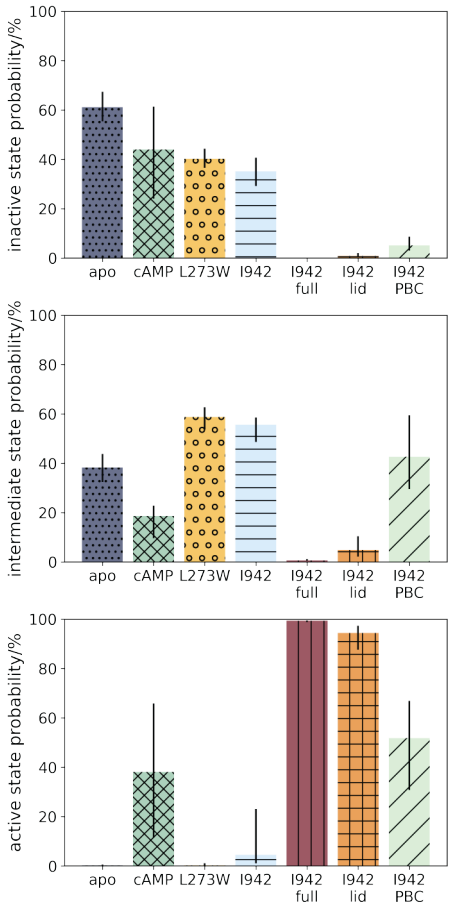

Figure 3.10: The bootstrapped probabilities of the inactive, intermediate and active EPAC1 states for each system described in this work. The bar values are the medians values of the boostrapped probability distributions, while the black vertical bars show the inter-quartile range between quartiles 1 and 3.

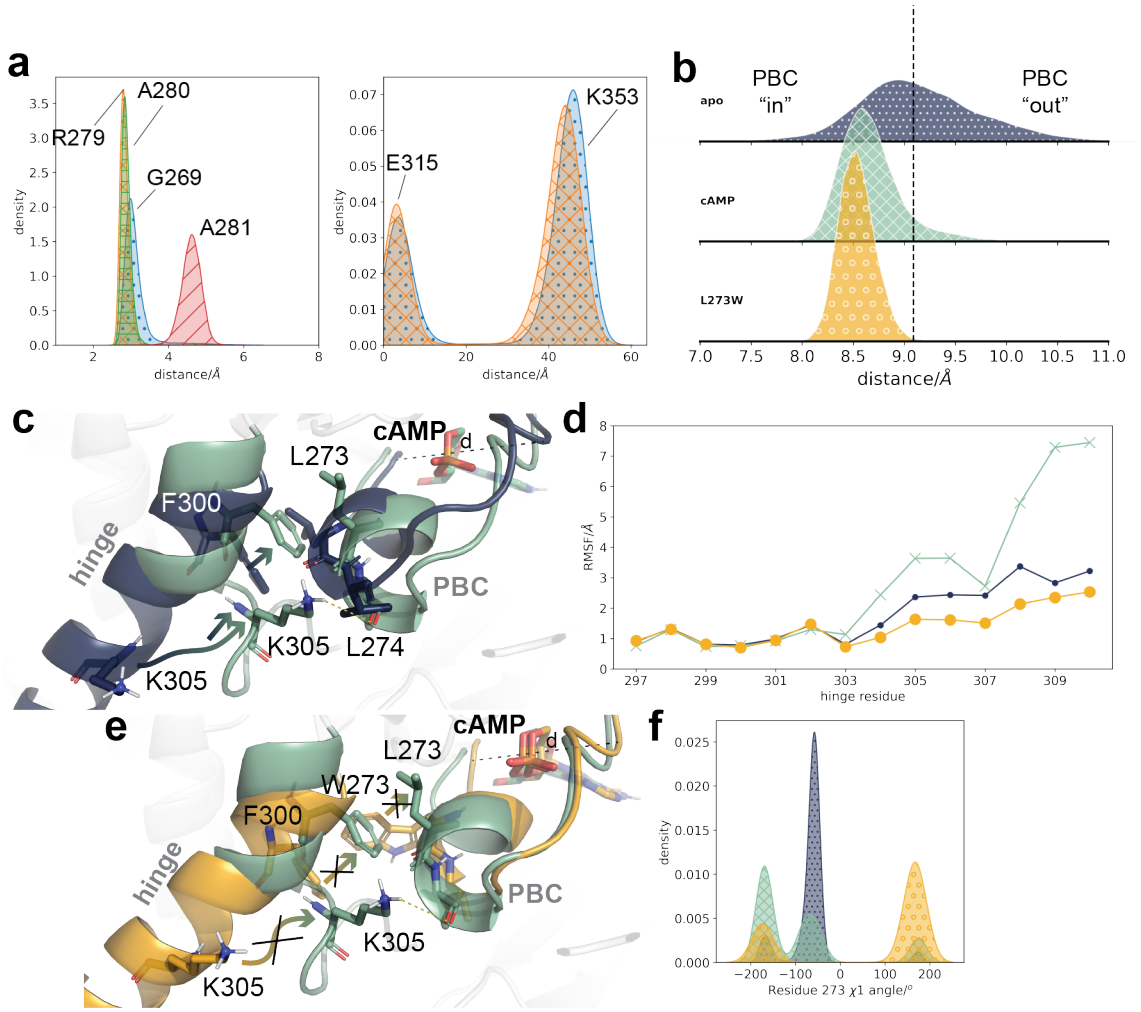

Figure 3.11: Comparison of conformational ensembles in *apo* EPAC1, EPAC1-cAMP, and EPAC1$_{L273W}$-cAMP. (a) The distribution of hydrogen bond distances shown in A. Left: the PBC region, G269 (blue dots), R279 (orange crosses), A280 (green horizontal lines), and A281 (red diagonal lines). Right: lid region, K353 (blue dots) and E315 (orange crosses). In the lid region, the large distances correspond to the inactive/intermediate conformations when the lid region is positioned away from cAMP. (b) The backbone atom distance between G269 and residues 279-281 in *apo* EPAC1 (blue dots), EPAC1-cAMP (green crosses), and EPAC1$_{L273W}$-cAMP (yellow circles). The median distribution value of *apo* EPAC1 is shown in a black dashed line. (c) The conformations of the hinge and the PBC in *apo* EPAC1 (dark blue) and when cAMP is present (green). Residue rearrangements are indicated by arrows. (d) The RMSF of each of the hinge residues (to the inactive conformation) for *apo* EPAC1 (small dots, dark blue), EPAC1-cAMP (crosses, green), and EPAC1$_{L273W}$-cAMP (large dots, orange). (e) The hinge and PBC conformation observed in WT EPAC1-cAMP (green) and EPAC1$_{L273W}$-cAMP (yellow). (f) The residue 273 χ1 angle in *apo* EPAC1 (dark blue dots), with cAMP present (green crosses) and the L273W mutant (yellow circles).

binding site. The change in the C-terminus of the hinge conformation is shown through the root-mean-square-fluctuations (RMSF) of each residue in Figure 3.11d. Residues 303-309 show increasingly higher fluctuations in the cAMP system. Figure 3.12 shows the distance between D750 and Q168, one residue pair making up the ionic latch interactions outlined in Figure 3.2, which lock the catalytic and regulatory regions together near the RAP binding site. There is a large increase in this distance observed when cAMP is present, where a peak at 40-60 Å corresponds to the active conformation. The higher energy conformation of the disordered hinge is stabilized by K305 reaching across and forming a hydrogen bond to the L274 of the PBC (Figure 3.11c, Figure 3.13). Additionally, when EPAC1 is activated, K353 and E325 residues of the "lid" region[159] are positioned near cAMP, creating a ligand-mediated hydrogen bond network between the catalytic and regulatory regions (Figure3.1c, Figure 3.11a). Therefore, cAMP appears to act in a dual fashion, both inducing the activation of EPAC1 and then stabilizing the active state, once it is reached. The importance of each action is investigated below by looking at the L273W mutant and I942.

3.3.3 The L273W Mutant Populates the Intermediate State

The L273W mutant was shown to prevent activation of EPAC1, even in presence of cAMP. It has been proposed that this mutation changes the interaction between L273 and F300 and stabilizes the helical structure of the hinge[160]. Therefore, the sMD/MSM protocol outlined above was applied to the L273W EPAC1, in complex with cAMP. In accordance with the Shao *et al.* interpretation[162], the hinge does not melt at its C-terminus, and the L273W mutant shows a 0% (IQR 0-1%) active state probability, even with cAMP present (Figure 3.10). Instead, the inactive and intermediate states are similarly populated (40%, IQR 37-44%, and 59%, IQR 54-63% respectively).

Comparing the conformation of the PBC in the MSM probability-weighted conformational ensemble shows that the PBC is slightly further "in" than in the WT

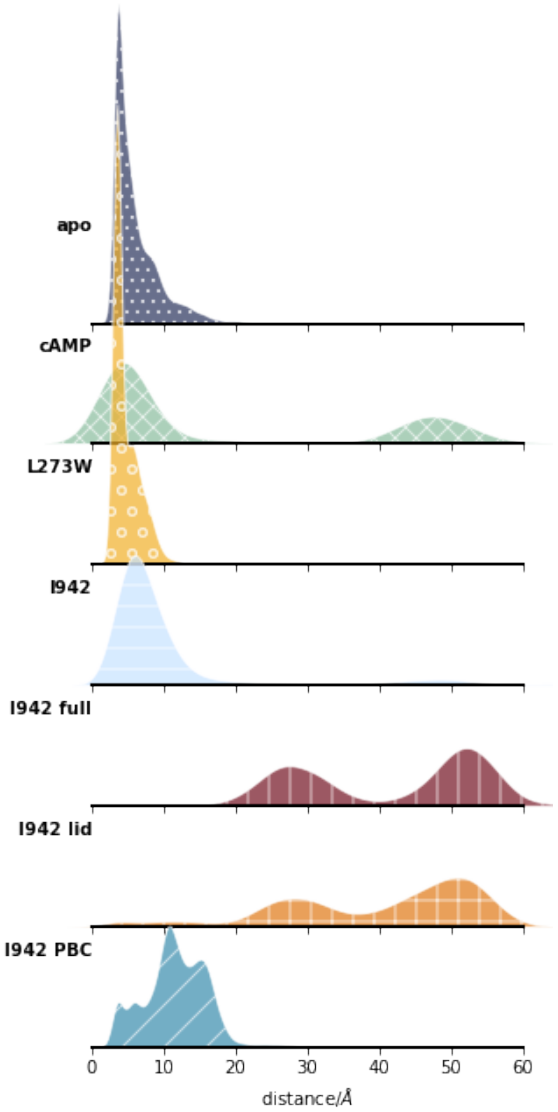

Figure 3.12: The distance between ionic latch residues D750(CG) (catalytic region) and Q168(NE2) (regulatory region), in all systems described in this chapter.

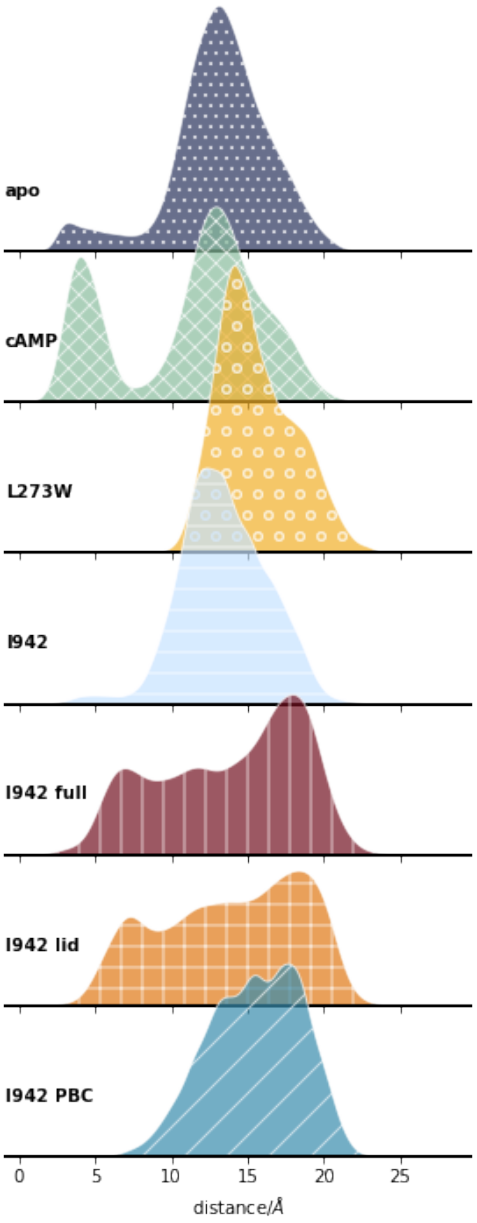

Figure 3.13: The K305(NZ) distance to L274(O), representing the hydrogen bonding between these two residues, for all systems discussed in this chapter.

cAMP complex (Figure 3.11b). However the mutant W273 does not adopt the same conformation as L273 in the presence of cAMP, indicated by the $\chi 1$ angle shown in Figure 3.11f. The larger tryptophan side-chain is too sterically hindered to move upwards, and instead remains in the space between the hinge and the PBC, shown in Figure 3.11e. This allows it to π-stack with F300, which in turn further stabilizes the helical conformation of the hinge. This causes the $EPAC1_{L273W}$-cAMP complex to favour the intermediate state - cAMP still acts on the PBC, pulling it into the "in" conformation (as well as the large W273 pushing it out), but the change in interaction between W273 and F300 means that the "out" hinge conformation is even further stabilized. In Figure 3.11d, the L273W hinge fluctuations are lower than even in *apo* EPAC1. This increased stabilization of the hinge also allows for stronger ionic latch interactions at the RR/CR interface, shown in Figure 3.12. The "inactive" microstates highly populated by this system are observed at higher PBC RMSD values (Figure 3.9a), which is caused by the larger W273 residue side chain separating the PBC and hinge regions further.

In the $EPAC1_{L273W}$-cAMP complex, cAMP maintains the same protein-ligand interactions as observed in the WT EPAC1-cAMP system (Figure 3.14). The hydrogen bonds to the "lid" region are not present in the full MSM weighted ensemble, as the active state is required to position the residues in place for the hydrogen bond interactions. To ensure that, when the system was in the active state, these stabilizing interactions were possible, only the active state ensemble was recreated using the MSM probabilities. The active state ensemble of $EPAC1_{L273W}$ shows that the stabilizing hydrogen bonds are indeed prevalent in these conformations (Figure 3.14). This confirms that the L273W mutation does not change the cAMP binding pose, but instead changes the hinge-PBC interactions, preventing activation by stabilizing the inactive hinge conformation.

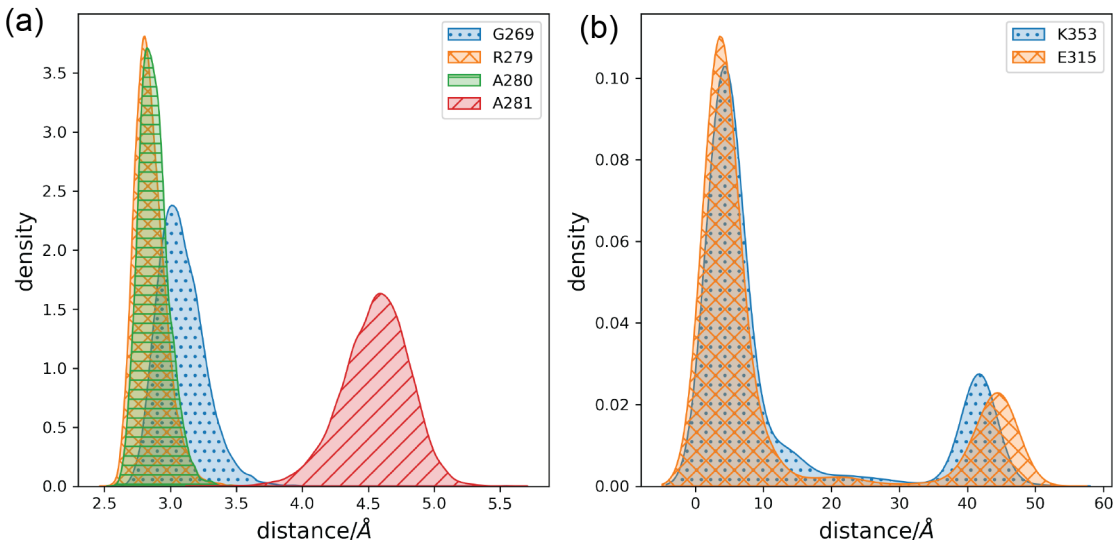

Figure 3.14: The hydrogen bond interactions of cAMP with EPAC1$_{\mathrm{L273W}}$. (a) hydrogen bonds to the PBC, in the full conformational ensemble of EPAC1$_{\mathrm{L273W}}$-cAMP. (b) hydrogen bonds to the lid region, in the active state conformational ensemble of EPAC1$_{\mathrm{L273W}}$-cAMP. As the active state population is very low in the general ensemble, the lid interactions are not apparent there, so only the active state is shown here.

3.3.4 I942 Does Not Induce the Same Conformational Changes in the PBC and Hinge Regions

I942 was shown to be a partial activator of EPAC1[173], and proposed to stabilize an intermediate conformation, that lies between active and inactive EPAC1[162]. Here we model EPAC1 with I942 as 5% (IQR 1-23%) active and 56% (IQR 49-59%) intermediate (Figure3.10). The slight increase in active state accounts for the partial activation observed experimentally, and there is also an increase in the intermediate state probability compared to the 38% observed with *apo* EPAC1. The partial activation of EPAC1 by I942 is also reflected by the distances between ionic latch residues of the catalytic and regulatory regions - the hydrogen bonding is interrupted, but the distance is only slightly larger than observed in *apo* EPAC1, and does not reach the 40-60 Å range observed with cAMP (Figure 3.12). While I942 forms hydrogen bond interactions with residues 279-281, similar to cAMP, the interaction with G269 is not always observed (Figure 3.15a). Consequentially, the distance between the two ends of the PBC remains higher than observed with cAMP (Figure 3.15b), corresponding to a more "out" conformation. Figure 3.15c compares the conformations of the PBC and the hinge with cAMP and I942, and illustrates how the gap required to accommodate the disordered hinge C-terminus between the hinge and the PBC is not present, preventing full activation of EPAC1. This is also reflected in the RMSF values of hinge residues 303-309 in EPAC1-I942, which show an increase in fluctuation but do not reach the same levels as cAMP (Figure 3.15d). Additionally, the plain napthyl group on I942 cannot form the same active state stabilizing interactions with K353 and E315 of the lid region. Therefore I942 fails both to fully activate EPAC1 and to stabilize the active state once it is populated.

In order to investigate the importance of the protein-ligand interaction differences between I942 and cAMP outlined above, I942 was remodelled with very weak distance restraints applied to I942-K353 (lid region) and I942-G269 (PBC), to mimic the hydrogen bonds seen with cAMP (Figure 3.3). Although the restraints to the lid

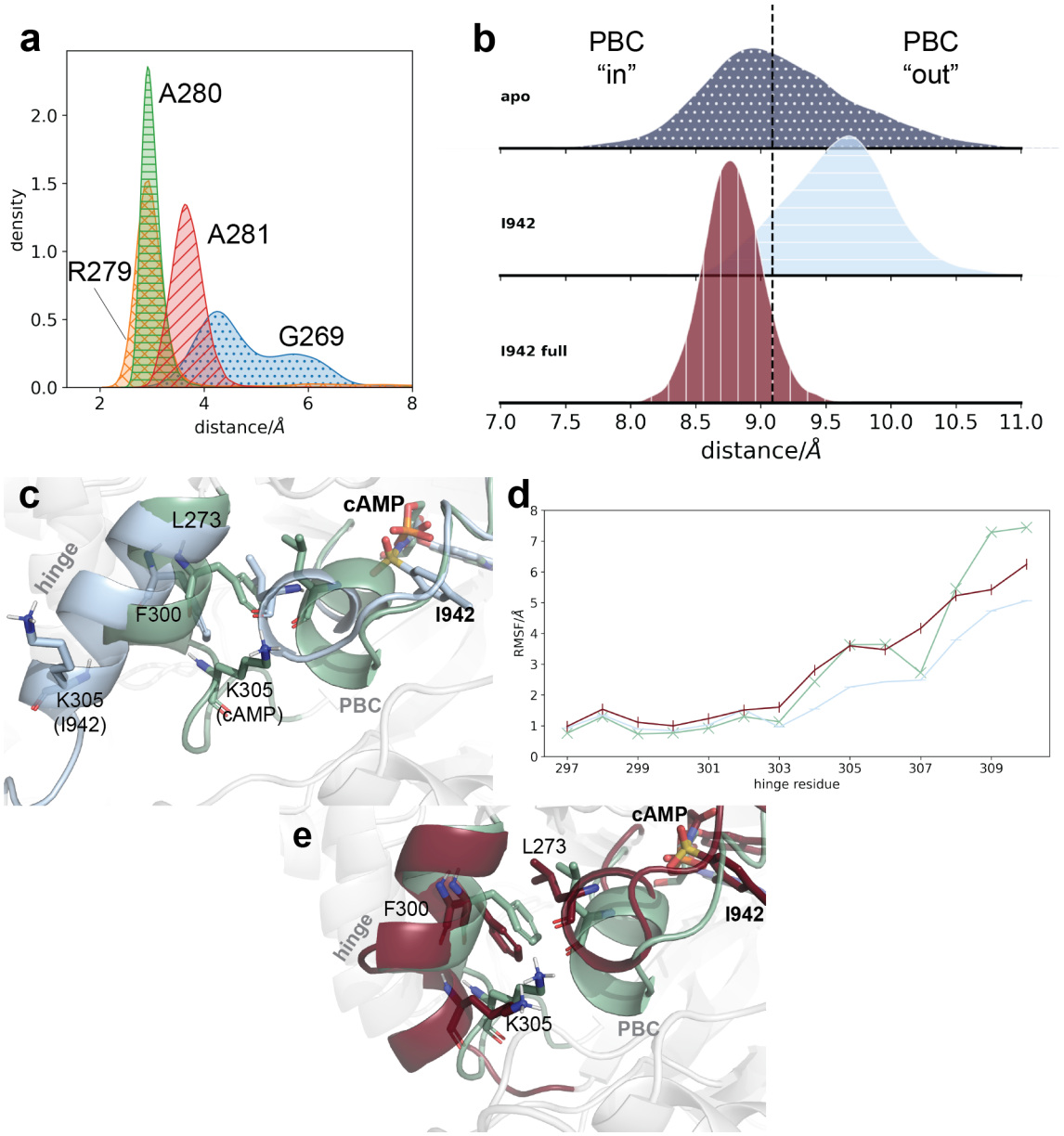

Figure 3.15: Interactions of unrestrained and restrained I942 with EPAC1. (a) The hydrogen bond distances between unrestrained I942 and G269 (blue dots), R279 (orange crosses), A280 (green horizontal lines), and A281 (red diagonal lines). (b) The backbone atom distance between G269 and residues 279-281 in *apo* EPAC1 (blue dots), EPAC1-I942 (blue horizontal lines), and EPAC1-I942 with I942 restrained to the lid region and PBC (red vertical lines). The median distribution value of *apo* EPAC1 is shown in a black dashed line. (c) The conformations of the hinge and the PBC in EPAC1-cAMP (green) and EPAC1-I942 (light blue) conformational ensembles. (d) The RMSF values of hinge residues, for EPAC1-cAMP (green, crosses), EPAC1-I942 (light blue, horizontal lines, and EPAC1-I942 restrained (red, vertical lines). (e) Representative conformations of the hinge and the PBC in conformational ensembles of EPAC1-cAMP (green) and EPAC1-I942, when I942 is restrained to the lid region and the PBC (red).

region directly stabilize the active state, they did not affect the sMD simulations, where the steering force was much stronger than the restraint. All restraints are summarized in Table 3.2. With these restraints in place and following the same protocol as above, the modelled active state probability of EPAC1-I942$_{\text{restrained}}$ increased to 99% (IQR 99-100%) (Figure 3.10b). The complete displacement of the regulatory region is also evident in the IL distance, where only distances above 20 Å are observed (Figure 3.12). The I942-G269 restraint was sufficient to decrease the distance between the N- and C-termini ends of the PBC, bringing it into the "in" conformation (Figure 3.15b). The increased space between the hinge and the PBC also allowed for greater disordering of the hinge, as reflected in the higher hinge RMSF values (Figure 3.15d). It is worth noting that the active state probability with I942$_{\text{restrained}}$ is significantly higher than with cAMP. This is due to the difference between a constant artificially introduced distance restraint used for I942 and the force field driven hydrogen bond interactions observed with cAMP. Since EPAC1 has to be in the active conformation for K353 of the lid region to interact with the ligand bound at the PBC, the two restraints applied to the distances between K353 and I942 directly stabilize the active conformation by $\sim$ 8 kcal mol^{-1}. This results in the large jump in active state probability seen when these restraints are applied. Additionally, while the cAMP active state probability is indeed lower, the hinge RMSD in the probability density plots (Figure 3.9a) and hinge RMSF values (Figure 3.15d) are higher than even when I942 is restrained.

As the cAMP interaction mimicking distance restraints showed strong positive effect on I942 activating EPAC1, further I942 models were constructed, this time applying only one restraint each. When only the I942-K353 restraints were applied, the active state probability of EPAC1 was 94% (IQR 88-97%) (Figure 3.10), which aligns with the direct active state stabilizing effect this interaction has. The ionic latch distances are also very similar to when both restraints are applied (Figure 3.12). As the I942-G269 as not applied in this case, the PBC still populates more of the "out" conformation (Figure 3.16a). However, the PBC N- to C-terminus

distance decreases even without a direct restraint - potentially the lid stabilization of the active state is so potent that the hinge is pulled into position and instead pushes on the PBC slightly.

Interestingly, only applying the I942-G269 restraint increased the active state probability from 5%, seen with unrestrained I942, to 52% (IQR 31-67%). The PBC behaviour in this case mostly corresponds to the "in" conformation, indicating that this interaction was indeed the missing link to fully shift the PBC conformation. This destabilized the hinge sufficiently to allow further destabilization of the ionic latch, as the distances observed are higher than observed in *apo* EPAC1 and EPAC1-I942, but not as high as when the active state stabilizing interactions with the lid region are present (Figure 3.12). The hinge RMSF values are similarly high when either of the restraints is applied (Figure 3.16b). However the lower active state probability when only the PBC is restrained (Figure 3.10) and the absence of K305-L274 hydrogen bonding seen in EPAC1-cAMP (Figure 3.13) suggests that the PBC shift from "out" to "in" allows space for a disordered hinge conformation, but the lid interactions specifically stabilize the active conformation. With the findings above, we propose that the most effective modification of I942 to develop it into a full activator would be to modify the naphthyl group to include both hydrogen bond donors and acceptors (Figure 3.16e). Additionally, engaging the G269 in addition to residues 279-281 would allow for larger PBC rearrangement and increase EPAC1 active state probability.

3.4 Discussion

In the results outlined above, we have constructed a three state - inactive, intermediate, and active - model of the dynamics of EPAC1. The MSMs reported here correctly capture the activation of EPAC1 by cAMP, as well as the interruption of said activation by the L273W mutation (Figure 3.10). Additionally, the modelled probabilities of states not only allow the quantification of the effects of ligands on the protein conformational ensemble, but can also be used to re-weigh the origi-

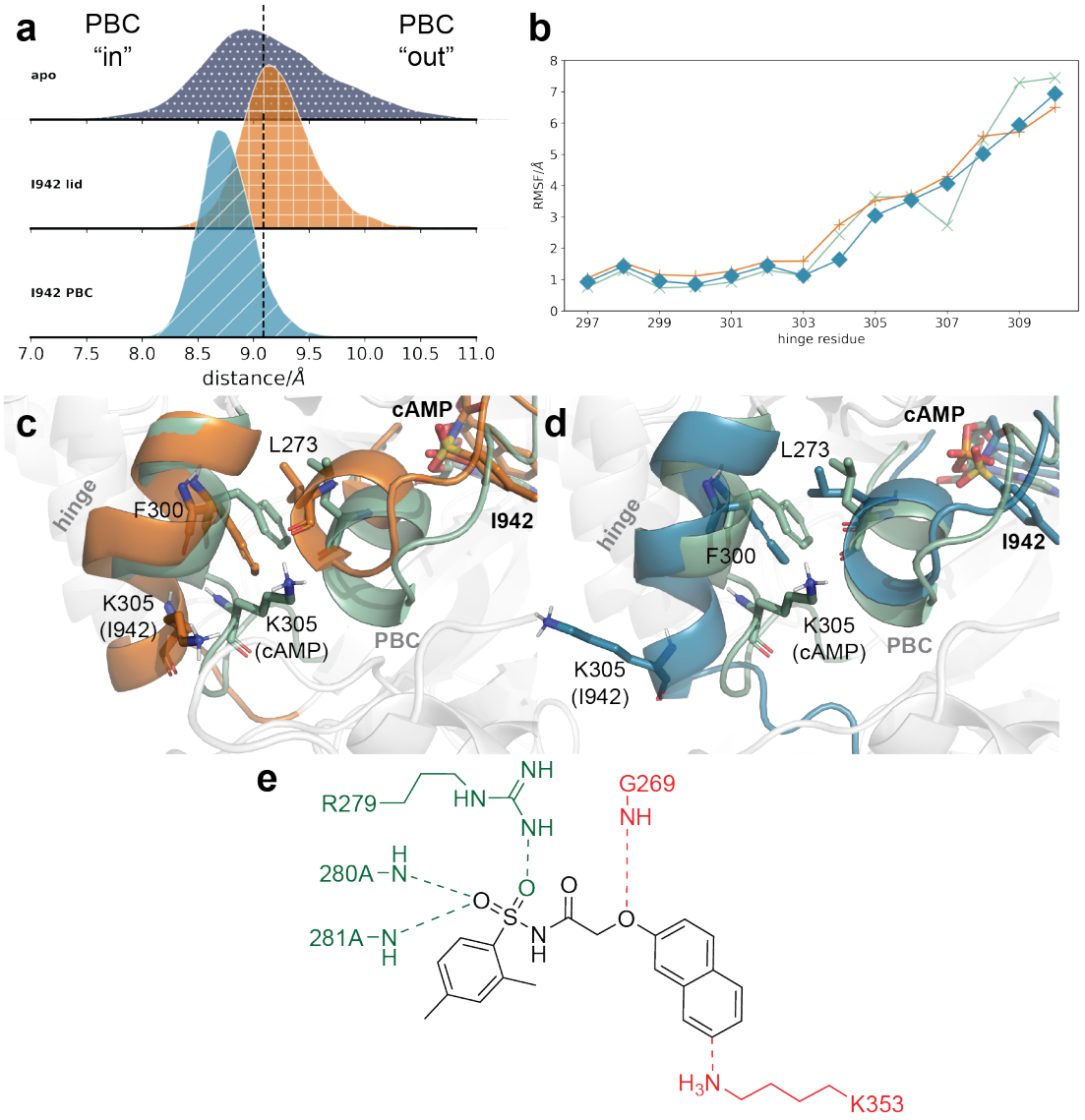

Figure 3.16: Interactions of I942, restrained only to the lid region or the PBC, with EPAC1. (a) The backbone atom distance between G269 and residues 279-281 in *apo* EPAC1 (blue dots), EPAC1-I942 restrained to the lid (orange squares), and EPAC1-I942 restrained to the PBC (teal diagonal lines). The median distribution value of *apo* EPAC1 is shown in a black dashed line. (b) The RMSF values of hinge residues, for EPAC1-cAMP (green, crosses), EPAC1-I942 restrained to the lid (orange, plus), and EPAC1-I942 restrained (teal, diamonds). (c) Representative conformations of the hinge and the PBC in conformational ensembles of EPAC1-cAMP (green) and EPAC1-I942, when I942 is restrained to the lid region only (orange). (d) Representative conformations of the hinge and the PBC in conformational ensembles of EPAC1-cAMP (green) and EPAC1-I942, when I942 is restrained to the PBC only (teal). (e) Interactions of I942 with EPAC1. Interactions that are already present are shown in green (residues 279-281), and potential interactions are shown in red (residues 269 and 353).

nal conformation pool and recreate a statistically representative example of protein dynamics. The dynamics of these ensembles confirm the previously reported dual action of cAMP, that is the change in PBC conformation and stabilization of active state via the lid region[159]. Comparing the WT EPAC1 to EPAC1$_{L273W}$ also confirms that the mutation decouples the PBC from the hinge, rather than preventing cAMP binding[160]. The mutated W273 residue not only does not shift out of the space between the hinge and the PBC to allow for stabilization of the shifted disoredered hinge C-terminus, but also actively stabilizes the helical conformation of the hinge through π-stacking interactions with F300 (Figure 3.11d, f, and g).

Our modelling also confirms the findings by Shao *et al.*[162] on the partial activation of EPAC1 by I942, that is, the presence of I942 increases the probabilities of both the active and intermediate states, compared to the *apo* EPAC1 model. Only a slight "in" conformation of the hinge is observed, and the PBC shift to the "in" conformation is incomplete (compare to that of EPAC1-cAMP complex). The utilization of *in silico* methods and the full-length protein model allowed us to trial imposing protein-ligand distance restraints on I942 to test hypotheses. The restraints to the lid region proved extremely effective in stabilizing the active conformation, yielding a model where the active state was the most populated (Figure 3.10). In this restrained model, no ionic latch interactions were observed, indicating a complete exposure of the RAP binding site (3.12). Such active state stabilizing interactions have been observed in the EPAC2-cAMP(Sp) X-Ray structure[159], and their presence in EPAC1-cAMP was confirmed by the simulations shown here. However, the EPAC1 construct used in the NMR work on comparing the action of I942 to cAMP did not include the lid region[162], and therefore a component of EPAC1 activation could not be evaluated. Further efforts to develop I942 into a full EPAC1 agonist so far have focused on further engagement of the PBC only, as the experimental methodology is limited by the stability of the isolated EPAC1 protein[180]. In contrast, the use of full EPAC1 in this work gives insight into possible steps for lead development of I942, engaging both the PBC and the lid regions.

Having validated the sMD/MSM approach applied here with cAMP, L273W mutant and I942, further modifications of I942 could be tested *in silico* before committing to experimental methods. Furthermore, as our EPAC1-cAMP model does capture the active state stabilization by interactions between cAMP and K353, it serves as a good baseline for developing agonists more potent than cAMP.

It is important to note that the inactive/intermediate state partition is not well defined, as the change in PBC conformation from "out" to "in" is only a value of 1.5 Å. Since the RMSD difference is so small, the distance between G269 and residues 279-281 was used here to describe the change in PBC conformation. However, it is also quite an indirect measure, and other descriptors, such as change in the activator pocket volume could have been used. The subtle change in the PBC also affects the state probabilities seen in the L273W model - while the inactive/intermediate state distribution in Figure 3.10 is quite similar, in the stationary distribution plot the inactive states that are highly populated have higher PBC RMSD values (Figure 3.9). A movement of the PBC towards the "in" conformation is required in order to accommodate the large mutant tryptophan side chain. This becomes less significant when the focus is on developing full activators, as the active/inactive state partition is much clearer. In the cases when a closer look at the state distributions is required, the equilibrium probability distribution plots such as in Figure 3.9 can be relied on for a closer look at the probabilities of the microstates.

Chapter 4

Modelling the Effects of a Gain-of-Function Mutation and PIP_2 Lipids on PKD2

4.1 Introduction

4.1.1 Ion Channels and the Lipid Bilayer

Every cell in the human body is surrounded by a lipid bilayer membrane. The inter- and intracellular environments, separated by this membrane, have many vary in concentrations of diverse proteins, nucleic acids, carbohydrates, and ions. Typically the concentration of Na^+ and Cl^- ions is higher in the extracellular matrix, whereas the concentration of K^+ ions is higher inside the cell[181]. The difference in concentrations of ions gives rise to a difference in the electrical potential between the cell and its outside, called the membrane potential[182]. Membrane potentials play a significant role in biological processes such as nervous system signalling and muscle contractions. The movement of ions across membranes is controlled in large part by ion channels, a class of transmembrane proteins that form pores in the membranes for ions to travel through via passive transport[183].

Since a large part of the environment of ion channels is the membrane they

are embedded in, the composition of the membrane affects the protein function. Different cells contain different types and proportions of lipids, which is crucial to the type and function of the cell. Cell signalling and metabolic processes can also affect the membrane composition[184]. There are two mechanisms through which protein regulation by membrane lipids occurs: "molecular" interactions between the membrane lipids and specific binding sites on the protein, and "physical" properties of the membrane[185].

Membrane physical properties

Cell membranes, while very thin (up to 5 nm), expand up to hundreds of μm in length. They are not rigid structures, but fluid and dynamic, changing shape as needed[186]. The length of the hydrophobic chains of the membrane lipids dictate the membrane thickness, as well as the size of the protein they can accommodate. A discrepancy between the hydrophobic transmembrane domain and the hydrophobic region of the protein can cause these hydrophobic regions to be exposed to water, or a distortion of the membrane to prevent it (Figure 4.1a). Transmembrane proteins may also cluster to form complexes to minimise the energy cost associated with membrane distortion[184]. Additionally, tightness of membrane packing is crucial to the channel conformation of transport proteins (Figure 4.1b)[187]. Very tight packing density can apply pressure on the protein, physically pushing it towards a closed conformation. Alternatively, a low packing density allows the protein to adopt an open conformation more easily. In return, the packing and aggregation of proteins affects the dynamics of the membrane[188].

Molecular interactions

The aggregation of membrane lipids and transmembrane proteins is not only driven by hydrophobicity, but also protein-lipid interactions. A well-studied lipid regulator example is phosphatidyl inositol bisphosphate ($PI(4,5)P_2$). One example of ion channels regulated by PIP_2 is inward rectifying K^+ (KIR) channels. PIP_2 phosphate

head interacts with the linker between transmembrane helices, inducing an open conformation (Figure 4.1c)[185, 189]. Another example is transmembrane protein 16F (TMEM16F), where PIP_2 plays a role in the activation of the protein. Even after TMEM16F has been desensitized by exposure to high calcium concentrations, protein activity can be recovered by introduction of PIP_2[190].

4.1.2 Transient Receptor Potential Protein Family

Transient receptor potential (TRP) channels are so named after their first dicovered mutation, which caused transient photoreceptor response in *Drosophila* flies[191]. They are Ca^{2+} voltage-gated ion channels mostly located in the plasma membrane of the cell. The TRP ion channels are separated into seven subfamilies of proteins, which are grouped into two groups based on sequence and structure. Group 1 of TRPs contains TRPC (canonical), TRPV (vanilloid), TRPM (melastatin), TRPN (NO-mechanopotential) and TRPA (ankyrin), while group 2 contains TRPP (polycystin) and TRPML (mucolipin)[192]. They are involved in various sensory pathways[192, 193], as well as homeostatic processes and muscle contraction[194].

Proteins of the TRP channel family have a similar general structure, which includes six transmembrane helices (S1-S6), and a smaller inter-pore region between helices S5 and S6. The channels primarily form homotetrameric structures, but have also been observed as heterotetramers[194]. Helices S1-S4 compose the membrane-adjacent voltage sensor-like domain, while helices S5 and S6 make up the pore domain. These two parts of the channels are connected by a linker between helices S4 and S5. The permeation of ions through TRP proteins is determined by two main structural features of the channel pore: the selectivity filter and the lower gate.

The selectivity filter in most TRP channels contains a negatively charged aspartic acid residue. It is responsible for cation selectivity and anion repulsion from the channel. Mutating these residues in TRPV6[195] and TRPM4[196] caused changes in ion selectivity, as well as increasing current conductance, indicating that they can play a significant role in gating certain TRP proteins. Towards the C-terminus

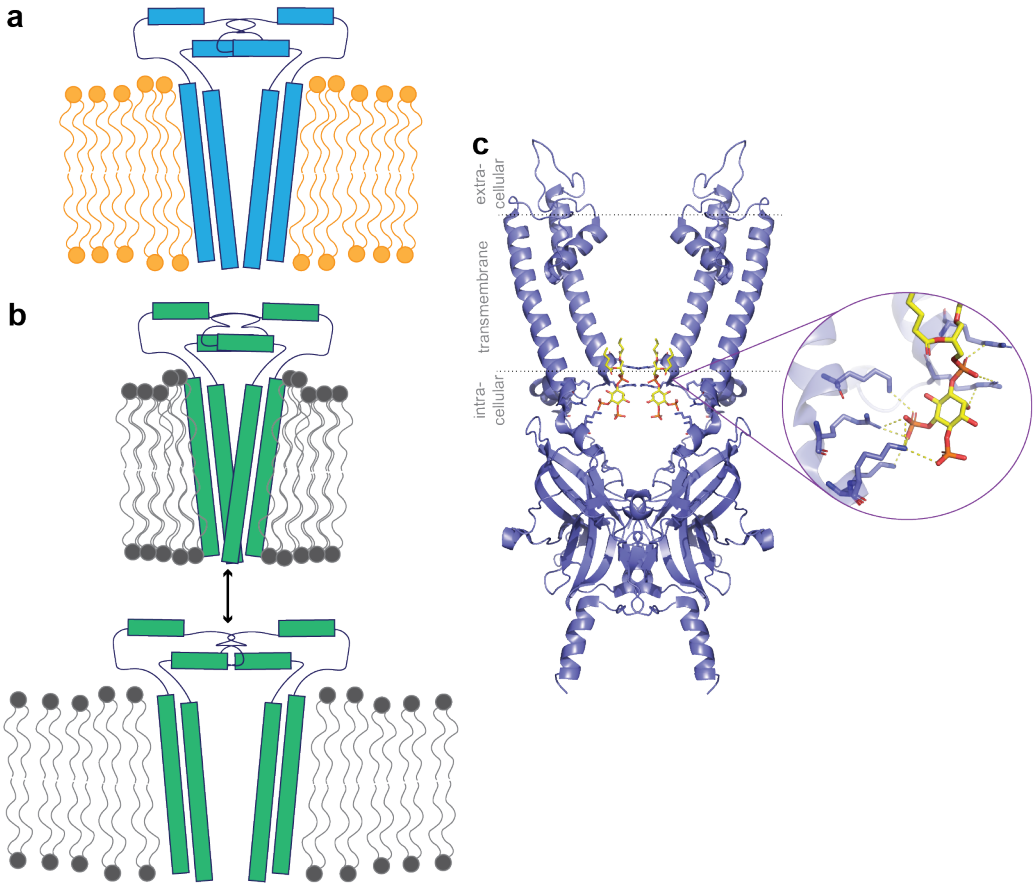

Figure 4.1: Examples of transmembrane protein regulation by membrane lipids. (a) Membrane thickness distorts to cover the hydrophobic transmembrane regions of proteins. (b) Membrane lipid packing density can affect the conformation of transporter proteins. (c) KIR2 protein interaction with PIP_2 (PDB ID 3SPI). The dotted lines indicate the the boundaries between the extracellular, transmembrane, and intracellular regions.

of the S6 helix, hydrophobic residue side chains extend into the pore, forming the lower gate. The opening and closing of this gate creates a physical barrier that controls the permeation of ions through the channels[197]. For instance, replacing hydrophobic L557 and A558 with hyrophilic asparagine residues increased channel activity as well as the population of an open conformation of TRPP3[198].

4.1.3 Polycystic Kidney Disease 2

Autosomal dominant polycystic kidney disease (ADPKD), as the name implies, involves the formation of cysts primarily in the kidneys, and can lead to renal failure or cyst infection[199]. ADPKD has been linked to the proteins of the TRPP subfamily: TRPP1 and TRPP2, also known as polycystic kidney disease 1 and 2 (PKD1 and 2). Decreased expression of PKD2, the protein described in this chapter, was shown to lead to hepatic cyst formation in mutant mice[200]. Additionally, mutations truncating large portions of PKD1 and PKD2 were observed in ADPKD patients[201]. Therefore, restoration of normal PKD2 function can lead to potential treatment to prevent further cyst formation and progression of ADPKD.

Similar to other proteins of the TRP family, PKD2 is a homotetramer, composed of six transmembrane helices and an extracellular TOP domain. The pore formed by the aggregation of the four monomers contains the characteristic selectivity filter (D643, L641, and F669), and the lower gate (L677 and N681)[202], all shown in Figure 4.2a. Mutations of lower gate residues identified that L677 forms the key hydrophobic gate regulating ion permeation through PKD2[198]. The significance of the lower gate was also illustrated by the F604P gain-of-function (GoF) mutation. Cryo-EM structures of the mutant revealed distortion of the S5, and in turn S6, helices at the lower gate, resulting in significantly increased conductance[198]. As mutations to alanine and isoleucine did not induce the same GoF effect, it is the α-helix distorting properties of proline that cause this change in channel conformation (Figure 4.2b)[203]. Additionally, the F604P/L677G mutant shows cooperation between the two mutations, exhibiting even further increase in activity[198].

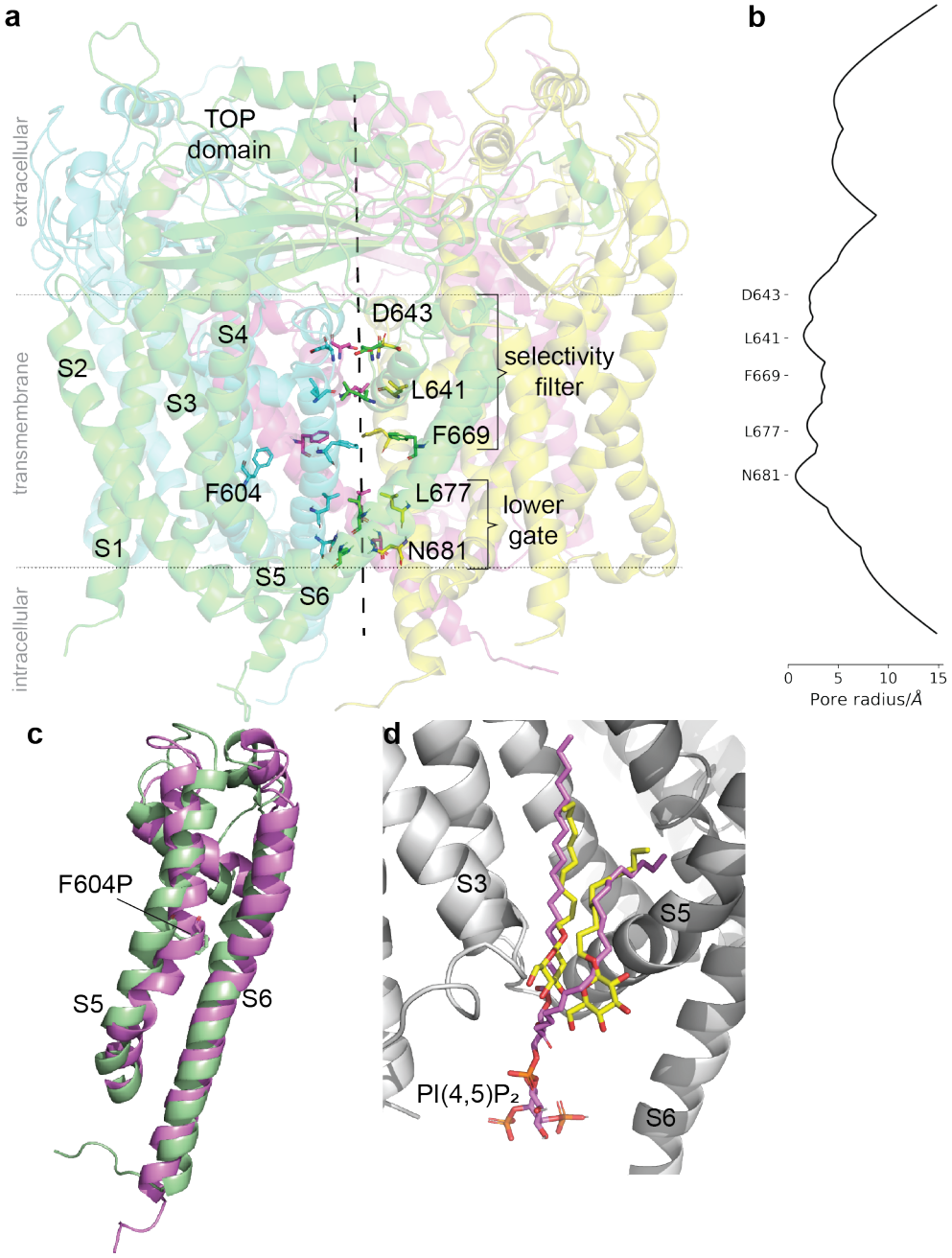

Figure 4.2: Structure of PKD2, GoF mutation and PIP_2 model. (a) The structure of PKD2 (PDB ID), with the 4 monomers shown in green, yellow, blue, and magenta. The relative selectivity filter and lower gate residues are also shown. Residue F604 is shown for one monomer in blue. The protein pore is indicated by a black dashed line. The dotted lines indicate the the boundaries between the extracellular, transmembrane, and intracellular regions. (b) The radius of the pore of PKD2 (PDB ID 5MKE), generated using the Hole software[206]. Key residues shown in (a) are highlighted. (c) The distortion of the S5 and S6 helices of $PKD2_{F604P}$ (magenta), compared to WT PKD2 (green). Structures are taken from MD simulations described in this chapter. (d) $PI(4,5)P_2$ modelled binding pose (pink), overlayed on the detergent molecules (yellow) placed in the PKD2-$PI(4,5)P_2$ cryo-EM structure (PDB ID 6T9N).

As outlined above, PIP_2 is a common transmembrane protein regulatory lipid. It also has been shown to have an effect on PKD2, acting as an inhibitor. The infusion of PKD2 enriched cells with PI(4,5)P_2 resulted in the reduction of observed current. Interestingly, PI(3,4,5)P_2 did not have any effect on the measured current[204]. Understanding the inhibitory mechanism of PIP_2 would lead to better understanding of allosteric modulation of PKD2, which in turn could be applicable to developing small molecule PKD2 modulators. More recently, cryo-EM structures of PKD2 with PI(3,5)P_2 and PI(4,5)P_2 have been resolved (Figure 4.2c), both showing a closed channel conformation, and interaction of PIP_2 lipids with PKD2 have been investigated, but the effects of lipid binding on the channel conformation were not extensively studied[205].

In this chapter, a truncated model of PKD2 and the sMD/MSM workflow are initially validated by capturing the effects of a GoF mutation F604P. The protocol is then applied to model inhibition by PI(4,5)P_2, and the lipid-protein interactions are investigated to identify key residues involved in modulation of PKD2 by PI(4,5)P_2. These results provide a new insight into the potential regulation of PKD2, which could be further applied towards drug design.

4.2 Methods

4.2.1 Membrane Building and System Setup

All systems with the channel in the open conformation were drived from protein coordinates from PDB ID 5MKE, and all systems with the closed channel were prepared from PDB ID 6T9N. Missing residues were modelled using the "LoopModel" class in MODELLER 10.1[207]. Residues 213-700 were modelled in the "full model" of PKD2 discussed in this chapter, and only residues 469-700 (excluding helix S1 and the extra-cellular TOP domain) were modelled in the truncated variant. The truncated variant was prepared to reduce system size and save on computation time. A disulfide bridge was modelled between residues 331 and 344, and E491 was mod-

elled as protonated, based on propka3[176] modelled pK_a value of 8.97. Although structures with Ca^{2+} ion-bound PKD2 are available, no Ca^{2+} ions were modelled in this case, to isolate the effects of the mutation and lipid binding described here.

Membrane building, system solvation and parameterization were carried out using the "Membrane Builder" functionality in CHARMM-GUI[208] in all cases. Protein structures were pre-aligned against the Orientations of Proteins in Membranes (OPM) database[209] entry for PDB ID 5MKE, and no additional protein alignment was carried out within CHARMM-GUI. Pore waters were generated using a cylindrical radius of up to 15 Å. A homogenous 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid bilayer was generated, with an XY extent of 140 Å for the full protein model and 135 Å for the truncated model (1:1 upper:lower leaflet ratio), and a 22.5 Å hydration layer was added to the top and bottom of the simulation box. The system was neutralized with 0.15 mM KCl salt. The force field used for all system atoms was the CHARMM36m all atom force field. The final system sizes for the full and truncated models were 305k and 225k atoms respectively.

After all inputs were generated, minimisation step and 6 equilibration steps were carried out with AMBER22[139] following the protocol provided by CHARMM-GUI, gradually removing restraints on the protein and the membrane lipids. Positional restraints were applied to all protein and membrane lipid P atoms (the keyword *ntr*=1 was used). Additionally, flat bottomed restraints were applied to each membrane lipid C1-C2-C3-O21 dihedral angle (keyword *nmropt*=1). The flat bottom bounds *r2* and *r3* were set as -122.5° and -117.5° respectively, while the parabolic restraint limits *r1* and *r4* were set as -132.5° and -107.5°. Forces at each step are given in Table 4.1. The systems were energy minimised with 2500 steepest descent steps, followed by 2500 conjugate gradient steps. Particle Mesh Ewald (PME) was used with a 12.0 Å non-bonded cutoff and a 10.0 Å force switching cutoff. All equilibration simulations were carried out at 300 K, using Langevin dynamics with a 1.0 ps^{-1} friction coefficient. The SHAKE algorithm was applied in all cases. Equilibration steps 1 and 2 were run in the NVT ensemble, and steps 3-6 were run in the

NPT ensemble, using the Monte Carlo barostat and semiisotropic pressure scaling. Surface tension was constant with interfaces in the *xy* plane. An additional NPT 50 ns equilibration simulation was performed with no restraints on the system.

F604P Mutant

For both the open and closed conformations of the F604P mutant, F604 in each of the domains of the previously modelled PKD2 structures was mutated to proline using Flare software[134].

PI(4,5)P_2 Modelling

To insert PI(4,5)P_2 into PKD2, the PDB ID 6T9N was used. In this structure, lipid-like density at a potential phospholipid binding site was observed, but it was not definitively interpreted as PI(4,5)P_2 and instead two detergent molecules were placed there for the PDB entry[205]. These two detergent molecules were used as alignment templates for a PI(4,5)P_2 structure in Flare. Once the lipid was aligned to one domain, that domain was aligned to the remaining three domains and the lipid coordinates copied. Although the PIP_2 structures available in the PDB have fully saturated 8-carbon carbonyl chains (PDB ID PIO), the structure used in this chapter is one with the 1-stearoyl-2-arachidonoyl (18:0/20:4) fatty acid chain, which is most common in mammals[210] and is the one available in CHARMM as a membrane lipid.

A CHARMM residue topology file (RTF) was prepared, using the PI(4,5)P_2 parameters from CHARMM36 force field (residue name in CHARMM - SAPI24). When preparing the system with CHARMM-GUI, these parameters were provided instead of re-parameterizing the lipid. After the system was built and equilibrated via CHARMM-GUI, an AMBER topology input was generated using ParmEd[211]. Box dimension and solvent pointer flags were added manually, and the periodic box conditions turned on (topology flag IFBOX=1). The further AMBER minimisation and equilibration steps were followed as outlined above.

Step	**Duration**	**Protein positional restraint force /kcal mol^{-1} Å^{-2}**	**Membrane positional restraint force /kcal mol^{-1} Å^{-2}**	**Dihedral restraint forces rk2 and rk3 /kcal mol^{-1} rad^{-2}**
Min	5000 steps	10.0	2.5	250.0, 250.0
Equil 1	125 ps	10.0	2.5	250.0, 250.0
Equil 2	125 ps	5.0	2.5	100.0, 100.0
Equil 3	125 ps	2.5	1.0	50.0, 50.0
Equil 4	500 ps	1.0	0.5	50.0, 50.0
Equil 5	500 ps	0.5	0.1	25.0, 25.0
Equil 6	500 ps	0.1	-	-

Table 4.1: PKD2 system equilibration restraint parameters for minimisation and each equilibration step.

4.2.2 Equilibrium MD simulations

Equilibrium MD simulations of full model (residues 213-700) and trucated (residues 469-700) PKD2 in the open conformation were carried out for a duration of 500 ns, using pmemd.cuda from AMBER22[139]. The temperature used was 300 K and the pressure was 1 bar. Langevin dynamics were used with a friction coefficient of 1 ps^{-1}, and the non-bonded cutoff was 12.0 Å, with force switching cutoff of 10.0 Å. A Monte Carlo barostat was used with semi-isotropic surface tension, constant in the xy plane. The average distances between Cα atoms of diagonally opposing domains (Figure 4.4a) were calculated using cpptraj from AmberTools[141], for residues L641, D642, F669, L677, and N681.

4.2.3 Steered MD

PKD2 was steered from open to closed and closed to open conformation, by using the average distance between Cα distances of diagonally opposing N681 atoms (pore diameter) as the collective variable (CV), with pmemd.cuda with PLUMED[177, 178]. A 3500 kJ mol^{-1} force constant was applied over 50 ns. The pore diameter value at N681 in the open and closed conformations were 15.0 Å and 11.0 Å, respectively. These values were observed in *apo* WT PKD2 after equilibration of each conformation. The temperature was 300 K, and coordinates were propagated with Langevin dynamics with a friction coefficient of 2 ps^{-1}. PME was used and a non-bonded cutoff of 8.0 Å was applied, with no force switching. The pressure was 1 bar, with a Berendsen barostat.

4.2.4 Seeded MD

After steering, 100 frames were evenly sampled from each trajectory (200 frames total per system). They were used as starting coordinates for an array of further equilibrium MD simulations, each of a duration of 50 ns and with a 2 fs timestep. Each seeded MD simulation was then reduced to the average of Cα distances of

opposite PKD2 domains at residues F669 and N681.

4.2.5 Markov State Modelling

The featurized data for all systems discussed in this chapter (*apo* WT PKD2, $PKD2_{F604P}$, and PKD2-PI(4,5)P_2) was pooled together and clustered into 100 microstates using k-means clustering. Each frame of each trajectory was assigned to a microstate, and implied timescales (ITS) were computed for each system (Figure 4.3). Based on the ITS, MSMs were built for each system with a lag time of 15 ns using PyEMMA version 2.5.7[142]. Perron Cluster-Cluster Analysis was used to partition the microstates into two metastable states based on the dynamics of the *apo* WT PKD2, and also into three metastable states. The probability of each metastable macrostate was taken as the sum of the stationary probabilities of the microstates belonging to it.

Further bootstrapping by resampling was carried out, selecting 200 random trajectories from the seeded MD data pool of each system. The micro- and macrostate assignments used in the initial MSMs were re-used here. The resampling was repeated for 100 iterations, and the state probability of a system was considered to be the mean of the resulting probability distribution, with a one standard deviation error.

Additionally, an equilibrium conformational ensemble was recreated from the seeded MD data for each system. 10,000 frames were sampled from the data pool, weighted by the MSM stationary probabilities of each state.

4.3 Results

4.3.1 Truncating the TOP Domain of PKD2

The tetramer organization of PKD2, as well as the requirement to simulate a lipid bilayer, yield very large protein systems (∼350,000 atoms), resulting in significantly longer simulation times. To reduce the simulation time, the TOP domain of PKD2

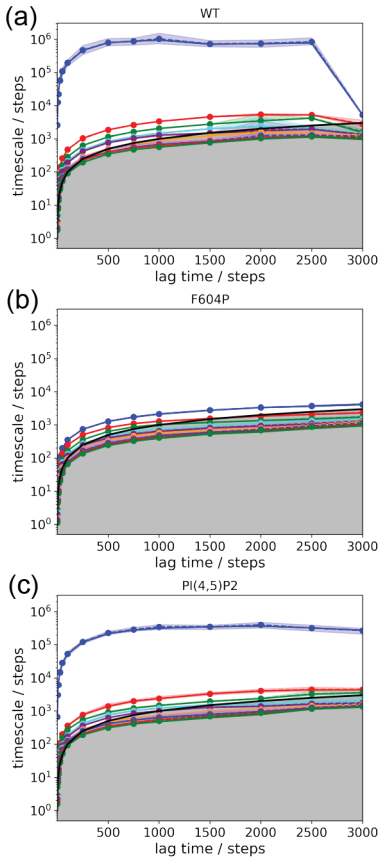

Figure 4.3: ITS plots for all systems modelled in this chapter (a) WT *apo* PKD2, (b) $PKD2_{F604P}$, and (c) PKD2-PI(4,5)P_2.

was truncated, yielding a final system size of ∼225,000 atoms (a 35% reduction in size). While the TOP domain has been shown to be relevant to PKD2 tetramer aggregation and mutations in that region can prevent the pore from opening[212, 213], it was expected that any effect this may have on the protein conformation ensemble will be equally present in all MSMs computed in this chapter. Therefore it should not affect the comparison of conformation probabilities, as also seen when truncating the α7 helix in PTP1B in Chapter 2.

However, in order to test for any pore behaviour differences between the full and truncated PKD2 models, 500 ns equilibrium MD simulations of the *apo* open conformation were carried out for both. On a GTX1080 Ti GPU card, the simulation rate using pmemd.cuda was 12 ns day^{-1} and 22 ns day^{-1} for the full and truncated models respectively. The average of the Cα distances between diagonally opposite residues were computed for the residues of the selectivity filter and the lower gate (Figure 4.2a), used as an approximation of the pore diameter at those points (Figure 4.4a). The pore diameter at residues L641 and D643 is higher in the truncated PKD2 than in the full model, and increases during the simulation. This increase is likely due to their proximity to the TOP domain, causing the largest effect. The pore at selectivity filter residue F669 remains mostly similar amongst the two protein models, although there a transient decrease in diameter at 100-300 ns. In a similar fashion, the diameter at L677 and N681 of the truncated PKD2 remains very close to the full length model (Figure 4.4b).

Mutations in the TOP domain cause a decrease in observed conductance, i.e. prevent the PKD2 channel from adopting the open conformation. Therefore, the main concern when removing the TOP domain in the truncated model described here was that PKD2 would spontaneously shift from the open conformation to the closed. In the 500 ns MD simulations outlined above, no significant decrease of the selectivity filter and lower gate regions was observed. Therefore, the protein models used to compute state probabilities in the following sections of this chapter are all of the truncated variant, which includes residues 469-700 only. Additionally,

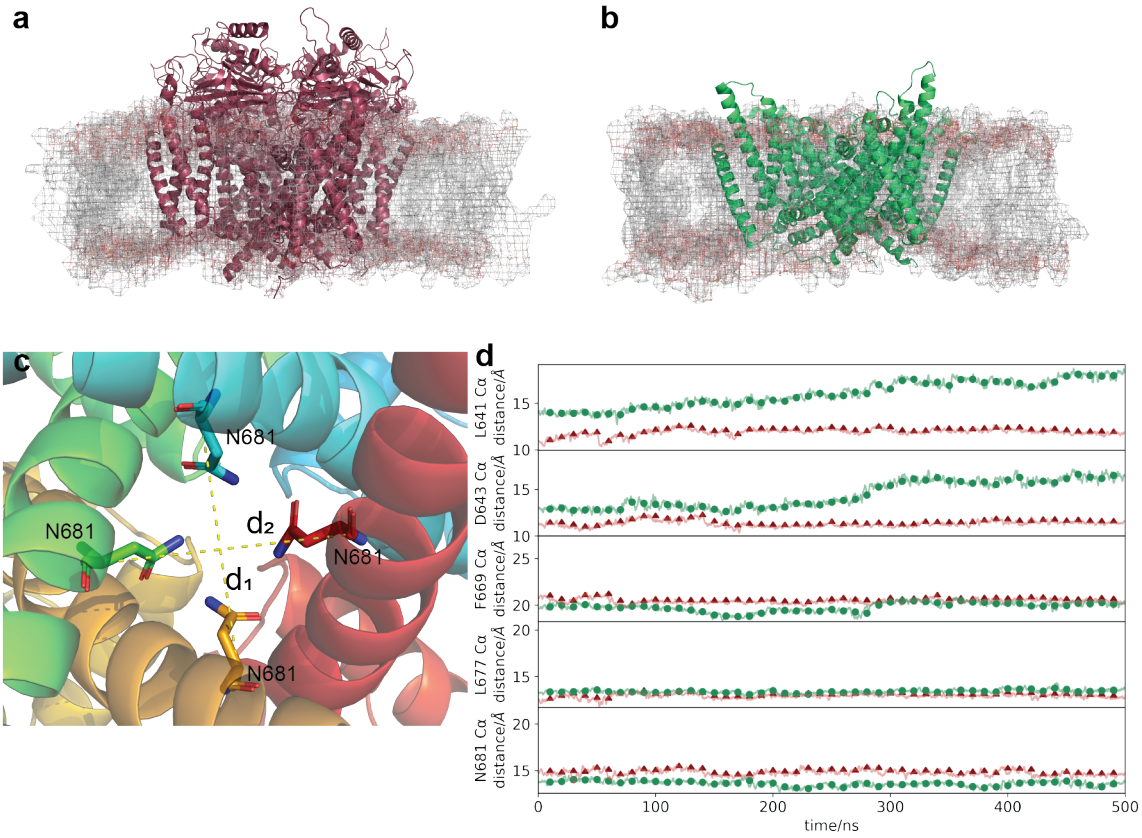

Figure 4.4: Monitoring PKD2 pore dynamics in the open conformations of the full and truncated protein models. (a) Full protein model, with the membrane shown as a black (hydrophobic tails) and red (hydrophilic heads) mesh. (b) Truncated protein model, with the membrane shown as a black (hydrophobic tails) and red (hydrophilic heads) mesh. (c) The residue distance is taken as the average of Cα distances d_1 and d_2, indicated by yellow lines. Residue N681 is shown as an example. Different monomers are coloured in green, yellow, red, and blue. (d) The residue distances for selectivity filter (L641, D643 and F669) and lower gate (L677 and N681) residues in the full PKD2 model (red, triangles) and the truncated variant (green, circles) over 500 ns. Note that while the ranges in these plots are different to showcase the different diameter values across the channel, the scale is the same.

due to the highest similarity between full and truncated models being observed in residues F669 and N681, the pore diameter at these points was chosen as the features for the Markov State Modelling. These residues represent the selectivity filter and lower gate respectively. N681 was chosen over L677 as it is further away from F669, allowing to capture pore conformational changes further along the pore.

4.3.2 The Effects of a Gain-of-Function Mutation Are Captured by MSMs

For initial validation that the AMMo workflow (section 1.5) can capture changes in the PKD2 conformational ensemble, the wild type (WT) PKD2 and $PKD2_{F604P}$ proteins were steered from pore open to pore closed conformations, and *vice versa*. Due to large system size, steered MD simulations were performed for 50 ns only, a shorter duration than one used in Chapters 2 and 3. As the lower gate of PKD2 is key to its function, only the Cα distance at N681 was steered. From each sMD trajectory, 100 snapshots were sampled, and used as seeds for 50 ns equilibrium MD simulations (10 μs total sampling time per model).

All seeded MD trajectories were featurized using pore diameters at residues F669 and N681, as illustrated in Figure 4.4. Data from all systems described in this chapter was pooled and clustered into 100 k-means clusters. Implied timescales were computed for a range of lag times between 0.01 ns and 30 ns (Figure 4.3). The lag time used for MSMs was 15 ns. After MSMs were computed, Perron Cluster Cluster Analysis (PCCA) of the WT MSM was used to further cluster the microstates into two metastable states. The stationary distributions are shown in Figure 4.5a-c. The state with lower Cα-Cα distances is considered closed, while the state with larger distances is considered open. The highest probability density regions in both the WT and F604P MSMs have high selectivity filter and lower gate distance values, with $PKD2_{F604P}$ having the highest. However, these regions are grouped into the same metastable state in the two state model, indicating no increased activation by a F604P mutation. The selectivity filter (i.e. residue F669) distance in the WT MSM

is similar to the pore dynamics observed in the long equilibrium MD simulations shown in Figure 4.4b), and the pore diameter at the selectivity filter in $PKD2_{F604P}$ is slightly larger. This increase in pore diameter compared to previous open channel simulations suggests that the F604P mutation may induce a hyper-open state rather than just stabilizing the same open conformation as observed in WT PKD2. This hyper-open conformation is the result of the distortion of the S5-S6 helices observed in cryo-EM structures of PKD2[198]. To this effect, PCCA was applied again, this time separating the protein conformational ensemble into 3 states: "closed", "open", and "hyper-open", each with increasing pore diameter. The partitioning of the conformational space is shown in Figure 4.5d-f, and the results of 100 iterations of bootstrapping are shown in Figure 4.5g.

WT PKD2 is modelled to have a closed state probability of 2% (IQR 1-4%), open state probability of 79% (IQR 75-82%), and hyper-open state probability of 17% (IQR 15-21%). The F604P mutation increases the hyper-open state probability to 94% (IQR 0-95%), leaving the closed and open state probabilities to be 0% (IQR 0-100%) and 6% (IQR 0-6%) respectively (Figure 4.5g). It was previously proposed that the increased current conductance observed with $PKD2_{F604P}$ comes from a kink induced in the S5 helix, which in turn distorts the S6 helix[203]. To confirm that this is captured by MD simulation, an MSM probability weighted conformational ensemble of 10,000 frames was resampled from the seeded MD data pool for WT PKD2 and $PKD2_{F604P}$. The conformational change observed in these ensembles is shown in Figure 4.6a.

The helix conformational change can also be represented by an angle value, with the mutation site as the central point. The WT PKD2 MSM-weighted conformational ensemble exhibits angle values of $\sim 150^{o}$, while the mutant angle values are lower, at 120-140^{o}, which indicates slight bending (Figure 4.6b). Additionally, the F604P mutation was previously proposed to change the conformation of the S6 helix (and therefore the lower gate), by π/α helix switching[198]. The ^{668}MFFIL^{672} section of the S6 helix it adopts a π helix conformation[198], which corresponds to a

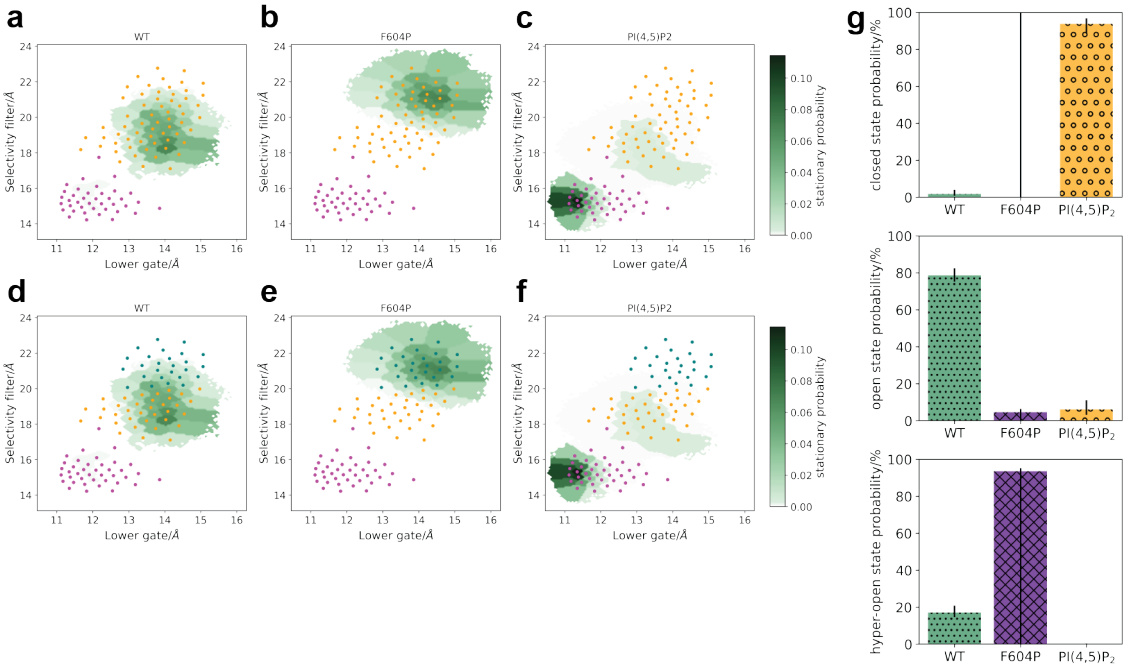

Figure 4.5: State probabilities of WT PKD2, PKD_{F604P} and PKD2-$PI(4,5)P_2$. (a-c) System equilibrium probability density, with two metastable states plotted on top: closed (magenta) and open (orange). (d-f) System equilibrium probability density with three mestastable states plotted on top: closed (magenta), open (orange), and hyper-open (teal). (g) Results of 100 bootstrapping iterations. The probability of each state is taken as the median of the resulting probability distribution, and the black vertical bars show the inter-quartile range between quartiles 1 and 3.

M668 ϕ backbone angle value of -60° in our simulations. In $PKD2_{F604P}$, we observe an increase in the M668 ϕ angle probability at -135° (Figure 4.6c). Furthermore, the movement of the N-terminus of the S5 helix was shown to affect the S4-S5 linker, shifting the R581 residue. R581 is generally conserved amongst TRPP proteins, and a R581A substitution was shown to decrease PKD2 activity[214]. Similarly, the $PKD2_{F604P}$ conformational ensemble shows an increase in the Cα distance between diagonally opposing R581 residues, compared to WT PKD2 (Figure 4.6d). These conformational differences between the WT and F604P PKD2 models described here confirm that the computational modelling is in good agreement with experiment.

4.3.3 Both the Hydrophobic Tail and Polar Head of PI(4,5)P_2 Lipid Stabilizes a PKD2 Closed Conformation

With the above results correctly capturing the GoF effect of a F604P mutation, and additional conformational changes observed in previous experimental work, the same workflow was applied to the PKD2-PI(4,5)P_2 complex. A lipid molecule was modelled bound to each of the PKD2 monomers, based on the cryo-EM structure of PKD2 with PI(4,5)P_2 (Figure 4.2c).

The presence of PI(4,5)P_2 significantly decreases pore diameter at both the lower gate and the selectivity filter (Figure 4.5f). It causes PKD2 to primarily adopt the closed conformation, with a probability of 94% (IQR 89-97%), in agreement with the closed conformation resolved by cryo-EM[205]. This leaves an open state probability of 6% (IQR 3-11%) and no hyper-open state is populated (Figure 4.5g). To further analyse why PI(4,5)P_2 has such an effect on the pore conformation, an MSM probability weighted equilibrium ensemble of 10,000 frames was constructed.

Throughout seeded MD, PI(4,5)P_2 adopts a stable conformation that has a ~ 3 Å heavy atom RMSD from the ensemble average (Figure 4.8a), held stable by the interactions between the lipid polar head and residues 504-507 in the S2-S3 linker of PKD2 (Figure 4.7). The hydrophobic tail extends into the space between helices S4 and S5 of one monomer, and helices S5 and S6 of the next monomer, adjacent to the

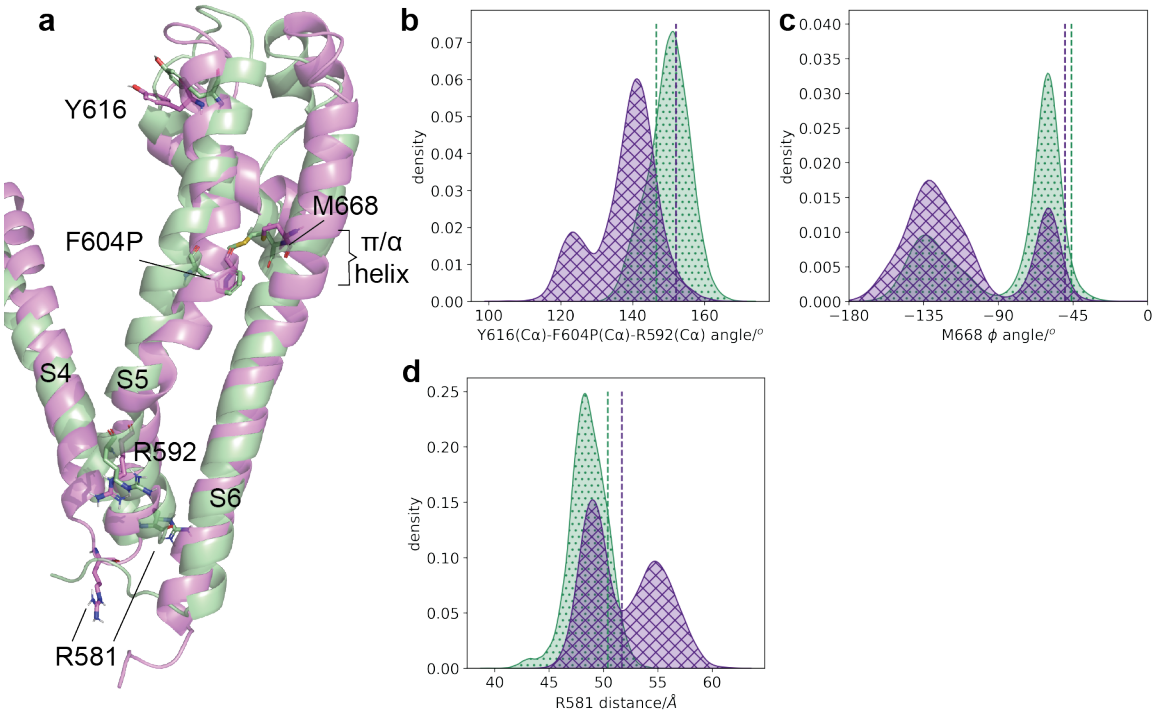

Figure 4.6: Comparison of the WT and F604P mutant PKD2 conformational ensembles. (a) The kink in the S5 helix, induced by the F604P mutation (violet) compared to WT PKD2 S5 helix (green). (b) The S5 helix angle, with F604P as the center point, represents the kink illustrated in (a). The values observed in cryo-EM structures for WT PKD2 (PDB ID 5MKE) and $PKD2_{F604P}$ (PDB ID 6D1W) are shown in dashed lines. (c) The M6881 ϕ backbone angle, which relates to the pi/α helix conformational changes in WT (green) and F604P mutant (violet) PKD2. The values observed in cryo-EM structures for WT PKD2 (PDB ID 5MKE) and $PKD2_{F604P}$ (PDB ID 6D1W) are shown in dashed lines. (d) The mean R581(Cα) distances between diagonally opposing domains in WT PKD2 (green) and the F604P mutant (violet). The values observed in cryo-EM structures for WT PKD2 (PDB ID 5MKE) and $PKD2_{F604P}$ (PDB ID 6D1W) are shown in dashed lines.

selectivity filter (Figure 4.8c). The tail remains buried in the protein throughout the simulations, as illustrated by the distances between the two terminal hydrophobic chain carbons and PKD2 (Figure 4.8b and c). This restricts the movement of the selectivity filter region of the pore, leaving no space for opening and pushing it into a closed conformation (Figure 4.8d).

In addition to forming hydrogen bonds to the S2-S3 linker of PKD2, PI(4,5)P_2 also interacts with S591 in helix S5, as shown in Figure 4.9a and f. This perturbs the conformation of the adjacent R592, which forms a hydrogen bond with E686 of the S6 helix in *apo* PKD2 (Figure 4.9b and c). The side chain of R592 extends further downward, instead forming a hydrogen bond with D690 (Figure 4.9a and d). This allows E686 to then form an ionic interaction with K688 of the previous domain (Figure 4.9a and e). This brings the domains of PKD2 closer together, stabilizing a closed pore conformation. The domain linking effect in PKD2-PI(4,5)P_2, and its absence in *apo* protein is illustrated in Figures 4.9g and h respectively. While the interactions between PKD2 and the hydrophobic tail shown in Figure 4.8 would be quite difficult to target by conventional small molecule drug design, the interactions formed between the polar lipid head and PKD2 are easier to exploit. Forming similar interactions with S591 or otherwise engaging R592 to free up E686 to interact with K688 would inhibit PKD2. On the other hand, ligands that could form interactions with E686 that mimic the R592-E686 hydrogen bonding would potentially act as allosteric activators of PKD2. These interactions were not observed in previously report MD simulations of PKD2 with PIP_2 lipids, potentially because of the shorter hydrophobic chain modelled. In previous work, PIP_2 was modelled as interacting with the S4-S5 linker, which would position it too far away from the S591 interactions observed here[205]. The hydrophobic tail length modelled in this chapter was 18 and 20 carbons, making the PI(4,5)P_2 model too elongated to reach the S4-S5 linker. This is illustrated by the difference in the structures of the detergent molecules to the lipid in Figure 4.2c. Therefore the use of the longer PIP_2 structure, which is most commonly found in mammalian cells[210], provided new insights into the regulation

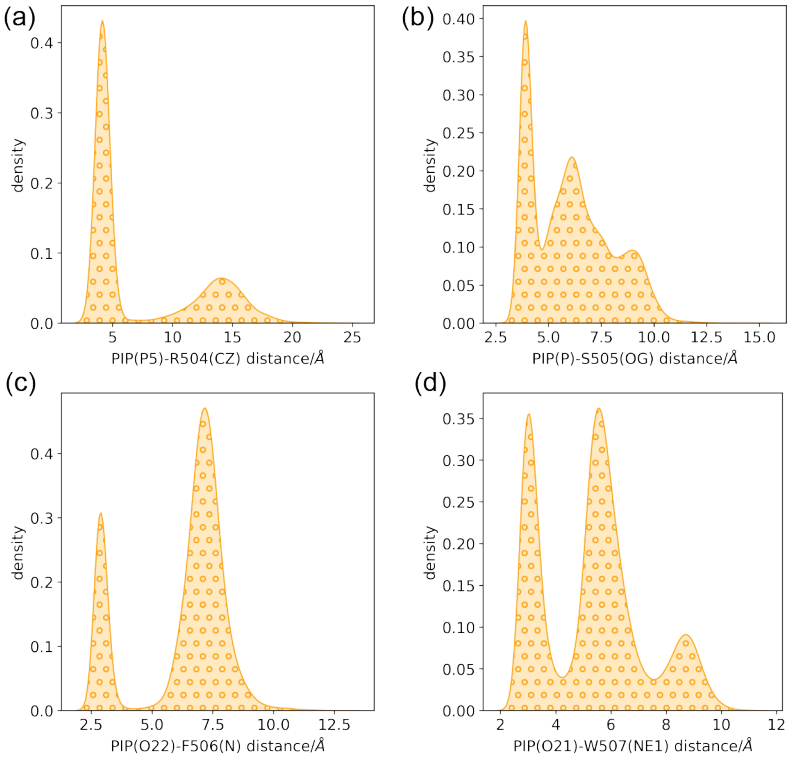

Figure 4.7: $PI(4,5)P_2$ hydrogen bonding with PKD2 S2-S3 linker residues: (a) R504, (b) S505, (c) F506, (d) W507

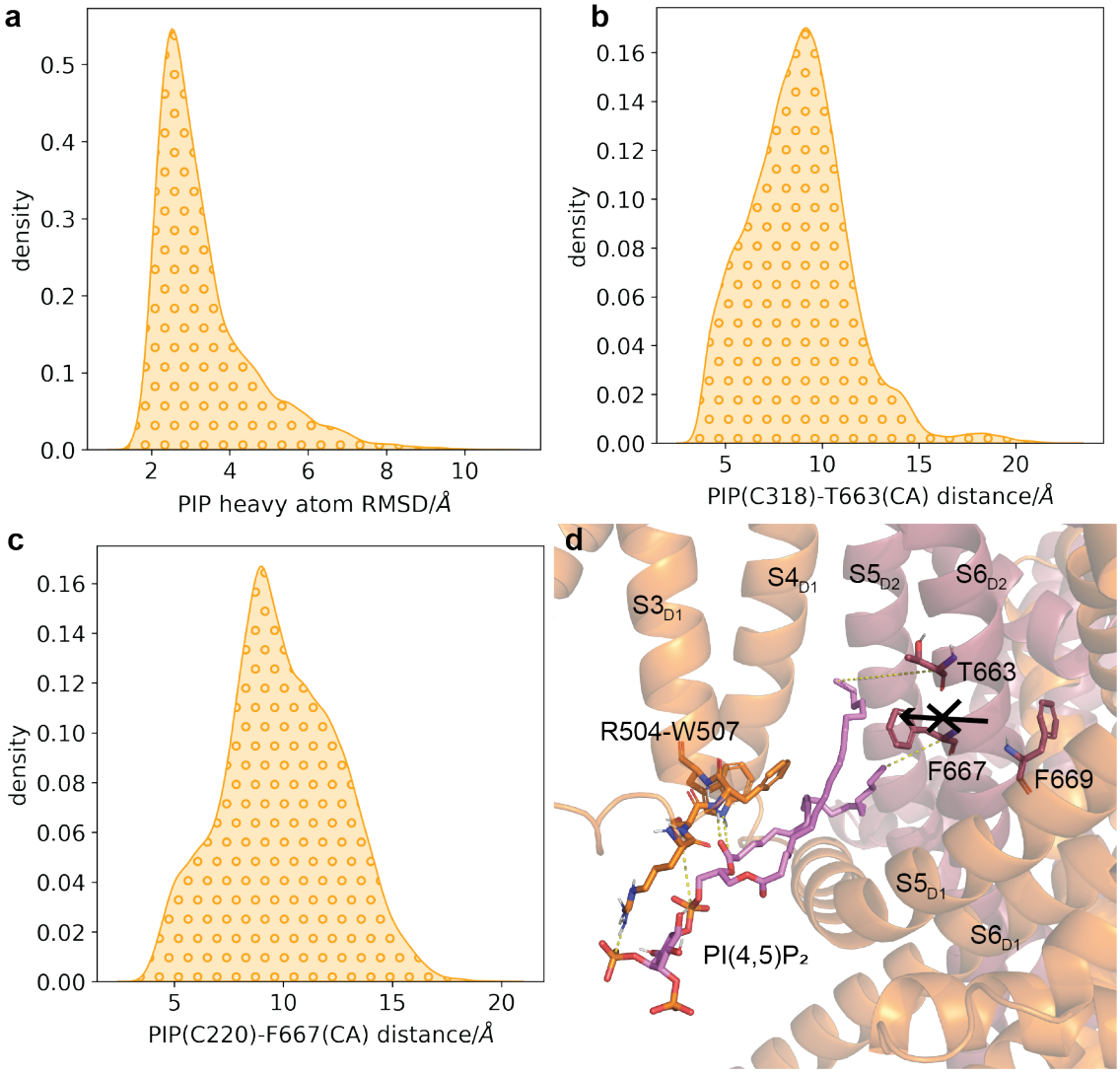

Figure 4.8: $PI(4,5)P_2$ hydrophobic tail interactions with PKD2. (a) $PI(4,5)P_2$ heavy atom RMSD to the average ensemble pose. (b) The distance between $PI(4,5)P_2$ hydrophobic tail terminal carbon C318 and T663. (c) The distance between $PI(4,5)P_2$ hydrophobic tail terminal carbon C220 and F667. (d) A representative conformation of $PI(4,5)P_2$ buried into PKD2. The black arrow indicates the direction the helices would have to move in for the ion channel to open, which is blocked by the presence of the $PI(4,5)P_2$ hydrophobic tail. F669 of the selectivity filter is also shown. The hydrogen bond distances between $PI(4,5)P_2$ and residues 504-507 are shown in Figure 4.7.

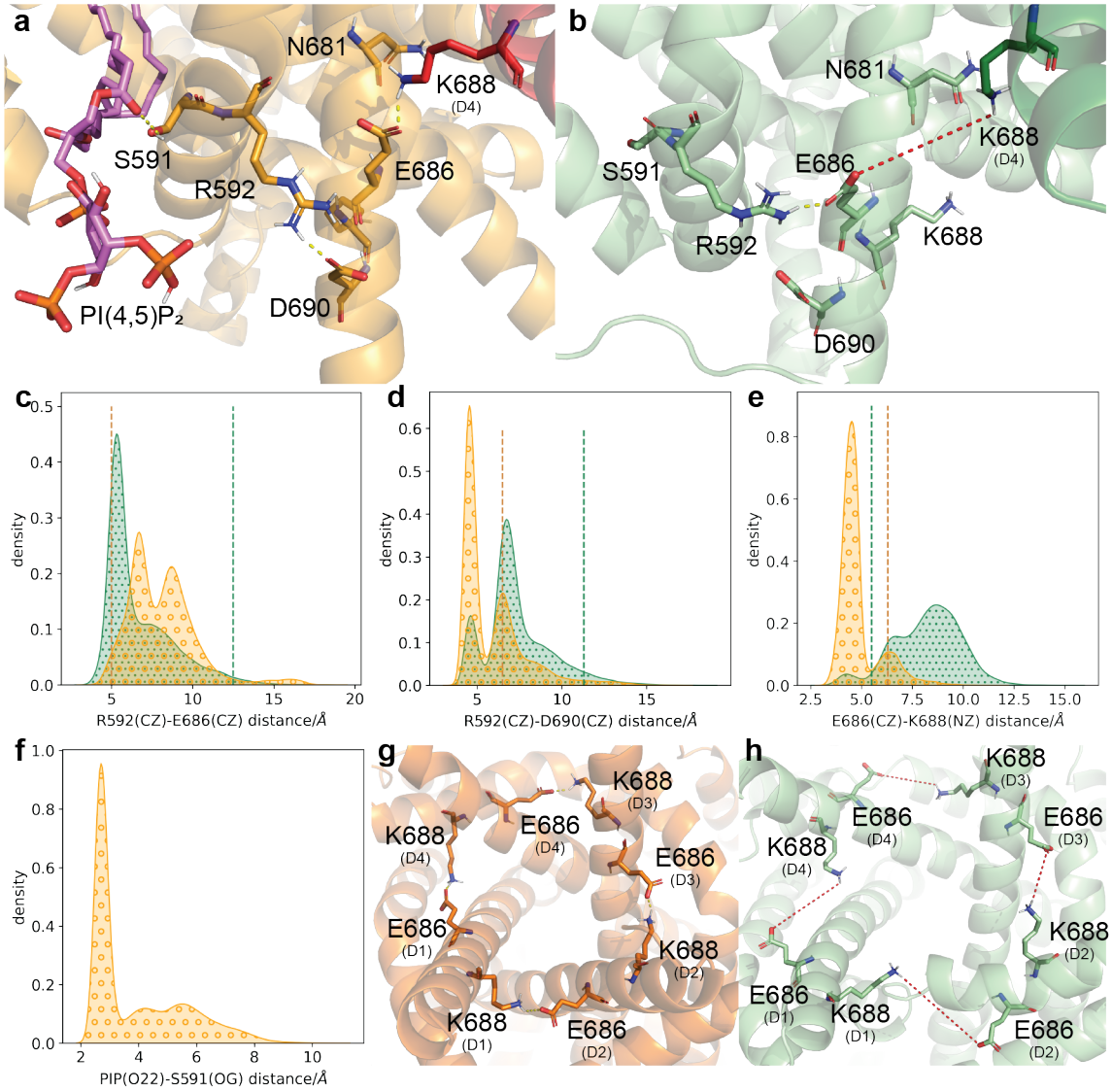

Figure 4.9: The stabilization of a PKD2 closed conformation at the lower gate by $PI(4,5)P_2$. (a) The interactions between $PI(4,5)P_2$ and S591, R592 and D690, and E686 and K688 in PKD2-$PI(4,5)P_2$. Domain 4 is highlighted in red, while the rest of the protein is shown in orange. $PI(4,5)P_2$ is shown in pink. Lower gate N681 is also shown. (b) The interaction between R592 and E686, and the missing bonding between E686 and K688 in *apo* PKD2. Domain 4 is shown in darker green to the rest of the protein. Lower gate N681 is also shown. (c) The R592(CZ)-E686(CZ) distance in *apo* PKD2 (green dots) and PKD2-$PI(4,5)P_2$ (yellow circles). The cr(d) The R592(CZ)-D690(CZ) distance in *apo* PKD2 (green dots) and PKD2-$PI(4,5)P_2$ (yellow circles). (e) The E686(CZ)-K688(NZ) distance in *apo* PKD2 (green dots) and PKD2-$PI(4,5)P_2$ (yellow circles). Here K688 is of the previous monomer to E686. (f) The distance between $PI(4,5)P_2$ O22 and OG atom of S591. (g) A bottom-up view of the interlocking interactions between E686 and K688 of adjacent monomers in PKD2-$PIP(4,5)P_2$. (f) A bottom-up view of the interlocking interactions between E686 and K688 of adjacent domains in *apo* PKD2.

of PKD2 by PIP_2.

4.4 Discussion

The results described above identify a hyper-open state of PKD2, which is largely populated in the F604P mutant. This increased pore diameter comes from the helix-breaking properties of the proline mutation, which cause a bend in the S5 helix, in turn distorting the S6 helix as well. Additionally, the presence of PI(4,5)P_2 causes the ion channel to adopt a primarily closed conformation. This lipid effect is in agreement with PKD2 conformation observed in cryo-EM structures in presence of PIP_2. A closer look at the PKD2-PI(4,5)P_2 conformational ensemble revealed a previously unknown lipid inhibition mechanism, changing the interactions between helices S5 and S6 at the lower gate. This could be exploited for the development of small molecule PKD2 modulators. Future work could also include the modelling of PKD2-PI(3,4,5)P_2, which was shown to have no effect on PKD2 activity[204]. This could be due to the additional phosphate group causing a change in the position of PIP_3 binding to PKD2, preventing hydrogen bonding to S591. Further understanding of the differences between PIP_2 and PIP_3 modulation would provide useful insight into the regulation of PKD2.

This chapter presents an example of what application the sMD/MSM workflow might look like of a less well studied, more complex system. To reduce computational cost, the protein TOP domain was removed, and the steered MD simulations were significantly shorter than ones seen in the previous chapters. At a total sampling time of 10 μs per model (50 ns $\times$ 200 seeded MD trajectories), simulating the full protein model using a GTX1080 GPU card would have taken $\sim$ 20k GPU hours for the work shown here (12 ns day^{-1} simulation rate). With the truncation of the protein, the simulation rate increased to 22 ns day^{-1}, and only 11k GPU hours were required.

Furthermore, the steered MD approach remained simple, only biasing the pore at the lower gate, and did not attempt to identify and steer any allosteric network

residues. These sort of compromises might be crucial in more time sensitive drug discovery studies, in order to deliver results at the required pace. However, the truncation of PKD2 was tested by comparing relevant residue dynamics to the full length protein model, and collective variables for steering and MSMs were chosen accordingly. Most importantly, the whole protocol was initially validated on a GoF mutant F604P. Had the results not shown increased activation of $PKD2_{F604P}$, other protocols would have had to be tested. This can include returning to full length PKD2 model, changing the sMD CV set or MSM features to better describe protein conformational changes, or simply increasing simulation time for improved sampling. A proof-of-concept validation step is crucial in determining the reliability of such computational results when modelling novel systems.

Chapter 5

Conclusions

With the large number of pharmaceuticals already on the market, drug discovery is an increasingly difficult and expensive endeavour. Focus on developing allosteric modulators of target proteins opens up new avenues in drug design. However, predicting the effects of ligands binding at an allosteric site remains non-trivial, as ligand binding does not always equate to a change in protein activity. This thesis outlines a combined steered MD (sMD) and Markov State Modelling (MSM) workflow, designed to quantify and study the effects of potential allosteric modulators on protein conformational ensembles.

5.1 Insights into Allosteric Modulation of Proteins Discussed in this Thesis

This thesis focused on three protein systems - Protein Tyrosine Phosphatase 1B (PTP1B), Exchange Protein Activated by cAMP 1 (EPAC1), and Polycystic Kidney Disease 2 (PKD2). While the work carried out on PTP1B mainly involved initial validation of the sMD/MSM methodology, it also provided insight into the two allosteric sites studied. The comparison of inhibitors binding to the same pocket, but exhibiting different effects, allowed for better understanding of the interactions determining allosteric modulation of PTP1B.

The research carried out on EPAC1 revealed protein-ligand interactions crucial to the action of the native agonist, cAMP, but not formed by the partial activator I942. Simulations with various distance restraints provided basis for proposed modifications of I942 and key residues that should be targeted by a full agonist. Additionally, modelling the agonistic effects of cAMP, as well as the partial agonism of I942, confirmed that the sMD/MSM method outlined in this thesis can be used to model activators, as well as inhibitors. Furthermore, the activation of EPACs is characterized by rearrangement of the regulatory domain, which is a much larger-scale conformational change than the movement of the WPD loop of PTP1B.

Finally, the modelling of the PKD2-PI(4,5)P_2 complex revealed a lipid induced change in the hydrogen bond network of PKD2, causing the inhibition of the ion channel. These interactions could be exploited to potentially drug PKD2 with small molecules. The model of the gain-of-function (GoF) F604P mutant being in agreement with previous experimental observations also lends validity to the lipid modulation mechanism proposed here.

5.2 Development of the sMD/MSM Workflow

Applying the sMD/MSM workflow to the variety of drug target classes outlined here has highlighted the potential pitfalls and complications of this methodology. Firstly, the main challenge of the approach is the dimensionality reduction, from collective variable (CV) selection for sMD, to space partitioning into discrete states for the MSMs. Describing complex changes in just a few dimensions risks losing key details and misrepresenting the protein conformational ensemble. For instance, focusing on just the active site of the protein might give an oversimplified view of the transition between active and inactive states, such as illustrated by the work on PTP1B. Therefore, a careful description of protein activation is necessary, however, this data might not be readily available. In these cases, a simpler approach may be applied, or an allosteric network may be proposed based on prior simulations.

Additionally, microstate clustering to biologically relevant states may also be

non-trivial. Part of the benefit of using the sMD to sample a wider conformational space is the reduction in overall computational time, resulting from using multiple shorter seeded MD simulations. However, this means that each of the trajectories samples a small subset of the conformational space. In some cases, such as seen with EPAC1, data driven approaches for dimensionality reduction may not be applicable in such a low data regime, and careful consideration of the protein system is required to assign which conformations are "active" or "inactive". Once again, such data may not be available in novel drug design endeavours. However, enhanced sampling approaches requiring less prior knowledge, such as accelerated MD (aMD), could be employed to explore conformational space where structural data is lacking[115].

A lot of the insight into the protein targets studied in this thesis came from not just the modelled active/inactive state probabilities, but also the MSM-weighted conformational ensembles. Using the MSM-computed probabilities to resample the seeded MD data pool and recreate a statistically representative ensemble of the relevant conformational space allows to take this approach beyond a "black box" method that simply yields some state probabilities. Relating those probabilities directly to protein-ligand interactions observed in the MD simulations allows to draw more meaningful conclusions on the ligand effects on the protein conformational ensembles. This can be taken even further by restraining the ligand in different ways, to test how changing the ligand to form different interactions with the protein would affect the protein conformational ensembles.

5.3 Future Work in Applying Enhanced Sampling and MSMs to Modelling Allostery

While our sMD/MSM approach was shown to be a useful tool to predict and quantify the effects of allosteric modulators, a large part of the results described here relied on previous trial and error, before a suitable protocol for each system was identified. As mentioned above, the selection of low-dimensional features to describe protein

conformations can be an involved process, requiring significant knowledge of the chemistry and dynamics of the target system. While an ideal solution would be full automation of the workflow, where simulations could be analysed to reveal the slowest dynamic processes and select relevant descriptors without any user input, this would require amounts of simulation data that are not currently routinely accessible.

However, further work can be carried out to make the "trial and error" process easier and more directed. For instance, initial longer equilibrium MD simulations at each allosteric end state were already employed in this thesis to investigate protein dynamics, such as to confirm that the protein model reflects information available from experimental data, or to find suitable descriptions of the relevant conformations. Systematic comparison of individual residue dynamics during these simulations could yield information on the allosteric network or the main conformational changes that characterize the active/inactive states of the protein. Computational tools for this sort of analysis are already available and could be integrated into the workflow, to allow easier CV and feature selection for dimensionality reduction.

Similarly, trialling various MSM parameters could also be made semi-automated. Features, micro- and macrostates, model lag time - all of these currently require careful selection by the modeller. If an example is selected where the result is already known (such as the activation of EPAC1 by cAMP, or the GoF mutant $PKD2_{F604P}$), a high throughput test of hyper parameters can be tested to find the MSM protocol that yields results in agreement with experimental conclusions. Selection of reasonable inputs would still be required, but reasonable starting points could also be suggested - for instance, the MSM lag times used in this thesis all ranged between 15 and 25 ns. The computational cost of building MSMs is trivial compared to the MD data generation, which would allow more extensive testing.

The above examples are only a couple of ways that the AMMo workflow could be made easier to apply. With the increase of computing power routinely available, it will also be easier to generate large amounts of data and subsequently less prohibitive to use such computationally expensive methods for drug discovery. Further

development of approaches such as the sMD/MSM workflow described in this thesis could make a significant impact on the way allosteric modulator drugs are being developed, saving on time and costs.

Bibliography

(1) Newman, D. J.; Cragg, G. M.; Snader, K. M. *Nat. Prod. Rep.* **2000**, *17*, 215–234.

(2) Drews, J. *Science* **2000**, *287*, 1960–1964.

(3) Pina, A. S.; Hussain, A.; Roque, A. C. A. *Methods Mol. Biol.* **2009**, *572*, 3–12.

(4) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug. Deliv. Rev.* **2001**, *46*, 3–26.

(5) Ngo, H. X.; Garneu-Tsodikova, S. *Med. Chem. Commun.* **2018**, *9*, 757–758.

(6) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. *Br. J. Pharmacol.* **2011**, *162*, 1239–1249.

(7) Scannell, J. W.; Blanckley, A.; Boldon, H.; Warrington, B. *Nat. Rev. Drug Discov.* **2012**, *11*, 191–200.

(8) Anderson, A. C. *Chem. Biol.* **2003**, *10*, 787–797.

(9) Sadybekov, A. V.; Katritch, V. *Nature* **2023**, *616*, 673–685.

(10) Ferreira, L. G.; Dos Santos, R. N.; Oliva, G.; Andricopulo, A. D. *Molecules* **2015**, *20*, 13384–13421.

(11) Sabe, V. T.; Ntombela, T.; Jhamba, L. A.; Maguire, G. E.; Govender, T.; Naicker, T.; Kruger, H. G. *Eur. J. Med. Chem.* **2021**, *224*, 113705.

(12) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(13) O'Boyle, N. M. *J. Cheminformatics* **2012**, *4*, 22.

(14) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(15) Wigh, D. S.; Goddman, J. M.; Lapkin, A. A. *WIREs Comput. Mol. Sci.* **2022**, *12*, e1603.

(16) Bajusz, D.; Rácz, A.; Héberger, K. *J. Cheminformatics* **2015**, *7*, 20.

(17) Zoete, V.; Daina, A.; Bovigny, C.; Michielin, O. *J. Chem. Inf. Model.* **2016**, *56*, 1399–1404.

(18) Castleman, P.; Szwabowski, G.; Bowman, D.; Cole, J.; Parrill, A.; Baker, D. *J. Mol. Graph.* **2022**, *111*, 108107.

(19) Śledź, P.; Caflisch, A. *Curr. Opin. Struct. Biol.* **2018**, *48*, 93–102.

(20) Kalliokoski, T. *Mol. Inform.* **2021**, *40*, 2100089.

(21) Clyde, A.; Liu, X.; Brettin, T.; Yoo, H.; Partin, A.; Babuji, Y.; Blaiszik, B.; Mohd-Yusof, J.; Merzky, A.; Turilli, M.; Jha, S.; Ramanathan, A.; Stevens, R. *Sci. Rep.* **2023**, *13*, 2105.

(22) Wong, K. M.; Tai, H. K.; Siu, S. W. I. *Chem. Biol. Drug Des.* **2021**, *97*, 97–110.

(23) Wang, L. et al. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(24) Kuhn, M.; Firth-Clark, S.; Tosco, P.; Mey, A. S. J. S.; Mackey, M.; Michel, J. *J. Chem. Inf. Model.* **2020**, *60*, 3120–3130.

(25) Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; Jr., T. D. M.; Pearlman, D. A.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Xu, H.; Sherman, W.; York, D. M. *J. Chem. Inf. Model.* **2020**, *60*, 5595–5623.

(26) Ertl, P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.

(27) Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H. *Nat. Rev. Drug Discover.* **2016**, *15*, 605–619.

(28) Doak, B. C.; Norton, R. S.; Scanlon, M. J. *Pharmacol. Ther.* **2016**, *167*, 28–37.

(29) Sun, D.; Gao, W.; Hu, H.; Zhou, S. *Acta Pharm. Sin. B* **2022**, *12*, 3049–3062.

(30) Hopkins, A. L.; Groom, C. R. *Nat. Rev. Drug Discov.* **2002**, *1*, 727–730.

(31) Agoni, C.; Olotu, F. A.; Ramharack, P.; Soliman, M. E. *J. Mol. Model.* **2020**, *26*, 120.

(32) Bowes, J.; Brown, A. J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S. *Nat. Rev. Drug Discov.* **2012**, *11*, 909–922.

(33) Schneider, G. *Nat. Rev. Drug Discov.* **2018**, *17*, 97–113.

(34) Gao, Z.-G.; Jacobson, K. A. *Drug Discov. Today* **2006**, *11*, 191–202.

(35) Swain, J. F.; Gierasch, L. M. *Curr. Opin. Struct. Biol.* **2006**, *16*, 102–108.

(36) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. *Nature* **2014**, *508*, 331–339.

(37) Verkhivker, G. M.; Agajanian, S.; Hu, G.; Tao, P. *Front. Mol. Biosci.* **2020**, *7*.

(38) Bohr, C.; Haseelbalch, K.; Krogh, A. *Skand. Arch. Physiol.* **1904**, *15*, 401–412.

(39) Gunasekaran, K.; Ma, B.; Nussinov, R. *Proteins* **2004**, *57*, 433–443.

(40) Changeux, J.-P.; Christopoulos, A. *Cell* **2016**, *166*, 1084–1102.

(41) Chatzigoulas, A.; Cournia, Z. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1529.

(42) Monod, J.; Wyman, J.; Changeux, J. P. *J. Mol. Biol.* **1965**, *12*, 88–118.

(43) Koshland, D. E.; Nemethy, G.; Filmer, D. *Biochemistry* **1966**, *5*, 365–385.

(44) Hilser, V. J.; Wrabl, J. O.; Motlagh, H. N. *Annu. Rev. Biophys.* **2012**, *41*, 585–609.

(45) Orozco, M. *Chem. Soc. Rev.* **2014**, *43*, 5051–5066.

(46) Weber, G. *Biochemistry* **1972**, *11*, 864–878.

(47) Grover, A. K. *Med. Princ. Pract.* **2013**, *22*, 418–426.

(48) Lu, S.; He, X.; Ni, D.; Zhang, J. *J. Med. Chem* **2019**, *62*, 6405–6421.

(49) Han, B.; Salituro, F. G.; Blanco, M.-J. *ACS Med. Chem. Lett.* **2020**, *11*, 1810–1819.

(50) Wilhelm, A.; Lopez-Garcia, L. A.; Busschots, K.; Fröhner, W.; Maurer, F.; Boettcher, S.; Zhang, H.; Schulze, J. O.; Biondi, R. M.; Engel, M. *J. Med. Chem.* **2012**, *55*, 9817–9830.

(51) Zhang, Z.-Y. *Acc. Chem. Res.* **2017**, *50*, 122–129.

(52) Ni, D.; Lu, S.; Zhang, J. *Med. Res. Rev.* **2019**, *39*, 2314–2342.

(53) Amamuddy, O. S.; Veldman, W.; Manyumwa, C.; Khairallah, A.; Agajanian, S.; Oluyemi, O.; Verkhivker, G. M.; Bishop, Ö. T. *Int. J. Mol. Sci.* **2020**, *21*, 847.

(54) Kelly, M. J. et al. *J. Med. Chem.* **2014**, *57*, 3912–3923.

(55) Novel Drug Approvals for 2024, https://www.fda.gov/drugs/new-drugs-fda-cders-new-molecular-entities-and-new-therapeutic-biological-products/novel-drug-approvals-2024, Accessed: 2024-03-19.

(56) Deligiannidis, K. M.; Meltzer-Brody, S.; Maximos, B.; Peeper, E. Q.; Freeman, M.; Lasser, R.; Bullock, A.; Kotecha, M.; Li, S.; Forrestal, F.; Rana, N.; Garcia, M.; Leclair, B.; Doherty, J. *Am. J. Psychiatry* **2023**, *180*, 668–675.

(57) Novel Drug Approvals for 2023, https://www.fda.gov/drugs/new-drugs-fda-cders-new-molecular-entities-and-new-therapeutic-biological-products/novel-drug-approvals-2023, Accessed: 2024-03-19.

(58) Dogra, R.; Bhatia, R.; Shankar, R.; Bansal, P.; Rawal, R. K. *Anticancer Agents Med. Chem.* **2018**, *18*, 1936–1951.

(59) Taylor, A. M. et al. *J. Med. Chem.* **2023**, *66*, 13384–13399.

(60) Kuzmanic, A.; Bowman, G. R.; Juarez-Jimenez, J.; Michel, J.; Gervasio, F. L. *Acc. Chem. Res.* **2020**, *53*, 654–661.

(61) Smith, R. D.; Carlson, H. A. *J. Chem. Inf. Model.* **2021**, *61*, 1287–1299.

(62) Sun, Z.; Wakefield, A. E.; Kolossvary, I.; Beglov, D.; Vajda, S. *Structure* **2020**, *28*, 223–235.

(63) Borsatto, A.; Akkad, O.; Galdadas, I.; Ma, S.; Damfo, S.; Haider, S.; Kozielski, F.; Estarellas, C.; Gervasio, F. L. *eLife* **2022**, *11*, e81167.

(64) Wood, M. R.; Hopkins, C. R.; Brogan, J. T.; Conn, P. J.; Lindsley, C. W. *Biochemistry* **2011**, *50*, 2403–2410.

(65) Sadowsky, J. D.; Burlingame, M. A.; Wolan, D. W.; McClendon, C. L.; Jacobson, M. P.; Wells, J. A. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 6056–6061.

(66) Del Torrent, C. L.; Casajuana-Martin, N.; Pardo, L.; Tresadern, G.; Pérez-Benito, L. *J. Chem. Inf. Model.* **2019**, *29*, 2456–2466.

(67) Ni, D.; Chai, Z.; Wang, Y.; Li, M.; Yu, Z.; Liu, Y.; Lu, S.; Zhang, J. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *12*, e1585.

(68) Huang, W.; Lu, S.; Huang, Z.; Liu, X.; Mou, L.; Luo, Y.; Zhao, Y.; Liu, Y.; Chen, Z.; Hou, T.; Zhang, J. *Bioinformatics* **2013**, *29*, 2357–2359.

(69) Jendele, L.; Krivak, R.; Skoda, P.; Novotny, M.; Hoksza, D. *Nucleic Acids Res.* **2019**, *47*, W345–W349.

(70) Amor, B.; Yaliraki, S. N.; Woscholski, R.; Barahona, M. *Mol. BioSyst.* **2014**, *10*, 2247–2258.

(71) Guvench, O.; Jr, A. D. M. *PLOS Comput. Biol.* **2009**, *5*, e1000435.

(72) Oleinikovas, V.; Saladino, G.; Cossins, B. P.; Gervasio, F. L. *J. Am. Chem. Soc.* **2016**, *138*, 14257–14263.

(73) Cuchillo, R.; Pinto-Gil, K.; Michel, J. *J. Chem. Theory Comput.* **2015**, *11*, 1292–1307.

(74) Kokh, D. B.; Czodrowski, P.; Rippmann, F.; Wade, R. C. *J. Chem. Theory Comput.* **2015**, *12*, 4100–4113.

(75) Paola, L. D.; Ruvo, M. D.; Paci, P.; Santoni, D.; Giuliani, A. *Chem. Rev.* **2013**, *113*, 1598–1613.

(76) Yao, X.-Q.; Hamelberg, D. *J. Chem. Theory Comput.* **2022**, *18*, 1173–1187.

(77) Yao, X.-Q.; Momin, M.; Hamelberg, D. *J. Chem. Inf. Model.* **2018**, *58*, 1325–1330.

(78) Li, S.; Shen, Q.; Su, M.; Liu, X.; Lu, S.; Chen, Z.; Wang, R.; Zhang, J. *Bioinformatics* **2016**, *32*, 1574–1576.

(79) Ferraro, M.; Moroni, E.; Ippoliti, E.; Rinaldi, S.; Sanchez-Martin, C.; Rasola, A.; Pavarino, L. F.; Colombo, G. *J. Phys. Chem. B* **2021**, *125*, 101–114.

(80) Wang, Y.; Li, M.; Liang, W.; Shi, X.; Fan, J.; Kong, R.; Liu, Y.; Zhang, J.; Chen, T.; Lu, S. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 628–639.

(81) Zhang, H.; Ni, D.; Fan, J.; Li, M.; Zhang, J.; Hua, C.; Nussinov, R.; Lu, S. *J. Chem. Inf. Model* **2022**, *62*, 4222–4231.

(82) Hafner, J. *J. Comput. Chem.* **2008**, *29*, 2044–2078.

(83) Karplus, ¡ McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(84) Hollingsworth, S. A.; Dror, R. O. *Neuron* **2018**, *99*, 1129–1143.

(85) Ozer, G.; Valeev, E. F.; Quirk, S.; Hernandez, R. *J. Chem. Theory Comput.* **2010**, *6*, 3026–3038.

(86) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786–798.

(87) Cuendet, M. A.; van Gunsteren, W. F. *J. Chem. Phys.* **2007**, *127*, 184102.

(88) Badar, M. S.; Shamsi, S.; Ahmed, J.; Alam, M. A. In *Transdisciplinarity. Integrated Science, vol 5.* Rezaei, N., Ed.; Springer: Cham, 2022, pp 131–151.

(89) Karplus, M.; Petsko, G. A. *Nature* **1990**, *347*, 631–639.

(90) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.

(91) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; Jr, A. D. M. *Nat. Methods* **2017**, *14*, 71–73.

(92) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(93) Monticelli, L.; Tieleman, D. P. In *Biomolecular Simulations: Methods and Protocols*, Monticelli, L., Emppu, S., Eds.; Humana Press: Totowa, NJ, 2013, pp 197–213.

(94) Vymětal, J.; Vondrášek, J. *J. Chem. Theory Comput.* **2013**, *9*, 441–451.

(95) Vitalini, F.; Mey, A. S. J. S.; Noé, F.; Keller, B. G. *J. Chem. Phys.* **2015**, *142*, 084101.

(96) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. *Chem. Rev.* **2021**, *121*, 10142–10186.

(97) Rosenberger, D.; Smith, J. S.; Garcia, A. E. *J. Phys. Chem. B* **2021**, *125*, 3598–3612.

(98) Noé, F.; Fabritiis, G. D.; Clementi, C. *Curr. Opin. Struct. Biol.* **2020**, *60*, 77–84.

(99) Hansson, T.; Oostenbrink, C.; van Gunsteren, W. F. *Curr. Opin. Struct. Biol.* **2002**, *12*, 109–196.

(100) Sagui, C.; Darden, T. A. *Annu. Rev. Bioph. Biom.* **1999**, *28*, 155–179.

(101) Hünenberger, P. H. In *Advanced Computer Simulation*, Holm, C., Kremer, K., Eds.; Springer: Berlin, 2005, pp 105–149.

(102) Ke, Q.; Gong, X.; Liao, S.; Duan, C.; Li, L. *J. Mol. Liq.* **2022**, *365*, 120116.

(103) Barducci, A.; Bonomi, M.; Parrinello, M. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 826–843.

(104) Grubmuiller, H.; Heymann, B.; Tavan, P. *Science* **1996**, *271*, 997–999.

(105) Vivo, M. D.; Masetti, M.; Bottegoni, G.; Cavalli, A. *J. Med. Chem.* **2016**, *59*, 4035–4061.

(106) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

(107) Wang, J.; Arantes, P. R.; Bhattarai, A.; Hsu, R. V.; Pawnikar, S.; Huang, Y.-m. M.; Palermo, G.; Miao, Y. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *11*, e1521.

(108) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schutte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*.

(109) Konovalov, K. A.; Unarta, I. C.; Cao, S.; Goonetilleke, E. C.; Huang, X. *JACS Au* **2021**, *1*, 1330–1341.

(110) Chodera, J. D.; Noé, F. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135–144.

(111) Wang, W.; Cao, S.; Zhu, L.; Huang, X. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *8*, e1343.

(112) Husic, B. E.; Pande, V. S. *J. A. Chem. Soc* **2018**, *140*, 2386–2396.

(113) Swope, W. C.; Pitera, J. W. *J. Phys. Chem. B* **2004**, *108*, 6582–6594.

(114) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99–105.

(115) Juarez-Jimenez, J.; Gupta, A. A.; Karunanithy, G.; Mey, A. S. J. S.; Georgiou, C.; Ioannidis, H.; Simone, A. D.; Barlow, P. N.; Hulme, A. N.; Walkinshaw, M. D.; Baldwin, A. J.; Michel, J. *Chem. Sci.* **2020**, *11*, 2670–2680.

(116) Choy, M. S.; Li, Y.; Machado, L. E.; Kunze, M. B.; Connors, C. R.; Wei, X.; Lindorff-Larsen, K.; Page, R.; Peti, W. *Mol. Cell* **2017**, *65*, 644–658.

(117) Cui, D. S.; Beaumont, V.; Ginther, P. S.; Lipchock, J. M.; Loria, P. *J. Mol. Bio.* **2017**, *426*, 2360–2372.

(118) Cui, D. S.; Lipchock, J. M.; Brookner, D.; Loria, J. P. *J. Am. Chem. Soc.* **2019**, *141*, 12634–12647.

(119) Keedy, D. A.; Hill, Z. B.; Biel1, J. T.; Kang, E.; Rettenmaier, T. J.; Brandao-Neto, J.; Pearce, N. M.; von Delft, F.; Wells, J. A.; Fraser, J. S. *eLife* **2018**.

(120) Mehlman, T. S.; Biel, J. T.; Azeem, S. M.; Nelson, E. R.; Hossain, S.; Dunnett, L.; Paterson, N. G.; Douangamath, A.; Talon, R.; Axford, D.; Orins, H.; von Delft, F.; Keedy, D. A. *eLife* **2023**, *12*, e84632.

(121) Deribe, Y. L.; Pawson, T.; Dikic, I. *Nat. Struct. Mol. Biol.* **2010**, *17*, 666–672.

(122) Mann, M.; Jensen, O. N. *Nat. Biotecnol.* **2003**, *21*, 255–261.

(123) Stanford, S. M.; Bottini, N. *Nat. Rev. Drug Discov.* **2023**, *22*, 273–294.

(124) Liu, X.; Zhang, Y.; Wang, Y.; Yang, M.; Hong, F.; Yang, S. *Biomolecules* **2021**, *11*, 1009.

(125) Grimes, M.; Hall, B.; Foltz, L.; Levy, T.; Rikova, K.; Gaiser, J.; Cook, W.; Smirnova, E.; Wheeler, T.; Clark, N. R.; Lachmann, A.; Zhang, B.; Hornbeck, P.; Ma'ayan, A.; Comb, M. *Sci. Signal.* **2018**, *11*, eaaq1087.

(126) Perluigi, M.; Barone, E.; Domenico, F. D.; Butterfield, D. A. *Biochim. Biophys. Acta. Mol. Basis Dis.* **2016**, *1862*, 1871–1882.

(127) Liu, R.; Mathieu, C.; Berthelet, J.; Zhang, W.; Dupret, J.-M.; Lima, F. R. *Int. J. Mol. Sci.* **2022**, *23*, 7027.

(128) Brandão, T. A. S.; Hengge, A. C.; Johnson, S. J. *J. Biol. Chem.* **2010**, *285*, 15874–15883.

(129) Singh, J. P.; Lin, M.-J.; Hsu, S.-F.; Peti, W.; Lee, C.-C.; Meng, T.-C. *Int. J. Mol. Sci.* **2021**, *60*, 3856–3867.

(130) Wiesmann, C.; Barr, K. J.; Kung, J.; Zhu, J.; Erlanson, D. A.; Shen, W.; Fahr, B. J.; Zhong, M.; Taylor, L.; Randal, M.; McDowell, R. S.; Hansen, S. K. *Nat. Struct. Mol. Biol.* **2004**, *11*, 730–737.

(131) Shrestha, S.; Bhattarai, B. R.; Cho, H.; Choi, J.-K.; Cho, H. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2728–2730.

(132) Liang, S.; Tran, E.; Du, X.; Dong, J.; Sudholz, H.; Chen, H.; Qu, Z.; Huntington, N. D.; Babon, J. J.; Kershaw, N. J.; Zhang, Z.-Y.; Baell, J. B.; Wiede, F.; Tiganis, T. *Nat. Commun.* **2023**, *14*, 4524.

(133) Khan, S.; Bjij, I.; Soliman, M. E. S. *Cell Biochem. Biophys.* **2019**, *77*, 203–211.

(134) Flare, v. 5.0.0, Cresset, Litlington, Cambridgeshire, UK, http://www.cresset-group.com/flare/.

(135) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

(136) Hedges, L. O.; Mey, A. S.; Laughton, C. A.; Gervasio, F. L.; Mulholland, A. J.; Woods, C. J.; Michel, J. *J. Open Source Softw.* **2019**, *4*, 1831.

(137) Steinbrecher, T.; Latzer, J.; Case, D. A. *J. Chem. Theory Comput.* **2012**, *8*, 4405–4412.

(138) Abraham, M.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.; Hess, B.; Lindahl, E. *SoftwareX* **2015**, *1*, 19–25.

(139) Case, D. et al., AMBER 2020, 2020.

(140) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

(141) Roe, D. R.; Cheatham, T. E. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.

(142) Wehmeyer, C.; Scherer, M. K.; Hempel, T.; Husic, B. E.; Olsson, S.; Noé, F. *Living J. Comp. Mol. Sci* **2019**, *1*.

(143) Thayer, K. M.; Lakhani, B.; Beveridge, D. L. *J. Phys. Chem.* **2017**, *121*, 5509–5514.

(144) Röblitz, S.; Weber, M. *Adv. Data Anal. Classif.* **2013**, *7*, 147–179.

(145) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. *Nat. Commun.* **2018**, *9*.

(146) Jr, T. E. K.; Pardo, A.; Selman, M. *Lancet* **2011**, *378*, 1949–1961.

(147) Cao, X.; Li, Y.; Shi, J.; Tang, H. *J. Inflamm. Res.* **2021**, *14*, 611–619.

(148) Raghu, G.; Chen, S.-Y.; Yeh, W.-S.; Maroni, B.; Li, Q.; Lee, Y.-C.; Collard, H. R. *Lancet Respir. Med.* **2014**, *2*, 566–572.

(149) Patrucco, F.; Solidoro, P.; Gavelli, F.; Apostolo, D.; Bellan, M. *Microorganisms* **2023**, *11*, 895.

(150) Rooij, J. D.; Zwartkruis, F. J.; Verheijen, M. H.; Cool, R. H.; Nijman, S. M.; Wittinghofer, A.; Bos, J. L. *Nature* **1998**, *396*, 474–477.

(151) Boettner, B.; Aelst, L. V. *Curr. Opin. Struct. Biol.* **2009**, *21*, 684–693.

(152) Li, Q.; Teng, Y.; Wang, J.; Yu, M.; Li, Y.; Zheng, H. *Pathol. Res. Pract.* **2018**, *214*, 1045–1050.

(153) Rooij, J. D.; Rehmann, H.; Triest, M. V.; Cool, R. H.; Wittinghofer, A.; Bos, J. L. *J. Biol. Chem.* **2000**, *275*, 20829–20836.

(154) Pereira, L.; Cheng, H.; Lao, D. H.; Na, L.; van Oort, R. J.; Brown, J. H.; Wehrens, X. H.; Chen, J.; Bers, D. M. *Circulation* **2013**, *127*, 913–922.

(155) Huang, S. K.; Wettlaufer, S. H.; Chung, J.; Peters-Golden, M. *Am. J. Respir. Cell Mol. Biol.* **2008**, *39*, 482–489.

(156) Zieba, B. J.; Artamonov, M. V.; Jin, L.; Momotani, K.; Ho, R.; Franke, A. S.; Neppl, R. L.; Stevenson, A. S.; Khromov, A. S.; Chrzanowska-Wodnicka, M.; Somlyo, A. V. *J. Biol. Chem.* **2011**, *286*, 16681–16692.

(157) Conrotto, P.; Yakymovych, I.; Yakymovych, M.; Souchelnytskyi, S. *J. Proteome Res.* **2007**, *26*, 287–297.

(158) Bos, J. L.; Rehmann, H.; Wittinghofer, A. *Cell* **2007**, *129*, 865–877.

(159) Rehmann, H.; Arias-Palomo, E.; Hadders, M. A.; Schwede, F.; Llorca, O.; Bos, J. L. *Nature* **2008**, *455*, 124–127.

(160) Rehmann, H.; Prakash, B.; Wolf, E.; Rueppel, A.; de Rooij, J.; Bos, J. L.; Wittinghofer, A. *Nat. Struct. Biol.* **2003**, *10*, 26–32.

(161) Rehmann, H.; Das, J.; Wittinghofer, A.; Bos, J. L. *Nature* **2006**, *439*, 625–628.

(162) Shao, H.; Mohamed, H.; Boulton, S.; Huang, J.; Wang, P.; Chen, H.; Zhou, J.; Luchowska-Stańska, U.; Jentsch, N. G.; Armstrong, A. L.; Magolan, J.; Yarwood, S.; Melacini, G. *J. Med. Chem.* **2020**, *63*, 4762–4775.

(163) Kraemer, A.; Rehmann, H. R.; Cool, R. H.; Theiss, C.; Rooij, J. D.; Bos, J. L.; Wittinghofer, A. *J. Mol. Biol.* **2001**, *306*, 1167–1177.

(164) Enserink, J. M.; Christensen, A. E.; de Rooij, J.; van Triest, M.; Schwede, F.; Genieser, H. G.; Døskeland, S. O.; Blank, J. L.; Bos, J. L. *Nat. Cell Biol.* **2002**, *4*, 901–906.

(165) Kang, G.; Joseph, J. W.; Chepurny, O. G.; Monaco, M.; Wheeler, M. B.; Bos, J. L.; Schwede, F.; Genieser, H. G.; Holz, G. G. *J. Biol. Chem.* **2003**, *278*, 8279–8285.

(166) Vliem, M. J.; Ponsioen, B.; Schwede, F.; Pannekoek, W.-J.; Riedl, J.; Kooistra, M. R. H.; Jalink, K.; Genieser, H.-G.; Bos, J. L.; Rehmann, H. *ChemBioChem* **2008**, *9*, 2052–2054.

(167) Schwede, F.; Bertinetti, D.; Langerijs, C. N.; Hadders, M. A.; Wienk, H.; Ellenbroek, J. H.; de Koning, E. J. P.; Bos, J. L.; Herberg, F. W.; Genieser, H.-G.; Janssen, R. A. J.; Rehmann, H. *PLOS Biology* **2015**, *13*, e1002038.

(168) Yuan, Y.; Engler, A. J.; Raredon, M. S.; Le, A.; Baevova, P.; Yoder, M. C.; Niklason, L. E. *Biomaterials* **2019**, *200*, 25–34.

(169) Rehmann, H.; Schwede, F.; Døskeland, S. O.; Wittinghofer, A.; Bos, J. L. *J. Biol. Chem.* **2003**, *278*, 38548–38556.

(170) Métrich, M.; Berthouze, M.; Morel, E.; Crozatier, B.; Gomez, A. M.; Lezoualc'h, F. *Pflüg. Arch. Eur. J. Physiol.* **2010**, *459*, 535–546.

(171) Herfindal, L.; Nygaard, G.; Kopperud, R.; Krakstad, C.; Døskeland, S. O.; Selheim, F. *Biochem. Biophys. Res. Commun.* **2013**, *437*, 603–608.

(172) Barker, G.; Parnell, E.; van Basten, B.; Buist, H.; Adams, D. R.; Yarwood, S. J. . *Cardiovasc. Dev. Dis.* **2017**, *4*, 22.

(173) Parnell, E.; Mcelroy, S. P.; Wiejak, J.; Baillie, G. L.; Porter, A.; Adams, D. R.; Rehmann, H.; Smith, B. O.; Yarwood, S. J. *Sci. Rep.* **2017**, *7*, 294.

(174) Hardie, A.; Cossins, B. P.; Lovera, S.; Michel, J. *Commun. Chem.* **2023**, *6*, 125.

(175) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. *Nucleic Acids Res.* **2018**, *46*, 296–303.

(176) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

(177) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

(178) The PLUMED Consortium *Nat. Methods* **2019**, *16*, 670–673.

(179) Wehmeyer, C.; Scherer, M. K.; Hempel, T.; Husic, B. E.; Olsson, S.; Noé, F. *Living J. Comp. Mol. Sci* **2019**, *1*.

(180) Mohamed, H.; Shao, H.; Akimoto, M.; Darveau, P.; MacKinnon, M. R.; Magolanb, J.; Melacini, G. *RSC Chem. Biol.* **2022**, *3*, 1230–1239.

(181) Wright, S. H. *Adv. Physiol. Educ.* **2004**, *28*, 139–142.

(182) Benarroch, J. M.; Asally, M. *Trends Microbiol.* **2020**, *28*, 304–314.

(183) Maffeo, C.; Bhattacharya, S.; Yoo, J.; Wells, D.; Aksimentiev, A. *Chem. Rev.* **2012**, *112*, 6250–6284.

(184) Levental, I.; Lyman, E. *Nat. Rev. Mol. Cell Biol.* **2023**, *24*, 107–122.

(185) Thompson, M. J.; Baenziger, J. E. *Nat. Chem. Biol.* **2020**, *16*, 1331–1342.

(186) Simunovic, M.; Prévost, C.; Callan-Jones, A.; Bassereau, P. *Philos. Trans. A Math. Phys. Eng. Sci.* **2016**, *374*, 20160034.

(187) Cornelius, F. *Biochemistry* **2001**, *40*, 8842–8851.

(188) Duncan, A. L.; Reddy, T.; Koldsø, H.; Hélie, J.; Fowler, P. W.; Chavent, M.; Sansom, M. S. P. *Sci. Rep.* **2017**, *7*, 16647.

(189) Hansen, S. B.; Tao, X.; MacKinnon, R. *Nature* **2011**, *447*, 495–498.

(190) Ye, W.; Han, T. W.; Nassar, L. M.; Zubia, M.; Jan, Y. N.; Jan, L. Y. *PNAS* **2018**, *115*, E1667–E1674.

(191) Minke, B.; Wu, C.-F.; Pak, W. L. *Nature* **1975**, *258*, 84–87.

(192) Zhang, M.; Ma, Y.; Ye, X.; Zhang, N.; Pan, L.; Wang, B. *Sig. Transduct. Target. Ther.* **2023**, *8*, 261.

(193) Rosenbaum, T.; Morales-Lázaro, S. L.; Islas, L. D. *Nat. Rev. Neurosci.* **2022**, *23*, 596–610.

(194) Nilius, B.; Owsianik, G. *Genome Biol.* **2011**, *12*, 218.

(195) Voets, T.; Janssens, A.; Droogmans, G.; Nilius, B. *J. Biol. Chem.* **2004**, *279*, 15223–15230.

(196) Nilius, B.; Prenen, J.; Janssens, A.; Wang, C.; Zhu, M. X.; Voets, T. *J. Biol. Chem.* **2005**, *280*, 22899–22906.

(197) Cao, E. *J. Gen. Physiol.* **2020**, *152*, e201811998.

(198) Zheng, W.; Yang, X.; Hu, R.; Cai, R.; Hofmann, L.; Wang, Z.; Hu, Q.; Liu, X.; Bulkley, D.; Yu, Y.; Tang, J.; Flockerzi, V.; Cao, Y.; Cao, E.; Chen, X.-Z. *Nat. Commun.* **2018**, *9*, 2302.

(199) Everson, G. T. *Mayo Clin. Proc.* **1990**, *65*, 1020–1025.

(200) Wu, G.; D'Agati, V.; Cai, Y.; Markowitz, G.; Park, J. H.; Reynolds, D. M.; Maeda, Y.; Le, T. C.; Jr., H. H.; Kucherlapati, R.; Edelmann, W.; Somlo, S. *Cell* **1998**, *93*, 177–188.

(201) Qian, F.; Watnick, T. J.; Onuchic, L. F.; Germino, G. G. *Cell* **1996**, *87*, 979–987.

(202) Wilkes, M. et al. *Nat. Struct. Mol. Biol.* **2017**, *24*, 123–130.

(203) Pavel, M. A.; Lv, C.; Ng, C.; Yang, L.; Kashyap, P.; Lam, C.; Valentino, V.; Fung, H. Y.; Campbell, T.; Møller, S. G.; Zenisek, D.; Holtzman, N. G.; Yu, Y. *PNAS* **2016**, *113*, E2363–E2372.

(204) Ma, R.; Li, W.-P.; Rundle, D.; Kong, J.; Akbarali, H. I.; Tsiokas, L. *Mol. Cell Biol.* **2004**, *25*, 8285–8298.

(205) Wang, Q.; Corey, R. A.; Hedger, G.; Aryal, P.; Grieben, M.; Nasrallah, C.; Baronina, A.; Pike, A. C.; Shi, J.; Carpenter, E. P.; Sansom, M. S. *Structure* **2020**, *28*, 169–184.e5.

(206) Hole, https://www.holeprogram.org/.

(207) Šali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779–815.

(208) Wu, E. L.; Cheng, X.; Jo, S.; Rui, H.; Song, K. C.; Dávila-Contreras, E. M.; Qi, Y.; Lee, J.; Monje-Galvan, V.; Venable, R. M.; Klauda, J. B.; Im, W. *J. Comp. Chem.* **2014**, *25*, 1997–2004.

(209) Orientations of Proteins in Membranes (OPM) database, https://opm.phar.umich.edu/, Accessed: 2024-03-04.

(210) Borges-Araújo, L.; Fernandes, F. *Molecules* **2020**, *25*, 3885.

(211) Shirts, M. R.; Klein, C.; Swails, J. M.; Yin, J.; Gilson, M. K.; Mobley, D. L.; Case, D. A.; Zhong, E. D. *J. Comput. Aided Mol. Des.* **2016**, *31*, 147–161.

(212) Luo, L.; Roy, S.; Li, L.; Ma, M. *Trends. Mol. Med.* **2023**, *29*, 268–281.

(213) Vien, T. N.; Wang, J.; Ng, L. C. T.; Cao, E.; DeCaen, P. G. *PNAS* **2020**, *117*, 10329–10338.

(214) Zheng, L.; Fan, J.; Mu, Y. *ACS Omega* **2019**, *4*, 15956–15965.