Abstract

Computer aided drug design (CADD) plays a crucial role in the drug discovery pipeline e.g. in virtual screening of chemical databases, *de novo* drug design, and lead optimization. Due to the increased numbers of protein structures elucidated, structure-based methods for developing pharmacophore models have started gaining in popularity and are becoming of particular importance. There have been a number of studies combining such methods with the use of molecular dynamics (MD) simulations to model protein flexibility. In this project, the development and application of a new methodology, based on coarse grained (CG) MD, through the use of the MARTINI forcefield, and employed to explore protein ligand interactions, will be presented. An overview of the history of CADD is presented, along with current computational fragment based methods available for exploring protein-ligand interactions. An overview of the theory and methods behind MD simulations both all atom and CG is also provided. In the first results chapter, the parametrization of MARTINI beads as pharmacophoric probes, the analysis protocol and the application of this method to a data set of water soluble targets of pharmacological interest is described. The results suggest that the pharmacophoric probes have the ability to identify protein-ligand interactions on the targets of interest. The probes are also able to identify the residues involved in forming ligand binding interactions, showing a particular accuracy in identifying "hotspot" interactions. In the second results chapter, the extension of the initial data set to a range of GPCRs is described. The results suggest that the pharmacophoric probes have the ability to accurately explore both the orthosteric and allosteric binding sites of the GPCR targets and accurately identify the interactions and residues involved in ligand binding. This is done without the need to embed the protein in a lipid bilayer. In the final results chapter, the application of the dynamic pharmacophoric probes to identifying PIP_2 and cholesterol binding sites,

on membrane proteins, is presented. The results suggest that the probes can indeed identify these binding sites, along with identifying the residues that are involved in binding the PIP_2 head groups. This was also done without embedding the protein in a lipid bilayer, which is the usual practice for identifying lipid binding interactions, reducing the computational resource requirements.

Acknowledgements

Firstly, my thanks go to my academic supervisors, Prof. Mark Sansom, Dr. Phillip Stansfeld and Dr. Zara Sands, for their support and advice over these past four years. This project would not have been a reality without their guidance. Most importantly, I am grateful to them for creating an environment that allowed me to grow both as a person and as a scientist. They have brought together a group of extraordinary and talented people, making SBCB an ideal place to work and generate new ideas.

The lab in its current form is much different from the lab I joined in 2014. From the "old guard", a special mention to Dr. Tere Paramo, Marc Daemgen, Dr. George Gerogiokas, Dr. Maria Musgaard, Dr. Fiona Naughton, Dr. Tom Newport, Dr. Anna Duncan, Dr. Amanda Buyan, Dr. Ole Andersen and Dr. George Hedger. Since my first day, they made me feel especially welcome, along with providing hours of entertainment, tonnes of baked goods, numerous coffee breaks and exceptional project management skills.

From the current lab members, I have the daily privilege to interact with and learn from an extremely gifted and motivated group of people. My days would be much more mundane, and my simulations would run much more slowly, without Dr. Owen Vickery, Shanlin Rao, Tom Dixon, Max Epstein, William Glass, Dr. Charlotte Lynch, Dr. Irfan Alibay, Dr. Aphroditi Zaki, Josh Sauer, Dr. Robin Corey, Gianni Klesse and Dr. Wanling Song. A particular shout out to Michael Horell, for his excellent printing skills, and to my thesis writing comrades, Naushad Velgi and Viwan Jarerattanachat. My thanks also go to to Lizzie Jefferys, for making sure my computer is always busy and running efficiently.

Outside of the lab, I have been blessed with an incredible group of friends. Firstly, a huge thank you to all my Greek and adopted-Greek friends. I would not have managed to cope with the stress and pressure of this DPhil without the football

games, nights out, lunch breaks and all other activities. Secondly, to all my Wolfson and non-Wolfson friends and especially to Marlene Spangeberg, Jack Tan and Camilla Mella San Martin, for all the memories that I will always cherish.

Last but not least, I would like to thank my parents and my girlfriend for their continuous love and support in every endeavour in my life, whether this is my DPhil or going scuba diving.

List of Abbreviations

CADD	Computer Aided Drug Design
Cryo-EM	Cryo Electron Microscopy
CG	Coarse Grain
CRD	Cysteine Rich Domain
ECL	Extracellular
ENM	Elastic Network Model
FB	Fragment Based
FEP	Free Energy Perturbations
GCMC	Grand Canonical Monte Carlo
GPCR	G Protein-Coupled Receptor
GROMACS	Groningen Machine for Chemical Simulations
HTS	High Throughput Screening
ICL	Intracellular
LB	Ligand Based
MCSS	Multiple Copy Simultaneous Search
MD	Molecular Dynamics
MM	Molecular Mechanics
MSCS	Multiple Solvent Crystal Structure
NMR	Nuclear Magnetic Resonance

PBC	Periodic Boundary Conditions
PCA	Principal Component Analysis
PIP_2	Phosphatidylinositol-4,5-bisphosphate
PME	Particle Mesh Ewald
PPI	Protein Protein Interactions
QSAR	Quantitative Structure-Activity Relationship
RILB	Residue Important for Ligand Binding
SB	Structure Based
TM	Transmembrane

Contents

1 Introduction **15**

1.1 Overview 15

1.2 CADD in the Drug Discovery Pipeline 16

1.3 Ligand Based Drug Design 19

1.4 Structure Based Drug Design 19

1.4.1 Identifying Binding Sites on Protein Structures 20

1.5 Fragment Based Drug Design 23

1.5.1 Mixed Solvent Molecular Dynamics Based Methods 25

1.5.2 MDmix 26

1.5.3 SILCS 27

1.5.4 MixMD 32

1.5.5 Yang and Wang Method 33

1.5.6 Bahar and GlaxoSmithKline 34

1.5.7 Caflisch Approach 35

1.5.8 Tan and Abell Approach 36

1.5.9 Fersht Approach 37

1.5.10 pMD 37

1.5.11 Overall Conclusions from the co-solvent MD Approaches . . . 38

1.5.12 Arcon Approach 38

1.6 Pharmacophore Models 39

1.6.1 Pharmacophore Models in Drug Design 41

1.6.2 Pharmacophore Model Creation 42

1.7 Aims and Motivations 43

2 Theory and Methods **45**

2.1 Molecular Dynamic Simulations 45

2.2 Forcefields 46

2.2.1 Bonded Interactions 47

2.2.2 Non-bonded Interactions 49

2.2.3 Energy Minimization 50

2.2.4 Integration Methods 51

2.2.5 Choice of Time Step 52

2.2.6 Bond Constraints 52

2.2.7 Periodic Boundary Conditions 53

2.2.8 Temperature and Pressure Coupling 54

2.2.9 Long Range Interactions 56

2.3 Coarse Grained Simulations 57

2.3.1 The MARTINI forcefield 58

2.3.2 Secondary and Tertiary Structure in MARTINI 61

2.3.3 Water Model 62

3 Dynamic Pharmacophore Probes **64**

3.1 Introduction 64

3.2 Methods 66

3.2.1 Target dataset 66

3.2.2 Molecular Dynamics simulations 66

3.2.3 Forcefield 68

3.2.4 MARTINI beads as pharmacophoric probes 69

3.2.5 Simulation protocol and analysis 70

3.2.6 Convergence 71

3.2.7 Clustering and pharmacophoric points 77

3.2.8 Contact analysis 77

3.2.9 Protein scoring 81

3.3 Results 82

3.3.1 Comparing the interactions identified through the use of the pharmacophoric probes with the ligand bound structures . . . 82

3.3.2 Quantifying the identification of interactions 85

3.3.3 Identification of residues involved in binding 88
3.3.4 Quantitative analysis of the predictive powers of the probes . 93
3.4 Discussion 95
3.4.1 Reasons for probe failure 95
3.4.2 Advantages and overcoming the limitations 97
3.5 Conclusions 98

4 Dynamic Pharmacophore Probes: Application to GPCRs 100
4.1 Introduction 100
4.1.1 GPCR families and structure 101
4.1.2 Rhodopsin like receptors (class A) 101
4.1.3 Secretin and Adhesion receptors (class B) 103
4.1.4 Glutamate receptors (class C) 103
4.1.5 Frizzled and Taste-2 receptors (class F) 103
4.1.6 GPCR signalling 104
4.1.7 Drugging GPCRs 104
4.2 Methods 108
4.3 Results 111
4.3.1 Identification of protein-ligand interactions 111
4.3.2 M2 Muscarinic Acetylcholine Receptor 111
4.3.3 Chemokine Receptor 9 111
4.3.4 β2-Adrenoceptor 113
4.3.5 Sphingosine 1-phosphate receptor 114
4.3.6 Quantifying the identification of interactions 117
4.3.7 Identifying the residues involved in ligand binding 118
4.3.8 Quantitative evaluation of the predictive abilities of the probes 122
4.3.9 Discussion 127
4.3.10 Conclusions 128

5 Identification of Lipid Binding Sites 129
5.1 Introduction 129
5.1.1 Phosphoinositides and Cholesterol 131
5.1.2 Inward Rectifying Potassium (Kir) Channel 2.2 132

5.1.3 Neurotensin Receptor 1 . . . 132
5.1.4 Adenosine A_2A Receptor . . . 135
5.1.5 Smoothened . . . 136
5.1.6 Polycystin-2 . . . 136
5.2 Methods . . . 137
5.2.1 Modelling and System Set Up . . . 137
5.3 Results . . . 139
5.3.1 Kir 2.2 . . . 139
5.3.2 Neurotensin Receptor 1 . . . 142
5.3.3 A_2A Receptor . . . 148
5.3.4 Smoothened . . . 153
5.3.5 Polycystin-2 . . . 157
5.4 Discussion . . . 159
5.5 Conclusions . . . 161

6 Future Directions **162**
6.1 Parametrization Upgrade . . . 162
6.2 Conformational Flexibility . . . 163
6.3 Identifying Cryptic Pockets . . . 163
6.4 Targeting Protein Protein Interactions . . . 164
6.5 Application to CADD Projects . . . 164
6.6 PIP_2 Interaction Patterns on GPCRs . . . 164
6.7 Final Remarks . . . 165

List of Figures

1.1 The role of CADD in Drug Design 18
1.2 The Use of Grids in MD 26
1.3 An Example of Interaction Densities Produced by the MDMix Approach 28
1.4 Repulsive vs No Repulsive Forces in SILCS 29
1.5 LGFE Maps in SILCS 30
1.6 An Example of a Pharmacophore Model 41

2.1 The Typical Energy Functions Used in a Biomolecular Forcefield . . . 47
2.2 A Two Dimensional Representation of the PBC Concept 53
2.3 Amino Acids in Martini 59
2.4 Amino Acids in Martini 2 60

3.1 The Ichihara Dataset 67
3.2 Simulation Work Flow 69
3.3 Convergence 72
3.4 Convergence 2 73
3.5 Clustering of Interaction Densities 80
3.6 Aurora A Interactions 82
3.7 Pantothenate Synthetase Interactions 83
3.8 UPA Interactions 84
3.9 Aurora A Residues 89
3.10 Pantothenate Synthetase Residues 90
3.11 UPA Residues 91

4.1 Schematic of a Typical GPCR 102
4.2 Signalling Cascades Controlled by GPCRs 105

4.3 An Example of a Class B GPCR 106
4.4 The GPCR Dataset Used 109
4.5 M2 Allosteric Binding Site Interactions 112
4.6 CCR9 Allosteric Binding Site Interactions 113
4.7 β2-Adrenoceptor Allosteric Binding Site Interactions 114
4.8 β2-Adrenoceptor Orthosteric Binding Site Interactions 115
4.9 S1P1 Orthosteric Binding Site Interactions 115
4.10 M2 Allosteric Binding Site Residues 120
4.11 CCR9 Allosteric Binding Site Residues 121
4.12 β2-Adrenoceptor Allosteric Binding Site Residues 123
4.13 β2-Adrenoceptor Orthosteric Binding Site Residues 124
4.14 S1P1 Orthosteric Binding Site Residues 125

5.1 PIP_2 and Cholesterol 133
5.2 Kir2.2 Structures with PIP_2 bound 134
5.3 The Structures of the Targets Used 138
5.4 PIP_2 Binding Site Identified on Kir2.2 141
5.5 Residues Identified Being in Contact with PIP_2 on Kir2.2 141
5.6 PIP_2 Headgroup and Tail Identified on Kir2.2 143
5.7 PIP_2 Binding Sites on NTS1 Between TM1 and TM2 144
5.8 PIP_2 Binding Sites on NTS1 on TM4 145
5.9 PIP_2 Binding Sites on NTS1 Between on TM6 146
5.10 PIP_2 Binding Sites on NTS1 on TM7 and TM8 147
5.11 PIP_2 Binding Sites on A_2A Between TM2, TM3 and TM4 148
5.12 PIP_2 Binding Sites on A_2A Between TM3 and TM4 149
5.13 PIP_2 Binding Sites on A_2A Between TM4 and TM5 150
5.14 PIP_2 Binding Sites on A_2A Between TM5, TM6 and TM7 151
5.15 Possible Cholesterol Binding Sites on A_2A 152
5.16 Smoothened CRD Domain 153
5.17 Smoothened Hydrophobic Cage 154
5.18 Possible Cholesterol Binding Sites on Smoothened 155
5.19 Suggested Tunnel Region on Smoothened 156
5.20 [Possible PIP_2 Binding Sites on PC2 158

5.21 Possible Cholesterol Binding Site on PC2 160

List of Tables

1.1 Summary of the Different Computational Fragment Based Approaches 40

3.1 The Ichihara Dataset and its Characteristics 68
3.2 The KS scores and p values of the Convergence Analysis 75
3.3 The Residues Involved in Ligand Binding 77
3.4 Fragment Scores 86
3.5 Lead Scores 87
3.6 Quantitative Analysis of the Predictive Capabilities of the Probes . . 93

4.1 A Summary of the Simulations Run 110
4.2 Orthosteric Binding Sites Interactions 116
4.3 Allosteric Binding Sites Interactions 117
4.4 Residues Involved in Ligand Binding in Orthosteric Binding Sites . . 118
4.5 Residues Involved in Ligand Binding in Allosteric Binding Sites . . . 119
4.6 Quantitative Analysis of the Predictive Powers of the Probes in Orthosteric Binding Sites 126
4.7 Quantitative Analysis of the Predictive Powers of the Probes in Allosteric Binding Sites 126

5.1 Summary of the Simulations Run 137

Chapter 1

Introduction

1.1 Overview

Historically, "endogenous bioactive ligands were discovered long before the structure of their biomolecular receptor could be identified" [1]. Information on the chemical structures of the ligands was also scarce and rudimentary, with early discoveries such as the polycyclic carbon framework of steroids being based on the stepwise pyrolytic breakdown of their structure. With the introduction of X-ray diffraction and the revelation of the three dimensional arrangement of atoms and ligands, a new era in drug discovery begun [1]. Nowadays, drug discovery has a plethora of tools in its disposal, with structures of pharmacological targets and ligands playing a prominent role amongst them. In exploiting these target-ligand interactions, the use of computers revolutionised the field of drug discovery and opened a great realm of possibilities for the development of new pharmaceuticals [2, 3].

An early mention of the widespread use of computers in drug discovery came on October 5, 1981, when *Fortune* magazine published an article with the title "Next Industrial Revolution: Designing Drugs by Computer at Merck" [4]. This has been credited as the beginning of an intense interest in Computer Aided Drug Design (CADD). However, with the development of high throughput screening (HTS), CADD was temporarily overtaken as the preferred method for the identification of novel therapeutics [5].

HTS entails the automated screening of large numbers of molecules against a target of interest, aiming to identify the compounds that elicit a desired biological

response. HTS has the advantage of well developed and efficient methods being readily available, along with a minimal need for prior knowledge and compound design. Nonetheless, the available HTS methods provide an extremely low hit ratio for potential leads, limiting them to projects where the screening of libraries with huge numbers of compounds are available [5, 6]. Furthermore, development of new HTS methods requires considerable development and validation before they can be utilised. These drawbacks allowed for the re-emergence of CADD as a potential way of reducing the number of compounds needed for screening, while maintaining the same hit ratio and level of lead compound discovery [5, 6]. Moreover, CADD can be used in conjunction or in parallel with HTS. In the first case, it can be used to identify compounds which are predicted to be inactive, significantly lowering the workload and cost of HTS. In the latter case, it can be run whilst HTS is being prepared or already running, allowing for the consensus of the resulting compounds to be identified and selected for further development.

A few examples of pharmaceuticals developed through the use of CADD include: the carbonic anhydrase inhibitor dorzolamide, approved as a drug in 1995 [7]; the ACE inhibitor captopril, approved in 1981; [8, 9] and the antiretrovirals saquinavir, ritonavir and indinavir, which were approved in 1995, 1996 and 1996 respectively for the treatment of HIV [10].

1.2 CADD in the Drug Discovery Pipeline

CADD tools and methods offer the possibility of increasing the hit ratio of new compounds due to their much more targeted search, compared to the brute force approach of HTS methods. In a drug development project, CADD methods are used with three main purposes in mind: firstly, to generate initial hits from the large compound libraries available, which will then be tested experimentally; secondly, to assist and guide the optimization of lead compounds by increasing their affinity or improving their pharmacokinetic properties, or reducing toxicity and increasing selectivity; thirdly, to being used in *de novo* drug design, either by synthesising or growing initial compounds or by linking and growing fragments which are known to bind to the target [11, 12].

CADD methods can be categorised into two main areas, Ligand Based (LB) and Structure Based (SB). LB methods depend on the knowledge of characterised active or inactive compounds. This knowledge is applied either through the use of similarity and fitting algorithms, or the construction of quantitative structure-activity relationship (QSAR) models, for the prediction of potential hit or lead compounds [12, 13]. SB methods rely on the availability of the structure of the target of interest, preferably of a high resolution, in order to extract the chemical information of potential binding sites, or calculating the interaction energies of possible compounds. SB methods are usually preferred when the structure of the target is available, such as in cases of small, soluble proteins, whilst LB methods are chosen when no target structures, or only structures of inadequate resolution, are available [13]. This has been the case for G-Protein Coupled Receptor targets, up until recently. A successful CADD campaign will lead to the generation of hit and lead compounds, which can then be optimised and subsequently turned into potential drug candidates. Figure 1.1 shows the role of CADD in the drug development pipeline, along with the different methods which fall under its umbrella.

The most prevalent uses of CADD are firstly virtual screening, also known as virtual high throughput screening (vHTS) and secondly, *de novo* drug design. With regards to vHTS, the resources needed for a traditional HTS can be significantly reduced, due to the ability of vHTS to discard compounds that have been predicted to be inactive. This approach reduces the number of compounds to be put through the HTS, without affecting the number of hits. An early mention of vHTS was in 1997 by Horvath *et al* [14]. In the case of *de novo* drug design, the guiding idea is that the synthesis of potential novel compounds which will possess the desired properties for binding to the target. Given the huge search space, 10^{60} potential compounds, a variety of algorithms can be used to try and ensure that the compounds predicted have high activity, no undesirable properties, and screen for a variety of other qualities [15]. Compounds which are already shown to have the desired properties can be used as scaffolds for the synthesis of new molecules. Two types of algorithms exist for this application. Linking algorithms work by docking small functional groups in binding sites and then linking the groups from adjacent sites. Growing algorithms begin from a single fragment which is then grown and modified to fulfil the desired

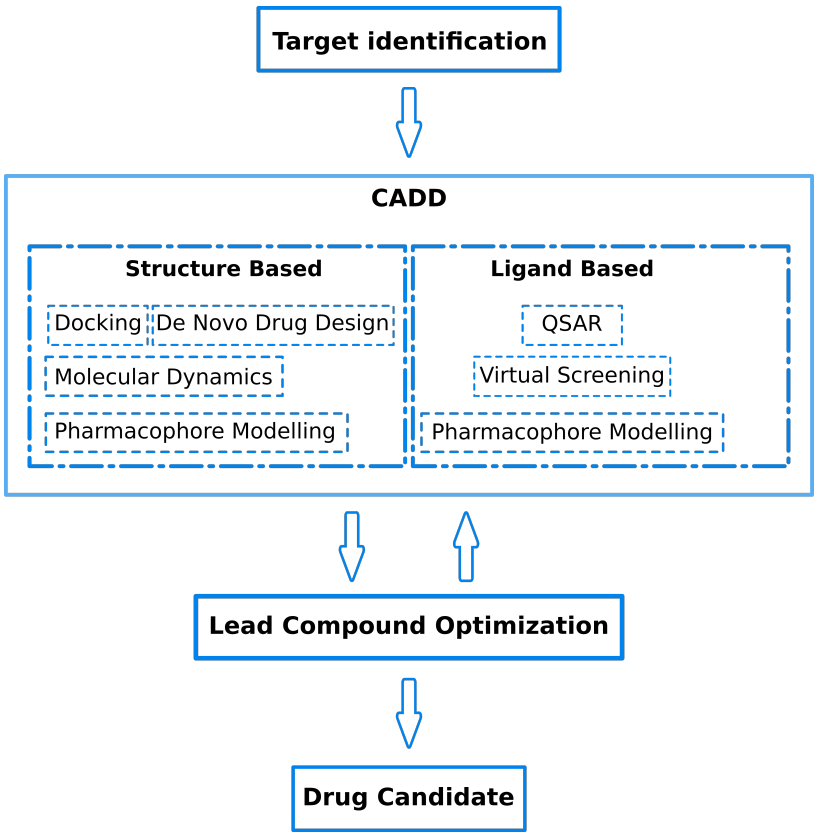


Figure 1.1: The role of CADD in the drug discovery pipeline. The therapeutic target is first identified and the best suitable CADD approach, depending on the structural information available, is selected. Assuming a successful CADD attempt, lead compounds will be generated and optimised, leading to potential drug candidates.

criteria [16, 17].

1.3 Ligand Based Drug Design

Ligand based CADD methods involve analysing ligands which are known to interact with a target of interest [18]. A set of compounds which conform to this assumption are selected and their 2D or 3D structures are analysed [19]. The aim is to present them in such a manner that the information, common amongst these compounds and important for the desired property, is retained while any non-relevant information is discarded [18, 19].

One of the challenges of LB methods is finding a way to extract the optimum amount of chemical information from the aforementioned compounds. The amount and type of information required varies, and as a result a range of algorithms that use different molecular descriptors has been developed [19, 20]. These descriptors can be either structural or chemical, and of varying complexity. The most common properties used are molecular weight, volume, surface area, geometry and atom types, amongst others [20]. As can be expected, selecting the most suitable descriptors can be quite challenging, affecting the resulting predicted molecules.

One important application of LB CADD methods is the creation of pharmacophore models. A pharmacophore model shows the common binding properties between different ligands which are capable of binding to the same binding site of a receptor. Through the alignment of the set of known binders, the chemical information is extracted and turned into a pharmacophore model [21]. The topic of pharmacophore models will be more thoroughly examined in subsequent sections.

1.4 Structure Based Drug Design

As the number of elucidated crystal structures increased over the years, SB drug design approaches gained in popularity within the field of CADD [5]. SB-CADD methods depend on the ability to detect and analyse potential ligand binding sites on the 3D structures of the target of interest [21]. This approach is based on the hypothesis that a molecule will only interact with a specific target, and exert the

desired biological effect, if it can favourably interact with one or more of the target's binding sites [22].

1.4.1 Identifying Binding Sites on Protein Structures

Identifying potential binding sites is therefore of paramount importance in SB-CADD. Experimentally, binding sites can be identified by the co-crystallization of ligands on the structure of interest, or on a structure related to that of the structure of interest [23]. In the cases where co-crystal structures are not available, mutational studies can be used to pinpoint the location of potential binding sites [24]. By mutating the residues thought to be important for ligand binding, and measuring the binding of the ligand, hypotheses can be developed for the location of a binding site. In the cases where no ligand binding data is available, or a novel binding site needs to be identified, e.g., in the cases of allosteric molecules, computational methods can be used to explore the possibilities [24].

Computational methods used to detect binding sites can be split into three categories, based on the principle used: (1) geometric methods (2) energy based methods and (3) dynamics based methods.

Geometric methods work by detecting cavities on the surface of the protein, which could then be classified as potential binding sites [25, 26]. A 3D grid is usually placed over the protein to describe the surface or the structure of the target. The dimensions of the pocket can then be defined by "rolling" a spherical probe over the surface of the grid [25]. When the probe does not "touch" any protein atoms, once the whole surface has been covered, smaller radius probes can then be used to map out shorter distances [25, 26]. The 3D shape of a pocket can be described by a combination of overlapping spheres. Software which are dependent upon this principle include POCKET [27], LIGSITE [28], SURFNET [29] and SPHGEN [30]. These methods suffer from the drawback of not taking into account protein dynamics, due to the protein being rigid during the cavity detection, along with the potential of missing unusually shaped binding sites, such as those found in channels or protein-protein interfaces [31].

More recently, in an attempt to overcome these drawbacks, a variety of pocket finding algorithms, using geometric methods, have been published. Some new

approaches include the use of voronoi tesselation, such as in the case of fpocket [32], or the alpha shape theory, like in the cases of CAVER [33] and MolAxis [34]. Other tools, combining elements from methods based on MD simulations, have been implemented in order to deal with the lack of protein dynamics. These include OMD [35], CAVER [33] and trj_cavity [36].

Energy based methods usually depend on the calculation of Van der Waals, electrostatic, hydrophobic, hydrogen bonding and solvent interactions [37, 38]. This calculation is achieved through the use of 3D grids and probes with different properties which are utilised to identify favourable binding regions [37, 38]. One of the first methods following this approach was the GRID algorithm [39, 37, 40, 41]. In this case, a 3D grid is placed over the structure of interest and attempts to identify the favourable binding positions of different probe types. The interaction energies of the probes are calculated as the sum of hydrogen bond, Lennard-Jones and Coulombic interactions. A more recent method, the Q-SITEFINDER algorithm [42], follows the same principle. However in this case, the interaction energies are calculated as the sum of the aliphatic carbon probes only. Similar approaches include SITEHOUND [43], where carbon and polar probes are used. Other methods employ metrics in a variety of approaches. SiMMAP [44] utilises three types of metric: pocket and ligand interaction types; conserved interacting residues of the pocket, and the composition of the ligand. POCKETPICKER [45] ranks potential binding sites by using a scale of buriedness.

Energy based methods can be as fast as geometric based methods with the added advantage of being quite sensitive and specific, given the chemical information included [5, 46]. However, one of their most serious limitations is the high number of false positives generated. This is due to the fact that, given the small size of the probes, a lot of local energy minima can be detected on the surface of the protein, making it difficult to filter out the false positives and identify the real binding site [5, 46]. One way of addressing these drawbacks is to include a solvation term in the energies calculated, and therefore the scoring, of potential binding sites. In a more complex solution, solvent inaccessible and solvent accessible surfaces can be identified, helping distinguish between potential binding sites. A method applying these solutions is CS-MAP [47, 48, 49].

The last type of binding site identification methods is based on molecular dynamic simulations. All the types of methods mentioned above use static structures to predict protein binding sites. These single structures have usually been elucidated using x-ray crystallography [50, 51], nuclear magnetic resonance (NMR) [52, 53] or cryo electron microscopy (Cryo-EM) [54, 55] methods. The structure of proteins under in vitro conditions is not always a good representation of the native state of the protein, leading to uncertainties regarding the orientation of side chains, the conformation of unstructured loops or water and/or ligand binding into the binding site [56]. These are all important features which can affect the accessibility and the size of a binding site [56]. Moreover, there are cases where the binding of a ligand or other moiety can induce conformational changes, allowing for the opening of an alternative pocket [57, 58]. Therefore, predicting cryptic, transient or allosteric binding sites on proteins can be impossible using static structures, due to the dynamic nature of these binding sites and the structure itself [59].

In order to tackle all the above problems, methods based on molecular dynamics were developed, to take into account protein flexibility and conformational plasticity. These methods have been shown to help us understand large scale conformational changes, induced by ligand binding or unbinding [60, 61, 62], along with predicting the most likely ligand binding modes [63, 64]. Furthermore, MD based alchemical transformation methods have been shown to be able to accurately predict binding energies [65, 66].

Classic molecular dynamic simulations have the drawback of potentially being trapped in local energy minima [67]. As a result, a variety of enhanced statistical, MD methods have been developed to try and overcome this. In the past, targeted-MD [68], SWARM-MD [69] and replica exchange MD [70]) were developed. More recent methods include temperature accelerated MD [71] along with the development of PLUMED, which opened the door for a variety of enhanced sampling methods, also known as metadynamics [72].

Consequently, a variety of tools have started incorporating MD approaches into their existing protocols, in the effort to detect protein binding sites. OMD [35] uses a MD protocol to identify solvent accessible areas and solvent binding sites, which are then used to improve the detection of the cavity from the static structure. Another

tool, dxTuber [73], uses the volmap plugin of VMD [74] to calculate a spatial density map of a cavity, from a MD simulation. Newer versions of tools, such as MolAxis [34], Caver [33] and MDpocket [75] now offer the option of using trajectories in their cavity searches by being able to process a series of PDB files. Finally, The POCKETANALYSER [76] version which uses principal component analysis (PCA) is using a molecular dynamics approach along with a clustering algorithm and a 3D grid to detect binding pockets on proteins.

1.5 Fragment Based Drug Design

A subcategory of SBDD which deserves special mention is fragment based drug design (FBDD).

Once a suitable binding site on a target has been identified, the search for compounds which possess the ability to interact with this binding site, with a high enough affinity, begins. A variety of methods, such as high throughput screening [77], QSAR [78], combinatorial chemistry [79] and high content screening [80] can be applied to identify compounds of high potency, which will then be optimized into leads. However, compounds coming out of these methods, especially molecules originating from large industry libraries through the use of HTS and combinatorial chemistry, have been shown to fail when it comes to lead optimization [5, 81, 82]. The reason for that has been attributed to the molecules being generated by these methods being potent but very large. From these failures, Lipinski's rule of five [83] was born to enable the identification of compounds that have a higher chance of surviving the lead development stage and reach the market. Lipinski's rule recommends the selection of compounds with a lower molecular weight, ideally ranging from 180 to 300 daltons and not exceeding 500 daltons [83].

The search for compounds of lower molecular weight led to the development of FBDD approaches. The characteristics and features of small fragments, such as their low molecular weight, higher polarity and solubility, make them more attractive from the perspective of drug discovery [84, 85]. They are also thought to allow for their translation onto compounds with more favourable physical and chemical properties, pharmacokinetics and toxicology [86]. Furthermore, fragments fit very well with the

"hot spot" driven model of ligand binding interactions [87, 88].

The "hot-spot" model predicts that the largest portion of the free energy required in protein-ligand binding comes from ligand interactions with a small number of key residues, also known as "hot spot residues" [88]. This implies that ligands binding to a target with higher affinity do so by optimally exploiting these interactions through properly orienting key functional groups, which can be matched by the smaller compound fragments which exhibit high ligand efficiency [89].

For all these reasons, a host of experimental and computational methods have been developed that try to exploit a variety of fragment binding data. The most commonly used experimental methods include 1D or 2D NMR [52, 53], surface plasma resonance (SPR) [90] and x-ray crystallography [50, 51]. Automated NMR pipelines allow for the testing of up to 1000 compounds in one day with the 1D approach allowing for an examination of the ligand and the 2D approach allowing for the protein conformation and ligand binding constants to be determined [91].

When it comes to X-ray crystallography, multiple solvent crystal structure (MSCS) techniques, and especially the keystone work of Ringe [92, 93], provided the experimental basis for the mixed solvent computational techniques. In MSCS, multiple protein structures are solved in combination with different co-solvents. Co-solvents found in the same location indicate a site of biological significance. Initial validation studies for this method were performed using elastase as the target and acetonitrile as the co-solvent probe [92, 93]. It was shown that acetonitrile has the ability to identify and map the active site along with crystal packing interfaces. When the range of probes was extended, it was found that many different probes bind to the active site and crystal packing surfaces as well, in contrast to other regions, suggesting that these regions correspond to important binding sites [92, 93].

Finally, SPR techniques have the capability of testing thousands of compounds in a day, but with a higher number of artefacts due to the higher concentration of fragments required, to detect weak binding, along with the need of target immobilization on an SPR chip [94].

Despite the high automisation and throughput of the experimental methods mentioned above, a significant amount of resources and time is needed to obtain useful results. Therefore, computational methods were developed in an effort to

reduce the initial load of the experimental methods [5]. A great variety of tools has been developed for this purpose, the first one being GRID [37, 40, 41], also mentioned above, in 1985, followed by MCSS [95], LUDI [96], Sprout [97], Superstar [98, 99] and MUSIC [100]. More recent methods include CS-Map [47, 48, 49] and FTMap [101]. Most of these methods employ force fields and molecular mechanics models of protein-ligand interactions to try and predict the binding poses of probes or fragments. In multiple copy simultaneous search (MCSS) simulations, developed by Martin Karplus, a variety of organic probes are dispersed on the sites of interest. They are energy minimised on the surface of this site, over 3000-6000 steps, and they are then sampled on a grid and presented as density maps [38, 102].

Superstar is an exception in that it is a knowledge based computational approach where a database of frequently observed intermolecular interactions in X-ray crystallography is used to predict the functional groups with the highest propensity for binding to a protein pocket [98, 99]. Based on Superstar, a method by Radoux *et al.* was recently developed where a range of small probe fragments are fitted into the propensities calculated, in an attempt to find the best binding pose for these fragments [88].

Further to the previously mentioned methods, virtual screening and molecular docking have also been extensively used in the initial stages of a lot of drug discovery projects. All these computational techniques, however, rely on static structures or snapshots in order to identify potential fragment-protein interactions, which means they suffer from the same intrinsic limitations as the pocket identification methods using static protein snapshots, which were mentioned previously [5, 103]. Again, a few mixed solvent, MD based methods have been developed in order to deal with the exploration of the protein dynamics and plasticity which can lead to the opening of otherwise inaccessible cryptic or allosteric binding sites [5, 59]. Most of these methods are based on the use of probes during the simulation and the sampling of these probes on a grid afterwards, as shown in Fig 1.2 [59].

1.5.1 Mixed Solvent Molecular Dynamics Based Methods

The most recent mixed solvent molecular dynamic based methods, presented below, have also been recently reviewed in an excellent paper by Ghanakota et al.[104]. The

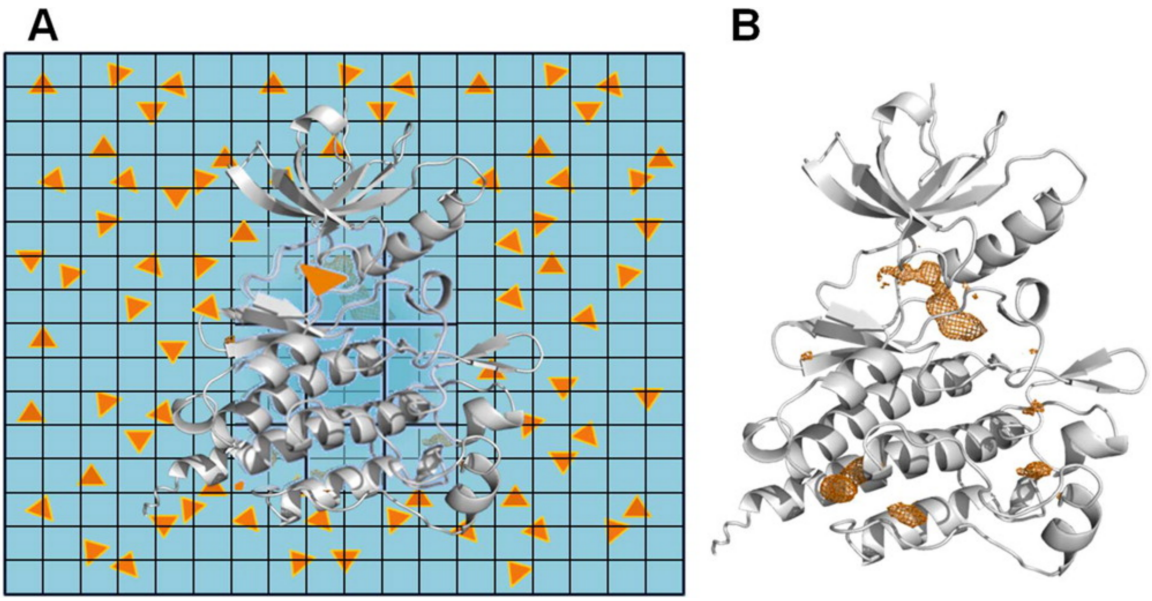


Figure 1.2: A figure showing a schematic of a mixed solvent molecular dynamics simulation and the use of a grid. (A) The protein (CDK2) is shown in a grey cartoon representation while the probes are represented as orange triangles. A grid is overlaid on top of the protein and the simulation box. Each box represents a voxel (B) The same protein shown in a cartoon representation (gray) along with the sampled ligand occupancy, shown as an orange mesh. Figure reproduced from [104].

existence of certain common pitfalls amongst most of these methods was observed, along with the identification of areas in need of further improvements.

1.5.2 MDmix

The first co-solvent, simulation based method reported was that of Barril and colleagues, named MDmix [105]. In this approach, a 20% concentration of isopropanol in water was used to run the simulations for at least 16ns. The results were evaluated based on the ability of the simulations to recapitulate the isopropanol molecule positions from the MSCS structures of thermolysin, p53 (core domain) and elastase. Separate occupancy grids were created for the hydroxyl oxygen and methyl carbons of isopropanol. Although it was reported that the method was successful in identifying the location of the isopropanol molecules from the MSCS experiments, there were additional sites mapped on the protein surface which were not discussed in the paper. These sites may further complicate the identification of potential ligand binding sites on the protein surface (figure 1.3). Moreover, no explanation was offered for not comparing the grid map of the whole isopropanol molecule to that of the MSCS

experiments [104].

A further study by Barril and colleagues looked into the effect of protein flexibility on the binding free energy of the probe molecules (ΔG_{bind}) [106]. A logarithmic relationship was determined between flexibility and its effects on ΔG_{bind}. It was concluded that if the protein has been restrained and the binding site has been preformed, ΔG_{bind} would be more favourable due to the fact that the entropic cost of restraining the protein has already been covered. If this is not the case, then the ΔG_{bind} would be less favourable due to the clashes with the binding site.

In more recent studies, an updated protocol has been suggested for MDmix. This consists of two separate co-solvent simulations, one with 20% ethanol in water and one with 20% acetamide in water [107]. Three simulation runs, of 20 ns each, have also been proposed, with the heavy atoms in the protein being weakly restrained. For the validation of the new protocol, heat shock protein 90 N-terminal (Hsp90) and HIV protease (HIVp) were used. Holo structures were used to create pharmacophore models and these models were then compared with the results of the density maps from the simulations. Some key pharmacophoric features of HIVp were not identified, prompting the authors to suggest that an extension of the technique with more probe types might be required. Furthermore, it was commented that using the densities from single atoms, which are part of the probes, to define pharmacophoric features is limited by the assumption that these atoms behave independently.

1.5.3 SILCS

This method, developed by MacKerell and colleagues, was named site-identification by ligand competitive saturation (SILCS) [108]. It has expanded to include techniques for the creation of pharmacophore models, free energy perturbation and the detection and sampling of cryptic binding pockets. The initial protocol called for simulations with 1 M propane and benzene solutions and the BTB domain of BCL-6 as the target. The benzene probes were used to identify potential aromatic interactions while the propane probes were used to identify the potential aliphatic interactions. The water molecules were also examined with respect to their properties as probes for reporting on potential hydrogen bond donor or acceptor interactions. 10 runs of 5 ns each were used with a weak restraint placed on the C_α atoms during the simulations and

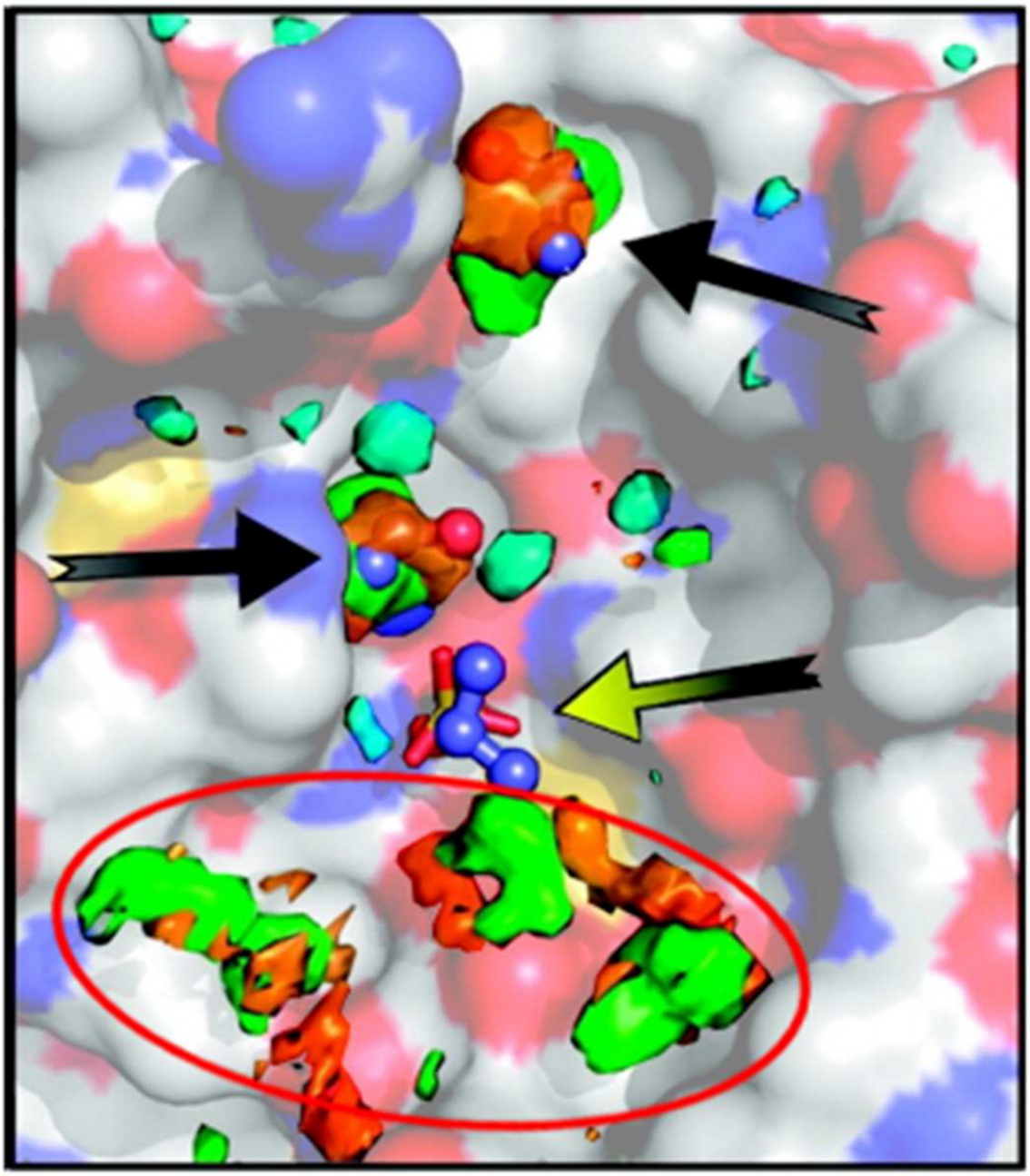


Figure 1.3: A figure showing the MDmix interaction densities (orange and green isosurfaces) compared to the experimentally determined isopropanol sites (balls and sticks). The orange isosurfaces correspond to the density of the OH group of isopropanol and the green isosurfaces correspond to the Me group of isopropanol. The black arrows pinpoint to areas where the densities match the experimentally identified binding sites of isopropanol and the yellow arrow pinpoints sites that were not identified by this method but are determined experimentally. The red eclipse points to the "spurious" probe densities that could correspond to false positives. Figure reproduced from [104].

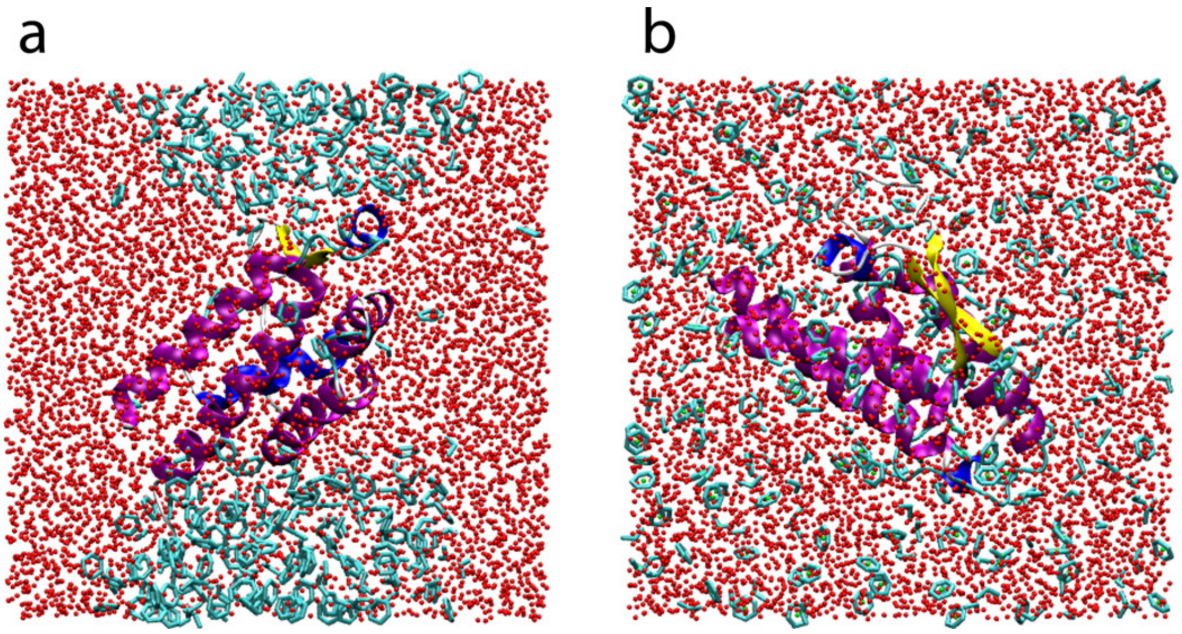


Figure 1.4: A figure showing the differences between the application and the lack of the repulsive force between the probes. (A) Is showing the simulation box, after a few ns of simulation, without the use of the repulsive force between the probes. (B) Is showing the same with the use of the repulsive force. In both cases, the probes are shown in cyan and water molecules in red. The protein is shown as a cartoon representation. Figure taken from [109].

a repulsive force introduced between the probe particles. The repulsive force was used in order to avoid aggregation of the hydrophobic probes (figure 1.4). Snapshots from each repeat were combined and used to create the density maps for each probe which were name "FragMaps". The method was validated based on the ability of the SILCS probes to identify biologically relevant binding sites on the target.

The study was later extended to include a greater variety of test systems [110]. These were trypsin, α-thrombin, HIVp, FK506-binding protein 12 (FKBP), factor Xa (FXa), NadD and RNAse A. The length of the simulations was also extended to 20 ns per repeat and the "FragMaps" generated were converted to grid free energies (GFE). This was done using the inverse Boltzmann equation. Each atom in the ligand was then assigned to one of the following types: aromatic, aliphatic, hydrogen bond donor or hydrogen bond acceptor. These correspond to the types of probes used in the simulation. The ligands from the holo structures were then aligned to the same frame of reference as the simulations and the atom type of the ligand and its position in the grid were used to obtain the value for its free energy from the GFE grid. The ligand grid free energy (LGFE) was then calculated by summing these free

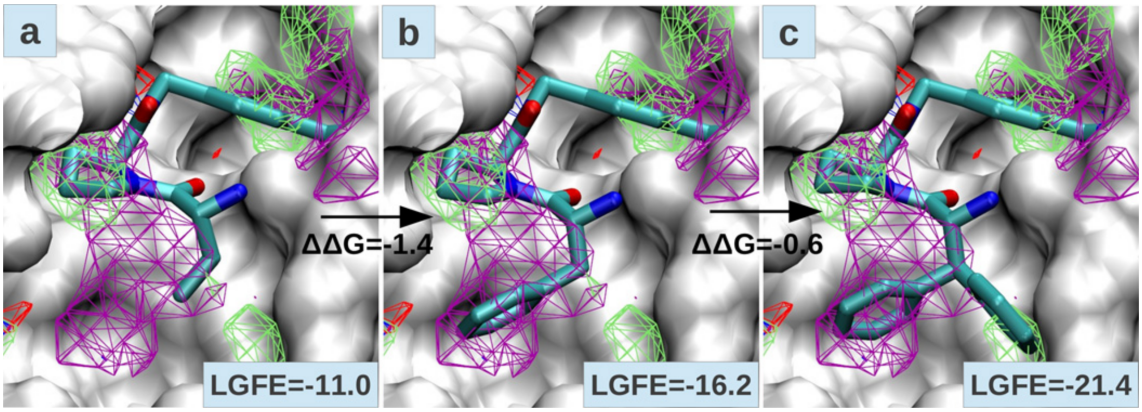


Figure 1.5: A figure showing thrombin - α along with three ligands of progressively increasing affinity, from a to c. The LGFE score is shown at the bottom left corner. The FragMaps for benzene are shown in purple and for propane in green. Figure taken from [109].

energy values. Finally, LGFE was used to score different binding poses of the ligands and it was shown that ligands known to bind score higher than decoy compounds.

Follow-up studies included the use of free energy perturbations (FEP) to assess the potential of a wider range of different fragments to bind to the site of interest [111]. α-thrombin and p38 MAPK were used as initial targets and benzene was the first probe whose relative energy of hydration, for moving to monosubstitued benzene, was calculated. A comparison between the single step FEP of benzene to its analogues, with varying experimental binding free energies, was made. Promising results were obtained for α-thrombin but that was not the case for p38 MAPK.

Another subsequent study also looked at the effects of protein restraints and denaturation in mixed co-solvent simulations [112]. Using interleukin-2 (IL-2) as the target, along with positional restraints of different strength, it was observed that there were cases when no restraints were used where IL-2 would denature and unfold in the simulation. Two possible solutions were suggested, removing trajectories in which the protein has denatured or completely restraining the protein backbone. The second option could prevent the identification of cryptic binding pockets so it was considered less favourable.

In the updated version, SILCS Tier-II [113], the list of co-solvents used was extended. The new probes introduced included formamide, acetaldehyde, methanol, acetate and methylamonium, with a concentration of 0.2 M for each one. The simulations were still run for 20 ns with a weak restraint on the C_α atoms. In this

protocol, the densities of the probes were combined based on the properties of each one. Benzene and propane carbons were combined into generic non-polar, methanol and formamide polar hydrogens were combined into neutral donors, methanol, formamide and acetaldehyde hydrogens were combined into neutral acceptors, methylammonium polar hydrogens were positive donors and acetate oxygens were defined as negative receptors. GFE were again created based on these densities and they were overlaid with ligands from a variety of targets. The targets used in this iteration were FXa, p38 MAPK, RNase A and HIVp. Along with visually inspecting the densities, a scoring method based on the LGFE scores from before was also developed. This method, in conjunction with a Monte Carlo based sampling of the ligands on the GFE grid was shown to be the best method to correlate the ligands with their experimental binding affinity. This method was shown to work for all the cases except for HIVp.

An attempt was also made to turn the simulations and densities calculated into pharmacophore models [114, 115]. In the initial study, a GFE cut-off of -1.2 kcal/mol was used for aromatic/aliphatic "FragMaps" and -0.5 kcal/mol was used for water based hydrogen bond donor and acceptor "FragMaps". The points below these cutoffs were then clustered, based on different distances for each "FragMap", and the sum of each GFE is reported as the feature grid free energy (FGFE). Pharmacophoric features were then generated for combinations of overlapping "FragMap" densities, which were ranked based on the cumulative FGFE of the "FragMaps" that were used to create them. HIVp, FXa abd DHFR were used to assess this approach by screening decoys from the DUD-E data set [116] and comparing them with results from MOE [117], Dock [30], Autodock [118] and SB pharmacophore models based on hydration data [119]. It was reported that this approach provides equally or better results from the aforementioned methods.

The updated protocol, SILCS Tier-II, was used, along with a variety of new targets, to repeat the pharamcophore analysis. The results were also compared with the results from an increased range of pharmacophore building software. It is also reported that the methods performs equally or better for most of the targets, with the exception of AmpC β-lactamse [114, 115]. Finally, a Grand Canonical Monte Carlo (GCMC) approach was developed, where a reservoir of water and co-solvents is

connected to the system and co-solvent or water molecules are either added or deleted based on metropolis criteria [120]. After a certain number of steps, approximately 100000, an MD simulation is initiated. T4 Lysozyme L99A was used to validate the approach which was then extended to a range of targets with occluded binding pockets along with two GPCR targets. It was reported that the occluded binding sites were successfully identified, however the results of the GPCR targets were not as good as expected [121, 122].

1.5.4 MixMD

Mixed solvent MD (MixMD) was developed by Carlson and colleagues. It follows the same approach to the previously mentioned co-solvent simulation methods with the difference that emphasis is placed on the use of water miscible organic probes to avoid probe aggregation [123]. The probe used was acetonitrile and the target used for validation was HEWL. A 50% concentration of acetonitrile was used, with 5 repeats of 10 ns each. The last 2 ns of each repeat were used to obtain the binding location of the probe on the HEWL surface. It was reported that acetonitrile can recapitulate binding areas on HEWL, compared to MSCS experiments. It was also reported that when running the simulations with a varying degree of protein restraints, only the repeats allowing full protein flexibility would faithfully reproduce acetonitrile binding sites. Using restraints would result in a lot of incorrect local minima.

In a follow-up study, the range of co-solvents was expanded, with isopropanol being added [124]. New targets were also examined with elastase, p53, RNase and thermolysin being simulated. The number and duration of the simulations were also adjusted, putting an emphasis on the use of a higher number of short simulations. In this case, 10 runs of 20 ns each were used. It was reported that the isopropanol sites, as seen in MSCS studies of these targets, were accurately reproduced.

In more recent studies, the quality of the parameters was examined with the parameters used by Barril et al. shown to cause isopropanol to be separated in two phases, when simulated at a 50% concentration with thermolysin [125]. In contrast, the parameters used by Carlson et al., based on OPLS, were shown to prevent this separation and keep the probes well mixed. It was reported that the authors used radial distribution functions of the solvent to monitor phase separation, which they

found to be equally important as the use of RMSD for protein denaturation. The protocol was repeated for eleven different co-solvents and six of them were identified as having good enough parameters to provide even mixing with the TIP3P water model. These were acetonitrile, isopropanol, acetone, N-methylacetamide, imidazole and pyrimidine.

The concentration of the co-solvents used was also examined in a subsequent study [126]. Using HIVp and a 5% concentration of co-solvents, in an effort to approximate the concentrations used in experimental data, was shown to improve the accuracy and the signal to noise ratio of the binding of the probes.

Finally, in their most recent studies, Carlson and colleagues have brought all the previous studies together with the development of a protocol for the identification and ranking of binding sites [127]. Using 5% acetonitrile, isopropanol and pyrimidine with a large number of shorter simulations, it was reported that they were able to identify the top four main and allosteric binding sites of ABL kinase, AR, CHK1 kinase, glucokinase, PDK1 kinase PTP1B and farnesyl pyrophosphate synthase. Their results were compared with FTMap and it was shown that for ABL, AR, PKD1 kinase and PTP1B MixMD outperforms FTMap. However, in both methods, spurious hotspots not corresponding to a known binding site, were also identified.

1.5.5 Yang and Wang Method

In this approach, developed by Yang and Wang, co-solvent simulations using the Barril method were compared with MSCS experiments of thermolysin using isopropanol, phenol and acetone. The inverse Boltzmann equation was used to compare the free energies obtained from the simulations with the decoupling method [87]. The authors focused on the two primary isopropanol binding sites of thermolysin. The second site was also identified by the other co-solvents. The free energy values calculated by both methods, for each thermolysin site, were shown to be within 1 kcal/mol of each other, with the double decoupling method [128, 129] being slightly more accurate.

In their follow-up study, protein-protein interfaces were examined. Different conformations of BCL-xL and Mcl-1 were simulated in a mixture of water and isopropanol [130]. One apo and three holo BCL-xL were used for 32 ns simulations in water which revealed that the protein underwent hydrophobic collapse, preventing its

interaction with its binding partner proteins. Addition of 20% isopropanol allowed the target to retain conformations that would allow binding with other proteins. It was also reported that the hotspots identified varied, depending on the starting conformation. Using co-solvent simulations of 20% isopropanol, 10% phenol and 2 M trimethylamine N-oxide, on BCL-xL and Mcl-1 in continuation of their previous work, the authors reported both similarities and differences on the locations of the hotspots identified, based on the starting conformation [131]. It was then recommended that the differences identified could be exploited to increase the selectivity of potential compounds.

In their more recent studies, the ectodomain of the interleukin 1 receptor type 1 (IL-1R1) was examined using co-solvent simulations of 10% phenol. The results of these simulations were compared with the three potentially druggable sites identified by SiteMap ([132]. Their efforts were focused on the second binding site identified, designated P2, which had not been previously identified. Co-solvent simulations were then used to assess the effects of protein flexibility on the identification and druggability assessment of this site and it was observed that only two of the initial conformations used allowed for this site to be targeted by the phenol probes.

Finally, using Bcl-xL again, the effect of different simulation types and co-solvent set-ups were investigated. Pure water MD simulations, co-solvent simulations, accelerated MD simulations and accelerated MD co-solvent simulations were tested. It was found that the last set up was the best one for generating protein conformations which would be the most useful in docking [133].

1.5.6 Bahar and GlaxoSmithKline

The method developed by Bahar and colleagues at GlaxoSmithKline (GSK) uses co-solvent simulations to address the druggability potential of targets' binding sites [134]. Two separate co-solvent simulations are run, one with isopropanol and one with a mixture of acetamide, acetic acid and isopropylamine. The concentration is about 2.3 M for all of the co-solvents and various repeats are run, ranging from 32 to 40 ns. The free energies of the *whole* probes, a major difference from the other methods, are then calculated using the inverse Boltzmann equation. These free energies were then defined as "interaction spots". Constraints were placed on

the definition of these interaction spots, the most important of them being that they were not allowed to overlap with other "interaction spots" and that only the spots with a free energy of -1 kcal/mol or lower were considered. The energy central grid point of the "interaction spot" was used as the free energy of the whole spot and the charge of the spot was determined based on the fraction of the total time that a spot was occupied by a charged probe. The spots were then clustered to identify druggable sites [134].

Using this setup, MDM2, PTPB1, LFA-1, kinesin Eg5 and p38 MAPK were tested. It was reported that the maximum free energies calculated for these targets agree with the experimentally calculated values of the best known binding ligands for these targets [134]. The detection of a cryptic binding site on MDM2 and allosteric binding sites on LFA-1 and Eg5 were reported. These were attributed to the equilibration of the targets at 600 K, with heavy atoms restraints to prevent denaturation. It was also observed that the use of co-solvents prevented the hydrophobic collapse of the targets, in agreement with Yang and Wang. The importance of using a mixture of co-solvents in the simulations was highlighted, due to the opening of the allosteric pockets of Eg5 and P38 MAPK only in the presence of a mixture of probes. The sugestion that charged co-solvents would also be required was made due to the charged or zwitteronic nature of a lot of drug compounds.

1.5.7 Caflisch Approach

In this approach, developed by Caflisch and colleagues, the binding affinities and on/off rates of DMSO on the binding site of FKBP were estimated using kinetics. A concentration of DMSO of 440 mM was used, along with ten simulations, each one 70 ns long. The DMSO molecules were able to map the active site of FKBP and a dissociation constant for DMSO of around 300 mM was computed, which agreed with the experimental results. Increasing or decreasing the concentration of DMSO by a factor of two did not have any effect on the results [135].

In their follow-up studies, two bromodomains were used as targets, along with a slightly expanded variety of co-solvents [136]. The zinc finger domain 2B (BAZ2B) and CREB binding protein were simulated in DMSO, methanol and ethanol, of a 440 mM concentration. Two 0.5 μs simulations for each co-solvent were run. It was

reported that the simulations were able to map the acetyl-lysine site of the CREBB binding protein and that the position of the DMSO agreed perfectly with that of the DMSO in the crystal structure of the protein [137]. The same was reported for the binding site acetyl-lysine site of BAZ2B, by different co-solvents.

Analysing the kinetics of the co-solvents, it was observed and reported that DMSO and ethanol were slower to bind and unbind than methanol, potentially due to their larger size and hydrophobicity. By analysing the water molecules in the acetyl-lysine domains during the simulations, the authors were able to identify water molecules that remained stably bound in the active site during the whole of the simulation duration, as well as transient water molecules which were replaced by co-solvents. Based on these observations, the suggestion of including the stable/non-replaceable in docking studies was made. For the transient waters, it was suggested that compounds with hydroxyl substituents could be used to target these areas [136].

1.5.8 Tan and Abell Approach

In this approach, by Tan and Abell, a lower concentration (0.2 M) of benzene was used. This was done to avoid the aggregation problems reported in previous studies. The polo-box domain of the polo-like kinase 1 was used in this study to demonstrate the need of co-solvents to prevent hydrophobic collapse and detect a cryptic binding site [138]. A drug compound was subsequently designed to target that cryptic binding site.

In more recent studies, the authors tested the use of chloro benzene as a potential co-solvent, due to the presence of halogens in drug molecules [139]. A lower chloro benzene concentration of 0.15 M was needed due to aggregation issues, increasing the simulation length to 10 ns, to ensure enough conformational space had been explored. MDM2, Mcl-1, IL-2 and BCl-xL were used to test this approach. These targets were selected due to the presence of halogenated ligands known to bind to them. The apo structures were selected as starting points of the simulation. It was reported that the cryptic binding sites of these targets were able to be identified by chloro benzene, except in the case of Bcl-xL.

In more recent studies, benzene was used to identify potential protein protein interaction sites. Aurora Kinase A, RAD51 and MDM2 were used as test targets.

It was reported that the benzene co-solvents were successful in identifying these hydrophobic sites and that RAD51 and MDM2 cryptic binding pockets were not open in water simulations, in contrast with co-solvent simulations [140].

1.5.9 Fersht Approach

In this study, by Fersht and colleagues, the cancer causing mutant p53-Y220C was used in co-solvent simulations [141]. A set-up similar to the Barrill approach was used but with a 20% isopropanol. A 600 K equilibration, with heavy atom restraints, was followed by a 1 ns equilibration at 300 K along with a 19 ns simulation. It was reported that a three potential druggable sites were identified, the one with the highest binding of propanol being on the dimer interface. The pocket binding site, caused by the formation of a pocket due to the mutation, was included in the sites identified. However, during the experimental fragment screening, only the mutation induced binding site was identified.

1.5.10 pMD

This final approach was developed by Gorfe and colleagues. K-ras was used as a target in probe MD (pMD), where isopropanol was used as the co-solvent probe [142]. A simulated annealing procedure, similar to that of Bahar and GSK, was used, starting with a restrained equilibration at 650 K, followed by cooling to 310 K for three production simulations, varying from 30 to 100 ns in length. The three runs were then combined and the inverse Boltzmann equation was used to calculate the free energy value for each grid point. As in Bahar et al., maximum free energies were used to identify which grid points would be kept, the ones with the most favourable grid energies being clustered. Clusters with four or more points were defined as "Druggable Sites" and clusters with three or two interaction points were defined as "subsites". Five druggable sites and three subsites were identified on K-ras, which were shown to match known allosteric sites [142].

In a recent study, the authors have repeated the same protocol, this time including a mutant K-ras structure, embedded in a lipid bilayer. Lipid membranes present an exceptional challenge for the co-solvent simulation approaches due to the fact that co-solvents can disrupt the integrity of the membrane and alter non-bonded

interactions between the membrane and the probes. The identification of certain allosteric sites of K-ras was reported on this case, but it was concluded that a lot of work still needs to be done [143].

1.5.11 Overall Conclusions from the co-solvent MD Approaches

Table 1.1 offers a summary of all the co-solvent MD approaches examined here. These methods offer a promising, less time and resources-consuming approaches to identifying potential protein-ligand or protein-protein interaction sites, which can produce results comparable to experimental methods, such as MSCS [104]. However, it is evident that the current approaches suffer from a range of limitations that leave a lot to be desired. Most of the current methods result in the identification of spurious hotspots, sometimes within the binding site of interest, which most likely correspond to false positives. These hotspots could skew any drug design efforts, leading to the selection of compounds which do not have the required physicochemical properties and which would not be able to efficiently bind to the target and exert the required effect.

Moreover, although a wide range of targets has been examined so far, by different methods, there is no quantitative validation or examination of these targets, provided by any of these methods, except for in the form of binding free energies compared to experimental results, for different co-solvents. Given that not all methods use the same co-solvents, and that obtaining accurate experimental free energies of binding can be quite tricky, a quantitative comparison between the methods available is very challenging. Furthermore, the qualitative assessment and simulation protocol followed varies between each method. This makes the comparison of the results even more difficult. Finally, with the exception of SILCS, no method has systematically examined a large number of targets with the same simulation protocol or analysis.

1.5.12 Arcon Approach

One newer study, and the only one we have identified so far, which deals with the issues of systematically examining a wide range of targets and quantitatively measuring the predictive powers of the co-solvent probes, was performed by Arcon and colleagues [144].

In this approach, which is based on the Barrill method, the following co-solvent probes were used in a 20% concentration: ethyl alcohol, acetamide, methylammonium, acetate and acetonitrile. Each co-solvent was simulated separately. 18 targets were simulated, in order to cover a wide range of different binding sites and interactions, with three repeats of 20 ns each. Holo structures of these targets were then selected and analysed and the interactions driving the binding of each different ligand were identified [144].

The interaction densities and binding free energies were calculated, in the same manner as in the Barril approach, and then the interactions identified for each binding site were compared with those identified from the holo structures were compared. For each probe atom, the type of interactions that it can fulfil were identified and, based on that, four metrics were defined: true positive, true negative, false positive and false negative. Based on these metrics, the sensitivity, precision, specificity and accuracy of each probe atom were determined. It was reported that the overall values for all co-solvents, including water, range in the 0.55 region, with the exception of precision which is around 0.40. Furthermore, the problem of spurious sites identified still persists. Overall, although this is a very good effort in tackling some of the issues mentioned previously, there is still room for improvement [144].

1.6 Pharmacophore Models

A pharmacophore model is defined as an abstract description of all the chemical features required by a ligand for recognition and binding to a receptor [145]. They can be used to explain how structurally different ligands bind to the same receptor. Pharmacophore models can have a plethora of features. The most important of them being hydrogen bond donor or acceptor moieties, aromatic or hydrophobic moieties, positively or negatively charged moieties and hydrogen atoms [145]. In Figure 1.6 an example of a phramacophore model, with all the features mentioned above, can be seen.

Table 1.1: A table showing the co-solvent MD methods currently published along with the co-solvents used and the targets examined and validated.

Developer (Method)	**co-solvent**	**Target**
Barril (MDmix)	Isopropanol, ethanol, acetonitrile, methanol, acetamide	Thermolysin, p53, elastase, MDM2, LFA-1, PTPB1, p38 MAPK, AR, HEWL, Hsp90, HIVp
MacKerell (SILCS)	Benzene, propane, water, acetonitrile, methanol, formamide, acetaldehyde, methylammonium, acetate, imidazole	BCL-6, tryprin, α-thrombin, HIVp, FKBP, FXa, NadD, RNase A, IL-2, p38 MAPK, DHFGR, FGFr1 kinase, adenosine deaminase, ERα, AmpC β-lactamase, T4-L99A, AR, PPARγ, mGluR5, β2AR
Carlson (MixMD)	Acetonitrile, isopropanol, pyrimidine, imidazole, N-methylacetamide, acetate, methylammonium	HEWL, elastase, p53, RNase A, thermolysin, HIVp, ABL, AR, CHK1, glucokinase, CHK1, PDK1, PTPB1, farnesyl pyrophosphate synthetase
Yand and Wang	Isopropanol, phenol, trimethylamine, N-oxide	Thermolysin, Bcl-xL, Mcl-1, IL-1R1
Bahar and GSK	Isopropanol, isopropylamine, acetic acid, acetamide	MDM2, PTP1B, LFA-1, Eg5 kinesin, p38 MAPK
Caflisch et al.	Dimethyl sulfoxide, methanol, ethanol	FKBP, BAZ2B, CREBBP
Tan and Abell	Benzene, chlorobenzene	Polo-like kinase 1, MDM2, MDMX, IL-2, Mcl-1, Bcl-xL, Aurora Kinase A, RAD 51, ERα, ERβ
Fersht et al.	Isopropanol	p53-Y220C
Gorfe et al. (pMD)	Isopropanol	K-ras
Arcon et al.	Ethanol, acetamide, water, acetonitrile, methylammonium, acetate	Acetylcholinesterase, aminotransferase, β-lactamase, cGMP, FXa, CDK2, CP450, dihydrofolate reductase, DNA gyrase B, estrogen receptor α, FGFR1, glucocorticoid receptor, glycogen phosporylase b, lysozyme, integrin α 1, p38 MAPK, PPRG, trypsin

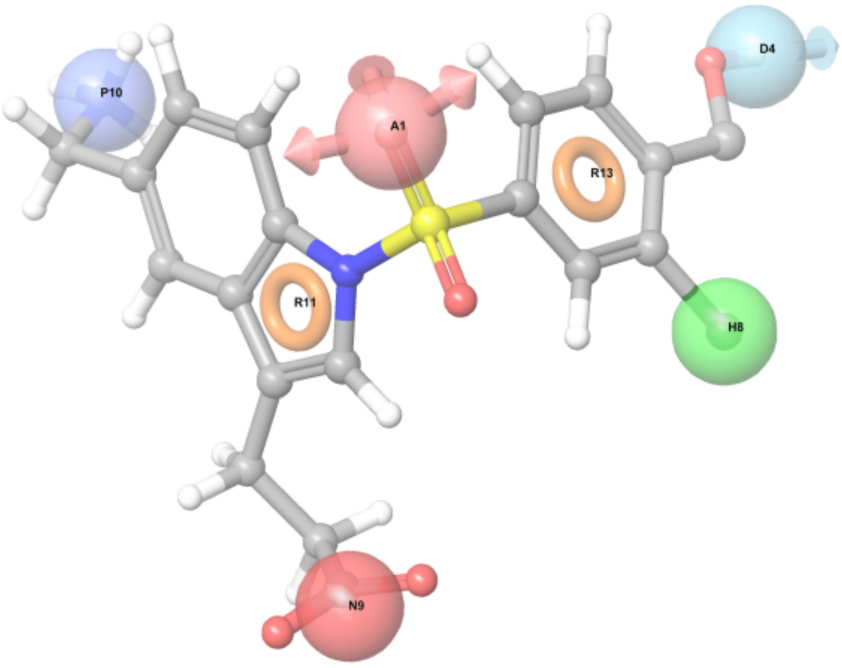


Figure 1.6: A ligand with an example pharmacophore model overlaid. Pharmacophore models can include a plethora of different features, but there are six that are of particular importance. These are positively or negatively charged moieties (P10 and N9 respectively), hydrogen bond donor or acceptor moieties (A1 and D4), hydrogen atoms (H8) and aromatic or hydrophobic moieties (R11 and R13). The arrows on the features indicate the directionality of potential bonds.

1.6.1 Pharmacophore Models in Drug Design

Pharmacophore approaches constitute one of the major tools available for modern drug design [146, 11]. Their most important application is their use in high throughput virtual screening where they can be used to screen huge libraries of compounds and generate initial hit or lead compounds [147, 145]. Moreover, assuming there is enough confidence in the accuracy of the model, they can be used for *de novo* drug design. In this case, a novel compound is synthesized to fit the model with the hope that it will then fit and bind the active site on which the model is based. Finally, pharmacophore models can also be used in lead optimization. As in the previous case, initial hit or lead compounds can be extended to fit the pharmacophore model of a target, with the hope that the new compounds will exhibit the desired properties [147, 145].

1.6.2 Pharmacophore Model Creation

There are two ways in which a pharmacophore model can be established. In ligand based (LB) pharmacophore modelling, all the physiological conformations of the ligands need to be accounted for and be properly aligned . In order to handle ligand flexibility, three possible solutions have been proposed [147, 148, 149]:

- The generation of all possible ligand conformations and their storage in a database. This solution allows for reduced computational time when it comes to searching the database to retrieve these conformations, but increases the amount of storage required.

- The design of specialised search queries that account for the flexible parts of the compounds. This approach places the burden of designing "smart" queries on the user.

- The generation of ligand conformations "on the fly" during the pharmacophore modelling process. This approach does not require as much storage space as the first approach but increases the computational time required.

In structure based pharmacophore modelling, the structure of the target is necessary, either in the apo form or in complex with a bound ligand.

In the target-ligand complex case, the ligand binding site and the key interactions between the ligand and the target are identified and then turned into a pharmacophore model. However, in the case of only one complex being available, no quantitative structure activity relationships (QSAR) are reflected [147, 150]. A multi-complex based approach has been suggested to overcome this, when more than one target-ligand complexes is available. When no ligand bound complex is available but ligands known to bind to the active site of the target available, molecular docking can be used to produce a ligand-target complex [147, 151]. However, when no compounds are known to bind to the target, this approach cannot be used.

In the case of using the apo structure of the target, LUDI interaction maps of the ligand binding site can be converted into pharmacophoric features. Using the LUDI approach, small probes are fitted into clefts of the target protein and the probe-protein interactions are then translated into maps [96]. However, this approach,

as well as the complex approach mentioned above, leads to the identification of too many un-prioritised pharmacophoric features. A knowledge based, hotspot guided method has been proposed in order to prioritise cavity atoms that are important for interactions. This is based on the use of machine learning algorithms trained by fingerprints of known ligand binding pockets. In the case of a multi complex approach, the most frequently identified pharmacophoric features can be prioritised [152].

1.7 Aims and Motivations

In this thesis, the development and application of a new, mixed solvent, coarse grained molecular dynamics based method will be described.

The aim of this method is to help in identifying protein interactions which drive ligand binding, turning them into interaction densities and pharmacophoric points, with the hope of facilitating and guiding the initial stages of drug design. The method is based on the use of CG probe beads, in place of co-solvents, which correspond to the important features of a pharmacophore model, described previously. Using moieties with these properties opens the door to the possibility of the pharmacophoric points being developed into a pharmacophore model.

In chapter 3, the development process of the method will be described and the choice of the selected parameters and simulation set-up will be explained. The test and validation of the method on a series of water soluble targets, of pharmacological interest, known as the Ichihara data set will also be presented. This will be done both through a qualitative and quantitative approach. In chapter 4, a set of more challenging targets, consisting of membrane proteins, will be examined. This set includes G-Protein Coupled Receptors (GPCR) for which the main binding site will be examined. Furthermore, the ability of this approach to detect allosteric binding sites will also be demonstrated. Finally, in chapter 5, an alternative application of this method will be presented. Using kir2.2, GPCRs and transient receptor potential cation (TRPV) channels, the ability of the coarse grained probes to identify lipid binding sites will be demonstrated.

The function of a wide range of membrane proteins, including GPCRs and TRPVs,

is known to be modulated by their lipid environment [153]. A lot of different aspects of protein function can be affected by this modulation, such as protein activity, the formation of protein protein interactions and the cellular localisation of a protein by restricting its movement to a specific region of the cell membrane [154, 153]. Currently available computational methods for predicting lipid-protein interactions require up to a few hundred microseconds of simulation time. Therefore, a method which only requires a few microseconds (usually less than 10) of simulation time and can prospectively predict lipid interaction sites on membrane proteins is of great value [153].

Chapter 2

Theory and Methods

This chapter provides a description of the theory behind key computational techniques used in this thesis. All atom and coarse grained (CG) molecular dynamics (MD) simulations will be discussed. A high level overview of the topics will be given guided by more detailed discussions in a number of established texts [155, 156].

2.1 Molecular Dynamic Simulations

Molecular dynamic simulations were first developed in the 1950's [157, 158] and since then they have not only been used in the characterisation of biological systems but have also found a wide range of applications in material science, engineering and chemistry. The first simulation of a protein molecule was reported in the 1970's by McCammon and Karplus [159]. Since then, theoretical developments, as well as the increase of the hardware capabilities, have allowed for an exponential increase in the complexity of the simulated systems and time scales available [160].

The current capabilities of molecular simulations allow for the modelling of biological systems at a range of resolutions, scales and time durations. The Schrödinger equation allows for an exact and fundamental description of molecular processes through the use of quantum mechanics simulations. These simulations can be used in understanding detailed chemical processes, like the binding of ligands or behaviour of substrates in an enzyme active site [161]. The drawback of quantum mechanics calculations, however, is their very high demand for computational resources, making the simulation of larger systems (5000 electrons or more) intractable, even through

the use of hybrid quantum mechanic simulations [162]. For that reason, in order to explore large scale biological phenomena, which take place over longer time scales, bio-molecular simulations which use classical Newtonian physics are conducted. In these cases, atoms are treated as individual particles without explicit electrons (the Born-Oppenheimer approximation). The instantaneous interactions between the particles are described using a potential energy function. Molecular mechanics (MM) are then used to calculate the forces experienced by the particles, as well as the acceleration, and the new coordinates of the particles. A time-step, δt is used for the calculations, which is usually in the range of femto to pico second. This is iterated over millions of times, allowing for the propagation of the particles, leading to dynamical motions of the system observed. These principles were first created with molecular dynamic simulations of systems of atomistic resolution in mind, however, they can also be applied to CG models.

2.2 Forcefields

In the context of this thesis, the term forcefield refers to the functional form of the potential energy function that is used to simulate particle interactions, as mentioned previously. It also covers the set of parameters that are associated with it. In general, these energy functions contain a term governing the interactions of the non-bonded particles and a term for the bonded interactions.

$$U(r^n) = \sum U_{bonded} + \sum U_{non-bonded} \tag{2.1}$$

Where U is the potential energy of n particles, with positions r^n.

These components can be further subdivided into the terms shown in figure 2.1. In order for these energy terms to be able to reproduce the real dynamics of a system, as accurately as possible, the force constants, bond lengths and partial charges are to fit theoretical and experimental calculations. In the case of theoretical calculations, quantum mechanics are usually used, while in the case of experimental metrics, NMR and EPR measurements can be used, along with crystallographic data. A variety of forcefields has been developed, with slightly different values and approximations for

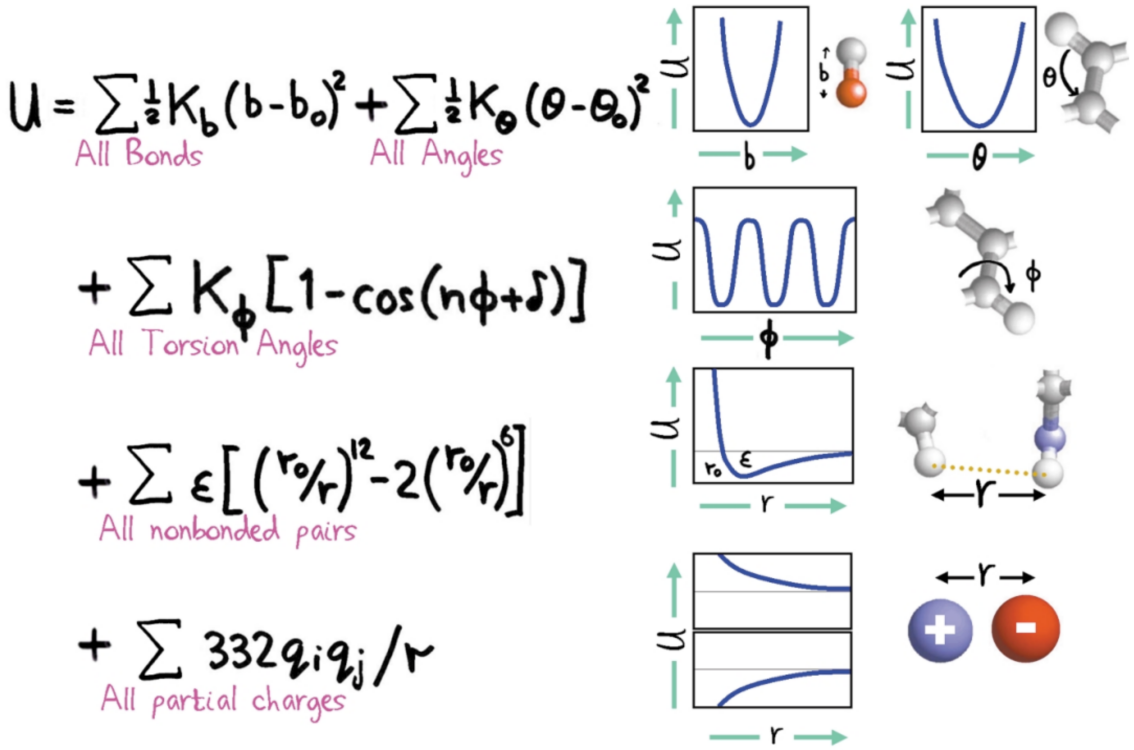


Figure 2.1: A figure showing the typical energy functions used in a biomolecular forcefield, along with a graphical representation. The figure was reproduced from [167]. It should be noted that for certain forcefields additional energy terms can be used, such as harmonic terms, for a more accurate representation of the properties of the system. For certain, high-level forcefields, Morse potentials [168] can be used to replace the harmonic terms in order to describe the anharmonicity of bond stretching [169, 170].

the parameters just mentioned such as the Amber forcefield [163], CHARMM [164], OPLS [165], GROMOS [166] and many more. However, all forcefields, ultimately, follow the same functional forms.

2.2.1 Bonded Interactions

Bonded interaction terms are used to describe the forces which act between covalently bonded particles. They include terms to describe parameters such as angles, bond stretching and bond rotation.

$$U_{bonded} = \sum U_{bonds} + \sum U_{angles} + \sum U_{dihedrals} \tag{2.2}$$

In order to describe the stretching of bonds, a potential derived from Hooke's Law

is usually used where the bond length between two particles, j and i, is described by b_{ij}. The b_{ij} corresponds to the instantaneous length of the bond between particles i and j, b^0_{ij} is the reference bond length and k^b_{ij} is the spring force constant.

$$U_{bonds}(b_{ij}) = \frac{k^b_{ij}}{2}(b_{ij} - b^0_{ij})^2 \tag{2.3}$$

In order to describe the bond angles (θ) between three particles, i, j and k, a similar potential is used. The same definitions as in equation 2.3 are used.

$$U_{angles}(\theta_{ijk}) = \frac{k^\theta_{ij}}{2}(\theta_{ij} - \theta^0_{ij})^2 \tag{2.4}$$

In some higher level forcefields, Morse potentials [168] can be used, instead of the above parameters, to more accurately represent the anharmonicity of bond stretching [169, 170]. Morse potentials have the advantage of describing more accurately the vibrational structure of a molecule. This is due to the fact that they contain explicit effects for bond anharmonicity and bond breaking [169, 170].

Dihedral angles are defined as the angles between two planes. Taking the coordinates of four particles, i, j, k and l, one plane corresponds to the ijk and the second one to jkl. Proper dihedral potentials are periodic in nature, with energy minima corresponding to the canonical 'eclipsed' conformations, such as in the case of butane, and minima corresponding to sterically favoured, staggered conformations [171].

$$U_{dihedrals}(\Phi_{ijkl}) = k_\Phi(1 + \cos(\eta\Phi_{ijkl} - \Phi^{ref}_{ijkl}) \tag{2.5}$$

Where Φ_{ijkl} is the instantaneous dihedral angle, Φ^{ref}_{ijkl} is the reference value, k_Φ is the force constant and η is the multiplicity which indicates the number of minima the bond rotates around 360°.

In the cases where planar structures need to be conserved, such as benzene, or the chirality of tetrahedral groups needs to be maintained, improper dihedrals can also be used. These are usually harmonic in nature, similar to the functions describing the bond and angle terms.

2.2.2 Non-bonded Interactions

In a classical forcefield, a Van der Waals term is used to describe the interactions of non bonded particles. This is modelled through the use of a Lennard-Jones potentia. Coulomb's law is used to described the electrostatic terms.

$$U_{non-bonded} = \sum U_{VdW} + \sum U_{Electro} \tag{2.6}$$

The Lennard-Jones potential is applied to all particles in the system. The 6-12 form is used in most forcefields where the $(\frac{1}{r_{ij}})^{12}$ term describes repulsive forces at short range and the $(\frac{1}{r_{ij}})^{6}$ describes attractive forces at long range.

$$U_{VdW}(r_{ij}) = 4\epsilon_{ij}\left(\left[\frac{\sigma_{ij}}{r_{ij}}\right]^{12} - \left[\frac{\sigma_{ij}}{r_{ij}}\right]^{6}\right) \tag{2.7}$$

Where r_{ij} represents the instantaneous inter atomic distance between particles i and j, ϵ_{ij} represents the well depth of the Lennard-Jones potential and σ_{ij} is the distance between particles i and j at which the energy is zero. The Lorentz-Berthelot rules are used to compute the values for σ_{ij} and ϵ_{ij}.

$$\sigma_{ij} = \frac{1}{2}(\sigma_{ii} + \sigma_{jj}) \tag{2.8}$$

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \tag{2.9}$$

Where σ_{ii} and σ_{jj} represent the collision diameters for particles i and j and ϵ_{ii} and ϵ_{jj} represent the well depth specific for these particles.

As mentioned previously, a Coulombic term is also used to describe the electrostatic interactions of charged particles. This is not only used to describe charged particles, such as ions, but is also applied to atoms with different charge distributions, due to the difference in the electronegativity of their constituent atoms, which have been assigned partial charges.

$$U_{Electro}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.10}$$

Where $U_{Electro}(r_{ij})$ is the Coulombic potential between two atoms, i and j, that have partial charges of q_i and q_j respectively and are separated by a distace of r_{ij}. The ϵ_0 term represents the permittivity of free space.

As it can be seen from the equations described previously, Lennard-Jones potentials have a rather steep rate of decay, compared to the electrostatic potentials. Therefore, a cut-off distance is usually set, after which the Lennard-Jones interactions are ignored. However, electrostatic interactions, which have a less steep decay and therefore a longer range, are treated quite differently. This will be discussed in a subsequent section.

2.2.3 Energy Minimization

The first step, before initiating a molecular dynamics simulation, is to subject the system ton an energy minimization step. This is done to alleviate incorrect geometries and steric clashes that might be present in the initial coordinates of the system. This is achieved by adjusting the coordinates of the system until the potential energy reaches a local minimum, based on the forcefield parameters described above. If this step is skipped, the incorrect geometries or steric clashes can lead to the accumulation of extremely high forces during the simulation run, causing it to crash. In the simulations described in this thesis, the steepest decent algorithm [155], as implemented in GROMACS [172], was used to energy minimize the systems.

Energy minimization is the numerical procedure for finding a local or global minimum on the potential energy surface. In the majority of energy minimization algorithms, these minima are achieved by determining the slope between the starting and end point of the energy surface. If the slope is positive it is an indication that the new coordinates are too high on the energy surface and if the slope is negative, it is an indication that the coordinates are too small. The numerical procedure is then used to adjust the coordinates and the slope calculations are repeated until a slope of zero is achieved.

$$\chi_{new} = \chi_{old} + correction \tag{2.11}$$

Where χ_{new} corresponds to the new energy surface coordinates and χ_{old} corresponds to the starting coordinates on the energy surface. The correction term corresponds to the correction applied by the numerical procedure.

In the Newton-Raphson method, a taylor series expansion of the potential energy at the current geometry is used to calculate the correction term. The first and second derivatives of the potential energy, also known as slope and curvature respectively, are used, making it a computationally expensive method. In the steepest descend minimization algorithm, the second derivative is approximated to a constant value, making it a much faster method.

2.2.4 Integration Methods

The forcefields described previously are used to govern the instantaneous interactions of the of the particles within the system. Finite difference methods are then utilised to integrate Newton's equations of motion and produce the updated coordinates of the system, which resulted from those interactions. This is done by splitting the integration into smaller parts, which are separated by a time step of δt. The force experienced by each particle, at a time t, is calculated as the vector sum of all its interactions with the other particles. From this, the acceleration of the particle can be calculated and, combined with the velocity and position of the particle, at time t, can be used to calculate its updated coordinates at time $t+\delta t$. This assumes that the particle experiences a constant force during the δt period. As a result, care should be taken to ensure the selection of an appropriate time step.

A variety of different algorithms have been proposed for the integrations of Newton's equation during a MD simulation run. For the simulations run in this thesis, the leap-frog integrator [173], as implemented within GROMACS, was chosen. It is an extension of the Verlet algorithm [174], taking its name from the fact that particle positions are updated at integer time steps while velocities are updated at half integer time steps, thus essentially 'leap-frogging' over one another. The integration is described by:

$$r(t+\delta t) = r(t) + \nu\big(t+\frac{\delta t}{2}\big)\delta t \tag{2.12}$$

$$\nu\big(t+\frac{\delta t}{2}\big) = \nu\big(t-\frac{\delta t}{2}\big) + a(t)\delta t \tag{2.13}$$

Where $a(t)$ is the acceleration of particle i at time t, as it was calculated from

$$\frac{F_i(t)}{m_i} \tag{2.14}$$

Where F_i is the force exerted on particle i at time t, and m_i is the mass of particle i.

This process depends on knowing the particles' velocities from the previous integration step. In order to generate initial velocities for the first step of the algorithm, a set of random initial velocities, canonically taken from a Maxwell-Boltzmann distribution at a user specified temperature, are assigned to the particles.

2.2.5 Choice of Time Step

The frequency at which Newton's equations of motion are integrated, is defined by the time step δt, as stated previously. This parameter is selected by the user. Choosing too small a value is computationally inefficient while choosing too big of a value will fail to adequately sample bond vibrations. This can lead to their deviation from their equilibrium values, resulting in too high forces and simulation errors. The choice of the best time step is dependent upon the integrator, the forcefield, the composition of the system and the bond constraints used. In the simulations performed in this thesis, a 20 fs time step was used for the simulations employing the MARTINI forcefield.

2.2.6 Bond Constraints

Covalent bonds with high vibrational frequency are typically constrained to their equilibrium values during a MD simulation. The constraint algorithm is applied after the integrator has updated the particles coordinates and is used to correct bonds

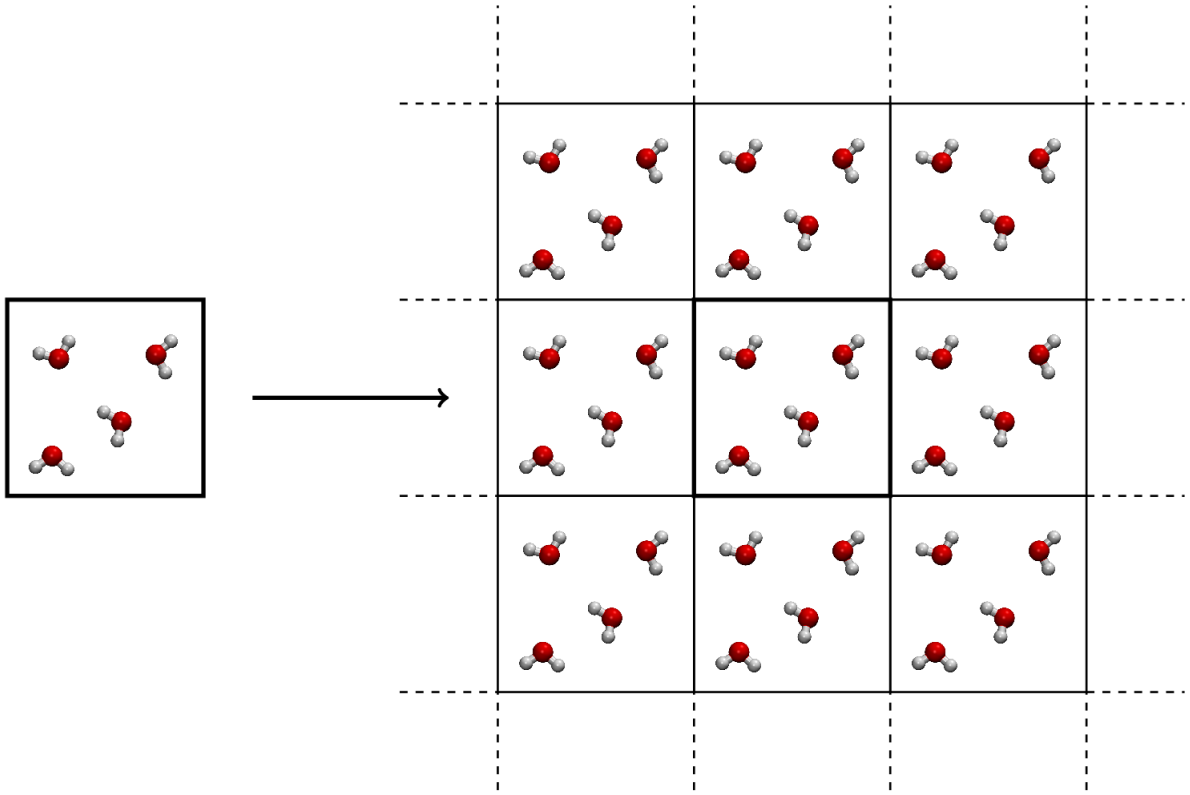


Figure 2.2: A two dimensional representation of the PBC concept. The unit cell, shown in the center, is surrounded by replicas of itself, in such a manner that if a particle diffuses across the periodic boundary of one side it will appear from the other side.

that deviate from their equilibrium values. This is particularly useful for bonds with high frequency vibrations, such as between hydrogen and heavy atoms, and allows for the use of a higher time step. For the simulations performed in this thesis, the LINCS algorithm [175] was used to restraint covalent bonds.

2.2.7 Periodic Boundary Conditions

One of the first questions raised regarding simulations of a finite number of particles, was how to treat the system at its boundaries. Treating the boundaries as a hard surface that the particles are not allowed to cross would lead to system perturbations and non-physical behaviours. Periodic boundary conditions (PBC) is a method to better approximate the behaviour of particles as if they were in the bulk phase. A way to visualise this approach would be to imagine a central simulation box with replica boxes around it, extending in all directions (figure 2.2).

Particles have the ability to diffuse from one side of a boundary and reappearing on the opposite side. Particle interactions are also calculated across the box. As a result, care needs to be taken to ensure that the system is big enough to ensure that the particles, or the protein molecules, cannot interact with themselves across the PBC, leading to artefacts. The minimum image convention, along with a distance cut-off, is applied to non-bonded interactions, meaning that a particle can only interact with other particles and image particles that are located within half a unit cell dimension. Despite this, the PBC approach still comes with certain limitations, especially in the case of membrane simulations, such as undesirable system ordering, influences on the wavelength of membrane fluctuation and challenges in dealing with finite size issues [176, 177, 178].

2.2.8 Temperature and Pressure Coupling

Given the simulation protocol described so far, it would lead to an NVE ensemble. This means that the number of particles (N), volume (V) and energy (E) would be constant during the simulation. Usually, one would like to simulate a NPT or NVT ensemble, since controlling the temperature or pressure during the simulation is of particular importance in order to approximate experimental or biological conditions as closely as possible. To this end, a variety of algorithms have been produced in order to control the temperature (thermostats) and the pressure (barostats) of a simulation.

Thermostats

The kinetic energy of a system is described by the following equation.

$$E_{Kinetic} = \frac{1}{2}\sum_{i=1}^{N} m_i v_i^2 = \frac{3}{2} N K_B T \tag{2.15}$$

Where m_i is the molecular mass, v_i is the velocity of a particle i, T is the instantaneous system temperature and K_B is the Boltzmann constant. The temperature is maintained at the desired value through the application of a thermostat algorithm. This works by coupling the system to an external heat bath, which is maintained at the required temperature, T_{ref}. Particle velocities are scaled as a function of the

deviation of the systems temperature from the reference temperature.

$$\frac{dT(t)}{dt} = \frac{1}{\tau_t}(Tref - T(t)) \tag{2.16}$$

The coupling constant, τ_t, determines the strength of the coupling and hence, the possible magnitude of temperature fluctuations. This approach is used in the Berendsen thermostat [179]. One of the drawbacks of the Berendsen thermostat is the fact that it 'suppresses the fluctuation of the kinetic energy' leading to an improper canonical ensemble, if examined rigorously, which would lead to incorrect sampling. Alternative thermostats are also available, such as the Nose-Hoover [180], Andersen [181] and velocity rescaling methods [182]. In the simulations run for this thesis, the velocity rescaling temperature coupling method was used to maintain the temperature at 323 K. This is essentially a Berendsen thermostat with an additional stochastic term that ensures a correct kinetic energy distribution.

Barostats

The same principle is applied in order to control the pressure of the system, through the coupling to an external 'pressure bath'.

$$\frac{dP(t)}{dt} = \frac{1}{\tau_p}(P_{ref} - P(t)) \tag{2.17}$$

Where *P(t)* us the pressure at time *t*, P_{ref} is the reference pressure and τ_p is again the coupling constant. This approach is used in the Berendsen barostat, which was used in the simulations of this thesis, to maintain a pressure of 1 bar. This also allows for the change of the simulation box size and shape, while the pressure is maintained at the desired value. In an anisotropic scheme, this can be an important consideration when setting up systems which are sensitive to a change of shape or size of the simulation box. The Berendsen barostat [179, 183] also suffers from the drawback of not yielding the correct NPT ensemble. For CG simulations, the Berendsen barostat appears to have minimal or no effect in the resulting ensemble, making it a viable choice. Nevertheless, alternative barostats are also available, such as the Parrinello-Rahman [184], and MTTK pressure control [185, 186].

The pressure of the system is usually calculated via the virial theorem of Clausius [155]. The virial is defined as the expectation value of the sum of the products of the coordinates of the particles and the forces acting on them. The virial theorem states that the virial is equal to $-3NK_BT$ [155].

In an ideal gas, the only forces experienced by the particles is due to the interactions with the container. As a result the virial is equal to -3PV. This result can also be obtained directly from $PV = NK_BT$ [155].

In the case of real gases or liquids, the forces experiences by the particles is also due to the interactions between them, hence a contribution due to these forces should also be taken into account when calculating the pressure [155]. As a result, the pressure of a system can be expressed as:

$$P = \frac{1}{V}\left(NK_BT - \frac{1}{3}\sum_{i=1}^{N}\sum_{j=i+1}^{N} r_{ij}f_{ij}\right) \tag{2.18}$$

Where f_{ij} is the force acting between particles i and j and r_{ij} is the distance between these particles [155].

2.2.9 Long Range Interactions

As mentioned previously, Lennard-Jones (LJ) interactions, being the most computationally demanding to compute, are usually ignored after a certain cut-off. This is also supported by the fact that LJ interactions decay with an r^{-6} dependence, as particle separation increases, quickly approaching zero (figure 2.1). A variety of cut-off schemes have been developed, ranging from a single, 'hard' cut-off to shifting function cut-off schemes, which modify the potential energy function around the cut-off distance in order to smoothly shift it to zero [187, 188]. A 'neighbours-list' is employed to keep track of the particles found within the cut-off distance that is updated every few steps, as the coordinates of the system evolve [188].

Coulombic interactions are more long ranged, compared to LJ, decaying with an r^{-1} dependence (figure 2.1). This allows for the use of a single cut-off scheme. The Particle Mesh Ewald (PME) method [189, 190] is usually used to treat coulombic interactions. In this approach, particle interactions are divided into short-range and long-range potentials, which are calculated in real space and Fourier space,

respectively, allowing for rapid convergence.

2.3 Coarse Grained Simulations

The principles and methods described here, were developed with simulations of atomic resolution in mind. Since then, a range of forcefields that employ a more abstract representation of a system have been developed. This is achieved by 'coarse graining' multiple atoms into groups represented by a single particle. The main advantage of such an approach is the reduction of the total number of particles in the system, allowing access to higher time scales, due to the reduced computational demands.

The appropriate level of abstraction is dependent of the biological question which is being examined. Investigation of highly detailed processes, such as studying the interactions of the binding of glycine to the glycine receptor, required atomistic detail simulations. In this aspect, an atomistic forcefield, such as [163], CHARMM [164] or OPLS [165] would be suitable. Examination of slower, larger scale phenomena, such as the clustering of proteins on a membrane, are more suited to a coarse-grained approach. A wide range of coarse-graining models have been developed, with different mapping schemes, ranging from 'mild' coarse-graining, where two to three heavy atoms are mapped to a larger group, to 'super' coarse-grained, where entire lipids are represented by three particles and residues are grouped into a single particles [191, 192, 193, 194, 195, 196].

Mild coarse-graining allows for the retention of some of the structural and molecular details of the biological system, however, these are lost in the case of super CG. In an attempt to maintain the speed of CG simulations while preserving a higher level of detail, hybrid models have been developed [197]. The idea is to try and preserve the principle component of interest, such as the active site of a target, in as much detail as possible, while the components considered of less importance are represented in a more abstract manner. Some of these models might even incorporate a dynamic CG approach, with particles changing resolution depending on their crossing of a barrier [198]. However, such models are still at their infancy and have not been widely adopted by the community yet.

2.3.1 The MARTINI forcefield

The MARTINI forcefield is 'milder' version of coarse-graining, first developed in 2004, by Marrink et al., for simulations of lipids [194]. Since then, it has gone through extensive overhauling in 2007 [199], expanding to include parameters for proteins [200], carbohydrates [201], nucleic acids [202, 203] and non-biological structures [204]. It has gained great popularity within the field and has been used in applications such as the simulations of entire virion envelopes [205, 206], vesicles [207], lipid-protein membranes [208] and nanotechnology [209]. The principle idea behind MARTINI is that a collection of well parametrised building blocks can be mixed and matched to create higher-order structures, without the need for continuous re-parametrisation of the resulting structure.

The forcefield uses, on average, a four to one mapping scheme, where four heavy atoms are CG into one larger bead. The CG beads are divided into four major types: apolar (C), non-polar (N), polar (P) and charged (Q). Apolar beads represent groups of heavy atoms with net hydrophobicity, non-polar beads represent mixed hydrophobic/hydrophilic groups, polar beads represent hydrophilic groups and charged beads represent groups with an ionic charge. P and C bead types are further divided into five categories, ranging from 1 to 5, based on their polarity, with P1 denoting low polarity and P5 denoting high polarity. Furthermore, N and Q type beads are broken down into four categories, d, a, da and 0, based on their ability to mimic hydrogen bond forming. d stands for a hydrogen bond donor, a for an acceptor, da for a donor/acceptor and 0 for neither. S and T type particles have also been introduced in order to model the ring geometry of structures such as aromatic residues, sterols and nucleic acids. In these cases, a mapping scheme of three or two to one is employed, providing higher resolution.

In the case of protein residues, each residue is assigned a single particle, representing the back bone heavy atoms, with each type based on the local secondary structure. Figure 2.3 shows the CG models of the twenty natural amino-acids, along with the bead type that is used for each one. Figure 2.4 shows two amino-acids, arginine (A) and Tryptophan (B) in their all atom and CG representations.

Bonded interactions between the beads are treated in the same manner as described previously, with harmonic potentials used to mimic bonds and angles.

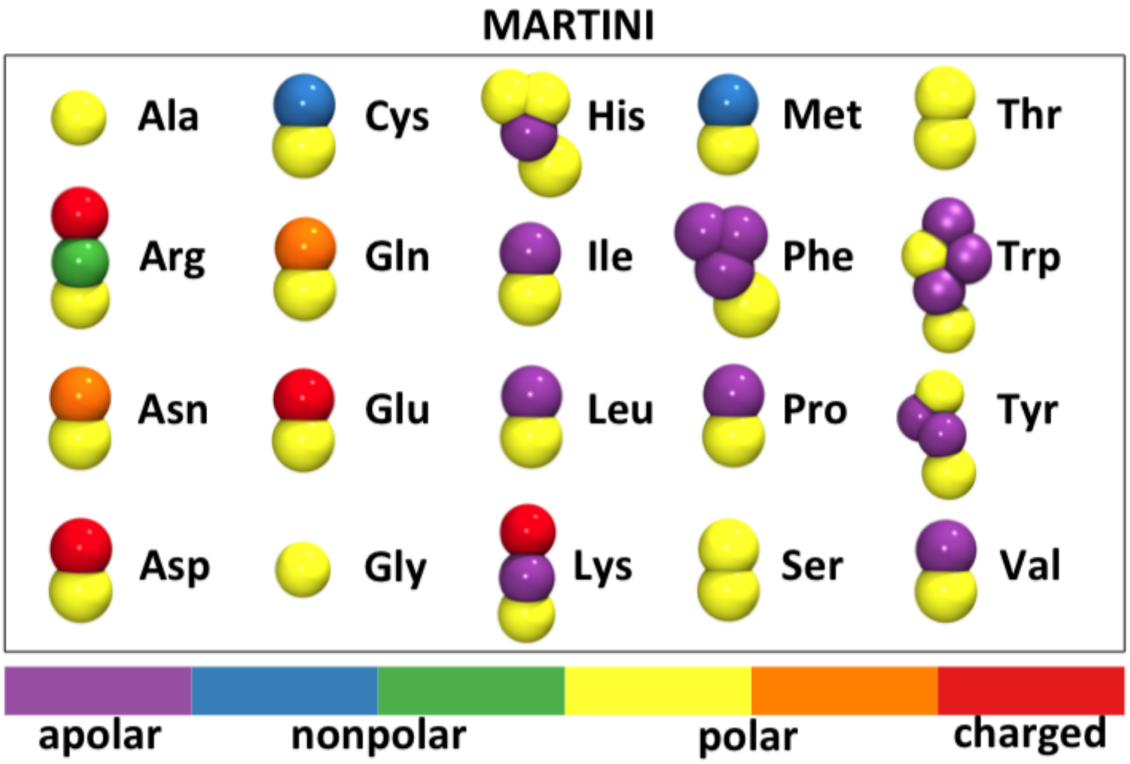


Figure 2.3: A figure showing the CG mapping of the twenty natural amino-acids into their MARTINI representations, along with the type of beads that are used to make them. Figure reproduced from [199].

Dihedrals are maintained through the use of harmonic and periodic potentials. The non-bonded interactions are also treated with a LJ potential, with different well depth ϵ values for each pair wise interaction. The σ parameter, denoting the distance at which the energy is zero is set to 0.47 nm for all beads except for the S-type beads ($\sigma = 0.43$ nm), the T-type beads ($\sigma = 0.32$ nm) and the interactions of Qx beads with C1 and C2 ($\sigma = 0.62nm$). Due to their more detailed mapping, the ϵ value for the S and T type beads ar reduced to 75% of their parent bead values. Partial charge interactions are taken into account through the LJ potentials while the ionic interactions are modelled through the use of a coulombic potential. A shifted cut-off scheme is used for the LJ interactions where they are smoothly turned off between 0.9 and 1.2 nm while the coulombic interactions are shifted to zero between 0.0 and 1.2 nm, to avoid noise from energy discontinuity.

The parametrization of the MARTINI model is based on comparison to data coming from experiments and atomistic simulations. Bond lengths and angles within protein structures were derived based on a survey of over 2000 structures from the PDB [200]. Moreover, the experimental partitioning free energies of small molecules

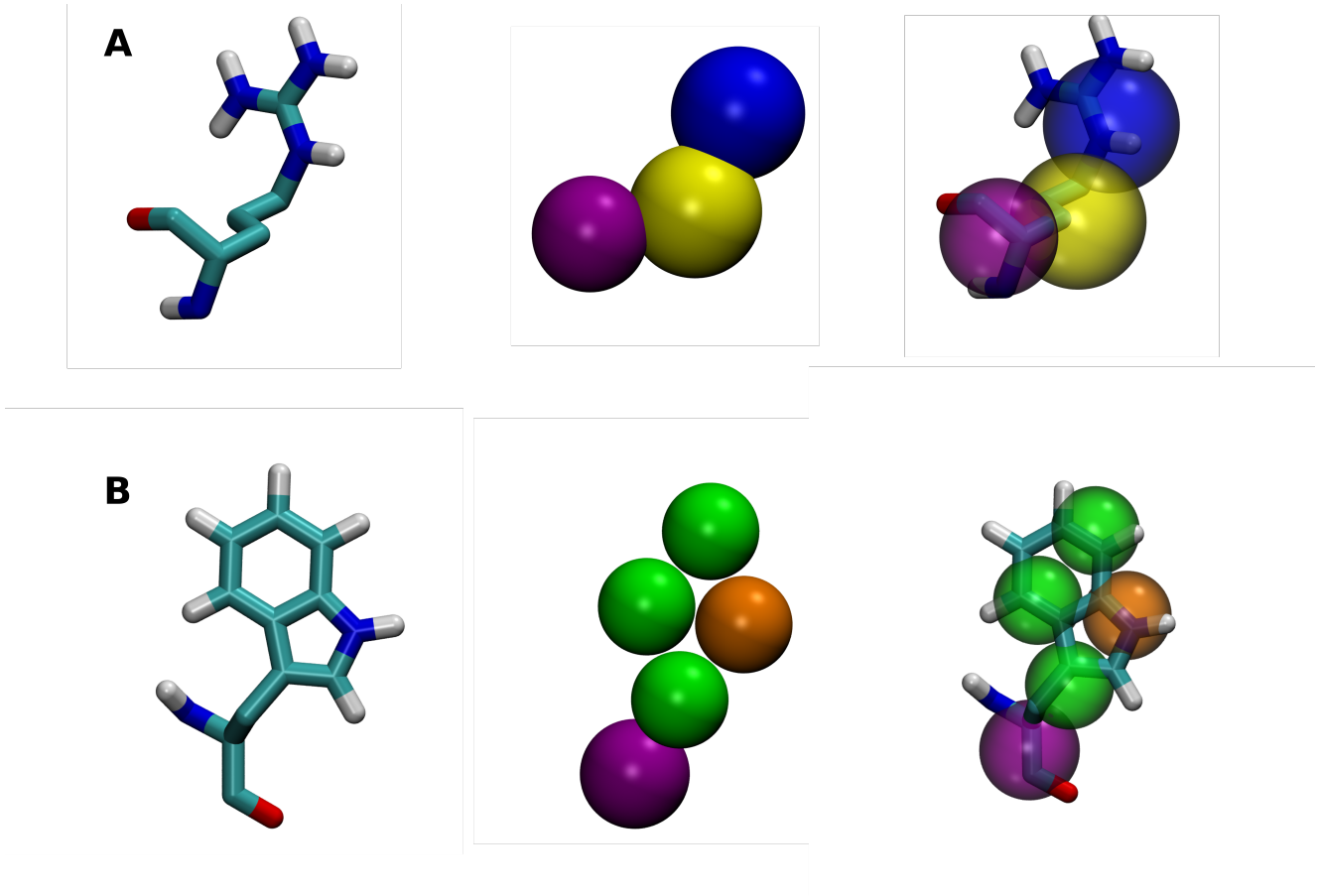


Figure 2.4: A figure showing two amino-acids (A) arginine and (B) Tryptophan in their all atom (left) and CG representations (middle). The superimposed representations are also shown (right).

analogues of amino acid side chains between aqueous solutions and organic phases, derived from atomistic simulations, were used to parametrize those in MARTINI. The values were optimized within 1-2 kT for most compounds. Similarly, in the original model, the free energy of partitioning of water in hexane was calculated, through simulations, to be 24.2 kJ/mol, compared to the experimental value of 25 kJ/mol [194]. Other system properties, such as the area per lipid in a lipid bilayer, can be readily compared with the experimental values. A full discussion of all the parametrization methods used in MARTINI is beyond the scope of this thesis, and can be found in the original MARTINI papers [194, 199, 210].

2.3.2 Secondary and Tertiary Structure in MARTINI

As it can be understood, MARTINI lacks any explicitly represented hydrogen bonds. Therefore, additional protocols are needed to maintain the secondary and tertiary structure of a protein.

In the case of secondary structure, DSSP [211] is first used to assign a structure type to each particle. Afterwards, the structural dynamics of these particles are restrained around an equilibrium which is determined based on their structure type and maintained through the use of harmonic and periodic potentials. One disadvantage of the model is the fact that it does not strictly follow the DSSP classification. For instance, every helical structure is merged into a single, generic type of helix, made to mimic an a-helix. As a result, it is hard to model conformational transitions of the secondary structure with this model.

In order to maintain the tertiary structure of a protein, two methods are most commonly used, a traditional Elastic Network Model (ENM) or the most recently developed Elnedyn [212] model. The principle idea behind both of these models is to apply a harmonic bond potential on the backbone beads that are found within certain distance of each other. This creates a web of bonds that ensures the initial tertiary structure is maintained, while allowing a small degree of local backbone dynamics. Side chain beads are usually not restrained. In the traditional ENM, harmonic bonds are applied to all backbone beads within 0.5 - 0.9 nm from each other, with a force constant of 500 kj/mol/nm^2. In the case of the Elnedyn model, a cut-off of 0.9 nm is used. The main difference between the two methods is the fact that in the case

of the ENM, the backbone neads are positioned on the center of mass of the heavy atoms that were clustered, while in the case of the Elnedyn model, the back bone beads are centred on the C_α atom of the residue. As a result, the Elnedyn model requires the alteration of the underlying mapping and the non-bonded parameters, as well as applying the harmonic potential.

Although the force constant and distance cut-off are highly tunable variables, with the idea being that they should be tuned to the ideal parameters for each target, in practice, the default parameters are widely used. This is done due to the fact that they have been shown to be adequate for a wide range of targets and scales. In the simulations performed for in this thesis, the ENM was used.

2.3.3 Water Model

In the standard MARTINI model, four atomic water atoms are mapped onto a larger water bead, which is of the P4 subtype. As a result, these water beads do not feel the electric fields of charge particles and, in common with most all atom water models, they cannot reproduce polarization effects. In an effort to address this drawback, some effort was put in developing a polarizable MARTINI water model [213]. Two further atoms were created, name W_P and W_M, and were attached to the water bead via harmonic bonds. Their orientation was fixed by a harmonic angle potential. They were bearing charges of +q and -q respectively and they were parametrized to only interact via coulombic interactions, meaning they were only visible to Q subtype beads. This model has shown to more accurately model the dielectric screening properties of bulk water, along with its freezing point. It has also shown to perform better in modelling certain protein-lipid interactions, such as in the case of glycolipids clustering around a protein [214]. This can be explained by the fact that glycolipids posses large, solvent exposed head groups and water might play a more prominent role in their interactions with the protein.

The main limitations behind this polarizable water model is the uncertainty surrounding its parameters, due to the fact that they were developed through simulations with the standard water model, and its inability to interact with non-polarizable molecules e.g. proteins and lipids. Moreover, the use of additional particles, results in an increase in the computational resources needed and the time

needed to run a simulation. In certain cases, such as a pure water system, this model can be up to three times slower compared to the standard water model.

An implicit water model has also been developed for MARTINI [215]. This attempt is going to the opposite direction of the polarizable water model, towards high levels of abstraction and approximation. Its main use is for macro-scale simulations of large systems, where bulk water forms a large proportion of the total particles of the system. A lot of time time is spent simulating these waters which are not of primary interest. In such cases, an implicit water model can offer significant speed ups while still maintaining the general properties of the system. The drawback of this model is that proteins have been shown to not behave well as well as having the same parameters concerns as the polarizable water model.

For the simulations run for this thesis, the standard MARTINI water model was used.

Chapter 3

Dynamic Pharmacophore Probes

Chapter Outline

In this chapter, the details of the simulations and the analysis behind the dynamic pharmacophore probes approach will be presented along with a justification for the use of the selected parameters. The test targets and methodology for the validation of this method will then be described, followed by a qualitative and quantitative analysis of the results. Finally, the advantages and limitations of this method will be described, along with its potential use and application in computer aided drug design efforts.

3.1 Introduction

The majority of proteins have pockets which have evolved to bind a variety of ligands, ranging from small molecules to long and complex lipids. The binding of ligands to these pockets can exert a variety of effects, from inhibition to activation [216]. As a result, ligand specificity and determining the key interactions driving the formation of protein-ligand complexes, is of paramount importance [217]. One area where this knowledge provides invaluable insights is the design of new and more effective pharmaceuticals [144, 217].

Techniques evolved to explore the propensities of small molecular fragments to bind targets, together with X-ray crystallography [218] and NMR [219], can be used to identify specific types of protein-ligand interactions. To this end, the use

of fragment library screening has been successfully utilised for a variety of targets [220, 221]. Furthermore, additional experimental methods, such as multiple solvent crystal structure (MSCS), have also been developed [92, 93]. However, as mentioned in the introduction section, they require a considerable amount of time and resources, limiting the number of targets on which they can be applied.

CADD methods, and more particular SB methods, encompass a variety of approaches, such as GRID [39, 37, 40, 41] based or "Hotspot" identification methods [88], to explore the chemical space of a potential binding site and translate the results into useful information.Mixed Solvent MD methods have also made substantial contributions to this area. Methods such as SILCS [108], MDmix [105] or MixMD [123] allow for the fragment probes to explore the binding pocket of the target protein, in what could be described as a computational equivalent of MSCS. Although these methods require less time and resources than their experimental siblings, they still require a considerable amount of computational resources, with the simulation times currently being limited to the ns scale. Furthermore, they suffer from the inherent limitation of identifying "spurious" interactions, that could impede drug design efforts.

Here we present a new method, based on Coarse Grained (CG) MD simulations, through the use of the MARTINI forcefield [194], in an attempt to overcome the drawback of limited simulation time and high demand for computational resources, of the methods based on traditional MD. The use of CG allows for a significant, fourfold to tenfold, speed-up, opening the door to simulation times on the μs time-scale. Moreover, the need for computational resources is greatly reduced, with a simulation time of 2μs, requiring only 12 hours on a workstation.

The use of the MARTINI forcefield also allows for dealing with the identification of "spurious" interactions that plague the previously mentioned approaches. In this method, the MARTINI beads are used as pharmacophore probes with the aim of exploring protein ligand interactions. The parametrisation of the MARTINI beads allows for a very low number of false positives, due to the fact that the different types of beads are only parametrised to interact with bead types that are supposed to be interacting with.

One more limitation of a few of the previously mentioned methods is the assump-

tion that the interactions identified are due to the entirety of the probe molecule used. This can lead to errors when calculating the energies of the probe interaction maps, such as in the case of SILCS' "grid maps", complicating drug design efforts. In this approach, this is mitigated by the nature of the MARTINI beads/probes.

Overall, in this chapter, the capacity of the MARTINI beads to be used as pharmacophoric probes will be demonstrated, along with their potential for identifying protein-ligand interactions that are important for binding. This will be done through the use of a variety of test targets, curated by Ichihara et al, which include apo, fragment and lead bound structures of the same target. Finally, the transformation of the interaction densities of the probes into pharmacophoric points, or models, and the predictive power of these points will be demonstrated.

3.2 Methods

3.2.1 Target dataset

The protein-ligand complexes and apo structures used in this study were selected from the protein dataset collated by Ichihara and colleagues [222]. It includes targets of pharmacological importance for which apo, fragment bound and ligand bound structures, of high resolution, are available. This dataset was also selected in order to achieve a high diversity in the shape, size, accessibility and buriedness of the binding sites. Table 3.1 lists the PDB accession codes of the protein complexes used, along with their binding site characteristics.

3.2.2 Molecular Dynamics simulations

Molecular dynamic simulations were performed using the GROMACS software package (version 5.0.2) [188] with a time step of 20 fs. This time step was chosen based on the standard MARTINI protocol for CG simulations. When choosing the appropriate time step for the simulation, a time step that is much lower, usually a tenfold decrease, than the bond vibrations of the atoms in the simulation, is required, otherwise numerical instability and errors can occur. Due to the increased masses of the beads in the MARTINI forcefield, as a result of their coarse grained nature,

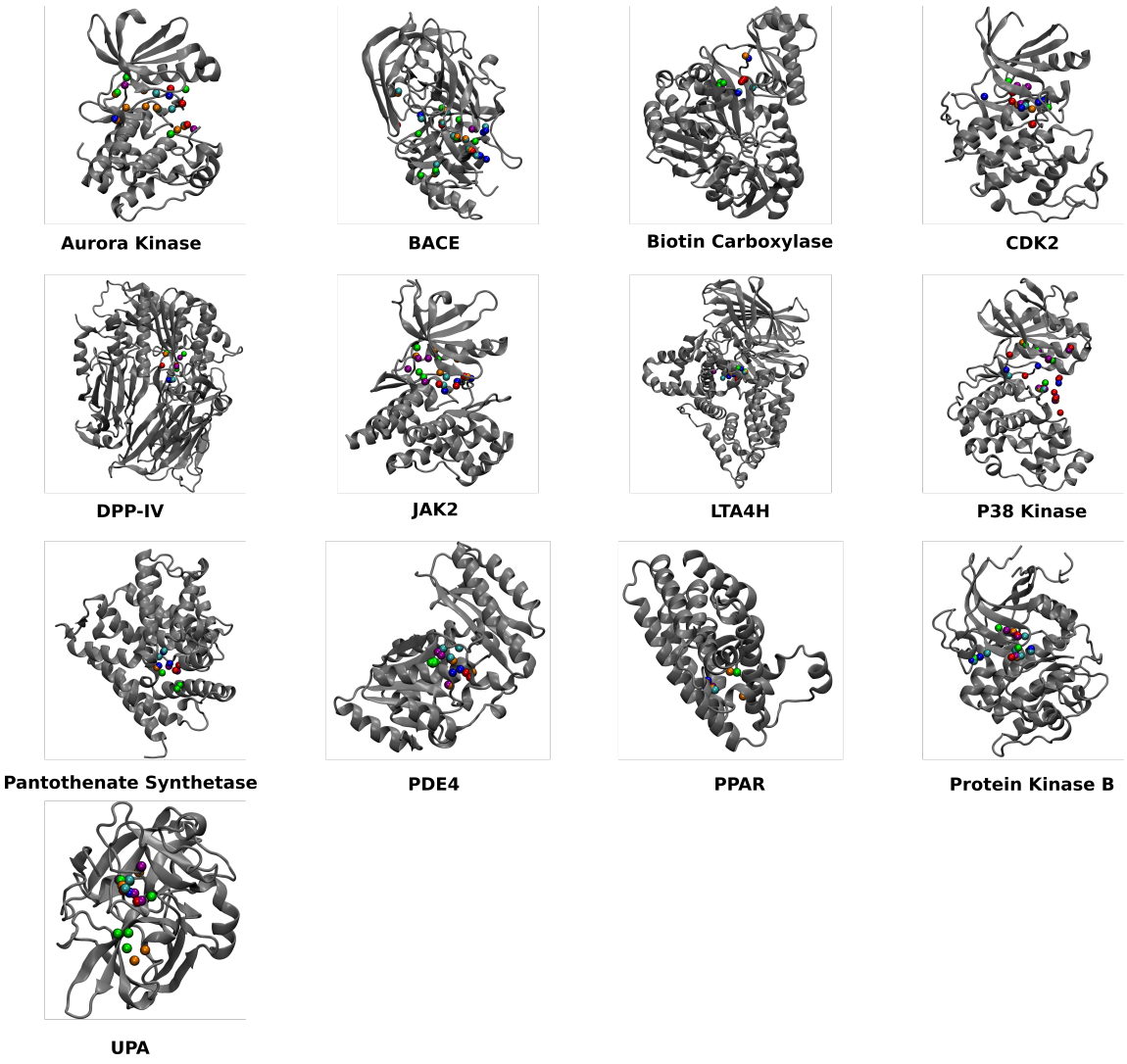


Figure 3.1: A figure of all the targets, from the Ichihara dataset, used in the simulations, along with the resulting pharmacophoric points. Green points correspond to the pharmacophoric points generated by the hydrophobic interaction densities, orange to hydrogen bond acceptors, purple to hydrogen bond donors, cyan to polar, blue to positively charged and red to negatively charged.

Table 3.1: A table of the targets used, their PDB accession codes, their binding site characteristics and their binding site volume. First column of accession codes corresponds to the apo structure while the second and third columns correspond to fragment and lead bound structures respectively.

Target	PDB Accession Code	Binding Site Description	Volume (Å²)
Aurora Kinase A	4J8N, 2W1D, 2W1G	Accessible	1300
BACE	1W50, 2V00, 2VA7	Accessible	700
Biotin Carboxylase	2J9G, 2W70, 2W71	Accessible	800
CDK2	1HCL, 2VTA, 2VU3	Accessible	1300
DPP-IV	1J2E, 3CCB, 3CCC	Partially Accessible	900
JAK2	4ZIM, 3E62, 3E64	Accessible	1500
LTA4H	3B7S, 3FU0, 3FH7	Buried	2700
P38 Kinase	1WFC, 1W84, 1W83	Accessible	2900
Pantothenate Synthetase	3COV, 3IMG, 3IUE	Buried	1600
PDE4	3SL3, 1Y2B, 1Y2K	Partially Accessible	800
PPAR	1PRG, 3ET0, 3ET3	Buried	1600
Protein Kinase B	4C33, 2UVX, 2VO6	Partially Accessible	1000
UPA	3OY5, 2VIN, 2VIW	Accessible	3600

these bond vibrations are much slower, allowing for a higher time step [210].

Isotropic Berendsen pressure coupling was used [179] with a coupling constant of 12 ps. The reference pressure was set at 1 bar for the XY and Z directions. A stochastic rescaling temperature control was used [182], with a characteristic time step of 1 ps, keeping the temperature constant at approximately 323 K. A cut-off of 1.2 nm was used for Lennard-Jones (LJ) interactions with the potential being smoothly switched off between 0.9 nm and 1.2 nm. The Coulombic interactions were calculated using the Reaction Field method [188] with a cutoff of 1.1 nm. These settings were consistently used throughout the simulations, across all the dataset targets.

3.2.3 Forcefield

The CG MARTINI forcefield (version 2.2) [200] was used for the simulations of all the protein targets. The software and the parameters supplied with the forcefield were used to coarse grain the structures and run the simulations respectively.

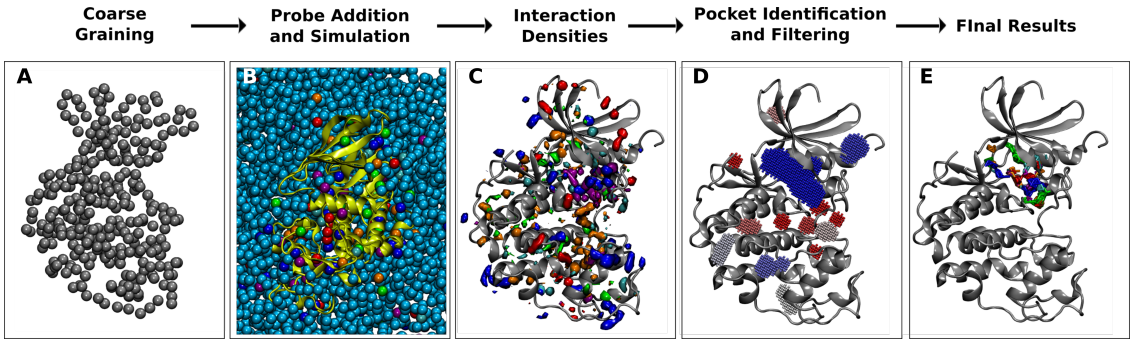


Figure 3.2: Work Flow of the Simulation and Initial Analysis Protocol. A schematic describing the simulation and initial analysis protocol followed. (Panel A) The protein is first coarse grained and placed in a simulation box. (Panel B) The simulation box is solvated and water beads are replaced by pharmacophoric probes. The simulation is run for 2μs. (Panel C) The MDAnalysis software is used to calculate the interaction densities of the probes. (Panel D) The GHECOM cavity finder software is used to identify potential binding sites. (Panel E) The interaction densities are filtered based on the cavities identified.

3.2.4 MARTINI beads as pharmacophoric probes

Six different types of MARTINI beads were used in the simulations to emulate pharmacophoric probes. The property of each type of bead was selected in order to correspond to one of the most important features found in a pharmacophore model. As a result, the six different types of pharmacophoric probes used in the simulations were: hydrophobic probes, probes acting as hydrogen bond donors or hydrogen bond acceptors, polar probes and probes bearing a positive or negative charge of one. The repulsive element of the LJ potential was increased between probes of the same and different types to prevent aggregation. This was done by reducing the ϵ value of their LJ interactions to 1.725 kj/mol. The non-bonded interaction parameters between the probe beads were also modified from the original MARTINI forcefield, resulting in the probes interacting with each other with a reduced Lennard-Jones σ value of 3.2 Å, effectively reducing their size, when they are interacting with other probes. This allows for the easier insertion of the beads in less accessible cavities or pockets. It also allows for an easier comparison to the available pharmacophore models, due to the fact that most pharmacophore models use a radius of approximately 3 Å for each pharmacophoric feature they include. These parameters were selected after extensive testing showed that they produced the optimum simulation conditions.

3.2.5 Simulation protocol and analysis

The ligand free structure of each target was selected and used as a starting point for each simulation. The MODELLER software [223] was used to fill in missing residues or atoms, when necessary. The structures were coarse grained, using the script provided by the MARTINI forcefield, placed in the centre of a cubically shaped simulation box, with a distance of at least 3 nm from each side of the box, and solvated. A number of water beads were then replaced with a number of pharmacophoric probes until a concentration of 100 pharmacophoric probes, of each type, per 16000 water beads, an approximate concentration of 0.86 mM, was reached. Concentrations higher than this resulted in probe aggregation. The simulations were run for 2 µs. The MDAnalysis [224] software package was finally used to calculate the interaction densities of the different probe types (figure 3.2). Three repeats, of 2 μs each, were run for each target.

The interaction densities correspond to the probability of finding a pharmacophoric probe of a certain type in a particular area of the simulation box, with this area being defined as the space region of a cubic voxel with dimensions of 1 Å. These probabilities were evaluated by dividing the total number of pharmacophoric probe occurrences in a particular voxel, over the course of the whole simulation, with the total number of simulation frames.

As it can be seen from panel C, figure 3.2, the resulting interaction densities were quite noisy. In order to further filter the densities and identify potential ligand binding sites, the cavity finding algorithm GHECOM [225, 226] was used on the target proteins. The interaction densities were then filtered based on the cavities identified by discarding any points of a density that were found outside an identified cavity. A complete schematic representation of the workflow can be seen in figure 3.2.

The afore mentioned simulations were repeated, minus the protein. The interaction densities of the probes were again calculated. The inverse Boltzmann equation (eq 1) was then used to assign an energy of interaction (ΔG_i) value to every point of the interaction densities with k_B corresponding to the Boltzmann constant and $\rho(r)$ and ρ_{bulk} corresponding to the interaction densities values obtained from the production and bulk simulations respectively. T corresponds to temperature. As

a result, the ΔG_i can be thought as the difference in energy between finding a pharmacophoric probe in a binding site and finding the same probe in the bulk solvent. These values were then used to further filter the densities and focus on the interaction hotspots.

$$\Delta G_i = -k_B T \ln \left(\frac{\rho(r)}{\rho_{bulk}(r)} \right) \tag{3.1}$$

3.2.6 Convergence

In order to ensure that the 2 µs simulation time would allow for the probes to explore adequate conformational space, the interaction densities of the binding site were calculated for each repeat and each probe type, using Aurora Kinase A as an example. This was done by selecting one of residues known to be located in the binding site, GLU 211 in this case, and using it as a center for a 40 x 40 x 40 Å box. The size of the box was selected to ensure it encompassed the entirety of the binding site. The interaction densities of the probes entering the box were then calculated, for the whole duration of the simulation. These were then plotted in Figure 3.3. Figure 3.3 the same interaction patterns emerging for each repeat, for all the pharmacophoric probe types, indicating good agreement between the resulting interaction densities.

The Kernel Density Estimation (KDE) [227, 228] method was then used to estimate the probability density function of its interaction density value, for each probe type and for each repeat (Figure 3.4). As it can be seen from figure 3.4, the probability density functions for each probe type and for each repeat are almost identical, indicating that the interaction density values from each repeat are the results of the extremely similar probability functions, suggesting that the simulations have converged.

Finally, in an attempt to quantitatively show that the simulations have explored enough conformational space, the two sample variation of the Kolgomorov-Smirnov test [229, 230] was utilised. In this variation of this statistical test, the distance between the empirical distribution of the two samples is quantified. The null hypothesis states that the two samples originated from the same distribution. A low resulting KS value, accompanied by a high p value, indicates that the null hypothesis

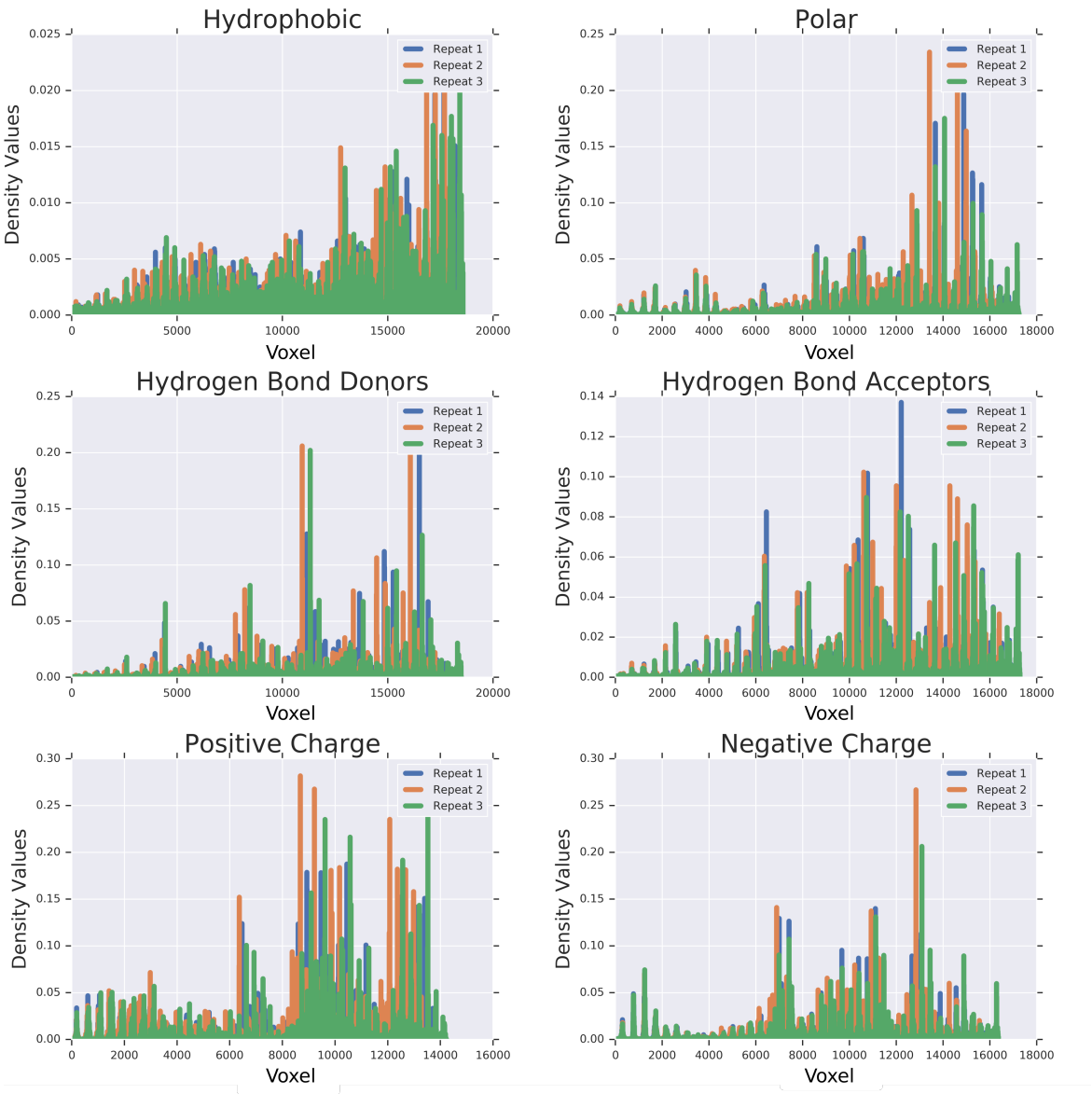


Figure 3.3: Graphs showing the plotted density interaction values, for each pharmacophoric probe type, for each repeat. The voxel number is plotted on the x axis while the density value for this voxel is plotted on the y axis. The interaction patterns for each repeat appear to be matching, meaning that the density is found at the same, or nearby, voxels, for each repeat, supporting the claim that the densities have converged.

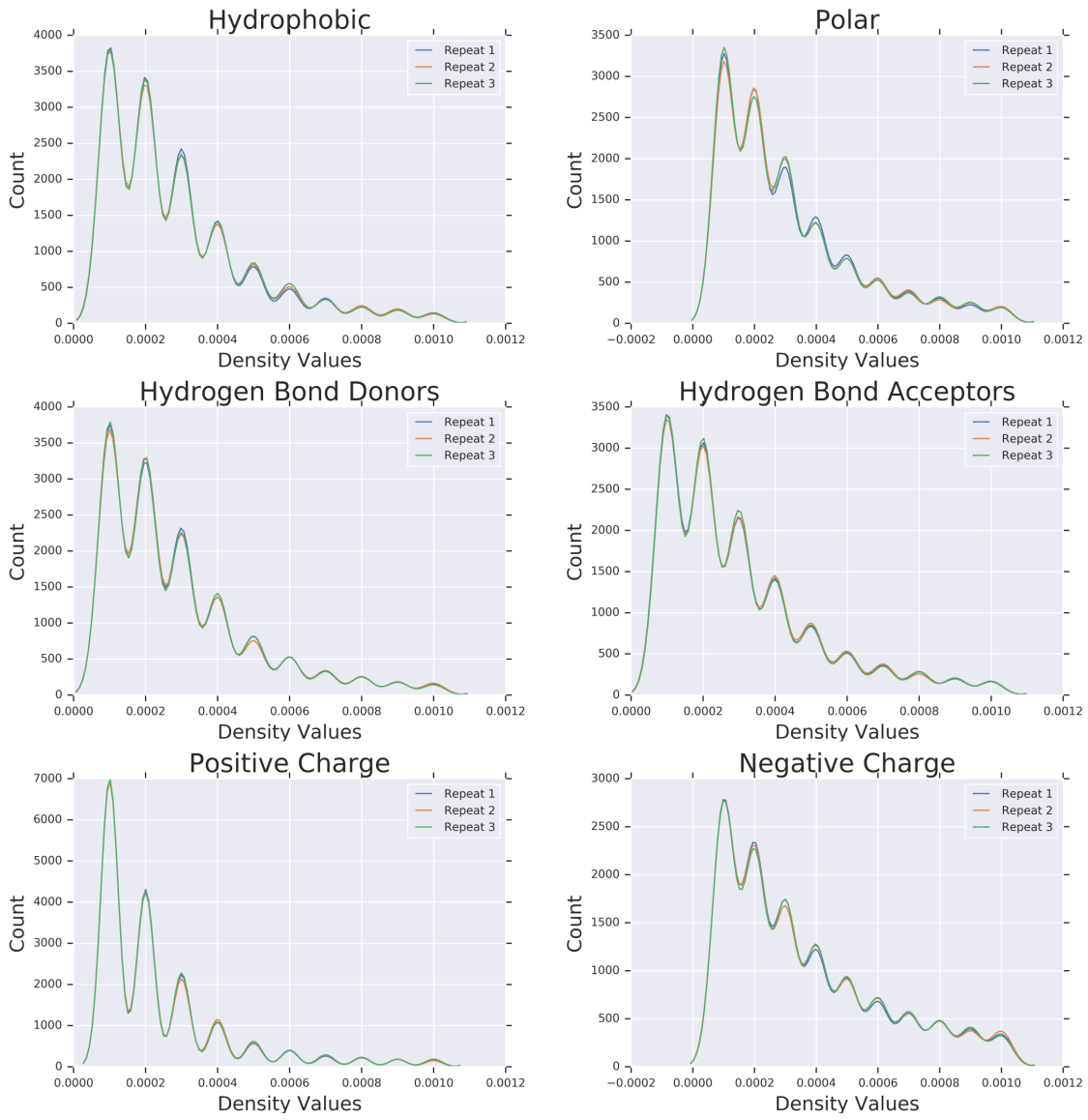


Figure 3.4: Graphs of the plotted probability density function, for each pharmacophoric probe type, for each repeat. The density values are shown on the x axis while the count for each value is shown on the y axis. The high degree of similarity between the probability density functions indicate that the densities have converged.

cannot be rejected. In table 3.2, the KS and p values, between each repeat and for each probe type, were calculated and presented. Examining these values, it can be seen that for every case, the null hypothesis cannot be rejected, meaning that the two samples can be considered as originating from the same distribution, which in turn leads to the assumption that the simulations have converged.

Table 3.2: A table of the calculated KS and p values for each probe type and for each repeat.

Probe Type	Repeats Compared	KS and p Values
Hydrophobic	1st - 2nd	KS = 0.01
		p = 0.56
	1st - 3rd	KS = 0.01
		p = 0.49
	2nd - 3rd	KS = 0.008
		p = 0.58
Hydrogen Bond Acceptor	1st - 2nd	KS = 0.02
		p = 0.67
	1st - 3rd	KS = 0.007
		p = 0.80
	2nd - 3rd	KS = 0.01
		p = 0.88
Hydrogen Bond Donor	1st - 2nd	KS = 0.01
		p = 0.42
	1st - 3rd	KS = 0.04
		p = 0.99
	2nd - 3rd	KS = 0.01
		p = 0.54
Polar	1st - 2nd	KS = 0.02
		p = 0.69
	1st - 3rd	KS = 0.008
		p = 0.49
	2nd - 3rd	KS = 0.02
		p = 0.70
Positive Charge	1st - 2nd	KS = 0.01
		p = 0.47
	1st - 3rd	KS = 0.004
		p = 0.99

Probe Type	**Repeats Compared**	**KS and p Values**
	2nd - 3rd	KS = 0.01
		p = 0.58
Negative Charge	1st - 2nd	KS = 0.08
		p = 0.58
	1st - 3rd	KS = 0.08
		p = 0.49
	2nd - 3rd	KS = 0.07
		p = 0.73

3.2.7 Clustering and pharmacophoric points

The filtered interaction densities were further clustered into pharmacophoric points using the HDBSCAN [231] clustering algorithm. The weighted geometric centres of the clusters were then calculated and defined as pharmacophoric points, with the ΔG_i values calculated previously being used as weights. The residues of the targets, within 4 Å of the pharmacophoric probes were identified and defined as being in contact with the pharmacophoric points. The above analysis was performed through the use of an in-house script. The resulting pharmacophoric points for all the targets in the data set can be seen in figure 3.1.

3.2.8 Contact analysis

Table 3.3: A table showing the PDB accession codes, that were used to identify the residues involved in ligand binding, for each target, along with the residues identified as playing a role in protein-ligand interactions. The residues that were shown to be important for ligand binding (RILB), identified in more than 75% of the cases, are shown in bold.

Target	PDB Accession Codes	Residues Involved in Binding
Aurora Kinase A	5ORV, 5ORL, 5ORS, 5OS0, 5OS1, 5OSD, 5G1X, 5EYK, 2W1G, 5K3Y, 5DNR, 5L8J, 5L8K, 4J8M, 3ZTX	R137, K143, F144, G145, K162, K166, R179, Q185, Y199, **E211**, **A213**, T217, E260, N261, D274
BACE	6EJ3, 6C21, 6BFD, 6BFE, 5MCQ, 5UYU, 5T1U, 5HTZ, 5HU0, 5HU1, 5ENM, 5KQF, 5I3V, 5I3X, 5HD0	D32, G34, Y71, T72, Q73, W76, **D93**, F108, W115, R128, W137, W176, **D228**, G229, G230, T323, N233, R235, D276, D289, G291, T329
Biotin Carboxylase	4XTV, 5MLK, 4XTU, 4XTY, 4XU0, 4MV1, 4MV4, 3RUP, 3RV3, 3OUZ, 2W6N, 2W6P, 2W70, 2W71, 3G8D	S21, Q42, H46, R48, R76, S77, K11, K116, K117, K159, G166, **E201**, **K202**, Y203, **L204**, H209, Q233, H236, Q237, K238, N290, R292, V295, R338, E393, T437

CDK2	5MHQ, 4FKL, 4FKU, 4EK4, 4GCJ, 2R3I, 1GZB, 5JQ5, 5K4J, 5IEY, 5ANE, 5D1J, 4D1Z, 4RJ3, 4BZD	I10, K33, F80, **E81**, **L83**, D86, N132, D145
DPP-IV	6B1E, 6B1O, 5T4B, 5T4E, 5I7U, 4PNZ, 4N8D, 4A5S, 2AJC, 5YP1, 5YP2, 5LLS, 5KBY, 5ISM, 4N8E	R125, **E205**, **E206**, F357, R358, Y547, K554, W627, W629, S630, Y631, Y662, Y666, N710, H740
JAK2	6BBV, 5UT1, 5UT2, 5TQ6, 5CF8, 5AEP, 4D0W, 4D0X, 4D1S, 4JIA, 4IVA, 4IVD, 4BBE, 4E6Q, 3E64	K581, Q626, E627, V629, S633, K677, S698, G856, K882, **E930**, **L932**, S936, R980, D994
LTA4H	6ENB, 6ENC, 5NI2, 5NI4, 4RSY, 4RVB, 4R7L, 5AEN, 4L2L, 4MKT, 4MS6, 3FH5, 3FTU, 3FU0, 3FUN	Q136, Y267, G268, G269, E271, E296, **W311**, **F314**, E318, F362, K364, P374, Y378, S379, **Y383**
P38 Kinase	2YIS, 2YIW, 5WJJ, 5XYY, 5N67, 5MTX, 5ML5, 5LAR, 4KIN, 4F9Y, 4DLI, 4EHV, 4E5A, 3ZSG, 4AA0	A34, Y35, K53, R70, E71, L104, T106, H107, **M109**, G110, D112, R149, D168, **F169**, G170, A172, H174, W197, Y200, S251, S252, S254, Y258, L291, S293, D294
Pantothenate Synthetase	4DE5, 4MQ6, 4MUG, 4MUH, 4G5Y, 4FZJ, 4EF6, 4EFK, 4DDK, 4DDM, 3UY4, 3AG6, 3LEB, 3IOB, 3IOE	P38, T39, M40, **H44**, H47, Q61, S65, Y82, Q155, F156, G158, K160, D161, Q164, **V187**, **M195**, S196, S197, R198
PDE4	5K1L, 3LY2, 3D3P, 4KP6, 5OHJ, 3FRG, 3GWT, 3O0J, 3SL3, 3SL5, 3W5E, 3WD9, 4MYQ, 2QYK, 5K32	H160, R163, H234, Y274, D318, F340, **Q369**, N395, Y405, T407, F414, **Q443**, **F446**, E476, Q615, F618

PPAR	3OZ0, 5Y2T, 5GTN, 5JID, 5U3Q, 5U3R, 5U41, 5U42, 5TWO, 4Y29, 4CI5, 3WJ4, 3WJ5, 4F9M, 3TKM	K263, F264, H266, R280, R288, **S289**, **H323**, **Y327**, S342, K367, **H449**, **Y473**
Protein Kinase B	4GV1, 4ZB3, 4ZB4, 3L9M, 2X39, 2VO3, 2UVX, 2UVZ, 2JDO, 1XH6, 1XHB, 1O6K, 1O6L, 2VNW, 2UW0	T51, K72, E121, **A123**, **E127**, E170, N171, D184, E236, D293
UPA	3MHW, 2OBT, 4FUB, 4FUC, 3QN7, 3OX7, 2VIN, 2VIW, 3IG6, 2FD6, 1O5E, 1OSC, 1SQT, 1OWE, 1OWH	D60, H99, **D189**, S190, G193, S195, V213, S214, **G216**, **G219**, P225, V227, R229, G230

For each target, the top 15 highest resolution, unique, holo structures were selected. They were then, through the use of the MAESTRO [232] software package, prepared and the ligand interaction plots generated, an example can be seen in figure 3.6 C. The protein ligand interactions involved in binding were then identified. This was repeated for the available fragment bound structures. It was noted that throughout the data set, for each target, there was a small number of residues, varying from two to five, which appeared to be involved in ligand binding for the great majority of different ligands, 75% of cases or more. These residues were designated as Residues Important for Ligand Binding (RILB). Table 3.3 shows the PDB accession codes of the structures used along with all the residues identified playing a role in ligand binding. The RILB residues are shown in bold.

The following parameters were defined in order to assist in quantitatively measuring the capacity of each probe type to identify protein-ligand interactions:

True Positives (TP) correspond to protein-ligand interactions, with a residue shown in Table 3.3, that are being picked up by a corresponding pharmacophoric probe type.

True Negatives (TN) correspond to protein-ligand interactions, with a residue shown in Table 3.3, that are not picked up by a non corresponding probe type.

False Positives (FP) correspond to protein-ligand interactions, with a residue

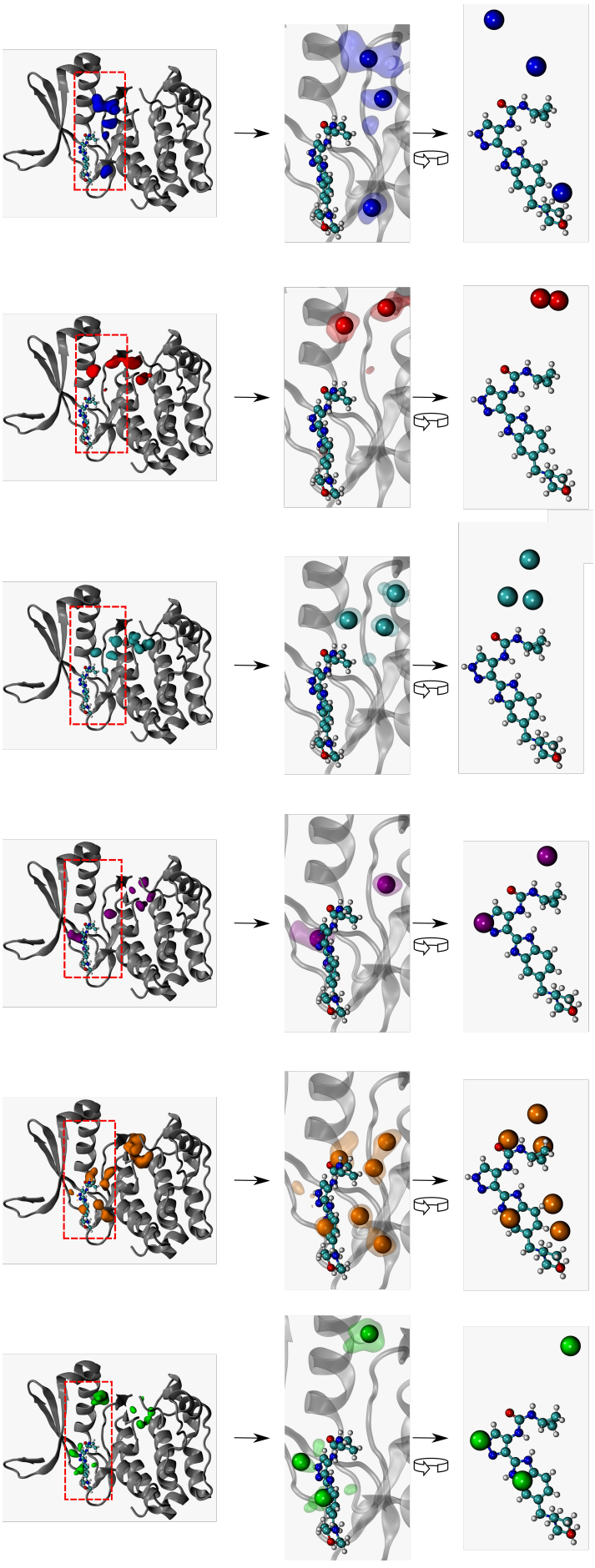


Figure 3.5: A figure of the clustering of the interaction densities into pharmacophoric points. Green points correspond to the pharmacophoric points generated by the hydrophobic interaction densities (shown as solid and semi transparent surfaces), orange to hydrogen bond acceptor densities, purple to hydrogen bond donor densities, cyan to polar densities, blue to positively charged densities and red to negatively charged densities.

shown in Table 3.3, being identified by a non corresponding probe type.

False Negatives (FN) correspond to the protein ligand interactions, with a residue included in Table 3.3, not being identified by a corresponding probe type.

Using the definitions described above, the following metrics were calculated:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

3.2.9 Protein scoring

The 15 holo structures for each target, as well as the fragment bound structures, were then aligned with the apo structure. The protein-ligand interactions were determined using the ligand interaction plots generated previously. The aligned holo structures and the pharmacophoric points were then inspected to determine whether the protein ligand interactions shown in the plots were satisfied by the pharmacophoric points. A protein-ligand interaction was considered as identified if a pharmacophoric point was found within 4 Å of the residue satisfying that interaction. For each interaction identified, a score of 1 was assigned while a score of 0 was assigned for the interactions that were not identified. The total scores, for each target, were then added together and divided by the total number of interactions for each target (eq. 1.6). The resulting score for each target, showing the ability of the probes to recapitulate ligand functional groups and interactions, was tabulated in Table 3.4 .

$$\text{protein score} = \frac{\text{total score of interactions}}{\text{total number of interactions}} \tag{3.6}$$

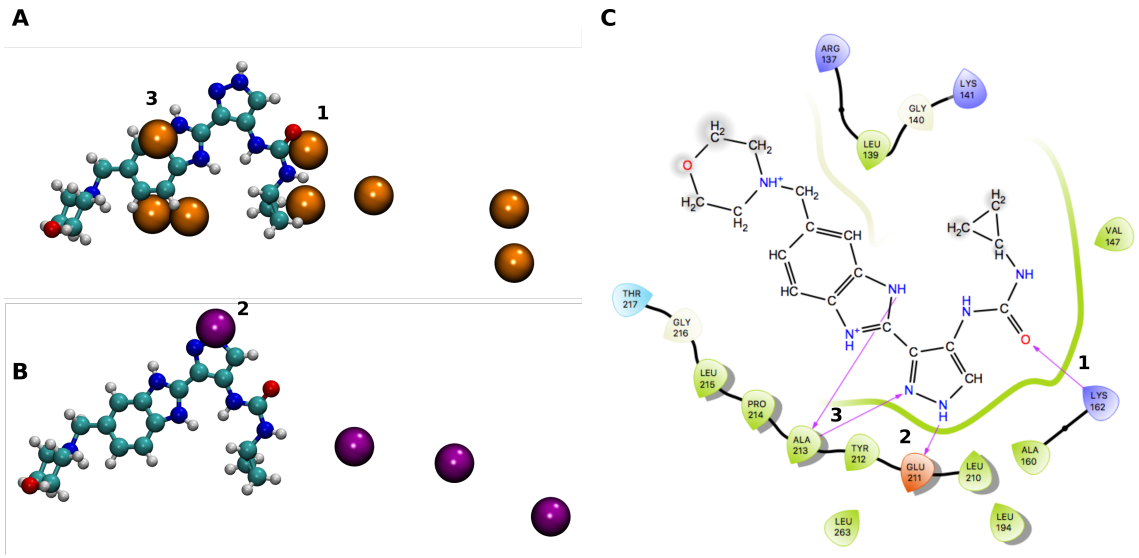


Figure 3.6: A figure of the pharmacophoric points identified in the case of Aurora Kinase A, orange for hydrogen bond acceptors (A) and purple for hydrogen bond donors (B), along with the ligand interaction plot (C). The interactions identified are numbered, along with the pharmacophoric points identifying them. The ligand shown is Pyrazol-4-Yl Urea, a kinase inhibitor.

3.3 Results

3.3.1 Comparing the interactions identified through the use of the pharmacophoric probes with the ligand bound structures

The first step of the analysis was to determine whether the pharmacophoric probes had the ability to predict the protein-ligand interactions which mediate ligand binding. To this end, the targets selected from the Ichihara data set were prepared using the MAESTRO software suite and the ligand interaction plots were generated. These plots were used to identify the interactions which support ligand binding to the target.

The holo targets were then aligned and overlaid with the apo structures used to run the production simulations. The pharmacophoric points were then overlaid on the aligned holo structures and compared with the bound ligands, in order to check if these points can identify the interactions involved in ligand binding. Moreover, it was observed that the pharmacophoric points have the ability to recapitulate the positions of the functional groups of the ligands and the protein residues involved in

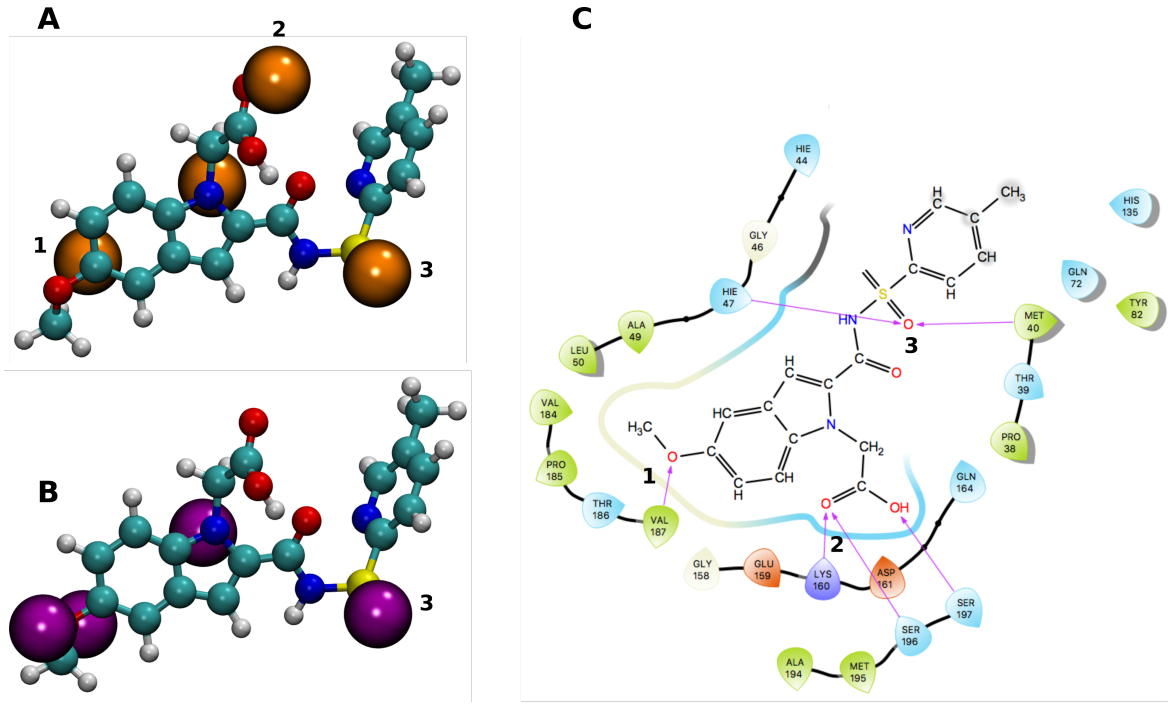


Figure 3.7: As described in figure 3.6, for the case of Pantothenate Synthetase. The ligand shown is (5-methoxy-2-[(5-methylpyridin-2-yl)sulfonyl]carbamoyl-1H-indol-1-yl)acetic acid, an inhibitor.

ligand binding.

Comparing the ligand interaction plots with the spatial arrangement of the pharmacophoric points we can see that in all cases, the moieties shown to interact in the plots are satisfied by either a hydrogen bond donor or acceptor probe, or both. In the case of Aurora Kinase (Fig 3.6), the hinge region interactions which have been known to be important for binding in the Kinase family, marked 1, 2 and 3, are all identified by a hydrogen bond donor or acceptor pharmacophoric points. More specifically, the hydrogen bond interaction with LYS 162, marked as 1, is identified by a hydrogen bond acceptor point while the hydrogen bond interaction of the ligand with GLU 211, marked as 2, is identified by a hydrogen bond donor probe. Finally, the interactions with ALA 213, marked as 3, are identified by a hydrogen bond acceptor point again.

In the case of Pantothenate Synethetase (Fig 3.7), the majority of interactions that are important for ligand binding, marked as 1, 2 and 3, are identified by the pharmacophoric points. The interaction with VAL 187, number 1, is identified by the hydrogen bond acceptor points, as well the hydrogen bond interactions with LYS 160 and SER 196, marked as 2. Finally, interactions with MET 40 and HIS 47,

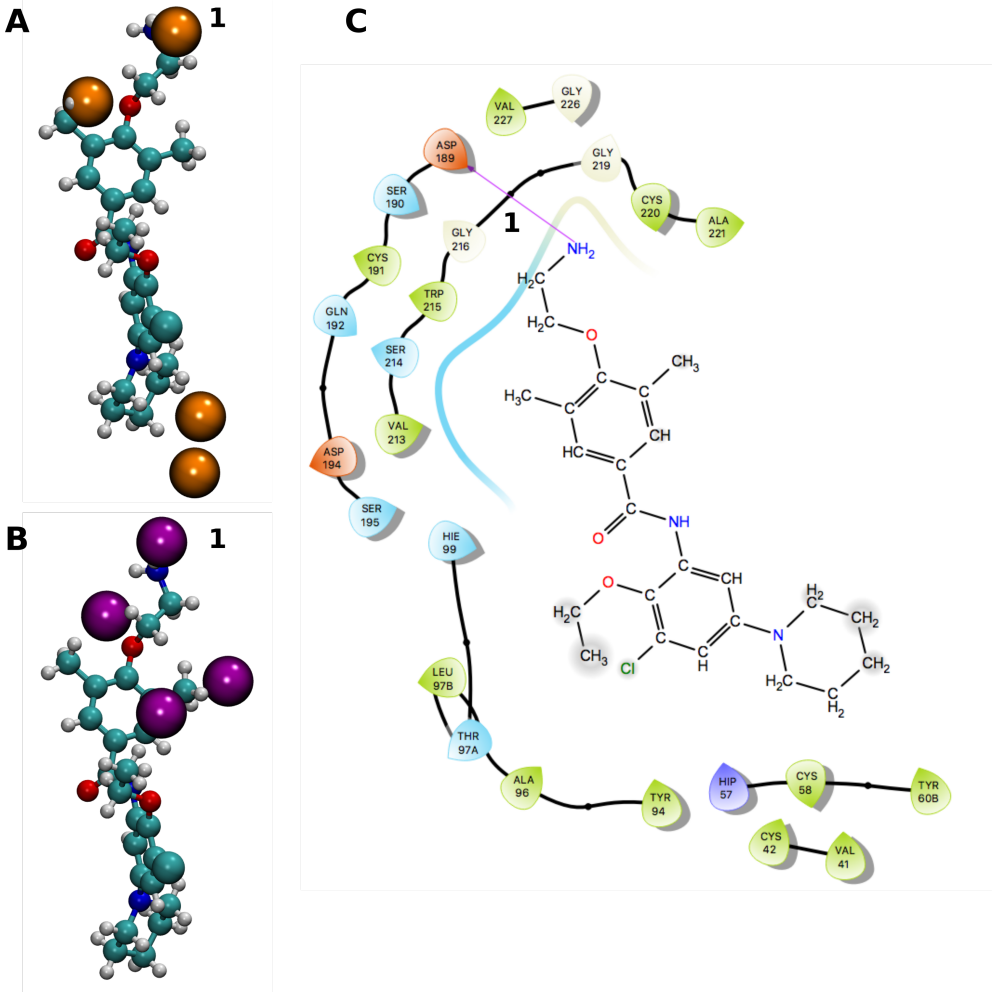


Figure 3.8: As described in figure 3.6, for the case of UPA. The ligand shown is 4-(2-aminoethoxy)-N-(3-chloro-2-ethoxy-5-piperidin-1-ylphenyl)-3,5-dimethylbenzamide, a mexiletine derivative, acting as an inhibitor.

marked as 3, are identified both by hydrogen bond donor and acceptor probes. The one interaction not identified in this case is that with SER 197. This could be due to the fact that SER 196, which is oriented towards the inside of the ligand binding site in order to interact with ligand, is shielding its neighbour, SER 197, from interacting with the pharmacophoric probes.

Finally, for the case of UPA (Fig 3.8), the one interaction important for binding, that with ASP 189, marked as 1, is satisfied by both a hydrogen bond donor and acceptor pharmacophoric point.

One observation from figures 3.6, 3.7 and 3.8 is that the pharmacophoric points acting as hydrogen bond donors, shown in purple, and the pharmacophoric points acting as hydrogen bond acceptors, shown in orange, can sometimes be found in the same region, identifying the same interactions. As mentioned previously, the pharmacophoric probes were adapted from the beads parametrised in the MARTINI forcefield. The hydrogen bond donor and the hydrogen acceptor beads have been parametrised to have very similar properties, interacting in exactly the same way with all of the the other types of beads, with very few exceptions, which explains the identification of the same region by different types of probes. As a result, both types of beads were used as pharmacophoric probes, in order to ensure that all hydrogen bond interactions would be accounted for, with some redundancy for the region which are identified by both types of probes. In fact, as it can be seen from figure 3.6, interactions number one and three are only identified by the hydrogen bond acceptor probes while interaction number two is only identified by the hydrogen bond donor probes. This is also the case for interactions number one and two in figure 3.7, which are only identified by the hydrogen bond acceptor probes. On the other hand, interactions number three (fig 3.7) and number one (figure 3.8) are identified by both types of hydrogen bond probes.

3.3.2 Quantifying the identification of interactions

In an effort to quantify the ability of the pharmacophoric probes to identify the interactions driving ligand binding, the 15 holo structures with the highest resolution in the protein Data Bank, for each test target, were selected, prepared with MAESTRO, the ligand interaction plots generated and the structures inspected to

Table 3.4: A table of the fragment bound targets along with the interaction score calculated for each one. In the case of PDE4, a score of 0 was shown due to the fact that the binding site cavity could not be detected by GHECOM.

Target	Score
Aurora Kinase A	1.0
BACE	0.65
Biotin Carboxylase	1.0
CDK2	0.50
DPP-IV	1.0
JAK2	1.0
LTA4H	1.0
P38 Kinase	1.0
Pantothenate Synthetase	1.0
PDE4	0.0
PPAR	0.50
Protein Kinase B	1.0
UPA	1.0

determine whether the pharmacophoric probes could identify the interactions shown in the plots. If an interaction was correctly identified, a score of 1 was assigned while if an interaction was not or incorrectly identified, a score of 0 was assigned. The total score for each target was then calculated by adding all the scores of its structures and divided by the total number of interactions from the ligand plots (Table 3.5). This was also repeated for all the fragment bound structures available (Table 3.4).

Table 3.4 shows that for the majority of the fragment bound targets, a score of 1.0 is achieved, meaning that all the interactions fulfilled by fragments, are also identified by the pharmacophoric points. Moreover, table 3.5 shows that for the majority of the lead bound targets, a score of 0.65 or higher is achieved, with some cases reaching 0.94. This corresponds to the probes being able to identify 65% of the lead compounds binding interactions or more. Fragments have been shown to bind to areas that can be defined as hotspots [88], leading to the conclusion that the pharmacophoric probes have the ability to identify hotspot interactions. Furthermore, lead compounds are usually based on the extension or linking of bound fragments, meaning that this approach has the potential of identifying regions where bound fragments can be extended.

There are three cases which stand out. Firstly, in the case of PDE4, a score of

Table 3.5: A table showing the ligand bound targets along with the interaction score calculated for each one. In the case of PDE4, a score of 0 was shown due to the fact that the binding site cavity could not be detected by GHECOM.

Target	Score
Aurora Kinase A	0.94
BACE	0.45
Biotin Carboxylase	0.70
CDK2	0.40
DPP-IV	0.82
JAK2	0.90
LTA4H	0.57
P38 Kinase	0.60
Pantothenate Synthetase	0.80
PDE4	0.0
PPAR	0.53
Protein Kinase B	0.65
UPA	0.85

0 is assigned. This is due to the fact that the binding cavity of the ligands could not be identified by the cavity detection software, resulting in no interactions being identified. The second case is CDK2, were a score of 0.40 is observed, much lower compared to the other kinases. This was observed due to the fact that the hinge regions residues which are important for ligand binding, in this case GLU 211 and LEU 210, are surrounded by phenylalanine and tyrosine residues whose bulky rings shield the backbones and side chains of GLU 211 and LEU 210 respectively, during the simulation, preventing their identification by the hydrogen bond donor or acceptor probes. Finally, in the case of BACE, the lower interaction identification score can be explained by the fact that a lot of the residues involved in ligand binding are found on loop of the flap that encompasses the binding site. As a result, given the fast moving nature of the probes, which is characterised by a lot of transient interactions with the protein rather than continuous binding, and the flexibility of loops, it is difficult for the probes to bind the region with enough frequency to identify the binding interactions. A lower threshold for filtering the interaction densities could be used which would allow for detection of these flexible regions but would come with an increased noise, making it harder to differentiate between true and false positives.

3.3.3 Identification of residues involved in binding

It was shown in the previous section that the pharmacophoric points have the ability to identify residues and interactions involved in ligand binding. Figures 3.6, 3.7 and 3.8 show all the pharmacophoric points generated for Aurora Kinase, Pantothenate Synthetase and Urokinase Plasminogen Activator. As it was seen, some of them were involved in the binding of the ligands as detected. However, there are several pharmacophoric points identified than the ones fulfilling the interactions shown in the plots. In an effort to assess the specificity of the pharmacophoric probes, the residues identified as being in contact with each pharmacophoric point were elucidated.

Using the interaction plots generated from the holo structures previously, and selecting the protein residues which are seen to participate in these interactions, the residues that play a role in ligand binding were identified. There were a few key residues which seemed to be involved in ligand binding in 75%, or more, of the cases. These were designated as residues important for ligand binding (RILB) (Table 3.3).

The pharmacophoric points and residues identified, in the case of Aurora Kinase A, can be seen in figure 3.9. LYS 162, ALA 213, THR 217, ASP 256, GLU 260, ASN 261 and THR 292 are identified by the hydrogen bond acceptor probes while GLU 211, ASP 256, ASN 261 and GLY 291 are also identified by the hydrogen bond donor points.

For the case of Pantothenate Synthetase (figure 3.10) HIS 47, ASP 161, VAL 187 and SER 197 are identified by hydrogen bond acceptor probes with the same residues, excepto for VAL 184 instead of SER 197, being identified by the hydrogen bond donor probes.

In the case of Urokinase Plasminogen Receptor (figure 3.11) a larger number of residues is identified. More specifically, CYS 58, ASP 60, ASP 189 and GLY 219 are identified by hydrogen bond acceptor probes while SER 214, GLY 216, GLY 219 and PRO 225 are identified by the hydrogen bond donor probes.

Looking at figures 3.9, 3.10 and 3.11 it can be seen that the pharmacophoric probes and points have the capability to correctly identify the protein residues which are involved in ligand binding, coloured spheres, and more importantly, the RILB residues, coloured spheres with a shadow around them. In all three cases all the residues identified participate in ligand binding with the only exception

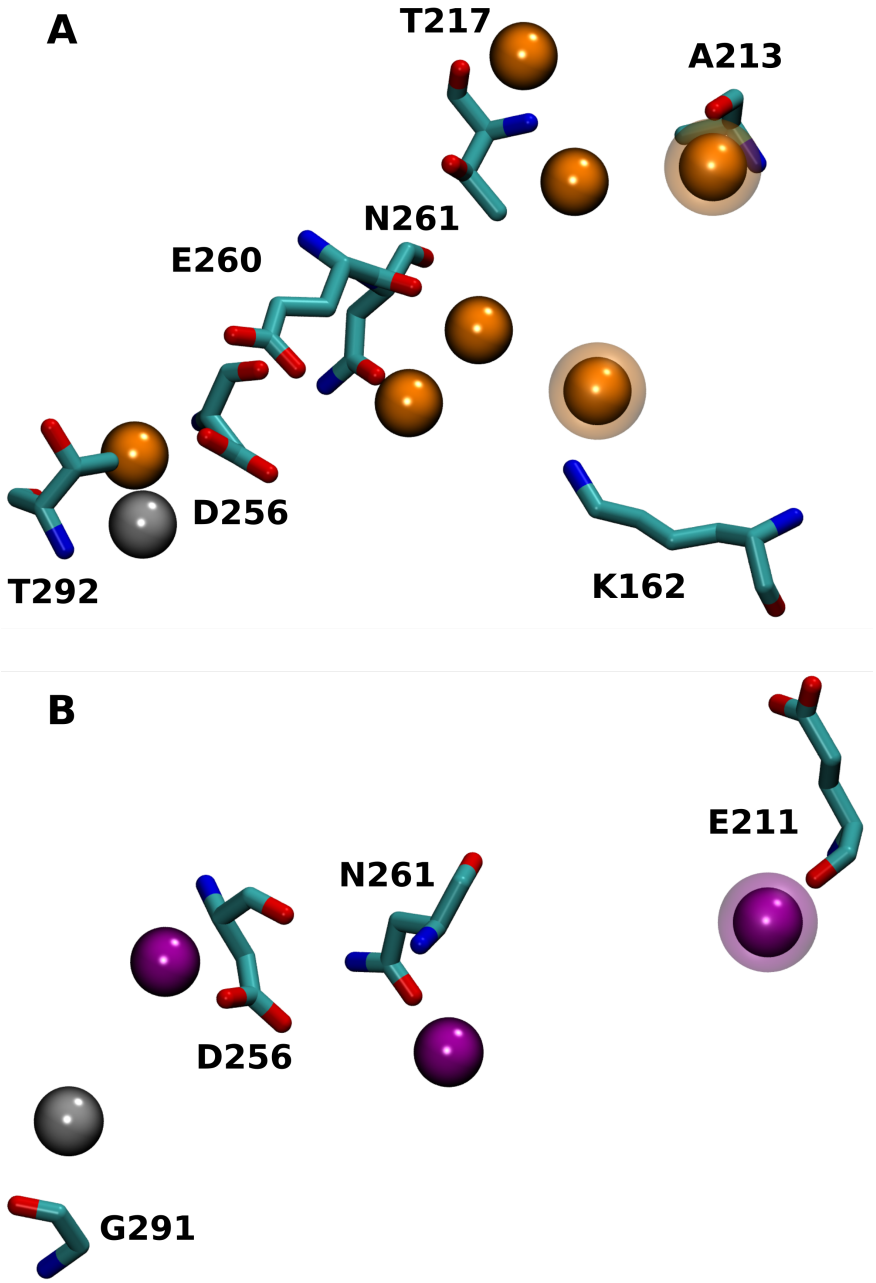


Figure 3.9: A figure of the pharmacophoric points identified in the case of Aurora Kinase A, orange for hydrogen bond acceptors (A) and purple for hydrogen bond donors (B), along with the residues being in contact with them. Residues shown with a shadow around them are residues shown to be important for ligand binding (RILB) while those shown in grey are identified as false positives. Those coloured are residues shown to participate in binding.

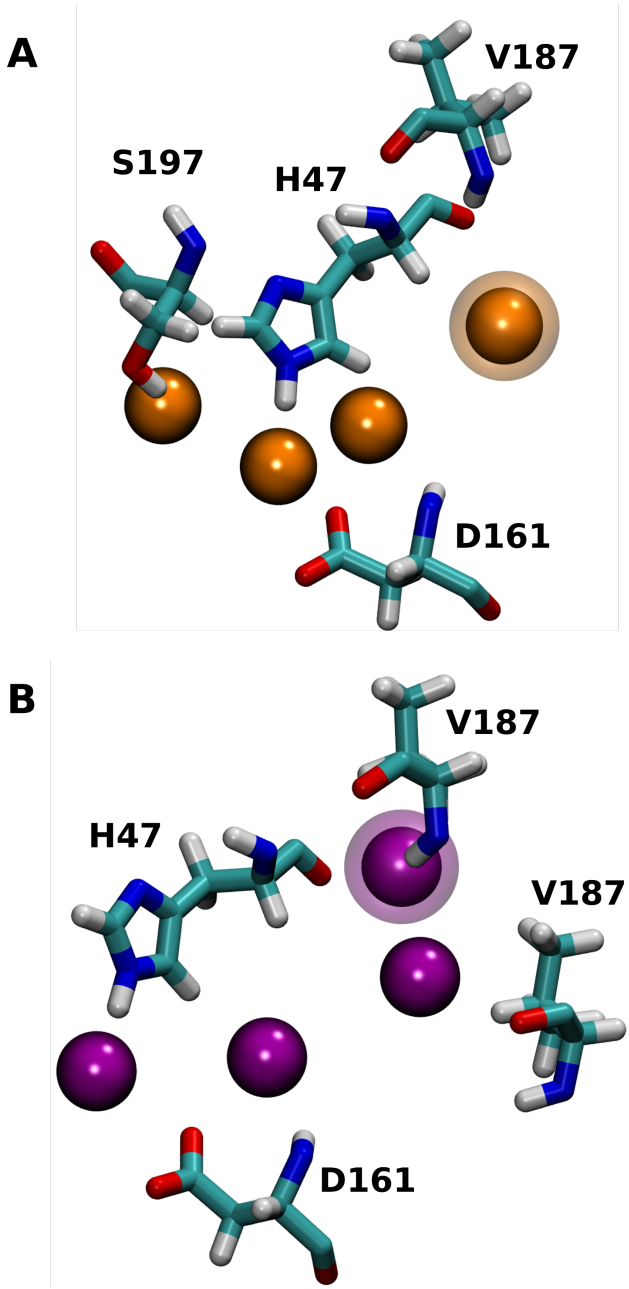


Figure 3.10: Same as figure 3.9 for the case of Pantothenate Synthetase.

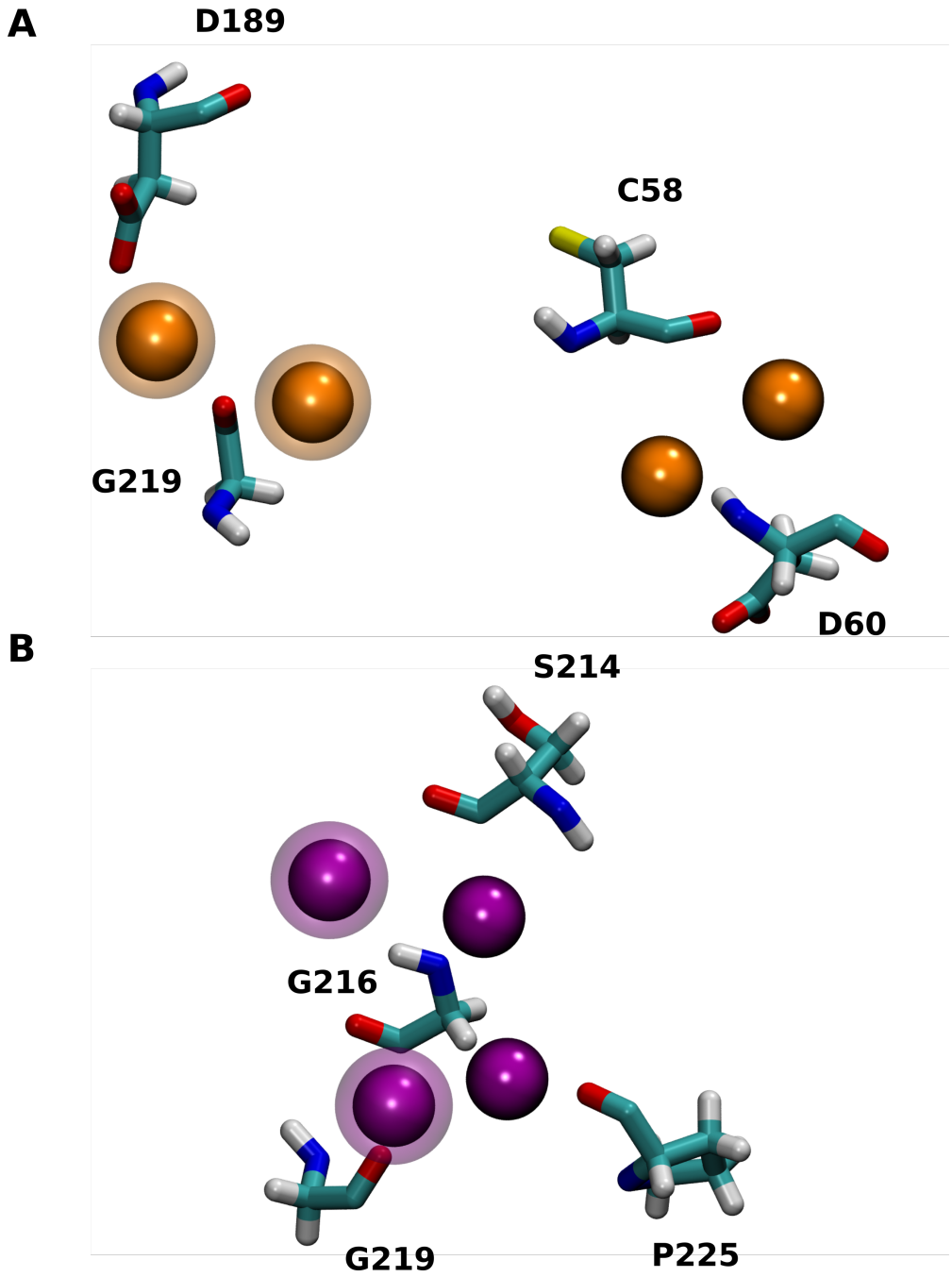


Figure 3.11: Same as figure 3.9 for the case of UPA.

being GLY 291 and THR 292 in the case of Aurora Kinase (figure 3.9), shown in grey. These residues are located on the activation loop of Aurora Kinase, nearby a phosphorylation activation site, which could explain their identification by the pharmacophoric probes.

The rest of the residues of Aurora Kinase identified by the hydrogen bond probes correspond to structurally important regions or regions already known to play a role in binding ligands. LYS 162 along with GLU 211, ALA 213 and THR 217 form part of the ADP binding site, with GLU 211 and ALA 213 being RILB, while ASP 256, GLU 260 and ASN 261 are located within the catalytic active site of the protein. In particular, the role of ASP 256 is to transfer is to transfer the γ-phosphate of ATP to the hydroxyl group of a substrate serine or threonine residue.

In the case of Pantothenate Synthetase, all the residues identified by the hydrogen bond probes have been shown to form part of the active site of the protein. HIS 47 has been seen to play a role in directly binding and stabilising ligands, therefore being extremely well conserved, while ASP 161 and Serine 197 have been shown to participate in the formation of hydrogen bond networks with water molecules that, along with magnesium ions, are also used to stabilise ATP. VAL 184 and VAL 187 can also be seen to bind the tail of ATP in order to orient the ligand towards the right conformation.

Finally, in the case of UPA, ASP 189, SER 214, GLY 216 and GLY 219 have been shown to play a role in binding ligands, either directly (ASP 189, GLY 216, GLY 219) or through the formation of water bridges (SER 214). ASP 189 is located at the base of the S1 pocket while SER 214, GLY 216 and GLY 219 can be found at the south. PRO 225 is also located at the back of the same pocket, explaining why it has been identified as a potential interaction forming residue. Finally, CYS 58 and ASP 60 are located next to or very nearby HIS 57 which forms part of the triad of catalytic residues, together with ASP 102 and SER 195. As a result, they also have the potential to form ligand binding interactions.

Table 3.6: A table showing the the sensitivity, precision, specificity and accuracy values calculated for each type of pharmacophoric probe. The same is shown in the parentheses values but when the RILB are used to calculate these values.

	Sensitivity	**Precision**	**Specificity**	**Accuracy**
Hydrophobic	0.5 (0.6)	0.4 (0.75)	0.79 (0.91)	0.72 (0.82)
H-Bond Acceptors	0.40 (0.51)	0.97 (0.94)	0.95 (0.96)	0.51 (0.65)
H-Bond Donors	0.41 (0.5)	0.97 (0.97)	0.95 (0.95)	0.52 (0.59)
Polar	0.37 (0.45)	0.94 (0.96)	0.88 (0.92)	0.5 (0.56)
Positive Charge	0.45	0.86	0.83	0.55
Negative Charge	0.41	0.93	0.91	0.56
Overall	0.42 (0.49)	0.85 (0.91)	0.89 (0.91)	0.56 (0.63)

3.3.4 Quantitative analysis of the predictive powers of the probes

In an effort to quantitatively evaluate the predictive powers of the pharmacophoric probes, the metrics mentioned in equations two to five were used. TP were defined as protein-ligand interactions that are identified by the corresponding probe type while FP correspond to protein-ligand interactions being identified by a non corresponding probe type. In this context, precision can be described as the measurement of the ability of the probes to identify protein-ligand interactions of the correct type. With FN being defined as the protein-ligand interactions which are not reproduced by a corresponding probe type, sensitivity can be described as the measurement of the ability of the probes to correctly sample the observed protein-ligand interactions. TN are being defined as the protein-ligand interactions which are not reproduced by the wrong probe type, therefore specificity can be described as the capability of each probe type to avoid detecting protein-ligand interactions of a non-corresponding type. Finally, accuracy can be described as the summary measurement of the correlation of positive results (TP and TN) within the context of all the the results identified (TP, TN, FP and FN).

Using the data from 3.3, on the ligand binding residues, and the residues identified as being in contact with the pharmacophoric points, the sensitivity, precision, specificity and accuracy values for each type of probe were calculated (Table 3.6). As it can be seen from Table , the majority of the values is around 0.5 or higher,

with a few exceptions, while the precision and specificity for each type of probe, except for the hydrophobic ones, is well above 0.80. These findings indicate that most types of probes exhibit a good level of accuracy. The extremely high values of precision and specificity for most types of probes indicate that the probes have a very high capacity for identifying interactions of the correct type with a very low number of false positives. The lower values of the hydrophobic probes, mainly due to the high number of false positives, can be expected due to their promiscuous nature. During the course of the simulation, they can be seen to penetrate and explore every potential orifice on the surface of the protein (figure 3.1, panel C). Furthermore, although Maestro is very capable in identifying hydrogen bond and ionic interactions, its ability to identify hydrophobic interactions, the main driving forces behind ligand binding, is somewhat limited. This would result in an increased number of false positives, in the case of hydrophobic probes, which in turn would result in lower sensitivity and accuracy values. The lower sensitivity values of the probes can also be explained by the high number of potential interaction forming residues identified in 3.3 for each protein target. As it is expected, not all of these residues can be identified which leads to a higher number of FN, resulting in the lower sensitivity values.

The values in Table 3.6 were calculated using all the identified residues from Table 3.3. However, as shown from our ligand interaction plots previously and from others [88], there are certain residues which contribute more to ligand binding, compared to others, designated as RILB or "hotspots". When the above analysis was repeated but only taking the RILB residues into account, we can see an increase in the sensitivity and accuracy values of all types of probes, with most of them being between 0.6 and 0.8, compared to between 0.3 to 0.6 when taking all of the interaction forming residues into account (Table 3.6, values shown in parentheses). The hydrophobic probes exhibit the highest increase in their precision and specificity values, due to the decrease in the numbers of FP.

The analysis including the RILB residues could not be performed for the positive and negative charge probes due to the low number of RILB residues forming ionic interactions, meaning no statistic parameters could be derived. It should also be mentioned that ionic probes have the capability to predict hydrogen bond interactions,

to a much lower degree than the hydrogen bond donor and acceptor probes though, and that polar probes also have the ability to detect ionic interactions. However, the presence of the hydrogen bond donor, acceptor and charged probes, makes these capabilities redundant. Overall, it can be concluded that the probes have a high capacity for identifying residues important for ligand binding on a target of interest.

3.4 Discussion

As it was mentioned before, the increasing number of available target structures and the popularity of SB drug design, has resulted in the development of a variety of mixed solvent MD based methods, with some excellent work up to date. However, looking through the literature, we have been hard pressed to find a study were a thorough quantitative analysis of the probes predicting capabilities was performed, with the exception of Arcon et al. As a result, performing a direct comparison with other methods was difficult. When possible, it was seen that the hydrogen bond donor and acceptor probes offer a significantly increased predictive power, while a small but still significant increase in the predictive power of the positive and negative charge probes is also observed. The predictive power of the hydrophobic probes is comparable, if somewhat lower, to what was observed in other studies, mainly due to the promiscuous nature of the probes, the coarse grained approach and the Maestro limitations. The most important advantage provided by the use of the CG MD approach, however, is its very low demand for computational resources, reaching time-scales of µs, allowing the exploration of a much larger conformational space.

3.4.1 Reasons for probe failure

As it was already seen previously, there were certain cases where this method results in a wrong prediction. An incorrect prediction, or no prediction at all, can be attributed to the following reasons.

First, this method is dependent on the use of the GHECOM cavity finding software in order to identify potential binding sites, which will then be used to filter the interaction densities, allowing to focus on the areas of interest. In the case of the NS5 polymerase structure used, the ligand was bound in an allosteric binding

site. The allosteric binding site is too shallow to be identified by the cavity detection software. As a result, the correct binding site was not identified and no prediction was made. This is the same reason why no prediction could be made in the case of PDE4.

Second, due to the coarse graining of the protein by MARTINI, the amino acids are represented as larger beads, which usually consist of four heavy atoms, condensed into one bigger bead. This can lead to smaller amino acids, such as GLU or LEU being coarse grained in two bead amino acids, making them smaller compared to their ring bearing neighbours. Residues such as TRP or PHE or TYR, will be coarse grained in a bigger entity due to the fact that the MARTINI forcefield uses a smaller size bead but a lower mapping of three or two heavy atoms into one bead, to create rings. In cases such as that of Cyclin Dependent Kinase 2 (CDK2), mentioned previously, GLU 81 and LEU 83, both RILB, are found next to PHE 80 and PHE 82 respectively. This leads to the hydrophobic rings shielding the backbone of GLU 81 or LEU 83, which form the hydrogen bonds important for ligand binding, preventing the hydrogen bond donor or acceptor or polar probes from interacting with them. As a result, very low interaction densities are detected at these points, which are then filtered out due to their resulting ΔG_i value.

Third, although the pharmacophoric probes have been modified to have a smaller size compared to the beads making up the protein residues, when interacting with other probes, they still interact as beads of a normal size when interacting with the protein beads. As a result, due to the large size of the normal beads, two probes will not be able to interact simultaneously with the same residue. Furthermore, in cases such as CDK2, where both hydrogen bonds and hydrophobic interactions are required for ligand binding, the resulting densities and pharmacophoric points end up overlapping, leading to, if not necessarily wrong, difficult to interpret predictions. Moreover, as mentioned above, the lower prediction capabilities of Maestro, when it comes to hydrophobic interactions, would also affect the values observed.

Finally, it is known that water plays a very important role in a lot of protein-ligand interactions, offer acting as a bridge for the formation of stabilising hydrogen bond networks. The initial computational methods used in CADD, struggled with taking into account the effects of water, leading to the development of specialised tools, such

as WaterMap. This is particularly true for this method due to the representation of water in the MARTINI forcefield, where four water atoms are condensed into one bigger water bead. As a result, water beads will suffer from all the limitations mentioned in the previous paragraph, making it even more difficult to correctly evaluate the effect of water in ligand binding.

3.4.2 Advantages and overcoming the limitations

When it comes to the limitations of this method, there are certain steps that can be taken in order to overcome or minimise them. First, regarding the failure to detect the cavity of a potential binding site, different cavity detection software could be used to detect a cavity of interest. Second, in order to deal with the probe beads - protein beads interaction issues and their size, the parametrisation of the beads could be modified to alter their interactions to the desired size. Afterwards, two, or more, beads bonded together, could be used as pharmacophoric probes to explore the directionality of the identified interactions, which is missing from the current approach. Third, in regards to the role of water in protein ligand interactions, an improved water model would be the only possible solution. Although a polarisable water model is currently available for the MARTINI forcefield, it still suffers from the limitation of the water beads being too big to adequately penetrate the cavities of the potential binding sites. Finally, in order to overcome the limitation of the Maestro hydrophobic interactions, other software suites or methods could be used to detect these interactions, which can then be exploited to assess the capabilities of this method.

The pharmacophoric points identified have the ability to recapitulate fragment and ligand functional groups and interactions important for binding. Being able to identify fragment interactions provides a quick and accurate way for identifying "hotspot" interactions, while accurately identifying ligand binding interactions allows for the identifications of regions towards which fragments can be grown or linked. Both of these features are of paramount importance in drug design. Furthermore, these features also allow the use of the pharmacophoric points as potential pharmacophoric features in the creation of a pharmacophore model, which can then be used to identify new leads for drug discovery. Finally, the speed of this method, with the simulations

taking approximately 12 hours on a workstation, allows for the testing of a very large number of different targets, or different conformations of the same target, in a very short amount of time. All of the above offer the ability to increase the speed of the initial phases of a new drug design project, along with reducing the cost, due to the lower demand for computational resources.

Another advantage of this approach is the low number of false positives. As described in the introduction section of this thesis, a limitation from which a lot of similar methods suffer, and which can impeded drug design efforts, is the identification of "spurious" interactions that likely correspond to false positives. The low number of false positives, as shown by the high number of the precision and specificity values for all types of probes, with the exception of the hydrophobic ones, allows for drug design efforts to proceed with a greater degree of confidence in the results. Moreover, when the overall values for the metrics shown in Table 3.6, are compared to those reported by Arcon et al. [144], similar numbers for sensitivity and accuracy are observed with this method showing values for precision and specificity which are 40% and 30% higher, respectively.

One final advantage of this method is its ability to be re-purposed to fulfil more roles. The potential of its application on membrane proteins is evident and, moreover, the hydrophobic and charged probes offer an excellent tool for the discovery of lipid binding sites on targets of interest. Currently, long coarse grained or all atom simulations are used to identify lipid binding sites on proteins. This method could offer an initial set of potential sites which could then be further explored.

3.5 Conclusions

The use of CG MD simulations with MARTINI beads as pharmacophoric probes, on a large data set of targets, and their ability to correctly identify protein-ligand binding interactions has been demonstrated. Moreover, we have quantitatively evaluated the predictive power of the pharmacophoric probes, with most of them having values of 0.6 or higher, similar or higher to other methods published in the literature. Finally, we have shown that the probes and the pharmacophoric points have the ability to identify the residues which are important for ligand binding, with a low number of

false positives, both for fragment bound and lead bound targets. This allows for the use of this approach in the identification of both "hotspot" (fragment binding) interactions and regions where these fragments can be expanded or linked, in binding sites of interest.

Chapter 4

Dynamic Pharmacophore Probes: Application to GPCRs

Chapter Outline

In this chapter the dynamic pharmacophore probes approach will be applied to a test set of GPCRs. These targets were selected as a more challenging data set due to their more buried and narrow orthosteric binding sites, compared to the targets used in the previous chapter, and the existence of potential allosteric binding sites. The same simulation and analysis protocols, as the previous chapter, were followed. The analysis was split based on orthosteric and allosteric binding sites on the targets examined.

4.1 Introduction

Membrane proteins mediate the transduction of a wide range of extracellular stimuli into an even wider range of signalling cascades, acting as an interface between the external environment and the cell. The extraordinary diversity of membrane protein allows them to be sensitive to nearly every environmental stimulus possible, such as light, pressure, pH and many others [233].

It is no surprise, therefore, that membrane proteins constitute drug targets of exceptional importance. Amongst them, a family of particular interest are the G-Protein coupled receptors (GPCRs) which take their name by their ability to bind

to, recruit and modulate the activity of the intracellular heterotrimeric G-proteins. They consist of more than 800 members, making them the largest super family of trans-membrane receptors, involved in many aspects of cell signalling [234]. As a result, signal transduction by GPCRs is fundamental in many physiological processes, from taste and smell to endocrine regulation, making them the target of more than 30% of all therapeutic drugs. Due to these properties, and the wide range of known ligands, they were selected as the second class of test cases on which the dynamic pharmacophore probes approach was tested on [235].

4.1.1 GPCR families and structure

There are over 800 distinct genes encoding GPCR in humans and they are classified based on phylogenetic criteria. GPCRs are categorised into six classes, A, B, C, D, E and F, based on the classification system introduced by Kolakowski in 1994 [236]. It should be noted that the D and E classes are only found in invertebrates. This system was revised in 2003, by Fredriksson and colleagues, with the main update being the splitting of class B into different classes, resulting into five main GPCR classes [237]. These are: Rhodopsin (class A), Adhesion (class B), Secretin (class B), Glutamate (class C) and Frizzled/Taste-2 (class F). Class A is the sub-family that includes, by far, the highest number of members, compared to the other classes. At the most basic level, all GPCRs share certain common structure characteristics, consisting of seven trans-membrane a-helical segments which are connected to each other by alternating intracellular and extracellular loop regions [234] (figure 4.1A).

4.1.2 Rhodopsin like receptors (class A)

As mentioned above, class A, is by far the largest of all five classes, with more than 700 receptors [234]. It can be further divided into four phylogenetic groups, α, β, γ and δ. Although there is no specificity with regards to the receptor and the endogenous ligand, meaning the same type of ligands bind to different subfamilies, it has also been seen that some of the subfamilies have been shown to bind ligands of the same type. Subgroup α, for instance, consists mainly of amine binding receptors while all the β subgroup members bind peptides [237].

Class A also includes the majority of current drug targets [234]. The histamine,

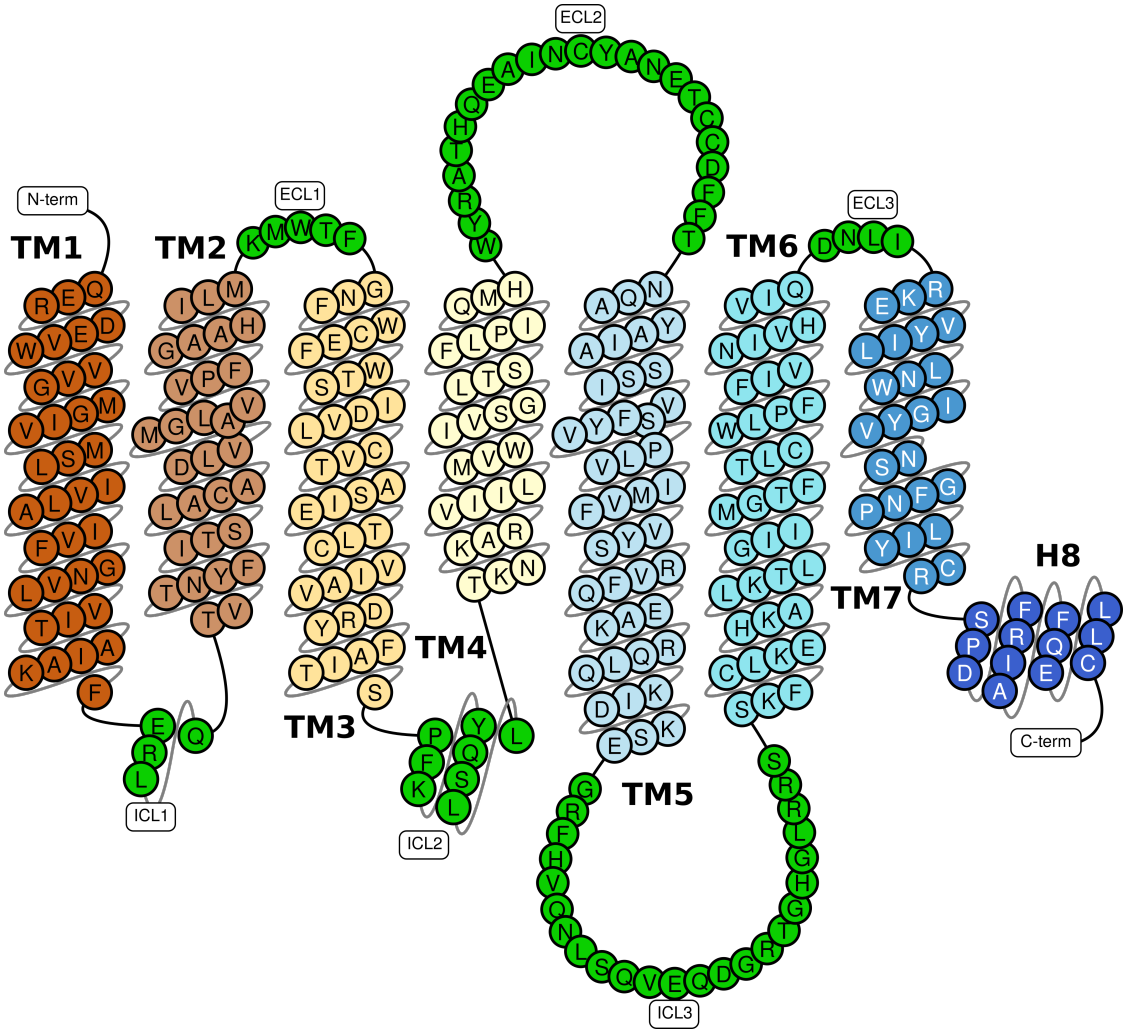


Figure 4.1: Schematic of a typical GPCR showing the seven TM helices, in different colours, connected by the intracellular (ICL) and extracellular loops (ECL), through the membrane and the amphipathic helix eight (H8).

dopamine, adrenoceptor and aminergic receptos are all included in the α subgroup while the β subgroup receptors are targets for a variety of hormones, such as oxytocin, gonadotropin-releasing hormone and endothelin. The γ subgroup has the opioid, somatostatin and angiotensin receptors. Finally, the purinergic, protease activated, glycoprotein binding and olfactory receptors are contained within the δ subgroup [234].

It is also characterised by the presence of highly conserved orthosteric binding site, located in the extracellular, trans-membrane domains (figure 4.3B). Moreover, a highly conserved Na^+ binding pocket is located underneath the orthosteric binding site of class A structures [234].

4.1.3 Secretin and Adhesion receptors (class B)

Class B contains a total of 48 members, 15 of them coming from the secretin family and 33 of them coming from the adhesion family. The secretin family receptors are predominantly involved in homeostatic functions while the adhesion family receptors are mediate the cytoskeletal formation and organisation. As a result, the adhesion family members are responsible for the architecture of the cell, such as the formation of dendrites in neuronal cells [238].

4.1.4 Glutamate receptors (class C)

Class C is composed of the metabotropic glutamate receptors, gamma-aminobutyric acid receptors, calcium sensing receptors, sweet and umami taste receptors. It numbers 22 members in total. Their mechanism of action is similar to the one observed in the venus fly trap, where glutamate is captured in the cavity formed between the two lobes of the extracellular N-terminal domain [239].

4.1.5 Frizzled and Taste-2 receptors (class F)

Class F consists of two separate groups, the Frizzled and Taste-2 receptors. The Frizzled receptors are responsible for mediating signals from the secreted Wnt glycoproteins and are involved in embryonic and tissue development and cell polarity [240]. Their function has been linked to the formation of neuronal circuits, such

as the formation of neuronal synapses and the development of dendrites and axons [241].

4.1.6 GPCR signalling

The original GPCR activation model, according to which GPCRs activate the heterotrimeric G-proteins and they are themselves down regulated by β-arrestin, is oversimplified with GPCR signalling having shown to be extremely complex. It has recently been observed that individual ligands can bias the signalling cascade taken by the receptor [242].

The activation of GPCRs thus triggers a varied collection of signal cascades which can proceed down either a G-protein-dependent or a G-protein-independent route. In turn, each of these activated signal patterns modulate a range of downstream processes, based on their intracellular binding targets [243]. These downstream processes include phospholipases A2 and C, phosphodiesterases, adenylyl and guanylyl cyclases and various ion channels. These then activate an even greater variety of secondary messengers like cyclic adenosine monophosphate (cAMP), cyclic guanosine monophosphate (cGMP), diacylglycerol (DAG), inositol (1,4,5)triphosphate, phosphatidylinositol(3,4,5)triphosphate, arachidonic acid and phospatidic acid [243, 244]. They can also affect the intracellular concentration of ions such as K^+, Na^+ and Ca^{2+} (figure 4.2) [243].

From all the above, the importance of being able to successfully drug GPCRs is quite evident. However, there is still a plethora of challenges that need to be overcome.

4.1.7 Drugging GPCRs

Computational methods have been used to model a wide aspect of GPCRs ranging from their active states to the activation pathway, conformational changes, dimerization, hydration, the ionic lock, ligand binding, protonation, and sodium binding. These efforts have been extensively covered by Taddese et al. [245]. When it comes to GPCR drug design, the same drug design approaches mentioned in the previous chapters have also been used to design new compounds and pharmaceuticals that would successfully modulate these receptors [246, 247, 248]. However, due to their

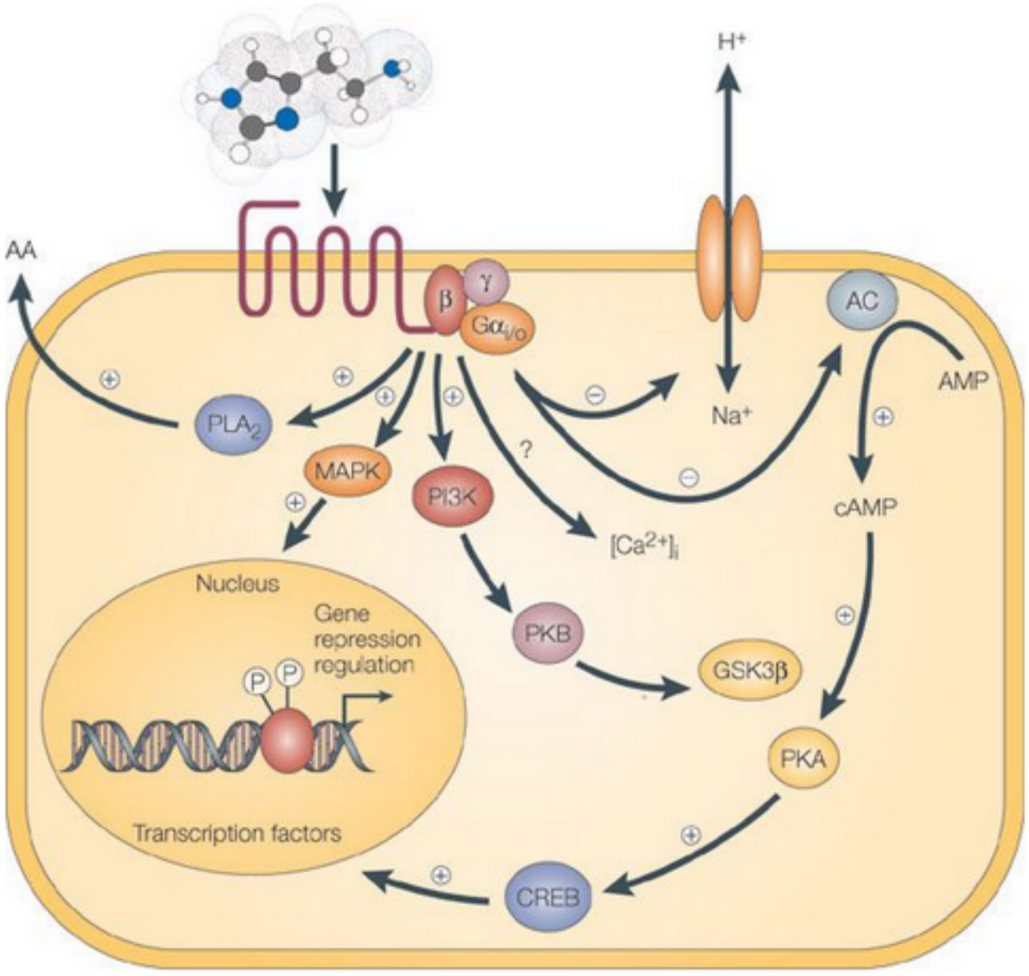


Figure 4.2: A schematic of the different signalling cascades that are controlled by GPCRs and G-proteins. Figure reproduced from [243]

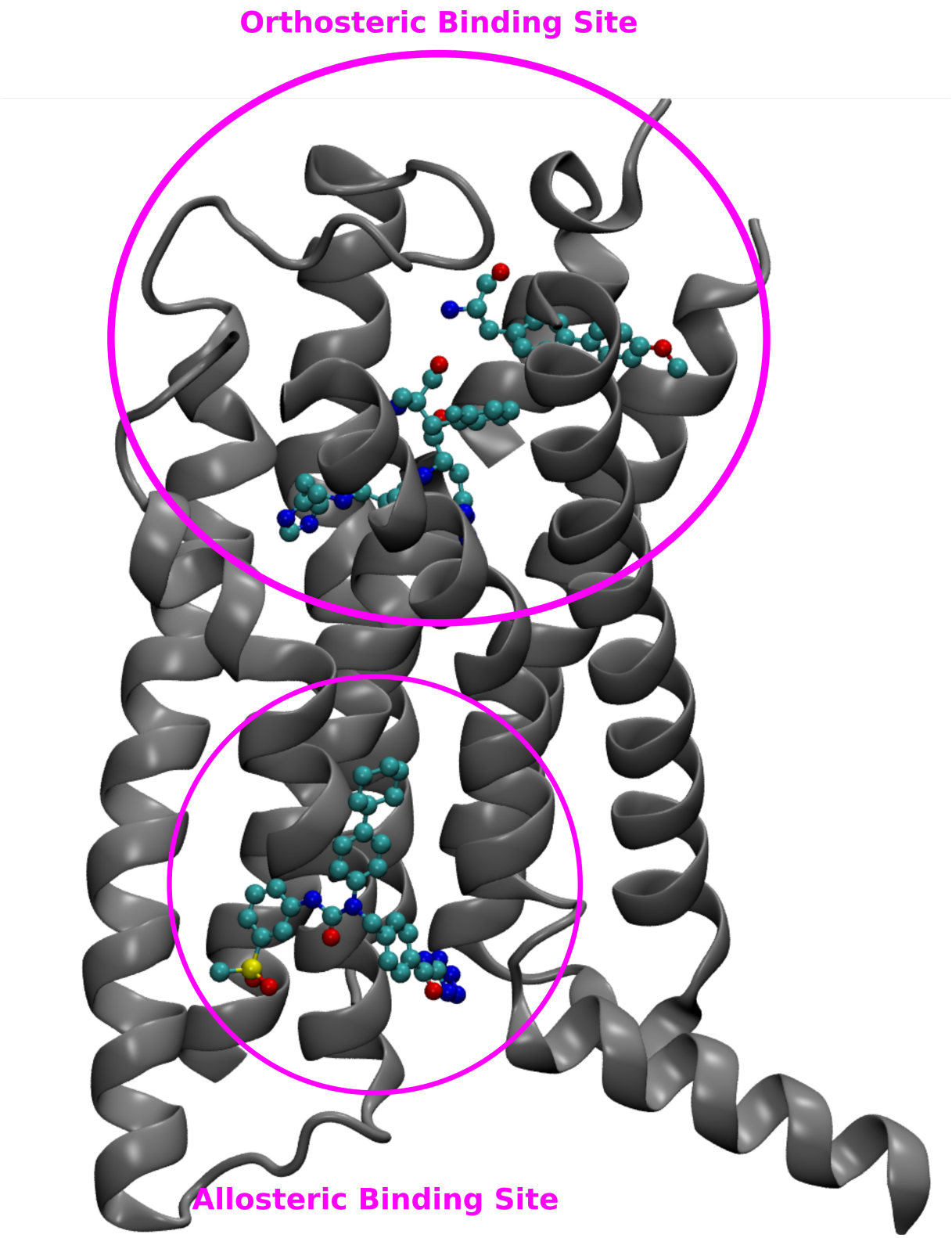


Figure 4.3: An example of a class B GPCR (GLP1), with ligands bound both in the orthosteric and allosteric binding site.

nature, a great number of caveats had to be overcome, and still need to be tackled, to achieve this [249].

Structure based drug design approaches had, until the last decade, been hampered by the lack of available GPCR structures [246, 247]. This was due to these receptors being very hard to crystallize. With the development of new crystallization techniques, better data collection, through the use of X-ray free electron lasers, and, more recently, cryo electron microscopy, more structures have been elucidated and the efforts to understand and target GPCRs have been intensified [250].

The orthosteric binding site of a protein is defined as the canonical binding site of the recognised natural ligand of the receptor. Although well conserved within each class, its location can greatly vary between classes, having been localised in different positions within the TM bundle, within the extracellular domain (class A) or in the interface of the two [251, 252]. Although this might allow for a certain degree of selectivity between classes, each ligand requiring different properties to access each location, it leaves no space for intraclass selectivity and very little that can be done to avoid unwanted side effects. Furthermore, in the case of class B receptors, where the endogenous ligands are comprised of peptides, small molecules cannot bind to the flat, large orthosteric binding sites (figure 4.3).

A potential way of avoiding these drawbacks is the targeting of allosteric binding sites on GPCRs. Allosteric binding sites are spatially distinct from the orthosteric ligand binding sites, even allowing for the simultaneous occupation of both sites by endogenous and allosteric ligands (figure 4.3) [253, 254, 255]. They can be located anywhere on the protein ranging from the intracellular region, between the trans-membrane helices and all the way to the extracellular domain, above the orthosteric binding site. They offer the advantage of greater selectivity by the reduced conservation of allosteric binding sites, compared to the orthosteric ones, along with offering an alternative binding such as the cases where the orthosteric one is not tractable (class B GPCRs) [253, 254, 255]. Furthermore, there are cases where the effect of an allosteric modulator is only exterted in the presence of the endogenous ligand, known as the probe dependence effect, which allows for the effects of the endogenous ligand without the accompanying side effects [253, 254, 255].

One final advantage of the use of an allosteric binding site is the cooperative

effect between an orthosteric and allosteric ligand. Allosteric modulators can act as agonists, positive allosteric modulators (PAM), or antagonists, negative allosteric modulators (NAM) to the effects exerted by the endogenous ligand. This allows for a greater modulation of the effects produced along with a protection mechanism against overdose [256, 257].

Although we have briefly touched upon the basics of GPCR drug design and the effects of orthosteric and allosteric ligand binding, the complexity of these effects, as well as GPCR activation, is enormous with many aspects of it still poorly understood. It is also beyond the scope of this chapter. What will be presented here is the qualitative and quantitative evaluation of the ability of the dynamic pharmacophore probes approach to map and evaluate orthosteric and allosteric binding sites on a selction of GPCR targets. This is done through the use of CG MD simulations, making it extremely fast, compared to the currently available methods.

4.2 Methods

In order to test the ability of the dynamic pharmacophore probes approach to identify orthosteric and allosteric binding sites on GPCRs, a test set of GPCRs of pharmacological interest was selected. These targets were also selected based on the availability of published structures with endogenous and allosteric ligands bound. This data set is shown in figure 4.4.

The simulations in this chapter were performed using the GROMACS v5.0.2 simulation package with GPU acceleration [188]. The initial protein coordinates used in the simulations were obtained from the pdb. A2A (pdb id :3eml), B2-adrenoceptor (pdb id: 2rh1), CCR2 (pdb id: 5t1a), CCR5 (pdb id: 4mbs), CCR9 (pdb id: 5lwe), CRF1R (pdb id: 4k5y), D3 receptor (pdb id: 3pbl), GCGP (pdb id: 5ee7), GLP1 (pdb id: 5vex), GPR40 (pdb id:4phu), M2 receptor (pdb id: 4mqt), P2Y1 (pdb id: 4nxv), S1P1 (pdb id: 3v2y) and smoothened (pdb id: 5l7d). Any missing atoms in the structures were completed using MODELLER [223]. The structures were energy minimised using the steepest descent algorithm and the MARTINI forcefield v2.2 [200] was used to convert the structures into their CG representations. The simulation protocol described in results chapter 1 was used. Table 4.1 shows a

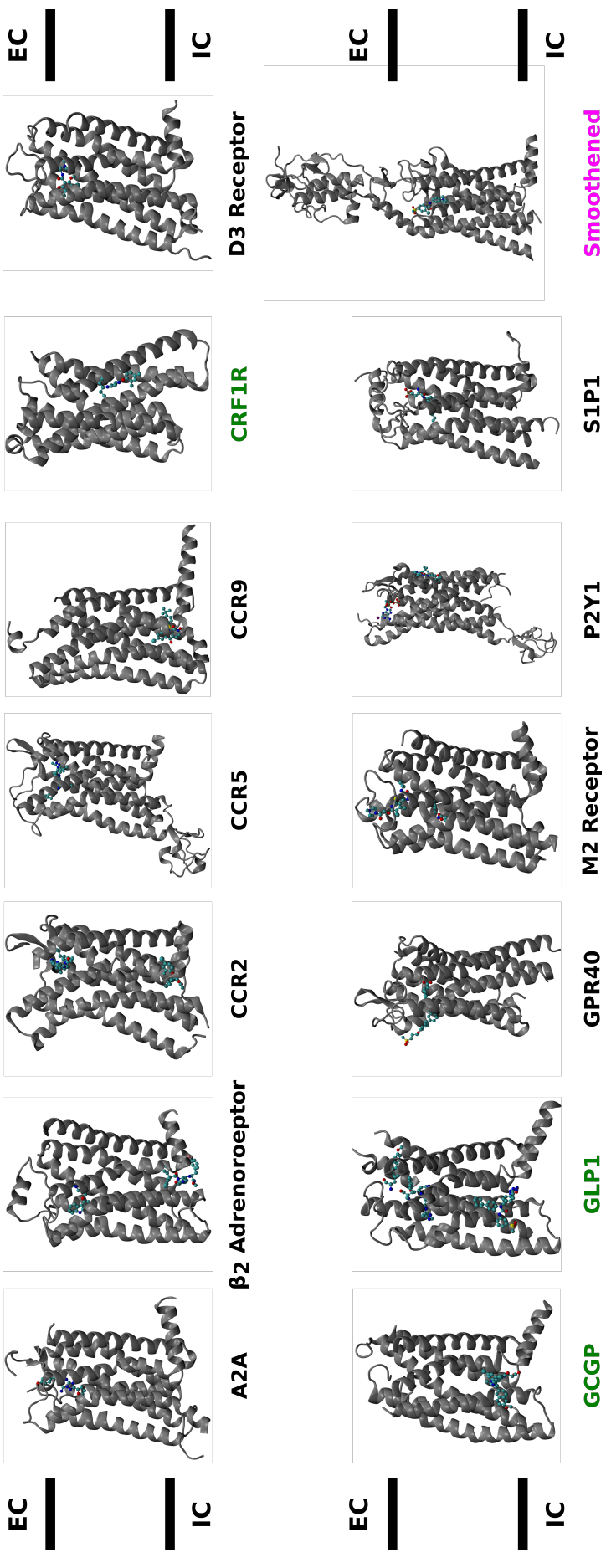


Figure 4.4: A figure showing the GPCR targets used in this chapter, in a cartoon representation, with bound endogenous and allosteric ligands shown as coloured balls and sticks. The horizontal black lines denote the cell membrane and the extracellular (EC) and intracellular (IC) space. Structures whose name is shown in black belong to **class A**, structures whose name is shown in green belong to **class B** and those shown in magenta belong to **class F**.

Table 4.1: A table showing the targets used, the number of simulations performed and the total simulation time.

Target	Number of Simulations	Total Simulation Time	Resolution
A2A (3EML)	3	3 x 2μs	2.60 Å
B2-Adrenoceptor (2RH1)	3	3 x 2μs	2.40 Å
CCR2 (5T1A)	3	3 x 2μs	2.80 Å
CCR5 (4MBS)	3	3 x 2μs	2.70 Å
CCR9 (5LWE)	3	3 x 2μs	2.80 Å
CRF1R (4K5Y)	3	3 x 2μs	2.90 Å
D3 Receptor (3PBL)	3	3 x 2μs	2.90 Å
GCGP (5EE7)	3	3 x 2μs	2.50 Å
GLP1 (5VEX)	3	3 x 2μs	3.0 Å
GPR40 (4PHU)	3	3 x 2μs	2.30 Å
M2 Receptor (4MQT)	3	3 x 2μs	3.7 Å
P2Y1 (4NXV)	3	3 x 2μs	2.20 Å
S1P1 (3V2Y)	3	3 x 2μs	2.80 Å
Smoothened (5L7D)	3	3 x 2μs	3.20 Å

summary of the simulations run.

The MARTINI forcefield v2.2 was used to perform the CG molecular dynamics simulations, with a time step of 20 fs. Isotropic Berendsen pressure coupling was used [179] with a coupling constant of 12 ps. The reference pressure was set at 1 bar for the XY and Z directions. A stochastic rescaling temperature control was used [182], with a characteristic time step of 1 ps, keeping the temperature constant at approximately 323 K. A cut-off of 1.2 nm was used for Lennard-Jones (LJ) interactions with the potential being smoothly switched off between 0.9 nm and 1.2 nm. The Coulombic interactions were calculated using the Reaction Field method with a cutoff of 1.1 nm. These settings were consistently used throughout the simulations, across all the targets.

The MDAnalysis [224] software package was used to calculate the interaction densities of the different probe types. Analysis of the simulation data was conducted using the VMD [74] and PyMol [258] visualisation packages along with tools implemented in GROMACS and in house scripts. The same analysis protocol as described in the previous chapter was followed in order to transform the binding densities into pharmacophoric points and identify the surrounding residues. The same protocol,

as described in results chapter 1, for qualitatively and quantitatively evaluating the predictive capabilities of the pharmacophoric probes was also followed. It should be noted that the number of published holo structures for the targets used here, and GPCRs in general, is significantly smaller compared to the targets of pharmacological interest used in the previous chapter.

4.3 Results

4.3.1 Identification of protein-ligand interactions

The first step in evaluating the ability of the probes to identify protein-ligand interactions was to qualitatively evaluate their ability to recapitulate the interactions which drive ligand binding. With this in mind, like in the previous chapter, the Maestro software suite was used to prepare a selection of the targets, which included both allosteric and orthosteric bound ligands, and to generate the ligand interaction plots. The resulting pharmacophoric points were then superimposed with the holo structures and examined, in an attempt to see whether the interactions identified in the ligand interaction plots were also able to be identified by the superimposed pharmacophoric points.

4.3.2 M2 Muscarinic Acetylcholine Receptor

The first target to be examined was the M2 receptor with a bound allosteric ligand (figure 4.5). Three hydrogen bond interactions were identified between the ligand and the receptor, one with TYR 80, one with ASN 410 and one with ASN 419 (panel C). Looking at panels A and B of figure 4.5, we can see that the interaction with TYR 80, marked as number 2, was identified by the hydrogen bond acceptor points (panel A). The interaction with ASN 419, marked as number 3, was identified by the hydrogen bond donor points (panel B). Finally, the interaction with ASN 410, marked as number 1, was identified by both the hydrogen bond donor and acceptor points (panel A and B).

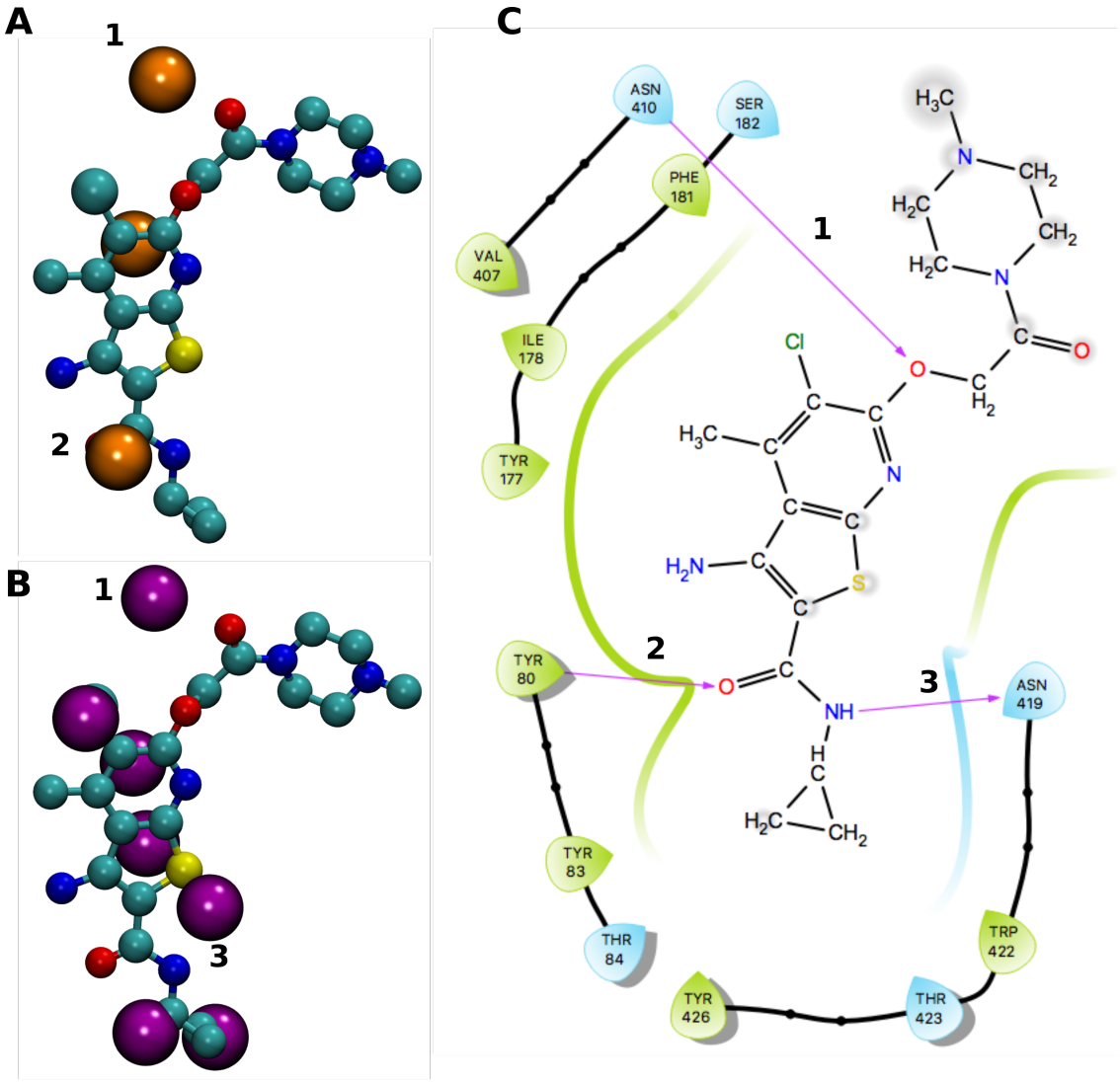


Figure 4.5: A figure of the bound positive allosteric modulator (LY2119620) of the holo structure of the M2 receptor with the overlaid pharmacophoric points. (A) Hydrogen bond acceptors are shown in orange and (B) Hydrogen bond donors are shown in purple. (C) The interaction plot of the ligand is shown in panel C. The numbers on panels A, B and C correspond to the interactions identified in the interaction plot and by each pharmacophoric point.

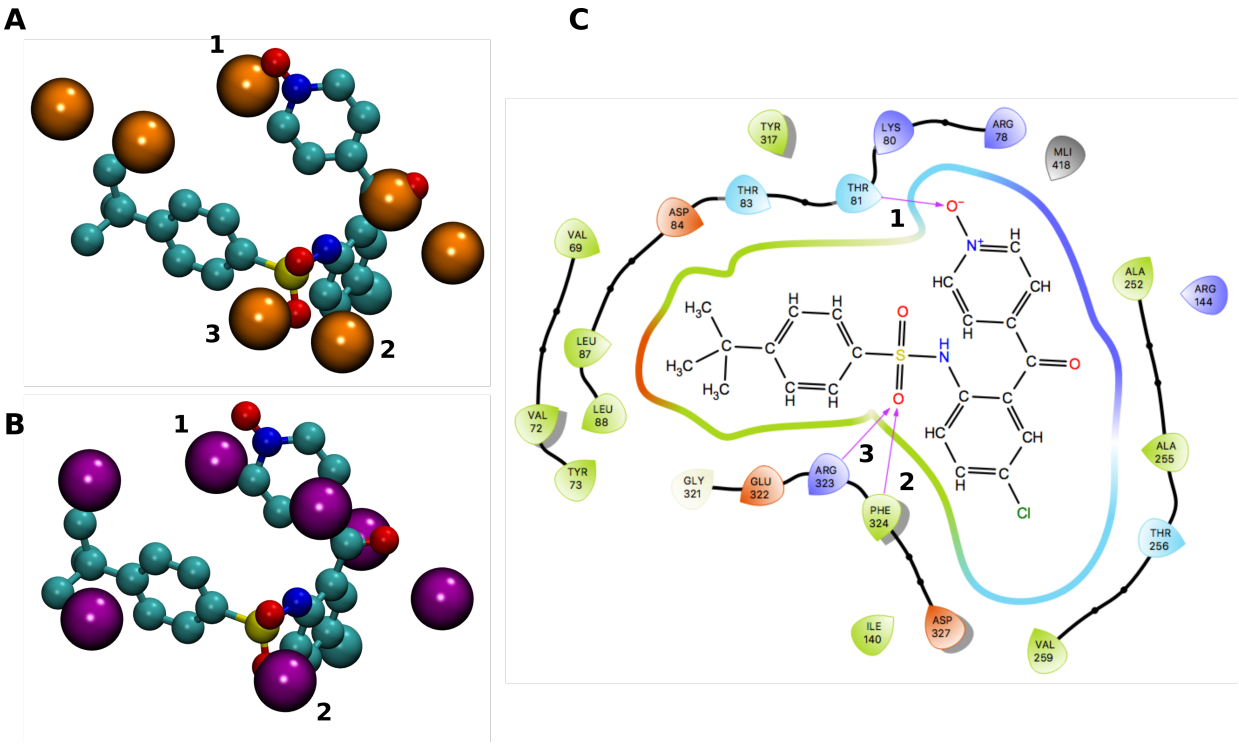


Figure 4.6: As described in figure 4.5, for the case of the CCR9 receptor with a bound allosteric agonist (vercicoron).

4.3.3 Chemokine Receptor 9

The next target to be examined was Chemokine receptor 9 (CCR9), which also has an allosteric ligand bound (figure 4.6). In this case, three hydrogen bonds between the ligand and the target were also identified, one with THR 81, one with ARG 323 and one with PHE 324. Looking at panels A and B of figure 4.6, we can see that the interaction with ARG 323, marked as number 3, was identified by the hydrogen donor points (panel B). The interactions with PHE 324, marked as number 2, and THR 81, marked as number 1, were identified by both hydrogen bond donor and acceptor points (panel A and B).

4.3.4 β2-Adrenoceptor

The target selected next, which has both an allosteric and an orthosteric ligand bound, was β2-Adrenoceptor. Starting from the allosteric ligand (figure 4.7), a plethora of hydrogen bond interactions, between the ligand and the receptors, were identified (panel C). These include one with ARG 63, one with ASN 69, one with THR 274, one with SER 329 and one with ASP 331. Looking at panels A and B

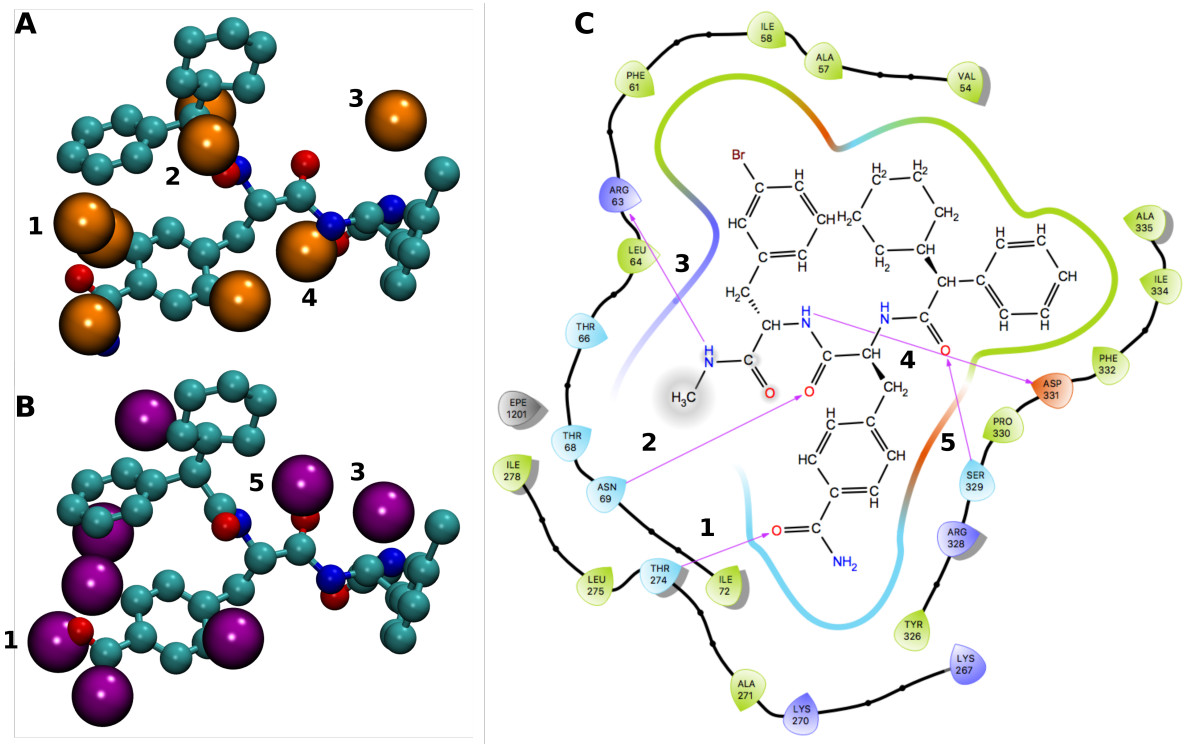


Figure 4.7: As described in figure 4.5 for the case of the β2-Adrenoceptor.

it can be seen that the interactions between the ligand and THR 274, marked as number 1, and between the ligand and ARG 63, marked as number 3, are both identified by hydrogen bond donor and acceptor points. The interaction between the ligand and Ser 329, marked as number 5, is only identified by the hydrogen bond donor points (panel B). Finally, the interactions between the ligand and ASN 69, marked as number 2, and between the ligand and ASP 331, marked as number 4, are only identified by hydrogen bond acceptor points (panel A).

Moving on to the ligand bound to the orthosteric binding site of β2-Adrenoceptor (figure 4.8), three hydrogen bond interactions were identified between the ligand and the receptor. One was with SER 203, marked as number 1, one with ASP 113, marked as number 2, and one with ASN 312, marked as number 3. Looking at panels A and B, it can be seen that all three interactions are identified by both hydrogen bond donor and hydrogen bond acceptor points.

4.3.5 Sphingosine 1-phosphate receptor

The final target to be examined was the sphingosine 1-phosphate receptor (S1P1) which has a ligand bound in its orthosteric binding site. Looking at figure 4.9, it can

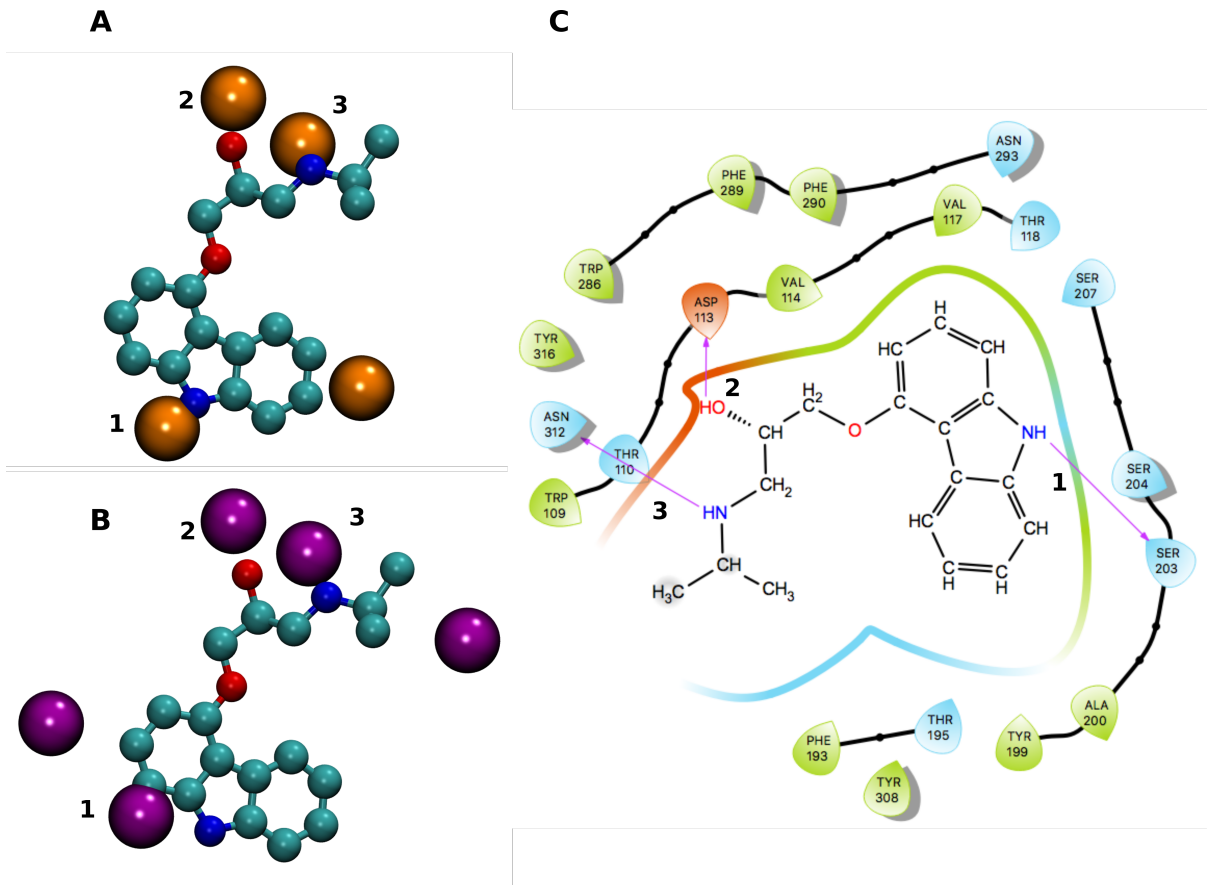


Figure 4.8: As described in figure 4.5 for the case the holo structure of β2-Adrenoceptor with a bound orthosteric inhibitor (carazolol).

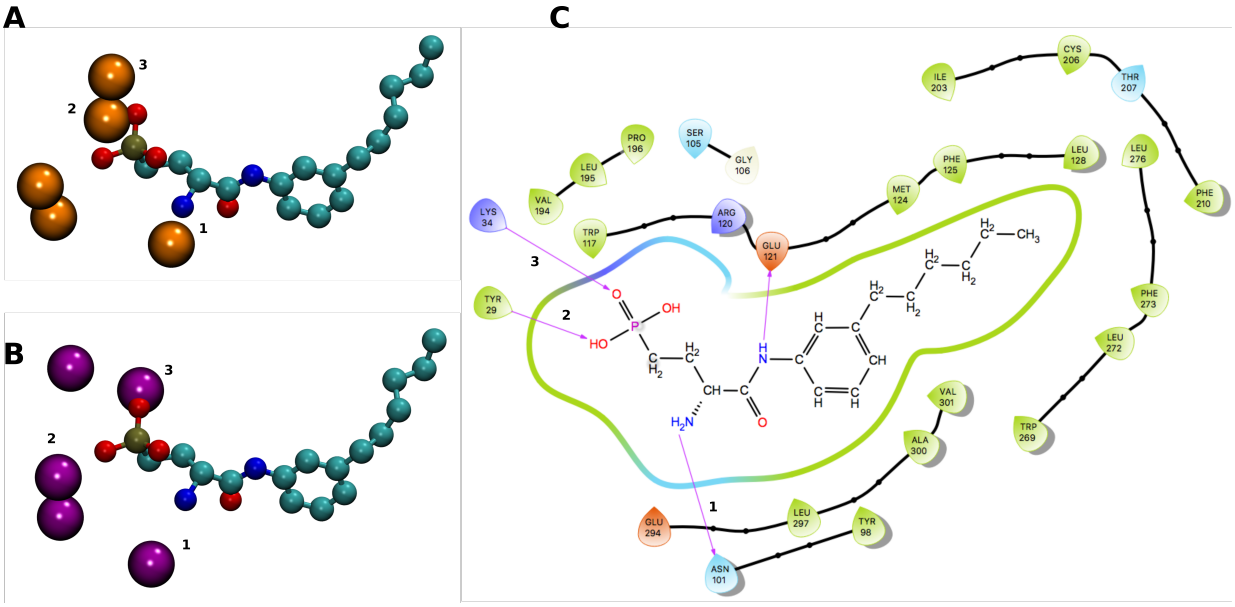


Figure 4.9: As described in figure 4.5 for the case of the holo structure of the S1P1 receptor with a bound orthosteric antagonist (a sphingo-lipid mimic).

be seen that three hydrogen bond interactions, between the ligand and the receptor, were identified. One with ASN 101, marked as number 1, one with TYR 29, marked as number 2, and one with LYS 34, marked as number 3. Examining panels A and B, it can be seen that all three interactions are identified by both hydrogen bond donor and hydrogen bond acceptor points. It should be noted that one more hydrogen interaction between the ligand and GLU 121 was identified in the ligand interaction plot, that was not identified by either the hydrogen bond donor or acceptor points. A possible explanation for this could be the high concentration of charged or polar residues at the top site of the orthosteric binding site, towards the extracellular part of the receptor, where the phosphate tail of the ligand binds. These residues could cause a higher concentration of binding of the probes in these areas, making it hard to penetrate further in the orthosteric binding site, where GLU 121 is located. Moreover, GLU 121 is located next to the charged ARG 120, which could also play a role in shielding it from the probes. In fact, examining the positively charged probes, it can be seen that they have the ability to identify the interaction with GLU 121, supporting the idea that the charged residues act as a screening barrier, only allowing the probes that form the strongest interactions through. Finally, the hydrophobic probes are seen to to penetrate all the way to the hydrophobic part of the binding site, in contrast with the other probe types, further supporting the idea of the screening by the charged residues.

4.3.6 Quantifying the identification of interactions

In order to quantitatively evaluate the ability of the probes to identify the interactions that drive ligand binding, the same analysis protocol as described in chapter 3 was repeated. Tables 4.2 and 4.3 show the results for the orthosteric and allosteric binding sites of the data set, respectively.

Tables 4.2 and 4.3 show that for the majority of the orthosteric and allosteric binding sites, a score of 0.6 or higher is achieved, indicating that most of the interactions driving ligand binding are identified. In the few cases, such as in the case of A2A and P2Y1, a lower score of 0.50 is obtained. A possible explanation for this, as it was suggested previously, could be the existence of a lot of residues, which are involved in ligand binding, towards the extracellular domain of the binding site,

Table 4.2: A table of the targets with orthosteric binding sites and the score for each target. A score is assigned to each target based on the interactions identified by the pharmacophoric points. For each of the structures used in Table 4.1, the ligand interaction plot was calculated. The structures were then checked against the pharmacophoric points to determine whether the interactions can be identified. If the interaction was able to be identified a score of 1 was assigned and if not, a score of 0, for each interaction. The total score was then divided by the total number of interactions to produce the score shown here.

Target	Score
A2A (3EML)	0.50
B2-Adrenoceptor (2RH1)	0.87
CCR2 (5T1A)	0.66
D3 Receptor (3PBL)	1.0
GLP1 (5VEX)	0.66
M2 Receptor (4MQT)	1.0
P2Y1 (4NXV)	0.50
S1P1 (3V2Y)	0.80

Table 4.3: A tables of the targets with allosteric binding sites and the score for each target. A score is assigned to each target based on the interactions identified by the pharmacophoric points. For each of the structures used in Table 4.1, the ligand interaction plot was calculated. The structures were then checked against the pharmacophoric points to determine whether the interactions can be identified. If the interaction was able to be identified a score of 1 was assigned and if not, a score of 0, for each interaction. The total score was then divided by the total number of interactions to produce the score shown here.

Target	Score
B2-Adrenoceptor (2RH1)	0.92
CCR2 (5T1A)	0.92
CCR5 (4MBS)	0.65
CCR9 (5LWE)	0.84
CRF1R (4K5Y)	0.93
GCGP (5EE7)	0.75
GLP1 (5VEX)	0.55
GPR40 (4PHU)	0.61
M2 Receptor (4MQT)	1.0
P2Y1 (4NXV)	0.66
Smoothened (5L7D)	0.80

Table 4.4: A table showing all the residues involved in ligand binding in the orthosteric binding sites of the targets examined.

Target	Residues Involved in Ligand Binding
A2A (3EML)	A63, S67, T88, F168, E169, F177, E178, N181, H250, N253, Y271, S277, H278, F289, H355, N358
B2-Adrenoceptor (2RH1)	D113, F193, S203, S207, F289, N293, N312
CCR2 (5T1A)	Y49, Q288, T292
D3 (3PBL)	F346, F345
GLP1 (5VEX)	W33, Y148, Y205, N300, D372, R380
M2 (4MQT)	D103, Y104, Y403, N404, W426
P2Y1 (4NXV)	K46, R195, T201, T205, N283, Y303, R310
S1P1 (3V2Y)	Y29, K34, N101, T109, E121, F125, W269, L272

trapping the probes there and preventing them from penetrating further. In addition to that, the interactions not identified by the probes are with the residues which are found between the TM helices of the binding site. The bulkier CG representation of amino acids in the MARTINI forcefield, compared to an all atom model, could mean that there is simply not enough space in that crowded environment, even for the tiny probes, to penetrate. Nevertheless, the dynamic pharmacophore probes show to have the ability to accurately detect the majority of the ligand binding interactions.

4.3.7 Identifying the residues involved in ligand binding

In the previous section, it was shown that the pharmacophoric points have the ability to identify the interactions driving ligand binding. However, looking at the figures, it can be seen that the number of interactions shown on the ligand interaction plots are less than the number of pharmacophoric points identified. In an attempt to assess whether these points were pointing to potential interactions that are fulfilled by different moieties than the ones shown in these examples or whether they were false positives, all the residues involved in ligand binding were identified (Tables 4.4 and 4.5). These were then compared with the residues identified by the

Table 4.5: A table showing all the residues involved in ligand binding in the allosteric binding sites of the targets examined.

Target	Residues Involved in Ligand Binding
B2-Adrenoceptor (2RH1)	V54, I58, R63, I64, N69, I72, T274, L275, Y326, S329, D331, F332
CCR2 (5T1A)	V63, L67, L81, L134, A241, V244, I245, Y305, E310, K311, F312
CCR5 (4MBS)	Y37, W86, Y89, Y108, F109, F112, Q194, T195, I198, W248, Y251, L255, T259, M279, E283, M287
CCR9 (5LWE)	V69, V72, Y73, R78, T81, T83, D84, L87, I140, R144, R233, R252, T256, V259, Y317, E322, R323, F324
CRF1R (4K5Y)	F203, M206, V279, L280, N283, F284, L287, I290T 316,L319, L323, Y327
GCGP (5EE7)	R199, L329, F345, R346, K349, S350, L352, T353, Q392, L399, N404, K405
GLP1 (5VEX)	F324, I328, V331, V332, L335, F347, K351, S352, L354, S355, L359, F390, M397, L401, V405, N406
GPR40 (4PHU)	Y44, A 83, V84, F87, A99, A102, L106, Y114, F117, Y122, S123, V126, I130, L133, V134, L135, L138, F142, L171, W174, R183, L190, L193, ILE197, Y240, R258
M2 Receptor (4MQT)	Y80, Y83, Y177, N410, N419, W422, Y426
P2Y1 (4NXV)	F62, F66, L102, T103, A109, F119, M123, L126, N127
Smoothened (5L7D)	L221, M230, W281, D384, V386, S387,I389, F391, Y394, R400, H470, Q477, W480, F484, P513, E518, L522, M525

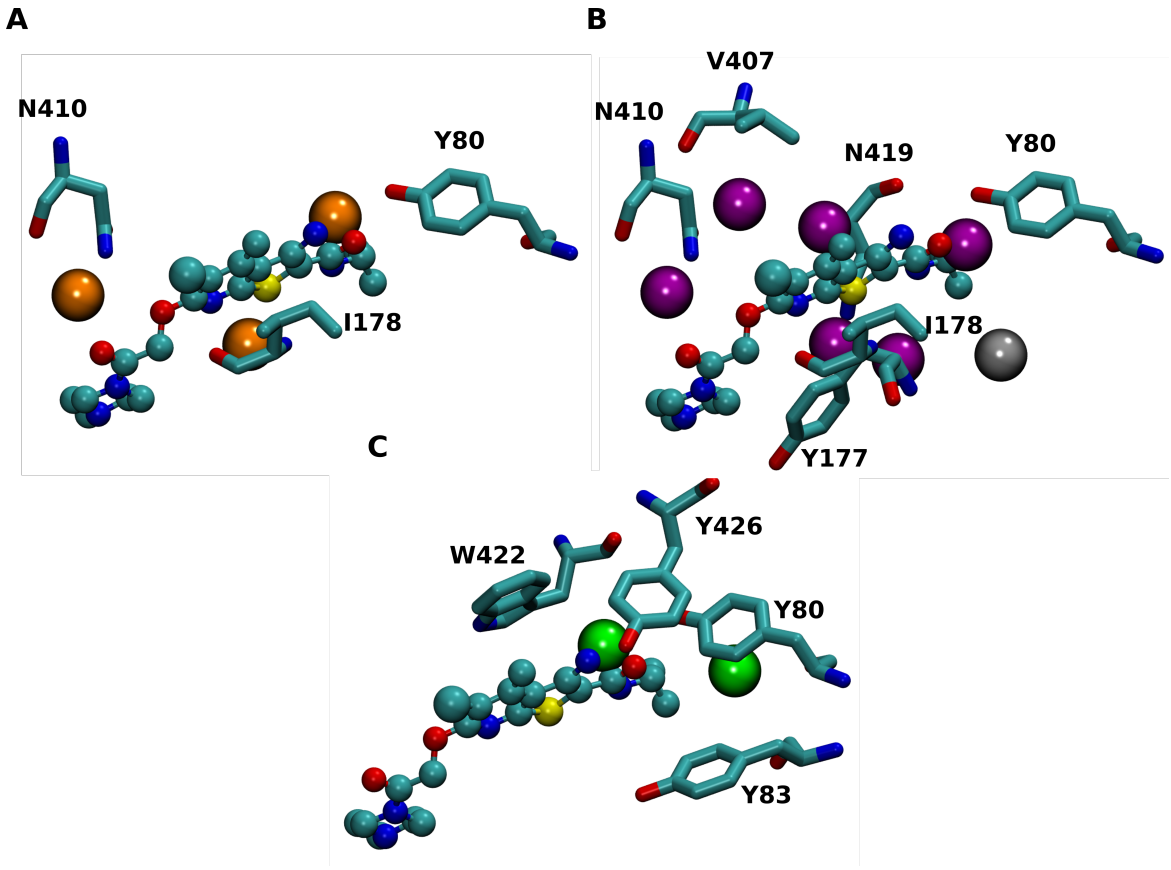


Figure 4.10: A figure of the bound allosteric ligand of the holo structure of the M2 receptor with the overlaid pharmacophoric points along with the residues identified by the points. (A) Hydrogen bond acceptors are shown in orange and (B) Hydrogen bond donors are shown in purple. (C) Hydrophobic ones are shown in green.

pharmacophoric points in order to determine whether the residues identified by the points corresponded to residues involved in ligand binding or false positives.

Examining the same targets as before, figure 4.10 shows the allosteric ligand of the M2 receptor along with the residues identified in ligand binding. The residues identified by the hydrogen bond acceptors were Y80, I178 and N410 (figure 4.10 A), while the ones identified by the hydrogen bond donor points were Y80, Y177, I178, V407, N410 and N419 (figure 4.10 B). Finally, the residues identified by the hydrophobic points were Y80, Y83, W422 and Y426 (figure 4.10 C). Looking at table 4.5, it can be seen that all these residues belong to those identified as being involved in ligand binding in the allosteric binding site of M2. The exception to this were I178 and V407. Their identification by the probes can be explained by their proximity to the residues involved in ligand binding, as well their orientation towards the binding site.

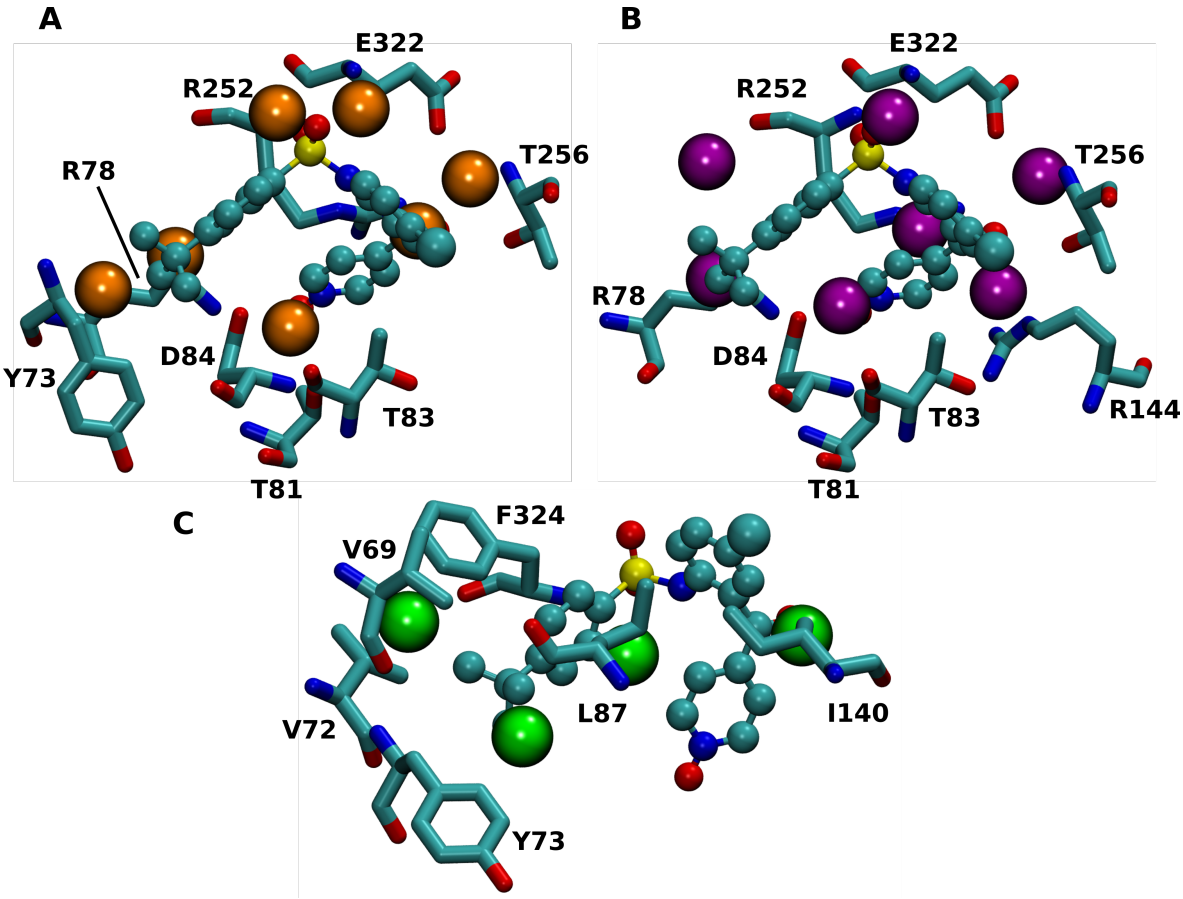


Figure 4.11: As described in figure 4.10 for the case of the allosteric binding site of the CCR9 receptor.

The next target was CCR9. Looking at figure 4.11, the residues identified by the hydrogen acceptor poinys were Y73, R78, T81, T83, D84, R252, T256 and E322 (figure 4.11 A). The residues identified by the hydrogen bond donor points were R78, T81, T83, D84, R144, R252, T256, E322 (figure 4.11 B). Finally, the ones identified by the hydrophobic points were V69, V72, Y73, L87, I140 and F324. Looking at table 4.5, it cab be seen that all of these residues were identified as being involved in ligand binding.

Next, the allosteric binding site of the B2-Adrenoceptor was examined (figure 4.12). The residues identified by both the hydrogen bond acceptor and donor points were R63, N69, T274, S329 and D331 (figure 4.12 A and B). The ones identified by the hydrophobic points were V54, I58, L64, I72, L275 and F332 (figure 4.12 C). Again, comparing these residues with the ones in table 4.5, it can be seen that all of them were identified as residues that are involved in ligand binding. However, in the case of the hydrophobic pharmacophoric points, two of them were not found to interact with any of the residues involved in ligand binding (figure 4.12 C, shown as grey spheres). These were classified as false positives.

Following that, the orthosteric binding site of the β2-Adrenoceptor was examined (figure 4.13). The residues identified by the hydrogen bond acceptor points were D113, S203, N293 and N312 (figure 4.13 A). The residues identified by the hydrogen bond donor residues were D113, S203, S207 and N312 (figure 4.13 B). Finally, the ones identified by the hydrophobic points were V114, F193, N293 and F289 (figure 4.13 C). Examining table 4.4, it can be seen that all these residues were identified as being involved in ligand binding, with the exception of V114. Its identification by the hydrophobic probes could be explained by its close proximity to D113.

The final target to be examined was the orthosteric binding site of the S1P1 receptor (figure 4.14). The residues identified by the hydrogen bond acceptor probes were Y29, K34, N101, T109 and G106 (figure 4.14 A). The same residues were also identified by the hydrogen bond donor probes (figure 4.14B). Finally, the hydrophobic probes identified M124, F125, W269 and L272 (figure 4.14 C). Comparing these residues with those in table 4.4, it can be seen that they were all identified as being involved in ligand binding, except for M124 and G106. As before, their identification by the probes could be explained by their close proximity to residues playing a role

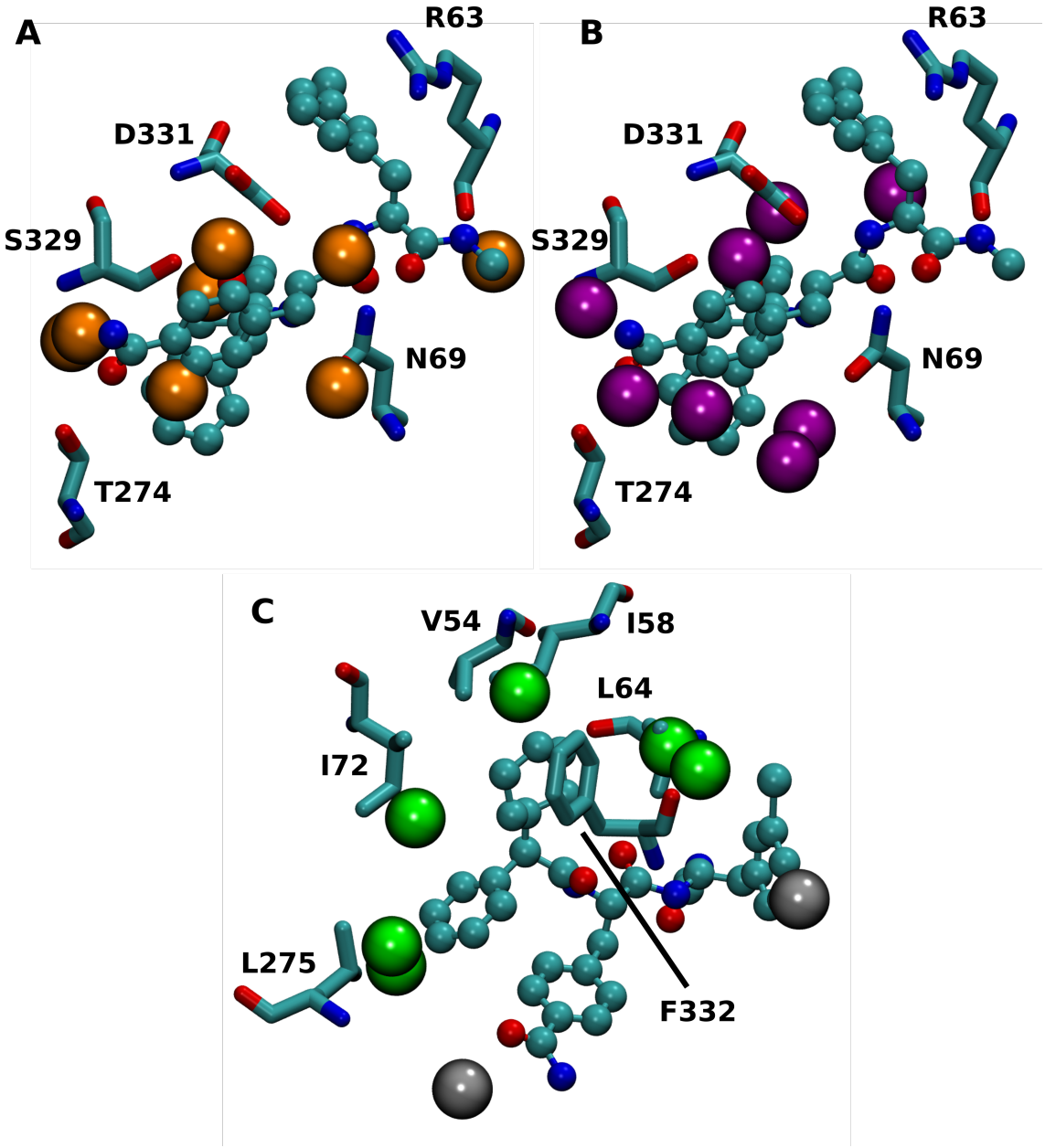


Figure 4.12: As described in figure 4.10 for the case of the allosteric binding site of the β2-Adrenoceptor.

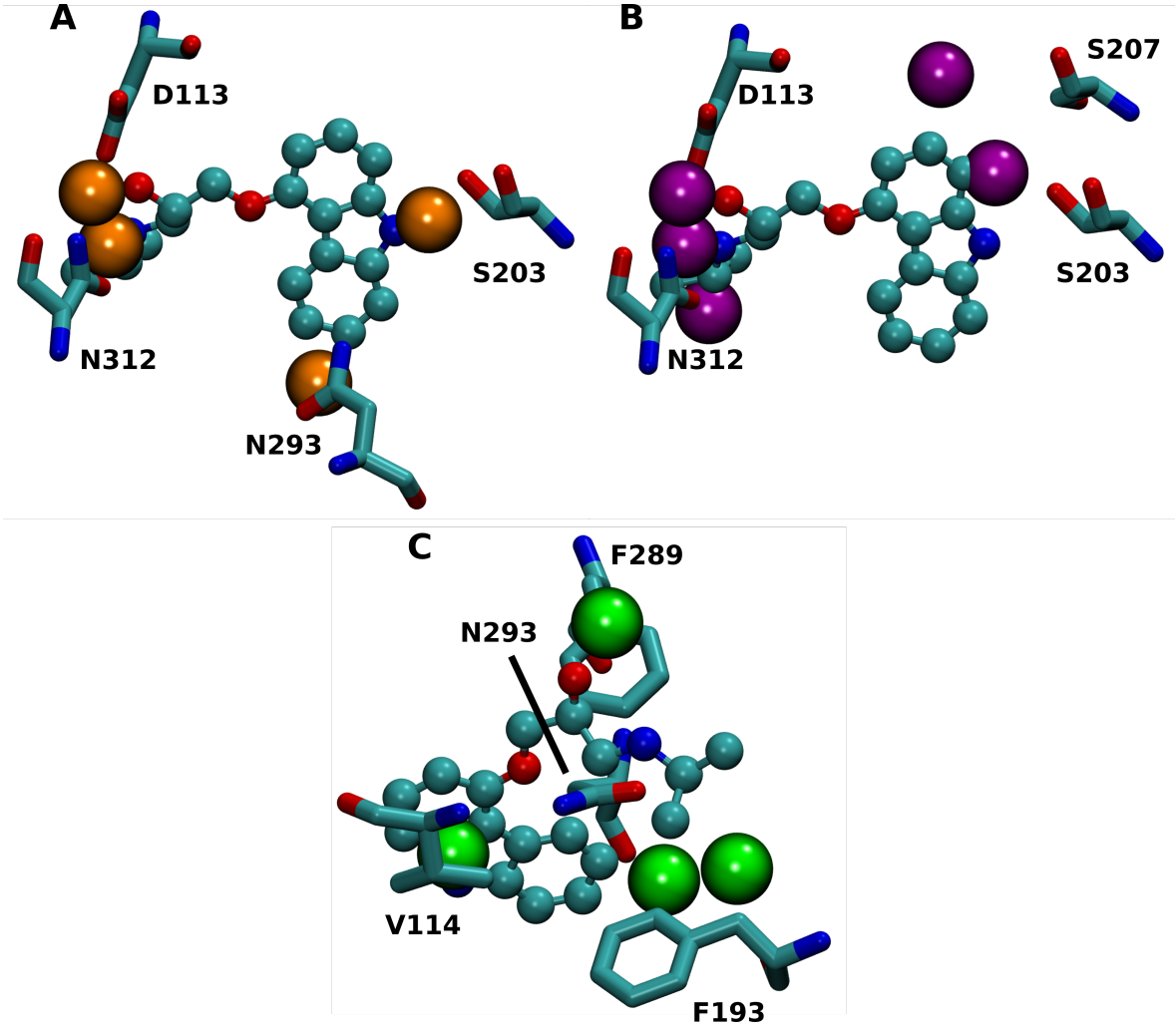


Figure 4.13: As described in figure 4.10 for the case of the orthosteric binding site of the β2-Adrenoceptor.

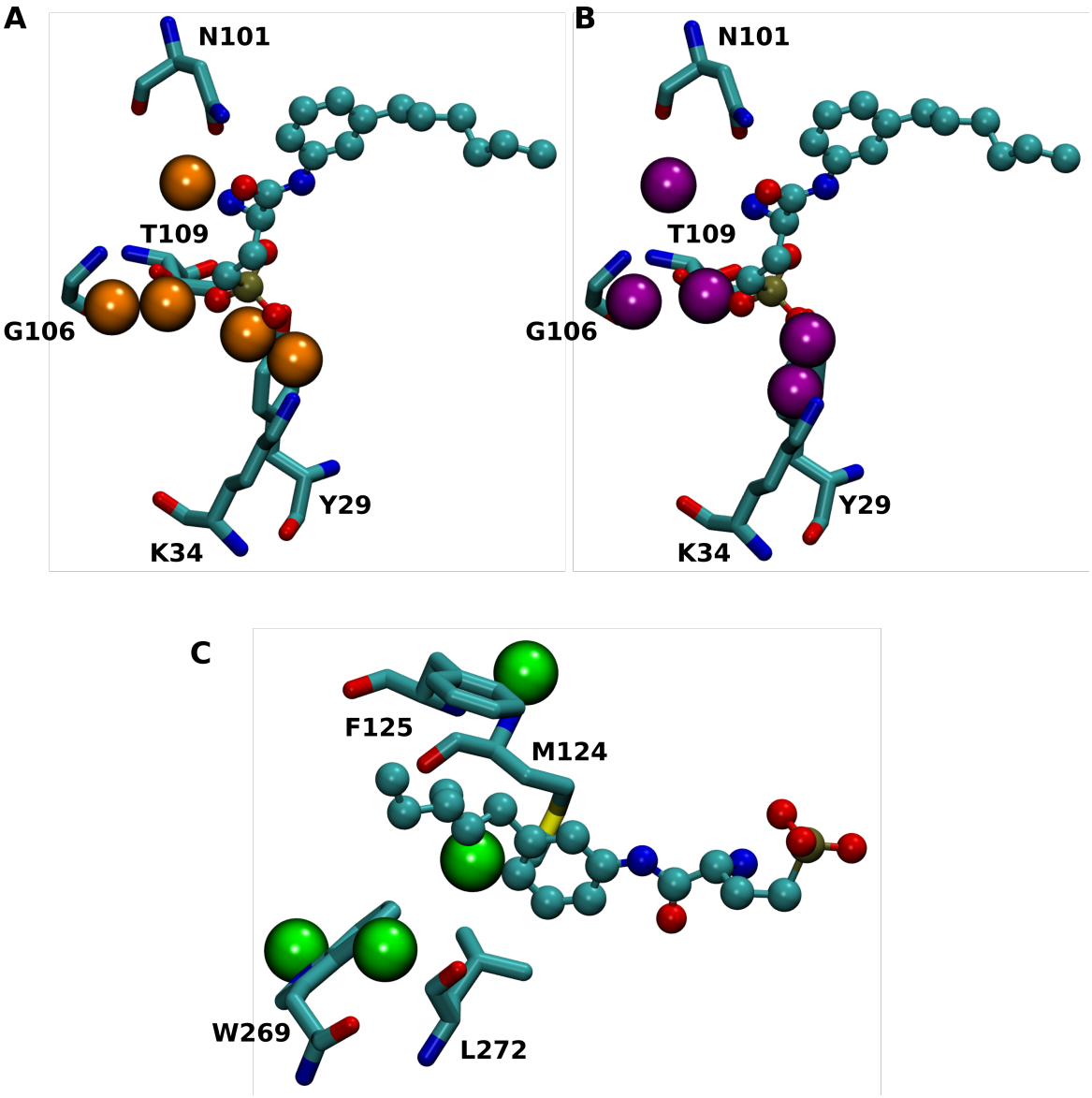


Figure 4.14: As described in figure 4.10 for the case of the orthosteric binding site of the S1P1 receptor.

Table 4.6: A table showing the sensitivity, precision, specificity and accuracy values calculated for each type of pharmacophoric probe for the orthosteric binding sites.

	Sensitivity	**Precision**	**Specificity**	**Accuracy**
Hydrophobic	0.75	0.52	0.71	0.66
H-Bond Acceptors	0.70	0.90	0.80	0.62
H-Bond Donors	0.62	0.95	0.93	0.68
Polar	0.51	0.86	0.80	0.60
Positive Charge	0.48	0.87	0.86	0.61
Negative Charge	0.38	0.70	0.69	0.48

Table 4.7: A table showing the sensitivity, precision, specificity and accuracy values calculated for each type of pharmacophoric probe for the allosteric binding sites.

	Sensitivity	**Precision**	**Specificity**	**Accuracy**
Hydrophobic	0.70	0.95	0.90	0.76
H-Bond Acceptors	0.86	0.83	0.92	0.90
H-Bond Donors	0.84	0.67	0.80	0.81
Polar	0.76	0.82	0.90	0.87
Positive Charge	0.62	0.78	0.91	0.82
Negative Charge	0.60	0.79	0.91	0.82

in ligand binding and their orientation towards the binding.

4.3.8 Quantitative evaluation of the predictive abilities of the probes

From all the above, it can be seen that the dynamic pharmacophore probes have the ability to accurately identify residues involved in ligand binding. To further evaluate the predictive abilities of the probes, the same metrics as in chapter 1 were used. The same protocol was followed and the sensitivity, precision, specificity and accuracy for each type of probe, and each type of binding site, were calculated (Tables 4.6 and 4.7).

The meaning and significance of each parameter has been explained in the previous chapter, in the methods section. Looking at tables 4.6 and 4.7, the high capacity of the probes to correctly predict residues that play a role in ligand binding is evident. In the great majority of cases, the values for each parameter are 0.60 or higher. The

exemption to this can be seen in the case of the sensitivity values of the positively and negatively charged probes, when examining the orthosteric binding sites of the ligands (table 4.6) which are 0.48 and 0.38 respectively. These lower values could be explained by the fact that a very low number of ionic interactions were observed in the cases of the targets examined, making the calculation of accurate statistic parameters difficult. This was also exasperated by the low number of available structures for each target, as mentioned previously. This might not be the case for the allosteric binding sites due to the fact that a lot of them are located on the outer surface of the TM region of the target, regions which are also involved in lipid head group binding and are, therefore, richer in charged residues, forming a higher number of ionic interactions. Overall, it can be said, that the dynamic pharmacophore probes have the ability to accurately identify and map both orthosteric and allosteric ligand binding sites on a variety of GPCR targets.

4.3.9 Discussion

The advantages and limitations of this approach, as well as ways to overcome them, have been elaborated in the discussion section of chapter 1. One interesting observation, comparing the results from tables 4.6 and 4.7 with those from chapter 1, it can be seen that the sensitivity, precision, specificity and accuracy values are better for the GPCR targets, for both orthosteric and allosteric binding sites, compared to the water soluble targets of the Ichihara data set. This is interesting due to the fact that GPCR targets have been known as a notoriously difficult target to explore. One potential explanation is that the better results could be due to the smaller size of the GPCR binding sites, compared to the water soluble targets. The size of the probes allows them to penetrate the binding sites, despite the sites' smaller size, and in return, the smaller size of the binding sites allows for greater local sampling of the regions, leading to better results.

Orthosteric binding sites on GPCRs are well localised and very well conserved in most classes. However, allosteric binding sites can be widespread throughout the structure of a receptor. The dynamic pharmacophore probes approach offers the potential of exploring the area of interest in a very quick manner, taking a certain degree of dynamics into account, something not currently offered by the

methods available. One drawback would be the limited ability of this approach to detect allosteric binding sites that are not located within a cavity. Even though this implies that no "false" allosteric sites would be detected, it also increases the chances of allosteric or cryptic binding sites being missed. Combining this approach with established allosteric binding site [259, 260] and cryptic binding site detection methods [59, 261], would greatly improve the predicting capabilities of this approach.

Finally, this method offers the advantage of being able to explore the protein-ligand interactions of membrane proteins, in the absence of a membrane, reducing the computational cost considerably. However, it is known that the conformations of membrane proteins are also affected by interactions with lipids. The lack of a membrane might not allow for the exploration of conformations that are only achievable in the presence of one, exasperating the loss of conformational space that results from CG the protein. On the other hand, the low computational costs of this approach allows for the simulation of different conformations, resulting from atomistic simulations or other methods offering a way around this limitation.

4.3.10 Conclusions

In this chapter we have demonstrated the ability of the dynamic pharmacophore probes approach to explore orthosteric and allosteric binding sites on a wide range of different GPCR targets of pharmacological interest. The ability of the probes to identify the interactions and the moieties driving ligand binding has been shown, along with its ability to identify the residues which are involved in the binding. Moreover, the predictive power of the probes, on this set of targets, has been quantitatively evaluated and has been shown to perform equally on better than the current available methods, being the first study to comprehensively examine these targets. Finally, we have shown the potential of this approach in CADD along with possible improvements for further work.

Chapter 5

Identification of Lipid Binding Sites

Chapter Outline

In this chapter the dynamic pharmacophore probes approach will be applied to a range of membrane proteins with the aim of testing the ability of the method to recapitulate known lipid binding sites. These targets were chosen based on their ability to bind PIP_2 and/or cholesterol. The same simulation parameters and analysis protocols, as described in the previous chapters, were used. A comparison between the lipid binding sites detected by this method and those lipid binding sites detected through experiments or long time scale simulations will then be presented.

5.1 Introduction

The function of a great number of membrane proteins is known to be modulated by the surrounding lipid environment. Such proteins include G-protein coupled receptors [262, 263], potassium channels [264, 265], receptor tyrosine kinases [266], redox proteins such as cytochrome c oxidase [267] and certain P-type ATPases [268]. Lipid modulation can have an effect on different aspects of protein function including modulating protein-protein interactions, having an effect on the activity of a membrane protein and altering the cellular localisation of a protein by restraining a protein to different sections of the cell membrane [153]. There are also cases where a

lipid acts as the native ligand for a protein, rather than an allosteric modulator, such as in the case of sphingosine-1-phosphate (S1P) group of lipid activating GPCRs [269].

Given the importance of lipid binding, there are cases where structural, functional and biophysical assays have been used to provide a detailed view of lipid modulation. In the case of eukaryotic inward rectifying potassium ion channels (Kir), functional assays revealed their activation to be dependent on the presence of anionic lipid phosphatidyl inositol-4,5-bisphosphate (PIP2). Crystal structures [270] and MD studies [271, 272] later revealed the presence of four specific PIP_2 binding sites on Kir channels and allowed the elucidation of the PIP_2 channel modulation mechanism.

Taking a broader view, there have been both experimental and computational techniques to help resolve lipid binding sites and understand their mechanism of action [273, 274]. In applying any of these techniques, it is important to keep in mind the limitations suffered by each one. In the case of X-ray crystallography, the protein structures are determined at low temperatures, in the absence of a membrane, with the exception of lipid cubic phase, and in the presence of detergents. As a result, along with the sometimes limited resolution available, a degree of uncertainty surrounds the "lipid" molecular identity and physiological relevance. The same is true for mass spectrometry approaches, where the protein-lipid complex is solubilized through the use of detergents [275]. Furthermore, in the case of assays which rely on the chemical modification of the lipid, like EPR or certain fluorescent assays, a great degree of perturbation is introduced in the system. Finally, in the case of cryo-electron microscopy (cryoEM), protein structures can be embedded in lipid nano-discs, resulting in "blobs" of density being identified. These densities might correspond to lipids, however they cannot always be assigned with confidence.

Computational methods, such as MD, provide a powerful tool for exploring protein-lipid interactions [274]. However, with the current computing power available, time-scales of more than a few microseconds are not yet available [276]. The use of CG-MD [210] allows for a more approximate representation of the system simulated which in turn opens the door to extending the simulations to hundreds of microseconds [277] and including hundreds of protein molecules [208]. However, even with the use of CG-MD, a significant amount of computational resources and human time is

required before an answer can be obtained.

From all the above, it is evident that a method that can provide quick and accurate predictions for potential lipid binding sites would be of great value. In this chapter, the successful application of the dynamic pharmacophore probes approach to membrane proteins, with the aim of identifying potential lipid binding sites, will be demonstrated. Firstly, Kir2.2 will be used as an example due to the plethora of experimental and computational data pinpointing the location and function of the four PIP_2 binding sites. Furthermore, three GPCR targets, for which lipid binding data from MD simulations and experiments are available, will be used. These three GPCRs include two class A GPCRs, the adenosine A_2A receptor (class Aα) and Neurotensin Receptor 1 (class Aβ) and the class F GPCR Smoothened. Finally, the Polycystin-2 (PC2) channel will be used in an attempt to demonstrate the capability of the method to predict potential lipid binding sites.

5.1.1 Phosphoinositides and Cholesterol

The phosphoinositides, also known as phosphatidylinositides, are a major family of eukaryotic membrane lipids [278]. They consist of the phospho-inositol head group which is attached to the glycerol backbone which is, in turn, covalently attached to two fatty acid tails (Figure 5.1 A). The hydroxyl group of the inositol part can be modified with further phosphoryl groups attached to form a wide range of different moieties [279]. This is how the formation of PIP_2, the most abundant phosphoinositide in the plasma membrane, takes place.

Cholesterol is one of the most important lipids in the cell membrane, being the principal sterol synthesized by mammals and the major precursor for various hormones and vitamins such as Vitamin D and cortisol [280]. It is known that it has the ability to condense and rigidify lipid bilayers [281, 280] as well as fluidizing them, depending on the lipid composition of the bilayer [281, 280]. It is helped in this role by its biophysical properties which allow it to be very easily inserted and extracted from membranes [280]. It consists of four hydrocarbon rings, linked together, with a hydrocarbon tail linked to one end and a hydroxyl group linked to the other end of the steroid (Figure 5.1 B).

Both cholesterol and PIP_2 are known to also interact and modulate the function

of proteins [270, 280]. Furthermore, the charged head group of PIP_2 and the absence of any charged moieties on cholesterol means that their binding is mainly driven by electrostatic and hydrophobic interactions respectively. Given their importance in lipid membranes and protein modulation, along with the less complicated interactions driving their binding to proteins, compared to other moieties, and the data available PIP_2 and cholesterol were chosen as the lipid examples to be studied through the use of the dynamic pharmacophores probes method.

5.1.2 Inward Rectifying Potassium (Kir) Channel 2.2

Kir channels are integral membrane proteins, composed of four identical monomers forming a tetramer, which control the selective permeation of potassium ions across the cellular membranes. They control a range of critical functions such as membrane resting potential and insulin secretion [282]. It has been shown that PIP_2 lipids can act as agonists for Kirs [283] due to PIP_2 being necessary for maximal activation of these channels. The location of PIP_2 binding on these channels has been accurately predicted and confirmed through both experimental and computational methods, for a range of structures [270, 271, 272].

With the release of a high resolution Kir2.2 structure [270] with four short chain (dioctanoyl) PIP_2 lipids bound, the binding sites and interaction forming residues, that were determined through the use of MD simulations, were verified [271]. Further multiscale simulation studies, using the PIP_2 bound, apo and phospatidic acid bound structures, unambiguously identified the four lipid binding sites also seen on the crystal structure (figure 5.2) [272]. The good agreement between the experimental approaches and the MD simulation predictions suggests the high accuracy of the PIP_2 binding site predictions, along with the accuracy of the lipid head and residue interaction predictions.

As a result of the above, Kir2.2 offers an excellent first example on which the accuracy of the lipid binding predictive powers of the dynamic pharmacophore probes approach can be tested.

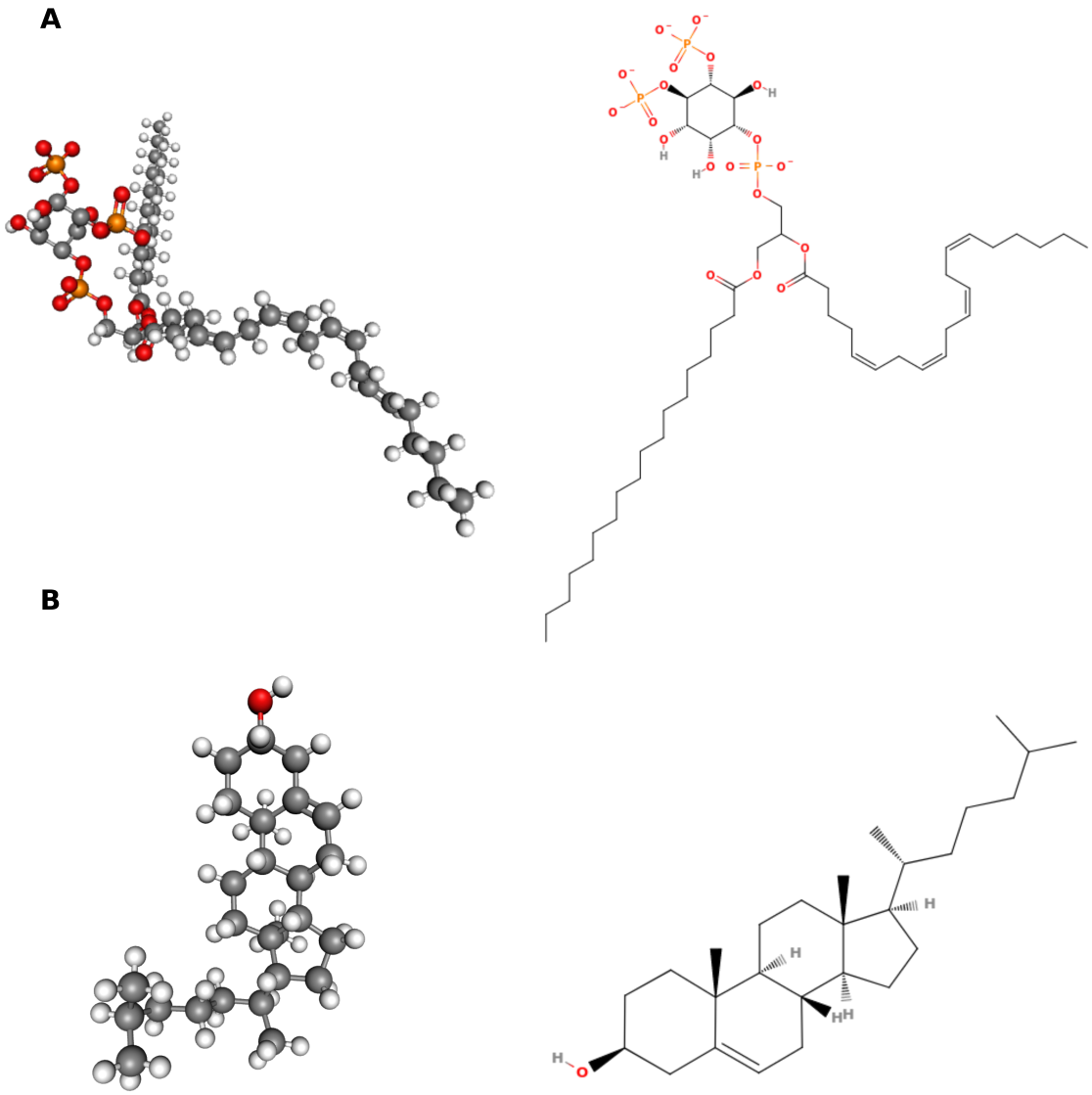


Figure 5.1: The structures of phosphatidylinositol 4,5-bisphosphate (A) and cholesterol (B) along with their chemical formulas.

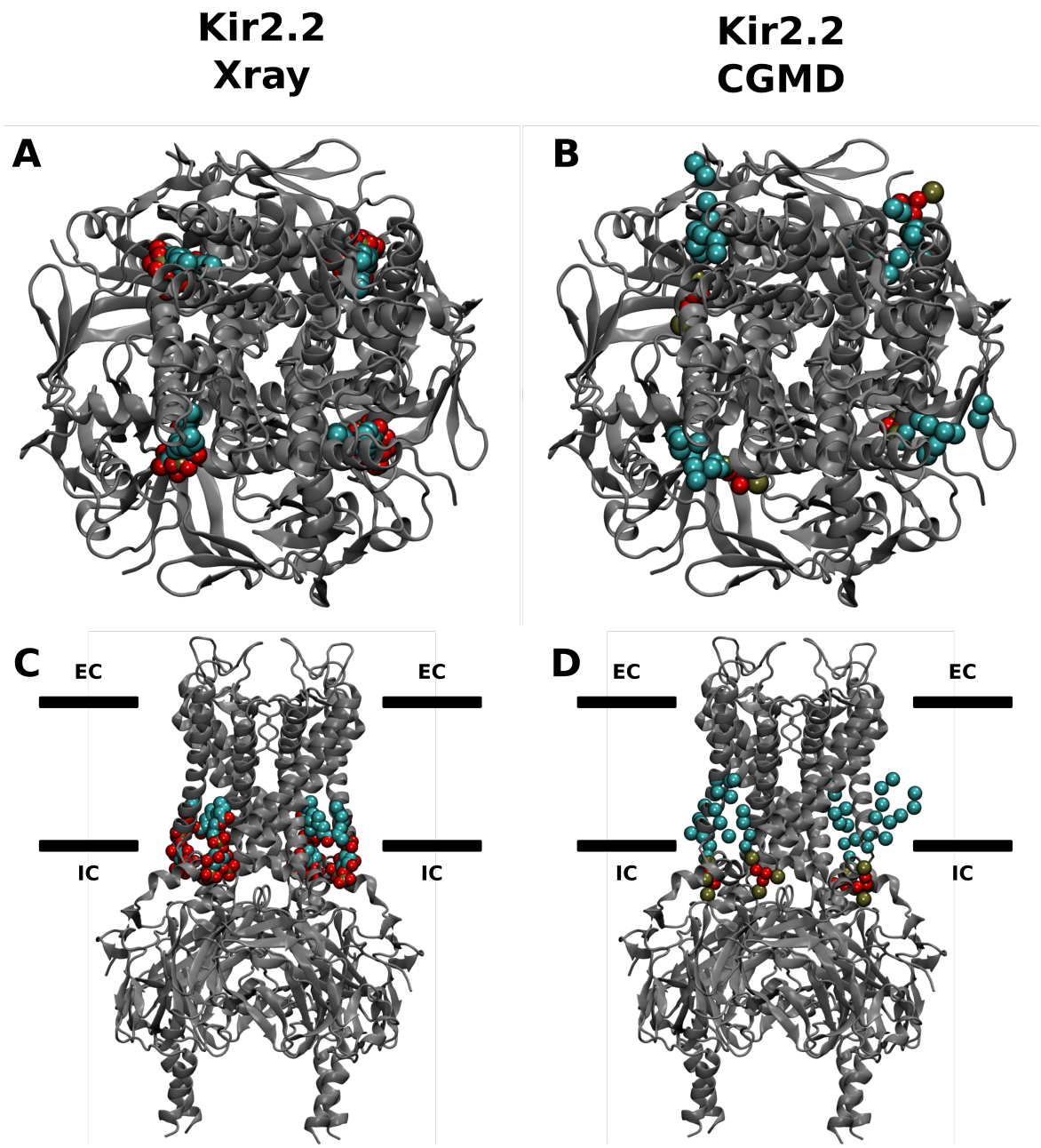


Figure 5.2: Similarities between lipids identified in X-ray crystallography and CG-MD simulations. Panels A and B offer a top down view of the lipid binding sites identified by X-ray crystallography (3SPI) (A) and CG-MD simulations by Stansfeld et al. (B). Panels C and D show the binding sites of the lipid tails in the X-ray structure (C) and after the simulation (D).

5.1.3 Neurotensin Receptor 1

Neurotensin Receptor 1 (NTS1) is a class Aβ GPCR which is mainly found in the central and peripheral nervous system, along with a range of other tissues [284]. Its role is to bind neurotensin [285], leading to the activation of GPCRs through the exchange of GDP for GTP and the subsequent activation of the intracellular signalling cascades. It has the typical class A GPCR structure, consisting of seven trans-membrane helices, linked together by linker regions, with an eighth, amphipathic helix sitting horizontally on the intracellular side of the plasma membrane. NTS1 interactions with spin-labelled lipids have been observed as well as the modulation of GPCRs by lipids [286, 287]. The influence of lipids on ligand binding has also been studied [287]. However, the binding of PIP_2 lipids to NTS1 had not been studied up until recently.

Recent results by the Robinson laboratory [288], obtained through mass spectrometry, and complemented by MD simulations, indicate that PIP_2 lipids bind directly and specifically to NTS1 through a head dominant binding mode, mediated by the terminal phosphoryl groups of PIP_2. A number of interaction hotspots for the PIP_2 head groups have been identified along with the residues involved in binding [288].

The wealth of experimental and computational data on lipid binding sites on NTS1, along with the importance of GPCRs as drug targets, prompted the use of NTS1 as a second test case example for the lipid binding site detection capabilities of the dynamic pharmacophore probes approach.

5.1.4 Adenosine A_2A Receptor

Adenosine A_2A receptor is a class Aα GPCR which also plays a major role in the central nervous system. It is responsible for the binding of adenosine and the activation of adenylyl cyclase which in turn induces the intracellular synthesis of cAMP [289]. It has also been shown to have an important role in glutamate and dopamine release in the brain, making it an important drug target for diseases like depression and Parkinson's [289].

In a recent study by Song et al. [290], the binding sites of three different lipid species, PIP_2, GM3 and cholesterol were identified using MD simulations. Potential

of mean force (PMF) calculations were also used to try and elucidate the energy landscape of PIP_2 binding in the hope of identifying binding sites of potential allosteric modulators.

The importance of A_2A as a potential drug target, along with the lipid binding site data available, recommended its use as the third test case example. In this chapter, the dynamic pharmacophore probes approach will be used to try and recapture the binding sites identified for cholesterol and PIP_2 lipids on the A_2A receptor by Song et al. and other computational studies.

5.1.5 Smoothened

Smoothened is a class F GPCR which plays a role in controlling developmental processes through the hedgehog signalling pathway [291]. Despite its very low sequence identity to class A GPCRS, Smoothened still shows remarkable structural conservation, the canonical domain of seven trans-membrane helices, found in GPCRs. It also includes an extracellular cysteine rich domain (CRD) which is connected to the trans-membrane region through a small linker domain [292].

Experimental data, and more specifically the elucidation of the full structure, showed the binding of cholesterol to the CRD domain [293]. Moreover, there is data (unpublished) suggesting that cholesterol has the ability to bind to the trans-membrane domain due to the fact that cholesterol has been co-purified with constructs lacking the CRD and linker regions. Given these indications, Hedger et al. [294] undertook a series of all atom and CG simulations which identified potential binding sites for cholesterol on smoothened and observed lower flexibility of the CRD domain in the presence of cholesterol.

As a more "difficult" test case for the dynamic pharmacophore probes approach, we tried to recapitulate the cholesterol binding sites mentioned previously, using this method.

5.1.6 Polycystin-2

Polycystin-2 (PC2) is a TRP-nonselective, Ca^{2+}-permeable cation channel [295]. It is composed of four monomers, forming a homotetramer. It is regulated by a variety of stimuli, including cations [296], pH [297], voltage [297] and phosphorylation [298].

Mutations in the PKD2 gene, which encodes PC2, are responsible for the 15% of ADPKD cases, a cyst form, end stage renal disease, affecting 1 in 800 births [298]. This makes PC2 a prime drug target for the treatment of ADPKD.

Recent data by Wang et al. (manuscript in preparation) suggest the existence of a PIP_2 and a cholesterol binding site within the interface of each subunit of PC2, for a total of four PIP_2 and cholesterol binding sites in total. This example prompted the use of PC2 as the last example for this chapter on which to use the dynamic pharmacophore probes approach for the detection of lipid binding sites. Figure 5.3 shows the structure of all test case examples used in this chapter, after Kir2.2.

5.2 Methods

5.2.1 Modelling and System Set Up

The simulations in this chapter were performed using the GROMACS v 5.0.2 simulation package with GPU acceleration [188], as described in the previous chapters. The same set up simulation was used in order to examine the ability of this approach to recapitulate PIP_2 binding sites on membrane proteins.

The initial protein coordinates used in the simulations were obtained from the PDB. Kir2.2 (pdb id: 3spi), NTS1 (pdb id: 4buo), A_2A (pdb id: 3eml), smoothened (pdb id: 5l7d). The polycystin-2 structure was provided by the authors of the paper. Any missing atoms in the structures were completed using MODELLER [223]. The structures were energy minimised using the steepest descent algorithm and the MARTINI forcefield v2.2 [210] was used to convert the structures into their CG representations. The simulation protocol described in chapter 3 was used. It should be noted that the simulations were run without the target being embedded in a lipid bilayer, which is the usual set up when it comes to identifying lipid binding regions on membrane proteins. Table 5.1 shows a summary of the simulations run.

The MARTINI forcefield v2.2 was used to perform the CG molecular dynamics simulations, with a time step of 20 fs. Isotropic Berendsen pressure coupling [179] was used with a coupling constant of 12 ps. The reference pressure was set at 1 bar for the XY and Z directions. A stochastic rescaling temperature control was used [182], with a characteristic time step of 1 ps, keeping the temperature constant at approximately

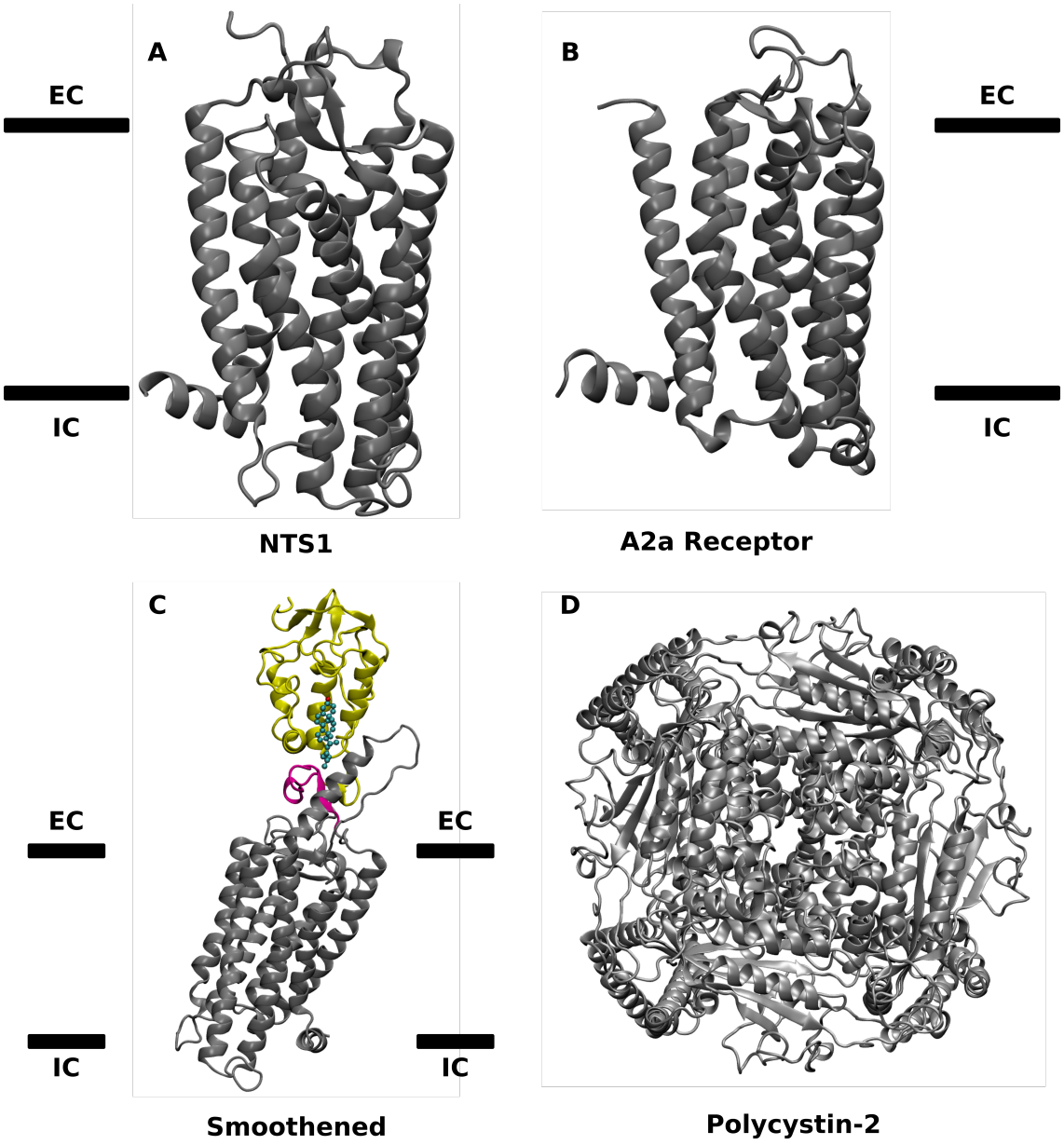


Figure 5.3: The structures of each of the test case examples used. (A) Neurotensin 1 (NTS1) receptor (4BUO) (B) A_2A Receptor (3EML) (C) Smoothened (5L7Dd) and (D) Polycystin-2 (in house structure). In panel (C) the CRD domain is coloured in yellow with the linker domain coloured in pink and the trans-membrane domain coloured in silver.

Table 5.1: A table of the targets used, the number of simulations performed and the total simulation time.

Target	Number of Simulations	Total Simulation Time
Kir2.2 (3SPI)	3	3 x 2μs
NTS1 (4BUO)	3	3 x 2μs
A_2A (3EML)	3	3 x 2μs
Smoothened (5L7D)	3	3 x 2μs
Polycystin-2 (In-house)	3	3 x 2μs

323 K. A cut-off of 1.2 nm was used for Lennard-Jones (LJ) interactions with the potential being smoothly switched off between 0.9 nm and 1.2 nm. The Coulombic interactions were calculated using the Reaction Field method with a cutoff of 1.1 nm. These settings were consistently used throughout the simulations, across all the targets.

The MDAnalysis [224] software package was used to calculate the interaction densities of the different probe types. Analysis of the simulation data was conducted using the VMD and PyMol visualisation packages along with tools implemented in GROMACS and in house scripts. The same analysis protocol as described in chapter 3 was followed in order to transform the binding densities into pharmacophoric points and identify the surrounding residues.

Due to the negatively charged head group of PIP_2 and the hydrophobic nature of its tail, the interactions of the negatively charged and hydrophobic probes were deemed to be of particular interest. Therefore, even though a mixture of probes was used in the simulations, they will be the ones presented here.

One difference compared to the analysis protocol described in the previous chapters, is the use of a cavity detection algorithm. In this case, the already predicted PIP_2 binding sites were used to define an area of interest. The already bound lipids were used as probes and an area of $4\mathring{A}$ around them, was defined, instead of a cavity.

5.3 Results

5.3.1 Kir 2.2

As mentioned previously, Kir 2.2 PIP_2 binding sites have been accurately identified both by experimental methods and multiscale molecular dynamic simulations. More specifically, the regions binding the head groups of the PIP_2 lipids have been shown to be located on the N-terminal end of the first trans-membrane helix (TM1) and the C-linker which connects the trans-membrane domain (TMD) to the cytoplasmic domain (CTD).

These PIP_2 head group binding regions form a cluster of basic residues, which is to be expected due to the charge on the phospho-inositol head group. Using the 3spi crystal structure, the residues forming these clusters were shown to be R78, W79 and R80, on the TM1 helix and K183, R186, K188 and K189 on the C-linker region.

The dynamic pharmacophore probes which best match the properties of the PIP_2 head groups would be the negatively charged probes with the hydrophobic probes matching the hydrophobic properties of a lipid tail. As a result, these two types of probes were the ones used to identify lipid binding regions.

Looking at the Kir 2.2 structure which was solved with PIP_2 lipids bound, the first step in the analysis was to determine whether the negatively charged probes identified the correct regions of PIP_2 head group binding. This was done by overlaying the binding densities of the probes with the lipid bound structure (Figure 5.4). As it can be seen from figure 5.4, the binding densities of the probes match with the positions of the charged PIP_2 head groups and have the ability to recapitulate the binding positions of the phospho-inositol moieties.

To further focus into the PIP_2 head group binding region, the analysis protocol described previously was used to transform the binding densities of the probes into pharmacophoric points and identify the residues being in contact with them. As before, residues found within 4 Å of a pharmacophoric point were considered to be in contact with it. This was done in order to identify the residues found to be in contact with the negatively charged pharmacophoric points. Panel A of figure 5.5 shows the residues identified along with the binding densities of the probes and a PIP_2 lipid bound on Kir 2.2 while panel B shows a zoomed in version of the binding

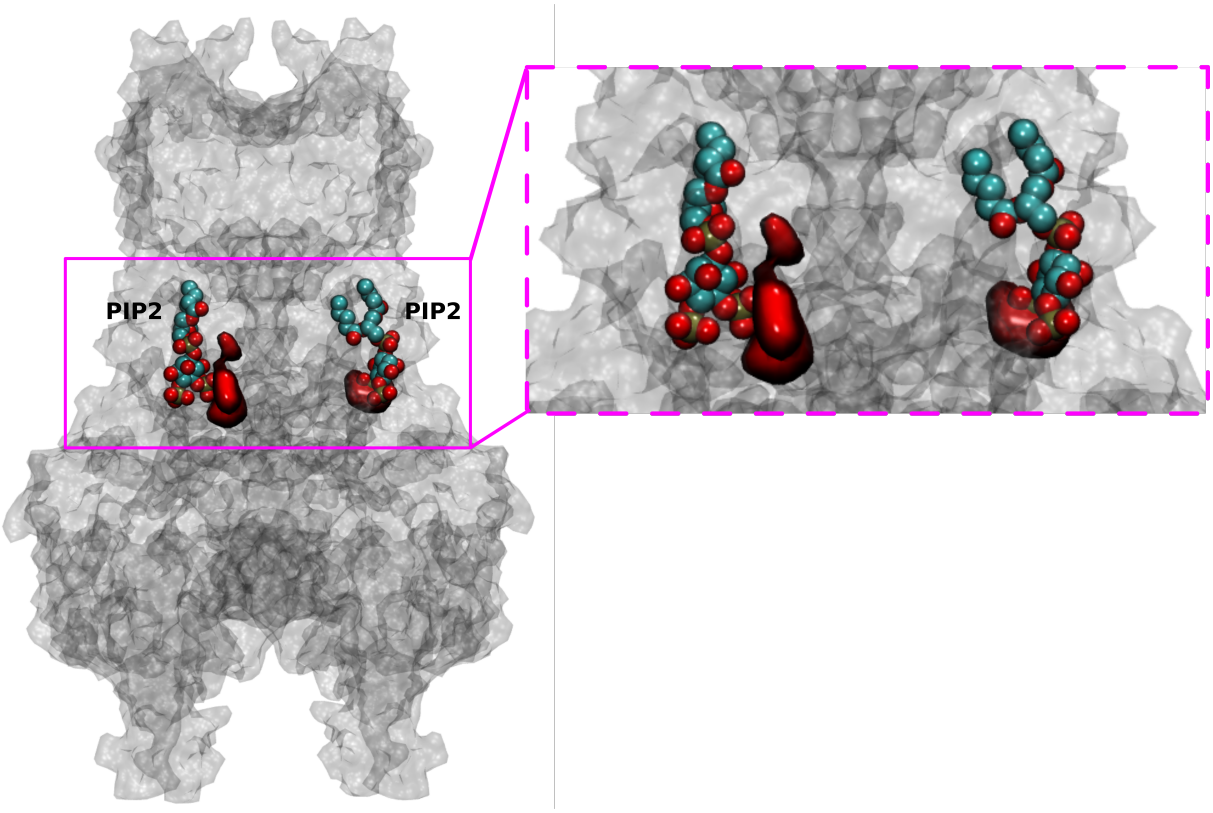


Figure 5.4: A figure of the structure of Kir 2.2 with two of the four PIP_2 lipids bound and the binding densities of the negative charged probes overlaid on top of the PIP_2, coloured in red. A zoomed in version is shown on the right panel.

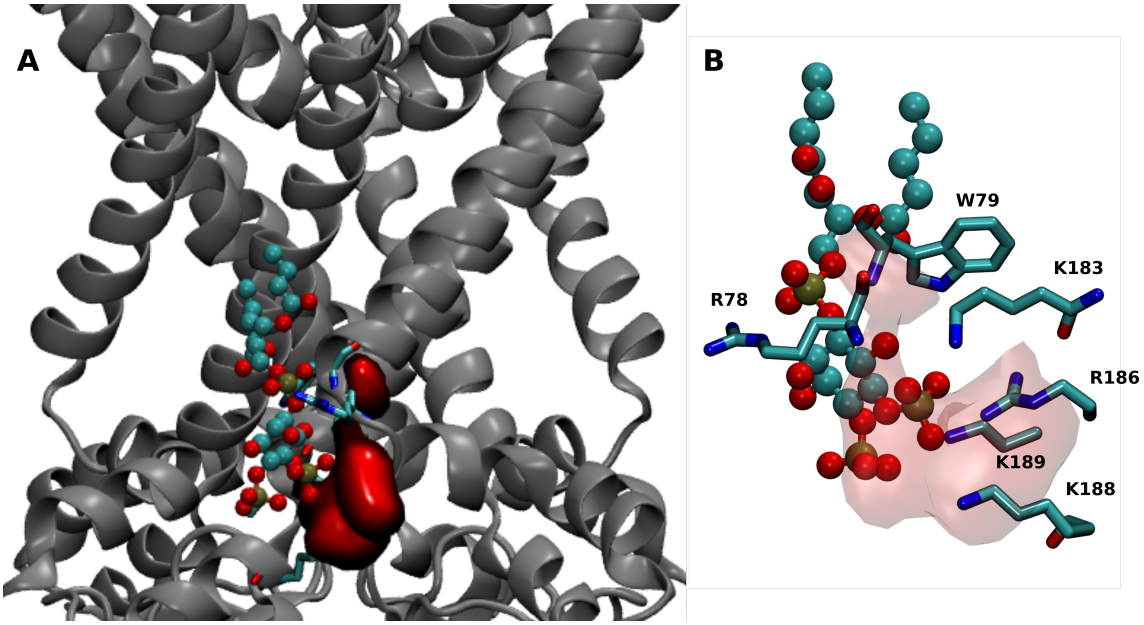


Figure 5.5: A figure of the (A) Kir 2.2 residues identified as being in contact with the negative charged probes' binding densities, along with the binding densities of the probes and a PIP_2 lipid bound on Kir 2.2 (B) A zoomed in version of the bound PIP_2 and the residues identified, with the binding density of the probes being shown in a semi transparent shade. The protein structure has been omitted for clarity.

site, again with the residues identified and the lipid. The protein has been omitted for clarity in this panel.

As can again be seen from figure 5.5 (A) and (B), the binding density of the probes matches the PIP_2 head group moieties. Furthermore, the residues identified as being in contact with the probes were R78, W79 on the TM1 region and K183, R186, K188 and K189 on the C-linker region. These residues are the same basic residues identified by Schmidt et al. in their multi scale simulations, suggesting that the dynamic pharmacophore probes have the capacity to identify PIP_2 head group binding regions.

The only residue identified by the multi scale MD simulations but not identified by this approach was R80. This can be explained by the fact that R80 is located on the outer part of TM1, facing outwards, and being oritented into position by interacting with the charged phosphate group on the first position of the phosphatidylinositol when PIP_2 binds to the structure. Although the structure used for the simulations was the one with PIP_2 bound, meaning R80 was correctly oriented, the way this approach is currently parametrized, it relies on the identification of potential binding cavities. The location of R80 on the outside of TM1 and not in the more enclosed

cluster, therefore, could be the reason it was not identified. Furthermore, the 4 Å distance cut off was set in order to focus on the immediate vicinity of potential binding sites. With R80 being located slightly further up on TM1, a higher cut off might be necessary in this case.

PIP_2 binding on Kir 2.2 is mainly driven by its head group, with the tails exhibiting greater flexibility. They do, however, bind to the interface between the TM1 of one monomer and the rest of the trans-membrane helices of its neighbouring monomer. With that in mind, the hydrophobic probes were also used to examine the binding of the lipid tails of PIP_2. As shown in figure 5.6, the binding densities of the hydrophobic probes match the binding crevice of one of the tails of PIP_2, suggesting that the probes have the potential to identify the tail moieties of lipids along with potential binding sites of the lipid tails, as well as the head group binding sites. Figure 5.6 B shows the hydrophobic and negatively charged probe binding densities and how they match the bound PIP_2 lipid.

Kir 2.2 provided an initial, well studied test case, on which the lipid binding site prediction capabilities of the dynamic pharmacophore probes were validated. Based on these results, further targets were subsequently explored and tested.

5.3.2 Neurotensin Receptor 1

Moving on to another test case, NTS1 was used to evaluate the ability of the probes to predict lipid binding sites. As mentioned previously, a range of mass spectrometry data and long MD simulations, either with a single structure or multiple structures embedded in membrane patches of different compositions, of the 5 μs to 100 μs range, for PIP_2 binding on NTS1, are available. These data mainly focus on the identification of PIP_2 head group binding sites, so the negatively charged probes were again used to identify these sites.

The sites pinpointed by the extended MD simulations as having the highest number of PIP_2 head group contacts were located between TM1, TM2 and TM4, TM3 and TM4 and TM6 and TM7. The residues with the highest number of contacts were found to be F39, R43, K44 and K45, on TM1 and K128, K130, L132, M133, S134, R135, S136, R137, K139, K140, and F141 on TM4. R239 and L305 and R311 were also identified on TM6 and TM7 respectively.

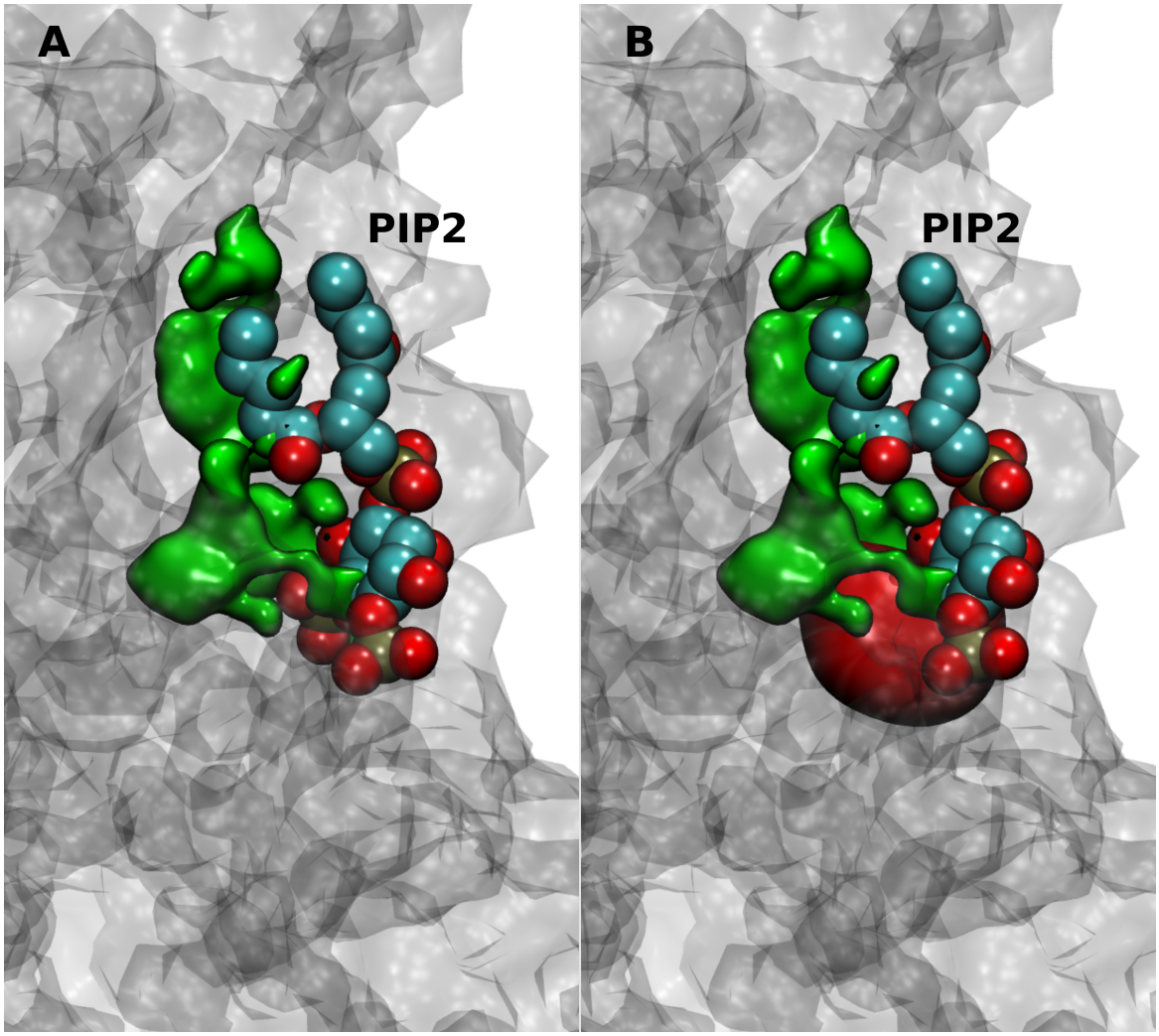


Figure 5.6: A figure of a close up view of Kir2.2 with a bound PIP_2 lipid and the binding density of the hydrophobic probes coloured in green (A) along with the binding density of the negatively charged probes coloured in red (B). The protein structure is shown in grey.

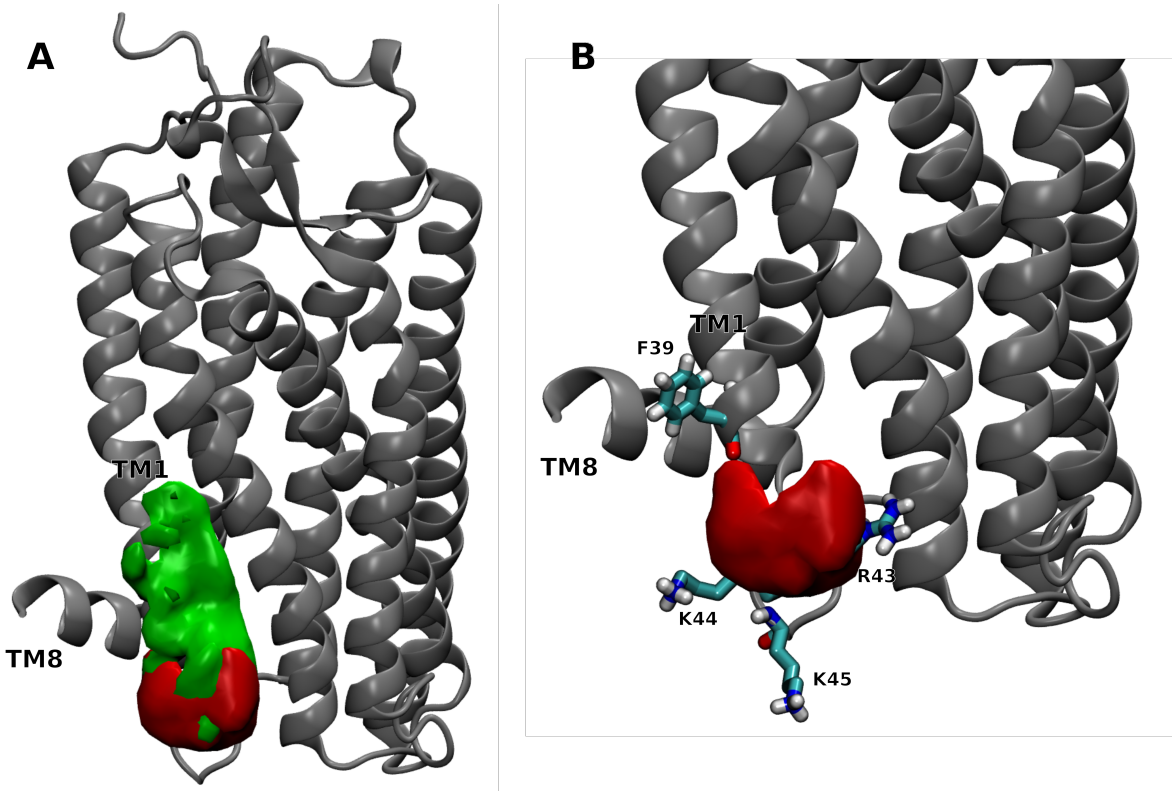


Figure 5.7: A figure of (A) the binding densities of the hydrophobic (green) and negatively charged (red) probes on the NTS1 receptor between TM1 and TM2 (B) a close up view with the residues identified as being in contact with the negatively charged binding densities of the probes.

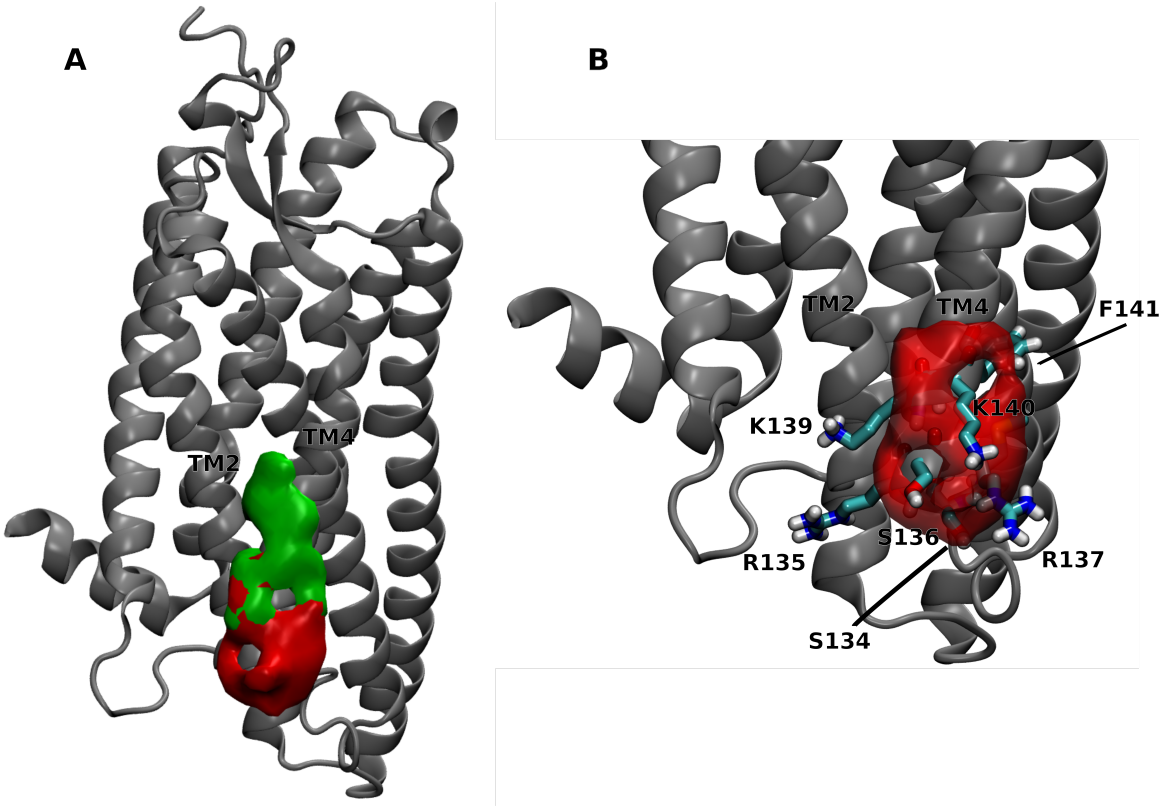


Figure 5.8: A figure of (A) the binding densities of the hydrophobic (green) and negatively charged (red) probes on the NTS1 receptor on TM4 (B) a close up view with the residues identified as being in contact with the negatively charged binding densities of the probes.

Figure 5.7 B shows the residues of TM1 that were identified as being in contact with the negatively charged probes' binding densities (shown in red). These were F39, R43, K44 and K45. These residues agree with the residues identified by the extended MD simulation. Moreover, figure 5.7 A shows the binding density of the hydrophobic probes, coloured in green, which is located in the crevice between TM1 and TM2, which could act as a binding site for a PIP_2 tail. Snapshots from the long MD simulations mentioned previously has shown this region to be occupied by the lipid tails of bound PIP_2s. It is also worth noting that R43, K44 and K45 are basic residues, forming a basic cluster, similar to the one observed in the case of Kir 2.2.

After examining TM1, TM4, the area with most residues shown to form PIP_2 head group contacts, was examined. Figure 5.8 shows the residues identified as being in contact with the the negatively charged probes, which in this case are S134, R135, S136, R137, K139, K140 and F141. As previously, these residues match the residues identified by the extended MD simulations, in these regions. However, in

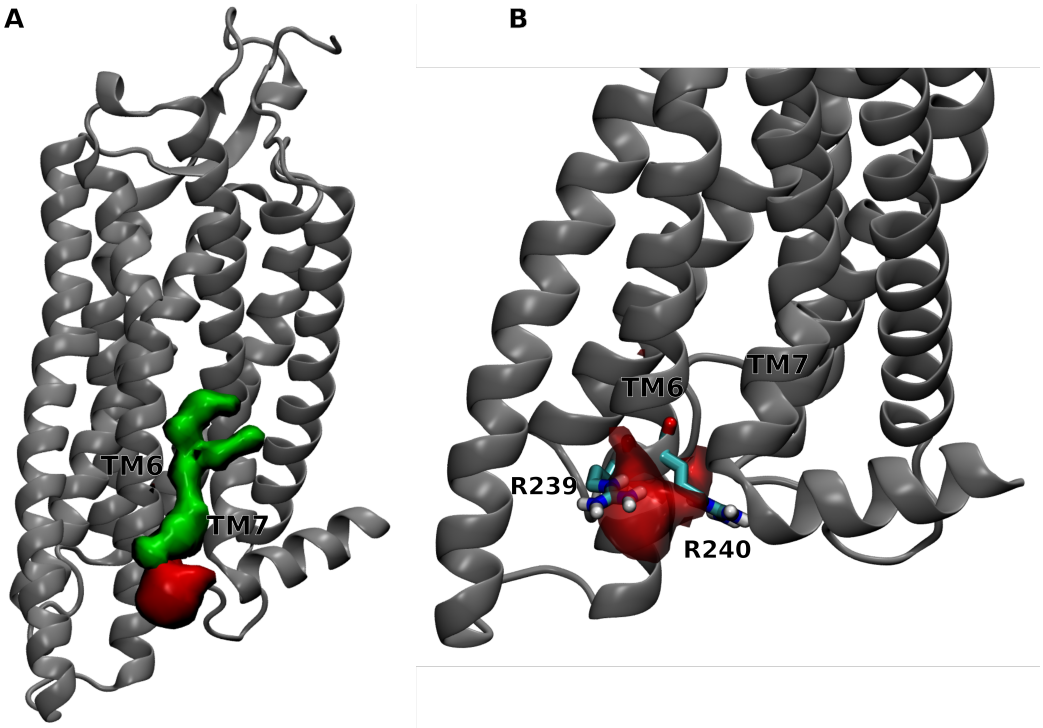


Figure 5.9: A figure of (A) the binding densities of the hydrophobic (green) and negatively charged (red) probes on the NTS1 receptor on TM6 (B) a close up view with the residues identified as being in contact with the negatively charged binding densities of the probes.

this instance, K128, K130, L132 and M133 were not identified by the probes, in contrast with the extended simulations. Examining the structure, it was observed that these residues are located on the linking, intra-cellular loop which connects TM3 and TM4. The increased flexibility of the intra-cellular loop regions, compared to the rest of the structure, even in a CG MD simulation, could explain the lower binding of the probes in this region, which would in turn result into a reduced binding density in that area, leading to the residues not being identified. Furthermore, as in the case of R80 of Kir 2.2, an extended distance cut-off might be required. Finally, figure 5.8 shows the binding density of the hydrophobic probes for this region, again suggesting a potential binding crevice for the lipid tail of PIP_2, which has been shown to bind PIP_2 tails, based on the extended MD simulation data. Once more, the regions binding the PIP_2 head group appear to contain a higher number of positive charged residues, clustered together.

The next PIP_2 head group binding site was pinpointed on TM6, R239 more specifically. As shown in figure 5.9, the residues identified by the negatively charged probe binding densities were R239 and R240, matching the residue identified by

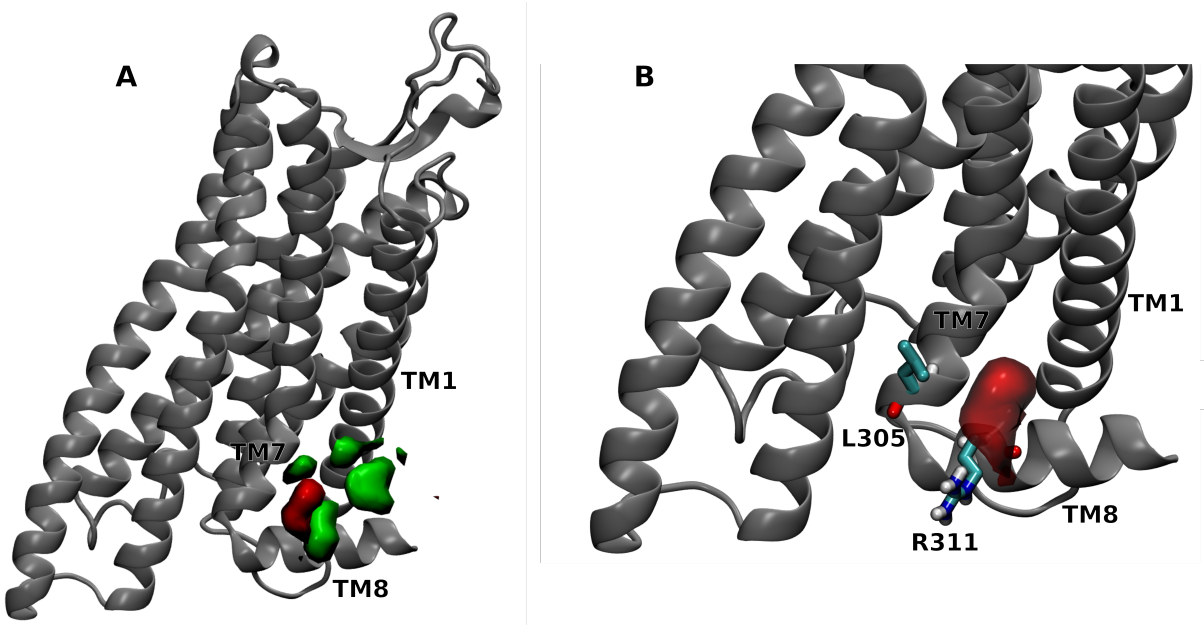


Figure 5.10: A figure showing (A) the binding densities of the hydrophobic (green) and negatively charged (red) probes on the NTS1 receptor between TM7 and TM8 (B) a close up view with the residues identified as being in contact with the negatively charged binding densities of the probes.

the extended MD simulations. R240 was not identified by these simulations but its selection by the probes can be explained due to its proximity to R239 and it's orientation towards the inside of the crevice, where the probes can penetrate, in contrast with a bulkier PIP_2 head group. Furthermore, the hydrophobic binding densities, shown in green in figure 5.9 A have identified a potential binding crevice between TM6 and TM7, for the PIP_2 lipid binding tail. This crevice has also been shown, in snapshots of the extended MD simulations, to be occupied by a PIP_2 tail.

The final PIP_2 head group binding region on NTS1 was shown to be on the region between TM7 and TM8 with residues L305 and R311 being the ones forming the most contacts with PIP_2 head groups. Looking at figure 5.10 B it can be seen that L305 and R311 were the residues identified to be in contact with the binding densities of the negatively charged probes. Furthermore, the binding crevice on top of TM8, has been shown to bind the second lip tail of the PIP_2 with the head group binding to the TM6, TM7 and TM8 basic residues cluster, with the tail binding in a parallel manner to TM8. In figure 5.10 A, some binding density of the hydrophobic probes is observed, indicating this region as a tail binding spot, but with less frequency compared to the previous examples. As mentioned before, this could be due to the

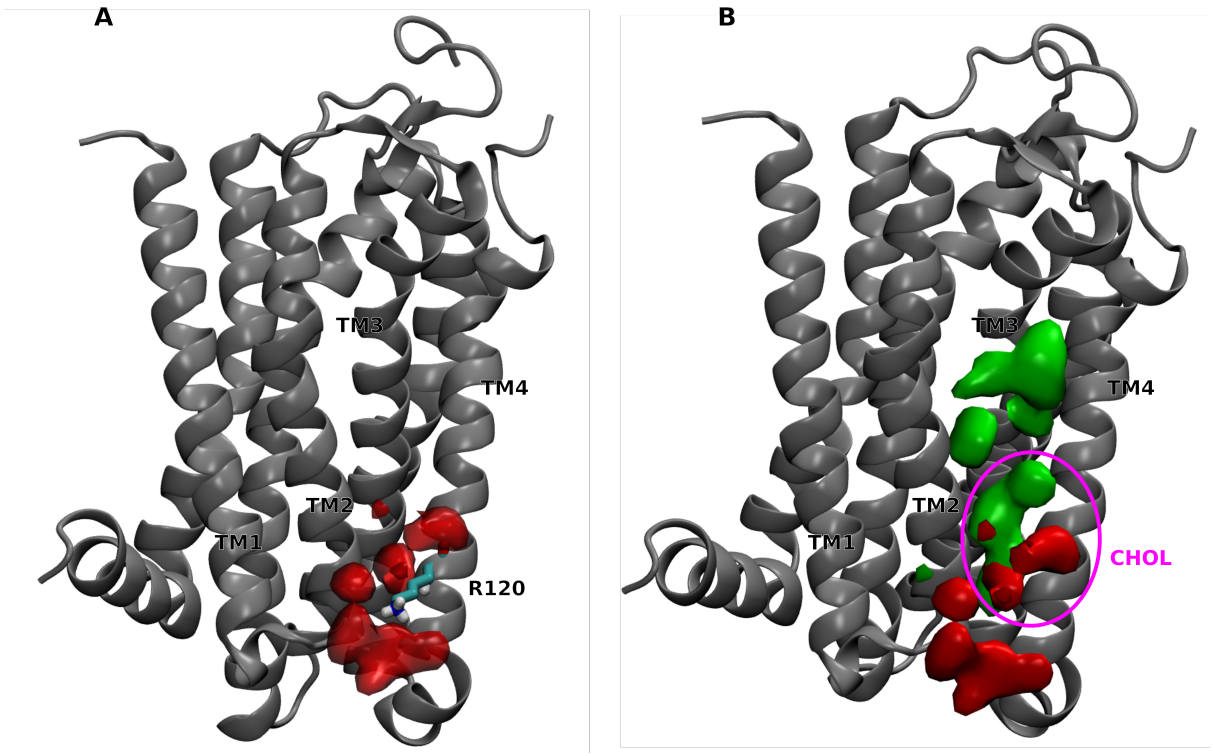


Figure 5.11: A figure showing (A) a close up view with the residues identified as being in contact with the negatively charged binding densities of the probes (B) the binding densities of the hydrophobic (green) and negatively charged (red) probes on the A_2A receptor between TM2, TM3 and TM4.

fact that this binding site is located on the edge of the truncated TM8, which could exhibit higher flexibility during the simulation, leading to lower probe binding.

Overall, from the examples presented above, the dynamic pharmacophore probes appear to have the potential to predict PIP_2 head group binding hotspots with almost the same degree of accuracy as extended MD simulations. Even more, they can also identify potential binding sites for the lipid tails although there is still progress to be made in order to be able to accurately identify these regions.

5.3.3 A_2A Receptor

The next example, including both PIP_2 and cholesterol binding sites, is the A_2A receptor. Extended CG MD simulation data (around 80 μs), of the receptor in a membrane, were used to predict PIP_2 and cholesterol binding sites, against which the predictive powers of the dynamic pharmacophore probes were compared.

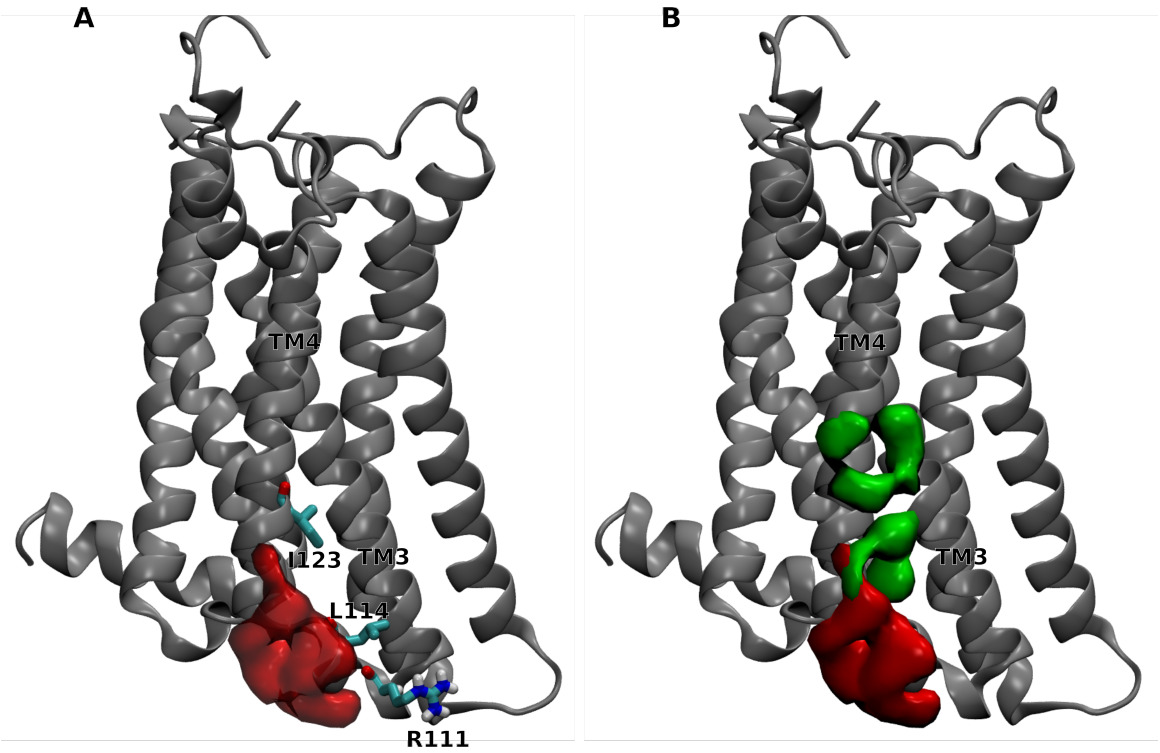


Figure 5.12: A figure showing (A) a close up view with the residues identified as being in contact with the negatively charged binding densities of the probes (B) the binding densities of the hydrophobic (green) and negatively charged (red) probes on the A_2A receptor between TM3 and TM4.

PIP_2 Binding Sites

The first spot identified by the dynamic pharmacophore probes approach as a potential PIP_2 head group binding region was located on TM4, being R120 (figure 5.11 A). This region matches quite well with the region identified in the longer time-scale simulations. Moreover, the potential tail binding region identified by the hydrophobic probes binding densities also matches quite well with the region shown by the extended simulations to be binding the PIP_2 tails. Furthermore, the region marked by the magenta circle in figure 5.11 B, also matches a cholesterol binding region identified in the longer simulations.

The next PIP_2 head group binding spot identified was located between TM3 and TM4. The residues identified as being in contact with the binding probe densities were R111, L114 and I123 (figure 5.12 A). R111 matches the lysine identified in that region by the extended simulations while the hydrophobic L114 and I123 were probably identified due to their close proximity to R111 and the binding probe densities, along with their orientation, which put them within the radius of detection.

It should be noted that the hydrophobic binding densities (figure 5.12 B) match exceptionally well with the densities of the PIP_2 lipid tails shown to be binding to that crevice, by the extended MD.

The next region of PIP_2 binding identified was between TM4 and TM5. Figure 5.13 is showing the residues identified by the dynamic pharmacophore probes approach, R107, R187 and R195. As before, these residues match the basic cluster identified by the long time scale membrane simulations in this region. The hydrophobic probes do not seem to have identified the binding crevice of the PIP_2 lipid tails in this case (figure 5.13 B). As seen in previous examples, the region identified by the hydrophobic probes is found on the outside of TM5, which could explain the lower density values.

The final PIP_2 binding region identified on A_2A was located between TM5, TM6 and TM7. The contact residues identified were R195, K204, V206, K210 and R268. These residues form another basic cluster, suitable for the binding of the charged PIP_2 head group, and again match the cluster identified by the extended simulations (figure 5.14 A). The exception in this case is V206 which was most likely identified due to its proximity to the cluster and its orientation. The hydrophobic probe densities (figure 5.14 B) accurately identify the potential PIP_2 tail binding site, when compared with the extended MD. This site also matched another cholesterol binding site, shown in the magenta circle in figure 5.14 B, identified by the extended simulations.

Cholesterol Binding Sites

There were seven potential cholesterol binding sites identified on A_2A, by the extended CG MD simulations. Out of the seven, three were located on the surface of transmembrane helices, which means that they would not have been able to be recaptured due to the binding cavity detection aspect of the dynamic pharmacophore probes approach. Of the remaining four sites, two were shown to be recapitulated by the hydrophobic pharmacophore probes previously (figures 5.11 B and 5.14 B).

The remaining two cholesterol binding sites were located between TM1 and TM2 and between TM3, TM4, TM5 and TM6. Figure 5.15 A, shows that the hydrophobic probe densities identified in the TM1 and TM2 region match the density of the

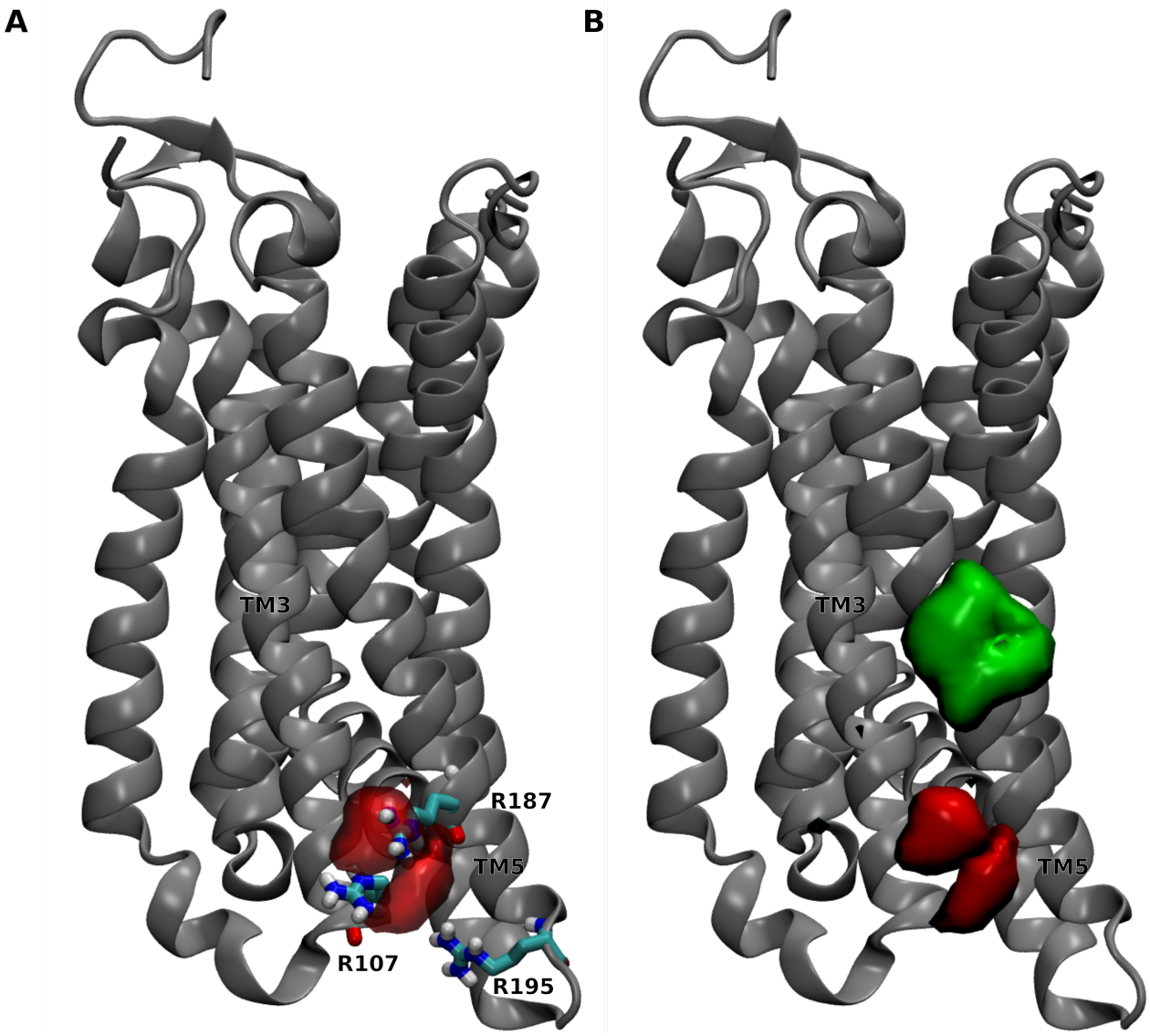


Figure 5.13: A figure showing (A) a close up view with the residues identified as being in contact with the negatively charged binding densities of the probes (B) the binding densities of the hydrophobic (green) and negatively charged (red) probes on the A_2A receptor between TM4 and TM5.

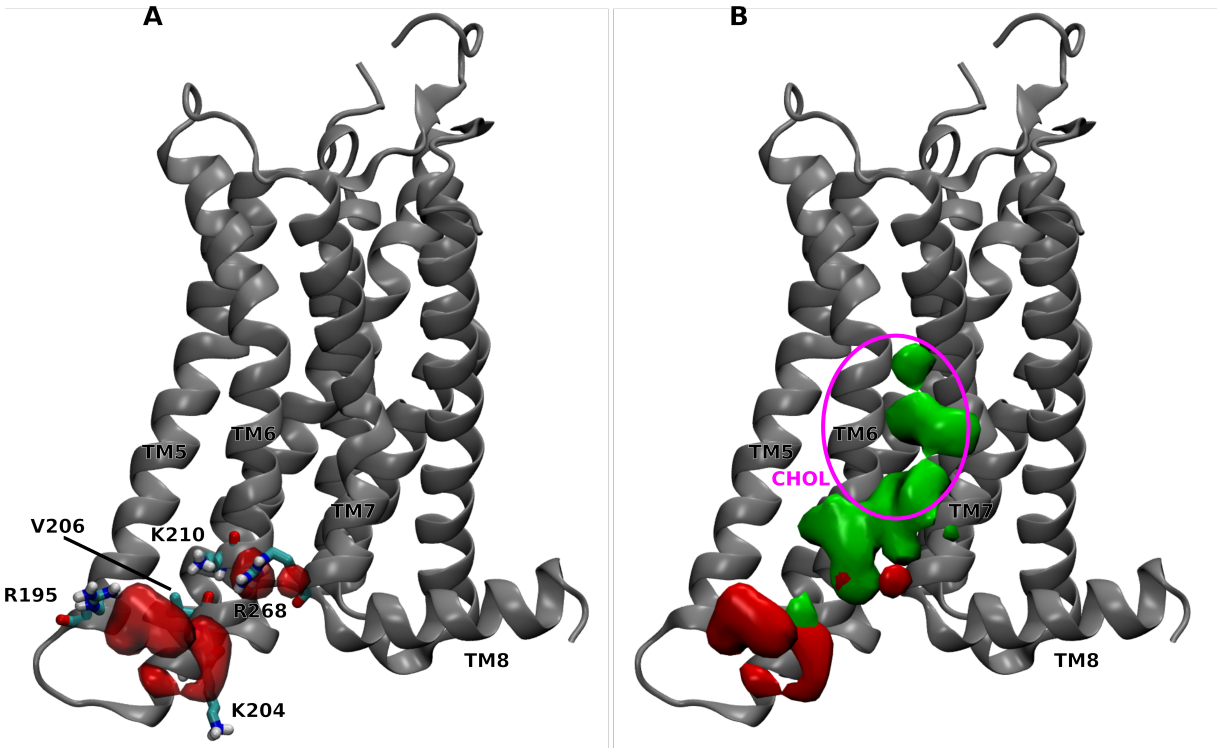


Figure 5.14: A figure showing (A) a close up view with the residues identified as being in contact with the negatively charged binding densities of the probes (B) the binding densities of the hydrophobic (green) and negatively charged (red) probes on the A_2A receptor between TM5, TM6 and TM7. The magenta circle indicates a potential cholesterol binding site.

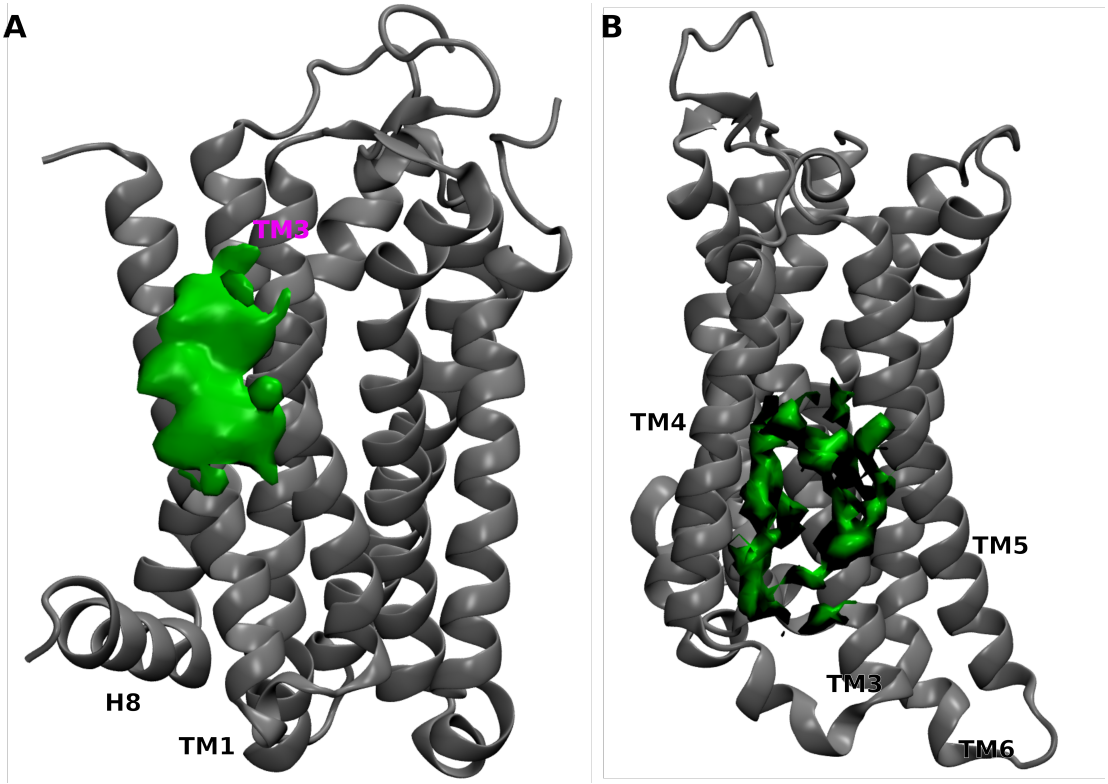


Figure 5.15: A figure showing (A) a potential cholesterol binding site between TM1 and TM2 (B) another potential cholesterol binding site between TM3, TM4, TM5 and TM6 on A_2A. Hydrophobic probe binding densities are shown in green.

cholesterol molecule shown to bind in that area, by the long CG MD simulations. In the TM4, TM5 and TM6 region, the hydrophobic density was more widespread and less well defined in the crevice. Looking at the extended simulations, the cholesterol seems to be bound in the area between TM3 and TM4 with the head group anchored on TM3 and the hydrophobic part extending down to the intracellular edge between TM3 and TM4. The binding density seen in figure 5.15 B, therefore, agrees with the observed binging site, however, it does not recapitulate the correct binding pose. The widespread hydrophobic density could be due to the large number of hydrophobic residues lining the inside of the two crevices, making it hard for the probes to accurately differentiate between them.

5.3.4 Smoothened

As described in the introduction section of this chapter, smoothened consists of a CRD domain, a linker region and a trans-membrane domain (figure 5.16). The CRD has been known to bind cholesterol, with a cholesterol bound structure, on the

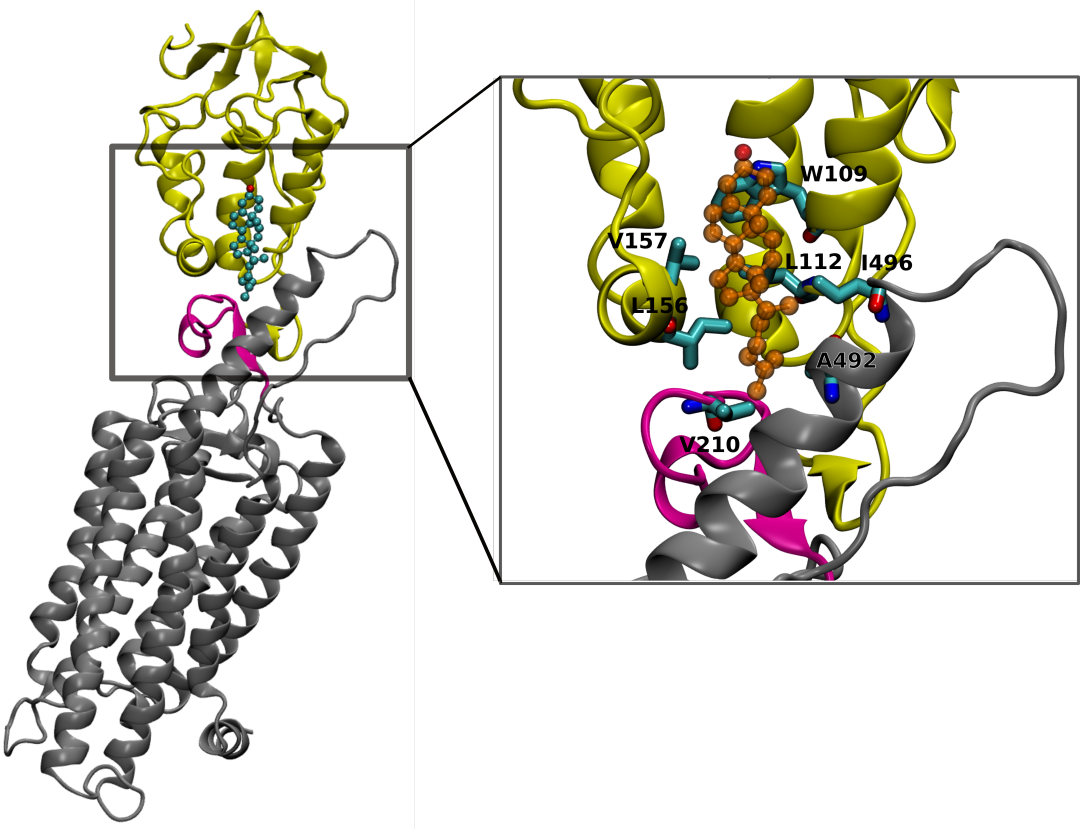


Figure 5.16: A figure showing the different domains of smoothened. CRD domain is coloured in yellow, with a cholesterol molecule bound, linker region is coloured in pink and the rest of the receptor is coloured in silver. On the right panel, a close-up of the bound cholesterol molecule (orange) is shown, along with the residues forming the hydrophobic "cage" of the CRD domain.

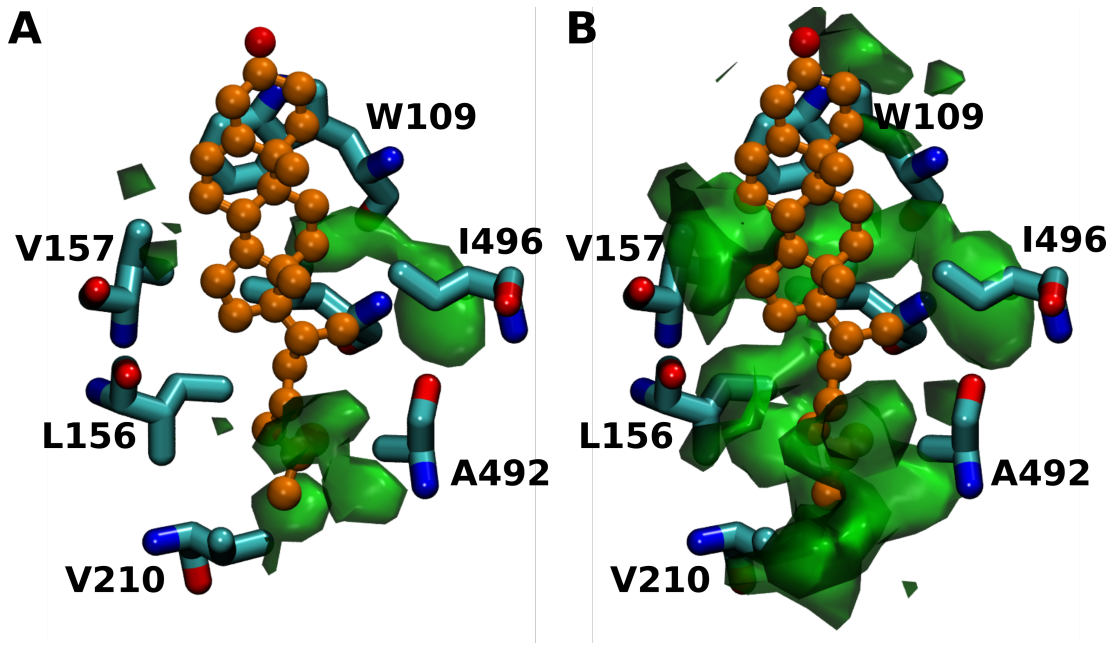


Figure 5.17: A figure showing the hydrophobic "cage" residues of the CRD domain, identified by the binding densities of the hydrophobic probes (green) at a (A) higher and (B) lower contour level. The bound cholesterol is shown in an orange ball and stick representation.

CRD, having been elucidated in the past. Cholesterol is bound to the CRD domain through the hydrophobic interactions formed by a series of hydrophobic residues, shown in the right panel of figure 5.16, also known as the hydrophobic "cage" of CRD. These residues include W109, L112, L156, V157, V210, A492 and I496. Data coming from atomistic simulations (500 ns) and CG MD simulations (80 μs) of the receptor embedded in membranes of different compositions, have identified four cholesterol binding sites on the trans-membrane domains of smoothened.

Firstly, the CRD domain was examined and the binding densities of the hydrophobic probes were overlaid with the bound cholesterol. Figure 5.17 shows the residues identified by the hydrophobic probe densities exactly matched those shown to form the CRD hydrophobic cage. In addition, using a lower contour, the binding densities were shown to recapitulate certain of the hydrophobic moieties of the cholesterol molecule. When a higher contour was used, the region and the cage residues were accurately identified.

Of the four trans-membrane cholesterol binding sites, two were found to be on the surface of trans-membrane helices, meaning they could not be identified using this approach. The remaining two were found to be between TM2 and TM3 and TM3 and

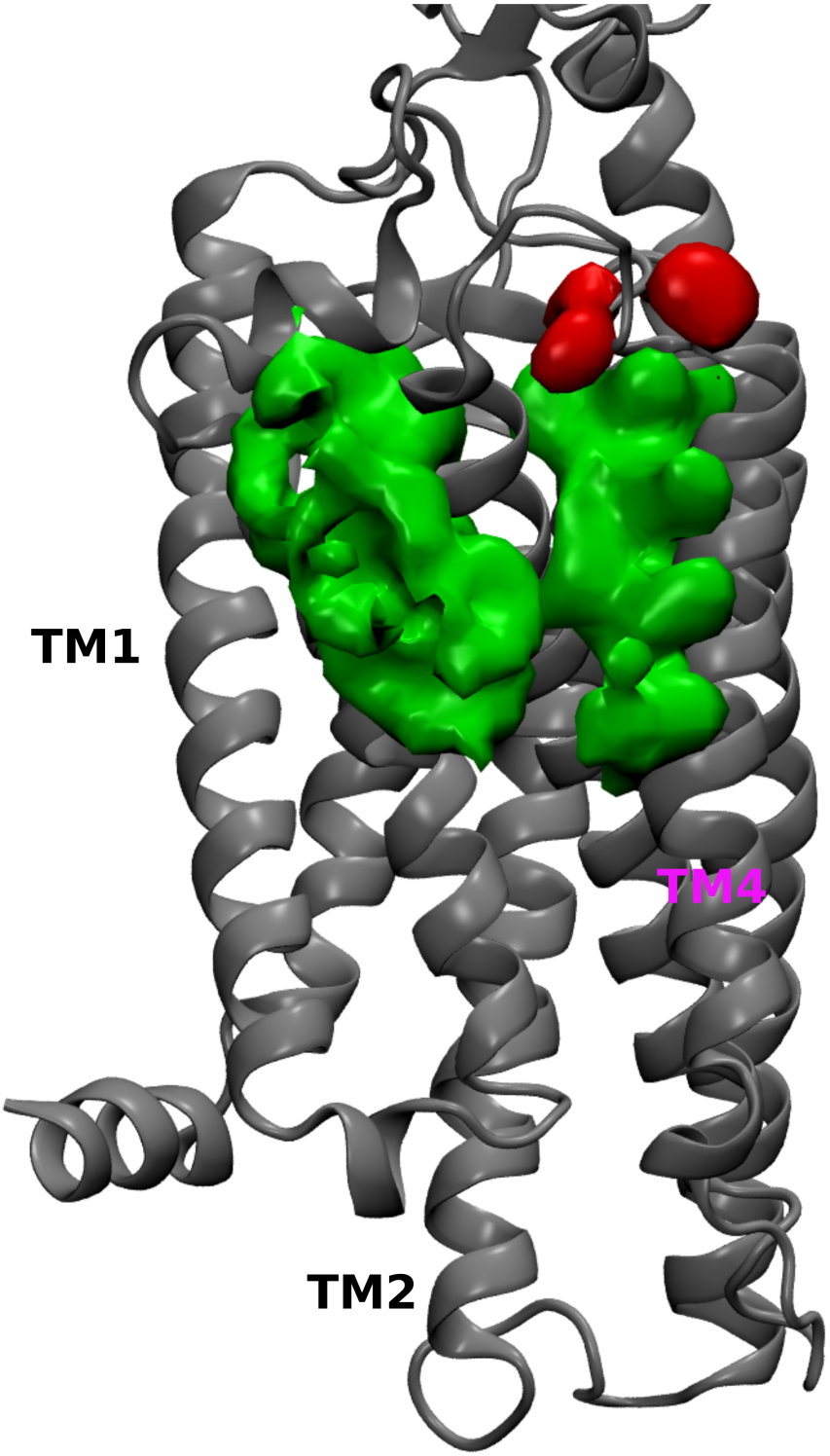


Figure 5.18: A figure showing the cholesterol binding spots on the TMD of smoothened, between TM2-TM3 and TM3-TM4. Hydrophobic probes are shown in green while negatively charged probes are shown in red.

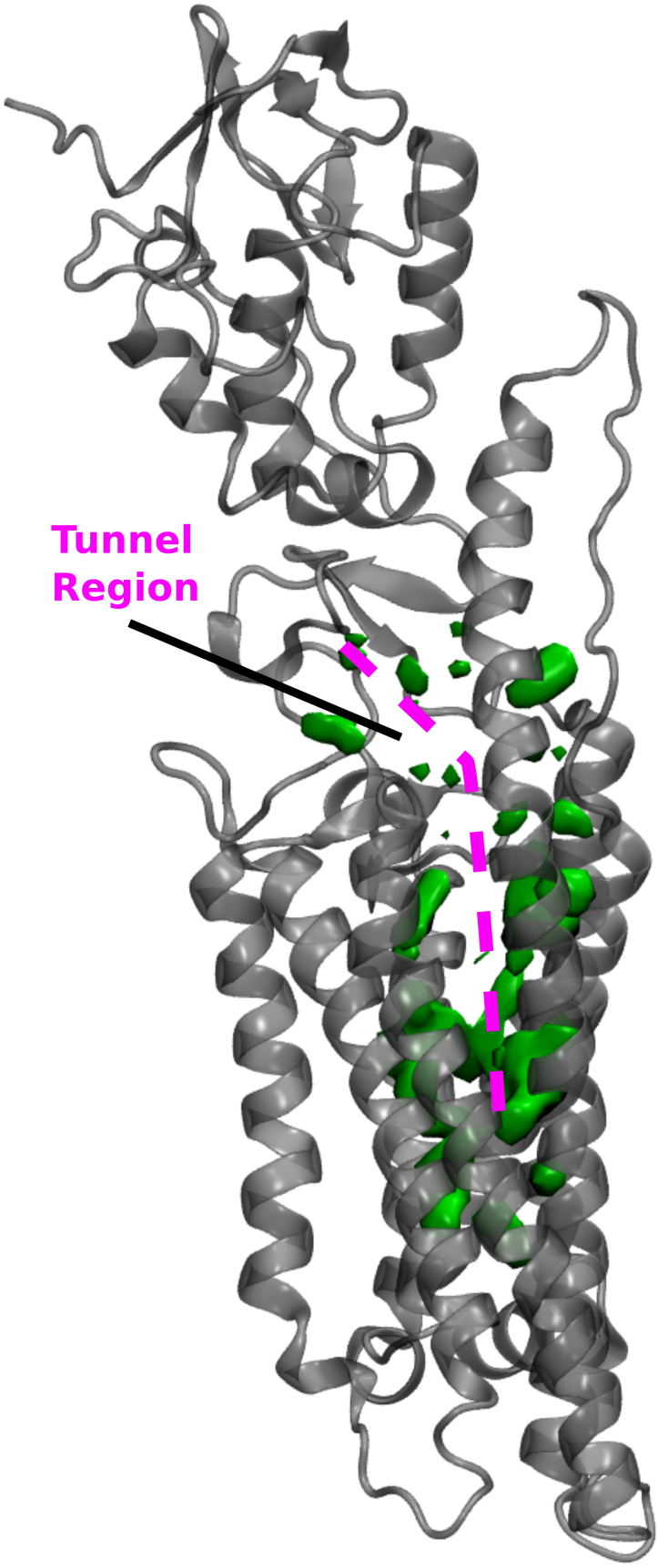


Figure 5.19: A figure showing the suggested tunnel region within smoothened, along with the interaction densities of the hydrophobic probes (green).

TM4. looking at figure 5.18 A, the regions identified by the binding densities of the hydrophobic probes match those identified by the extended AT and CG membrane simulations. Furthermore, on the extracellular region of TM4, a small region is identified by negatively charged probes. This region could serve as an anchoring point for the head of the cholesterol molecule.

The results of the pharmacophoric probes simulations were compared to those seen by Huang et al. [299]. Huang et al. propose the existence of a hydrophobic tunnel region, as shown in figure 5.19, that would allow the binding of hydrophobic ligands, such as sterols. To try and validate the existence of such a feature, the densities of the hydrophobic probes, in that region, were examined. Figure 5.19 shows that the hydrophobic probes, in green, are mainly found within the TM, ligand binding region of smoothened. However, they do not seem to be extending towards the extracellular region that is covered by the suggested tunnel nor towards the intracellular region between TM3 and TM6.

These results suggest that the existence of a purely hydrophobic tunnel might be unlikely. The lack of hydrophobic densities towards the extracellular region of the tunnel could be explained by the fact that this is a more solvent exposed region and the pharmacophoric probes might not be bulky enough to fill that region. However, the lack of hydrophobic densities towards the narrower, intracellular region between TM5 and TM6 supports an argument against the existence of the tunnel. As it is suggested by Huang et al., the movement of a wide range of residues on the intracellular region between TM3 and TM6 is required for the formation of this tunnel. It could therefore be argued that the structure used in these simulations, being different from the crystal structure presented in the paper (4O9R and 4NW4), and the CG nature of the simulations, might be the reason for the lack of hydrophobic densities in this region. As a result, extended all atom or CG simulations would be required as an initial step towards answering this question. Energy calculations of the proposed cholesterol tunnel would also be a good approach in order to try and determine the energy landscape of the proposed tunnel.

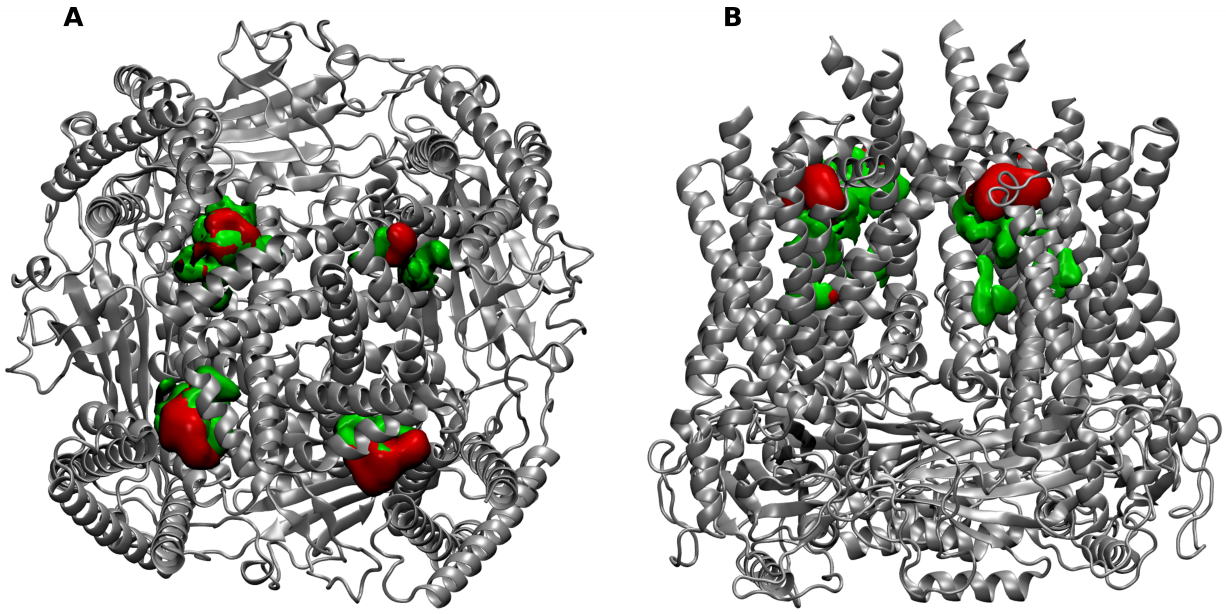


Figure 5.20: A figure showing (A) a top down view of PC2 and its potential PIP_2 binding sites (B) an inverted side view of two PIP_2 binding sites of PC2. Hydrophobic probe densities are shown in green and negatively charged probe densities are shown in red.

5.3.5 Polycystin-2

PC2 is a homo-tetramer with experimental and computational data (manuscript in preparation) suggesting potential PIP_2 binding sites on the interfaces between the monomers. Although limited data are currently available with regards to identifying the residues involved in the binding, PC2 was selected as an example to test the success of the method when limited information is available. A piece of evidence supporting the use of PC2 as a potential target is the fact that the PIP_2 binding site identified on PC2 corresponds to the same site seen to be binding PI lipids on TRPV1 channel structures.

When the regions of interest were examined, four sites matching the interaction patterns of a PIP_2 lipid binding site, seen in the previous examples, were identified, as shown in figure 5.20. Figure 5.20 A shows a top down view of the regions identified while figure 5.20 is showing a side view of two of the regions. These potential lipid binding sites match the areas identified by the initial simulation and experimental data (Wang, personal communications). The difference in the amount of binding densities observed in these regions could be explained by the fact that not all four binding sites have identical size and side chain orientations, leading to small

differences in their accessibility to the dynamic pharmacophore probes. Nevertheless, in all four cases, a PIP_2 head group region was identified with a potential crevice for the binding of the lipid tail(s).

Moreover, the existence of a cholesterol binding site has also been suggested (Wang, personal communications), very close to the PIP_2 binding site. Looking at figure 5.21, it is observed that the interaction densities of the hydrophobic probes match the suggested cholesterol binding site, shown in a magenta circle.

5.4 Discussion

Through the five example test cases demonstrated here, the dynamic pharmacophore probes have the ability to identify potential PIP_2 and cholesterol binding sites, in a range of different targets. As is the case with every method, there are certain limitations which would need to be addressed in the future.

Firstly, in the current examples, binding sites for PIP_2 and cholesterols were examined, due to the charged head group of PIP_2 and hydrophobic interactions of cholesterol, that drive their binding, which makes them ideal targets for the dynamic pharmacophore probes. However, there are currently more than 40000 known lipid species. Even with the generalization of a CG approach, where a lot of different species are coarse grained to the same molecule, there is an extremely large number of molecules to be accounted for. It is therefore expected that more work and a variety of different probe types will be needed to account for a wider range of lipids.

Secondly, in the case of cholesterol binding sites, it was observed that sometimes the binding site for a cholesterol molecule falls on to the same region for the binding of another lipid tail, PIP_2 in this case, due to the hydrophobic interactions driving the binding of both moieties. In such instances, it might be difficult to differentiate between lipid tail and cholesterol binding sites. As a result, a combination of the dynamic pharmacophore probes approach with another method, such as long time-scale membrane simulations or umbrella sampling with potential of mean force simulations or meta-dynamics, can be used to further explore a potential binding site, once it has been identified by this method.

Thirdly, it was observed that for PIP_2 head group binding, at least a positive

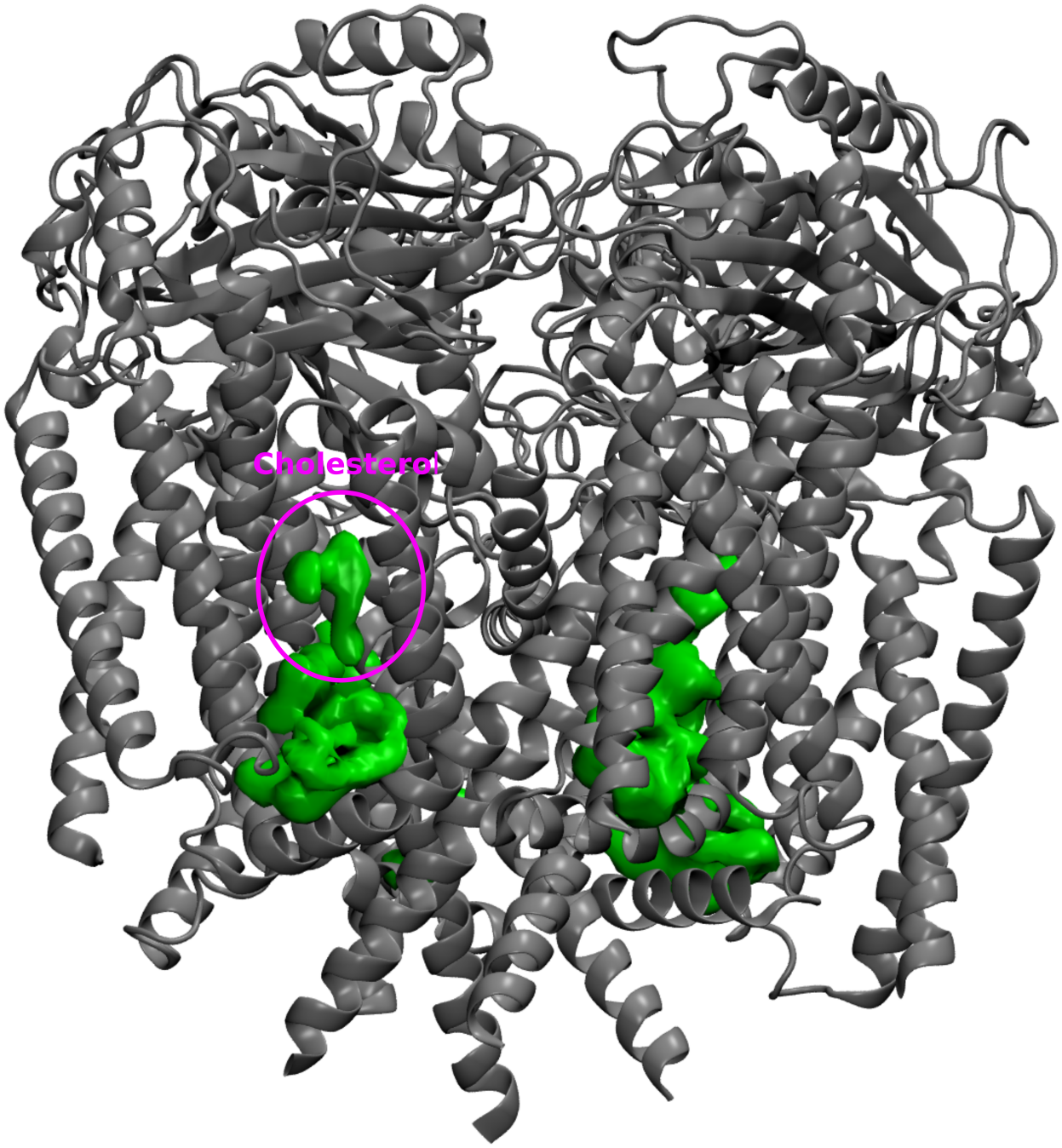


Figure 5.21: A figure showing the interaction densities of the hydrophobic probes (green), along with the suggested cholesterol binding site (magenta circle).

charged residue, or cluster of basic residues, were necessary. This is unsurprising, given the negatively charged moieties of the PIP_2, but still agrees with what was observed through experimental and computational studies.

Finally, when examining the locations of PIP_2 binding sites, it was observed that there were a few common sites between the two class A GPCRs. More specifically, a PIP_2 binding site was observed between TM2, TM3 and TM4 and between TM6 and TM7 in the cases of NTS1 and A_2A. These common PIP_2 binding sites have also been observed between NTS1 and another two class A GPCRs, more specifically the S1P1 receptor and the β1AR receptor. Moreover, initial in house data, from long time scale membrane simulations, of a wide range of class A GPCRs, suggest the existence of a few common PIP_2 binding sites on this family. As a result, a more extended search of these targets, through the use of the dynamic pharmacophore probes approach, would be useful in trying to validate the existence of such pattern.

5.5 Conclusions

In this chapter we have demonstrated the ability of the dynamic pharmacophore probes approach to also be used in identifying charged lipids (PIP_2) and hydrophobic lipids (cholesterol) binding regions on a variety of membrane target proteins. This was achieved in the absence of a lipid bilayer, offering a significant reduction in the computational resources required, compared to long time scale simulations. In particular, a total simulation time of 6 μs was required, compared to the tens or hundreds of microseconds that long time scale simulations require. In addition to that, we have shown that the probes have the ability to identify the particular residues which are involved in the binding of these lipids. Finally, we have shown initial data which suggest the existence of a PIP_2 binding site pattern on class A GPCRs, which would need to be further explored.

Chapter 6

Future Directions

6.1 Parametrization Upgrade

As could be seen from the Figures in chapters 1 and 2, the pharmacophoric probes are particularly effective in identifying certain protein ligand interactions, such as hydrogen bond interactions, but perform less well when it comes to certain other types of interactions, such as pi-pi interactions. In an upgraded version of this method, it would be worth trying to parametrize the hydrophobic probes in order to be able to accurately identify the pi-pi interactions due to the fact that aromatic rings are an important feature of many drug compounds [300]. Moreover, as was evident from the Figures in the previous chapters, the hydrogen bond and acceptor probes would sometimes identify the same interactions. This could be explained through the way these beads were parametrized in the MARTINI forcefield, where the same interaction parameters where used for both types of beads, with slight adjustments for each type, depending on the beads they are interacting with. Although the current configuration offers redundancy when it comes to identifying hydrogen bond interactions, it requires the user to examine the interactions identified to differentiate between hydrogen bond donor and acceptor interactions. It would be beneficial to change the parameters of these probes as well, to ensure they identify distinct interactions. The forcefield currently used is MARTINI version 2.2. With the release of the new version (MARTINI 3.0), currently in its beta version, we hope that some of the drawbacks mentioned above will be resolved by accordingly adjusting the newly-released parameters.

One more drawback of the current application of this approach is the lack of directionality, due to the CG nature of the beads. It is observed that in certain cases, such as CDK2, the amphiphilic environment of the binding site might not be adequately explored by the probes in their current form, due to their smaller size. In order to overcome this limitation, we believe that a new set of probes could be used. These would include two, or potentially more, probes of different types bonded together that would then be used in place of the current pharmacophoric probes, allowing for a certain degree of directionality.

6.2 Conformational Flexibility

The MARTINI forcefield uses an elastic network to maintain the tertiary structure of the protein. Although this allows for a degree of conformational flexibility, it does not allow for the observation of large conformational changes. As a result, phenomena such as the activation of a GPCR, would not be observed during the simulation, irrespective of the time scales used.

Results by Yang and Wang and Ghanakota et al. [104], on Bcl-xL and ABL kinase respectively, indicate the difference of the resulting interaction maps based on the starting conformations of the targets used. The low requirements for computational resources of this approach makes it easy to run multiple repeats with different conformations of the same target, minimising the impact of the low conformational flexibility. However, the incorporation of Go potentials [301] instead of an elastic network, as is the case in MARTINI 3, would allow for a greater movement and structural rearrangement of the CG structures simulated. This, in turn, would allow for greater dynamics in the binding sites of the targets examined, which would lead to more accurate protein-ligand interaction predictions.

6.3 Identifying Cryptic Pockets

We have shown that this method has the potential to explore orthosteric as well as allosteric binding sites on protein targets. However, it might not always be possible to identify an allosteric binding site or cryptic binding pocket using this approach alone, especially if the lower conformational flexibility of CG simulations is taken

into account. Methods to identify cryptic binding pockets, through the use of replica exchange simulations [59] have been developed. Combining these methods with the CG dynamic pharmacophore probes approach would provide an extremely powerful tool to explore potential drug targets, by allowing the identification of the cryptic binding sites which can then be explored with this approach.

6.4 Targeting Protein Protein Interactions

Protein protein interactions (PPI) constitute a very difficult drug target, as they are usually spread over wide, shallow interfaces. Being able to identify which PPI can be prioritised would offer a new avenue for rational drug design. The small size of the dynamic pharmacophore probes, along with the probes' properties, offer a great tool for examining such interfaces. The current methods employed for identifying binding hotspots could be used and adapted for identification of PPI hotspots [302].

6.5 Application to CADD Projects

One of the biggest challenges in drug design is the identification of the few compounds that have the properties and the potential to become prime drug candidates. Here, we present a quantitative validation of the predictive capabilities of the method and the pharmacophoric probes used. Therefore, the next step would be the application of this method to real-world drug design projects. An additional feature which would make this approach even more for suitable for such an application is the ability to rank potential ligands. SILCS has already made headway in this direction through the identification of co-solvent binding preferences on a grid and the use of ligand scores [110]. In our approach, the ΔG_i values calculated during the analysis could be used in a similar manner to rank potential ligands.

6.6 PIP_2 Interaction Patterns on GPCRs

As was shown before, this method has the ability to identify PIP_2 binding interactions on a range of GPCRs. The conformational conservation of GPCRs of known structure raises the question of whether the interactions observed here are also conserved on

other GPCRs, of the same or different subfamilies. Data from initial long time scale CG MD simulations suggest a conserved pattern of PIP2 interactions on class A structures. However, this study did not include a large enough number of structures belonging to different families in order to draw conclusions for the other subfamilies. Initial data from simulations of the smoothened receptor with cholesterol, suggest a pattern of interactions distinct from class A GPCRS [294]. Using a data set which includes as wide a range of different GPCR structures as possible would offer the opportunity to draw meaningful conclusions for the rest of the GPCR subfamilies. The low computational cost of this method, 2 μs of simulation time require about 12 hours on a workstation, also facilitates such an endeavour.

6.7 Final Remarks

In chapter 3 the use of the MARTINI beads as pharmacophoric probes was demonstrated. The parametrisation process was explained and the analysis protocol was described. The results of the application of the method to a data set of water soluble targets, of pharmacological interest, were then presented. It was shown that the pharmacophoric probes have the ability to identify the majority of protein-fragment interactions, along with the protein-lead with slightly reduced accuracy. A quantitative evaluation of the predictive capabilities of the probes was then performed, showing results equal or better to the currently available methods. Comparing the probes' sensitivity, precision, specificity and accuracy values with those presented by Arcon et al. [144] higher values, of approximately 30-40%, are observed for this approach, with the exception of the hydrophobic probes, the reasons for which have been elaborated on in previous chapters.

In chapter 4, the dataset was expanded to a range of membrane proteins, and more specifically GPCRs. The orthosteric and allosteric binding sites of a range of GPCR targets were examined and the same analytical framework as the previous chapter was applied. It was shown that in this case the probes were able to predict the protein ligand interactions with reasonably high accuracy, both for the orthosteric and allosteric binding sites.

In chapter 5, this approach was applied to a range of membrane proteins in an

effort to identify PIP_2 and cholesterol binding sites. It was seen that the method has the capability to predict these sites, along with identifying the residues involved in binding the head groups of PIP_2.

Since its conception, CADD has rapidly evolved to have the capabilities which are useful today. We are certain that with the current rate of advancement, CADD will be able to tackle a lot of the major pharmaceutical challenges that still remain unsolved. We believe this thesis is a positive contribution to the field and will provide another useful tool in the battle against disease.

Bibliography

[1] V. Lounnas, T. Ritschel, J. Kelder, R. McGuire, R. P. Bywater, and N. Foloppe, "Current progress in structure-based rational drug design marks a new mindset in drug discovery," *Computational and structural biotechnology journal*, vol. 5, no. 6, p. e201302011, 2013.

[2] W. Richards, D. Aschman, and J. Hammond, "Conformational flexibility in physiologically active amines," *Journal of theoretical biology*, vol. 52, no. 1, pp. 223–230, 1975.

[3] W. Richards, R. Clarkson, and C. R. Ganellin, "Molecule-receptor specificity," *Phil. Trans. R. Soc. Lond. B*, vol. 272, no. 915, pp. 75–85, 1975.

[4] F. K. Brown, E. C. Sherer, S. A. Johnson, M. K. Holloway, and B. S. Sherborne, "The evolution of drug design at merck research laboratories," *Journal of computer-aided molecular design*, vol. 31, no. 3, pp. 255–266, 2017.

[5] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe, "Computational methods in drug discovery," *Pharmacological reviews*, vol. 66, no. 1, pp. 334–395, 2014.

[6] A. Kumar, A. Voet, and K. Zhang, "Fragment based drug design: from experimental to computational approaches," *Current medicinal chemistry*, vol. 19, no. 30, pp. 5128–5147, 2012.

[7] R. Vijayakrishnan *et al.*, "Structure-based drug design and modern medicine," *Journal of postgraduate medicine*, vol. 55, no. 4, p. 301, 2009.

[8] J. Bryan, "From snake venom to ace inhibitor–the discovery and rise of captopril," *Pharmaceutical Journal*, vol. 282, no. 7548, p. 455, 2009.

[9] T. T. Talele, S. A. Khedkar, and A. C. Rigby, "Successful applications of computer aided drug discovery: moving drugs from concept to the clinic," *Current topics in medicinal chemistry*, vol. 10, no. 1, pp. 127–141, 2010.

[10] J. H. Van Drie, "Computer-aided drug design: the next 20 years," *Journal of computer-aided molecular design*, vol. 21, no. 10-11, pp. 591–601, 2007.

[11] I. Kapetanovic, "Computer-aided drug discovery and development (caddd): in silico-chemico-biological approach," *Chemico-biological interactions*, vol. 171, no. 2, pp. 165–176, 2008.

[12] R. G. Hill, *Drug Discovery and development-e-book: technology in transition*. Elsevier Health Sciences, 2012.

[13] K. M. Merz Jr, D. Ringe, and C. H. Reynolds, *Drug design: structure-and ligand-based approaches*. Cambridge University Press, 2010.

[14] D. Horvath, "A virtual screening approach applied to the search for trypanothione reductase inhibitors," *Journal of medicinal chemistry*, vol. 40, no. 15, pp. 2412–2423, 1997.

[15] D. Prada-Gracia, S. Huerta-Yépez, and L. M. Moreno-Vargas, "Application of computational methods for anticancer drug discovery, design, and optimization," *Boletín Médico Del Hospital Infantil de México (English Edition)*, vol. 73, no. 6, pp. 411–423, 2016.

[16] P. J. Hajduk and J. Greer, "A decade of fragment-based drug design: strategic advances and lessons learned," *Nature reviews Drug discovery*, vol. 6, no. 3, p. 211, 2007.

[17] R. J. Bienstock, *Library design, search methods, and applications of fragment-based drug design*. American Chemical Society, 2011.

[18] P. Aparoy, K. Kumar Reddy, and P. Reddanna, "Structure and ligand based drug design strategies in the development of novel 5-lox inhibitors," *Current medicinal chemistry*, vol. 19, no. 22, pp. 3763–3778, 2012.

[19] C. Acharya, A. Coop, J. E Polli, and A. D MacKerell, "Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach," *Current computer-aided drug design*, vol. 7, no. 1, pp. 10–22, 2011.

[20] C.-H. Lee, H.-C. Huang, and H.-F. Juan, "Reviewing ligand-based rational drug design: the search for an atp synthase inhibitor," *International journal of molecular sciences*, vol. 12, no. 8, pp. 5304–5318, 2011.

[21] O. Guner, "History and evolution of the pharmacophore concept in computer-aided drug design," *Current topics in medicinal chemistry*, vol. 2, no. 12, pp. 1321–1332, 2002.

[22] G. D. Geromichalos, C. E. Alifieris, E. G. Geromichalou, and D. T. Trafalis, "Overview on the current status of virtual high-throughput screening and combinatorial chemistry approaches in multi-target anticancer drug discovery; part i," *J Buon*, vol. 21, no. 4, pp. 764–779, 2016.

[23] A. C. Anderson, "The process of structure-based drug design," *Chemistry & biology*, vol. 10, no. 9, pp. 787–797, 2003.

[24] D. E. Danley, "Crystallization to obtain protein–ligand complexes for structure-aided drug design," *Acta Crystallographica Section D: Biological Crystallography*, vol. 62, no. 6, pp. 569–575, 2006.

[25] Z. Zhang, Y. Li, B. Lin, M. Schroeder, and B. Huang, "Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction," *Bioinformatics*, vol. 27, no. 15, pp. 2083–2088, 2011.

[26] T. Simões, D. Lopes, S. Dias, F. Fernandes, J. Pereira, J. Jorge, C. Bajaj, and A. Gomes, "Geometric detection algorithms for cavities on protein surfaces in molecular graphics: a survey," in *Computer Graphics Forum*, vol. 36, pp. 643–683, Wiley Online Library, 2017.

[27] D. G. Levitt and L. J. Banaszak, "Pocket: a computer graphies method for identifying and displaying protein cavities and their surrounding amino acids," *Journal of molecular graphics*, vol. 10, no. 4, pp. 229–234, 1992.

[28] M. Hendlich, F. Rippmann, and G. Barnickel, "Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins," *Journal of Molecular Graphics and Modelling*, vol. 15, no. 6, pp. 359–363, 1997.

[29] R. A. Laskowski, "Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions," *Journal of molecular graphics*, vol. 13, no. 5, pp. 323–330, 1995.

[30] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, "Dock 4.0: search strategies for automated molecular docking of flexible molecule databases," *Journal of computer-aided molecular design*, vol. 15, no. 5, pp. 411–428, 2001.

[31] R. Laurie, T. Alasdair, and R. M. Jackson, "Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening.," *Current Protein and Peptide Science*, vol. 7, no. 5, pp. 395–406, 2006.

[32] V. Le Guilloux, P. Schmidtke, and P. Tuffery, "Fpocket: an open source platform for ligand pocket detection," *BMC bioinformatics*, vol. 10, no. 1, p. 168, 2009.

[33] M. Petřek, M. Otyepka, P. Banáš, P. Košinová, J. Koča, and J. Damborskỳ, "Caver: a new tool to explore routes from protein clefts, pockets and cavities," *BMC bioinformatics*, vol. 7, no. 1, p. 316, 2006.

[34] E. Yaffe, D. Fishelovitch, H. J. Wolfson, D. Halperin, and R. Nussinov, "Molaxis: efficient and accurate identification of channels in macromolecules," *Proteins: Structure, Function, and Bioinformatics*, vol. 73, no. 1, pp. 72–86, 2008.

[35] A. Bhinge, P. Chakrabarti, K. Uthanumallian, K. Bajaj, K. Chakraborty, and R. Varadarajan, "Accurate detection of protein: ligand binding sites using molecular dynamics simulations," *Structure*, vol. 12, no. 11, pp. 1989–1999, 2004.

[36] T. Paramo, A. East, D. Garzoén, M. B. Ulmschneider, and P. J. Bond, "Efficient characterization of protein cavities within molecular simulation trajectories:

trj_cavity," *Journal of chemical theory and computation*, vol. 10, no. 5, pp. 2151–2164, 2014.

[37] P. J. Goodford, "A computational procedure for determining energetically favorable binding sites on biologically important macromolecules," *Journal of medicinal chemistry*, vol. 28, no. 7, pp. 849–857, 1985.

[38] A. Miranker and M. Karplus, "Functionality maps of binding sites: a multiple copy simultaneous search method," *Proteins: Structure, Function, and Bioinformatics*, vol. 11, no. 1, pp. 29–34, 1991.

[39] C. A. Reynolds, R. C. Wade, and P. J. Goodford, "Identifying targets for bioreductive agents: using grid to predict selective binding regions of proteins," *Journal of molecular graphics*, vol. 7, no. 2, pp. 103–108, 1989.

[40] R. C. Wade, K. J. Clark, and P. J. Goodford, "Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. ligand probe groups with the ability to form two hydrogen bonds," *Journal of medicinal chemistry*, vol. 36, no. 1, pp. 140–147, 1993.

[41] R. C. Wade and P. J. Goodford, "Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. ligand probe groups with the ability to form more than two hydrogen bonds," *Journal of medicinal chemistry*, vol. 36, no. 1, pp. 148–156, 1993.

[42] A. T. Laurie and R. M. Jackson, "Q-sitefinder: an energy-based method for the prediction of protein–ligand binding sites," *Bioinformatics*, vol. 21, no. 9, pp. 1908–1916, 2005.

[43] M. Hernandez, D. Ghersi, and R. Sanchez, "Sitehound-web: a server for ligand binding site identification in protein structures," *Nucleic acids research*, vol. 37, no. suppl_2, pp. W413–W416, 2009.

[44] Y.-F. Chen, K.-C. Hsu, S.-R. Lin, W.-C. Wang, Y.-C. Huang, and J.-M. Yang, "Simmap: a web server for inferring site-moiety map to recognize interaction

preferences between protein pockets and compound moieties," *Nucleic acids research*, vol. 38, no. suppl_2, pp. W424–W430, 2010.

[45] M. Weisel, E. Proschak, and G. Schneider, "Pocketpicker: analysis of ligand binding-sites with shape descriptors," *Chemistry Central Journal*, vol. 1, no. 1, p. 7, 2007.

[46] J. R. Wagner, C. T. Lee, J. D. Durrant, R. D. Malmstrom, V. A. Feher, and R. E. Amaro, "Emerging computational methods for the rational discovery of allosteric drugs," *Chemical reviews*, vol. 116, no. 11, pp. 6370–6390, 2016.

[47] T. Kortvelyesi, S. Dennis, M. Silberstein, L. Brown III, and S. Vajda, "Algorithms for computational solvent mapping of proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 51, no. 3, pp. 340–351, 2003.

[48] M. Silberstein, S. Dennis, L. Brown III, T. Kortvelyesi, K. Clodfelter, and S. Vajda, "Identification of substrate binding sites in enzymes by computational solvent mapping," *Journal of molecular biology*, vol. 332, no. 5, pp. 1095–1113, 2003.

[49] T. Kortvelyesi, M. Silberstein, S. Dennis, and S. Vajda, "Improved mapping of protein binding sites," *Journal of computer-aided molecular design*, vol. 17, no. 2-4, pp. 173–186, 2003.

[50] B. E. Warren, *X-ray Diffraction*. Courier Corporation, 1990.

[51] B. Rupp, *Biomolecular crystallography: principles, practice, and application to structural biology*. Garland Science, 2009.

[52] G. M. Clore and A. M. Gronenborn, "Structures of larger proteins in solution: three-and four-dimensional heteronuclear nmr spectroscopy," *Science*, vol. 252, no. 5011, pp. 1390–1399, 1991.

[53] J. Keeler, *Understanding NMR spectroscopy*. John Wiley & Sons, 2011.

[54] W. Kühlbrandt, "Microscopy: cryo-em enters a new era," *Elife*, vol. 3, p. e03678, 2014.

[55] E. Callaway, "The revolution will not be crystallized: a new method sweeps through structural biology," *Nature News*, vol. 525, no. 7568, p. 172, 2015.

[56] D. L. Mobley and K. A. Dill, "Binding of small-molecule ligands to proteins:"what you see" is not always "what you get"," *Structure*, vol. 17, no. 4, pp. 489–498, 2009.

[57] D. Beglov, D. R. Hall, A. E. Wakefield, L. Luo, K. N. Allen, D. Kozakov, A. Whitty, and S. Vajda, "Exploring the structural origins of cryptic sites on proteins," *Proceedings of the National Academy of Sciences*, vol. 115, no. 15, pp. E3416–E3425, 2018.

[58] S. Vajda, D. Beglov, A. E. Wakefield, M. Egbert, and A. Whitty, "Cryptic binding sites on proteins: definition, detection, and druggability," *Current opinion in chemical biology*, vol. 44, pp. 1–8, 2018.

[59] V. Oleinikovas, G. Saladino, B. P. Cossins, and F. L. Gervasio, "Understanding cryptic pocket formation in protein targets by enhanced sampling simulations," *Journal of the American Chemical Society*, vol. 138, no. 43, pp. 14257–14263, 2016.

[60] D. Garzón, C. Anselmi, P. J. Bond, and J. D. Faraldo-Gómez, "Dynamics of the antigen-binding grooves in cd1 proteins reversible hydrophobic collapse in the lipid-free state," *Journal of Biological Chemistry*, vol. 288, no. 27, pp. 19528–19536, 2013.

[61] T. Paramo, T. J. Piggot, C. E. Bryant, and P. J. Bond, "The structural basis for endotoxin-induced allosteric regulation of the toll-like receptor 4 (tlr4) innate immune receptor," *Journal of Biological Chemistry*, pp. jbc–M113, 2013.

[62] P. J. Bond and J. D. Faraldo-Gómez, "Molecular mechanism of selective recruitment of syk kinases by the membrane antigen-receptor complex," *Journal of Biological Chemistry*, vol. 286, no. 29, pp. 25872–25881, 2011.

[63] Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger, and D. E. Shaw, "How does a drug molecule find its target binding site?," *Journal of the American Chemical Society*, vol. 133, no. 24, pp. 9181–9183, 2011.

[64] R. O. Dror, A. C. Pan, D. H. Arlow, D. W. Borhani, P. Maragakis, Y. Shan, H. Xu, and D. E. Shaw, "Pathway and mechanism of drug binding to g-protein-coupled receptors," *Proceedings of the National Academy of Sciences*, vol. 108, no. 32, pp. 13118–13123, 2011.

[65] J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande, "Alchemical free energy methods for drug discovery: progress and challenges," *Current opinion in structural biology*, vol. 21, no. 2, pp. 150–160, 2011.

[66] F. Samsudin, J. L. Parker, M. S. Sansom, S. Newstead, and P. W. Fowler, "Accurate prediction of ligand affinities for a proton-dependent oligopeptide transporter," *Cell chemical biology*, vol. 23, no. 2, pp. 299–309, 2016.

[67] K. C. Mundim and D. Ellis, "Stochastic classical molecular dynamics coupled to functional density theory: Applications to large molecular systems," *Brazilian journal of physics*, vol. 29, no. 1, pp. 199–214, 1999.

[68] J. Schlitter, M. Engels, and P. Krüger, "Targeted molecular dynamics: a new approach for searching pathways of conformational transitions," *Journal of molecular graphics*, vol. 12, no. 2, pp. 84–89, 1994.

[69] T. Huber and W. F. Van Gunsteren, "Swarm-md: searching conformational space by cooperative molecular dynamics," *The Journal of Physical Chemistry A*, vol. 102, no. 29, pp. 5937–5943, 1998.

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

[71] C. F. Abrams and E. Vanden-Eijnden, "Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics," *Proceedings of the National Academy of Sciences*, vol. 107, no. 11, pp. 4961–4966, 2010.

[72] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, "Plumed 2: New feathers for an old bird," *Computer Physics Communications*, vol. 185, no. 2, pp. 604–613, 2014.

[73] M. Raunest and C. Kandt, "dxtuber: detecting protein cavities, tunnels and clefts based on protein and solvent dynamics," *Journal of Molecular Graphics and Modelling*, vol. 29, no. 7, pp. 895–905, 2011.

[74] W. Humphrey, A. Dalke, and K. Schulten, "Vmd: visual molecular dynamics," *Journal of molecular graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[75] P. Schmidtke, A. Bidon-Chanal, F. J. Luque, and X. Barril, "Mdpocket: open-source cavity detection and characterization on molecular dynamics trajectories," *Bioinformatics*, vol. 27, no. 23, pp. 3276–3285, 2011.

[76] I. R. Craig, C. Pfleger, H. Gohlke, J. W. Essex, and K. Spiegel, "Pocket-space maps to identify novel binding-site conformations in proteins," *Journal of chemical information and modeling*, vol. 51, no. 10, pp. 2666–2679, 2011.

[77] P. B. Gupta, T. T. Onder, G. Jiang, K. Tao, C. Kuperwasser, R. A. Weinberg, and E. S. Lander, "Identification of selective inhibitors of cancer stem cells by high-throughput screening," *Cell*, vol. 138, no. 4, pp. 645–659, 2009.

[78] P. Gramatica, "Principles of qsar models validation: internal and external," *QSAR & combinatorial science*, vol. 26, no. 5, pp. 694–701, 2007.

[79] J.-M. Lehn, "Dynamic combinatorial chemistry and virtual combinatorial libraries," *Chemistry–A European Journal*, vol. 5, no. 9, pp. 2455–2463, 1999.

[80] F. Zanella, J. B. Lorens, and W. Link, "High content screening: seeing is believing," *Trends in biotechnology*, vol. 28, no. 5, pp. 237–245, 2010.

[81] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, *et al.*, "Impact of high-throughput screening in biomedical research," *Nature reviews Drug discovery*, vol. 10, no. 3, p. 188, 2011.

[82] E. Lionta, G. Spyrou, D. K Vassilatis, and Z. Cournia, "Structure-based virtual screening for drug discovery: principles, applications and recent advances," *Current topics in medicinal chemistry*, vol. 14, no. 16, pp. 1923–1938, 2014.

[83] C. A. Lipinski, "Lead-and drug-like compounds: the rule-of-five revolution," *Drug Discovery Today: Technologies*, vol. 1, no. 4, pp. 337–341, 2004.

[84] B. Lamoree and R. E. Hubbard, "Current perspectives in fragment-based lead discovery (fbld)," *Essays in biochemistry*, vol. 61, no. 5, pp. 453–464, 2017.

[85] M. J. Wasko, K. A. Pellegrene, J. D. Madura, and C. K. Surratt, "A role for fragment-based drug design in developing novel lead compounds for central nervous system targets," *Frontiers in neurology*, vol. 6, p. 197, 2015.

[86] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug discovery," *British journal of pharmacology*, vol. 162, no. 6, pp. 1239–1249, 2011.

[87] C.-Y. Yang and S. Wang, "Computational analysis of protein hotspots," *ACS medicinal chemistry letters*, vol. 1, no. 3, pp. 125–129, 2010.

[88] C. J. Radoux, T. S. Olsson, W. R. Pitt, C. R. Groom, and T. L. Blundell, "Identifying interactions that determine fragment binding at protein hotspots," *Journal of medicinal chemistry*, vol. 59, no. 9, pp. 4314–4325, 2016.

[89] A. L. Hopkins, G. M. Keserü, P. D. Leeson, D. C. Rees, and C. H. Reynolds, "The role of ligand efficiency metrics in drug discovery," *Nature reviews Drug discovery*, vol. 13, no. 2, pp. 105–121, 2014.

[90] J. A. Creighton, C. G. Blatchford, and M. G. Albrecht, "Plasma resonance enhancement of raman scattering by pyridine adsorbed on silver or gold sol particles of size comparable to the excitation wavelength," *Journal of the Chemical Society, Faraday Transactions 2: Molecular and Chemical Physics*, vol. 75, pp. 790–798, 1979.

[91] D. M. Dias and A. Ciulli, "Nmr approaches in structure-based lead discovery: recent developments and new frontiers for targeting multi-protein complexes," *Progress in biophysics and molecular biology*, vol. 116, no. 2-3, pp. 101–112, 2014.

[92] J. Ringe, C. Kaps, B. Schmitt, K. Büscher, J. Bartel, H. Smolian, O. Schultz, G. R. Burmester, T. Häupl, and M. Sittinger, "Porcine mesenchymal stem cells," *Cell and tissue research*, vol. 307, no. 3, pp. 321–327, 2002.

[93] D. A. Erlanson, W. Jahnke, R. Mannhold, H. Kubinyi, and G. Folkers, *Fragment-based drug discovery: lessons and outlook*, vol. 67. John Wiley & Sons, 2016.

[94] A. Chavanieu and M. Pugniere, "Developments in spr fragment screening," *Expert opinion on drug discovery*, vol. 11, no. 5, pp. 489–499, 2016.

[95] E. Evensen, D. Joseph-McCarthy, and M. Karplus, "Mcss version 2.1," *Harvard University, Cambridge, MA, USA*, 1997.

[96] H.-J. Böhm, "The computer program ludi: a new method for the de novo design of enzyme inhibitors," *Journal of computer-aided molecular design*, vol. 6, no. 1, pp. 61–78, 1992.

[97] V. Gillet, A. P. Johnson, P. Mata, S. Sike, and P. Williams, "Sprout: a program for structure generation," *Journal of computer-aided molecular design*, vol. 7, no. 2, pp. 127–153, 1993.

[98] M. L. Verdonk, J. C. Cole, and R. Taylor, "Superstar: a knowledge-based approach for identifying interaction sites in proteins," *Journal of molecular biology*, vol. 289, no. 4, pp. 1093–1108, 1999.

[99] J. W. M. Nissink and R. Taylor, "Combined use of physicochemical data and small-molecule crystallographic contact propensities to predict interactions in protein binding sites," *Organic & biomolecular chemistry*, vol. 2, no. 22, pp. 3238–3249, 2004.

[100] H. A. Carlson, K. M. Masukawa, K. Rubins, F. D. Bushman, W. L. Jorgensen, R. D. Lins, J. M. Briggs, and J. A. McCammon, "Developing a dynamic pharmacophore model for hiv-1 integrase," *Journal of medicinal chemistry*, vol. 43, no. 11, pp. 2100–2114, 2000.

[101] D. Kozakov, L. E. Grove, D. R. Hall, T. Bohnuud, S. E. Mottarella, L. Luo, B. Xia, D. Beglov, and S. Vajda, "The ftmap family of web servers for determining and characterizing ligand-binding hot spots of proteins," *Nature protocols*, vol. 10, no. 5, p. 733, 2015.

[102] C. M. Stultz and M. Karplus, "Mcss functionality maps for a flexible protein," *Proteins: Structure, Function, and Bioinformatics*, vol. 37, no. 4, pp. 512–529, 1999.

[103] D. Ramírez, "Computational methods applied to rational drug design," *The open medicinal chemistry journal*, vol. 10, p. 7, 2016.

[104] P. Ghanakota and H. A. Carlson, "Driving structure-based drug discovery through cosolvent molecular dynamics: Miniperspective," *Journal of medicinal chemistry*, vol. 59, no. 23, pp. 10383–10399, 2016.

[105] J. Seco, F. J. Luque, and X. Barril, "Binding site detection and druggability index from first principles," *Journal of medicinal chemistry*, vol. 52, no. 8, pp. 2363–2371, 2009.

[106] D. Alvarez-Garcia and X. Barril, "Relationship between protein flexibility and binding: Lessons for structure-based drug design," *Journal of chemical theory and computation*, vol. 10, no. 6, pp. 2608–2614, 2014.

[107] D. Alvarez-Garcia and X. Barril, "Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites," *Journal of medicinal chemistry*, vol. 57, no. 20, pp. 8530–8539, 2014.

[108] O. Guvench and A. D. MacKerell Jr, "Computational fragment-based binding site identification by ligand competitive saturation," *PLoS computational biology*, vol. 5, no. 7, p. e1000435, 2009.

[109] A. E. Klon, *Fragment-based Methods in Drug Discovery*. Springer, 2015.

[110] E. P. Raman, W. Yu, O. Guvench, and A. D. MacKerell Jr, "Reproducing crystal binding modes of ligand functional groups using site-identification

by ligand competitive saturation (silcs) simulations," *Journal of chemical information and modeling*, vol. 51, no. 4, pp. 877–896, 2011.

[111] E. P. Raman, K. Vanommeslaeghe, and A. D. MacKerell Jr, "Site-specific fragment identification guided by single-step free energy perturbation calculations," *Journal of chemical theory and computation*, vol. 8, no. 10, pp. 3513–3525, 2012.

[112] T. J. Foster, A. D. MacKerell Jr, and O. Guvench, "Balancing target flexibility and target denaturation in computational fragment-based inhibitor discovery," *Journal of computational chemistry*, vol. 33, no. 23, pp. 1880–1891, 2012.

[113] E. P. Raman, W. Yu, S. K. Lakkaraju, and A. D. MacKerell Jr, "Inclusion of multiple fragment types in the site identification by ligand competitive saturation (silcs) approach," *Journal of chemical information and modeling*, vol. 53, no. 12, pp. 3384–3398, 2013.

[114] W. Yu, S. K. Lakkaraju, E. P. Raman, and A. D. MacKerell, "Site-identification by ligand competitive saturation (silcs) assisted pharmacophore modeling," *Journal of computer-aided molecular design*, vol. 28, no. 5, pp. 491–507, 2014.

[115] W. Yu, S. K. Lakkaraju, E. P. Raman, L. Fang, and A. D. MacKerell Jr, "Pharmacophore modeling using site-identification by ligand competitive saturation (silcs) with multiple probe molecules," *Journal of chemical information and modeling*, vol. 55, no. 2, pp. 407–420, 2015.

[116] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet, "Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking," *Journal of medicinal chemistry*, vol. 55, no. 14, pp. 6582–6594, 2012.

[117] C. C. G. Inc., "Molecular operating environment (moe)," 2016.

[118] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "Autodock4 and autodocktools4: Automated docking with selective receptor flexibility," *Journal of computational chemistry*, vol. 30, no. 16, pp. 2785–2791, 2009.

[119] B. Hu and M. A. Lill, "Protein pharmacophore selection using hydration-site analysis," *Journal of chemical information and modeling*, vol. 52, no. 4, pp. 1046–1060, 2012.

[120] S. K. Lakkaraju, E. P. Raman, W. Yu, and A. D. MacKerell Jr, "Sampling of organic solutes in aqueous and heterogeneous environments using oscillating excess chemical potentials in grand canonical-like monte carlo-molecular dynamics simulations," *Journal of chemical theory and computation*, vol. 10, no. 6, pp. 2281–2290, 2014.

[121] X. He, S. K. Lakkaraju, M. Hanscom, Z. Zhao, J. Wu, B. Stoica, A. D. MacKerell, A. I. Faden, and F. Xue, "Acyl-2-aminobenzimidazoles: A novel class of neuroprotective agents targeting mglur5," *Bioorganic & medicinal chemistry*, vol. 23, no. 9, pp. 2211–2220, 2015.

[122] S. K. Lakkaraju, H. Mbatia, M. Hanscom, Z. Zhao, J. Wu, B. Stoica, A. D. MacKerell, A. I. Faden, and F. Xue, "Cyclopropyl-containing positive allosteric modulators of metabotropic glutamate receptor subtype 5," *Bioorganic & medicinal chemistry letters*, vol. 25, no. 11, pp. 2275–2279, 2015.

[123] K. W. Lexa and H. A. Carlson, "Full protein flexibility is essential for proper hot-spot mapping," *Journal of the American Chemical Society*, vol. 133, no. 2, pp. 200–202, 2010.

[124] K. W. Lexa and H. A. Carlson, "Improving protocols for protein mapping through proper comparison to crystallography data," *Journal of chemical information and modeling*, vol. 53, no. 2, pp. 391–402, 2013.

[125] K. W. Lexa, G. B. Goh, and H. A. Carlson, "Parameter choice matters: validating probe parameters for use in mixed-solvent simulations," *Journal of chemical information and modeling*, vol. 54, no. 8, pp. 2190–2199, 2014.

[126] P. M. Ung, P. Ghanakota, S. E. Graham, K. W. Lexa, and H. A. Carlson, "Identifying binding hot spots on protein surfaces by mixed-solvent molecular dynamics: Hiv-1 protease as a test case," *Biopolymers*, vol. 105, no. 1, pp. 21–34, 2016.

[127] P. Ghanakota and H. A. Carlson, "Moving beyond active-site detection: Mixmd applied to allosteric systems," *The Journal of Physical Chemistry B*, vol. 120, no. 33, pp. 8685–8695, 2016.

[128] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon, "The statistical-thermodynamic basis for computation of binding affinities: a critical review," *Biophysical journal*, vol. 72, no. 3, pp. 1047–1069, 1997.

[129] D. Hamelberg and J. A. McCammon, "Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method," *Journal of the American Chemical Society*, vol. 126, no. 24, pp. 7683–7689, 2004.

[130] C.-Y. Yang and S. Wang, "Hydrophobic binding hot spots of bcl-xl protein-protein interfaces by cosolvent molecular dynamics simulation," *ACS medicinal chemistry letters*, vol. 2, no. 4, pp. 280–284, 2011.

[131] C.-Y. Yang and S. Wang, "Analysis of flexibility and hotspots in bcl-xl and mcl-1 proteins for the design of selective small-molecule inhibitors," *ACS medicinal chemistry letters*, vol. 3, no. 4, pp. 308–312, 2012.

[132] C.-Y. Yang, "Identification of potential small molecule allosteric modulator sites on il-1r1 ectodomain using accelerated conformational sampling method," *PloS one*, vol. 10, no. 2, p. e0118671, 2015.

[133] A. Kalenkiewicz, B. J. Grant, and C.-Y. Yang, "Enrichment of druggable conformations from apo protein structures using cosolvent-accelerated molecular dynamics," *Biology*, vol. 4, no. 2, pp. 344–366, 2015.

[134] A. Bakan, N. Nevins, A. S. Lakdawala, and I. Bahar, "Druggability assessment of allosteric proteins by dynamics simulations in the presence of probe molecules," *Journal of chemical theory and computation*, vol. 8, no. 7, pp. 2435–2447, 2012.

[135] D. Huang and A. Caflisch, "Small molecule binding to proteins: affinity and binding/unbinding dynamics from atomistic simulations," *ChemMedChem*, vol. 6, no. 9, pp. 1578–1580, 2011.

[136] D. Huang, E. Rossini, S. Steiner, and A. Caflisch, "Structured water molecules in the binding site of bromodomains can be displaced by cosolvent," *ChemMedChem*, vol. 9, no. 3, pp. 573–579, 2014.

[137] T. Halgren, "New method for fast and accurate binding-site identification and analysis," *Chemical biology & drug design*, vol. 69, no. 2, pp. 146–148, 2007.

[138] Y. S. Tan, P. Śledź, S. Lang, C. J. Stubbs, D. R. Spring, C. Abell, and R. B. Best, "Using ligand-mapping simulations to design a ligand selectively targeting a cryptic surface pocket of polo-like kinase 1," *Angewandte Chemie International Edition*, vol. 51, no. 40, pp. 10078–10081, 2012.

[139] Y. S. Tan, D. R. Spring, C. Abell, and C. Verma, "The use of chlorobenzene as a probe molecule in molecular dynamics simulations," *Journal of chemical information and modeling*, vol. 54, no. 7, pp. 1821–1827, 2014.

[140] Y. S. Tan, D. R. Spring, C. Abell, and C. S. Verma, "The application of ligand-mapping molecular dynamics simulations to the rational design of peptidic modulators of protein–protein interactions," *Journal of chemical theory and computation*, vol. 11, no. 7, pp. 3199–3210, 2015.

[141] N. Basse, J. L. Kaar, G. Settanni, A. C. Joerger, T. J. Rutherford, and A. R. Fersht, "Toward the rational design of p53-stabilizing drugs: probing the surface of the oncogenic y220c mutant," *Chemistry & biology*, vol. 17, no. 1, pp. 46–56, 2010.

[142] P. Prakash, J. F. Hancock, and A. A. Gorfe, "Binding hotspots on k-ras: Consensus ligand binding sites and other reactive regions from probe-based molecular dynamics analysis," *Proteins: Structure, Function, and Bioinformatics*, vol. 83, no. 5, pp. 898–909, 2015.

[143] P. Prakash, A. Sayyed-Ahmad, and A. A. Gorfe, "pmd-membrane: a method for ligand binding site identification in membrane-bound proteins," *PLoS computational biology*, vol. 11, no. 10, p. e1004469, 2015.

[144] J. P. Arcon, L. A. Defelipe, C. P. Modenutti, E. D. Loépez, D. Alvarez-Garcia, X. Barril, A. G. Turjanski, and M. A. Martí, "Molecular dynamics in mixed

solvents reveals protein–ligand interactions, improves docking, and allows accurate binding free energy predictions," *Journal of chemical information and modeling*, vol. 57, no. 4, pp. 846–863, 2017.

[145] X. Qing, X. Y. Lee, J. De Raeymaecker, J. R. Tame, K. Y. Zhang, M. De Maeyer, and A. Voet, "Pharmacophore modeling: advances, limitations, and current utility in drug discovery," 2014.

[146] T. Kaserer, K. R. Beck, M. Akram, A. Odermatt, and D. Schuster, "Pharmacophore models and pharmacophore-based virtual screening: concepts and applications exemplified on hydroxysteroid dehydrogenases," *Molecules*, vol. 20, no. 12, pp. 22799–22832, 2015.

[147] S.-Y. Yang, "Pharmacophore modeling and applications in drug discovery: challenges and recent advances," *Drug discovery today*, vol. 15, no. 11-12, pp. 444–450, 2010.

[148] G. Wolber, T. Seidel, F. Bendix, and T. Langer, "Molecule-pharmacophore superpositioning and pattern matching in computational drug design," *Drug discovery today*, vol. 13, no. 1-2, pp. 23–29, 2008.

[149] K. Poptodorov, T. Luu, and R. D. Hoffmann, "Pharmacophore model generation software tools," *Methods and Principles in Medicinal Chemistry*, vol. 32, p. 17, 2006.

[150] J. Zou, H.-Z. Xie, S.-Y. Yang, J.-J. Chen, J.-X. Ren, and Y.-Q. Wei, "Towards more accurate pharmacophore modeling: Multicomplex-based comprehensive pharmacophore map and most-frequent-feature pharmacophore model of cdk2," *Journal of Molecular Graphics and Modelling*, vol. 27, no. 4, pp. 430–438, 2008.

[151] C. Barillari, G. Marcou, and D. Rognan, "Hot-spots-guided receptor-based pharmacophores (hs-pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores," *Journal of chemical information and modeling*, vol. 48, no. 7, pp. 1396–1410, 2008.

[152] S. Macalino, S. Basith, N. Clavio, H. Chang, S. Kang, and S. Choi, "Evolution of in silico strategies for protein-protein interaction drug discovery," *Molecules*, vol. 23, no. 8, p. 1963, 2018.

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

[154] J. L. Robertson, "The lipid bilayer membrane and its protein constituents," *The Journal of General Physiology*, pp. jgp–201812153, 2018.

[155] A. R. Leach, *Molecular modelling: principles and applications*. Pearson education, 2001.

[156] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*, vol. 1. Elsevier, 2001.

[157] B. Alder and T. Wainwright, "Phase transition for a hard sphere system," *The Journal of chemical physics*, vol. 27, no. 5, pp. 1208–1209, 1957.

[158] B. J. Alder and T. E. Wainwright, "Studies in molecular dynamics. i. general method," *The Journal of Chemical Physics*, vol. 31, no. 2, pp. 459–466, 1959.

[159] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," *Nature*, vol. 267, no. 5612, p. 585, 1977.

[160] H.-X. Zhou, "Theoretical frameworks for multiscale modeling and simulation," *Current opinion in structural biology*, vol. 25, pp. 67–76, 2014.

[161] R. Lonsdale, J. N. Harvey, and A. J. Mulholland, "A practical guide to modelling enzyme-catalysed reactions," *Chemical Society Reviews*, vol. 41, no. 8, pp. 3025–3038, 2012.

[162] H. M. Senn and W. Thiel, "Qm/mm studies of enzymes," *Current opinion in chemical biology*, vol. 11, no. 2, pp. 182–187, 2007.

[163] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," *Journal of computational chemistry*, vol. 25, no. 9, pp. 1157–1174, 2004.

[164] A. D. MacKerell Jr, B. Brooks, C. L. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus, "Charmm: the energy function and its parameterization," *Encyclopedia of computational chemistry*, vol. 1, 2002.

[165] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids," *Journal of the American Chemical Society*, vol. 118, no. 45, pp. 11225–11236, 1996.

[166] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: the gromos force-field parameter sets 53a5 and 53a6," *Journal of computational chemistry*, vol. 25, no. 13, pp. 1656–1676, 2004.

[167] M. Levitt, "The birth of computational structural biology," *Nature Structural & Molecular Biology*, vol. 8, no. 5, p. 392, 2001.

[168] F. Jensen, *Introduction to computational chemistry*. John wiley & sons, 2017.

[169] A. K. Rappé, C. J. Casewit, K. Colwell, W. A. Goddard III, and W. Skiff, "Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations," *Journal of the American chemical society*, vol. 114, no. 25, pp. 10024–10035, 1992.

[170] T. A. Halgren, "Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94," *Journal of computational chemistry*, vol. 17, no. 5-6, pp. 490–519, 1996.

[171] Y. Mo, "A critical analysis on the rotation barriers in butane," *The Journal of organic chemistry*, vol. 75, no. 8, pp. 2733–2736, 2010.

[172] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, "Gromacs: fast, flexible, and free," *Journal of computational chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[173] R. Hockney, S. Goel, and J. Eastwood, "Quiet high-resolution computer models of a plasma," *Journal of Computational Physics*, vol. 14, no. 2, pp. 148–158, 1974.

[174] L. Verlet, "Computer" experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules," *Physical review*, vol. 159, no. 1, p. 98, 1967.

[175] B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije, "Lincs: a linear constraint solver for molecular simulations," *Journal of computational chemistry*, vol. 18, no. 12, pp. 1463–1472, 1997.

[176] D. P. Tieleman, "Methods and parameters for membrane simulations," *Molecular Simulations and Biomembranes: From Biophysics to Function, RSC*, pp. 1–25, 2010.

[177] M. Voógele and G. Hummer, "Divergent diffusion coefficients in simulations of fluids and lipid membranes," *The Journal of Physical Chemistry B*, vol. 120, no. 33, pp. 8722–8732, 2016.

[178] E. A. Dolan, R. M. Venable, R. W. Pastor, and B. R. Brooks, "Simulations of membranes and other interfacial systems using p21 and pc periodic boundary conditions," *Biophysical journal*, vol. 82, no. 5, pp. 2317–2325, 2002.

[179] H. J. Berendsen, J. v. Postma, W. F. van Gunsteren, A. DiNola, and J. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of chemical physics*, vol. 81, no. 8, pp. 3684–3690, 1984.

[180] D. J. Evans and B. L. Holian, "The nose–hoover thermostat," *The Journal of chemical physics*, vol. 83, no. 8, pp. 4069–4074, 1985.

[181] E. Koopman and C. Lowe, "Advantages of a lowe-andersen thermostat in molecular dynamics simulations," *The Journal of chemical physics*, vol. 124, no. 20, p. 204103, 2006.

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

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

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

[185] M. E. Tuckerman, J. Alejandre, R. López-Rendón, A. L. Jochim, and G. J. Martyna, "A liouville-operator derived measure-preserving integrator for molecular dynamics simulations in the isothermal–isobaric ensemble," *Journal of Physics A: Mathematical and General*, vol. 39, no. 19, p. 5629, 2006.

[186] G. J. Martyna, M. E. Tuckerman, D. J. Tobias, and M. L. Klein, "Explicit reversible integrators for extended systems dynamics," *Molecular Physics*, vol. 87, no. 5, pp. 1117–1157, 1996.

[187] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, "Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of chemical theory and computation*, vol. 4, no. 3, pp. 435–447, 2008.

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

[189] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh ewald method," *The Journal of chemical physics*, vol. 103, no. 19, pp. 8577–8593, 1995.

[190] T. Darden, D. York, and L. Pedersen, "Particle mesh ewald: An nŃ log (n) method for ewald sums in large systems," *The Journal of chemical physics*, vol. 98, no. 12, pp. 10089–10092, 1993.

[191] S. Izvekov and G. A. Voth, "A multiscale coarse-graining method for biomolecular systems," *The Journal of Physical Chemistry B*, vol. 109, no. 7, pp. 2469–2473, 2005.

[192] A. Y. Shih, A. Arkhipov, P. L. Freddolino, and K. Schulten, "Coarse grained protein- lipid model with application to lipoprotein particles," *The Journal of Physical Chemistry B*, vol. 110, no. 8, pp. 3674–3684, 2006.

[193] P. J. Bond and M. S. Sansom, "Insertion and assembly of membrane proteins via simulation," *Journal of the American Chemical Society*, vol. 128, no. 8, pp. 2697–2704, 2006.

[194] S. J. Marrink, A. H. De Vries, and A. E. Mark, "Coarse grained model for semiquantitative lipid simulations," *The Journal of Physical Chemistry B*, vol. 108, no. 2, pp. 750–760, 2004.

[195] A. Srivastava and G. A. Voth, "Hybrid approach for highly coarse-grained lipid bilayer models," *Journal of chemical theory and computation*, vol. 9, no. 1, pp. 750–765, 2012.

[196] M. G. Saunders and G. A. Voth, "Coarse-graining of multiprotein assemblies," *Current opinion in structural biology*, vol. 22, no. 2, pp. 144–150, 2012.

[197] T. A. Wassenaar, H. I. Ingoélfsson, M. Prieß, S. J. Marrink, and L. V. Schaófer, "Mixing martini: electrostatic coupling in hybrid atomistic–coarse-grained biomolecular simulations," *The Journal of Physical Chemistry B*, vol. 117, no. 13, pp. 3516–3530, 2013.

[198] M. Praprotnik, S. Matysiak, L. Delle Site, K. Kremer, and C. Clementi, "Adaptive resolution simulation of liquid water," *Journal of Physics: Condensed Matter*, vol. 19, no. 29, p. 292201, 2007.

[199] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. De Vries, "The martini force field: coarse grained model for biomolecular simulations," *The journal of physical chemistry B*, vol. 111, no. 27, pp. 7812–7824, 2007.

[200] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, "The martini coarse-grained force field: extension to proteins," *Journal of chemical theory and computation*, vol. 4, no. 5, pp. 819–834, 2008.

[201] C. A. López, A. J. Rzepiela, A. H. De Vries, L. Dijkhuizen, P. H. Huónenberger, and S. J. Marrink, "Martini coarse-grained force field: extension to carbohydrates," *Journal of Chemical Theory and Computation*, vol. 5, no. 12, pp. 3195–3210, 2009.

[202] J. J. Uusitalo, H. I. Ingoélfsson, P. Akhshi, D. P. Tieleman, and S. J. Marrink, "Martini coarse-grained force field: extension to dna," *Journal of chemical theory and computation*, vol. 11, no. 8, pp. 3932–3945, 2015.

[203] J. J. Uusitalo, H. I. Ingólfsson, S. J. Marrink, and I. Faustino, "Martini coarse-grained force field: extension to rna," *Biophysical Journal*, vol. 113, no. 2, pp. 246–256, 2017.

[204] H. Lee, A. H. de Vries, S.-J. Marrink, and R. W. Pastor, "A coarse-grained model for polyethylene oxide and polyethylene glycol: conformation and hydrodynamics," *The journal of physical chemistry B*, vol. 113, no. 40, pp. 13186–13194, 2009.

[205] T. Reddy, D. Shorthouse, D. L. Parton, E. Jefferys, P. W. Fowler, M. Chavent, M. Baaden, and M. S. Sansom, "Nothing to sneeze at: a dynamic and integrative computational model of an influenza a virion," *Structure*, vol. 23, no. 3, pp. 584–597, 2015.

[206] J. K. Marzinek, D. A. Holdbrook, R. G. Huber, C. Verma, and P. J. Bond, "Pushing the envelope: dengue viral membrane coaxed into shape by molecular simulations," *Structure*, vol. 24, no. 8, pp. 1410–1420, 2016.

[207] H. J. Risselada and S. J. Marrink, "Curvature effects on lipid packing and dynamics in liposomes revealed by coarse grained molecular dynamics simulations," *Physical Chemistry Chemical Physics*, vol. 11, no. 12, pp. 2056–2067, 2009.

[208] H. Koldso and M. S. Sansom, "Organization and dynamics of receptor proteins in a plasma membrane," *Journal of the American Chemical Society*, vol. 137, no. 46, pp. 14694–14704, 2015.

[209] M. Lelimousin and M. S. Sansom, "Membrane perturbation by carbon nanotube insertion: pathways to internalization," *Small*, vol. 9, no. 21, pp. 3639–3646, 2013.

[210] S. J. Marrink and D. P. Tieleman, "Perspective on the martini model," *Chemical Society Reviews*, vol. 42, no. 16, pp. 6801–6822, 2013.

[211] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers: Original Research on Biomolecules*, vol. 22, no. 12, pp. 2577–2637, 1983.

[212] X. Periole, M. Cavalli, S.-J. Marrink, and M. A. Ceruso, "Combining an elastic network with a coarse-grained molecular force field: structure, dynamics, and intermolecular recognition," *Journal of Chemical Theory and Computation*, vol. 5, no. 9, pp. 2531–2543, 2009.

[213] S. O. Yesylevskyy, L. V. Schäfer, D. Sengupta, and S. J. Marrink, "Polarizable water model for the coarse-grained martini force field," *PLoS computational biology*, vol. 6, no. 6, p. e1000810, 2010.

[214] R.-X. Gu, H. I. Ingoélfsson, A. H. de Vries, S. J. Marrink, and D. P. Tieleman, "Ganglioside-lipid and ganglioside-protein interactions revealed by coarse-grained and atomistic molecular dynamics simulations," *The Journal of Physical Chemistry B*, vol. 121, no. 15, pp. 3262–3275, 2016.

[215] C. Arnarez, J. J. Uusitalo, M. F. Masman, H. I. Ingoélfsson, D. H. de Jong, M. N. Melo, X. Periole, A. H. de Vries, and S. J. Marrink, "Dry martini, a coarse-grained force field for lipid membrane simulations with implicit solvent," *Journal of chemical theory and computation*, vol. 11, no. 1, pp. 260–275, 2014.

[216] J. L. Lahti, G. W. Tang, E. Capriotti, T. Liu, and R. B. Altman, "Bioinformatics and variability in drug response: a protein structural perspective," *Journal of The Royal Society Interface*, vol. 9, no. 72, pp. 1409–1437, 2012.

[217] A. R. Bradley, A. Echalier, M. Fairhead, C. Strain-Damerell, P. Brennan, A. N. Bullock, N. A. Burgess-Brown, E. P. Carpenter, O. Gileadi, B. D. Marsden, *et al.*, "The sgc beyond structural genomics: redefining the role of 3d structures

by coupling genomic stratification with fragment-based discovery," *Essays in biochemistry*, vol. 61, no. 5, pp. 495–503, 2017.

[218] N. M. Pearce, T. Krojer, A. R. Bradley, P. Collins, R. P. Nowak, R. Talon, B. D. Marsden, S. Kelm, J. Shi, C. M. Deane, *et al.*, "A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density," *Nature communications*, vol. 8, p. 15123, 2017.

[219] J.-P. Renaud, C.-w. Chung, U. H. Danielson, U. Egner, M. Hennig, R. E. Hubbard, and H. Nar, "Biophysics in drug discovery: impact, challenges and opportunities," *Nature Reviews Drug Discovery*, vol. 15, no. 10, p. 679, 2016.

[220] D. A. Keedy, Z. B. Hill, J. T. Biel, E. Kang, T. J. Rettenmaier, J. Brandao-Neto, N. M. Pearce, F. von Delft, J. A. Wells, and J. S. Fraser, "An expanded allosteric network in ptp1b by multitemperature crystallography, fragment screening, and covalent tethering," *Elife*, vol. 7, p. e36307, 2018.

[221] L. Pellegrini, D. F. Burke, F. von Delft, B. Mulloy, and T. L. Blundell, "Crystal structure of fibroblast growth factor receptor ectodomain bound to ligand and heparin," *Nature*, vol. 407, no. 6807, p. 1029, 2000.

[222] O. Ichihara, Y. Shimada, and D. Yoshidome, "The importance of hydration thermodynamics in fragment-to-lead optimization," *ChemMedChem*, vol. 9, no. 12, pp. 2708–2717, 2014.

[223] A. Fiser and A. Šali, "Modeller: generation and refinement of homology-based protein structure models," in *Methods in enzymology*, vol. 374, pp. 461–491, Elsevier, 2003.

[224] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, "Mdanalysis: a toolkit for the analysis of molecular dynamics simulations," *Journal of computational chemistry*, vol. 32, no. 10, pp. 2319–2327, 2011.

[225] T. Kawabata and N. Go, "Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites," *Proteins: Structure, Function, and Bioinformatics*, vol. 68, no. 2, pp. 516–529, 2007.

[226] T. Kawabata, "Detection of multiscale pockets on protein surfaces using mathematical morphology," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 5, pp. 1195–1211, 2010.

[227] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, pp. 832–837, 1956.

[228] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[229] A. Kolmogorov, "Sulla determinazione empirica di una lgge di distribuzione," *Inst. Ital. Attuari, Giorn.*, vol. 4, pp. 83–91, 1933.

[230] N. Smirnov, "Table for estimating the goodness of fit of empirical distributions," *The annals of mathematical statistics*, vol. 19, no. 2, pp. 279–281, 1948.

[231] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[232] S. Release, "2: Maestro, schrodinger, llc, new york, ny, 2017," *Received: February*, vol. 21, p. 2018, 2016.

[233] D. Julius and J. Nathans, "Signaling by sensory receptors," *Cold Spring Harbor perspectives in biology*, vol. 4, no. 1, p. a005991, 2012.

[234] M. C. Lagerström and H. B. Schiöth, "Structural diversity of g protein-coupled receptors and significance for drug discovery," *Nature reviews Drug discovery*, vol. 7, no. 4, p. 339, 2008.

[235] A. Jazayeri, S. P. Andrews, and F. H. Marshall, "Structurally enabled discovery of adenosine a2a receptor antagonists," *Chemical reviews*, vol. 117, no. 1, pp. 21–37, 2016.

[236] J. L. Kolakowski, "Gcrdb: a g-protein-coupled receptor database.," *Receptors & channels*, vol. 2, no. 1, pp. 1–7, 1994.

[237] R. Fredriksson, M. C. Lagerström, L.-G. Lundin, and H. B. Schiöth, "The g-protein-coupled receptors in the human genome form five main families. phylo-

genetic analysis, paralogon groups, and fingerprints," *Molecular pharmacology*, vol. 63, no. 6, pp. 1256–1272, 2003.

[238] U. Grauschopf, H. Lilie, K. Honold, M. Wozny, D. Reusch, A. Esswein, W. Schäfer, K. P. Rücknagel, and R. Rudolph, "The n-terminal fragment of human parathyroid hormone receptor 1 constitutes a hormone binding domain and reveals a distinct disulfide pattern," *Biochemistry*, vol. 39, no. 30, pp. 8878–8887, 2000.

[239] L. Chun, W.-h. Zhang, and J.-f. Liu, "Structure and ligand recognition of class c gpcrs," *Acta Pharmacologica Sinica*, vol. 33, no. 3, p. 312, 2012.

[240] P. C. Salinas, "Wnt signaling in the vertebrate central nervous system: from axon guidance to synaptic function," *Cold Spring Harbor perspectives in biology*, vol. 4, no. 2, p. a008003, 2012.

[241] H.-C. Huang and P. S. Klein, "The frizzled family: receptors for multiple signal transduction pathways," *Genome biology*, vol. 5, no. 7, p. 234, 2004.

[242] R. Ranjan, H. Dwivedi, M. Baidya, M. Kumar, and A. K. Shukla, "Novel structural insights into gpcr–β-arrestin interaction and signaling," *Trends in cell biology*, 2017.

[243] R. Leurs, R. A. Bakker, H. Timmerman, and I. J. de Esch, "The histamine h 3 receptor: from gene cloning to h 3 receptor drugs," *Nature reviews Drug discovery*, vol. 4, no. 2, p. 107, 2005.

[244] B. M. Koeppen and B. A. Stanton, *Berne and levy physiology e-book*. Elsevier Health Sciences, 2017.

[245] B. Taddese, L. M. Simpson, I. D. Wall, F. E. Blaney, and C. A. Reynolds, "Modeling active gpcr conformations," in *Methods in enzymology*, vol. 522, pp. 21–35, Elsevier, 2013.

[246] M. Congreve, J. M. Dias, and F. H. Marshall, "Structure-based drug design for g protein-coupled receptors," in *Progress in medicinal chemistry*, vol. 53, pp. 1–63, Elsevier, 2014.

[247] B. K. Shoichet and B. K. Kobilka, "Structure-based drug screening for g-protein-coupled receptors," *Trends in pharmacological sciences*, vol. 33, no. 5, pp. 268–272, 2012.

[248] A. Grossfield, "Recent progress in the study of g protein-coupled receptors with molecular dynamics computer simulations," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1808, no. 7, pp. 1868–1878, 2011.

[249] B. K. Kobilka and X. Deupi, "Conformational complexity of g-protein-coupled receptors," *Trends in pharmacological sciences*, vol. 28, no. 8, pp. 397–406, 2007.

[250] M. Filizola, *G Protein-coupled Receptors in Drug Discovery: Methods and Protocols*. Springer.

[251] X. Zhang and U. S. Eggert, "Non-traditional roles of g protein-coupled receptors in basic cell biology," *Molecular bioSystems*, vol. 9, no. 4, pp. 586–595, 2013.

[252] G. Stewart, J. Kniazeff, L. Prézeau, P. Rondard, J.-P. Pin, and C. Goudet, "Metabotropic receptors for glutamate and gaba," in *Pharmacology*, InTech, 2012.

[253] N. J. Smith and G. Milligan, "Allostery at g protein-coupled receptor homo-and heteromers: uncharted pharmacological landscapes," *Pharmacological reviews*, vol. 62, no. 4, pp. 701–725, 2010.

[254] J. R. Lane, A. Abdul-Ridha, and M. Canals, "Regulation of g protein-coupled receptors by allosteric ligands," *ACS chemical neuroscience*, vol. 4, no. 4, pp. 527–534, 2013.

[255] C. J. Wenthur, P. R. Gentry, T. P. Mathews, and C. W. Lindsley, "Drugs for allosteric sites on receptors," *Annual review of pharmacology and toxicology*, vol. 54, pp. 165–184, 2014.

[256] L. Wang, B. Martin, R. Brenneman, L. M. Luttrell, and S. Maudsley, "Allosteric modulators of g protein-coupled receptors: future therapeutics for complex physiological disorders," *Journal of Pharmacology and Experimental Therapeutics*, vol. 331, no. 2, pp. 340–348, 2009.

[257] P. J. Conn, A. Christopoulos, and C. W. Lindsley, "Allosteric modulators of gpcrs: a novel approach for the treatment of cns disorders," *Nature reviews Drug discovery*, vol. 8, no. 1, p. 41, 2009.

[258] W. L. DeLano, "Pymol," 2002.

[259] O. Lichtarge, H. R. Bourne, and F. E. Cohen, "Evolutionarily conserved galphabetagamma binding surfaces support a model of the g protein-receptor complex," *Proceedings of the National Academy of Sciences*, vol. 93, no. 15, pp. 7507–7511, 1996.

[260] C. J. Illingworth, P. D. Scott, K. E. Parkes, C. R. Snell, M. P. Campbell, and C. A. Reynolds, "Connectivity and binding-site recognition: Applications relevant to drug design," *Journal of computational chemistry*, vol. 31, no. 15, pp. 2677–2688, 2010.

[261] S. R. Kimura, H. P. Hu, A. M. Ruvinsky, W. Sherman, and A. D. Favia, "Deciphering cryptic binding sites on proteins by mixed-solvent molecular dynamics," *Journal of chemical information and modeling*, vol. 57, no. 6, pp. 1388–1401, 2017.

[262] M. Zocher, C. Zhang, S. G. Rasmussen, B. K. Kobilka, and D. J. Müller, "Cholesterol increases kinetic, energetic, and mechanical stability of the human β2-adrenergic receptor," *Proceedings of the National Academy of Sciences*, vol. 109, no. 50, pp. E3463–E3472, 2012.

[263] R. Saxena and A. Chattopadhyay, "Membrane cholesterol stabilizes the human serotonin 1a receptor," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1818, no. 12, pp. 2936–2942, 2012.

[264] C. M. Lopes, H. Zhang, T. Rohacs, T. Jin, J. Yang, and D. E. Logothetis, "Alterations in conserved kir channel-pip2 interactions underlie channelopathies," *Neuron*, vol. 34, no. 6, pp. 933–944, 2002.

[265] V. G. Romanenko, G. H. Rothblat, and I. Levitan, "Modulation of endothelial inward-rectifier k+ current by optical isomers of cholesterol," *Biophysical journal*, vol. 83, no. 6, pp. 3211–3222, 2002.

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

[267] N. C. Robinson, J. Zborowski, and L. H. Talbert, "Cardiolipin-depleted bovine heart cytochrome c oxidase: binding stoichiometry and affinity for cardiolipin derivatives," *Biochemistry*, vol. 29, no. 38, pp. 8962–8969, 1990.

[268] Y. Li, M. Ge, L. Ciani, G. Kuriakose, E. J. Westover, M. Dura, D. F. Covey, J. H. Freed, F. R. Maxfield, J. Lytton, *et al.*, "Enrichment of endoplasmic reticulum with cholesterol inhibits sarcoplasmic-endoplasmic reticulum calcium atpase-2b activity in parallel with increased order of membrane lipids implications for depletion of endoplasmic reticulum calcium stores and apoptosis in cholesterol-loaded macrophages," *Journal of Biological Chemistry*, vol. 279, no. 35, pp. 37030–37039, 2004.

[269] M. A. Hanson and R. Peach, "Structural biology of the s1p 1 receptor," in *Sphingosine-1-Phosphate Signaling in Immunology and Infectious Diseases*, pp. 23–53, Springer, 2014.

[270] S. B. Hansen, X. Tao, and R. MacKinnon, "Structural basis of pip 2 activation of the classical inward rectifier k+ channel kir2. 2," *Nature*, vol. 477, no. 7365, p. 495, 2011.

[271] P. J. Stansfeld, R. Hopkinson, F. M. Ashcroft, and M. S. Sansom, "Pip2-binding site in kir channels: definition by multiscale biomolecular simulations," *Biochemistry*, vol. 48, no. 46, pp. 10926–10933, 2009.

[272] M. R. Schmidt, P. J. Stansfeld, S. J. Tucker, and M. S. Sansom, "Simulation-based prediction of phosphatidylinositol 4, 5-bisphosphate binding to an ion channel," *Biochemistry*, vol. 52, no. 2, pp. 279–281, 2012.

[273] P. L. Yeagle, "Non-covalent binding of membrane lipids to membrane proteins," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1838, no. 6, pp. 1548–1559, 2014.

[274] P. J. Stansfeld and M. S. Sansom, "Molecular simulation approaches to membrane proteins," *Structure*, vol. 19, no. 11, pp. 1562–1572, 2011.

[275] A. Laganowsky, E. Reading, J. T. Hopper, and C. V. Robinson, "Mass spectrometry of intact membrane protein complexes," *Nature protocols*, vol. 8, no. 4, p. 639, 2013.

[276] R. O. Dror, R. M. Dirks, J. Grossman, H. Xu, and D. E. Shaw, "Biomolecular simulation: a computational microscope for molecular biology," *Annual review of biophysics*, vol. 41, pp. 429–452, 2012.

[277] C. Arnarez, J.-P. Mazat, J. Elezgaray, S.-J. Marrink, and X. Periole, "Evidence for cardiolipin binding sites on the membrane-exposed surface of the cytochrome bc 1," *Journal of the American Chemical Society*, vol. 135, no. 8, pp. 3112–3120, 2013.

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

[279] S. McLaughlin, J. Wang, A. Gambhir, and D. Murray, "Pip2 and proteins: interactions, organization, and information flow," *Annual review of biophysics and biomolecular structure*, vol. 31, no. 1, pp. 151–175, 2002.

[280] Z. Huang and F. C. Szoka Jr, "Sterol-modified phospholipids: cholesterol and phospholipid chimeras with improved biomembrane properties," *Journal of the American Chemical Society*, vol. 130, no. 46, pp. 15702–15712, 2008.

[281] M. R. Krause and S. L. Regen, "The structural role of cholesterol in cell membranes: from condensed bilayers to lipid rafts," *Accounts of chemical research*, vol. 47, no. 12, pp. 3512–3521, 2014.

[282] C. A. Doupnik, N. Davidson, and H. A. Lester, "The inward rectifier potassium channel family," *Current opinion in neurobiology*, vol. 5, no. 3, pp. 268–277, 1995.

[283] B.-C. Suh and B. Hille, "Pip2 is a necessary cofactor for ion channel function: how and why?," *Annu. Rev. Biophys.*, vol. 37, pp. 175–195, 2008.

[284] J.-P. Vincent, J. Mazella, and P. Kitabgi, "Neurotensin and neurotensin receptors," *Trends in pharmacological sciences*, vol. 20, no. 7, pp. 302–309, 1999.

[285] S. E. Leeman and R. E. Carraway, "Neurotensin: discovery, isolation, characterization, synthesis and possible physiological roles," *Annals of the New York Academy of Sciences*, vol. 400, no. 1, pp. 1–16, 1982.

[286] P. M. Dijkman and A. Watts, "Lipid modulation of early g protein-coupled receptor signalling events," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1848, no. 11, pp. 2889–2897, 2015.

[287] J. Oates, B. Faust, H. Attrill, P. Harding, M. Orwick, and A. Watts, "The role of cholesterol on the activity and stability of neurotensin receptor 1," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1818, no. 9, pp. 2228–2233, 2012.

[288] H.-Y. Yen, K. K. Hoi, I. Liko, G. Hedger, M. R. Horrell, W. Song, D. Wu, P. Heine, T. Warne, Y. Lee, *et al.*, "Ptdins (4, 5) p 2 stabilizes active states of gpcrs and enhances selectivity of g-protein coupling," *Nature*, vol. 559, no. 7714, p. 423, 2018.

[289] M. de Lera Ruiz, Y.-H. Lim, and J. Zheng, "Adenosine a2a receptor as a drug discovery target," *Journal of medicinal chemistry*, vol. 57, no. 9, pp. 3623–3650, 2013.

[290] W. Song, H.-Y. Yen, C. V. Robinson, and M. Sansom, "State-dependent lipid interactions with the a2a receptor revealed by md simulations using in vivo-mimetic membranes," *bioRxiv*, 2018.

[291] D. J. Robbins, D. L. Fei, and N. A. Riobo, "The hedgehog signal transduction network," *Sci. Signal.*, vol. 5, no. 246, pp. re6–re6, 2012.

[292] C. Wang, H. Wu, V. Katritch, G. W. Han, X.-P. Huang, W. Liu, F. Y. Siu, B. L. Roth, V. Cherezov, and R. C. Stevens, "Structure of the human smoothened receptor bound to an antitumour agent," *Nature*, vol. 497, no. 7449, p. 338, 2013.

[293] E. F. Byrne, R. Sircar, P. S. Miller, G. Hedger, G. Luchetti, S. Nachtergaele, M. D. Tully, L. Mydock-McGrane, D. F. Covey, R. P. Rambo, *et al.*, "Structural basis of smoothened regulation by its extracellular domains," *Nature*, vol. 535, no. 7613, p. 517, 2016.

[294] G. Hedger, H. Koldsoe, M. Chavent, C. Siebold, R. Rohatgi, and M. Sansom, "Cholesterol interaction sites on the transmembrane domain of the hedgehog signal transducer and class f g protein-coupled receptor smoothened," *bioRxiv*, 2018.

[295] M. Grieben, A. C. Pike, C. A. Shintre, E. Venturi, S. El-Ajouz, A. Tessitore, L. Shrestha, S. Mukhopadhyay, P. Mahajan, R. Chalk, *et al.*, "Structure of the polycystic kidney disease trp channel polycystin-2 (pc2)," *Nature Structural and Molecular Biology*, vol. 24, no. 2, p. 114, 2017.

[296] S. Gonzalez-Perrett, K. Kim, C. Ibarra, A. E. Damiano, E. Zotta, M. Batelli, P. C. Harris, I. L. Reisin, M. A. Arnaout, and H. F. Cantiello, "Polycystin-2, the protein mutated in autosomal dominant polycystic kidney disease (adpkd), is a ca2+-permeable nonselective cation channel," *Proceedings of the National Academy of Sciences*, vol. 98, no. 3, pp. 1182–1187, 2001.

[297] S. Gonzalez-Perrett, M. Batelli, K. Kim, M. Essafi, G. Timpanaro, N. Moltabetti, I. L. Reisin, M. A. Arnaout, and H. F. Cantiello, "Voltage dependence and ph regulation of human polycystin-2-mediated cation channel activity," *Journal of Biological Chemistry*, vol. 277, no. 28, pp. 24959–24966, 2002.

[298] V. E. Torres and P. C. Harris, "Autosomal dominant polycystic kidney disease: the last 3 years," *Kidney international*, vol. 76, no. 2, pp. 149–168, 2009.

[299] P. Huang, S. Zheng, B. M. Wierbowski, Y. Kim, D. Nedelcu, L. Aravena, J. Liu, A. C. Kruse, and A. Salic, "Structural basis of smoothened activation in hedgehog signaling," *Cell*, 2018.

[300] R. D. Taylor, M. MacCoss, and A. D. Lawson, "Rings in drugs: miniperspective," *Journal of medicinal chemistry*, vol. 57, no. 14, pp. 5845–5859, 2014.

[301] A. B. Poma, M. Cieplak, and P. E. Theodorakis, "Combining the martini and structure-based coarse-grained approaches for the molecular dynamics studies of conformational transitions in proteins," *Journal of chemical theory and computation*, vol. 13, no. 3, pp. 1366–1374, 2017.

[302] A. Winter, A. P. Higueruelo, M. Marsh, A. Sigurdardottir, W. R. Pitt, and T. L. Blundell, "Biophysical and computational fragment-based approaches to targeting protein–protein interactions: applications in structure-guided drug discovery," *Quarterly reviews of biophysics*, vol. 45, no. 4, pp. 383–426, 2012.