THE UNIVERSITY OF CHICAGO
UNIVERSITÉ DE LORRAINE
LABORATOIRE DE PHYSIQUE ET CHIMIE THÉORIQUES

EFFICIENT SAMPLING OF COMPLEX BIOMOLECULAR ASSEMBLIES USING MOLECULAR SIMULATIONS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY
C2MP

BY
FLORENCE SZCZEPANIAK

CHICAGO, ILLINOIS
DECEMBER 2024

C2MP
Laboratoire de Physique et Chimie
Théoriques – UMR CNRS 7019

University of Chicago

<u>Thèse</u>

Présentée et soutenue publiquement pour l'obtention du titre de

DOCTEUR DE l'UNIVERSITE DE LORRAINE

Mention : Chimie

par **Florence SZCZEPANIAK**

Sous la direction de François DEHEZ et Benoît ROUX

Efficient sampling of complex biomolecular assemblies using molecular simulations

23 octobre 2024

Membres du jury :

Directeur(s) de thèse :	**Dr. François Dehez**	**Université de Lorraine, France**
	Pr. Benoît Roux	**University of Chicago, USA**
Président du jury :	**Dr. Christophe Chipot**	**Université de Lorraine, France**
Rapporteurs :	**Pr. Nathalie Reuter**	**University of Bergen, Norway**
	Pr. Bettina Keller	**Freie Universität Berlin, Germany**
Examinateurs :	**Pr. Aaron Dinner**	**University of Chicago, USA**
	Pr. Laura Gagliardi	**University of Chicago, USA**
	Dr. Emmanuelle Bignon	**Université de Lorraine, France**

ABSTRACT

Atomic-level information is essential to describe the structure and dynamics of biomolecular assemblies. The work presented in this thesis aims to explore and enhance computational techniques explaining the formation of complexes, quantifying binding free energies or describing the dynamics of multi-components systems.

I first developed a protocol to compute the binding free energy of a ligand buried in a membrane protein. It relies on alchemical transformations carried out in a rigorous statistical mechanical framework. The protocol is distributed within the BFEE2 plugin, a tool designed to assist the end user in preparing all the necessary input files and performing the post-treatment of the simulations towards the final estimate of the binding affinity.

Molecular Dynamics (MD) and alchemical simulations have been employed to provide insights into the formation of specific protein complexes in terms of structure and dynamics. The set of Dpr and DIP proteins, which play a key role in the neuromorphogenesis in the nervous system of *Drosophila melanogaster*, offer a rich paradigm to learn about protein-protein recognition. Many members of the DIP subfamily cross-react with several members of the Dpr family and vice-versa. While there exists a total of 231 possible Dpr-DIP heterodimer, only 57 "cognate" pairs have been detected by surface plasmon resonance (SPR) experiments, suggesting that the remaining 174 pairs have low or unreliable binding affinity. Here I assessed the performance of computational approaches to in quantifying the binding affinities between Dpr and DIP proteins and I identified by means of a series of point mutations, the interfacial residues governing the specificity of the recognition process.

Building on alchemical transformations, I developed a hybrid nonequilibrium molecular dynamics - Monte Carlo (neMD/MC) simulation method to aimed at enhancing the sampling of inhomogeneous membranes, circumventing the slow lateral diffusion of the various constituents. Randomly chosen lipid molecules are swapped to generate configurations that are subsequently accepted or rejected according to a Metropolis criterion based on the alchemical work associated to the attempted swap calculated via a short trajectory. The performance of the hybrid neMD/MC algorithm and its ability to sample the distribution of lipids near a transmembrane helix carrying a net charge are illustrated for a binary mixture of charged and zwitterionic lipids. To enforce equilibrium between a simulated system and an infinite surrounding bath, a modified version of the neMD/MC algorithm was developed, in which a randomly chosen lipid molecule in the simulated system is swapped with a lipid picked in a separate system standing as a thermodynamic "reservoir" with the desired mole fraction for all lipid components.

Membrane proteins function has been shown to depend on the lipid organization within the membrane either through averaged bulk effect or specific binding. A well-known class of protein exhibiting such a dependance is the family of pentameric ligand-gated ion channels (pLGICs). Upon the binding of a neurostransmitter, the conformation of these proteins changes establing a ionic current at the synapse junctions, transforming therby a chemical into an electric signal. Here, we generated several MD trajectories of various agonist-bound structures of nicotonic acethlycholine receptors solved by cryoEM, providing a molecular basis shedding light on the desensitization process. The conductivity and the stability of the pore of the pLGICs in a desensitized state are measured. The functions of these proteins have also been shown to depend on lipid composition. Finally, we employed alchemical tranformations to quantify the relative binding affinities of anionic and zwitterionic lipids at putative pLGIC binding sites, enlightening how lipids modulate the fonction of these proteins.

RÉSUMÉ EN FRANÇAIS

Les organismes vivants sont fait d'une ou de plusieurs cellules. Une cellule est un ensemble complexe de molécules - complexe dans sa composition mais aussi dans son fonctionement. Être capable de comprendre et d'étudier différents éléments d'une cellule est au coeur de beaucoup de sujets de recherche. Les informations au niveau atomique sont essentielles pour décrire la structure et la dynamique des complexes biomoléculaires. Les travaux présentés dans cette thèse visent à explorer et à améliorer les techniques informatiques expliquant la formation de complexes, quantifiant les énergies libres de liaison ou décrivant la dynamique de systèmes multi-composants.

J'ai d'abord développé un protocole pour calculer l'énergie libre de liaison d'un complexe protéine-ligand. Il s'appuie sur des transformations alchimiques réalisées dans un cadre mécanique statistique rigoureux. Le protocole est distribué au sein du plugin BFEE2, un outil conçu pour aider l'utilisateur à préparer tous les fichiers d'entrée nécessaires et à effectuer le post-traitement des simulations d'estimation d'affinité de liaison. Je me suis particulièrement intéressée à l'énergie libre de liason d'un complexe fait d'une protéine membranaire et d'un ligand enfoui dans la protéine. La structure de la membrane impose des changements dans l'écriture des fichiers d'entrée, qui ont du être implémenté dans le plugin. Le site de liaison peu accessible au solvant a également demandé des précautions supplémentaires, pour assurer la réversibilité des transformations ainsi qu'une hydratation suffisante du site de liaison dans l'état découplé.

La dynamique moléculaire (MD) et les simulations alchimiques ont été utilisées pour fournir des informations sur la formation de complexes protéiques spécifiques en termes de structure et de dynamique. L'ensemble des protéines Dpr et DIP, qui jouent un rôle clé dans la neuromorphogenèse du système nerveux de *Drosophila melanogaster*, offre un paradigme riche pour en apprendre davantage sur la reconnaissance protéine-protéine. De nombreux membres de la sous-famille DIP réagissent de manière croisée avec plusieurs membres de la famille Dpr et vice-versa. Bien qu'il existe un total de 231 hétérodimères Dpr-DIP possibles, seules 57 paires « apparentées » ont été détectées par des expériences de résonance plasmonique de surface (SPR), ce qui suggère que les 174 paires restantes ont une affinité de liaison faible ou peu fiable. Les complexes "apparentés" ont initialement été attribués à une complémentarité des structures des protéines. Or, deux isoprotéines, c'est à dire deux protéines avec une structure secondaire et une forme similaire, mais des variations de séquences ne vont pas se lier avec les mêmes partenaires. Des études plus récentes suggèrent une sélectivité induite par des interactions spécifiques entre résidues à l'interface du complexe DIP-Dpr. Ici, j'ai évalué les performances des approches informatiques pour quantifier les affinités de liaison entre les protéines Dpr et DIP et j'ai identifié, au moyen d'une série de mutations ponctuelles, les résidus interfaciaux régissant la spécificité du processus de reconnaissance. Les mutations de résidues par transformations alchimiques donnent des résultats proches des données expérimentales, mais sont trop coûteuses pour être calculées sur toutes les interfaces. Elles ont toutefois permit de mettre en évidence certains résidues clefs dans le méchanisme de sélectivité lors de la formation des complexes.

En m'appuyant sur les transformations alchimiques, j'ai développé une méthode de simulation hybride dynamique moléculaire hors équilibre - Monte Carlo (neMD/MC) visant à améliorer l'échantillonnage de membranes inhomogènes, en contournant la lente diffusion latérale des différents constituants. Les protéines membranaires représentent 50 % de la masse d'une membrane plasmique, 23 % des protéines et environ 60 % des cibles de médicaments. Elles représentent donc un important champ de recherche. Des études suggèrent que la composition de la membrane autour de ces protéines peut impacter leur dynamique et leur fonction. Améliorer l'échantiollonnage de la distribution des lipides dans une membrane permet donc une meilleure modélisation des protéines membranaires. Dans la méthode neMD/MC, deux molécules, choisies aléatoirement dans le système, sont échangées au cours d'une transformation alchimique. Le travail associé à cet échange est calculé par intégration

thermodynamique. Un critère de Métropolis, basé sur ce travail, permet de n'accepter que les échanges qui mènent vers l'équilibre thérmodynamique. Après chaque échange, une trajectoire de dynamique moléculaire est calculée pour équilibrer la membrane. Pour simplifier le développement de la méthodologie, j'ai tout d'abord étudié un système de particules de Lennard-Jones. Le but de cette étape est d'avor un système simplifié, uniquement fait de sphères toutes de même rayon, qui ne présente donc pas de difficulté stérique, et dont l'équilibre thermodynamique est connu par construction du champ de force. Pour appréhender les difficultés associées aux conformations des lipides, j'ai ensuite appliqué la méthode neMD/MC à l'échange de deux lipides dans l'eau. Un lipide est choisi anionique, l'autre est zwiterrionique, et les deux ont une chaine carbonnée similaire. Les échanges de ces deux lipides dans l'eau sont - en théorie - associé à un travail proche de 0 par construction du système. Cette étape a permit le développement d'une stratégie pour l'échange alchimique des lipides qui diminue la perturbation stérique. Enfin, la méthode est généralisée aux membranes. Les performances de l'algorithme hybride neMD/MC et sa capacité à échantillonner la distribution des lipides à proximité d'une hélice transmembranaire portant une charge nette sont illustrées pour un mélange binaire de lipides chargés et zwitterioniques. Différents systèmes sont étudiés pour explorer les forces et faiblesses de l'algorithme. Une première modification est proposée pour améliorer l'efficacité, en imposant qu'au moins un lipide impliqué dans l'échange provienne de l'environement de l'hélice transmembranaire. Cette modification permet d'accélere l'échantillonage des intéractions spécifiques péptide-lipides.

Pour maintenir l'équilibre entre un système simulé et un bain environnant infini, une version modifiée de l'algorithme neMD/MC a été développée, dans laquelle un lipide choisi au hasard dans le système simulé est échangée avec un lipide prélevé dans un système séparé faisant office de système thermodynamique. « réservoir » avec la fraction molaire souhaitée pour tous les composants lipidiques. Cette nouvelle version permet notamment d'échantilloner les configurations de membranes avec des lipides à très basses concentration à moindre coût. Une telle modification impose des changements dans la sélection des lipides, ainsi que des précautions pour conserver la charge du système et du reservoir au cours de tous les échanges. Pour évaluer la validité ainsi que l'efficacité de la version modifée de l'algorithme, deux membranes composées exclusivement d'un lipide zwiterionnique (une avec l'hélice transmembranaire précedement utilisée, et l'autre sans), sont équilibrées avec deux réservoirs composé de lipides zwiterionniques at anioniques, avec différentes concentrations. Cette méthode montre notamment que l'hélice améliore la convergence des simulations, et que les sites de liaisons peptide-lipides peuvent être élucidé.

En parallèle de ces simulations, la dynamique des canaux ioniques pentamères ligand-dépendants (pLGIC) est étudiée. PLGICs sont des recepteurs de neurotransmetteurs, présents dans le système nerveux ou aux jonctions neuromusculaires. Sans neurotransmetteur, les pLGICs sont dans une conformation fermée qui ne permet pas l'échange d'ion de part et d'autre de la membrane. Lors la liaison avec son agoniste, la conformation de la protéine change, pour ouvrir le pore et permettre un transfert d'ion. Après un long temps d'exposition à l'agoniste, la protéine est dans un état désensibilisé. Différentes études sur différents états désensibilisés de protéine recepteurs de nicotine montre que la dynamique de la protéine varie en fonction de la structure. Soit le pore est hydraté et toujours conducteur, à une intensité plus faible que si la protéine est dans l'état ouvert, soit le pore collapse, ne permettant plus le transfert d'eau ou d'ion. À l'aide de diverses structures liées à des récepteurs nicotiniques, des simulations MD sont calculées. La conductivité et la stabilité du pore des pLGICs à l'état désensibilisé sont mesurées. Ces différentes dynamiques des protéines sont observées dans les différentes trajectoires de MD générées.

Il a également été démontré que les fonctions de ces protéines dépendent de la composition lipidique. Par exemple, une membrane de POPC pure semble décorréler la liaison avec l'agoniste et le mechanisme d'ouverture du pore. Un protocole pour calculer précisement la différence d'affinité protéine-lipide est développé dans le cadre d'un recepteur nicotinique dans une membrane de POPC et POPA (ratio 3:2). Différents protocoles sont testés. L'usage de restraintes pour réduire l'entropie configurationnelle du système est une solution utilisée très fréquemment lors des transformations alchimiques pour améliorer la convergence des simulations. Deux protocoloes, un avec restraintes et l'autre sans, sont donc testés pour comparer la validité et l'efficacité des méthodes. Des précautions pour la conservation de la charge du système sont aussi testées. Des erreurs de calculs peuvent être créees si la charge totale du système n'est pas concervée. Une stratégie consiste à coupler un contre-ion au lipide chargé, pour maintenir la perturbation des charges à zéro. Cette route est également simulée dans le cadre du calcul de la différence d'affinité protéine-lipide. Ce projet n'est pas terminé au moment de l'écriture de ce manuscrit, les différentes routes sont donc détaillées mais aucune conclusion ne peut être déduite des simulations pour l'instant.

L'optimisation de la stratégie pour réaliser des simulations alchimiques a été un aspect majeur de mon doctorat. Dans les simulations à l'équilibre et hors équilibre, l'ajout de contraintes réduit le nombre de degrés de liberté, réduisant ainsi l'entropie configurationnelle à échantillonner le long du transformation. L'utilisation de restraintes était également essentielle pour l'échange de lipides dans la membrane avec l'algorithme neMD/MC. Dans le futur, il serait intéressant d'appliquer les méthodologies développé dans cette thèse pour explorer la modulation de la fonction des protéines membranaires par la membrane. La méthode neMD/MC peut générer une image plus complexe et plus réaliste des configuration des lipides entourant la protéine. La méthode pour calculer l'affinité protéine-lipide peut affiner le modèle en décrivant précisément les lipides dans les sites de liaison spécifiques. La réponse des protéines membranaires aux ligands, par exemple dans le contexte de la conception de médicaments, pourrait alors être prédite avec plus de précision par des méthodes informatiques, car elles seraient plus efficaces pour modéliser des phénomènes biologiques réalistes.

To my loved ones

"Science also refers to the type of knowledge that can be rationally explained and reliably applied. It’s like a philosophy. Kind of."

The Aquabats, *Doing Science!*

“Il y a deux réponses à cette question, comme à toutes les questions : celle du poète et celle du savant. Laquelle veux-tu en premier ?”

Pierre Bottero, *Ellana*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to thank everyone who helped and supported me through this PhD. I am very grateful for all of you.

Pour commencer, je voudrais remercier Elise Dumont et Paul, qui m'ont fait découvrir la chimie théorique, et m'ont plus généralement ouvert la porte vers le monde de la recherche. Sans cette impulsion, cette thèse n'existerait probablement pas.

I would like to thank Dragi Karevski and Robert J. Zimmer for accepting me in the LPCT and at the University of Chicago. I would like to extend my acknowledgments to Séverine Bonenberger and Vera Dragisich, for their compassion and help to navigate through the administrative maze that the dual PhD is.

I am grateful to my committee members Professor Aaron Dinner and Professor Nathalie Reuter, for their scientific discussions, their compassion and support.

I also want to thank all of the colleagues I met during this PhD. Lynn Bluchhoz, Haochuan Chen, Angela Barragan, Ramon Mendoza Uriarte, Francis Alipranti, Trayder Thomas, Emma Goulard Coderc De Lacam, Marharyta Blazhynska, Sarah Moe, Jonathan Harris, Spencer Guo, Jose Luis Guerra, Noah Schwartz, Rachel Yougworth, Johnathan Thirman, Yiwei He and Werner Treptow. I appreciate your help and your kindness through this journey.

Mon équilibre (pas toujours bien balancé) entre la vie personnelle et ma thèse se doit également de remercier Guillaume, Victor, Muriel, Ervan et Alex. Merci de m'avoir proposé des pauses et des escapades loin de mon quotidien et ma routine. Merci à ma famille pour leurs encouragements et leur soutien. Merci à Clément, qui ne m'a pas seulement supporté, mais qui m'a aussi appris que j'étais beaucoup plus forte et combattante que ce que je pensais. Thank you to Kevin, who made the roller coasters of emotions that is this last year of PhD so much smoother and easier to live. Enfin, il est impossible pour moi de ne pas remercier Lysis, pour être là, avec moi, toujours.

Je voudrais aussi remercier Chris, pour son temps, sa patience et son aide. Enfin, je ne peux pas remercier assez mes deux directeurs de thèse, Benoit et François. Merci à vous pour l'aide que vous m'avez apportée, pour m'avoir poussée et encouragée à me dépasser. Surtout, merci pour l'humanité dont vous avez fait preuve toute au long de cette thèse, merci d'avoir cru en moi, merci pour votre soutien lors de toutes mes périodes de doute.

CHAPTER 1

INTRODUCTION: MOLECULAR DYNAMICS SIMULATIONS OF BIOMOLECULAR ASSEMBLIES

1.1 Molecular assemblies in biology

1.1.1 Biological concepts

The definition of "life" does not find a consensus despite a lot of efforts to phrase it, but it is accepted that living organisms are made of one or several cells (Alberts et al. (2002); Trifonov (2011)). Cells are complex molecular assemblies, made of very different categories of molecules. Because of their roles in life, it is of major importance to understand what cells are made of and how they work, but due to the variety of the composition and of the phenomena, cells represent a large field of study.

There are two categories of cells: the prokaryotes and eukaryotes. Prokaryotes include bacteria and exclusively unicellular organisms, especially *Escherichia Coli*, or *E. Coli*, one of the most studied and well-known organism in biology (Blount (2015)). The prokaryote organisms can be found in various environment, from the hot sulfur springs to the human gut microbiome. Cell sizes range generally between 1 to 10 μm, and they are composed of billions of molecules. Figure 1.1 shows different length-scales of components of a cell.

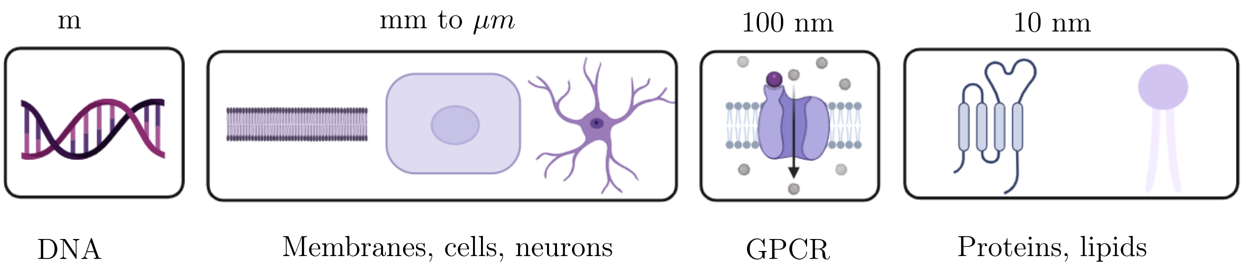


Figure 1.1: Length-scale of molecular assemblies in the cell. GPCR stands for G protein-coupled receptors, a family of membrane proteins.

The large variety of molecules can be sorted in four major categories: the sugars, the nucleotides, the fatty acids and the amino acids. The sugars are necessary to fuel cells with energy, and are components of the cell wall. Nucleotides are the components of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), coding the genomic identity of the organisms (Alberts et al. (2002)). DNA strands are macromolecules that can be extremely long, up to the order of the meter for the human genome. Fatty acids are involved in the plasma membrane, and the amino acids form the peptides and proteins that are found in the cell wall, the plasma membrane and the cells (Cooper (2000); Alberts et al. (2002)). Eukaryotic cells are similar to prokaryotes, yet they differ in some ways. Eukaryotes can form multicellular organisms, DNA is contained into a nucleus, and they can have a cytoskeleton. They are therefore usually larger and can reach up to one mm length. Because they have a more complex structure, they also involve additional biological phenomena, for example, generating energy through the oxydation of ATP, which is considered to be the main power source of cells (Bonora et al. (2012)).

In addition to a large variety of molecules, cells host myriad of biological processes spanning various time scales. A few nanoseconds is enough to observe the permeation of ions crossing membranes through channel proteins. On the other extreme, it can take minutes to hours to translate the DNA or divide a cell (Cooper (2000); Alberts et al. (2002)) (Figure 1.2).

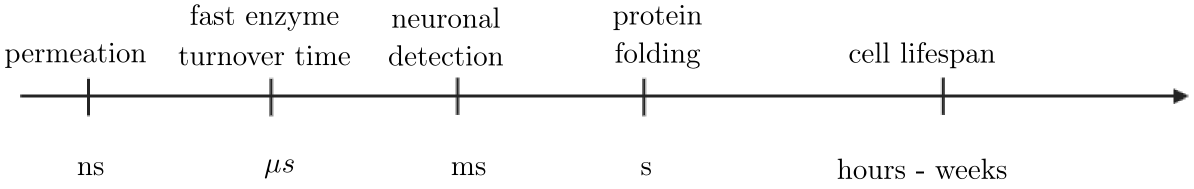


Figure 1.2: Range of typical timescales of cellular processes.

In my Ph.D., I focus essentially on two major categories of biomolecules: the proteins and the lipids. Lipids have three major functions. They capture triacylglycerol esters and steryl esters in droplets, ensuring a function of energy reservoir. Vesicules store carbon source

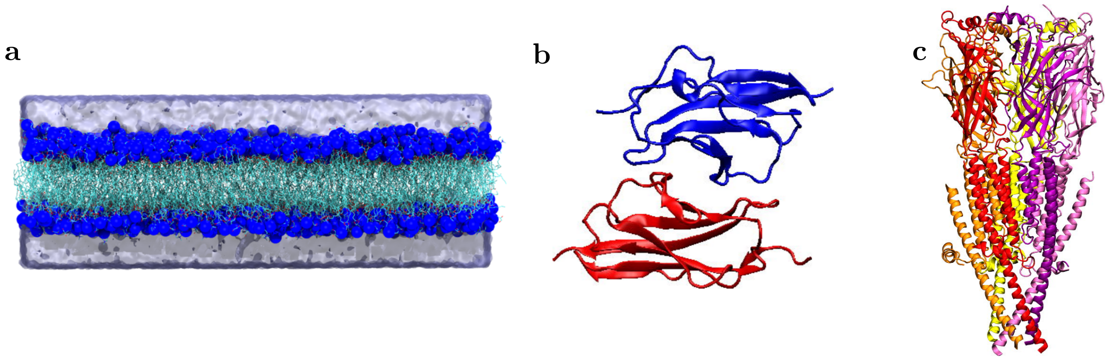


Figure 1.3: **a** Membrane. In dark blue, the polar heads, in cyan the carbon tails. **b** Protein-protein complex. **c** Membrane protein.

used for ATP production (Bonora et al. (2012)). They act in signal transduction and in molecular recognition processes. For example, the lipid rafts (assemblies of ordered lipids and proteins floating in the disordered bilayers of membranes) are in charge of signaling and protein trafficking (Sodt et al. (2015); Ghysels et al. (2019)). The main function of lipids is to form the plasma membranes, which are essential components of the cells (Cooper (2000); Van Meer et al. (2008)). Lipids are amphiphilic molecules that self-assemble into bilayers (see Figure 1.3 **a**). Membranes comprise a large variety of lipids (Wallin and Heijne (1998)) and membrane proteins (Watson (2015); Alberts et al. (2002)). These components assemble in different proportions, giving specific biological properties to membranes (Dowhan (1997); Harayama and Riezman (2018)). Mixing of membrane domains are processes requiring a few tens of microseconds for reorganizing of a few hundreds of nanometers (Sodt et al. (2015)). It is now well-established that lipid composition affects not only the mechanical properties of membranes but also modulates the function of membrane proteins through distinct mechanisms (Laganowsky et al. (2014); Hénault et al. (2019)). Despite recent advances in experimental studies of lipid-protein interaction, including high-resolution structures capturing lipid binding sites, the understanding of the spatial organisation of membranes around proteins remains limited (Corradi et al. (2019)).

Proteins are chains of amino acids, with specific sequences, conformations and biochemical properties (Alberts et al. (2002)). There is a wide diversity of proteins, with very different structures and functions. The human genome encodes at least 20,000 proteins (Ponomarenko et al. (2016)). To narrow it down, this thesis is focusing on protein-protein complexes and membrane proteins (see Figure 1.3 **b** and **c**), as they are mediating many intra- and inter-cellular communication processes (Zipursky and Sanes (2010); Marquart et al. (1983); Özkan et al. (2013, 2014)). Protein-protein complexes represent an interesting challenge in biology as most existing protein-protein "interactome" datasets lack information for extracellular interactions (Zipursky and Sanes (2010)), despite mediating cell-cell communication, adhesion, initiation of signaling events or the connection between neurons. Slight variation of the sequence of the surface proteins drive the specificity of the protein:protein complexes. Understanding the selectivity mechanism can be challenging. For example, in the case of two families of proteins involved in the neuromorphogenesis, the DIP and Dpr proteins, the specificity of the interactions has firstly been assigned to a shape complementarity between proteins (Carrillo et al. (2015)). Yet, as two isoproteins (with the same shape and secondary structure but some differences in the sequences) do no have the same partners, this hypothesis has been ruled out. Recent studies have suggested that the selectivity is rooted in the interactions between specific residues (Cosmanescu et al. (2018)).

Membrane proteins are not only a large group of proteins, but they also constitute about 50 % of the mass of a plasma membrane (Alberts et al. (2002)), are about 23 % of the proteins, and there are the target of almost 60 % of the drugs (Yin and Flynn (2016)). Among other functions, they can act as channels, allowing ions to cross the membrane. The ability to control the flow of ions into and out of the cells can be modulated by the lipids surrounding the channel, by association with ligands, etc (Polovinkin et al. (2018); Hénault et al. (2019); Zarkadas et al. (2022)). An atomistic insight is required to understand the functionnal dynamics of membrane proteins, but the study of such biological assemblies remains extremely challenging (Roux (2011)). Dynamic phenomena in membrane proteins happen over a large timescale, from a few nanoseconds to observe permeation events to a few milliseconds to observe conformational changes (Nury et al. (2010); Rao et al. (2021)).

1.1.2 Experimental techniques to study biomolecular assemblies

Experimental techniques have been developed to elucidate the structures and study the biophysical properties of molecular assemblies. An extensive description of all techniques can not be covered by this thesis. Here, I present a few techniques to which I will refer in the next chapters.

Crystallography (X-ray), Nuclear Magnetic Resonance (NMR) spectroscopy and cryo-electron microscopy (cryo-EM) are techniques commonly employed to determine the structures of proteins (Milne et al. (2013); Faraggi et al. (2018)). X-ray crystallography first requires protein crystallization which is often highly challenging, especially for membrane proteins. The spacial arrangement of the atoms is deduced from the diffraction of light on the crystal (Brünger (1997); Smyth (2000); Hassaine et al. (2014)). NMR do not require crystallisation, the proteins are studied in solution or in solid-state. A magnetic field is applied to the sample to determine the structure of the protein based on the response of each atom, the latter depending on the chemical environment. It is a dynamic process, enabling the observation of various protein conformations. NMR also provides insights into the interactions of protein-protein complexes or the evolution of membrane domains (Brünger (1997); Sodt et al. (2015); Faraggi et al. (2018)). In Cryo-EM, a flow of accelerated electrons go through a thin solution containing the protein sample, preferably in liquid nitrogen or liquid helium to limit the damages. 2D images of transmitted electrons are then collected, representing different side-views of the sample. 3D reconstruction is achieved by means of algorithms able to classify the various 2D orientation (Milne et al. (2013)).

In addition to determining the structures of complexes, experimental methods can also be used to determine their biophysical properties. Here I describe very brieflty the two techniques I will refer to in the next chapters: one devoted to measuring binding free energies of complexes, and one employed to measure the conductance of membrane channels. Surface Plasmon Resonance (SPR) is a method commonly employed to quantify binding affinities. One protein is fixed on a surface, and the other protein is in a solvent that flows on the surface (Yeh and Wallqvist (2011)). Depending on the affinity, protein-protein complexes form, changing the weight of the surface. The different weight indicates how many complexes are formed, and what is the affinity between the two proteins. The interface is scrutinized by performing alanine mutations of specific residues and measuring the variation of the binding affinity (Cheng et al. (2019)). The conductance of ion channels can be measured using electrophysiology techniques, a standard experimental approach for assessing the function of ion channels in excitable cells. In the voltage clamp technique, a transmembrane voltage is applied to a membrane patch, allowing for measuring the ionic currents flowing through the channel proteins (Grewer et al. (2013)).

The study of biomolecular assemblies also benefit from computational techniques. Molecular Dynamics (MD) is commonly employed to study the dynamics, the thermodynamics and the kinetics associated to complex molecular system mimicking conditions encountered *in vitro* or *in vivo* (Hollingsworth and Dror (2018)). Both computational and experimental techniques complement each other. In the following part of the thesis, computational techniques and their applications will be depicted. I focus mainly on the methods I used in my thesis.

1.2 Computational techniques to study of biomolecular system

1.2.1 Force Field

At the molecular level, biological systems can be described thanks to a Force Field, which is empirical expression of the potential energy of a system of interacting particles. Force Fields describe the interactions between the particles by the means of dedicated mathematical functions and physical parameters. A common formulation is:

$$\begin{aligned} E_{tot} &= E_{bonded} + E_{non-bonded} \\ &= E_{bond} + E_{angle} + E_{dihedral} + E_{elec} + E_{VdW} \end{aligned} \tag{1.1}$$

where E_{elec} and E_{VdW} are the electrostatic and the Van der Waals energies, respectively, and E_{bond}, E_{angle} and $E_{dihedral}$ are the internal energies (Frenkel and Smit (2002)). Force field parameters are designed to reproduce the structure, the dynamics and the thermodynamics of model systems. They are initially extracted from experimental observations, and improved and supplemented using Quantum Mechanics calculations and condensed phase simulations (Polêto and Lemkul (2022)). The energy of each particle depends on all the other particles of the system, through the bonded or the non-bonded terms.

Coarse-Grained (CG) and an all-atom (AA) representations are commonly employed to to describe molecular processes requiring no explicit description of the electrons. In CG models, group of atoms are represented by a single particle bearing specific properties. To represent a given molecular system, a CG representation has a reduced number of particles compared to AA representation, reducing thereby the computational cost of the model but also its precision (Marrink et al. (2007)).

In addition to the number of particles, the different level of accuracy in the description of the physical interactions can affect the cost and precision of the simulations. For example, common force fields employed for biomolecualr systems relies on a pair additive potential, taking polarization effects into account only in a average manner. Explicitly polarizable force-fields (AMOEBA, DRUDE, ..) have been optimized for biomolecular sytems, but their use is associated with with a higher computational cost (Polêto and Lemkul (2022)). Solvent can be described either explicitly or implicitly, reducing thereby the computational treatment. When described implicitly, the solvent is modeled by means of a dielectric continuum interacting with the system of interest as opposed to a description where every single solvent molecule interact explicitly with all the rest of the system (Jorgensen et al. (1983); Mark and Nilsson (2001); Abascal and Vega (2005)). In general, there is a trade-off between the accuracy required for describing a given molecular system and the associated computational cost.

1.2.2 Molecular Dynamics simulations

In the 1950s, Alder and Wainwright reported an algorithm aimed at modeling the behavior of an ideal gas by means of hard-spheres (Alder and Wainwright (1957)). The method was then extended to study more complex systems, like liquid argon (Rahman (1964)), liquid water (Rahman and Stillinger (1971)), a protein (McCammon et al. (1977)), an ion channel (Mackay et al. (1984)), a bilayer membrane (Egberts and Berendsen (1988)), and an ion channel in a membrane (Woolf and Roux (1994)). Since then, the number of publications on MD simulations keeps increasing, involving a large variety of systems and techniques. I will briefly present the principle underlying molecular dynamics in the context of classical simulations.

Following the Born Oppenheimer approximation, the movement of the electrons and the nuclei can be dissociated (Born and Oppenheimer (1927)). The equation of motion $\vec{F} = m\vec{A}$ is solved to generate the trajectory of the atoms, modeled as hard spheres (Frenkel and Smit (2002)). Initiating a MD simulations requires a set of coordinates and velocities for all the atoms of a given system. Initial MD setups are most often generated using structural data obtained by experimental techniques (Berman (2000)), by homology with known structures (Song et al. (2013)) or by Machine Learning (Jumper et al. (2021)). Then, all other molecules (such as ions, water or even lipids) are added using different softwares (Humphrey et al. (1996); Lee et al. (2016)). The equipartition theorem is used to attribute initial velocities to the atoms (Leach (2009)):

$$\left\langle \sum_{i=1}^{N} \frac{1}{2} m_i v_i^2 \right\rangle = \frac{3}{2} N k_{\mathrm{B}} T \tag{1.2}$$

where m_i and v_i are the mass and velocity of the particle i, respectively, N the number of particles, k_B the Boltzmann constant, and T the temperature of the system.

The forces applied on each particle derived from the potential energy function encoded in the force field (Huang and García (2014); Marrink et al. (2007)). The equations of motions are then integrated step wisely to get the displacement of the particles generating thereby the trajectory of the atoms in the configuration space. The integration is done numerically, imposing the use of a time step Δt small enough to sample all the fastest movements. 1 fs is a typical time-step employed in MD simulations, accounting for the fast vibrations of atoms. The very short time step and the large number of particles usually involved to model biomolecular systems limit the time scales that can be access even using powerful computers. A compromise to keep the precision of the all-atom representation with yet a larger timestep has been found with the SHAKE algorithm or even later with the Hydrogen Mass Repartitioning (HMR). SHAKE constraint movements, for example of the hydrogens of the water, and HMR changes the repartition of the masses between the hydrogens and the heavier atoms they are bound to. Doing so, the timestep can be of 2 to 4 femtoseconds (Andersen (1983); Miyamoto and Kollman (1992); Hopkins et al. (2015)). Common trajectories are on the order of a few microseconds with all-atoms representation. In CG simulations, because the particles are much heavier, the fastest movements to consider are much slower, allowing for a larger timestep ($\approx 30 fs$ in Marrink et al. (2007) paper).

At each time step, the integration of the equation of motions calculates the coordinates and velocities of all atoms in the system $\{\mathbf{r}_i, \mathbf{p}_i\}_{i \in N}$. Based on the Maxwell-Boltzmann distribution, the probability for a system of N atoms to be in the state $\mathbf{X} = \{\mathbf{r}_i\}_{i \in N}$ is:

$$P(\mathbf{X}) \propto \exp\left(-\frac{U(\mathbf{X})}{k_{\mathrm{B}}T}\right) \tag{1.3}$$

where $U(\mathbf{X})$ is the potential energy of the state $\mathbf{X}$ derivated from the potential function of the force field, k_{B} is the Boltzmann constant and T the temperature. The most stable states (with a lower energy) are visited with a higher probability. According to the principle of ergodicity, for a trajectory of length τ consistent with the thermodynamic equilibrium, the average of an observable A can be calculated according to the sampling:

$$\bar{A} = \frac{1}{\tau}\int_0^{\tau} A(t')dt' = \frac{1}{M}\sum_{k=1}^{M} A(t_k) \tag{1.4}$$

In a N-particles system, the forces applied on one particle i depend on the N-1 other particles, which significantly increases the computational cost to generate a trajectory when there are a lot of atoms. A first solution is to resort to a cutoff in the non-bonded term. The electrostatic and the Van der Waals decrease with the distance:

$$E_{elec} = \sum_j \sum_{i \neq j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{1.5}$$

$$E_{VdW} = \sum_j \sum_{i \neq j} \epsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right] \tag{1.6}$$

where r_{ij} describes the distance between two particles i and j, q_i and q_j describe the charges of i and j, ϵ_0 the electric constant. When using a cutoff, only the "short range" interactions are computed in the non-bonded terms (Braun et al. (2019)). Long-range interactions are usually treated by means of the Particle Mesh Ewald approach (PME). The interaction potential is calculated as the sum of the short-range interactions, and a long-range term calculated in the Fourier space. This second term requires an infinite system (Ewald (1921); Darden et al. (1993); Wang et al. (2012)).

To avoid limitations and computational errors due to artificial effects of edges in the system, a convention is to resort to Periodic Boundary Conditions (PBC). The system is approximated as being infinite by creating images of the box in all directions (See Figure 1.4) (Frenkel and Smit (2002)).

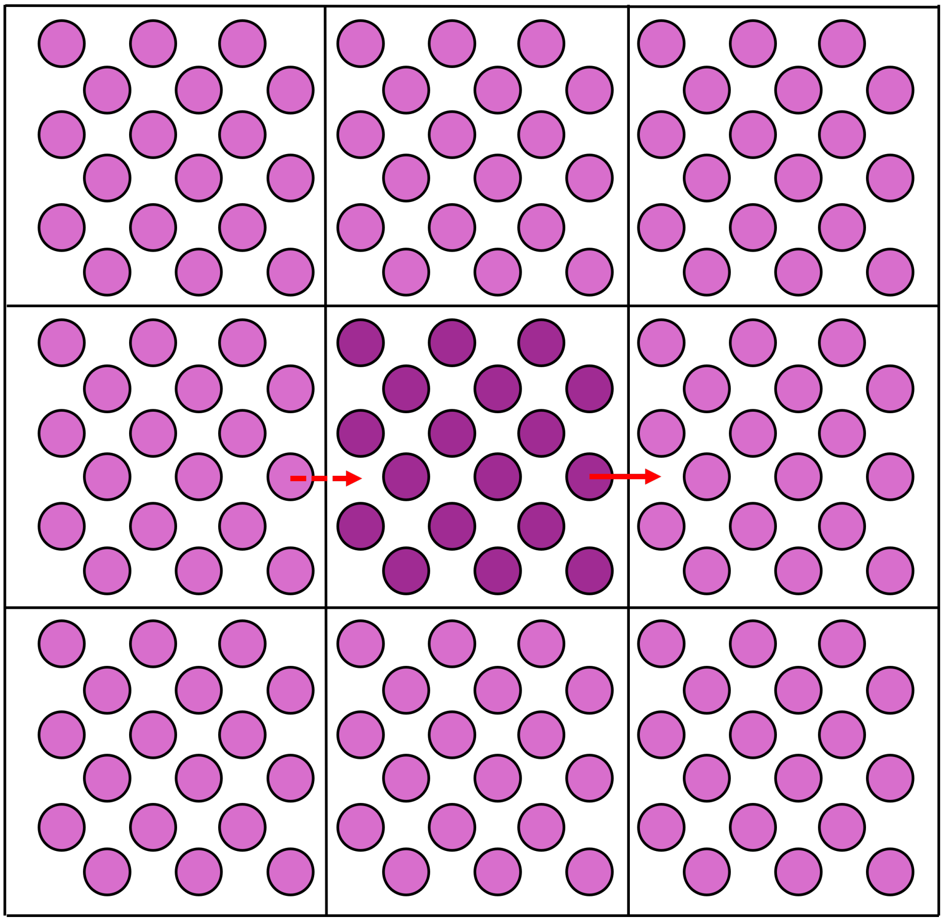


Figure 1.4: In purple, the system simulated. When the particle with the red arrow would reach the edge of the system, its image enters on the other side.

In addition to the terms derived from the Force Field, additional forces can be applied to the system, such as a piston and a thermostat to control the pressure and the temperature (Feller et al. (1995); Braun et al. (2019)). When the temperature is controlled by a thermostat, the velocities of the particles is being modified to keep the temperature constant, in the case of the pressure, the volume has to change to keep the pressure constant (Braun et al. (2019)). Thermostats can be either stochastic or deterministic. Berendsen and Nosé Hoover are two examples of deterministic thermostats. The Berendsen thermostat rescales velocities so that the instant temperature matches a targeted temperature, with a relaxation term to allow a few fluctuations avoiding any abrupt velocities modifications. The Nosé-Hoover thermostat introduces the thermal bath as a degree of freedom, with an artificial mass that control how much temperature fluctuations (Berendsen et al. (1984); Hünenberger (2005); Braun et al. (2019)). Examples of stochastic thermostats are the Andersen or Langevin thermostats. The Andersen thermostat randomly chooses particles and changes their velocity from the Maxwell-Boltzmann distribution. The Langevin thermostat adds a random and a friction forces. These two forces describe the coupling with the heat bath (Andersen (1980); Schneider and Stoll (1978)). Similarly to the Berendsen thermostat, a Berendsen barostat scales the volume to reach the target pressure with a little bit of fluctuation to avoid having a too strong piston (Berendsen et al. (1984)). The Anderson barostat couples the system to a pressure bath with a theory close to the Nosé-Hoover thermostat (Andersen (1980)), and the Langevin barostat, similarly to the Langevin thermostat, adds extra forces to describe the coupling with the pressure baths (Schneider and Stoll (1978)). The work presented in the thesis uses Langevin dynamics:

$$m\frac{dv}{dt} = F - \gamma v + f(t) \tag{1.7}$$

where F stands for the forces that are applied by the rest of the environment, γv is a friction force and $f(t)$ is a Gaussian random variable (Roux (2021)). In the context of the coupling with a bath (thermal or pressure), the friction term represents the viscosity of the bath, and the random variable adds collision between the bath and the particles (Braun et al. (2019)).

1.2.3 Beyond brute-force MD simulations

Solving the equation of motion and generating a trajectory can give insight into the conductivity of proteins, the stability of a complex or a specific conformation, or the mechanisms of molecular machines (Roux (2011); Karplus and McCammon (2002); Holzmann et al. (2016)). Trajectories generated by brute-force MD are limited to fast phenomenon (up to a few μs), preventing any slow biological processes to be observed. To circumvent this limitation, a strategy is to resort to enhance sampling algorithms. Below I present arbitrarily a few of these algorithms and discuss their use in computing thermodynamics quantities.

Replica Exchange:

This method creates several replica of the system, all simulated with some differences in a parameter. For example in the context of parallel tempering, each replica of the system is simulated at a given temperature. Then, different configurations $\mathbf{x}_i$ and $\mathbf{x}_j$ from two different replica i, j are exchanged with a certain probability:

$$P_{\text{accept}}(\mathbf{x}_i, i, \mathbf{x}_j, j) = \min\left[1, \frac{\exp(-[U_i(\mathbf{x}_j) + U_j(\mathbf{x}_i))}{\exp(-[U_i(\mathbf{x}_i) + U_j(\mathbf{x}_j))}\right] \tag{1.8}$$

Replicas at high temperatures provide access to configurations separated by large free energy barriers. However, to be accepted, exchanges between configurations must have energies that are close enough, requiring a very large number of replicas to cover a wide range of temperatures (Sugita and Okamoto (1999); Abrams and Bussi (2013); Hénin et al. (2022)).

Hybrid Monte-Carlo:

The goal of Monte-Carlo simulations is to perform a random sampling of the system. For example, particles can be added or removed from the system, exploring configurations not accessible in brute force MD (Kroese et al. (2014)). In a Monte–Carlo simulation, each configuration is generated randomly from the previous one, towards the equilibrium distribution (Metropolis et al. (1953); Roux (2021)). In the space of configurations $\mathbf{X}_1 \leftrightarrow \mathbf{X}_2 \leftrightarrow ... \leftrightarrow \mathbf{X}_k \leftrightarrow ..$, each step must respect the microscopic detailed balance:

$$\Pi(\mathbf{X}_i)T(\mathbf{X}_i \to \mathbf{X}_j) = \Pi(\mathbf{X}_j)T(\mathbf{X}_j \to \mathbf{X}_i) \tag{1.9}$$

with $\Pi(\mathbf{X}_i)$ the equilibrium probability to be in state $\mathbf{X}_i$ and $T(\mathbf{X}_i \to \mathbf{X}_j)$ the probability to go from i to j. The probability Π depends on the Boltzmann distribution:

$$\Pi(\mathbf{X}) \propto \exp\left(-\frac{U(\mathbf{X})}{k_\mathrm{B}T}\right) \tag{1.10}$$

The probability T is more complex to evaluate, as it considers the probability to try the exchange and the probability to do the exchange. In a symmetric and non–deterministic space of configuration, the probability to accept the exchange is given by: (Metropolis et al. (1953))

$$P = \min\left\{1, \exp\left(-\frac{\Delta U}{k_\mathrm{B}T}\right)\right\} \tag{1.11}$$

Hybrid Monte–Carlo methods, where the steps between two configurations are using non–equilibrium MD (Nilmeier et al. (2011); Chen and Roux (2015c)), overcome high energy barrier, and the space of configurations can be sampled more extensively. Applications such as constant pH simulations, in which the protonation state of some residues is modified, is an example of a process than cannot be sampled using classical brute-force MD (Chen and Roux (2015a); Radak et al. (2017); Martins De Oliveira et al. (2022)).

For long processes, such as the sampling of the configuration space of inhomogeneous bilayer using all-atom simulations, brute force MD are limited. Kindt and coworkers were the first to develop a neMD/MC strategy for sampling of configurations of lipids in the membrane (de Joannis et al. (2006); Coppock and Kindt (2009); Kindt (2011)). They were capable of modifying some lipids differing by a few atoms, modifying the composition of the bilayer itself and thereby generating configurations in the grand canonical ensemble. More recently, Fathizadeh and Elber developed the MDAS algorithm (Molecular Dynamics with Alchemical Steps) for all-atoms lipids mixtures. They exchanged lipids within a bilayer on the basis of alchemical work calculated from a nonequilibrium MD trajectory (Fathizadeh and Elber (2018)). They studied a two lipid mixtures, one composed of POPC and DOPC, two lipids differing by one unsaturation and a longer acyl chain, and one composed of DPPC and DLPC, two lipids differing by their acyl chain length only. They concluded that the efficiency of such neMD/MC hybrid approaches depends crucially on the mutation strategy (Fathizadeh et al. (2020)). In a CG lipid mixture, they also succeeded to exchange lipids differing by the polar heads. They however failed to extend their strategy to the exchanges of all-atom lipids bearing a different net charge (Cherniavskyi et al. (2020)).

End points approximations :

The Molecular Mechanics energies combined with the Poisson-Boltzmann Surface Area (MM/PBSA) or Generalized Born and Surface Area continuum solvation (MM/GBSA) are widely used (Wang et al. (2019)). The main reason for the success of these methods is that it is used on molecular configurations – so on snapshot – or on very short trajectories, making it a cheap method (Xu et al. (2013)). The principle of the calculations rely on these equations (Kollman et al. (2000); Genheden and Ryde (2015)):

$$\begin{aligned} \Delta G_{bind} &= \langle G_{PL} \rangle - \langle G_P \rangle - \langle G_L \rangle \\ &= E_{bond} + E_{elec} + E_{VdW} + G_{polar} + G_{non-polar} - TS \end{aligned} \tag{1.12}$$

where the indexes PL, P and L describe the complex, the protein alone and the ligand alone, respectively. All the contributions are evaluated based on physical properties and the force field – hence the success on snapshots only. Yet, if the methods are a good first approximation, they don't give very precise results, the error bars can be large (Nandigrami et al. (2022); Adelusi et al. (2023); Akash et al. (2023)). More details about these methods are going to be given in another chapter.

Umbrella Sampling:

The Umbrella Sampling Method was first theorized by Torrie and Valleau (1977). A bias potential is applied to modify the energy landscape along the reaction coordinate. Adding this potential allows to overcome high energy barriers in order to explore the whole space of configurations. The surface energy along the reaction coordinate is then deduced by unbiaising the energy:

$$E^b(\mathbf{r}) = E^u(\mathbf{r}) + w(\xi) \tag{1.13}$$

with $E^b(\mathbf{r})$ the biased energy, $E^u(\mathbf{r})$ the biased energy and $w(\xi)$ the bias potential added along the reaction coordinate ξ (Kästner (2011)). In order to access to the unbiased energy profile, the distribution must be computed as:

$$P^u(\xi) = P^b e^{\beta w(\xi)} \langle e^{-\beta w(\xi)} \rangle \tag{1.14}$$

Then, the energy is:

$$A(\xi) = -\frac{1}{\beta} ln\Big(P^b(\xi) \Big) - w(\xi) - \frac{1}{\beta} ln\Big\langle e^{-\beta w(\xi)} \Big\rangle \tag{1.15}$$

This method is exact and does not require any approximation (Kästner (2011)). An extension of the Umbrella Sampling is the Weighted Histogram Analysis Method (WHAM) (Kumar et al. (1992); Souaille and Roux (2001)). In this method, the simulation is separated in windows, the unbiased distribution is given as:

$$P^u(\xi) = \sum_{i}^{window} p_i(\xi) P_i^u(\xi) \tag{1.16}$$

where $P_i^u(\xi)$ is the unbiased distribution along ξ in the window i. The weights p_i are minimizing the errors on the distribution (Kumar et al. (1992); Souaille and Roux (2001); Kästner (2011); Zhu and Hummer (2012)). This method presents the interesting perk of being precise, although simulations may be long to converge (Kästner (2011)).

Metadynamics:

Metadynamics is also a method to overcome energetic barrier and explore the full energy landscape. In metadynamics, collective variables (CV) are chosen to describe the system. At specific values of the CV, a Gaussian Potential is added to the potential to "flood" the wells in the energy surface (Grubmüller (1995); Barducci et al. (2011)):

$$B_t(s) = w \sum_{t'<t} \exp\left(-\frac{(s_{t'} - s)^2}{2\sigma^2}\right) \tag{1.17}$$

where B is the bias potential, t is the time, s the CV, w and σ the height and width of the Gaussian, respectively (Bussi and Laio (2020)).

Unlike the Umbrella Sampling, the metadynamics does not require a prior estimation of the energy landscape, but requires an appropriate set of CV to reach a convergence. When converged, the estimation of metadynamics can be as exact and precise as Umbrella Sampling (Bochicchio et al. (2015)). To improve the convergence of metadynamics, it is possible to resort to well-tempered Metadynamics, that decreases the height of the "flooding" Gaussian potential over the simulations (Barducci et al. (2011); Fu et al. (2018)).

ABF:

Another biasing method is the Adaptive Biasing Force (ABF). Unlike Umbrella Sampling or metadynamics, biasing the Energy Potential, this method act on the forces. The distance between the two molecules is increased and the equilibrium constant along the reaction coordinate is:

$$K_{eq} = 4\pi \int_0^R drr^2 e^{(-\beta w(r))} \tag{1.18}$$

with $\beta = 1/k_B T$ and $w(r)$ the Potential of Mean Force (PMF) over the reaction coordinate r (Shoup and Szabo (1982)). The binding free energy is then computed as:

$$\Delta G = -k_\mathrm{B} T ln(K_{eq}.C^{(0)}) \tag{1.19}$$

where $C(0) = 1M$ in a standard state. This formula for the equilibrium constant is true only if the simulation is reversible. To obtain that in a reasonable amount of time, it is necessary to resort to restraints (Blazhynska et al. (2023)). Usually, the restraints are applied on the RMSD of the protein and the ligand, and different positional angles. The figure 1.5 shows the angles that are restraint. The cost of the restraints have to be remove from the final PMF to get the binding free energy of the complex (Chipot and Pohorille (2007); Pohorille et al. (2010); Gumbart et al. (2013b); Comer et al. (2015)). Although the simulations can be long and expensive, the method gives very accurate results (Comer et al. (2015); Blazhynska

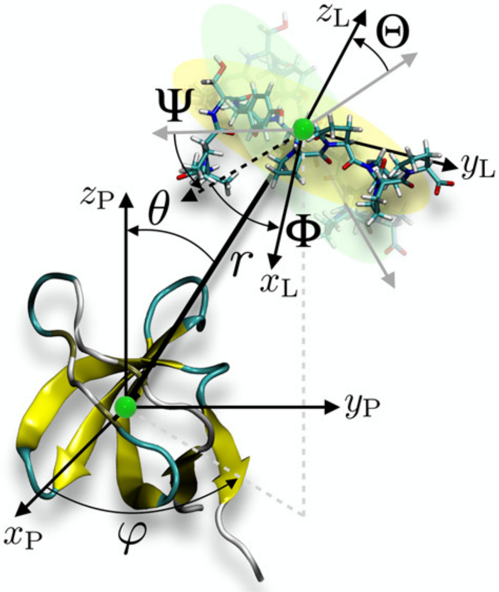


Figure 1.5: Graphical representation of the restraints used to have a reversible simulations. Figure from Fu et al. (2017)

et al. (2023)). To improve the convergence, it is possible to resort to extended ABF (eABF), where a fictitious degree of freedom is added to reduce the cost of the simulations, or even to combine several methods, for example well-tempered metadynamics and eABF (WTM-eABF) (Comer et al. (2015); Fu et al. (2018)).

Alchemical transformations

Another approach to calculate Binding Free Energy or to observe the response of a system to a perturbation is alchemical simulations. In alchemical simulations, the number of atoms do not have to remain constant during the simulation. The composition of the system is modified, at least locally, to create a mutation or to measure a binding free energy (Chipot and Pohorille (2007)). For example, if, in a protein-ligand complex, the ligand is buried inside the protein, the geometrical route is not optimal. Increasing the distance between the protein and the ligand would require a reaction coordinate not easy to define, and the convergence of such a simulation would be too expensive. By scaling the interactions of the ligand with the rest of the system, it can be decoupled (or removed) from the system. Then, the difference of energy between the initial state (the ligand is coupled to the protein) and the final state (the ligand is decoupled) is the binding Free Energy of the ligand to the protein.

To run an alchemical simulation, it is necessary to introduce a coupling parameter λ. During the trajectory, λ is increasing from 0 to 1, to scale the interactions of a set of atoms (or of molecules) during the simulation. The interactions of the *incoming* atoms (coupled to the system during the simulations) are coupled by λ, while the interactions of the *outgoing* atoms (decoupled during the simulations) are coupled by $(1 - \lambda)$. For example, if the goal is to mutate a residue of the protein into another amino acid, the interactions of the atoms of the initial residue are scaled by $(1 - \lambda)$, and the interactions of the atoms of the targeted residue are scaled by λ. This way, at the beginning of the simulation ($\lambda = 0$), the protein has its initial sequence and structure. At the end of the simulation ($\lambda = 1$), the initial set of atoms is decoupled from the rest of the system (the interactions are scaled by $1-1=0$), and the mutated residue is coupled with the rest of the system (Chipot and Pohorille (2007)).

To achieve that in practice, there are two main paths: the dual and the single topology. In a dual topology, the two sets of atoms (*incoming* and *outgoing*) are initially in the system (Chipot and Pohorille (2007); Mey et al. (2020)). On the other hand, on a single topology, the initial outgoing atoms are gradually modified to become the incoming atoms. The figure 1.6 shows this two topologies with the example of the mutation of a glycine to an alanine. For the dual topology (**a**), the outgoing H and the incoming CH_3 are both there through the whole simulation. Their interactions with the rest of the systems are scaled by λ (or $1-\lambda$), and an exclusion partition ensures that they do not interact with each other (the outgoing atoms can never interact with the incoming). Even though it appears as if the common carbon is linked to 5 groups, because of the scaling of the interactions, its environment evolves but it does not break any chemical rules with this topology. For the single topology (**b**), the common carbon is linked to the initial H at the beginning of the simulation. Throughout the transformation, the volume, the mass and other all of the other properties of the H are modified gradually to reach the parameters of the CH_3 at the end of the transformation. Some hybrid techniques are starting to be developed, with variations from these two topologies. The choices between these topologies can affect the simulations. In the case of non-equilibrium simulations for

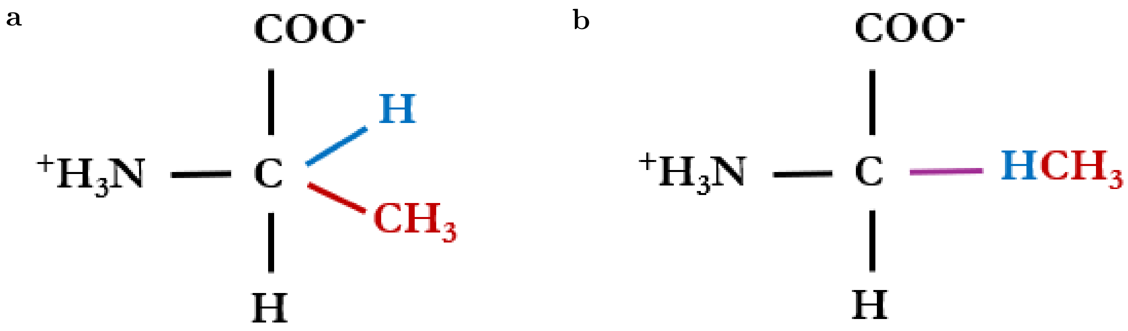


Figure 1.6: Representation of the dual (**a**) and the single (**b**) topologies.

example, the type of topology affects the path, and therefore affects the final result (Mey et al. (2020)).

Because everything has to be solved numerically, the increase of λ is discrete. Two strategies can be applied: to use a slow-growth, where λ is changing at each time step, or to define windows within the trajectory. At each new window, λ increases, and inside each window, λ remains constant such that the system relaxes after the perturbation induced by the change of λ (Chipot and Pohorille (2007)).

FEP and TI:

The difference of energy associated with the perturbation is measured through the transformation. One method to do that is the Free Energy Perturbation (FEP) theory. For a small perturbation between the state **A** and **B**, the Zwanzig equation (Zwanzig (1954)) is:

$$\Delta F(\mathbf{A} \rightarrow \mathbf{B}) = -k_{\mathrm{B}}T \ln \left\langle \exp\left(-\frac{E_{\mathbf{B}} - E_{\mathbf{A}}}{k_{\mathrm{B}}T}\right)\right\rangle \tag{1.20}$$

In the case of a Binding Free Energy computed over the variation of λ from 0 to 1 (Pohorille et al. (2010)), the equation is:

$$\Delta G = -k_{\mathrm{B}}T \ln \left(\frac{1}{N}\sum_{i=0}^{N} \exp\left(-\frac{\Delta U(\Gamma_i)}{k_{\mathrm{B}}T}\right)\right) \tag{1.21}$$

where N is the number of λ-windows used during the simulation, and $\Delta U(\Gamma_i)$ the variation of energy associated with the change of value of λ in the window i.

The other method to evaluate the perturbation is to compute the work using Thermodynamic Integration (TI) (Kirkwood (1935)):

$$\Delta G = \int_0^1 \left\langle \frac{\partial H(x, p_x, \lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda \tag{1.22}$$

Both methods should give an equal result if the simulations reach convergence. In the context of nonequilibrium, there are differences as the work depends on the path, and the energy depends on the states.

Bennett Acceptance Ratio:

Assessing the convergence of simulations is nontrivial. A common strategy is to perform the simulations in a *forward* (λ goes from 0 to 1) and *backward* (λ goes from 1 to 0) directions. These designations are used for example with NAMD software (Liu et al. (2012)), but can also be *forward* and *reversed*. Then, the overlap between the equivalent windows of the forward and backward trajectories is measured, using for example the Bennett Acceptance Ratio (BAR) (Bennett (1976); Liu et al. (2012)). The BAR estimator quantifies how similar the two trajectories are, indicating if the simulation is reversible, if there is hysteresis issue between the two trajectories, and therefore how trustworthy the results of the alchemical simulations are (Bennett (1976); Chipot and Pohorille (2007); Liu et al. (2012)).

Relationship between and nonequilibrium work and energy:

In some cases, for example with the Hybrid Monte-Carlo methods mentioned earlier, resorting to nonequilibrium simulations is prefered. The TI and FEP methods are true for simulations with $\tau \to \infty$, but for a finite simulation, it is not exactly accurate. To address this issue, Jarzynski proved a relationship between the converged value of the energy and the computed energy or work. Jarzynski's equation states that (Jarzynski (1997); Chipot and Pohorille (2007)):

$$\Big\langle \exp[-\beta W(\tau)] \Big\rangle = \exp(-\beta \Delta A) \tag{1.23}$$

So the accurate energy ΔA can be extrapolated from the computed work $W(\tau)$. Another important theorem is the Crooks Fluctuation Theorem (Crooks (1999); Chipot and Pohorille

(2007)):

$$\frac{p_f(w = W(\tau))}{p_b(w = -\underline{W(\tau)})} = \exp[\beta(w - \Delta A)] \tag{1.24}$$

where f and b are for the forward and backward trajectories, respectively, $\underline{W}$ is the work on the backward path.

1.3 Presentation of the thesis

This first chapter introduced the biological and methodological concepts used in the Ph.D. The "Chapter 2: Free Energy Calculations" and "Chapter 3: Protein-Protein Binding Specificity" present alchemical transformations as a method to compute binding free energy. In Chapter 2, the absolute free energy of binding of a membrane protein:ligand complex has been calculated in an effort to design a rigorous protocol for the plugin Binding Free Energy Estimator 2 (BFEE2). Chapter 3 presents an overview of computational techniques to elucidate the selectivity the interface of protein:protein complexes. Relative binding free energies have been calculated to assess the roles of residues in the binding mechanism. Both these chapters cover a wide range of applications of alchemical transformations, and introduce techniques to improve the convergence of the simulations.

Building on alchemical transformations, I developed a new methodology combining non-equilibirum Molecular Dynamics with Monte-Carlo (neMD/MC) to enhance the sampling of the configurations of inhomogeneous membrane by swapping lipids. The "Chapter 4: Nonequilibrium Monte–Carlo" and "Chapter 5: Hybrind neMD/MC Lipid Swapping Algorithm to Equilibrate Membrane Simulation with Thermodynamic Reservoir" explain the methodology and examples of applications. Chapter 4 details the technical aspects of the method, as well as its ability to sample the distribution of lipids near a transmembrane helix carrying a net charge are illustrated for a binary mixture of charged and zwitterionic lipids. In an attempt to optimize the methodology, the Chapter 5 presents a variation to the neMD/MC method. The membrane with the helix is equilibrated with another membrane, acting as a thermodynamic “reservoir” with the desired mole fraction for all lipid components.

A major application of the neMD/MC method is the sampling of lipid configuration around the pore of a membrane protein. The "Chapter 6: Pentameric Ligand-Gated Ion Channels" details MD simulations on nicotinic receptors, membrane proteins with specific protein:lipid interactions. It presents trajectories to study the dynamics and stability of the pores of several structures of nicotinic receptors, followed by an elaboration of a method to calculate the difference of affinity between a nicotinic receptor and anioninc and zwitterionic lipids using alchemical transformations.

Finally, the "Chapter 7: Conclusion and Discussion" concludes the thesis.

CHAPTER 2

FREE ENERGY CALCULATIONS

This chapter uses the results from the paper by Haohao Fu, Haochuan Chen, Marharyta Blazhynska, Emma Goulard Coderc de Lacam, **Florence Szczepaniak**, Anna Pavlova, Xueguang Shao, James C. Gumbart, François Dehez, Benoît Roux, Wensheng Cai, and Christophe Chipot: Accurate determination of protein:ligand standard binding free energies from molecular dynamics simulations, published in *Nature Protocols*, in April 2022.

Free energy calculations are an important application of Molecular Dynamics (MD) simulations. Several methods have been developed to perform such calculations. This chapter details applications of the geometric and alchemical routes. In the geometric route, the distance between the two components of the complex is increased along a reaction coordinate, and the binding free energy is computed as the difference between the energies of the bound state and the unbound state. In some cases, this method is not efficient, for example for the computation of the binding free energy between a protein and a ligand buried inside the protein. In the alchemical route, the ligand is coupled or decoupled from the system, and the binding free energy is computed as the difference of energy between the coupled and the decoupled states. This route allows the investigation of binding free energy when the distance coordinate between the protein and the ligand is be easily defined. The chapter presents applications of computations of free energy of binding, and how to establish a robust and extensive protocol to study protein-ligand complexes.

2.1 Presentation of the plugin BFEE2

Binding free energy calculations of protein-ligand complexes are used in several domains, such as drug design, and have been developed to give very precise results (Muegge and Hu (2023)).

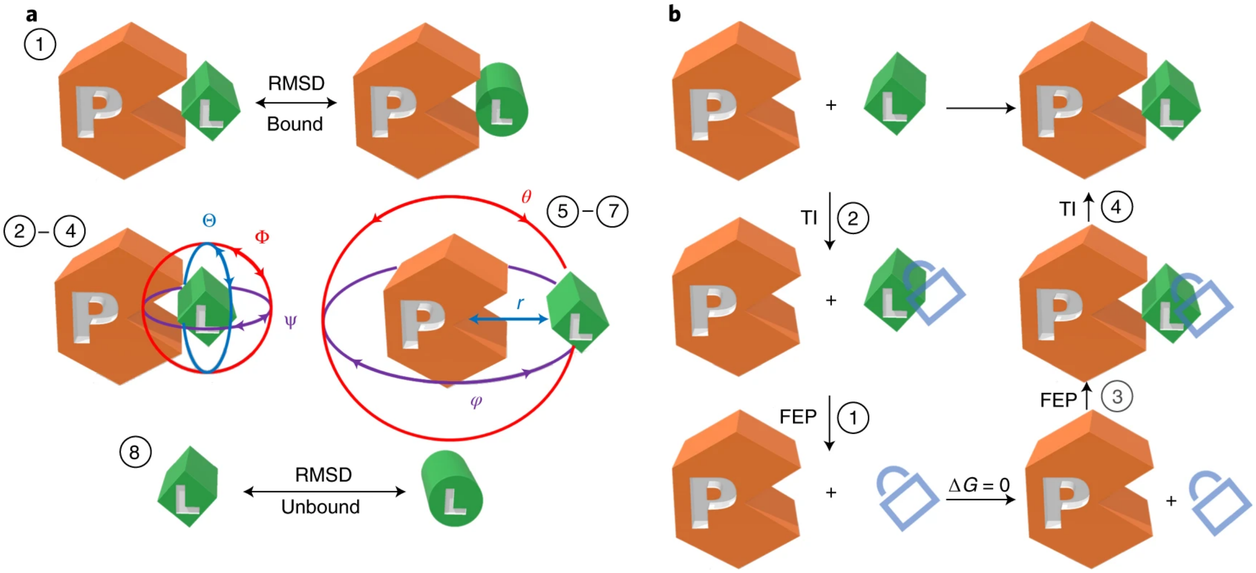


Figure 2.1: **a**, Geometrical route. **b**, Alchemical route. The numbers indicate the order of free-energy calculations set up using BFEE2. The lock represents the restraints applied to the conformational, orientational and positional degrees of freedom of the ligand with respect to the protein. Figure from Fu et al. (2022)

Despite the progress made in the methods to be more precise, running the simulations can remain a challenge. BFEE2 (Binding Free Energy Estimator) is a plugin developed for the software VMD (Humphrey et al. (1996)), to generate the input files to compute the binding free energies and to analyse the output files of the simulations (Fu et al. (2021)). BFEE provides input files for both NAMD (Phillips et al. (2020)) or GROMACS (Abraham et al. (2015)) for either the geometric and the alchemical routes. Figure 2.1 (from the paper Fu et al. (2022)) shows a representation of the two routes.

In order to ensure the convergence of the geometrical or the alchemical computations, a set of restraints is necessary (Blazhynska et al. (2023)). BFEE2 resorts to the Collective variables (colvars) module (Fiorin et al. (2013)) to handle the restraints. The plugin generates the files for the computation of the binding free energy as well as those to evaluate the contribution of the restraints. In the case of the geometrical route, the conformations of the protein and the ligand are restrained to their initial position by means of a soft harmonic potential acting on a Root-Mean-Square-Displacement (RMSD) variable. The angles translating the relative orientation of the ligand with respect to the protein are restrained, resulting in five additional restraint. The contribution of these restraints to the binding free energy is evaluated by means of a Potential of Mean Force (PMF) computed in the framework Well-Tempered Meta-extended Adaptive Biasing Force (WTM-eABF) (Comer et al. (2015); Fu et al. (2019)). The restraints are added one after another, PMFs being computed at each step. Finally the PMF associated to the dissociation of the complex is computed within the same theoretical framework. BFEE2 post-treatment reconciles all the individual contributions to provide a unbiased estimate of the binding free energies of the studied complex. The Alchemichal route follows a similar strategy. The difference lies essentially in the way free energies are calculated, instead of PMFs, we resort here to alchemical transformations. The plugin not only analyses the results to get numerical estimation of all the contributions of the simulations, but also gives an insight on the convergence of the simulations (Fu et al. (2021, 2022)). For the alchemical simulations, the Bennett Acceptance Ratio and the convergence of the simulations is estimated using the VMD Plugin ParseFEP (Liu et al. (2012)).

The plugin has been tested on a wide variety of systems, to ensure that it can support very different protein-ligand complexes. The automation of the creation of the files and of the analysis limits the risk of human mistakes and ensures the reproducibility of the results, and the computational methods WTM-eABF and TI or FEP give precise estimations of the binding free energy. Despise all of these strengths, there are some limitations in the applications of the plugin. All of the results depend on the accuracy of the force field. The whole evaluation of the energy can be computationally expensive because of the number of the simulations and their lengths. And finally, there are some cases where the simulations can be even harder to run, for example in the case of a initial unbound structure with no information on the binding motif or for deeply buried ligands (Fu et al. (2022)).

The paper presents a demonstration of the efficiency of the plugin and a protocol to use it. Thus, it was necessary to do a large screening of different protein-ligand complexes to explore

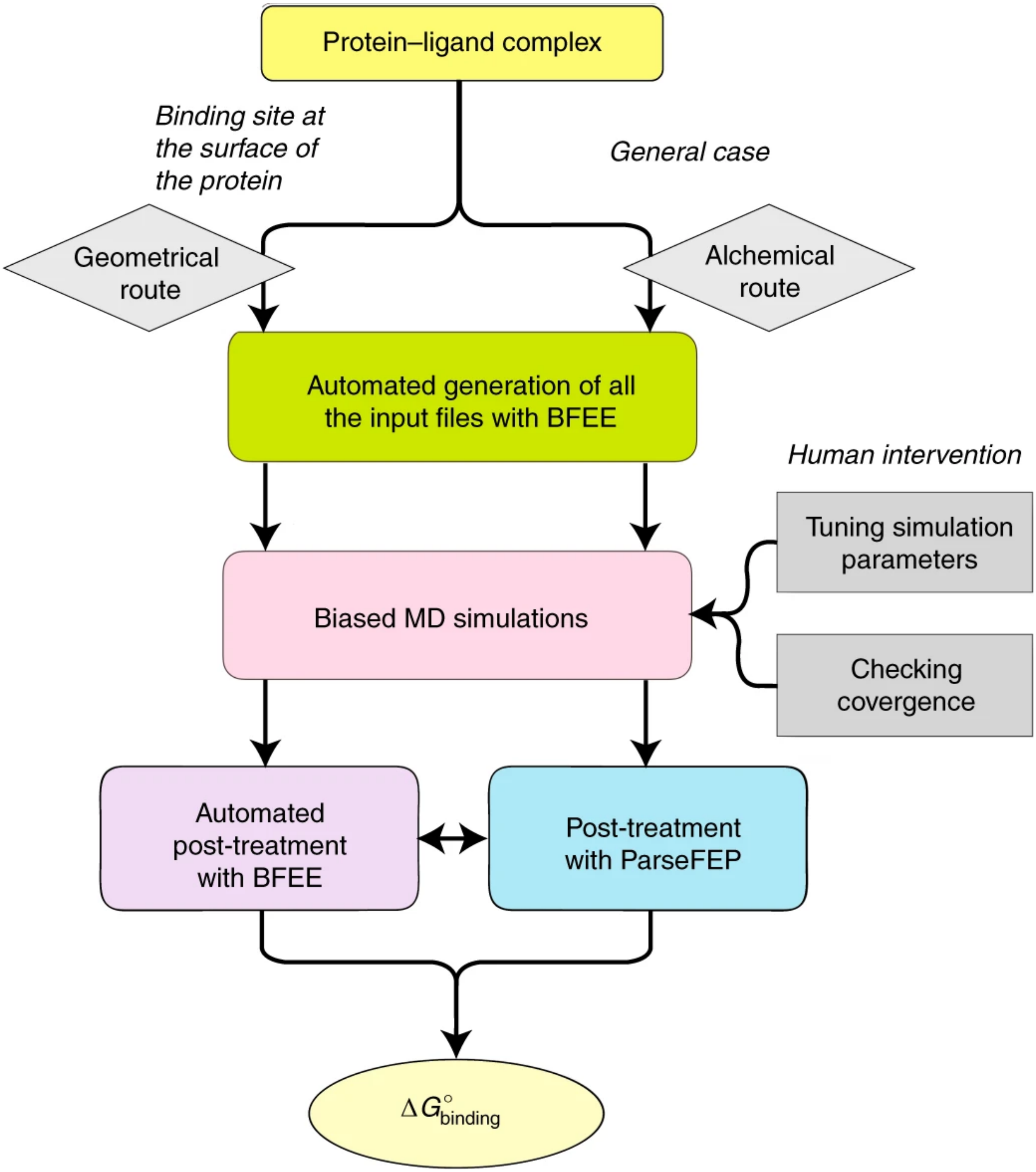

Protein–ligand complex
Binding site at the surface of the protein
General case
Geometrical route
Alchemical route
Automated generation of all the input files with BFEE
Human intervention
Tuning simulation parameters
Biased MD simulations
Checking covergence
Automated post-treatment with BFEE
Post-treatment with ParseFEP
$\Delta G^{\circ}_{\text{binding}}$

Figure 2.2: Workflow of the methodology of the plugin. Figure from Fu et al. (2022)

Table 2.1: Examples to assess the efficiency of the plugin. The * after MUP-I refers to a different ligand compared to the previous line. The energies are in kcal/mol. The details of the ligand and the references are in the text.

System	Route	$\Delta\Delta G_{\mathrm{BFEE2}}$	$\Delta\Delta G_{\mathrm{exp}}$	Interest
DIAP1-BIR1	Geometrical	-8.7 ± 0.7	-9.5	GROMACS
T4 lysozyme L99A	Alchemical	-6.0 ± 1.0	-5.2	Buried ligand
Trypsin	Geometrical	-7.8 ± 0.6	-7.2 to -6.3	Rigidity of ligand
Factor Xa	Geometrical	-8.7 ± 0.4	-9.0	FF and driving force
MUP-I	Alchemical	-7.8 ± 1.0	-7.8	Affinity ranking
MUP-I*	Alchemical	-5.5 ± 0.7	-6.0	Affinity ranking
β1-adrenergic receptor	Alchemical	-8.1 ± 1.0	-9.07	membrane protein
ATP	Alchemical	-11.6 ± 0.8	-4.1 ± 1.1	ATP
ADP + Pi	Alchemical	-8.3 ± 0.9	-4.3 ± 0.8	ATP

the limits and the capacities of the plugin. Table 2.1 summarizes all the systems studied and the binding free energies computed. Most of the examples have been computed using NAMD. To test the code on GROMACS, the protein-ligand complex of DIAP1-BIR1: Grim peptide has been studied, using the paper by Brown and Muchmore (2009) as reference. The complex T4 lysozyme L99A:benzene based on the paper by Morton and Matthews (1995) has been studied as the ligand is deeply buried in the protein. That assesses the convergence difficulty to solvate the binding site in the decoupled state. Depending on the rigidity of the ligand, the RMSD contribution can be more or less easy to evaluate. The complex Trypsin:benzamidine (Schwarzl et al. (2002)) has been studied to enlarge the screening of such property of the system. The complex Factor Xa:quaternary ammonium (Schärer et al. (2005)) requires a specific Force Field (FF) to translate the Cation-π interaction, and is an interesting complex to study the diving force of the formation of the complex. The two complexes MUP-I:2-methoxy-3-isopropylpyrazine and MUP-I:6-hydroxy-6-methyl-3-heptanone (Bingham et al. (2004); Timm et al. (2001)) prove that BFEE2 can be used to rank the relative affinity of different complexes consistently with experimental results. The complex β1-adrenergic receptor:4-methyl-2-(piperazin-1-yl) quinoline (Christopher et al. (2013)) is a membrane

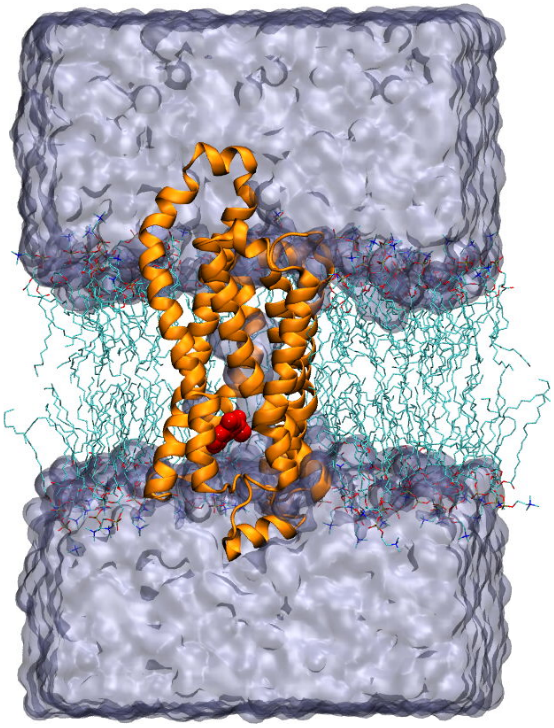


Figure 2.3: Representation of the system studied for the protocole of BFEE2. In orange, the protein, in red, the buried ligand, in cyan, the membrane and in blue, the bulk.

protein with a buried ligand. The problematic of the buried ligand is again tested, but more importantly the generation of inputs is different when there is a membrane. This will be detailed later, but implemented the specificity of a membrane protein is the purpose of the computation of this system with the plugin. The systems V1-ATPase:nucleotide (ATP) and V1-ATPase:nucleotide (ADP + Pi) (Adachi et al. (2012)) have been challenging, and are part of the large problematic of the study of the ATP mechanism. All of the energies computed and the comparison with the experimental values are in Table 2.1.

2.2 Contribution to the project

The system I studied is the β1-adrenergic receptor:4-methyl-2-(piperazin-1-yl) quinoline (code PDB: 3ZPR), represented in Figure 2.3 (Christopher et al. (2013)). This system is a G protein-coupled receptor (GPCR). It presents two advantages for the plugin: it is a membrane protein, and the ligand is buried in the protein. The plugin BFEE2 writes all the files for the simulations. Because the system is no longer isotropic, inputs files for hydrosoluble complexes need to be adapted (when the volume varies to keep the pressure constant, the direction perpendicular to the membrane can not be treated similarly to that in the plane of the bilayer). In addition to that, creating the structure of the decoupled states (for example for the RMSD contribution) requires not only to deal with the protein but also with the lipids, which calls for changes in the generation of inputs.

A situation difficult to handle is the buried ligand in alchemical simulations: When decoupling the ligand, the water still needs to flood the binding site, requiring much longer simulations to converge. The strategy used by the plugin for the alchemical is to divide the trajectory into λ-windows. In such a protein-ligand complex, it is important to do long equilibration per windows, and it can be interesting to divide the trajectory in more windows. The solvatation of the binding site is then more favorable, as the water has more time to populate the binding site, the decoupling of the ligand being smoother. A strategy that can help is to equilibrate the system at different intermediate λ states for a longer time, and to then use these equilibrated configurations as starting points for different windows.

2.2.1 Computational details and convergence issues

The whole structure of the system (protein, ligand, membrane and bulk) has been created using the PDB structure 3ZPR (Christopher et al. (2013)) and the CHARMM-GUI input generator (Jo et al. (2008)). The CHARMM36 force field has been used for the protein, the lipids and the bulk (Lee et al. (2016); Huang et al. (2017)). Not all molecules, and especially among the small molecules, have a parameterized force field. It was necessary to generate force field parameters for the ligand. A useful tool is CGENFF (Yu et al. (2012); Vanommeslaeghe et al. (2010)). By homology with already known force fields molecules, CGENFF suggests the parameters for the ligand, with an evaluation of the precision of

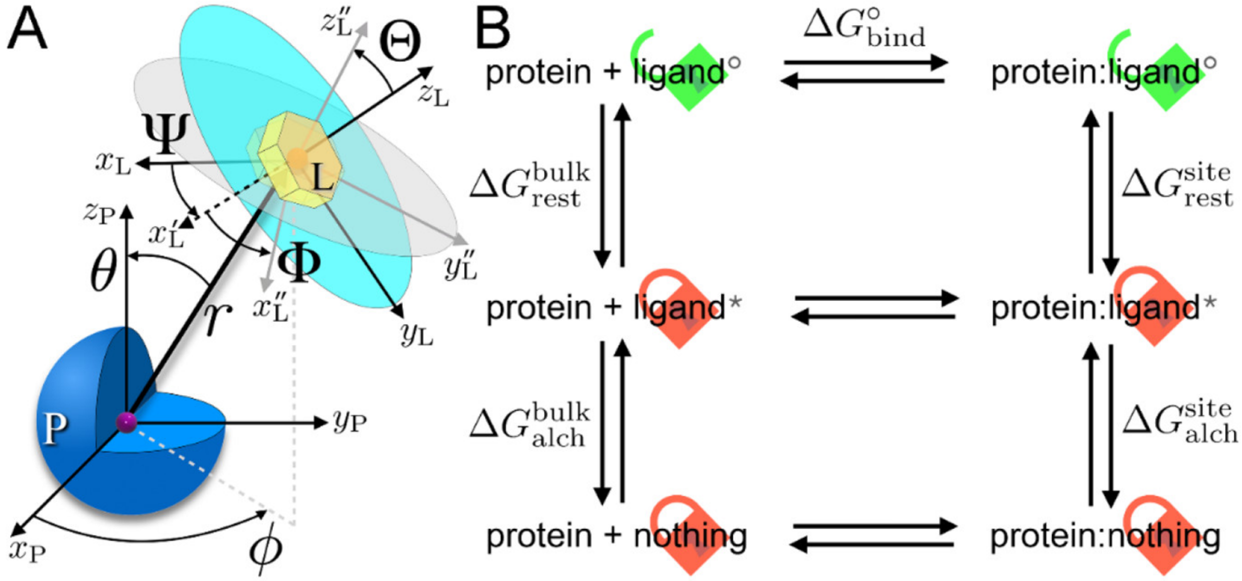


Figure 2.4: **A** Representation of the restraints applied in the alchemical route, **B** Thermodynamic cycle corresponding to the alchemical route. Figure from Fu et al. (2021)

these parameters.

The crystallographic structure is surrounded by a bilayer membrane (POPC molecules), and solvated in TIP3P water and NaCl ions (0.15 M). The total system consists of 65402 atoms in a box with the dimensions 74 Å × 74 Å × 116.7 Å. MD simulations have been carried out using NAMD 3.0. Damped Langevin dynamics and piston have been used to ensure the stability of the temperature at 300 K and the pressure at 1 atm. The Van der Waals interactions have been smoothly turned to zero between 10 and 12 Å. The pairlists parameter was set up to 14 Å. The long-range electrostatic interactions have been computed by the PME algorithm. Considering the size of the system, a mass repartition on the hydrogens was applied to use a timestep of 4 fs. The system has been equilibrated during 40 ns with the protein-ligand complex restrained to its crystallographic conformation. The latter restraints are smoothly decoupled during 7 ns. 80 extra ns are then produced to further equilibrate the system. This equilibrated configuration is used as the starting point of the BFEE2 simulations. To be consistent with the simulation of the other persons of the team, the timestep was set to 2 in this part (with the suitable psf file). If not said otherwise, the parameters by default are used. The figure 2.4 represents the set of restraints used during the alchemical simulations, and the route followed for the calculations of the binding free energy.

With the parameters by default, even if the simulation time or the stratification are increased, it may be difficult to reach a proper convergence. BFEE2 and ParseFEP give an estimation of the error and of the convergence of the simulation. And with the ParseFEP plugin, it is possible to see the convergence per window. The Figure 2.5 shows some windows from a calculation that did not converge. Errors and hysteresis are estimated for the whole trajectory, so a high hysteresis on some windows leads to a high error on the total estimate.

Once the plugin was modified to adapt to membrane protein, a protocol has been developed to compute the binding free energy on membrane protein-ligand complexes. First, the bound state (protein+ligand) is considered. The hydration of the buried binding site alongside the decoupling of the ligand is particularly slow. Therefore, instead of starting from the configuration with the ligand and decoupling it through the simulation, the ligand is being coupled in the hydrated binding site. That required running a long equilibration without the ligand, to let the water penetrate the protein. Here, $\lambda = 1$ corresponds to the system without ligand, and $\lambda = 0$ corresponds to the system coupled with the ligand. Starting from the unbound state improves the convergence and the reversibility of the transformation. Following a stratification strategy, the total trajectory has been divided in 10 smaller trajectories, each of the run for a total $\Delta\lambda = 0.1$. Each subtrajectory require a preliminary equilibration with the ligand in the binding site, coupled with the value of λ corresponding to the subtrajectory. Then, the forward and backward trajectories were calculated, with 5 windows of 10 ns ($\Delta\lambda = 0.02$). Once we had the total trajectory, some windows were splitted in 2 windows (10 ns, $\Delta\lambda = 0.01$) to lower the hysteresis. After the alchemical transformation, contribution of the restraints to the binding free energies are

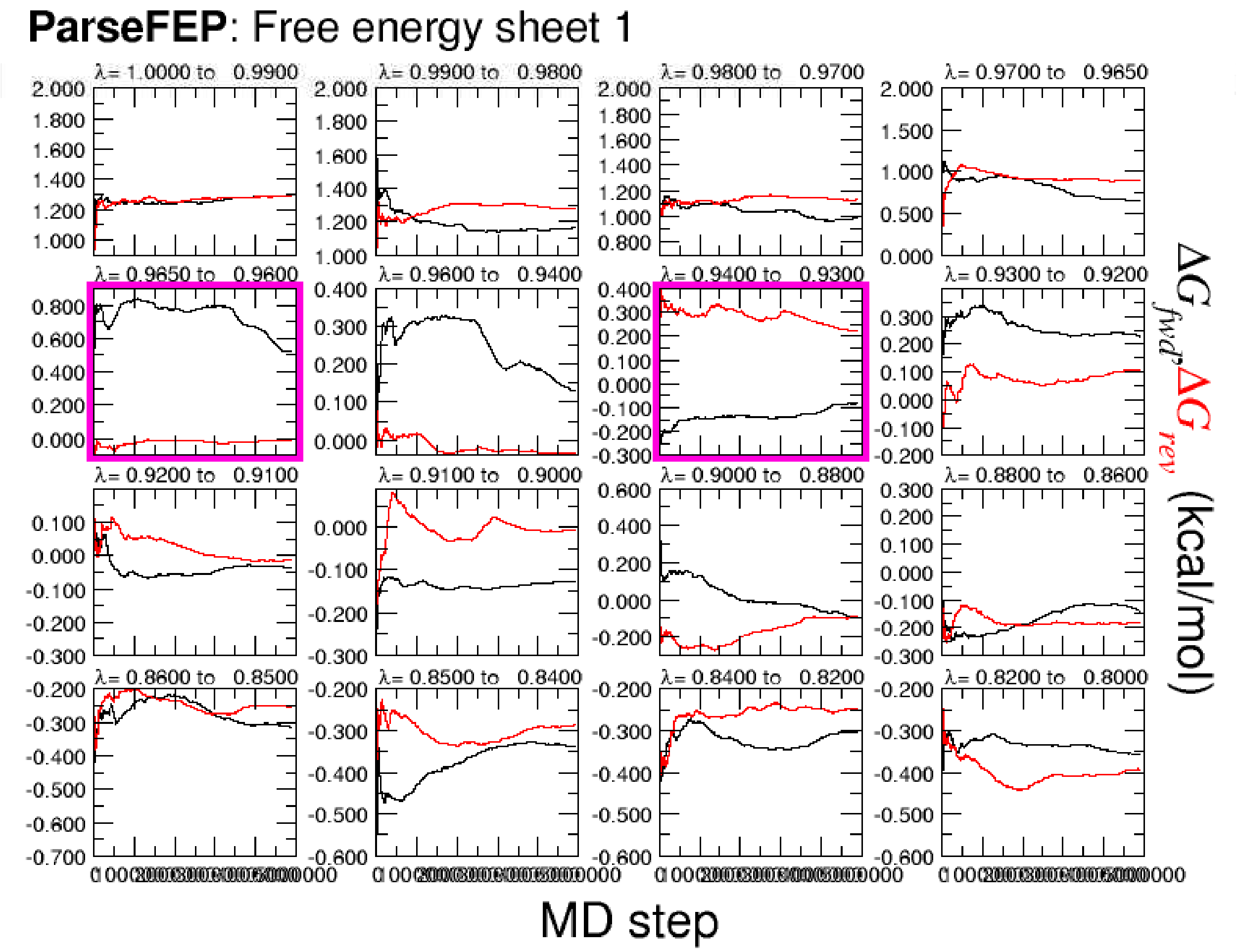


Figure 2.5: Example of windows obtained with a too short trajectories. The windows with the pink frame have a hysteresis between the forward and backward trajectories more than 0.2 kcal/mol.

Table 2.2: Results given by the plugin BFEE2 after analysis of the output files of the simulations

Contribution	Energy (kcal.mol^{-1})	Simu. Time	Stratification
Δ G (site,couple)	-21.1 ± 1.0	1240	62 windows
Δ G (site,c+o+a+r)	-2.0 ± 0.1	102	51 windows
Δ G (site,c)	-0.3 ± 0.0		
Δ G (site, Θ)	-0.7 ± 0.1		
Δ G (site, Φ)	-0.6 ± 0.0		
Δ G (site, Ψ)	-1.0 ± 0.1		
Δ G (site, θ)	-0.0 ± 0.0		
Δ G (site, ϕ)	-0.0 ± 0.0		
Δ G (site, r)	-0.3 ± 0.0		
Δ G (bulk, decouple)	3.4 ± 0.1	100	50 windows
Δ G (bulk, c)	1.6 ± 0.0	42	21 windows
Δ G (bulk, o+a+r)	11.52		
Δ G (total)	-8.1 ± 1.0		

estimated using backward and forward thermodynamic integration using 51 ns of sampling in each direction and 50 lambda windows.

The contribution from the unbound state is computed with no modifications to the standard protocol. After a 2 ns equilibration with the restraints, the decoupling of the ligand is simulated for 50 ns. The contribution of the restraints are estimated using thermodynamic integration, using 21 ns of sampling in each direction.

With this protocol, the hysteresis between the forward and backward trajectories are much smaller, as shown on Figure 2.6. The final results are shown in the Table 2.2. Δ G (site,couple) represents the contribution of the decoupling of the ligand from the binding site with the restraints on. Δ G (site,c+o+a+r), Δ G (site,c), Δ G (site, Θ), Δ G (site, Φ), Δ G (site, Ψ), Δ G (site, θ), Δ G (site, ϕ) and Δ G (site, r) represent the contribution of the restraints shown in Figure 2.4 A. Δ G (bulk, decouple) represents the contribution of the decoupling of the ligand from the bulk. Δ G (bulk, c), Δ G (bulk, o+a+r) are the contributions of the restraints on the ligand in the bulk.

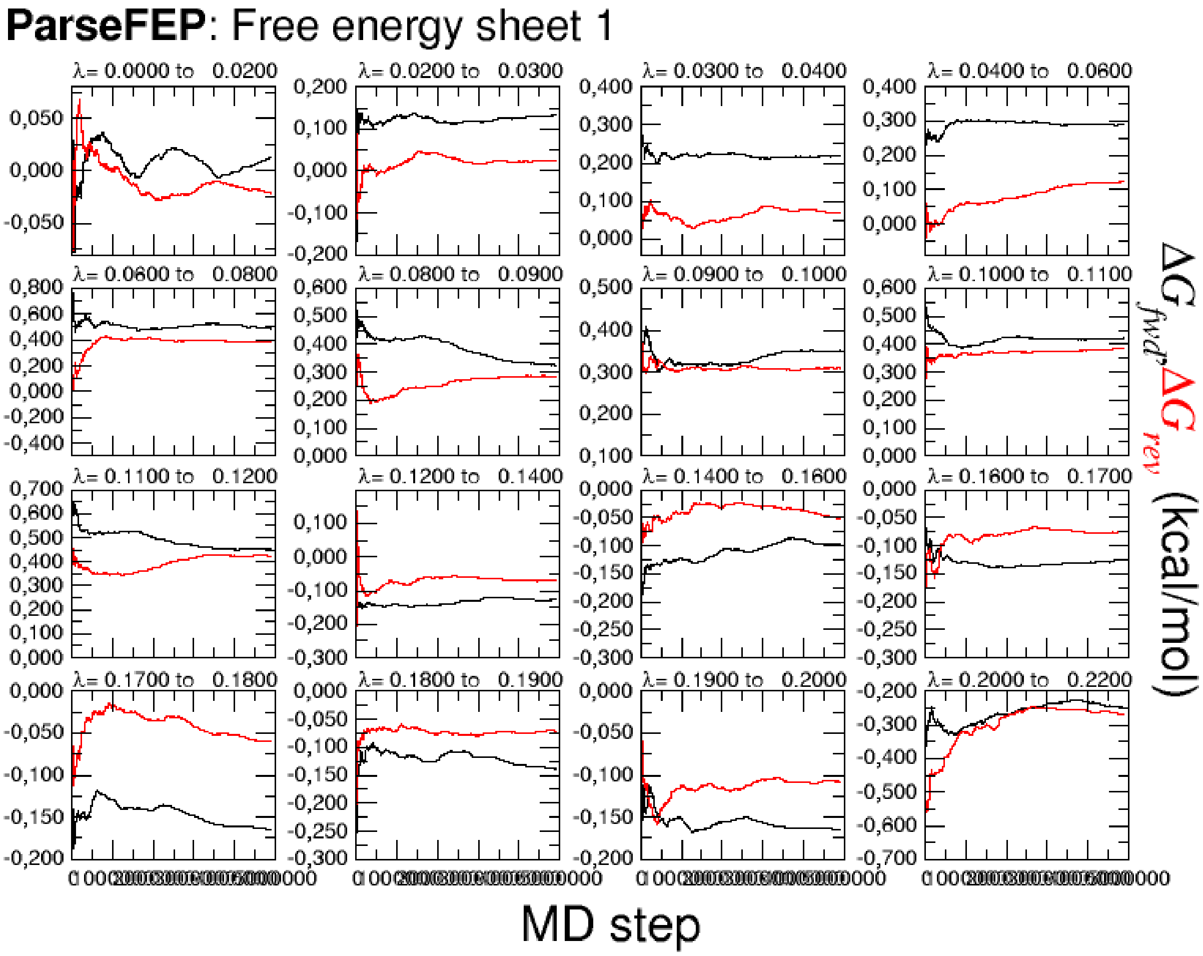


Figure 2.6: Example of windows obtained with the converged trajectory. All the windows have a hysteresis lower than 0.2 kcal/mol.

2.2.2 Concluding remarks

Table 2.2 shows the details of the simulation. Δ G(site,couple) required 1.24 μs of simulation. Without the set of restraints applied, it is very likely that it would not have been possible to compute this value. It shows the importance of the computation strategy. Applying the restraints require additional simulations to compute their cost, and slow down the immediate performance of the alchemical simulations. Yet, without them, it would not have been possible to compute the Δ G (total). Decreasing the degree of freedoms of the system facilitates the convergence of the simulation. It is then necessary to compute all the elements of the corresponding thermodynamic cycle (such as the one on Figure 2.4) to calculate the binding free energy.

CHAPTER 3

PROTEIN-PROTEIN BINDING SPECIFICITY

This chapter explores alternate computational techniques, such as end-points methods, to investigate the interface of binding in protein-protein complexes. It presents the results from the paper by Prithviraj Nandigrami, **Florence Szczepaniak**, Christopher T. Boughter, François Dehez, Christophe Chipot, and Benoît Roux: Computational Assessment of Protein – Protein Binding Specificity within a Family of Synaptic Surface Receptors. in *The Journal of Physical Chemistry B*, in October 2022. The goal of this work is to study binding mechanism, and how protein mutations give an insight on the binding specificity.

3.1 Dpr-DIP proteins

How living cells process information and make decisions at the molecular level is among the most important issues in biology. For example, many of the intra- and intercellular communications are mediated by protein–protein interactions and recognition processes involving receptors and ligands (Zipursky and Sanes (2010); Marquart et al. (1983); Özkan et al. (2013, 2014)). Atomic-level information is essential to explain the specific interactions between biological macromolecules in terms of structure and dynamics. Over the last two decades, there has been an accumulation of high-resolution protein structure data. Nevertheless, many proteins are orphans without identified homologues, and complexes with potential binding partners remain unknown. While protein–protein recognition is a core biophysical problem, the lack of reliable interaction data is a challenge when trying to understand large protein interactome databases. This points to a fundamental gap in our ability to understand the roles of structure and dynamics and how it controls the specificity of protein–protein complexes, a gap that can only be addressed through a multidisciplinary approach at the interface of biological, physical, and computational sciences.

A particularly rich paradigm to harden our understanding of protein-protein binding specificity is presented by the family of Dpr, Defective in Proboscis Extension Response, proteins (Nakamura et al. (2002); Zinn and Özkan (2017); Sergeeva et al. (2020)). The Dpr proteins belong to the immunoglobulin superfamily (IgSF), with 21 Dpr genes that can be found in *Drosophila melanogaster*. It has been shown that Dpr proteins interact with a family of previously uncharacterized proteins, the "Dpr Interacting Proteins" or named DIP, with 11 DIP genes (Özkan et al. (2013)). Utilizing a network of *Drosophila IgSF* cell surface proteins, synaptic connectivity and interaction specificity have been identified (Carrillo et al. (2015)), as well as interaction affinity and interaction specificity determinants of Dpr-DIP proteins (Cosmanescu et al. (2018)). Most members of the DIP subfamily cross-react with several members of the Dpr family and vice-versa (Carrillo et al. (2015); Cosmanescu et al. (2018)). For example, although Dpr6 and Dpr10 are homologous, only DIPα (CG32791) can bind to both while DIPβ (CG42343), which is homologous to DIPα (CG32791), only binds to Dpr6. Despite their importance in neuronal development, many of the cell surface receptor-ligands complexes related to synaptogenesis remain uncharacterized due to difficulties in crystallization and expression. These proteins have a highly specific, yet coevolutionary, drift of the receptor–ligand complex that can be observed in the case of Dpr and DIP protein pairs (Özkan et al. (2013); Tan et al. (2015)). The mechanism of such evolutionary drift in highly specific complexes is not yet clear.

The Dpr–DIP interactome provides a rich testing ground for the study of protein–protein interactions. The extensive characterization of each pair of homologous proteins in this interactome, both via measurements of binding affinity and phenotypic responses, provides baseline definitions of what constitutes an interacting or noninteracting pair. The most important question revolves around the identification, in a biological context, of the defining molecular features that discriminate between cognate and noncognate receptors for these homologous proteins. In the case of the Dpr–DIP interactome, the homology across the Dpr and DIP molecules necessitate the application of a range of physics-based approaches for deconvolution of the determinants of productive receptor pairs.

The Dpr and DIP family present a particularly interesting system – a network of homologous proteins with high sequence similarity whose inter-relations are reflected both with *in vivo* neuromorphogenesis and *in vitro* quantitative measurements of the binding affinities. While the former are revealed through knowledge-based approaches combining evolutionary and structural features, the latter require physics-based approaches relying on atomic models to pinpoint the molecular determinants controlling binding specificity. For example, experimental surface plasmon resonance (SPR) data have identified broad group of 57 Dpr–DIP cognate complexes showing interaction specificity while the remaining 174 pairs display no detectable binding affinity (Cosmanescu et al. (2018)). Computations based on atomic models are needed to clarify why, despite their high structural similarities, only a subset of these proteins bind together.

More generally, an understanding of the intricate details of Dpr-DIP system can help decipher the molecular code that governs the specific protein–protein interactions of such dynamical protein networks. Ultimately, having the ability to predict specific protein–protein association computationally could open the door to useful dissection of cell signaling pathways or the design of protein-based materials. In this context, the Dpr–DIP complexes offer a great opportunity to explore how different computational approaches can provide meaningful information about binding specificity and the overall structure of complex functional networks constructed from protein–protein interactions. An important objective of the present effort is to assess and contrast the ability of knowledge-based and physics-based computational approaches, to identify the relative strengths and weaknesses, to tackle the problem of binding specificity of the Dpr–DIP complexes, and ultimately, to see how they can be reliably used in efforts aimed at producing new knowledge about the molecular basis of protein–protein interactions. Through this analysis, we aim to characterize the global set of interactions between Dpr and DIP proteins and identify the specificity of binding between each DIP with their Dpr binding partners and characterize the influence of mutations on the specificity of these interactions.

The project relies on a series of experimental binding affinity data (Cosmanescu et al. (2018); Carrillo et al. (2015); Cheng et al. (2019)). To determine which protein can partner with another, Comanescu et al. and Cheng et al. have performed Surface Plasmon Resonance (SPR) (Cosmanescu et al. (2018); Cheng et al. (2019)). They proposed a list of cognate and non cognate complexes. Figure 3.1 shows the cognate complexes detected by SPR. Because of the experimental conditions, some Dpr may form dimers, preventing their interactions with DIP proteins and biasing the precise determination of K_ds of Dpr–DIP complexes. Moreover, the definition of cognate and non cognate complexes is ambiguous *in vitro*: The threshold separating interacting from non–interacting complexes is set arbitrarily and the list of interacting complexes varies depending on the experimental setup (Cosmanescu et al. (2018); Carrillo et al. (2015); Cheng et al. (2019)). Here we will rather refer to high affinity and low affinity complexes. For Dpr6–DIPα and Dpr6–DIPγ complexes, the reported affinities are quantitatively in lines in all studies, defining with no ambiguity a high affinity and low affinity complex respectively. The effect of mutations on the affinity of Dpr–DIP complexes have been quantified using SPR experiments (Cheng et al. (2019)). To investigate the role of individual residues at the origin of specificity, Cheng et al. performed *in vitro* alanine scanning. They determined whether a given residue stabilizes the interface and how much it contributes to the total free energy of binding (Figure 3.2) (Cheng et al. (2019)).

Among the 231 possible protein-protein complexes, we focused 11 complexes: the 11 DIP proteins interacting with the protein Dpr6. First of all, complexes with Dpr6 have been more studied experimentally, so there are more references to compare the results (Cosmanescu et al. (2018); Carrillo et al. (2015); Cheng et al. (2019)). Also, studying in details 231 complexes in detail would require very advanced techniques (such as hybrid Monte-Carlo)

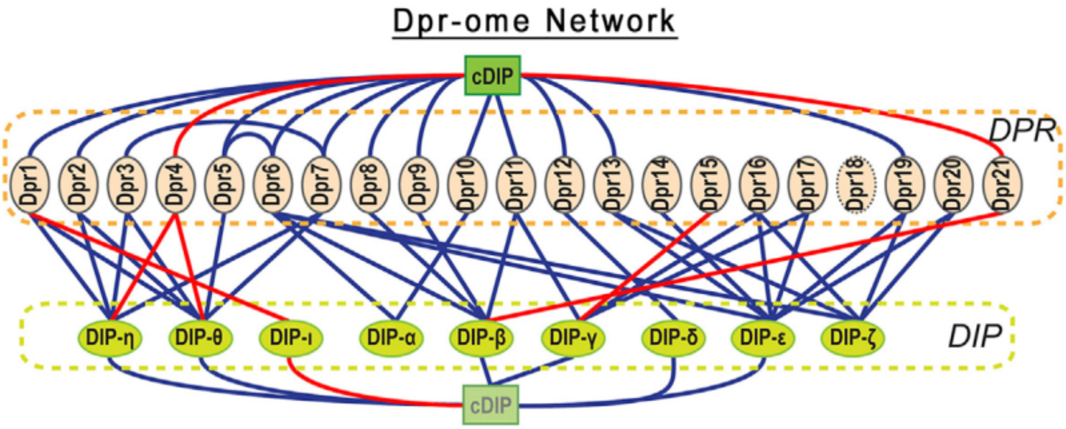


Figure 3.1: Map of the cognate partners between the DIP and the Dpr proteins. Figure taken from Ref Carrillo et al. (2015)

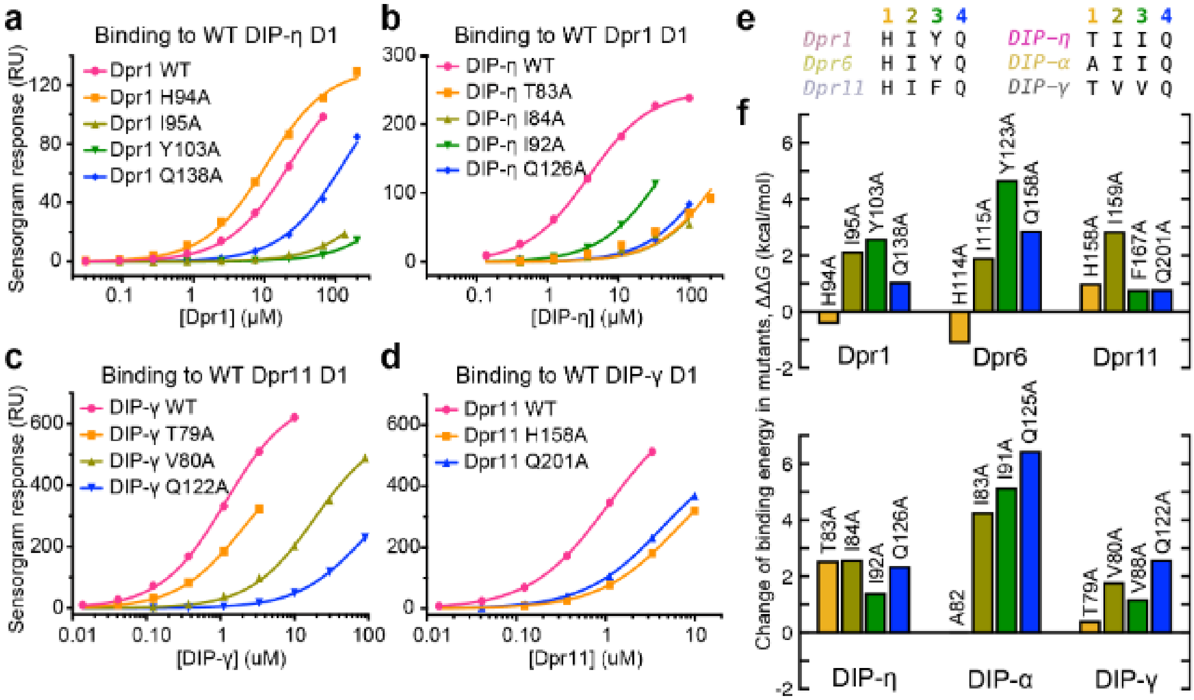


Figure 3.2: Results of the alanine scanning performed by Cheng et al. (2019)

and computational means. We resorted to fast methods to study the 231 complexes, but these methods don't give a detailed and precise insight of the binding selectivity.

3.2 Methods

3.2.1 Sequence Alignment and Homology Modeling

To prepare the similarity (homology) modeling (Šali and Blundell (1993); Šali et al. (1995)), multisequence alignment (MSA) were constructed of all 21 Dpr sequences and all 11 DIP sequences using the software CLUSTAL-W (Thompson et al. (1994)). The alignment was performed by setting a higher gap penalty mask at positions that correspond to secondary structure elements (α helices and β strands) and setting a lower gap penalty mask at positions that correspond to loops. This helps ensure that the secondary structure regions of the template crystal structures are locked in proper orientation in the homology models of the target Dpr–DIP sequences. The resulting alignment of each Dpr–DIP target sequences in PIR format was, then, used to generate structural models of all 231 possible Dpr–DIP complexes with the program MODELER (Šali and Blundell (1993); Šali et al. (1995)) using all available crystallographic structures as a set of templates (i.e.,Dpr6–DIPα, PDB id 5EO9 (Carrillo et al. (2015)); Dpr4–DIPη, PDB id 6EG0 (Cosmanescu et al. (2018)); Dpr2–DIPθ, PDB id 6EG1 (Cosmanescu et al. (2018)); Dpr1–DIPη (CG14010), PDB id 6NRW (Cheng et al. (2019)); Dpr10–DIPα, PDB id 6NRQ (Cheng et al. (2019)); Dpr11–DIPγ, PDB id 6NRR (Cheng et al. (2019))). Additional side chain refinement of the models were performed subsequently with the program ROSETTA (Rohl et al. (2004); Das and Baker (2008)). The reliability of the structural models was assessed by comparing the structural alignment of a model of Dpr6-DIPα (generated by excluding the crystal structure of this complex from the set of templates) and its corresponding crystal structure. A structural comparison of the X-ray structure and the model yields an overall root-mean-square deviation (RMSD) less

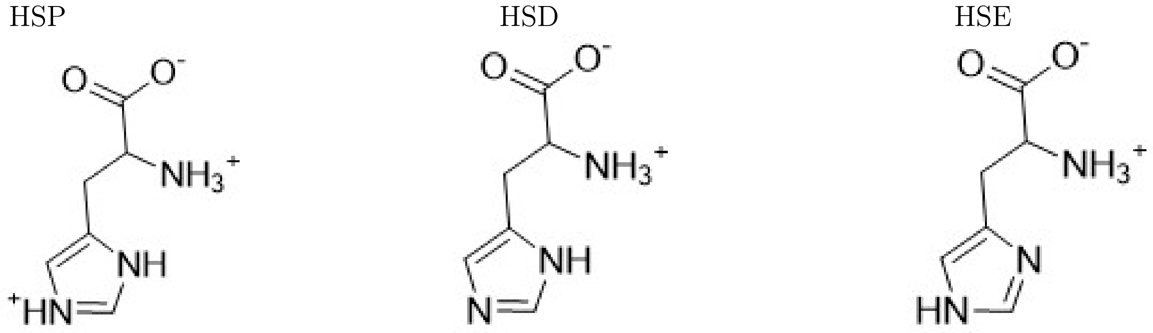
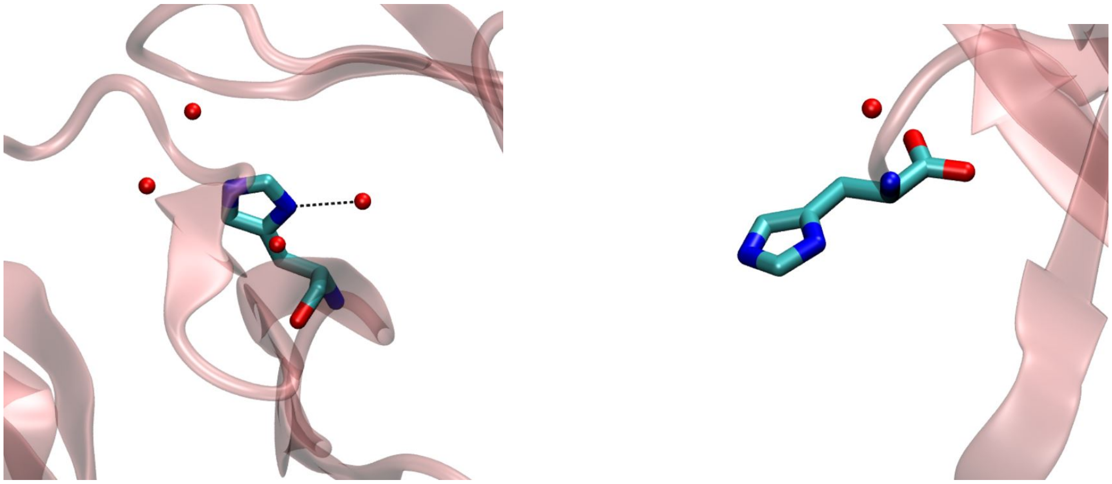


Figure 3.3: The three protonation states of the Histidines.

than 2 Å (all atoms), indicating that the alignment/modeling procedure is sound.

3.2.2 The Protonation states of the Histidines

To study the binding specificity, it is necessary to study the interface between the two proteins. Among the different kind of possible interactions between the proteins, there is the formation of hydrogen bonds between residues. In classical MD, except in specific cases like constant pH simulations, the protonation states of a residue is constant through the whole simulation. Yet, the protonation state of a residue is not always given by the experimental structure. For most residue, the protonation state at pH = 7 is constant and known, so the lack of information of the hydrogens is not a problem. But for the Histidine, it is less obvious. The protonation states of the Histidines have to be carefully chosen between the three possible forms shown in Figure 3.3. The protonation state of each histidine was assigned based on its chemical environment and its apparent pKa predicted by propKa (Berman (2000); Olsson et al. (2011)). With this software, the protonation state HSP was assigned when needed. Then, using VMD (Humphrey et al. (1996)), the structure of the Dpr6–DIPα complex is observed, and especially all the nitrogens of the Histidines and the surrounding oxygens within 4 Å. The pKa can not decipher between HSD and HSE, so there is a need to a more careful approach. The protonation state of the histidine was then decided depending on where a hydrogen bond could be formed (Figure 3.4). Whenever no



HIS114: HSD due to an O within 4 Å

HIS177: HSE because undetermined

Figure 3.4: The attribution of the protonation states of the Histidines in the complex Dpr6–DIPα.

hydrogen bond could be clearly defined, the protonation state chosen by default was HSE. The protonation state has been generalized to all complexes that have been simulated by homology.

3.2.3 Molecular Dynamics Simulations

The molecular systems were subsequently hydrated in 0.15 M NaCl using the solvate and autoionize plugins of VMD Humphrey et al. (1996). The resulting systems consist each of ∼48.000 atoms embedded in a cubic box of 80 Å side-length. All simulations were performed in the isothermal-isobaric ensemble using the program NAMD (Phillips et al. (2005)). The proteins and ions were described using the CHARMM36 force-field (Jo et al. (2017)) and TIP3P (Jorgensen et al. (1983)) was used for water. A Langevin thermostat and a Langevin piston were employed to maintain a temperature of 300K and a pressure of 1.0315 bar, respectively (Feller et al. (1995)). The Particle–Mesh Ewald algorithm was used to handle long–range interactions (Darden et al. (1993)). Above 12 Å, electrostatic and Van der Waals interactions were truncated with a switching distance of 14 Å. Integration was performed with a time step of 4 and 2 fs for long- and short-range interactions, respectively, employing the r-RESPA multiple time-stepping algorithm (Tuckerman et al. (1992)). The SHAKE/RATTLE (Ryckaert et al. (1977); Andersen (1983)). was used to constrain covalent bonds involving hydrogen atoms to their experimental lengths, and the SETTLE algorithm (Miyamoto and Kollman (1992)) was utilized for water. For Dpr6–DIPα, we employed the following protocol. First the energy of the system was minimized for 1000 steps. Next, the hydrated complex was equilibrated during 10 ns using soft-harmonic restraints (1 kcal/mol/Å) to maintain all backbone-atoms to their initial positions. Harmonic restraints were removed step-wisely over 1 ns followed by a production run of 200 ns. For all the other complexes resulting from homology modeling, we add an extra equilibration step to the latter protocol. Positions of backbone atoms were kept fixed and side chains were restrained harmonically. The energy of the system was minimized for 1000 steps and equilibrated for 10 ns. Side-chain restraints were removed step-wisely over 1 ns. Next unrestrained side chains were equilibrated during 20 ns while applying harmonic positional restraints to the backbone atoms. All harmonic restraints were eventually removed step-wisely over 1 ns followed by a production run of 200 ns.

3.2.4 Poisson–Boltzmann Calculations

A continuum solvent model based on the Poisson–Boltzmann (PB) equation provides a useful and computationally inexpensive approach for calculating the binding free energies. The binding free energy ΔG_b between two proteins A and B is expressed as eq 3.5. The nonpolar contribution to the binding free energy, ΔG_{VDW}, is empirically written as a fraction of the average van der Waals interactions λ ΔU_{VDW} upon formation of the complex, where $\lambda = 0.17$ is an empirical scaling factor meant to account for the missing protein–solvent van der Waals interaction in the implicit solvent representation (Eriksson and Roux (2002)). The value of λ (0.17) is similar to a VDW scaling factor (0.161) in the linear interaction energy (LIE) method (Åqvist et al. (1994); Aqvist and Marelius (2001)). (29,30) No term proportional to the solvent accessible surface area (SASA) is included. The electrostatic free energy contribution ΔG_{PB} is expressed as

$$\Delta G_{PB}^{AB} = G_{PB}^{AB} - G_{PB}^{A} - G_{PB}^{B} \tag{3.1}$$

G_{PB}^{AB} is the total electrostatic free energy of the complex AB, and G_{PB}^{A} and G_{PB}^{B} are the total electrostatic free energy of the isolated protein A and B, respectively. For each cases (AB, A, B), these free energies are calculated as

$$G_{PB} = \frac{1}{2} \sum_i q_i \phi(\mathbf{r}_i) \tag{3.2}$$

where $\phi(\mathbf{r}_i)$ is the electrostatic potential determined by solving the finite-difference PB equation (FDPB)

$$\nabla \cdot \epsilon(\mathbf{r}) \nabla \phi(\mathbf{r}) = -4\pi \rho_{PB}(\mathbf{r}) \tag{3.3}$$

The space-dependent dielectric $\epsilon(\mathbf{r})$ has a value of 80 in the bulk solvent, and a value of 12 in the protein interior. The dielectric constant of the protein interior is taken from previous work (Eriksson and Roux (2002)). Atomic Born radii used to define the protein–solvent dielectric interface were taken from ref Nina et al. (1997). The FDPB equation is solved using the PBEQ module (Im et al. (1998)) of the program CHARMM (Brooks et al. (2009)). A focusing algorithm with was used, starting with a first grid of 1 Å spacing that covers more than 2.6 times the length of the solute in the XYZ directions, then followed by a second FDPB calculations with a finer grid with a mesh spacing of 0.45 Å. In a first round, the continuum model was used to score directly the homology models. In a second calculation round, the binding free energy was recalculated and averaged using approximately 400 configurations taken from the 200 ns MD trajectories of the complexes. The last third of the trajectory was used for analysis (approximately 65 ns). The scoring of each MD trajectory using the continuum solvent PB model takes approximately 6 h for completion on a CPU core, a total of 1386 h for the 231 complexes.

3.2.5 Alchemical Free Energy Perturbation Calculations

A series of in-silico alanine scanning (ALA-scan) were performed over a subset of residues lying at the complex interface of Dpr6–DIPα and Dpr6–DIPγ, respectively. All necessary input files were generated using the alanine scanning plugin of VMD (Ramadoss et al. (2016)). The perturbations associated with the mutation have been computed using the free energy perturbation (FEP) theory (Chipot and Pohorille (2007); Zwanzig (1954)). For each system, a decoupling (forward) and a coupling (backward) transformation were performed. The ParseFEP VMD plugin (Liu et al. (2012)) was employed to analyze the resulting data and to compute the free-energy changes and their associated uncertainties using the maximum-likelihood Bennet Acceptance Ratio estimator (Bennett (1976)).

For each mutation, alchemical transformations were stratified in a series of λ windows. For each system, an initial backward-forward transformation involving 20 windows in each direction was performed. Each window consisted in a 200 ps equilibration trajectory followed by a 1 ns production run. To ensure proper convergence of the free-energy estimates, the level of stratification was increased in windows exhibiting a large hysteresis (higher than 0.2 kcal·mol^{-1}) between the backward and the forward transformation.

In the case of the residue H114 of the protein Dpr6 of the Dpr6–DIPα complex, the position of the histidine at the beginning of the forward trajectory and at the end of the backward trajectory (so in both cases when the histidine was coupled with the system) was totally different. The random movements of the histidine when it is decoupled from the rest of the system led to the loss of a hydrogen bond between the histidine and the residue Q125 on DIPα. To ensure that the histidine side chain remains in a relatively stable configuration during the free-energy simulation, restraints were applied to the $\chi 1$ dihedral angle (Cα–Cβ–Cγ–Nδ1) (Fiorin et al. (2013)). The cost of this restraint was, then, computed by thermodynamical integration (TI) and removed from the final result. The simulation of each backward-forward trajectory of the FEP takes approximately 6 h for completion on GPU core.

3.2.6 Simple Scoring Method for Amino Acid Interactions

The scoring of interaction in the sequence analysis is based on the Automated Immune Molecule Separator (AIMS) software (Boughter et al. (2020)). While the AIMS software is capable of characterizing amino acid sequences in physical terms, it lacks any ability to account for protein–protein interactions. For each biophysical property, there exist rules for how these properties constructively or destructively contribute to protein binding. We know that opposite charges attract, like charges repel, and hydrophobic groups "prefer" to be buried away from solution. However, programming in these physical rules is nontrivial, especially when we are attempting to minimize the structural information used in this particular approach. The calculation of Newtonian forces using for example Coulomb's law or the equation for the van der Waals interaction require explicit distance information not available for all sequences studied in the present work.

Instead, we can try to approximate these interaction rules using a simple scoring metric (table 3.1). The table tries to recapitulate the interactions between amino acids at the level of an introductory biochemistry course. Arginine and aspartate can form a salt bridge, so this interaction receives a score of +2. Lysine-lysine or lysine-leucine residue pairs are repulsive or entropically unfavorable, so these highly unfavorable interactions receive a score of -2. While further refinement of these interaction scores are certainly possible and indeed should be the focus of future studies, the performance of this first interaction scoring matrix was sufficient for the present study. All code for replicating this analysis using AIMS, ver.

Table 3.1: Simple scoring metric for AIMS

Res	A	G	L	M	F	W	K	Q	E	S	P	V	I	C	Y	H	R	N	D	T
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	-2	-1	-2	0
M	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	-2	-1	-2	0
F	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	-2	-1	-2	0
W	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	-2	-1	-2	0
K	0	0	-2	-2	-2	-2	-2	1	2	0	-1	-2	-2	$\frac{1}{2}$	1	1	-2	1	2	0
Q	0	0	-1	-1	-1	-1	1	1	1	0	-1	-1	-1	$\frac{1}{2}$	1	1	-2	1	2	0
E	0	0	-2	-2	-2	-2	-2	1	-2	0	-1	-2	-2	$\frac{1}{2}$	1	1	2	1	-2	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	1	1	1	1	-1	-1	-1	0	0	1	1	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	-2	-1	-2	0
V	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	-2	-1	-2	0
I	0	0	1	1	1	1	-2	-1	-2	0	1	1	1	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	-2	-1	-2	0
C	0	0	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	2	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0
Y	0	0	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	1	1	1	0	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	1	1	1	1	1	0
H	0	0	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	1	1	1	0	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	1	1	1	1	1	0
R	0	0	-2	-2	-2	-2	-2	1	2	0	-2	-2	-2	$\frac{1}{2}$	1	1	-2	1	2	0
N	0	0	-1	-1	-1	-1	1	1	1	0	-1	-1	-1	$\frac{1}{2}$	1	1	1	1	1	0
D	0	0	-2	-2	-2	-2	-2	1	-2	0	-2	-2	-2	$\frac{1}{2}$	1	1	2	1	-2	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

0.6, is freely available on GitHub.

3.3 Results and Discussion

A central hypothesis in trying to understand the cellular receptors controlling morphogenesis is that the equilibrium binding affinity of Dpr–DIP complexes is a key determinant of biological function (Sperry (1963); Barish et al. (2018); Ashley et al. (2019); Courgeon and Desplan (2019)). That is, cognate complexes form and drive morphogenesis when their binding affinity is above a certain threshold. Whether this hypothesis is correct or not remains unclear at this point. Nonetheless, the binding affinity is likely to be one important marker of biological function, even if additional factors play important roles. In this regard, the binding data from surface plasmon resonance (SPR) experiments play a central role in the present analysis (Cosmanescu et al. (2018)). The SPR data identified 57 Dpr–DIP cognate complexes (Cosmanescu et al. (2018)). One may note that the measured dissociation constants are reported for protein complexes only to some lower limit. It is tempting to interpret the interacting complexes with detectable binding affinity as “cognate” pairs, and consider that the remaining 174 undetected and weakly interacting complexes as “noncognate” pairs. However, it is unclear whether the biological neuromorphogenesis function of these Dpr–DIP complexes is simply correlated with the binding affinity or some other factors are at play. In that sense, it is possible that some of the complexes undetected by SPR measurements could still represent biologically meaningful cognate pairs. Alternatively, it is possible that in cellulo, complexes form only for binding affinities above a certain threshold and only a smaller subset of 57 complexes discriminated by SPR measurements represent true biologically meaningful cognate pairs. For the sake of simplicity, we will refer to the detected and undetected complexes as cognate and noncognate complexes in the following. Nonetheless, one should note that this nomenclature does not imply that there is unambiguous certainty about the true biological function. To pursue this analysis, it is useful to convert the experimental dissociation constant Kd into an effective binding free energy, as $\Delta G_{bind} = k_B T \ln(K_d)$ with a standard concentration of 1 M. We use these values to assign a binding free energy to those complexes and will specifically examine the impact of the cutoff value on the interaction analysis of Dpr–DIP complexes.

3.3.1 Parsing Molecular Differences in Amino Acid Sequences

A straightforward first-pass analysis of the Dpr–DIP interactome begins with an inspection of the primary structure of each individual protein in our data set. Multisequence alignments (MSA) were constructed for all 21 Dpr sequences and all 11 DIP sequences using the software CLUSTAL-W (Thompson et al. (1994)). As shown in Figure 3.5, DIP proteins have a higher degree of conservation than the Dpr proteins, that is, there are more insertions (gaps) in the alignment of Dpr protein sequences compared to the alignment of DIP protein sequences.

Given the large number of receptors in this interactome, a simple visual inspection of this alignment is insufficient to provide meaningful insights into the key features that

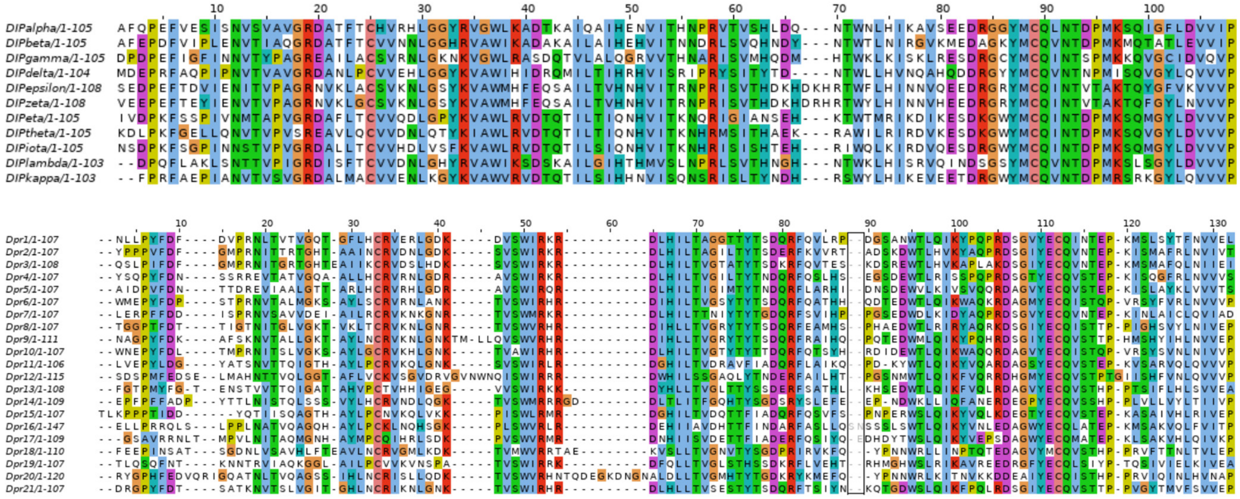


Figure 3.5: Alignment of 11 DIP sequences (top) and alignment of 21 Dpr sequences (bottom) performed using CLUSTAL (Thompson et al. (1994)) by conserving the secondary structure regions in template crystal structures. The dashed region (shown inside a box) for Dpr sequence alignment represents a long region of 36 residues that is present only in Dpr16 and absent in all other Dpr sequences

may be responsible for specificity of cognate pairs. To deepen our perspective, we used the Automated Immune Molecule Separator (AIMS) (Boughter et al. (2020)) software to quantitatively characterize these sequence pairs. While AIMS was originally developed for the analysis of immune molecules, such as antibodies, T cell receptors, and MHC-presented peptides, the analytical pipeline readily extends into the analysis of sequences preprocessed using multisequence alignment schemes like that shown in Figure 3.5. This multisequence alignment was encoded into an AIMS-compatible numeric matrix with each row in the matrix representing cognate (57 Dpr–DIP pairs) or noncognate (174 Dpr–DIP pairs) receptor pairs and each column corresponding to a location in the multisequence alignment.

Using this matrix generated by the MSA and AIMS, we can, then, calculate the average biophysical properties for the cognate and noncognate groups as a function of the position in the multisequence alignment. As expected, the average position-sensitive properties of the cognate and noncognate receptor pairs are nearly identical. The ability of AIMS to discern the biophysical differences in two distinct molecular groups is hampered in this application because of the significant overlap in the entries of each class. Each individual Dpr and DIP molecule is found in both the cognate and noncognate groups, albeit at different frequencies.

To overcome this deficiency, we can turn to an analysis centered primarily on the incorporation of interprotein interactions, beginning first with an information theoretic approach. The AIMS platform utilizes information theory, a framework classically applied to communication across noisy channels, to quantify biochemical interactions through calculations of the Shannon entropy and mutual information (Shannon (1948)). Shannon entropy, in its simplest form, can be used as a proxy for the diversity in a given input population. This entropy at position i is formulated as, $H_i = -\sum_x p_i(x) log_2 p_i(x)$, where $p_i(x)$ is the occurrence probability of a given amino acid at position i in the MSA of the two proteins, and the sum is over all possible 20 amino acids. Likewise, the joint entropy for two positions i and j is defined as,$H_{ij} = -\sum_{x,y} p_{ij}(x,y) log_2 p_{ij}(x,y)$, where $p_{ij}(x,y)$ is the joint probability of the amino acids at the positions i and j in the MSA, which relates back to the standard entropy by the equivalence $p_i(x) = \sum_y p_{ij}(x,y)$. We note that despite the increased degeneracy of each Dpr and DIP pair in the noncognate group, and the larger size of this group, the positional entropy is broadly similar for the cognate and noncognate receptors. Given the lack of strong discrepancies in the positional entropy, we can subsequently directly compare the mutual information of the cognate and noncognate pairs. Mutual information is capable of identifying residues within cognate Dpr–DIP receptors that covary in any meaningful way, whether it be a positive or negative correlation between residues or some nonlinear relation. The mutual information Iij associated with positions i and j is defined as the difference between the individual entropy of each site, Hi and Hj and the joint entropy Hij

$$I_{ij} = -H_{ij} + H_i + H_j = \sum_{x,y} p_{ij}(x,y) log_2 \frac{p_{ij}(x,y)}{p_i(x) p_j(y)} \tag{3.4}$$

In other words, given a pair of positions in the MSA, how much information is gained by accounting for the covariation of the amino acids in Dpr and DIP relative to that treating them as independent? Looking at the difference in the interprotein positional information between the cognate and noncognate pairs we see, again as we would expect, that there is a higher mutual information between the cognate Dpr–DIP receptors than the noncognate group at all positions. The mutual information between noncognate receptors is treated as noise and subtracted from the mutual information between cognate receptors, as these pairs generated via simple combinatorics are not expected to contain meaningful information.

Focusing on the mutual information between interacting residues on Dpr and DIP proteins (Figure 3.6A) clear hotspots can be identified, where there exists heightened information gain between specific regions of these proteins. Given that the coevolution of residues can lead to high mutual information between distal residues that likely do not play a role in the definition of cognate and noncognate pairs, (Lockless and Ranganathan (1999); Salinas and Ranganathan (2018)) we weighted the mutual information by the inverse distance between residues (Figure 3.6B). The distances were calculated using the crystal structure of the Dpr10–DIPα complex (Cheng et al. (2019)), which is expected to be representative of all the possible complexes for these homologous proteins. We then aim to identify how these proximal residues with high mutual information might be able to discriminate between cognate and noncognate receptors. The mutual information is a summary metric for groups of sequences. To generate a score from these residues, we begin with a simple classification of residues, which should interact positively or negatively in an interface (Table 3.1). The top five residue pairs with the highest mutual information and the associated scores with these pairs across the cognate and noncognate receptors is depicted in Figure 3.6 C and D.

A first key observation that can be made is that residue positions with a high mutual information are not always key interacting residues, which emphasizes that a simple extrapolation between knowing key residues at a given Dpr–DIP interface and inferring about their binding profile is not possible. Two of the top five highest mutual information sites in the Dpr–DIP interactome, the GLN93/ALA151 and VAL95/LYS150 residue pairs in the Dpr10–DIPα complex, appear to be relatively distant in this same structure. A second

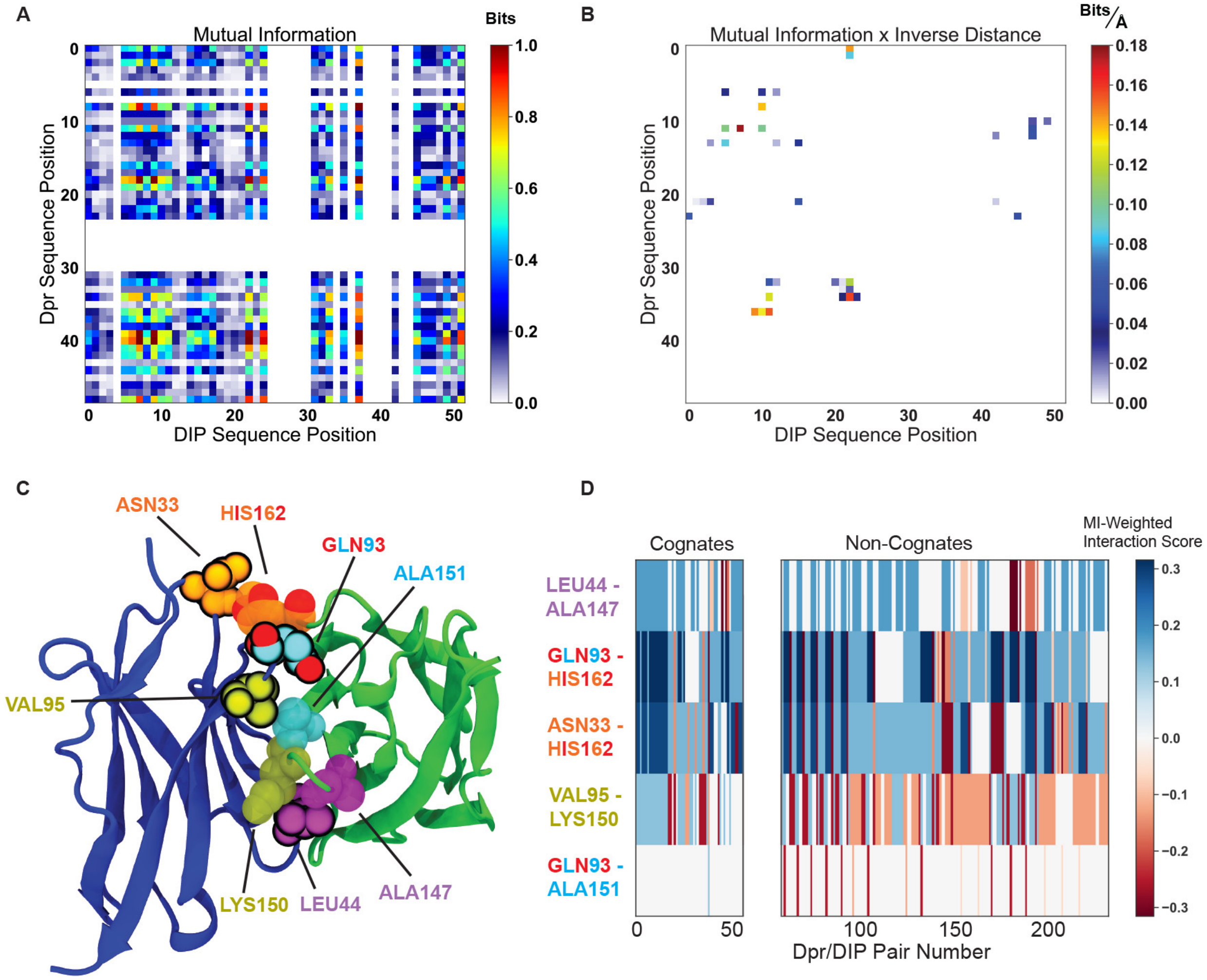


Figure 3.6: Mutual information analysis and AIMS scores of Dpr–DIP complexes. (A) Interprotein mutual information of the residues located at the interface of the Dpr–DIP complex. (B) Mutual information on panel A multiplied by the inverse distance between residues in the Dpr10–DIPα crystal structure. Residues further than 6 Å away were excluded from the analysis. (C) Representative Dpr10–DIPα structure shown, with the backbone trace of Dpr10 in blue and the backbone trace of DIPα in green. Dpr10 residues given opaque representation, while DIPα residues are translucent. (D) Residue interaction scores of these same residues for cognate and noncognate Dpr–DIP pairs. The score is weighted by the mutual information at that given site. Colors are matched across the structure, the y-axis of the interaction score plot, and the key.

observation is that the analysis does not identify specific key determinant residues that would be responsible for controlling the formation of cognate or noncognate Dpr–DIP complexes. Such key “gatekeeping” residues do not appear to be present in the Dpr–DIP complexes according to this analysis.

Instead, there exists a mosaic of interactions that define the cognate or noncognate receptor pairs. We observe, for instance, that while the Dpr10–DIPα pair has a quite unfavorable interaction (VAL95–LYS150), if we look at the corresponding residue pairs from the MSA, we find that this becomes an interaction beneficial to binding more frequently for cognate Dpr–DIP pairs than for noncognate pairs. Specifically, this VAL–LYS interaction in Dpr10–DIPα becomes LYS–GLN in cognate pairs Dpr1–DIPη, Dpr2–DIPθ, Dpr3–DIPι, and Dpr17–DIPγ, to name a few. Likewise, we see at the GLN93–ALA151 interaction site a mostly negligible interaction, but we find that it is only ever an interaction that may disrupt binding for noncognate pairs. Specifically, the disruptive interactions at this site comes from the methionine in DIPδ (MSA position 44) that interacts with the mostly hydrophilic residues at Dpr MSA position 117. The only “true” interaction partners with DIPδ, Dpr9 and Dpr12, form either a negligible (threonine) or beneficial (proline) interaction at this site.

3.3.2 Atomic Models of Dpr–DIP Complexes

Ultimately, to go beyond sequence analysis and attempt to identify key interactions missed by this bioinformatic approach, it is necessary to explicitly consider atomic models of the Dpr–DIP complexes and physics-based interactions. While there is a total of 231 possible Dpr–DIP complexes, only a relatively modest number of Dpr–DIP heterodimer crystal structures are available (Dpr6–DIPα, PDB id 5EO9 (Carrillo et al. (2015)); Dpr4–DIPη, PDB id 6EG0 (Cosmanescu et al. (2018)); Dpr2–DIPθ, PDB id 6EG1 (Cosmanescu et al. (2018)); Dpr1–DIPη (CG14010), PDB id 6NRW (Cheng et al. (2019)); Dpr10–DIPα, PDB id 6NRQ (Cheng et al. (2019)); Dpr11–DIPγ, PDB id 6NRR (Cheng et al. (2019))). It is interesting to compare the RMSD values of crystal structures of DIP proteins when they exist as monomers and when they are in complex with a Dpr partner. Comparing four DIP crystal structures, DIPα, (Cosmanescu et al. (2018)) DIPθ, ((Cosmanescu et al. (2018))) DIPη, (Cheng et al. (2019)) and DIPγ (Cheng et al. (2019)) between their monomer and heterodimer forms, we find that the RMSD of DIPα (PDB id 6EFY) as a monomer with DIPα in a complex to be around 0.65 Å, RMSD of DIPθ (PDB id 6EFZ) as a monomer with DIPθ in a complex to be around 1.5 Å, RMSD of DIPη (6NRX) as a monomer with DIPη in a complex to be around 0.62 Å, and RMSD of DIPγ (PDB id 6NS1) as a monomer with DIPγ in a complex to be around 0.74 Å. This analysis suggests that protein–protein interactions does not significantly alter the structures of the monomers in Dpr–DIP complexes. For this reason, it is justified to generate all the remaining cognate and noncognate complexes computationally using the known experimental structures as templates. Structural comparison indicate that available X-ray structures can serve as meaningful templates to generate accurate models of these homologous complexes. Structural models of all 231 possible Dpr–DIP complexes were built with the program MODELER (Šali and Blundell (1993); Šali et al. (1995)) using all available crystallographic structures as a set of templates. The multisequence alignment (MSA) shown in Figure 3.5 was used to match the target sequence to the template structures to generate the models (Šali and Blundell (1993); Šali et al. (1995)). The side chain rotamers were refined using ROSETTA Rohl et al. (2004); Das and Baker (2008)).

To further refine the models, all-atom molecular dynamics (MD) 200 ns trajectories were generated for each of the 231 possible Dpr–DIP complexes in explicit solvent. The primary objective of these simulations is to refine the models and ascertain their stability. The 200 ns trajectories are sufficiently long to allow the relaxation of the initial coordinates from the homology models toward stable complexes, and then, one can explore the small fluctuations within these stable complexes. Figure 3.7 shows a few representative sets of trajectories for several cognate and noncognate Dpr–DIP complexes over the simulated trajectory length of 200 ns. In fact, several of the models of Dpr–DIP complexes displayed instabilities marked by a progressive increase in RMSD. Further analysis indicated that these instabilities were correlated with the protonation state of histidine residues (detailed previously). It is noteworthy that these changes improved the stability of both the cognate and noncognate Dpr–DIP complexes. All the trajectories are stable in the simulation time scale of 200 ns. The results are illustrated in Figure 3.7 in the case of 4 cognate complexes (top row Dpr4–DIPι, Dpr4–DIPθ, Dpr7–DIPι, and Dpr7–DIPθ) and in the case of 4 noncognate complexes (bottom row Dpr1–DIPδ, Dpr2–DIPλ, Dpr8–DIPη, and Dpr10–DIPκ). These stable simulations provide an equilibrium configurational ensemble of all possible 231 complexes, which will subsequently be used to estimate the binding free energy.

In Figure 4, the global RMSD (in Å) of each of the 231 Dpr–DIP complexes is reported with the binding free energy extracted from the SPR measurements. It is apparent that the RMSD is insufficient to accurately distinguish between the cognate and noncognate complexes. Nonetheless, the RMSD are below 2.5 Å for most cognate complexes, and frequently up to 3.5 Å for a large number of noncognate complexes. The weak but nonzero correlation between the RMSD and its binding free energy is indicative of the possible importance of dynamical fluctuations in classifying the different Dpr–DIP complexes. Such information will be examined below in the linear discriminant analysis (LDA).

3.4 Computational Estimates of the Binding Free Energy and Scoring

3.4.1 Poisson–Boltzmann Continuum Model

The binding free energy of each Dpr–DIP complex was calculated using an implicit solvation continuum approximation, comprising a Poisson–Boltzmann (PB) electrostatic contribution

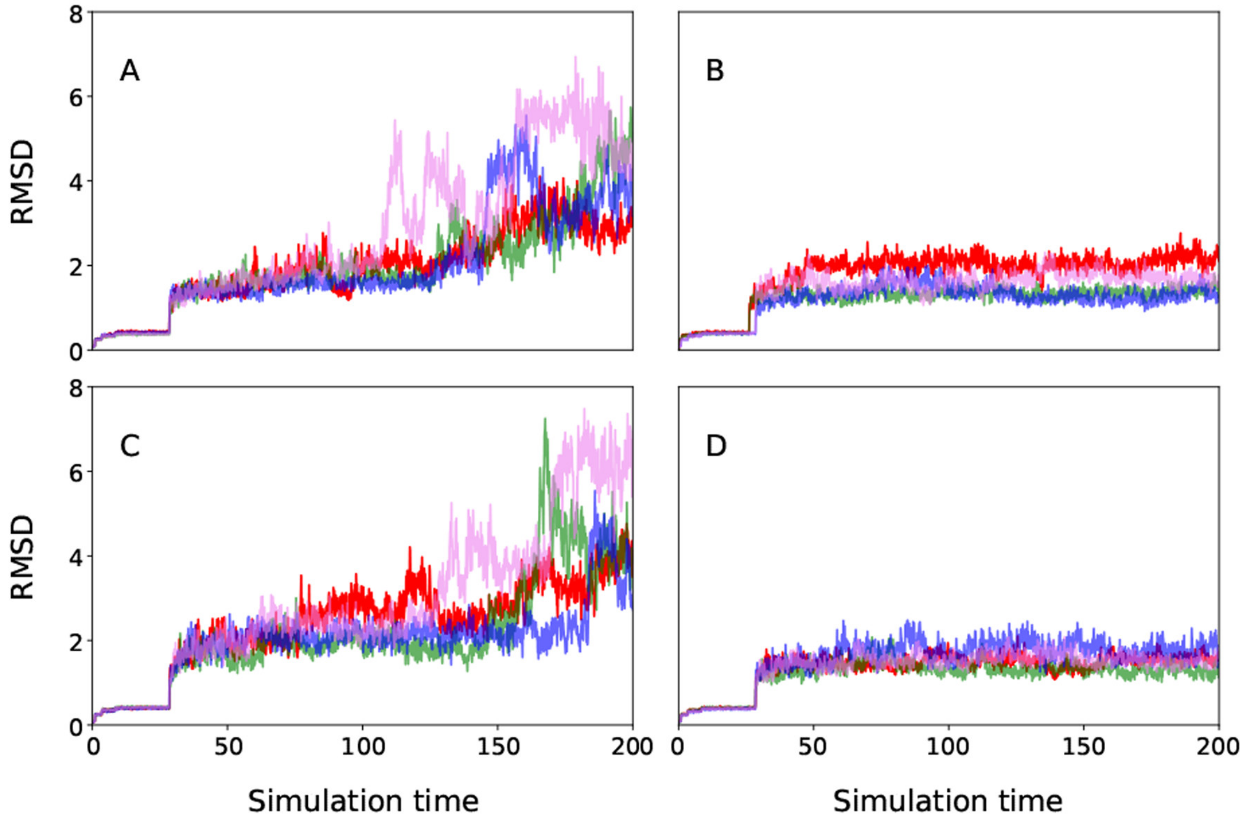


Figure 3.7: Global RMSD time-series plots for 4 cognate complexes (top row Dpr4–DIPι (red), Dpr4–DIPθ (green), Dpr7–DIPι (blue), and Dpr7–DIPθ (violet)) and 4 noncognate complexes (bottom row Dpr1–DIPδ (red), Dpr2–DIPλ (green), Dpr8–DIPeta (blue), and Dpr10–DIPκ (violet)). Unstable trajectories with incorrect assignment of histidine protonation states (left column, (A) unstable cognate complex and (C) unstable noncognate complex), and stable trajectories with accurate assignment of histidine protonation states (right column, (B) stable cognate complex and (D) stable noncognate complex). The x-axis is in nanoseconds, and the y-axis is in angstroms.

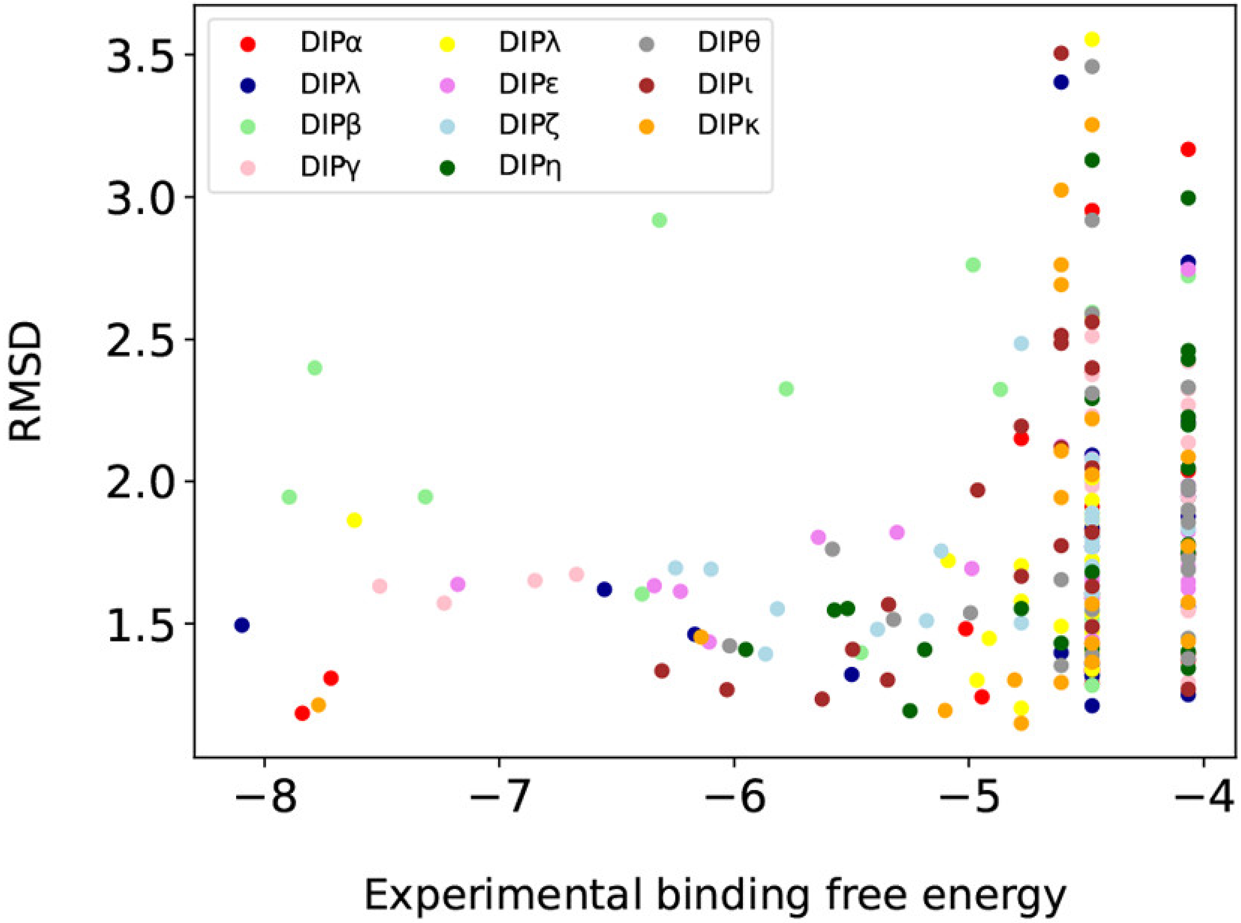


Figure 3.8: Comparison of the calculated average backbone RMSD obtained from MD with explicit solvent with the corresponding binding free energy extracted from the SPR measurements for all 231 Dpr–DIP complexes (Cosmanescu et al. (2018)). Different colors represent each of the 11 DIP protein bound to their 21 Dpr partners. The RMSD was calculated by averaging over the final 60 ns of the MD simulations. The x-axis is in kilocalories per mole, and the y-axis is in angstroms.

and a scaled van der Waals (VDW) dispersion contribution (Eriksson and Roux (2002)). Briefly, the binding free energy between two proteins A and B in this PB-VDW model is expressed as

$$\Delta G_b = \Delta G_{PB} + \Delta G_{VDW} \tag{3.5}$$

where $\Delta\mathrm{G}_{VDW}$ and $\Delta\mathrm{G}_{VPB}$ are the VDW nonpolar and PB electrostatic interaction upon formation of the protein–protein complex, respectively. The nonpolar VDW contribution $\Delta\mathrm{G}_{VDW}$ is empirically written as a fraction of the average protein–protein VDW interaction $\lambda\Delta\mathrm{U}_{VDW}$ upon formation of the complex, where $\lambda = 0.17$ is an empirical scaling factor meant to account for the missing protein–solvent VDW interaction in the implicit solvent representation (Eriksson and Roux (2002)). No term proportional to the solvent accessible surface area (SASA) is included. The scaling factor λ is designed to account for the change from protein–water to protein–protein VDW interactions. Briefly, both the average protein–protein VDW interactions in the complex and the (missing) protein–solvent VDW interactions in the dissociated system correspond to similar sums over short-range attractive 1/r6 over VDW centers. The small positive value of λ (0.17) reflects the fact that the favorable protein–protein dispersion interactions in the complex are not completely compensated by protein–water interactions in the dissociated state because proteins presents a higher density of VDW centers than liquid water. One may note that a very similar factor (equal to 0.161) appears in the Linear Interaction Energy (LIE) method of Åqvist (Åqvist et al. (1994); Aqvist and Marelius (2001)).

The PB-VDW estimates were calculated from eq 3.5 using a single trajectory of the complexes (no trajectories of the unbound proteins was used). Effects of induced structural shifts, and conformational entropy on the binding specificity were neglected. Ignoring the effect of induced structural shifts is justified, given that the backbone of the monomers is not significantly affected by the formation of the complex; see the comparison of the RMSD values of crystal structures of DIP proteins when they exist as monomers and when they are in complex with a Dpr partner in the Atomic Models of Dpr–DIP Complexes section. The issue of conformational entropy is more delicate. While the contribution from conformational entropy can affect the global binding affinity of two proteins (Gumbart et al. (2013b,a); Suh et al. (2019)), whether it could affect the binding specificity of the different Dpr–DIP complexes is unclear. Conformational entropy, in the context of the MM/PBSA end-state framework, is typically estimated on the basis of a quasi-harmonic approximation analysis of the complex and the unbound proteins (Kollman et al. (2000); Swanson et al. (2004); Minh et al. (2005)). Such a limited treatment, especially regarding its ability to account for the rotameric states of side chains and appears unlikely to provide meaningful information to understand binding specificity. Given these limitations, the PB-VDW computational analysis was kept as simple as possible.

Figure 3.9 shows the PB-VDW binding free energies averaged over the MD trajectories compared with the binding free energy extracted from the SPR measurements. Such a framework combines conformational sampling from MD trajectories with continuum solvent model (Kollman et al. (2000); Swanson et al. (2004)). There is a relatively poor correlation between the calculated and experimental binding free energies. While there is some uncertainty in the measured dissociation constants, it is clear that such a simple continuum solvent approximation is too limited. Going beyond the ambitious goal of matching the calculated binding free energies and experimental affinity of all the possible complexes one-for-one, we sought to find out if the computational model has at least the ability to distinguish between the cognate and noncognate set of complexes, that is, the ability of the computational model to correctly determine if a complex should be identified as cognate versus noncognate. Figure 6 documents the performance of the PB-VDW approximation regarding the distinguishability between the two main classes of complexes. To ascertain the utility of the conformational sampling provided by the MD trajectories, we compare the results obtained by relying solely on the static structures generated by homology modeling (Figure 3.10A and B), and the

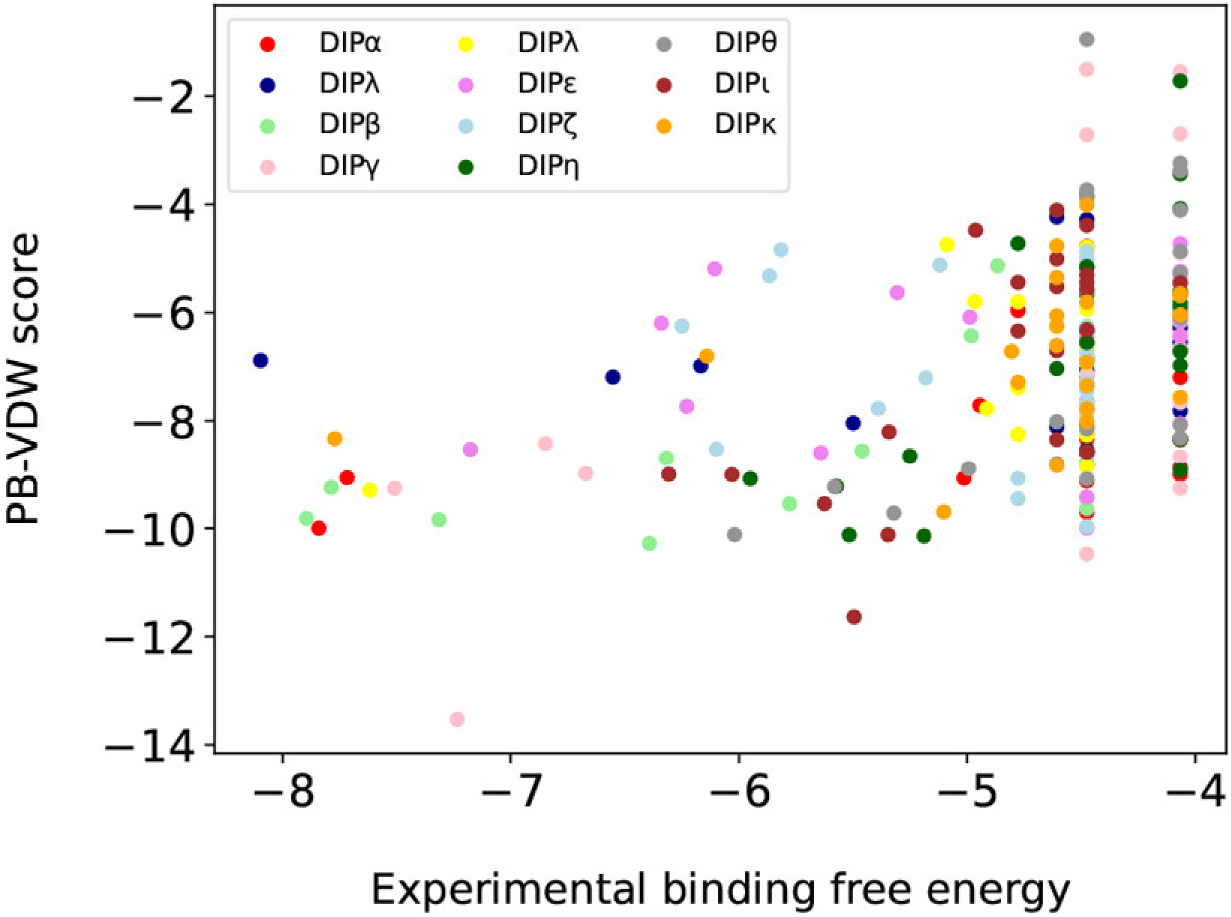


Figure 3.9: Comparison of the calculated binding free energy obtained from the continuum PB-VDW model averaged over MD trajectories with the corresponding binding free energy extracted from the SPR measurements for all 231 Dpr–DIP complexes. Different colors represent each of the 11 DIP protein bound to their 21 Dpr partners. The x- and y-axis are given in kilocalories per mole.

results obtained by averaging over MD trajectories (Figure 3.10C and D).

It is observed that the distribution of the binding free energies for the two classes have a significant overlap (Figure 3.10A and C). Although the performance is significantly improved by using the average from MD, this overlap severely undermines the distinguishability of the two classes by the computational model. The cumulative percentage of cognate and noncognate complexes (green and blue lines) indicates, however, that the cognate complexes are associated with a slightly stronger calculated binding free energy on average. The difference between the cumulative percentage of the cognate and noncognate complexes (red line in Figure 3.10B and D) indicates that the model reaches maximum distinguishability at

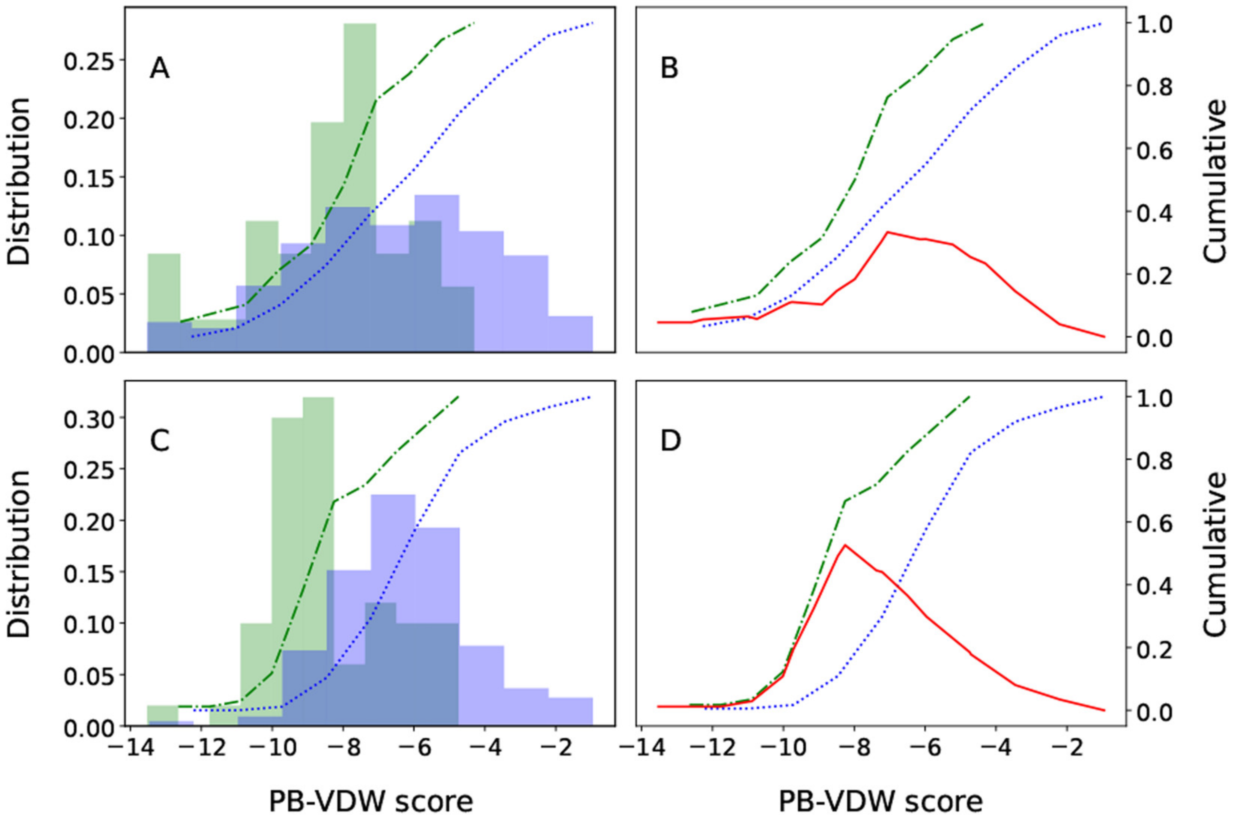


Figure 3.10: Sorting the cognate (green) and noncognate (violet) Dpr–DIP complexes using computed PB-VDW score. (A, B) Distribution of PB-VDW binding free energy for cognate and noncognate complexes calculated from the homology models. (C, D) Distribution of PB-VDW binding free energy for cognate and noncognate complexes averaged over the MD trajectories. (A, C) Normalized histogram and cumulative distribution of PB-VDW score. The green-shaded histogram represent the PB-VDW score of the cognate complexes. The violet-shaded histogram represent the PB-VDW score of the noncognate complexes. The green and blue lines represent the cumulative percentage of cognate and noncognate complexes, respectively. The red line (B, D) represents the difference between the running percentage of cognate and noncognate complexes, corresponds to the distinguishability index as a function of a binding free energy threshold. The x-axis is in kilocalories per mole.

some intermediate energy threshold.

Relying solely on the static structures generated by homology modeling for the binding free energy calculations, the maximum distinguishability between cognate and noncognate complexes, defined by the peak of the red curve, is approximately 30 % (Figure 3.10B). However, the performance is significantly improved by averaging the binding free energy from the continuum model over MD trajectories of the complexes in explicit solvent. The mean square deviation of the binding free energy computed using PB-VDW model for cognate complexes is around 1.2 kcal/mol, and for noncognate complexes, it is around 1.6 kcal/mol. As shown in Figure 6D, the maximum distinguishability between cognate and noncognate set of Dpr–DIP complexes reaches approximately 52 % (defined by the peak of the red curve). Therefore, relying on an ensemble of conformations from MD rather than a single conformation provides a larger degree of distinguishability between the cognate and noncognate complexes. For the six experimental crystal structures used as templates in generating homology models of Dpr–DIP complexes, we computed the binding free energy using continuum solvent PB-VDW model as well. We find reasonable agreement between the computational estimate and experimental value of binding free energy. Specifically, for Dpr6–DIPα, the values are -9.1 kcal/mol (PB-VDW) and -7.7 kcal/mol (experiment); for Dpr4–DIPη, we obtain -9.2 kcal/mol (PB-VDW) and -5.5 kcal/mol (experiment); for Dpr2–DIPθ, we obtain -9.2 kcal/mol (PB-VDW) and -5.8 kcal/mol (experiment); for Dpr1–DIPη, we obtain -10.1 kcal/mol (PB-VDW) and -6.5 kcal/mol (experiment); for Dpr10–DIPα, we obtain -9.9 kcal/mol (PB-VDW) and -7.8 kcal/mol (experiment); for Dpr11–DIPγ, we obtain -8.4 kcal/mol (PB-VDW) and -6.9 kcal/mol (experiment). As a comparison, the PB-VDW binding free energy obtained for homology models of the available experimental structures are in good agreement with the PB-VDW binding free energy of the experimental structures themselves. This seems to suggest that the accurate generation of homology models using proper templates, and the input sequence alignment is at the crux of obtaining reasonable binding free energy using the various computational approaches presented in this work, within the limits of expected accuracy of the various approaches. Averaging the binding free energy is most certainly one factor responsible for the improved performance. In addition, the improvement may also be attributed, in part, to the refinement of the initial structural homology models. For completeness, we also compared the scored PB-VDW values of the final frame from the simulated trajectories with the experimentally reported binding free energy for cognate complexes with available crystal structures. The agreement of time-averaged PB-VDW binding free energy with experimentally reported values are somewhat better than considering only the final snapshot from the MD trajectories.

While the analysis displayed in Figure 3.10 is informative, it is based upon a nonrigorous global classification of the Dpr–DIP complexes from SPR data declaring that all complexes with detectable binding (binding free energy better than -4.8 kcal/mol) are “cognate”, while the others are “non-cognate”. Which of the possible 231 Dpr–DIP pairs ought to be considered as putative cognate complexes in terms of their biological function is unclear. The binding cutoff to ascribe an interacting Dpr–DIP complex as “cognate” and a noninteracting Dpr–DIP complex as “non-cognate” based on the SPR measurement is, to some extent, arbitrary. Perhaps, the threshold needed to give rise to a true biological response must be higher than the weakest complexes detected by SPR measurements. Thus, different binding free energy threshold values could be used to classify the set of complexes based on the experimental SPR. From this perspective, we re-examined the performance of the computational model to see if different binding free energy threshold could also affect the distinguishability of the global set of Dpr–DIP complexes. As an illustration, Figure 3.11 shows the results using five different binding free energy thresholds to classify the cognate and noncognate pairs from the data extracted from the SPR measurements. On the basis of this comparison, it appears that the best distinguishability is realized with a threshold of -7.0 kcal/mol, although the change compared to Figure 3.10 is modest.

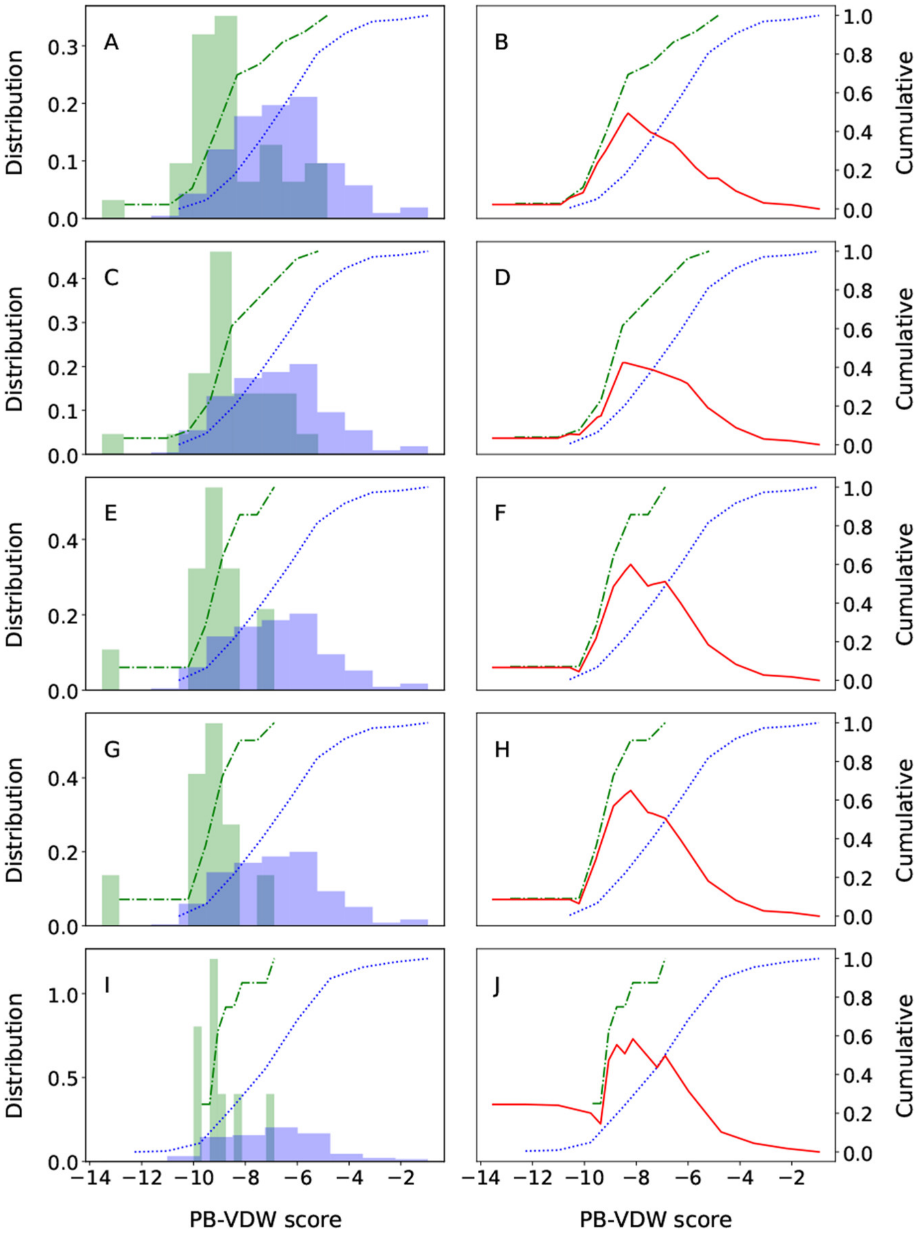


Figure 3.11: Impact of the binding free energy threshold on sorting criteria. The distribution of the PB-VDW binding free energy for cognate and noncognate complexes (left) and the cumulative distribution of scored binding free energy (right) are shown. The green-shaded histograms represent the binding free energy distribution of the cognate complexes. The blue-shaded histograms represent the binding free energy distribution of the noncognate complexes. As in Figure 3.10, the green and blue lines represent the cumulative percentage of cognate and noncognate complexes, respectively. while the red line (right panels) represents the difference between the running percentage of cognate and noncognate complexes, corresponding to the distinguishability index as a function of a binding free energy threshold. Distribution of scored binding free energy for global set of cognate (green) and noncognate (violet) set of Dpr–DIP complexes for threshold values of experimental binding free energy -5.5 (A, B), -6.0 (C, D), -6.5 (E, F), -7.0 (G, H), and -7.5 (I, J) kcal/mol (in sequential order from top row to bottom row) to define cognate complexes. The x-axis is in kilocalories per mole.

3.4.2 LDA Treatment of the PB-VDW Continuum Model

The maximum distinguishability from the PB-VDW continuum approximation averaged over MD trajectories is on the order of about 50 % at best, that is, we can classify about 50 % of the complexes into cognate and noncognate groups. It is important to understand that such an approximate treatment relies on an end-state approximation (Swanson et al. (2004)), rather than a more rigorous ensemble of all intermediates states along the dissociation of the protein complex (Gumbart et al. (2013b,a); Suh et al. (2019)). For example, issues of conformational entropy are only accounted for indirectly with such an end-state approximation. In this context, rather than a single number, it may be useful to retain a small number of representative computational features in the analysis in trying to classify the different Dpr–DIP complexes. In particular, one could consider simultaneously the information associated with averages, as well as the fluctuations of the individual PB and VDW contributions to the binding free energy. A standard approach to achieve maximum discrimination of the two main classes of complexes is to seek a linear projection of the multidimensional data onto an hyperplane via a linear discriminant analysis (LDA) (Barker and Rayens (2003)). Here, we consider five features that are strong markers of the overall binding free energy: the PB electrostatic contribution (ΔG_{PB}) and its variance, the nonpolar contribution (ΔG_{VDW}), corresponding to the scaled protein–protein VDW interaction and its variance, and finally, the variance of the total binding free energy (ΔG_b). For all 231 Dpr–DIP complexes, we generated a data set of these five features from the 200 ns MD trajectories. LDA was then carried on the combined data set to classify the cognate and noncognate complexes. The resulting weights of the five features shown in Table 3.2. It is interesting to note that, with the weight of 0.90 for ΔG_{VDW}, the overall scaling factor of the VDW protein–protein interactions becomes $0.17 \times 0.90 = 0.15$, which is slightly closer to the 0.161 value of the scaling factor in LIE (Åqvist et al. (1994); Aqvist and Marelius (2001)). On the other hand, the weight of 0.57 on ΔG_{PB} suggests that the pure electrostatic contributions needs to

LDA feature	weight
ΔG_{PB}	0.57
var (ΔG_{PB})	0.22
ΔG_{VDW}	0.90
var(ΔG_{VDW})	0.98
var(ΔG_{PB}+ΔG_{VDW})	0.10

Table 3.2: Linear Discriminant Analysis

be scaled down. Of additional interest is the weights on the variance of the PB and VDW contributions, with values near 1.0, indicating that this is important to indirectly incorporate the effect of conformational fluctuations.

The performance of PB-VDW-LDA is illustrated in Figure 3.12 by showing the cumulative and difference histogram of the linear Discriminant for the set of 231 cognate and noncognate Dpr–DIP complexes. A threshold of -4.8 kcal/mol was used for the binding free energies extracted from SPR data. Utilizing only the average binding free energy gives a maximum distinguishability between cognate and noncognate classes to approximately 52 %. This result is very similar to that obtained previous for the PB-VDW scoring (Figure 6D). Incorporating the global RMSD and its variance did not improve the results, possibly because this information is redundant with the variance in the PB and VDW contributions. On the basis of this criterion, the PB-VDW-LDA model is apparently not improved with regard to the maximum distinguishability between cognate and noncognate set of Dpr–DIP complexes. However, we will see below that the PB-VDW-LDA performs slightly better to predict the effect of point mutations.

3.4.3 LDA Treatment of the AIMS Interaction Scores

Given the ability of the AIMS analysis to identify key interacting residues in the Dpr–DIP interface, we reasoned that combining these pairwise interaction scores with the linear discriminant analysis may improve distinguishability between cognate and noncognate receptors. The interacting residues identified in Figure 3.6 were input in the LDA algorithm,

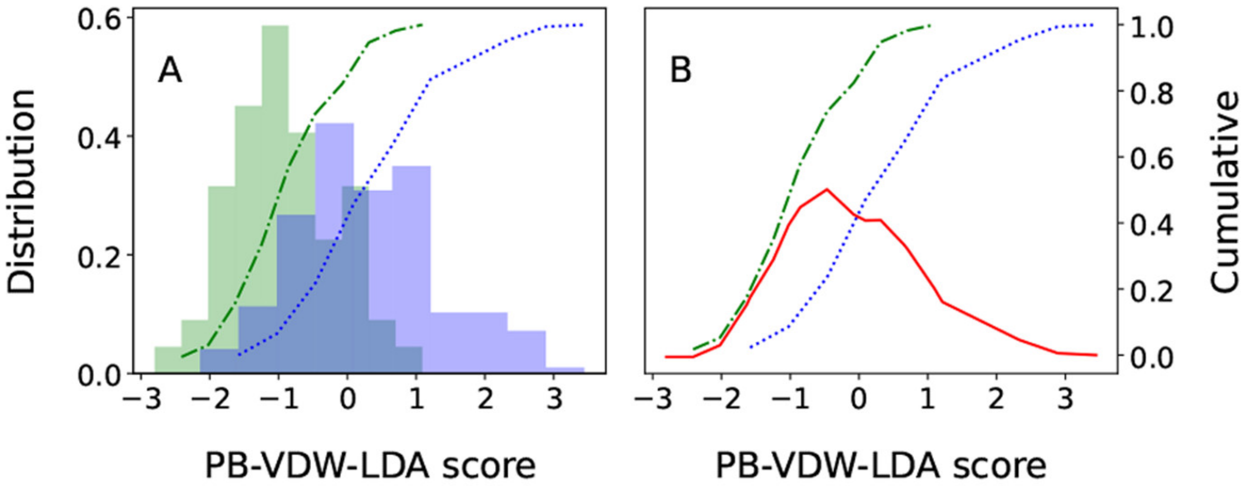


Figure 3.12: Scoring and distinguishability between cognate and noncognate complexes comprising a five feature set LDA model calculated using the PB-VDW continuum solvation model for cognate (green) and noncognate (violet) complexes. (A) Normalized distribution and (B) cumulative distribution of linear discriminant scores to classify the set of cognate and noncognate Dpr–DIP complexes with a threshold applied on the data at about -4.8 kcal/mol, estimated from experimental values of reported binding free energy. The x-axis is given in the dimensionless units of the PB-VDW-LDA score.

and the resultant classifier is able to differentiate cognate and noncognate receptors with an accuracy that rivals the LDA results of the previous section. As discussed in the Sequence Alignment and Homology Modeling section, however, it appears that cognate Dpr–DIP pairs rely on a mosaic of interactions to form productive dimers. Because of the lack of a clear set of residues necessary to generate cognate interactions, we inferred that the ability of LDA to differentiate cognate Dpr–DIP pairs using the AIMS outputs could be further improved by including all residues in the interface.

In the Sequence Alignment and Homology Modeling section, only those residue pairs with a nonzero mutual information were included in the analysis. In expanding the LDA input data set to include the full list of interacting residues, we include fully conserved residues in Dpr and DIP interfaces. In this case the entropy $H_i(x)$ is zero at the conserved site, leading to a mutual information of 0. However, such sites can still act as key components contributing to the delineation between cognate and noncognate pairs. The assumption is that complementary interacting residues on the binding partner may exhibit some diversity

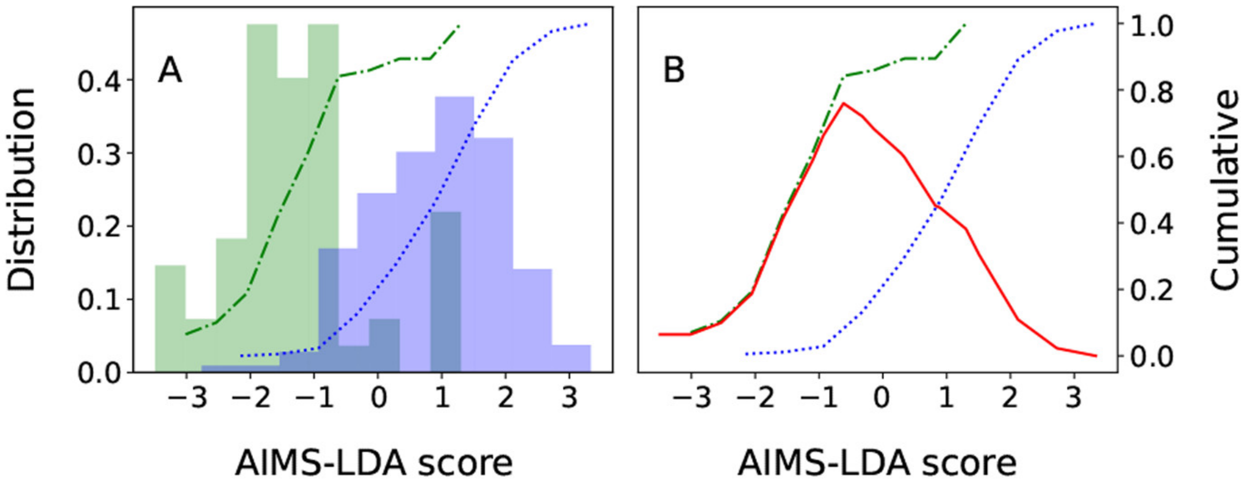


Figure 3.13: Scoring and distinguishability for cognate (green) and noncognate (violet) complexes derived from AIMS-LDA model calculated from residue interaction scores for cognate and noncognate Dpr–DIP pairs. (A) Normalized distribution and (B) cumulative distribution of AIMS-LDA score for cognate and noncognate complexes.

across the pool of available receptors. In other words, assuming Dpr has some fully conserved residue at a given site, there may be multiple nonconserved residues in the DIP interface which interact constructively in cognate receptor pairs and destructively in noncognate receptor pairs. Indeed, as shown in Figure 3.13, we find that this expanded set of residue pairs greatly improves the classification accuracy of the LDA. We can confirm that we are not overfitting with this increased number of input vectors by treating this linear discriminant as a classic machine learning problem, splitting the data into test and training sets. We find that performance of test data sets left out of the training set performs similarly to the results from training on the full data set.

As discussed above, one of the great strengths of LDA is in its simplicity and interpretability. We can use the linear weights of the classifier to identify some of the key residues necessary for a meaningful Dpr–DIP interaction. Importantly, we again find that there exist no key gatekeeping residues that are the strong determinants of cognate and noncognate complexes. However, we do see that one of the key interacting residues, Dpr GLN113 (MSA position), is fully conserved in all Dpr molecules. This Dpr GLN113 is in proximity to DIP MSA position 43, which is either a LYS, GLN, or SER residue. While clearly this conserved residue will not discriminate cognate and noncognate pairs, the LYS–GLN or GLN–GLN hydrogen bonding may be able to compensate for disruptive interactions elsewhere. This demonstrates that the inclusion of residue pairs with a mutual information of 0 has the potential to improve distinguishability of cognate pairs through the identification of compensatory residue interactions.

3.4.4 FoldX Model

FoldX is a popular algorithm aimed at predicting changes in free energy used to score the structural stability of mutations in proteins (Schymkowitz et al. (2005); Buß et al. (2018)). As this approach was used in a previous study of Dpr–DIP complexes (Sergeeva et al. (2020)), we tried to assess if it could distinguish the cognate and noncognate pairs from the 231 possible Dpr–DIP complexes. FoldX is based on an empirical potential function that includes various terms for polar and hydrophobic desolvation, as well as free energy change at protein interfaces of protein complexes. The results are shown in Figure 3.14. The calculated binding free energy distribution using FoldX does not separate the set of cognate and noncognate complexes accurately. Cognate and noncognate complexes can be misclassified because the scoring from FoldX shows a large degree of overlap. Using the classification of binding/nonbinding interaction pairs for the global set of Dpr–DIP complexes from reported SPR data or using different binding free energy thresholds did not improve the results.

3.4.5 Analysis of Point Mutations

The effect of point mutations on the binding affinity is an important experimental tool to finely dissect the molecular determinants of binding specificity. Arguably, alchemical free energy perturbation combined with all-atom MD simulations (FEP/MD) is the most

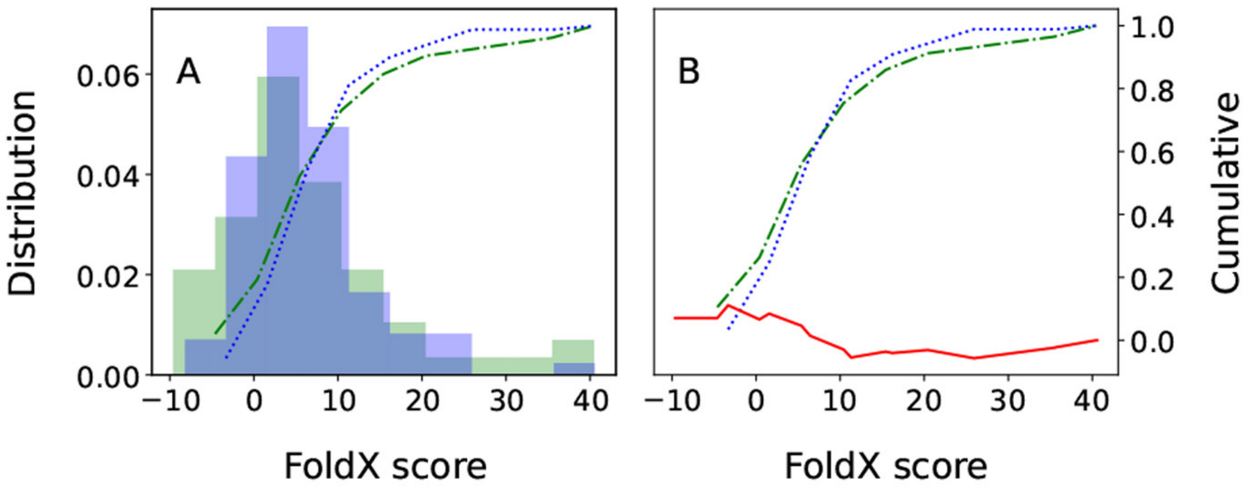


Figure 3.14: Scoring and distinguishability of cognate (green) and noncognate (violet) complexes with FoldX model. (A) Normalized distribution, and (B) Cumulative distribution of FoldX scores to classify the set of cognate and noncognate Dpr–DIP complexes with a binding free energy threshold applied on the data at -4.8 kcal/mol.

powerful computational approach to quantitatively ascertain the effect of mutations on the thermodynamic stability of molecular complexes (Chipot and Pohorille (2007)). We first examined the effect of alanine substitutions in the crystallographic Dpr6–DIPα complex (PDB id 5EO9) (Carrillo et al. (2015)). Using alchemical FEP/MD, we calculated the change in binding free energy, $\Delta\Delta G_b$ caused by an alanine substitution. We include the free energy difference predicted by the approximate and computationally inexpensive models, PB-VDW continuum solvent, PB-VDW-LDA, FoldX, and AIMS-LDA. To quantify the performance of the different approaches, the correlation coefficients between the computations and experimental values were calculated. The results are reported in Table 3.3.

With correlation coefficient of 0.14 and 0.10, respectively, FoldX and AIMS-LDA yield the worst performance. By comparison, the PB-VDW-LDA and the PB-VDW models perform moderately well for several of the mutants, with a correlation coefficient of 0.46 and 0.18, respectively. The agreement observed for the PB-VDW-LDA model is slightly better, suggesting that weighting multiple features improves the performance of the model although it does not improve the performance with respect to distinguishability as shown

Table 3.3: Effect of alanine substitution on the cognate Dpr6-DIPα complex

Mutation	$\Delta\Delta G^{a}_{\text{Expt}}$	$\Delta\Delta G_{\text{FEP}}$	$\Delta\Delta G^{b}_{\text{PB}}$	$\Delta\Delta G^{b}_{\text{PB}}$	$\Delta\Delta G_{\text{FoldX}}$	$\Delta\Delta G_{\text{AIMS}}$
			$-\text{VDW}$	$-VDW-LDA$		$-LDA$
Dpr6-H114A	-1	-1.6 ± 1.2	0.55	0.9	1.2	-0.6
Dpr6-I115A	2	2.9 ± 0.5	0.42	1.25	1.5	0.0
Dpr6-Y123A	4.5	4.2 ± 0.6	1.2	3.1	1.8	0.7
Dpr6-Q158A	3	0.8 ± 0.7	0.6	1.1	1.3	9.2
DIPαI83A	4	4.2 ± 0.7	0.4	2.9	1.7	0.0
DIPα-I91A	5	2.6 ± 0.5	0.5	2.3	2.1	0.0
DIPα-Q125A	6	2.3 ± 1.5	0.9	1.9	1.1	5.6
correlation coeff.	1.0	0.52	0.18	0.46	0.14	0.10

[a] Cheng et al. (2019) , [b] The fluctuation of computationally calculated binding free energy is within the range 0.8 - 1.2 kcal/mol. Correlation coefficient is calculated for ΔΔG obtained from each method in reference to the $\Delta\Delta G_{\text{Expt}}$.

above. The alchemical FEP/MD calculations appear to be the most reliable approach, with a correlation coefficient of 0.52 with respect to experiment.

For the sake of completeness, we also considered the impact of alanine substitutions at several positions in the noncognate (nonbinding) complex Dpr6–DIPγ as indicated by SPR data. Alanine-scan is a widely used experimental strategy to determine the importance of key residues in a protein–protein association (Kortemme et al. (2004)). We calculated the changes in binding free energy $\Delta\Delta G_b$ due to alanine substitution using alchemical FEP/MD. The results are given in Table 3.4. Again, we include the free energy changes predicted by the more approximate treatments, PB-VDW, PB-VDW-LDA, FoldX, and AIMS-LDA scores. With a correlation coefficient of 0.9, the PB-VDW-LDA is the approximate method that appears to best-match the $\Delta\Delta G_b$ determined from alchemical FEP/MD.

Lastly, we also determined the impact of 22 point mutations previously reported by Sergeeva et al. (Sergeeva et al. (2020)) using the computationally inexpensive approaches (PB-VDW, PB-VDW-LDA, FoldX, AIMS-LDA). Simple visual examination indicate that PB-VDW-LDA offers the best performance. This is confirmed from the correlation coefficients of computationally calculated ΔΔG with the reported experimental values of binding free energy. The model PB-VDW-LDA yields a correlation coefficient of 0.68, while the model

Table 3.4: Effect of alanine substitution on the non-cognate Dpr6-DIPγ complex

Mutation	$\Delta\Delta G_{\mathrm{FEP}}$	$\Delta\Delta G^{a}_{\mathrm{PB-VDW}}$	$\Delta\Delta G^{a}_{\mathrm{PB-VDW-LDA}}$	$\Delta\Delta G_{\mathrm{FoldX}}$	$\Delta\Delta G_{\mathrm{AIMS-LDA}}$
Dpr6-H114A	-0.2 ± 0.6	0.4	0.9	1.4	-0.6
Dpr6-I115A	0.0 ± 0.5	0.2	0.7	1.2	0.0
Dpr6-Y123A	2.1 ± 0.9	0.9	1.7	2.6	0.7
Dpr6-Q158A	4.8 ± 0.7	0.4	3.1	0.7	5.4
DIPγ-T80A	-0.6 ± 0.8	0.3	0.5	1.1	0.0
DIPγ-V81A	2.0 ± 0.5	0.2	1.4	1.6	0.0
DIPγ-V89A	1.5 ± 0.9	0.1	0.9	1.3	0.0
DIPγ-Q123A	1.7 ± 0.9	0.2	1.1	0.9	5.6
correlation coeff.	1.0	0.08	0.90	0.01	0.48

[a] The fluctuation of computationally calculated binding free energy is within the range 0.8 - 1.2 kcal/mol. Correlation coefficient is calculated for$\Delta\Delta$G obtained from each method in reference to the $\Delta\Delta G_{\mathrm{FEP}}$.

PB-VDW yields a correlation coefficient of 0.52. The FoldX model offers reasonable performance with a correlation coefficient of 0.31, while AIMS-LDA come last, with a correlation coefficient of about 0.04. While far from perfect, the PB-VDW-LDA model appears to provides useful semiquantitative agreement of the impact of mutations in these complexes based on the results from Tables 3.4 and 3.5.

3.5 Concluding Discussion

The Dpr–DIP interactome provides a rich testing ground for the study of protein–protein interactions. The extensive characterization of each pair of homologous proteins in this interactome, both via measurements of binding affinity and phenotypic responses, provides baseline definitions of what constitutes an interacting or noninteracting pair. The most important question revolves around the identification, in a biological context, of the defining molecular features that discriminate between cognate and noncognate receptors for these homologous proteins. Coevolution sequence analysis has proven to be an effective inference method for the prediction of possible protein complexes (Guo et al. (2008); Cheng et al. (2014); Bai et al. (2016); Morcos and Onuchic (2019)). However, in the case of the Dpr–DIP interactome, the aforementioned homology across the Dpr and DIP molecules confounds

Table 3.5: Effect of Point Mutations in Dpr6–DIPα, Dpr4–DIPη, and Dpr10–DIPα Complexes

protein	residue mutated	$\Delta\Delta G^{a}_{\mathrm{Expt}}$	$\Delta\Delta G_{\mathrm{FoldX}}$	$\Delta\Delta G^{b}_{\mathrm{PB}}$	$\Delta\Delta G^{b}_{\mathrm{PB}}$	$\Delta\Delta G_{\mathrm{AIMS}}$
				−VDW	−VDW−LDA	−LDA
Dpr6-DIPα	Dpr6H11K	1.91	0.46	0.62	1.44	-2.3
	DIPαK81Q	1.31	1.06	0.73	1.87	63.6
	DIPαS113D	0.21	0.52	1.1	1.55	0.7
	DIPαG74S	1.01	0.86	0.42	2.05	0.0
	DIPαA82T	0.85	1.90	0.7	1.6	0.0
	DIPαN94D	-0.45	0.12	0.3	0.9	0.0
	DIPαG74A	-.046	-0.28	0.14	0.55	0.0
	DIPαG74L	1.35	6.72	0.86	1.3	0.0
	DIPαA78K	-0.95	-0.26	-0.15	-0.82	-4.0
	DIPαI91A	2.18	1.79	0.5	2.3	0.0
Dpr4-DIPη	Dpr4K82H	0.12	0.12	0.3	0.95	0.0
Dpr10-DIPα	Dpr10Q138D	1.36	3.82	1.4	2.3	-9.3
	DIPαK81Q	1.85	1.49	0.8	1.7	-3.6
	DIPαD129S	0.16	-0.85	0.4	0.87	-1.8
	DIPαG74S	0.65	-1.07	1.3	1.65	0.0
	DIPαS133D	0.32	0.26	1.2	1.8	0.7
	DIPαI91A	2.11	1.27	0.75	1.72	0.0
	DIPαG74L	1.88	6.55	1.26	2.1	0.0
	DIPαA82T	1.33	0.87	0.9	1.55	0.0
	DIPαN94D	-.013	0.11	0.35	0.77	0.0
	DIPαG74A	-0.42	-0.25	0.1	0.3	0.0
	DIPαA78K	-1.09	-0.47	-0.22	-0.68	-4.0
Corr. coeff.		1.0	0.31	0.52	0.68	0.04

[a] All experimental values for the mutations shown are are taken from Sergeeva et al. (2020). [b] The fluctuation of computationally calculated binding free energy is within the range 0.8 - 1.2 kcal/mol. Correlation coefficient is calculated for∆∆G obtained from each method in reference to the $\Delta\Delta G_{\mathrm{Expt}}$.

such analysis, necessitating the application of a range of physics-based approaches for deconvolution of the determinants of productive receptor pairs.

Sequence analysis reveals the nuanced amino acid conservation inherent to the Dpr–DIP interactome (Figure 3.5), whereby the defining structural features are very well conserved, and the majority of sequence variability is concentrated in the protein–protein interaction interfaces. The situation is analogous to that found in the exceptionally diverse components of adaptive immunity: antibodies, T cell receptors, and major histocompatibility complexes. These immune molecules are likewise nearly structurally identical at the backbone level, save for regions of variability localized at their protein–protein interaction interfaces (Yin et al. (2012); Gras et al. (2012); Birnbaum et al. (2012)). Leveraging the molecular similarities of the Dpr–DIP interactome and adaptive immune receptors we reconfigured the Automated Immune Molecule Separator (AIMS) software, a software originally developed for the biophysical characterization of these immune molecules (Boughter et al. (2020)), to identify the key determinants of binding specificity in Dpr–DIP binding partners. AIMS provides a means to go beyond a simple genomic sequence analysis and help make the analysis more quantitative.

Specifically, we used AIMS to identify regions of high mutual information between cognate Dpr–DIP residue pairs. Coupling this mutual information with structural information from the Dpr10–DIPα crystal structure, we isolated key regions where residues are in close proximity and likely evolutionarily coupled. Through the creation of a new simple scoring scheme based on the fundamentals of amino acid interactions, we scored each receptor pair based on these identified regions (Figure 3.6). This new AIMS scoring metric shows that the basis for specificity appears to be somewhat diffuse, with no well-identified residues responsible for the cognate and noncognate complexes. A number of residues are identified as being partly responsible for the binding specificity, although the complex network of cognate complexes cannot be explained through the lens of a simple approach by identifying the interaction between a few pairs of residues.

The high structural similarity makes it possible to generate stable MD simulations for all 231 possible complexes (Figure 3.7). Generating atomic models of all possible complexes based on the sequence similarity is fairly straightforward. Furthermore, such initial homology models could be improved using MD simulations with explicit solvent. However, this information appears to be insufficient to identify the cognate and noncognate protein pairs among the family members. What is needed is the binding free energy of all the possible complexes. While there are computational strategies to calculate the absolute binding free energy of proteins from MD simulations with explicit solvent molecules (Gumbart et al. (2013a); Suh et al. (2019); Woo and Roux (2005)), these approaches are too computationally demanding to consider all 231 possible complexes. For this reason, it is necessary to use more approximate methods relying on a continuum solvent representation, such as PB-VDW. Averaging over an ensemble of configuration generated by MD simulation improves the results, possibly because this accounts for protein dynamics, which considers an ensemble of pairs of interacting residues at the interface of any given Dpr–DIP complex. However, the direct correlation between binding free energy estimated from a continuum solvent approximation (PB-VDW) and values extracted from SPR measurements remains too small to provide a quantitatively accurate ranking and scoring of all possible complexes (Figure 3.9).

A somewhat less ambitious goal is to try to predict the cognate and noncognate Dpr–DIP pairs of the interactome, that is, the distinguishability of the family members without attempting to quantitatively predict all the binding free energies of all complexes. Using averages from the MD trajectories improves the PB-VDW results from the simple homology models, but disappointingly, the PB-VDW model (Figure 3.10) and LDA-weighted PB-VDW model (Figure 3.12) continuum solvent approximations only lead to a maximum distinguishability of nearly 0.5, which is only modestly successful. The performance of FoldX is even worse (Figure 3.14). Remarkably, a simple interaction scoring metric of

LDA-weighted AIMS scoring interactions was sufficient to classify cognate and noncognate Dpr–DIP complexes, yielding a distinguishability of nearly 0.8 (Figure 3.13). It is likely that such a scheme could successfully decipher the protein encoding of similar Dpr–DIP interactomes from other organisms. However, the same LDA-weighted AIMS scoring scheme performed poorly when utilized to predict the effects of mutations to cognate receptor pairs (Tables 3.3, 3.4, 3.5). This is expected, as the application of LDA to the AIMS interaction scores takes into account only directly interacting residues, not those that may have some longer-range effects. Further, the first version of the scoring matrix disregards alanine, threonine, serine, and glycine as interacting residues, highlighting the substantial room for the improvement of this scoring through the inclusion of more physics-based rules for interactions. Along with this improvement to the scoring, explicit training on mutants may help to generalize the classifier, forcing the analysis to account for all residues in the interface, not just those that are key to determining cognate and noncognate pairs. According to Table 3.3, the predictions of the effects of mutations to cognate receptor pairs from alchemical FEP simulations are in best agreement with the experimental values. Approximate continuum solvent approaches, such as PB-VDW and, especially, PB-VDW-LDA, provide an acceptable accuracy, almost equivalent to alchemical FEP. On the other hand, the accuracy of FoldX in predicting the effect of mutations appears to be limited.

Protein–protein interactions and binding specificity reflect biological and biophysical components. Making progress to understand these complex systems requires a multiprone approach to be most effective. The present efforts shows that knowledge-based and physics-based approaches are complementary, with different strengths and limitations. While biological aspects are often best revealed through knowledge-based approaches combining evolutionary and structural features, physics-based approaches are needed to pinpoint the molecular determinants controlling binding specificity. It is important to pick the correct computational strategy that fits the questions at hand. For example, alchemical FEP calculations based on MD simulations with explicit solvent molecules represents the most accurate method to predict the effect of mutations on the binding affinity. Furthermore, experimental SPR data have identified a broad group of 57 Dpr–DIP cognate complexes showing interaction specificity, while the remaining 174 pairs display no detectable binding affinity with reasonable confidence (Cosmanescu et al. (2018)). Computations based on atomic models may clarify why some of these proteins bind together but not others despite their high structural similarities, although practical computational methods remain too approximate to perfectly correlate with experiment.

In closing, it is important to put the present work in the context of the recent progress from neural network artificial intelligence approaches, including AlphaFold 2 (Jumper et al. (2021), RoseTTAFold (Baek et al. (2021)), as well as additional machine learning approaches (Li et al. (2022); Zhou et al. (2022). These approaches combine evolutionary and structural features for structural predictions. Remarkably, while the fact that the methods were primarily trained to predict single-protein structures, there has been also success in predicting the structure of complexes. In particular, RoseTTAFold and AlphaFold 2 derived protocols and models have achieved great strides in protein–protein docking (Jumper et al. (2021); Baek et al. (2021); Bryant et al. (2022)). These advances raise questions about the role of more traditional physics-based approaches. These methods rely heavily on coevolutionary information, which obscures a clear interpretation in terms of physical interactions. Furthermore, deep neural networks, with their many interconnected layers, do not allow an easy identification of the molecular determinants responsible for a specific physical outcome (Jiménez-Luna et al. (2020)). While the approaches outlined in this work do not explicitly predict complex structures, their interpretability at each step allows for the direct identification of the key physical components necessary for protein–protein interactions.

In this context, it is likely that physics-based methods will continue to help better understand the physical principles behind protein–protein interactions. A completely different study of protein–protein complexes is that of Pan et al., who relied on all-atom explicit solvent simulations generated with the specialized Anton 2 (Pan et al. (2019)). Importantly, the authors noted that backbone torsional restraints around the bound native structures were needed to prevent conformational degradation at the (hundreds of) microsecond time scales simulate to successfully observe reversible association. This observation on the sensitivity of protein complexes on the accuracy of the structures is consistent with recent results by Faruk et al. (Faruk et al. (2022)) In the long term, further progress in improving the accuracy of atomic force fields and the efficiency of molecular dynamics approaches will be important for studying the thermodynamics and kinetics of protein association, and the associated conformational changes. Ultimately, having the ability to predict specific protein–protein association computationally could open the door to useful dissection of cell signaling pathways or the design of novel protein-based materials. In this context, systems like the Dpr–DIP interactome offer a great opportunity to explore how different computational approaches can provide meaningful information about binding specificity and the overall structure of complex functional networks constructed from protein–protein interactions.

CHAPTER 4

NONEQUILIBRIUM MONTE–CARLO

The previous chapters presented applications of alchemical simulations for binding free energy simulations. In this chapter, alchemical simulations are used in a context nonequilibrium Molecular Dynamics - Monte-Carlo (neMD/MC) methodology to enhance the sampling of the configurations of inhomogeneous membranes. This chapter uses the paper Configurational Sampling of All-Atom Solvated Membranes Using Hybrid Nonequilibrium Molecular Dynamics Monte Carlo Simulations by **Florence Szczepaniak**, François Dehez and Benoit Roux, published in 2024 in *The Journal of Physics and Chemistry Letters*.

4.1 Development of the methodology

4.1.1 Algorithm

The neMD/MC algorithm was initially developed for constant-pH simulations (Stern (2007); Chen and Roux (2015a); Radak et al. (2017)). Protons are coupled and decoupled from the rest of the system with alchemical simulations. Resorting to Metropolis test after the simulation leads to the sampling of protonation states according to their Boltzmann weight. More generally, algorithms combining Monte-Carlo exchanges and nonequilibrium MD have different names and different applications (Nilmeier et al. (2011); Suh et al. (2018); Fathizadeh and Elber (2018)). In all cases, the structure of the algorithm is similar. This chapter develops an application of the neMD/MC method to exchange lipids in a membrane. The method is schematized in Figure 4.1. Starting from an equilibrated configuration of a system $\mathbf{X}_0$, two lipids **a** and **b** are randomly chosen to be exchanged. Then, the exchange of the two lipids is performed with an alchemical transformation. The work **W** associated with the exchange is computed with Thermodynamic Integration (TI) (Kirkwood (1935)), and then is used in a Metropolis test to know if the exchange leads to a favorable configuration.

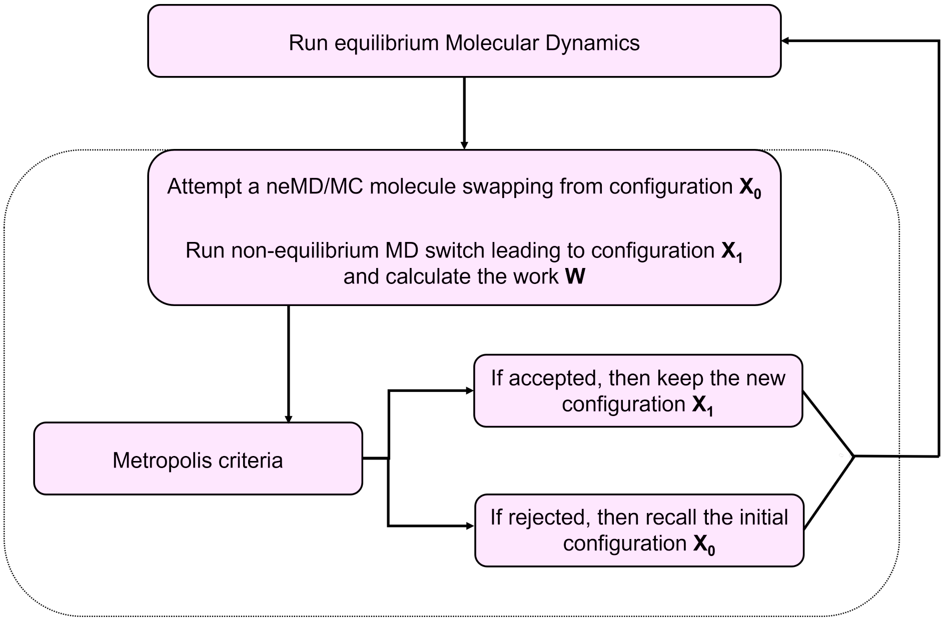

Run equilibrium Molecular Dynamics
Attempt a neMD/MC molecule swapping from configuration **X₀**
Run non-equilibrium MD switch leading to configuration **X₁** and calculate the work **W**
Metropolis criteria
If accepted, then keep the new configuration **X₁**
If rejected, then recall the initial configuration **X₀**

Figure 4.1: Summary of the neMD/MC methodology

If the exchange is accepted, the new configuration of the system $\mathbf{X}_1$ is equilibrated with Molecular Dynamics, before doing another attempt. If the exchange is rejected, the previous configuration $\mathbf{X}_0$ is equilibrated in MD, and another exchange is attempted with other molecules.

Let us consider the equilibrium Boltzmann distribution:

$$P_{\mathrm{eq}}(\mathbf{x}) \propto \exp\Bigg(-U(\mathbf{x})/k_{\mathrm{B}}T\Bigg) \tag{4.1}$$

In the Metropolis Monte Carlo method, a random walk is built in the configuration space to reach the equilibrium distribution (Metropolis et al. (1953)). To converge to the correct Boltzmann distribution, the condition is:

$$P_{\mathrm{eq}}(\mathbf{x})T(\mathbf{x} \rightarrow \mathbf{x'}) = P_{\mathrm{eq}}(\mathbf{x'})T(\mathbf{x'} \rightarrow \mathbf{x}) \tag{4.2}$$

where $\mathbf{x}$ and $\mathbf{x'}$ are two configurations, and $T(\mathbf{x} \rightarrow \mathbf{x'})$ is the probability of the transition from $\mathbf{x}$ to $\mathbf{x'}$. This condition is called the condition of microscopic detailed balance. Another way of presenting it would be:

$$\frac{T(\mathbf{x} \rightarrow \mathbf{x'})}{T(\mathbf{x'} \rightarrow \mathbf{x})} = \frac{P_{\text{eq}}(\mathbf{x'})}{P_{\text{eq}}(\mathbf{x})} = \exp\left(-\frac{U(\mathbf{x'}) - U(\mathbf{x})}{k_{\text{B}}T}\right) \tag{4.3}$$

The transition probability T($\mathbf{x} \rightarrow \mathbf{x'}$) is usually seen as $\text{T}^{P}(\mathbf{x} \rightarrow \mathbf{x'})\text{T}^{a/d}(\mathbf{x} \rightarrow \mathbf{x'})$, with $\text{T}^{P}(\mathbf{x} \rightarrow \mathbf{x'})$ the probability to propose the move from **x** to **x'**, and $\text{T}^{a/d}(\mathbf{x} \rightarrow \mathbf{x'})$ the probability to accept or deny the exchange. To simplify the equations, only perform reversible transformations can be run and the distribution of **x'** around **x** can be set up to be symmetric. Then, $\text{T}^{P}(\mathbf{x} \rightarrow \mathbf{x'}) = \text{T}^{P}(\mathbf{x'} \rightarrow \mathbf{x})$ (Chen and Roux (2015a); Roux (2021)). The notation **x** (or **x'**) represents the positions and the velocities. So to ensure that we are performing a Monte-Carlo switch, and that $\text{T}^{P}(\mathbf{x} \rightarrow \mathbf{x'}) = \text{T}^{P}(\mathbf{x'} \rightarrow \mathbf{x})$, a momentum reversal prescription needs to be applied. To conserve the equilibrium distribution, the velocities need to be reversed (Nilmeier et al. (2011)), preventing a deterministic situation to conserve the equilibrium distribution an to ensure that the move $\mathbf{x'} \rightarrow \mathbf{x}$ is possible. Chen and Roux showed that a symmetric two-ends momentum reversal, so a reversal of the momentum of the system both at the beginning and at the end of the switch with a probability of 1/2 generates the correct results and respects the microscopic detailed balance while limiting any complication in the dynamics (Chen and Roux (2014)). Based of these equations and these conditions, the Metropolis test to know if the exchange is accepted or rejected is:

$$P_{\text{acc}} = \min\left[1, e^{-W/k_{\text{B}}T}\right] \tag{4.4}$$

To verify the accuracy of the theory, and to develop the methodology, a simpler system made of Lennard Jones Particles is studied first.

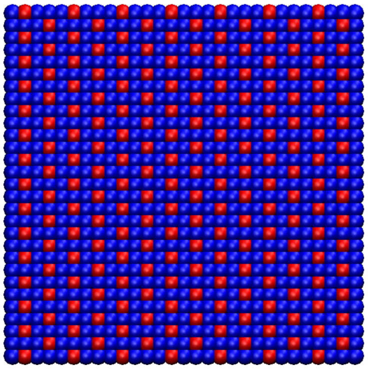


Figure 4.2: Initial configuration

4.1.2 Lennard-Jones Particles

The first application of the neMD/MC method to swap molecules is a planar distribution of Lennard-Jones particles. Two types of particles are introduced: blue and red. They all have the same radius, no charge and the same mass. The only difference lies in the Van der Waals parameters. Consider the formula for the Lennard–Jones interactions used in the CHARMM Force Fields (Brooks et al. (2009)):

$$V_{LJ} = \epsilon\Bigg((\frac{r_{min}}{r})^{12} - 2(\frac{r_{min}}{r})^{6} \Bigg) \tag{4.5}$$

the parameter ϵ has been set up such that: $\epsilon_{b-b} = \epsilon_{r-r} = -0.05$ kcal/mol $= 5 \times \epsilon_{b-r}$. The red and the blue want to regroup in two areas: one with only red particles, the other with only blue particles. The temperature is set at 1 K, the volume remains constant, and the particles are restraint in the plan. Then, a second test with the same system but at 300 K has been conducted, this time with $\epsilon_{b-b} = \epsilon_{r-r} = -1.05$ kcal/mol, and $\epsilon_{b-r} = -0.01$ kcal/mol. The ϵ was changed to be in a range of energy that is closer to k$_B$T ($\approx$0.59 kcal/mol at 300 K).

In Molecular Dynamics simulation at 1 K (Figure 4.3 left), there is almost no movements. With neMD/MC simulation, the system is sorted. Figure 4.3 (right) shows the configuration

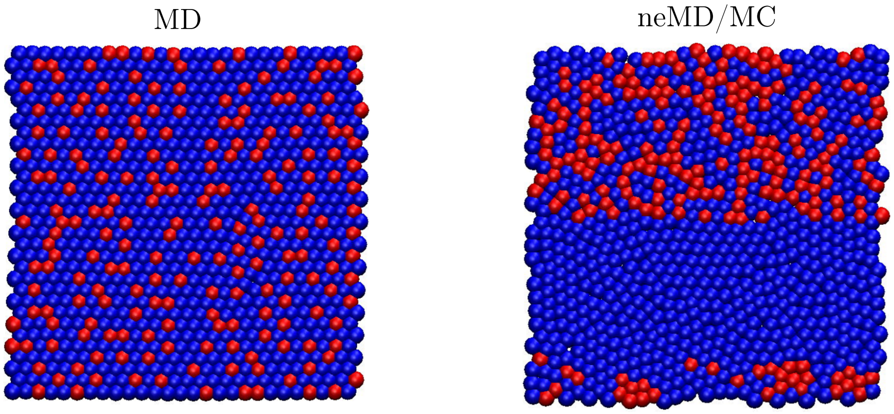


Figure 4.3: Lennard Jones particles after 1 ns of MD (left) and 5.7 ns of neMD/MC (right).

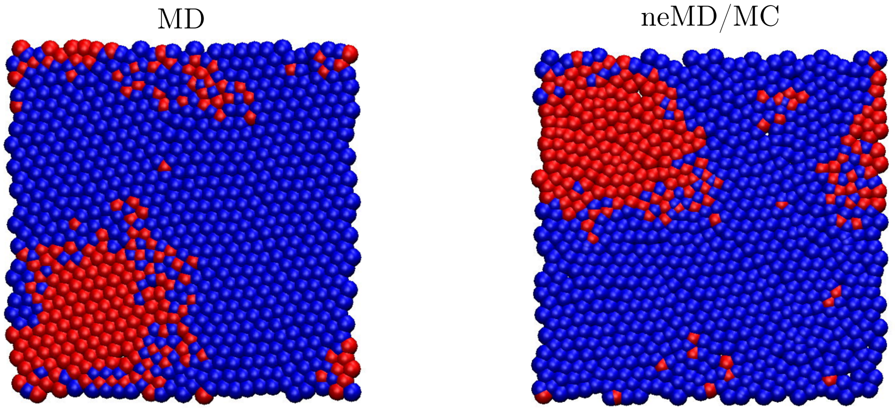


Figure 4.4: Lennard Jones particles after 1 ns of MD (left) and 627 ps of neMD/MC (right).

after 5.7 ns. About 20 % of the moves are accepted. The system is not completely sorted yet, but there is a separation between the red and the blue particles, and the system tends to create two bands, each with only one kind of particles.

At 300 K, the MD is much more efficient to sort the particles. The higher temperature allows more movements to sort the particles. Figure 4.4 shows the repartition of the particles after 1 ns of MD (left), and the configuration after 627 ps of neMD/MC (right), with about 20 % of the moves accepted.

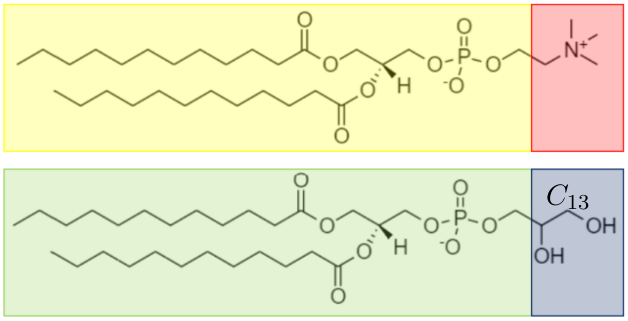


Figure 4.5: Structures of the DLPC (top) and DLPG (bottom). In yellow and green: common atoms, in red and blue: non common part.

4.1.3 *Polar heads in water*

A major difficulty when exchanging the lipids in a membrane is a sterical hindrance. The membrane being a dense phase, the reorganization of the environment around incoming atoms is slow. In order to lower the perturbation and the work associated with the exchange, it is helpful to have the incoming lipid in a conformation as close as possible to the one of the outgoing lipid. This problematic could not be explore with the 2D Lennard Jones system, so to simplify our approach of this problematic, I started with two lipids (one phosphocholine and one phosphoglycerol) in water. These two polar heads are solvated in water. The mean value of the work associated with the exchange of these two molecules should be zero, as the system remains the same, and the environment of the polar heads is the solvant, so easy to reorganize and without specific interactions. The goal of this step is to find a path to do exchanges with a work as close to zero as possible while having the incoming polar heads fitting the conformation of the outgoing polar heads. For creating the copies of the lipids, the common atoms (in yellow and green on Figure 4.5) are aligned. For the non common part (in red and blue on Figure 4.5), an approximate alignment between the carbons and nitrogen of the DLPC and the carbons, oxygens and hydrogens of carbon C_{13} of DLPG is done. Positions of the common atoms are then restrained, and an extrabond between the incoming/outgoing nitrogen and the outgoing/incoming carbon C_{13} is added. The positional restraints are strong (force constant of 100 kcal.mol^{-1}.Å^{-1}) whereas the extrabond is weak (force constant of 0.5 kcal.mol^{-1}.Å^{-1}) to do hamper proper equilibration of the conformation of the polar heads.

4.2 neMD/MC on membranes

4.2.1 Introduction

Membranes are essential components of living organisms. They consist of amphiphilic lipid molecules organized in a bilayer hosting a variety of intrinsic membrane proteins involving a wide range of biological functions (Watson (2015)). The lipidome of living organism comprises a considerable number of chemically different lipids (Wallin and Heijne (1998)). They assemble in various proportions to give the biological membranes specific properties that vary according to the species, cells of a given organism or organelles (Dowhan (1997); Harayama and Riezman (2018)). It is now well-established that lipid composition affects not only the mechanical properties of membranes but also modulates the function of membrane proteins through distinct mechanisms (Laganowsky et al. (2014); Hénault et al. (2019)). Despite recent advances in experimental studies of lipid-protein interaction, including high-resolution structures capturing lipid binding sites, our understanding of the spatial organization of membranes around proteins remains limited (Corradi et al. (2019)).

Molecular simulations emerged as an attractive approach to provide atomic-level information regarding both the structure and dynamics of biological membranes. All-atom molecular dynamics (MD) are often employed to study the structure and dynamics of membrane proteins in simple homogeneous phospholipid bilayers. It is however clear that the accessible simulation timescales are not sufficient to allow a satisfactory sampling of inhomogeneous multi-component membranes where lipid diffusion is very slow (Rose et al. (2015); Muller et al. (2019)). For this reason, many simulation studies of lipid-protein association in complex mixtures are based on coarse-grained (CG) MD simulations, in which groups of atoms are reduced to a single effective particle (Ingólfsson et al. (2014); Marrink et al. (2007)). The speed-up of the calculations puts the study of complex membrane dynamics within reach, yet at the cost of an approximate representation of the molecular system (Corradi et al. (2019); Muller et al. (2019)).

Very few methods aimed at sampling the configuration space of inhomogeneous bilayer using all-atom simulations have been proposed in recent years. Some of them relies on generalized-ensemble algorithms such as replica-exchange molecular dynamics (REMD) (Huang and García (2014); Mori et al. (2013)). Replicas running at higher temperature allows for a faster diffusion of the lipids, resulting in an enhanced sampling of bilayer configurations. Another category consists in driving the system via short nonequilibrium MD (neMD) trajectories to generate a new state of the system that are subsequently accepted or rejected via a Metropolis MC step (Nilmeier et al. (2011); Chen and Roux (2014, 2015c); Radak and Roux (2016)). These hybrid simulation methods combining the advantages of MC with the strengths of MD offer promising strategies to efficiently sample the configurations of complex molecular systems such as membranes. The concept of the neMD/MC algorithm was first introduced in the context of constant-pH simulations (Stern (2007); Chen and Roux (2015a); Radak et al. (2017)) whereby the protons are coupled and decoupled alchemically from the rest of the system, leading to sampling of all protonation states according to their proper Boltzmann weight. More generally, the ability of neMD/MC simulations to sample equilibrium configurations provide an important tool for studying complex biomolecule systems (Chen et al. (2016); Chen and Roux (2015b); Suh et al. (2018); Gill et al. (2018)).

Kindt and coworkers reported the first example of a hybrid MD/MC simulation of a bilayer involving lipid mutations (de Joannis et al. (2006); Coppock and Kindt (2009); Kindt (2011)). In their work, mutation was limited to lipids differing by a very few atoms. Every accepted mutation modified the bilayer composition thereby, generating configurations in the grand canonical ensemble. More recently, Fathizadeh and Elber developed the MDAS algorithm (Molecular Dynamics with Alchemical Steps) aimed at sampling efficiently multicomponent fluids by means of MC exchange between pair of molecules selected randomly on the basis of alchemical work calculated from a nonequilibrium MD trajectory (Fathizadeh and Elber (2018)). They assessed the efficiency of MDAS by studying the mixing of a binary lipid system made of POPC and DOPC, two lipids differing by one unsaturation and a slightly longer acyl chain. The nonequilibrium work associated with the alchemical swapping was evaluated through a single-topology paradigm, involving a single copy of all the common atoms of DOPC and POPC supplemented by the atoms belonging solely to each of these two lipids. They assessed further the efficiency by studying the phase transition in a membrane composed of DPPC and DLPC, two lipids differing by the acyl chain length. They concluded that the efficiency of such neMD/MC hybrid approaches depends crucially on the mutation strategy (Fathizadeh et al. (2020)). The MDAS algorithm has been extended to sample mixing of coarse-grained (CG) lipid mixtures (Cherniavskyi et al. (2020)). Cherniavskyi et al. showed that the exchange rate for lipids carrying different charges was high in the context of the CG Martini force field. Yet, attempts exchanges of the same lipids described in all-atom models were essentially rejected. While the previous work with hybrid neMD/MC approaches provided great advances, there remains unresolved challenges in the context of all-atom simulations.

Here, we designed a general hybrid neMD/MC method aimed at sampling detailed atomic models of biological membranes. Specifically, we propose a strategy based on a dual-topology paradigm for swapping lipids that differ in their molecular structure. The method is illustrated and its performance critically evaluated in the context of a binary mixture of both zwitterionic and negatively charged lipids, DLPC and DLPG respectively. We also demonstrate the potential of our approach towards sampling the specific association of lipids to a transmembrane peptide bearing negative charges.

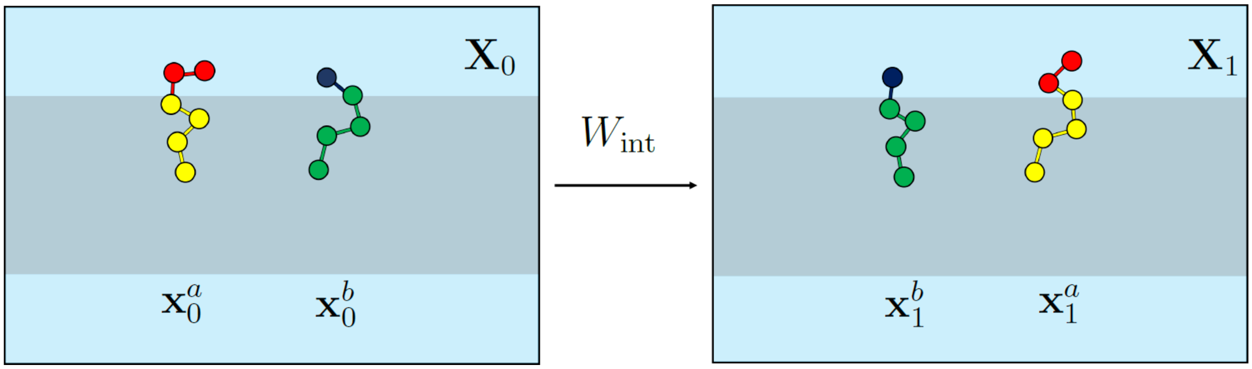


Figure 4.6: Schematic representation of a possible lipid exchange move. $\mathbf{x}_0^a$ and $\mathbf{x}_0^b$ are the coordinates and velocities of the lipids a and b at their initial sites and $\mathbf{x}_1^a$ and $\mathbf{x}_1^b$ in their final state after the exchange. $\mathbf{X}_o$ and $\mathbf{X}_1$ are the coordinates and velocities of the rest of the system. Red and blue spheres represent the non common atoms of the exchanging lipids, while yellow and green spheres stand for their common part. W_{int} is the nonequilibrium work associated to the lipid exchange.

4.2.2 Method

The objective is to sample configurations according to equilibrium Boltzmann statistics. The hybrid scheme is based on a series of equilibrium MD simulations interspersed with discrete stochastic transitions in which one randomly picks two chemically different lipid molecules, a and b and attempts to exchange their position according to a Monte Carlo (MC) algorithm. If the attempted exchange is rejected, another exchange is proposed. After a series of rejected exchanges, the equilibrium propagation can be resumed without disturbance, whereas the equilibrium propagation is continued from the new position if the attempted swap is accepted. The exchange process follows a hybrid neMD/MC scheme designed to allow configurational transition between two states (Stern (2007); Nilmeier et al. (2011); Chen and Roux (2014, 2015c)).

The goal of the exchange process is depicted schematically in Figure 4.6. Briefly, it is assumed that the two molecules have a core of identical chemical topology, for which atoms can be matched one-for-one. The coordinates of the molecules a and b are written as $\mathbf{r}^a = (\bar{\mathbf{r}}^a, \underline{\mathbf{r}}^a)$ and $\mathbf{r}^b = (\bar{\mathbf{r}}^b, \underline{\mathbf{r}}^b)$, respectively, where $\bar{\mathbf{r}}^a$ and $\bar{\mathbf{r}}^b$ represent the coordinates of the atoms of the identical core, while $\underline{\mathbf{r}}^a$ and $\underline{\mathbf{r}}^b$ represent the coordinates of the remaining "dangling" atoms. The number of dangling atoms in the two molecules is not necessarily the same. We denote the remaining coordinates of the surrounding environment as $\mathbf{R}$. Conceptually, the entire swapping process consists in decoupling molecules a and b from their surrounding environment, exchanging the atomic coordinates of their common core, generating proper equilibrium coordinates for their dangling atoms, and re-coupling the exchanged molecules to their surrounding environment. The nonequilibrium work, W_{int}, associated with the complete process is then used in a Metropolis MC to accept or reject the proposed exchange.

For the sake of clarity, we write the positions and velocities of molecule a, b and the surrounding environment as, $\mathbf{x}^a \equiv (\mathbf{r}^a, \mathbf{v}^a)$, $\mathbf{x}^b \equiv (\mathbf{r}^b, \mathbf{v}^b)$, and $\mathbf{X} \equiv (\mathbf{R}, \mathbf{V})$, respectively. The exchange transition begins at $(\mathbf{x}_0^a, \mathbf{x}_0^b, \mathbf{X}_0)$ and ends at $(\mathbf{x}_1^a, \mathbf{x}_1^b, \mathbf{x}_1)$. During the alchemical nonequilibrium simulation carried out within the dual topology framework of NAMD, the two original copies a and b are progressively switched off (selected as "outgoing" in the NAMD alchemical free energy syntax) as the coupling parameter λ varies from 0 to 1, while two new copies b' and a' are simultaneously switched on (selected as “incoming” in the alchemical free energy NAMD syntax). The complete transition $(\mathbf{x}_0^a, \mathbf{x}_0^b, \mathbf{x}_0 \rightarrow \mathbf{x}_1^a, \mathbf{x}_1^b, \mathbf{X}_1)$ is carried out in a step-wise fashion. Preliminary tests involving exchanges of lipids truncated to their polar head-groups in bulk water and exchanges of entire lipids in a membrane pointed out the need to resort to a dual topology strategy to reduce the statistical noise associated with the calculation of out-of-equilibrium work computed along a neMD/MC simulation. Employing the latter strategy becomes particularly crucial for estimating the work associated to exchanges in highly crowded molecular environment such as lipid bilayers.

The overall algorithm employed to evaluate the nonequilibrium work associated to a lipid exchange is illustrated in Figure 4.7. In a first step, a copy of molecules a and b is created in vacuum, referred to as a' and b'. This step aims at creating the copies of a and b in a conformation and position as close as possible to b and a, respectively, in order to limit the steric hindrance when inserted in the membrane. Therefore, the copies a' and b' have exchanged coordinates for the common core atoms, $\mathbf{r}_0^{a'} = (\overline{\mathbf{r}}_0^b, \underline{\mathbf{r}}_0^{a'})$ and $\mathbf{r}_0^{b'} = (\overline{\mathbf{r}}_0^a, \underline{\mathbf{r}}_0^{b'})$. The positions of the non-common dangling atoms of molecules a' and b' are initially constructed consistently with the position of the common core atoms in their new conformation, and are subsequently equilibrated via a simulation in vacuum. For this step in vacuum, all the atoms of molecules a and b as well as the common core atoms of a' and b' are fixed to their initial positions. An additional weak distance restraint between one dangling atom of a (or b) and one atom of b' (or a') is imposed to favor representative conformations in which the positions of the dangling atoms of molecules a' and b' overlap with those of the molecules b and a, respectively. The computationally inexpensive vacuum simulation is carried out as long as necessary to generate a conformation of a' and b' consistent with the equilibrium distribution of a dual topology system. Contributions arising from the dual topology framework are assumed to be negligible but could be taken into account if necessary.

In a second step, we calculate the nonequilibrium work W_{int} associated with exchanging the molecules a and b in the membrane environment. This is accomplished by alchemically switching off the nonbonded interactions of molecules a and b (outgoing) and switching on the non-bonded interactions of molecules a' and b' (incoming) as a function of λ during the alchemical free energy simulation. In terms of the coupling parameter λ, the outgoing energy $U_v(\mathbf{r}^a, \mathbf{r}^b\,\mathbf{R})$ is active when $\lambda = 0$, and the incoming part $U_v(\mathbf{r}^{a'}, \mathbf{r}^{b'}\,\mathbf{R})$ is fully active when $\lambda = 1$. The weak distance restraint operating between the outgoing and incoming atoms is unchanged during the transformation. The dual-topology potential energy $U_{int}(\mathbf{r}^a, \mathbf{r}^b, \mathbf{r}^{a'}, \mathbf{r}^{b'}, \mathbf{R}; \lambda) = \lambda U_v(\mathbf{r}^{a'}, \mathbf{r}^{b'}, \mathbf{R}) + (1-\lambda)U_v(\mathbf{r}^a, \mathbf{r}^b\,\mathbf{R})$ is constructed with the time-dependent coupling parameter $\lambda(t)$ to implement this transformation and calculate the work W_{int} as,

$$W_{int} = \int_0^{t_{int}} \left(\frac{\partial U_{int}}{\partial \lambda}\right) \dot{\lambda}(t)\, dt \tag{4.6}$$

The starting coordinates for this simulation are $\mathbf{r}_0^a$, $\mathbf{r}_0^b$, and $\mathbf{R}_0$ for molecules a, b and the environment, respectively. The initial velocities for a, b and the environment are the velocities at the end of the previous MD step. In addition, the coordinates $\mathbf{r}_i^{a'}$, and $\mathbf{r}_i^{b'}$, taken from the end of the previous vacuum simulation are used for molecules a' and b', respectively. The velocities of a' and b' (incoming) are randomly generated at the corresponding temperature. The cartesian coordinates of all the common core atoms (incoming and outgoing) are constrained during attempted lipid exchange. This choice, satisfies microscopic detailed-balance by construction, greatly simplifies the dual topology treatment and helps reduce the statistical noise in the alchemical nonequilibrium work W_{int} for the lipid exchange. The alchemical simulation for the attempted exchange consists in switching off the nonbonded interactions (intramolecular and intermolecular) of the lipids a, b, and switching on the nonbonded interactions (intramolecular and intermolecular) of the lipids copies a' and b'. All internal covalent terms are kept all along (bonds, angles, dihedrals, etc). The constraints fixing the cartesian coordinates of the common core atoms do not contribute to the transformation.

If the attempted exchange is accepted, the coordinates of the interacting molecules a', b' at the end of this simulation are copied onto $\mathbf{x}_1^a$ and $\mathbf{x}_1^b$ to continue the equilibrium dynamics. If the attempted exchange is rejected, we return to the initial coordinates and velocities of molecules a, b and of the environment ($\mathbf{x}_0^a$, $\mathbf{x}_0^b$, $\mathbf{X}_0$) to continue the equilibrium dynamics or to do another attempt. A two-ends momentum reversal with a probability of 0.5 is applied before and after slow-growth simulations to symmetrize the probability of attempted exchanges (Chen and Roux (2014)). Under these conditions, the Metropolis acceptance probability for the attempted exchange,

$$\begin{aligned} T^{(\text{a})}(\mathbf{x}_0^a, \mathbf{x}_0^b, \mathbf{X}_0 \rightarrow \mathbf{x}_1^a, \mathbf{x}_1^b, \mathbf{X}_1) &= T_{\text{int}}^{(\text{a})}(\mathbf{x}_0^a, \mathbf{x}_0^b, \mathbf{x}_0^{a'}, \mathbf{x}_0^{b'}, \mathbf{X}_0, \lambda_0 \rightarrow \mathbf{x}_1^a, \mathbf{x}_1^b, \mathbf{x}_1^{a'}, \mathbf{x}_1^{b'}, \mathbf{X}_1, \lambda_1) \\ &= \min\left[1, e^{-W_{int}/k_B T}\right] \end{aligned} \tag{4.7}$$

ensures that the process satisfies microscopic detailed balance, and thus converges toward the equilibrium Boltzmann distribution (Chen and Roux (2015c)).

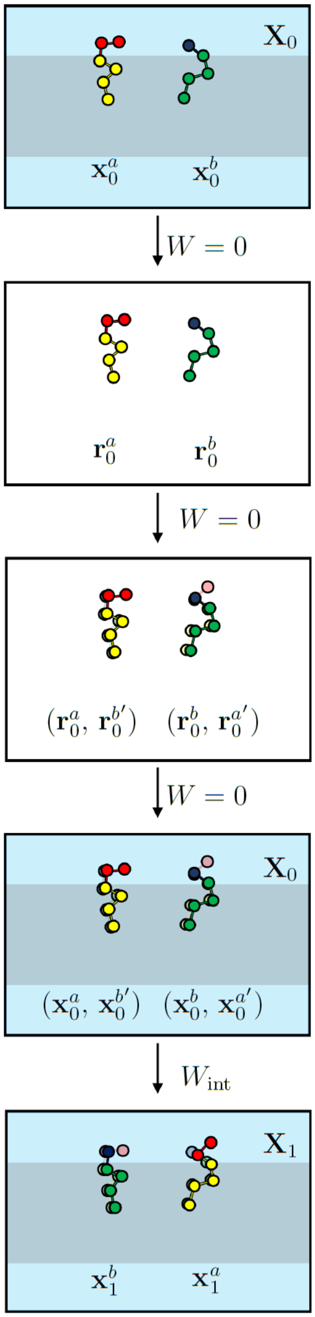


Figure 4.7: Schematic representation of the lipid exchange algorithm. The work associated with the exchange is evaluated in the step in the membrane. $\mathbf{r}_0^a$ and $\mathbf{r}_0^b$ are the coordinates of the lipids a and b at their initial sites and $\mathbf{r}_0^{a'}$ and $\mathbf{r}_0^{b'}$ at their final sites in vacuum. $\mathbf{x}_0^a$ and $\mathbf{x}_0^b$ and $\mathbf{x}_1^a$ and $\mathbf{x}_1^b$ are the coordinates and velocities of the lipids at their initial (0) and final (1) sites in the membrane. $\mathbf{X}_0$ and $\mathbf{X}_1$ are the coordinates and velocities of the environment at the initial and final states. Red and blue spheres represents the non common atoms of the exchanging lipids, while yellow and green spheres stand for their common part. Lighter colored spheres illustrate decoupled states.

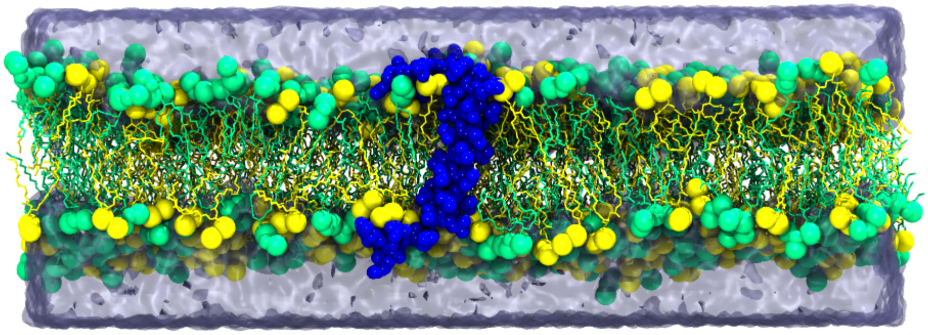


Figure 4.8: Structure of the homogeneous system. A peptide capped by arginines is embedded in the bilayer. The yellow and green spheres represents the polar heads of DLPG and DLPC, respectively.

4.2.3 Simulation details

The neMD/MC algorithm for equilibrating phospholipid bilayers is first tested in the case of a binary mixture of the zwitterionic lipid DLPC (1,2-Dilauroyl-sn-glycero-3-phosphocholine) and the negatively charged lipid DLPG (1,2-dilauroyl-sn-glycero-3-phosphoglycerol). These two lipids have the same acyl chains but possess two very different polar heads. By convention, the common atoms encompass all the atoms of the carbon chains and the phosphate groups, whereas the non-common atoms are those of the choline group of DLPC and those of the glycerol groups of DLPG. The procedure for generating the initial positions of the dangling atoms is detailed in the previous section (Figure 4.5).

A small patch, with 100 DLPG and 100 DLPC per leaflet is created to compare brute-force MD and the neMD/MC algorithm. The lipids are initially sorted (the DLPG in one half of the box, the DLPC in the other half). The initial dimension of this patch is 100 × 100 × 93 Å^3 containing around 110,000 atoms. It is hydrated by around 19,700 water molecules with a concentration of NaCl set to 0.1 M neutralizing the overall charge of the system.

Two large patches, with 300 DLPG and 300 DLPC per leaflet, are also studied. To mimic the effect of a membrane protein on spatial distribution of lipids, a transmembrane peptide of 40 amino acids is embedded in the bilayer (Figure 4.8). It consists of a hydrophobic core of 22 leucine residues capped by 9 positively charged arginine at both ends. The leucine segment is shaped as an α-helix and inserted perpendicular to the membrane plane. After a 10 ns equilibration period, its position is restrained during all simulations by means of an harmonic potential (100 kcal.mol^{-1}.Å^{-1}) applied to the C_α to avoid undesirable movements. The two arginine ends emerge into the solvent on both side of the bilayer. The presence of positively charged residues is intended to favor the enrichment of DLPG against DPLC around the peptide. In one patch, referred to as inhomogeneous, a central circular region of roughly 60 Å in radius contains only DLPC (225 per leaflet), surrounding the transmembrane peptide, while the remaining DPLC are randomly mixed with the DLPG outside the central region. In the second one, referred to as homogeneous, both DLPC and DLPG molecules are randomly positioned in each leaflets. Such a uniform configuration is representative of widely used membrane building procedures (Lee et al. (2016)). The molecular system is hydrated by 56,000 water molecules with a concentration of NaCl set to 0.1 M neutralizing the overall charge of the system. The size of the resulting simulation box is 190 × 190 × 75 Å^3 containing about 280,000 atoms. The three patches with different initial spatial distributions of DLPC and DLPG are generated using the CHARMM-GUI membrane builder (Lee et al. (2016)) (Figure 4.9).

All simulations are performed in the NPT ensemble using the program NAMD (Phillips et al. (2005)). The nonpolarizable CHARMM36 force field (Best et al. (2012)) is used for the peptide, ions, and lipids. Water is described with TIP3P (Jorgensen et al. (1983)). A Langevin thermostat and a Langevin piston are employed to maintain a temperature of 315K and a pressure of 1.0315 bar, respectively (Feller et al. (1995)). The Particle–Mesh Ewald algorithm is used to handle long–range interactions (Procacci et al. (1996)). Above 12 Å, electrostatic and Van der Waals interactions are truncated at a switching distance of 14 Å. The SHAKE/RATTLE Ryckaert et al. (1977); Andersen (1983)) algorithm is used to constrain covalent bonds involving hydrogen atoms and the SETTLE algorithmMiyamoto and Kollman (1992) is utilized for water molecules. In all simulations, the Hydrogen Mass

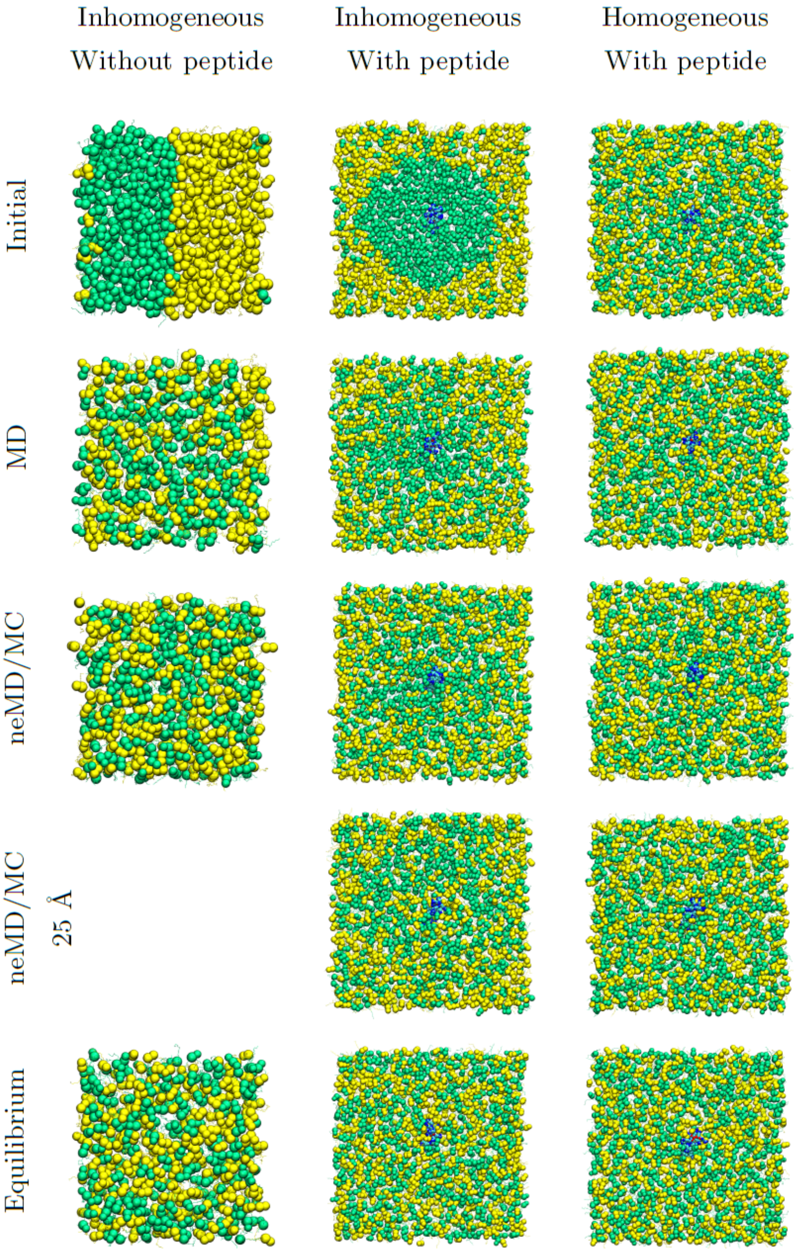


Figure 4.9: Lipid distribution in the different molecular systems. Bilayer without peptide (first column): The MD distribution is generated after 500 ns, the neMD/MC after 500 successes and the equilibrium spatial distribution after 1 μs of trajectory. Bilayer with the peptide (second and third columns): The MD distribution is generated after 360 ns, the neMD/MC after 360 successes and the equilibrium distribution after 3 μs of trajectory. Yellow and green spheres stand for DLPG and DLPC, respectively. First row: initial distribution, Second row: MD, Third row: neMD/MC, Fourth row: neMD/MC with a selection of at least one lipid within 25 Å of the peptide, fifth row: equilibrated distribution.

Repartitioning (HMR) scheme (Hopkins et al. (2015)) is used. All equilibrium MD and neMD/MC simulations were carried out with a time-step of 2 fs.

To serve as standard references, long equilibrium MD simulations exceeding 1 μs are generated for the different molecular systems. All the lipid exchanges are performed using an alchemical dual-topology strategy. No soft-core potential is applied (Zacharias et al. (1994); Beutler et al. (1994)). The electrostatics and the van der Waals interactions are decoupled linearly from the system for λ varying from 0 to 1. The neMD/MC simulations rely on the alchemical thermodynamic integration (TI) code previously developed by Radak in the context of constant-pH simulations (Suh et al. (2018)). The end goal of the method is to exchange lipids in the most efficient way. Rather than optimizing the acceptance probability, $P_{acc}(\tau)$, of lipid exchanges by increasing the sampling time, τ, of the alchemical transformation, we sought to maximize the exchange rate, defined as $k(\tau) = P_{acc}(\tau)/\tau$ (Radak and Roux (2016)). Evolution of $P_{acc}(\tau)$ and $k(\tau)$ as a function of the sampling time are reported in Figure 4.10 for DPLC/DLPG exchanges in a small inhomogeneous lipid patch. Whereas the acceptance probability increases continuously with τ, the optimal exchange rate is obtained for an alchemical transformation performed over 10 ps, the corresponding acceptance probability being 3 %. To improve the computational efficiency of the algorithm and maximize the computational time spent for the lipid exchanges, we limited the equilibrium MD to 100 ps, either when an exchange has been accepted or if 33 attempted exchanges have been successively rejected, ensuring that the exchange rate is on average about 1 lipid exchange per nanosecond. The length of the equlibrium MD is independent of the neMD switches and could be made longer if so desired. In practice, 30 and 40 % of the equilibrium MDs are performed after a successful exchange, whereas the others follow a series of 33 rejected exchanges. For each proposed exchange, we first randomly select one of the two leaflets, then pick at random a couple of DPLG and DPLC lipids. One may note that alchemical simulations often carry some computational overhead relative to straight MD.

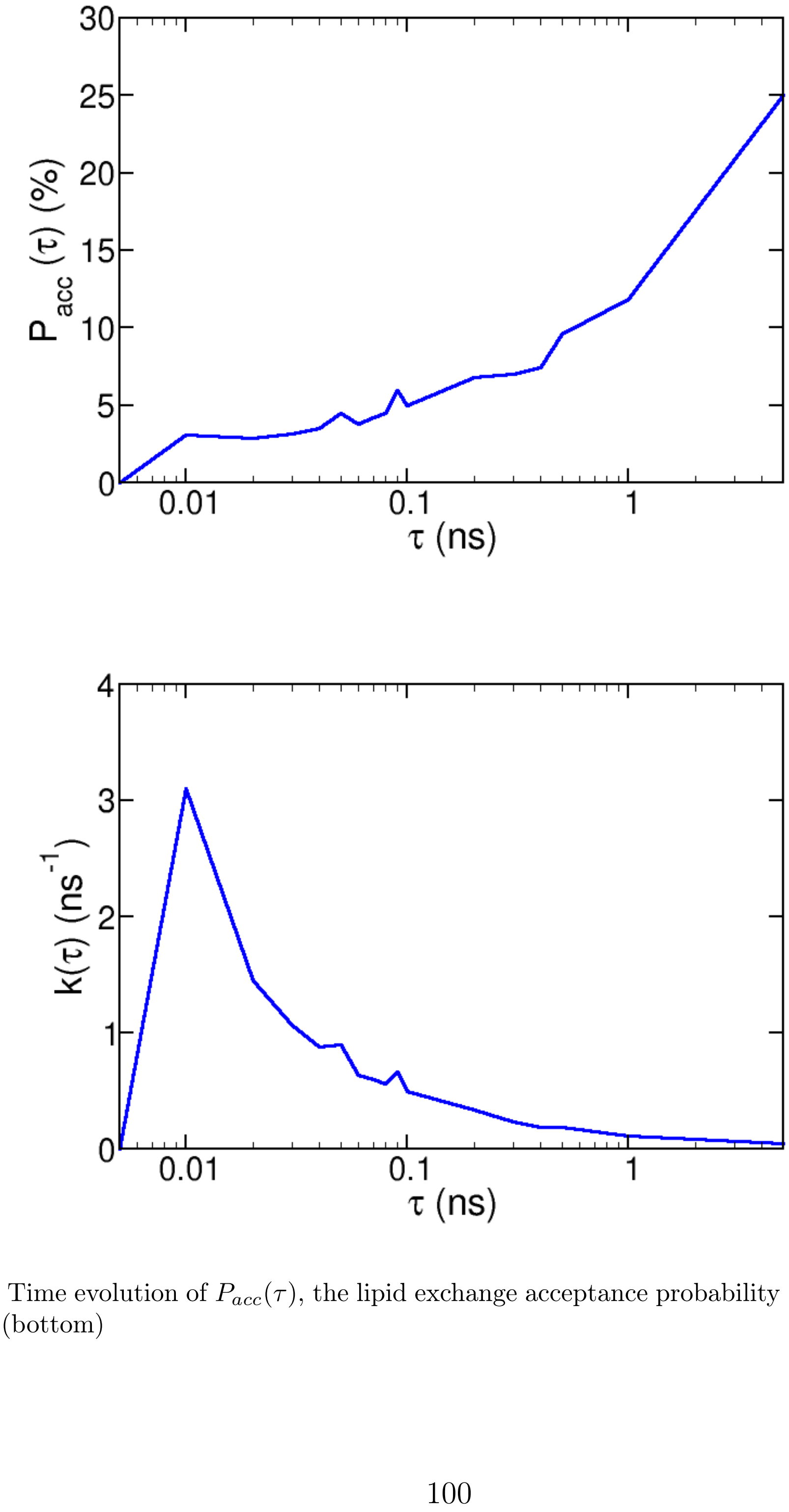


Figure 4.10: Time evolution of $P_{acc}(\tau)$, the lipid exchange acceptance probability (top) and $k(\tau)$, the lipid exchange rate (bottom)

To minimize the computational cost, we used the NAMD option with direct interactions "alchDecouple off" requiring the smallest number of PME grid calculations per step. The performance of neMD alchemical simulations compared to equilibrium MD decreases by 20% on average.

Two different exchange protocols are tested with the neMD/MC simulations (Lines 3 and 4 in Figure 4.9). In the first one, lipids are randomly selected from the entire membrane leaflet. In the second one, the choice of lipids is no longer completely random to increase the lipid exchanges attempts near the peptide to enhance the configurational sampling in the region of interest. The membrane is formally divided into an inner region, corresponding to a distance equal or smaller than 25 Å from the transmembrane helix, and an outer region, corresponding to a distance equal or greater than 25 Å from the helix. All attempted exchanges must always include at least one lipid from the inner region (2 inner or 1 inner and 1 outer). No exchanges between lipids in the outer region far from the peptide are attempted. When the attempted exchange involves lipids lying in the inner and outer region, the Metropolis criteria must be modified as,

$$k_{n_a,n_b \rightarrow n_a-1,n_b+1} = \left(\frac{n_a(N_b - n_b)}{(N_a - n_a + 1)(n_b + 1) + n_a(N_b - n_b)} \right) \tag{4.8}$$

where a represents the lipid initially in the inner region (DLPC or DLPG) and b represents the lipid initially in the outer region (DLPG or DLPC). We want to swap lipids of type a and b between an inner region (i) and an outer region (o). In the entire system, the total number of lipids of type a is N_a, and the total number of lipids of type b is N_b. The total number of lipids in the inner and outer regions does not change upon a swapping event. Let n_a and n_b be the number of lipids a and b in the inner region, and $N_a - n_a$ and $N_b - n_b$ be the number in the outer region. Partitioning these identical molecules between two region involves many equivalent configurations. We write the constrained partition function as,

$$
\begin{aligned}
\Xi = \sum_{n_a \geq 0} \sum_{n_b \geq 0} & \frac{N_a!}{n_a!(N_a - n_a)!} \frac{N_b!}{n_b!(N_b - n_b)!} \\
& \int_i d\mathbf{R}_i \int_o d\mathbf{R}_o \; e^{-\beta U(n_a, n_b, N_a - n_a, N_b - n_b)}
\end{aligned}
\tag{4.9}
$$

where N_a and N_b are the total number of molecules. Introducing the scaled coordinates $\mathbf{X} \equiv \{\mathbf{X}_a, \ldots, \mathbf{X}_n\}$ such that the coordinates become dimensionless and varies between 0 and 1, we re-write this as,

$$
\begin{aligned}
\Xi \;\; = & \sum_{n_a \geq 0} \sum_{n_b \geq 0} \frac{N_a!}{n_a!(N_a - n_a)!} \; (V_i)^{n_a} (V_o)^{N_a - n_a} \; \frac{N_b!}{n_b!(N_b - n_b)!} \; (V_i)^{n_b} (V_o)^{N_b - n_b} \\
& \int_i d\mathbf{X}_i \int_o d\mathbf{X}_o \; e^{-\beta U(n_a, n_b, N_a - n_a, N_b - n_b)}
\end{aligned}
\tag{4.10}
$$

The probability of a given configuration with n_a and n_b particles in the inner region is

$$
\begin{aligned}
P(n_a, n_b) \;\; = & \; \frac{1}{\Xi} \frac{N_a!}{n_a!(N_a - n_a)!} \; (V_i)^{n_a} (V_o)^{N_a - n_a} \; \frac{N_b!}{n_b!(N_b - n_b)!} \; (V_i)^{n_b} (V_o)^{N_b - n_b} \\
& \int_i d\mathbf{X}_i \int_o d\mathbf{X}_o \; e^{-\beta U(n_a, n_b, N_a - n_a, N_b - n_b)}
\end{aligned}
\tag{4.11}
$$

An important pre-requisite in constructing a valid nonequilibrium simulation algorithm is that the system relaxes to the correct statistical properties when the channel is submitted to equilibrium boundary conditions. Let us construct a Markov chain for the system in which the fraction of molecules a and b can vary by +1 (creation) or −1 (destruction) via random transitions. This random walk in the number of particles can be indicated schematically as,

$$
\cdots \leftrightarrow (n_a - 1, n_b + 1) \leftrightarrow (n_a, n_b) \leftrightarrow (n_a + 1, n_b - 1) \leftrightarrow \ldots \tag{4.12}
$$

There is an infinite number of Markov chains with transition probabilities $k_{n_a, n_b \rightarrow n_a + 1, n_b - 1}$

and $k_{n_a+1,n_b-1\rightarrow n_a,n_b}$ converging towards the equilibrium probabilities given by Eq. (4.11). A sufficient condition to insure that the transition probabilities will yield the correct GCE equilibrium probabilities is to impose the condition of detailed balance, i.e.,

$$\begin{aligned} P(n_a, n_b)\ k_{n_a,n_b\rightarrow n_a+1,n_b-1} &= P(n_a+1, n_b-1)\ k_{n_a+1,n_b-1\rightarrow n_a,n_b} \\ \frac{k_{n_a,n_b\rightarrow n_a+1,n_b-1}}{k_{n_a+1,n_b-1\rightarrow n_a,n_b}} &= \frac{P(n_a+1, n_b-1)}{P(n_a, n_b)} \end{aligned} \tag{4.13}$$

We need the ratio of the equilibrium probabilities, $P(n_a + 1, n_b - 1)/P(n_a, n_b)$ to proceed further. If we add a particle of type a and remove a particle of type b,

$$\begin{aligned} \frac{P(n_a+1, n_b-1)}{P(n_a, n_b)} &= \frac{\frac{N_a!}{(n_a+1)!(N_a-n_a-1)!}}{\frac{N_a!}{n_a!(N_a-n_a)!}}\ \frac{\frac{N_b!}{(n_b-1)!(N_b-n_b+1)!}}{\frac{N_b!}{n_b!(N_b-n_b)!}}\ \frac{\frac{(V_i)^{na+n_b}}{(V_o)^{na+n_b}}}{\frac{(V_i)^{na+n_b}}{(V_o)^{na+n_b}}} e^{-\Delta U/k_BT} \\ &= \frac{n_a!(N_a-n_a)!}{(n_a+1)!(N_a-n_a-1)!}\ \frac{n_b!(N_b-n_b)!}{(n_b-1)!(N_b-n_b+1)!}\ e^{-\Delta U/k_BT} \\ &= \frac{(N_a-n_a)}{(n_a+1)}\ \frac{n_b}{(N_b-n_b+1)}\ e^{-\Delta U/k_BT} \end{aligned} \tag{4.14}$$

where ΔU is the difference in energy between the configuration $U(n_a, n_b, N_a - n_a, N_b - n_b)$ to the configuration $U(n_a + 1, n_b - 1, N_a - n_a - 1, N_b - n_b + 1)$. Because the MC move is executed via a nonequilibrium MD exchanging the molecules a and b, ΔU is replaced by the nonequilibrium work to do the swapping according to eq. (1) in the main text. Therefore, we have that

$$\begin{aligned} \frac{k_{n_a,n_b\rightarrow n_a+1,n_b-1}}{k_{n_a+1,n_b-1\rightarrow n_a,n_b}} &= \frac{P(n_a+1, n_b-1)}{P(n_a, n_b)} \\ &= \frac{(N_a-n_a)}{(n_a+1)}\ \frac{n_b}{(N_b-n_b+1)}\ e^{-W_{\text{int}}/k_BT} \end{aligned} \tag{4.15}$$

Detailed balance provides only a constraint on the relative magnitude of the transition probabilities. To have a practical closed form expression, we write

$$k_{n_a,n_b\to n_a+1,n_b-1} = \left(\frac{(N_a-n_a)\,n_b}{C}\right)\min\Big\{1,e^{-W_{\mathrm{int}}/k_BT}\Big\} \tag{4.16}$$

and

$$k_{n_a+1,n_b-1\to n_a,n_b} = \left(\frac{(n_a+1)\,(N_b-n_b+1)}{C}\right)\min\Big\{1,e^{-W_{\mathrm{int}}/k_BT}\Big\} \tag{4.17}$$

Setting the constant, $C=(N_a-n_a)\,n_b+(n_a+1)\,(N_b-n_b+1)$, we get ,

$$\begin{aligned} k_{n_a,n_b\to n_a+1,n_b-1} = &\left(\frac{(N_a-n_a)\,n_b}{(N_a-n_a)\,n_b+(n_a+1)\,(N_b-n_b+1)}\right) \\ &\min\Big\{1,e^{-W_{\mathrm{int}}/k_BT}\Big\} \end{aligned} \tag{4.18}$$

for increasing n_a in the inner region, and

$$\begin{aligned} k_{n_a+1,n_b-1\to n_a,n_b} = &\left(\frac{(n_a+1)\,(N_b-n_b+1)}{(N_a-n_a)\,n_b+(n_a+1)\,(N_b-n_b+1)}\right) \\ &\min\Big\{1,e^{-W_{\mathrm{int}}/k_BT}\Big\} \end{aligned} \tag{4.19}$$

for decreasing n_a in the inner region. Equivalently, we get

$$\begin{aligned} k_{n_a,n_b\to n_a-1,n_b+1} = &\left(\frac{n_a\,(N_b-n_b)}{(N_a-n_a+1)\,(n_b+1)+n_a\,(N_b-n_b)}\right) \\ &\min\Big\{1,e^{-W_{\mathrm{int}}/k_BT}\Big\} \end{aligned} \tag{4.20}$$

for decreasing n_a by 1 and increasing n_b by 1, which can be recognized as the same expression derived for $k_{n_a,n_b\to n_a+1,n_b-1}$ when the subscripts a and b are swapped (adding a and decreasing b versus adding b and decreasing a).

4.2.4 Results and Discussion

Let us now describe the main results obtained from the neMD/MC method. A histogram of the alchemical work calculated from neMD/MC trajectories for the small lipid patch are shown in Figure 4.11. Most of lipid exchanges give rise essentially to positive works distributed around a central value of 40 kcal/mol. However, a tiny proportion of exchanges result in works that are negative or weak enough to pass the Metropolis test. It results in a lipid exchange probability of about 3 % corresponding to one exchange every nanosecond, in line with the data reported in Figure 4.10. The histograms for the other molecular systems are all similar to that plotted in Figure 4.11 (data not shown).

Although the conformation of the incoming lipid is by design very similar to that of the outgoing lipid, the alchemical transformation may sometimes lead to an end-point catastrophe due to an overlap between the incoming lipid atoms and the environment. Such events occur at the first time step of the alchemical transformation, leading automatically to an exchange rejection without any sampling. Alchemical transformations carried out over longer sampling times show that for this type of conformation the exchange work is always high, resulting in the systematic reject of lipid swapping (data not shown here). It is noteworthy that such rejects, which occur before sampling, do not affect the number of lipid exchange accepted per time unit. We have therefore not taken their number into account when estimating the overall exchange rate.

To assess the validity of the neMD/MC algorithm, we first sampled the mixing of a DLPC/DLPG mixture starting from an inhomogeneous configuration in which the two lipid components are separated along the x-axis in two adjacent reservoirs. Figure 4.12 compares the spatial distribution of DLPG obtained after 500 successful lipid exchanges to that observed after 500 ns and 1 μs of brute-force equilibrium MDs. The results indicate that in all three cases, the DLPG have diffused throughout the entire bilayer and their distribution is fairly homogeneous.

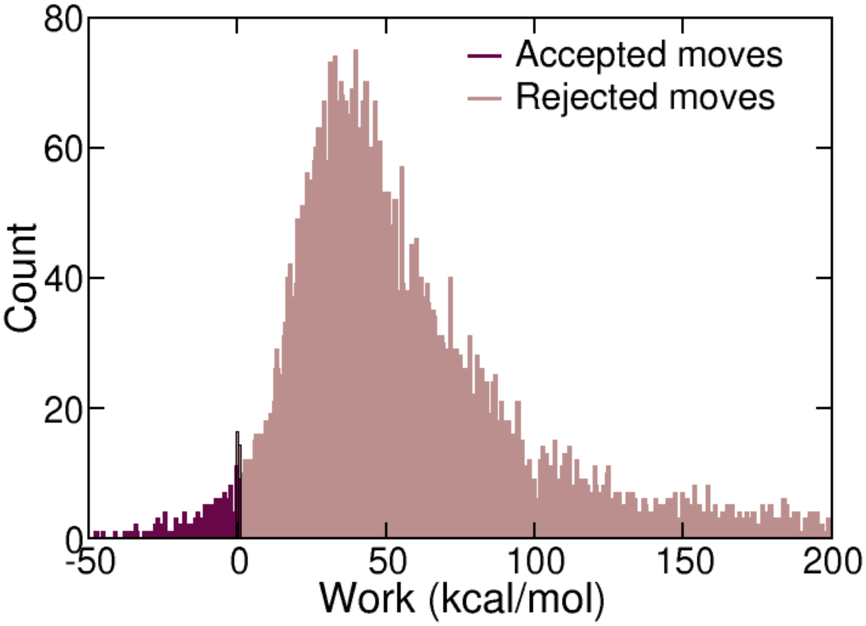


Figure 4.11: Histograms of the works associated with every exchange attempts in the small lipid patch. Brown and indigo stand for rejected (6814) and accepted (200) exchanges, respectively.

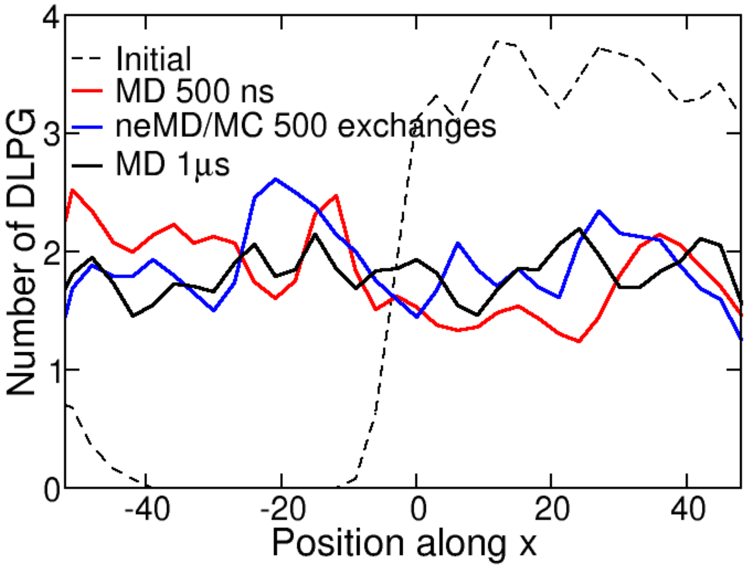


Figure 4.12: Spatial distribution of DLPG along the x-axis in the small bilayer starting from a fully sorted lipid distribution. Initial distribution (dotted line). Distribution after 500 ns of equilibrium MD (red line). Distribution after 500 successful exchanges with the neMD/MC method (sampling time equivalent to 500 ns) (blue line). Distribution after 1 μs of equilibrium MD (black line).

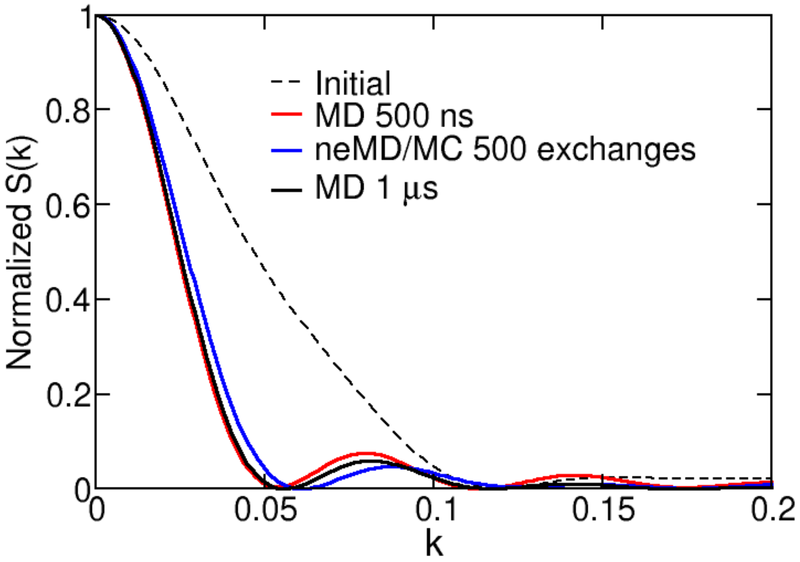


Figure 4.13: Structure factor of the distribution of the DLPG along the x-axis in the small bilayer starting from a fully sorted lipid distribution. Initial distribution (dotted line). Distribution after 500 ns of equilibrium MD (red line). Distribution after 500 successful exchanges with the neMD/MC method (sampling time equivalent to 500 ns) (blue line). Distribution after 1 μs of equilibrium MD (black line).

To better assess the homogenisation of the lipids, Figure 4.13 shows the structure factor of the lipids. It is observed that after 500 ns or 500 exchanges with neMD/MC, we obtain a homogeneous distribution of the lipids that is close to the result after 1 μs.

Simulation of membrane systems requires not only a correct description of the lipid mixing process, but also accurate sampling of the specific association of lipids with membrane proteins. We therefore carried out a series of neMD/MC and equilibrium molecular dynamics simulations for DLPC/DLPC bilayers embedding a transmembrane helix capped with positively charged residues. Figure 4.14 depicts the pair radial distribution functions (Levine et al. (2011)) (RDF) between DLPG headgroups and the first arginine linked to the the hydrophobic helix in the two-dimensional membrane plane generated from a neMD/MC simulation with 360 lipid exchanges together with those monitored after 360 ns and 3 μs of equilibrium MD for the different lipid/peptide systems described in Figure 4.9. Their

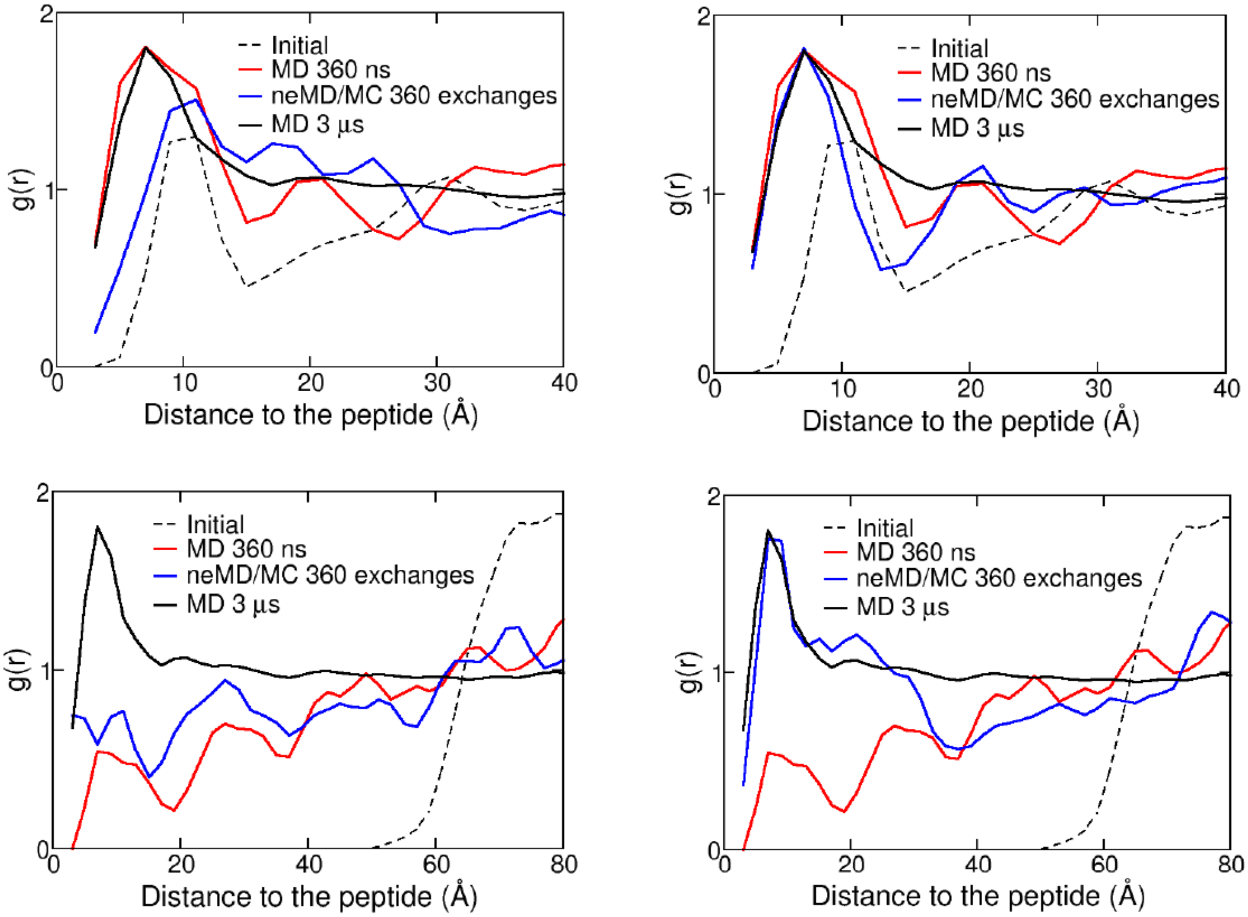


Figure 4.14: Pair radial distribution functions (g(r)) between the phosphorous of DLPG and the alpha carbon of the peptide arginine starting from a homogeneous (top row) and an inhomogeneous (bottom row) lipid distribution. The homogeneous distribution is shown on a smaller scale (only 40 Å around the peptide), as it is around one by construction far from the peptide. For the neMD/MC simulations, exchanged lipids are selected randomly over the entire system (left column) or at a distance lower than 25 Å from the peptide (right column). Initial distribution (dotted line). Distribution after 360 ns of MD (red line). Distribution after 360 successes with the neMD/MC method (sampling time equivalent to 360 ns) (blue line). Distribution after 3 μs of MD (black line).

corresponding integrals are provided in Figure 4.15.

RDFs obtained at different sampling times for equilibrium MD and at different lipid exchange number for neMD/MC as well as the detail of their calculation are provided in Figure 4.16.

The RDF generated from an initial DLPC/DLPG random mixture built with CHARMM-GUI (Lee et al. (2016)) show that, by construction, the arginine neighbourhood is already slightly enriched in negatively charged DLPG. The latter gradually increases to reach an

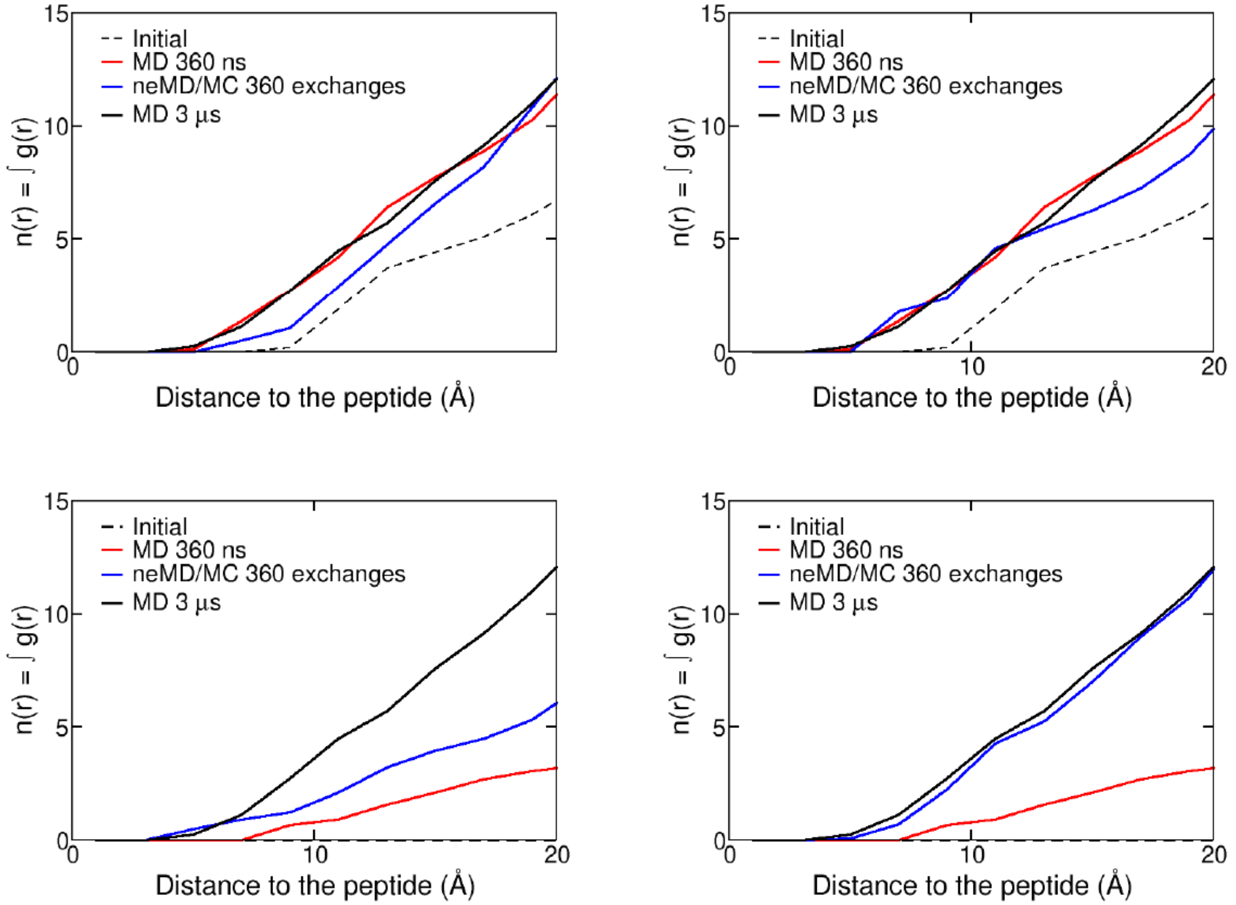


Figure 4.15: Integral of the pair radial distribution functions $g(r)$ between the phosphorous of DLPG and the alpha carbon of the peptide arginines. Initial distribution (dotted line). Distribution after 360 ns of MD (red line). Distribution after 360 successes with the neMD/MC method (sampling time equivalent to 360 ns) (blue line). Distribution after 3 μs of MD. First line: homogeneous patch with peptide. Second line: inhomogeneous patch with peptide. Left column: The neMD/MC method is applied on all the lipids of the systems. Right column: The neMD/MC method is applied on at least one lipid within 25 Å of the peptide.

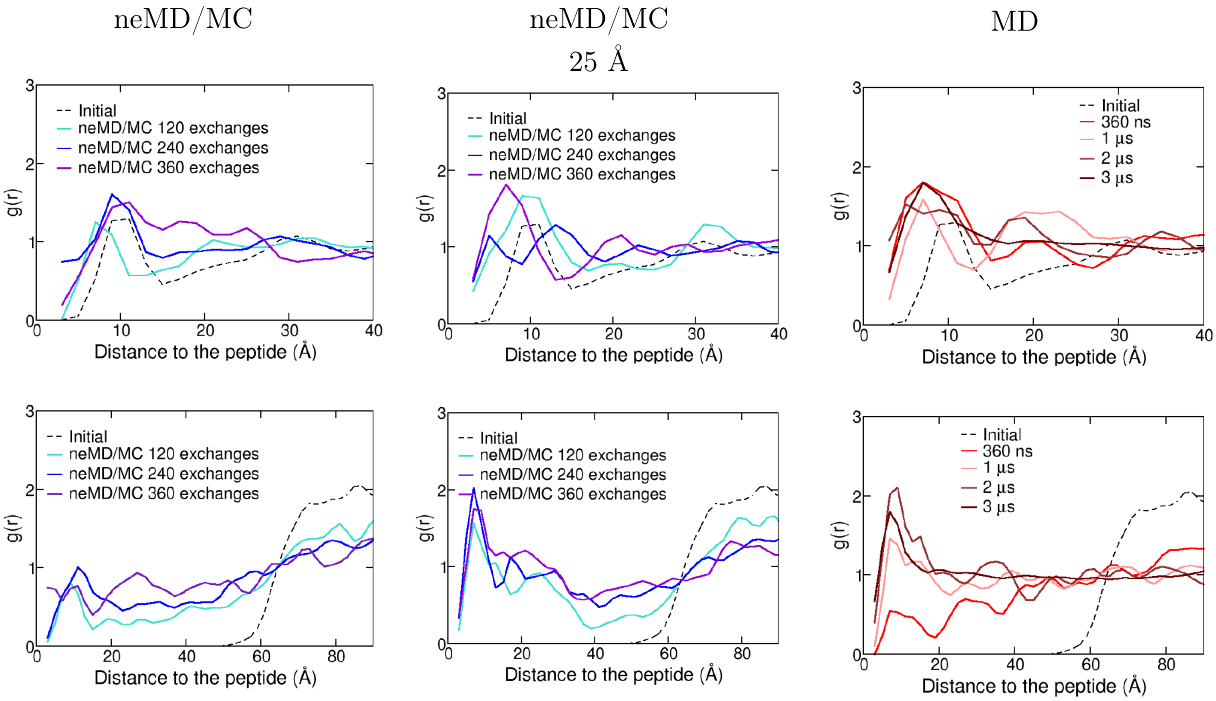


Figure 4.16: Radial distribution g(r) at different time of the neMD/MC method (left), neMD/MC method with at least one lipid within 25 Å of the peptide (center) and the MD (right). The first row shows the distributions in the homogeneous distribution, and the second row shows the inhomogeneous distribution with peptide.

equilibrium illustrated by the RDF at 3 μs./ The RDF obtained after 360 lipid exchanges demonstrates the capacity of the neMD/MC algorithm to generate configurations in line with those sampled after a few μs of equilibrium MD. Biasing the selection of lipids to be exchanged to a region close to peptide does not fundamentally influence the sampling, indicating that for this system, the initial lipid distribution generated by CHARM-GUI is suitable. For the innhomogenous lipid system, all the DPLGs are initially distributed at the periphery of the lipid patch at distances greater that 50 Å from the transmembrane peptide. The RDFs reported in Figure 4.14) show that in 360 ns of equilibrium MD the lipids have barely diffused to the central peptide, leaving the system far from the equilibrium observed after several microseconds of MD simulation. In contrast, a neMD/MC trajectory of 360 lipid exchanges (sampling equivalent to 360 ns of MD at equilibrium) generates configurations in which a few DLPGs are bound to peptide arginines. However, the system is still far from the equilibrium observed after 3 μs of simulation, indicating that for large systems the neMD/MC hybrid approach does not overtake equilibrium MD for lipid sampling. When choosing the lipids to be exchanged in a region close to the peptide, it is clear that the neMD/MC approach can, at least in this region, generate patterns close to those sampled after 3 μs of MD (see Figure 4.14 and Figure 4.15).

Simple considerations are helpful to clarify the conditions where the neMD/MC lipid swapping algorithm is expected to be the most useful compared to unbiased MD simulations. The most difficult situation to sample corresponds to a protein in membrane comprising a mixture in which one type of lipid is very dilute. In this case, it can take a very long time to recruit the dilute lipid near the protein and allow its overall spatial distribution to relax. Treating the problem the limit of a simple bimolecular association and dissociation process, the global relaxation time is expected to go as $\tau^* \propto 1/D(C + K_{\mathrm{d}})$, where C is the concentration per area of the dilute lipid, and K_{d} is the equilibrium dissociation constant of the lipid with the protein. This expression is obtained as the global relaxation time of a two-state system $1/(k_f + k_b)$ where forward rate is $k_f = k_{\text{ass}}C$ and the backward rate is $k_b = k_{\text{ass}}K_{\text{d}}$, with the assumption that the bimolecular association rate k_{ass} is proportional to the diffusion coefficient D. The relaxation time gets slower at low concentration and small diffusion coefficient. In unbiased MD simulations, the lateral diffusion coefficient D_{L} of a lipid molecule in the plane of the bilayer is on the order of 1.0 Å^2/ns. However, in the neMD/MC algorithm, the effective lateral diffusion process is completely artificial. Any given lipid could exchange with a uniform probability anywhere in the simulation box of length L. For each accepted move, the mean-square displacement along one axis increases by $L^2/12$. It follows that the effective diffusion coefficient arising from the lipid exchanges follows the relation, $2D_{\text{ex}}\tau = N_\tau(L^2/12)$, where N_τ is the average neMD/MC moves accepted using a nonequilibrium switching time τ (here, $N_\tau = 1$ when $\tau = 1$ ns). Accordingly, the global relaxation time of the spatial distribution of the dilute lipid around the membrane protein with the neMD/MC algorithm goes as, $\tau^*_{\text{ex}} = \tau^*_{\text{MD}}\,(D_{\text{L}}/D_{\text{ex}})$. For example, in the case of a 100×100 Å^2 membrane patch in a typical simulation system comprising a membrane protein the effective neMD/MC diffusion coefficient is about 417 Å^2/ns suggesting that, compared to unbiased MD, the the spatial distribution of a dilute lipid around the membrane protein could relax almost 500 times faster with the neMD/MC algorithm. The accelerated relaxation is expected to be particularly advantageous at low concentration C. It is important to note, however, that this argument pertains only to the translational diffusion. A complete sampling of a membrane also requires sampling the internal configuration of the lipid chains, which occurs on a timescale of about 20 ns.

The hybrid neMD/MD algorithm for exchanging different molecules with a common core is expected to be of practical value as long as their size and shape are sufficiently similar. Future work will seek to expand the algorithm to sample systems comprising mixtures of multiple lipids that differ from their carbon tails (Fathizadeh and Elber (2018)), as well as heterogeneous systems giving rise to the formation of liquid-disordered and liquid-ordered nanodomains (Park et al. (2023)).

4.2.5 Annexe

The algorithm for the selection of the lipids at a specific distance of the peptide is done as followed:

```
  dlpg = random DLPG selected in the membrane
  dlpc = random DLPC seleced in the same layer than DLPG
  indice_dlpg = 0
  indice_dlpc = 0

if $(x_{PG} * x_{PG} + y_{PG} * y_{PG}) \leq R^2$ then # Is the DLPG within R Å of the peptide?
    indice_dlpg = 1
end if
if $(x_{PC} * x_{PC} + y_{PC} * y_{PC}) \leq R^2$ then # And the DLPC?
    indice_dlpc = 1
end if
if indice_dlpg + indice_dlpc = 0 then
    Select new DLPG, DLPC # No lipid is within 25 Å of the peptide.
else if indice_dlpg + indice_dlpc = 2 then
    Try the exchange # The two lipids are within 25 Å of the peptide.
else# Only one lipid is within 25 Å of the peptide.
    # We calculate the number of the lipids within the considered layer.
    $N_{PG} = 300$
    $N_{PC} = 300$
    $n_{PG}$ = [ Number of DLPG within R Å ]
```

```
        $n_{PC}$ = [ Number of DLPC within R Å ]
    **if** indice_dlpg = 1 **then** # It is the DLPG.
        $P = \frac{n_{PG}(N_{PC}-n_{PC})}{(N_{PG}-n_{PG}+1)(n_{PC}+1)+n_{PG}(N_{PC}-n_{PC})}$
    **else**# It is the DLPC.
        $P = \frac{n_{PC}(N_{PG}-n_{PG})}{(N_{PC}-n_{PC}+1)(n_{PG}+1)+n_{PC}(N_{PG}-n_{PG})}$
    **end if**
    **if** $P \geq$ rand_numb(0,1) **then**
        Try the exchange
    **else**
        Select new DLPG, DLPC
    **end if**
**end if**
```

CHAPTER 5

HYBRID NEMD/MC LIPID SWAPPING ALGORITHM TO EQUILIBRATE MEMBRANE SIMULATION WITH THERMODYNAMIC RESERVOIR

This chapter uses the submitted paper Hybrid neMD/MC Lipid Swapping Algorithm to Equilibrate Membrane Simulation with Thermodynamic Reservoir by **Florence Szczepaniak**, François Dehez, Benoît Roux.

5.1 Introduction

Membranes in living cells are are complex inhomogeneous systems, literally comprising hundreds of chemically distinct lipid species (Watson (2015)). This presents an outstanding challenge to efforts aimed at simulating these systems using detailed atomistic models. The lipidome of living organism comprises a considerable number of chemically different lipids (Wallin and Heijne (1998)). Lipid composition affects not only the mechanical properties of membranes but also modulates the function of membrane proteins through distinct mechanisms (Dowhan (1997); Harayama and Riezman (2018); Laganowsky et al. (2014); Hénault et al. (2019)). The wide ranging composition of biological membranes is such that some components dominate the overall structure of the membrane, while others are present at extremely low abundance (Harayama and Riezman (2018)). Because of the rich compositional diversity, some molecules with low abundance nonetheless find themselves involved in local organizational structures such as lipid rafts (Sezgin et al. (2017)), or affect the properties of cancer cells (Szlasa et al. (2020)). The partitioning of hundreds of lipid species a very low abundance may be associated with bilayer asymetries and various functional complexities (Kopitz (2017); Lorent et al. (2020); Symons et al. (2021)).

From a computational point of view, the sampling challenges of biological membrane using all-atom molecular dynamics (MD) simulations has long been recognized (Goossens and De Winter (2018)). However, the accessible simulation timescales are not sufficient to allow a satisfactory sampling of inhomogeneous multi-component membranes where lateral diffusion of the lipid molecules is very slow (Rose et al. (2015); Muller et al. (2019)). Efforts were dedicated to develop and adapt various enhanced sampling algorithms to more efficiently explore the accessible configurations of membranes (Mori et al. (2016)). Very few methods aimed at sampling the configuration space of inhomogeneous bilayer using all-atom simulations have been proposed in recent years (Huang and García (2014); Mori et al. (2013)). Alternatively, the use of simplified models is one approach that can be used to simulation large and complex cellular membranes (Marrink et al. (2019)). To decrease the computational time, many simulation studies of lipid-protein association in complex mixtures are based on coarse-grained (CG) models, in which groups of atoms are reduced to a single effective particle (Ingólfsson et al. (2014); Marrink et al. (2007); Corradi et al. (2019); Muller et al. (2019)).

A particularly difficult situation is encountered when simulating a system in which some lipid components are meant to be at very low abundance. To illustrate the situation, let us consider a concrete system in which one lipid type is supposed to be at 0.1% mole fraction. In setting up a simulation, one might want to include one copy of this molecule in a system with 1000 lipids. While this may seem reasonable, the fixed number does not account for the considerable fluctuations that could occur at low mole fraction. Furthermore, this naive treatment may become invalid if some protein or other component is present that recruits and increase the present of the lipid at low abundance. Typically, the solution is then to simulate a much larger representation of the system, with say 1000 lipids including 10 copies of the lipid component at low abundance. However, while this may offer a practical solution in some cases, resorting to increasingly large system rapidly becomes computationally costly. Furthermore, it may still run into the same problems if the lipid at low abundance is strongly recruited by some component in the system.

Fundamentally needed is a representation of the simulation box as an open system, in which the number of lipid can naturally fluctuate in equilibrium with an infinite bath or "reservoir" with the desired mole fraction for all lipid components. In principle, Grand Canonical Monte Carlo (GCMC) simulation algorithms can address this type of issue (Woo et al. (2004); Deng and Roux (2008)). However, generating attempted insertion or annihilation attempts of a whole lipid with a non-zero acceptance Metropolis probability in the context of an all-atom membrane simulation is nearly impossible. These practical difficulties can be partly alleviated by substituting the insertion/annihilation with a swapping process, in which the two different types of lipids are exchanged. This is the essence of the hybrid nonequilibrium MD Monte Carlo algorithm (neMD/MC) for lipid exchange that has been proposed recently to better sample the configurations of all-atom membrane models (Szczepaniak et al. (2024)). The neMD/MC approach consists in driving the system via short nonequilibrium trajectories to generate a new state of the system corresponding to the attempted lipid exchange that are subsequently accepted or rejected via a Metropolis MC step (Nilmeier et al. (2011); Chen and Roux (2014, 2015c); Radak and Roux (2016)). These hybrid simulation methods combining the advantages of MC with the strengths of MD offer promising strategies to efficiently sample the configurations of complex molecular systems such as membranes. The ability of neMD/MC simulations to sample equilibrium configurations provide an important tool for studying complex biomolecule systems (Chen et al. (2016); Chen and Roux (2015b); Suh et al. (2018); Gill et al. (2018)). Kindt and coworkers reported the first example of a hybrid MD/MC simulation of a bilayer involving lipid mutations (de Joannis et al. (2006); Coppock and Kindt (2009); Kindt (2011)), followed by Fathizadeh and Elber with the MDAS algorithm (Molecular Dynamics with Alchemical Steps) (Fathizadeh and Elber (2018); Fathizadeh et al. (2020); Cherniavskyi et al. (2020)).

To enforce equilibrium between a simulated system and an infinite surrounding bath,

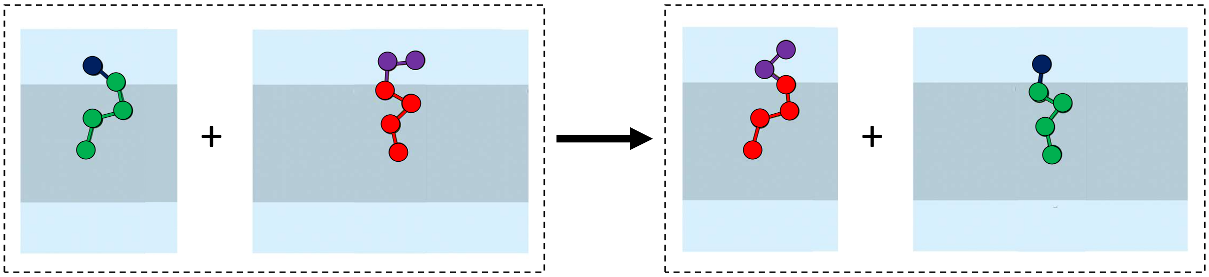


Figure 5.1: Schematic representation of lipid swapping within a combined system (dashed line box) comprising a finite simulation system (left) and a thermodynamic reservoir (right). Depicted is a lipid exchange of the polar head group of a lipid of type a (blue-green) with a lipid of type b (red-yellow) between a simulation box (left) and a large external thermodynamic reservoir (right).

we propose a hybrid neMD/MC algorithm, in which a randomly chosen lipid molecule in the simulated system is swapped with a lipid picked in a separate system serving as a thermodynamic "reservoir" with the desired mole fraction for all lipid components. Theoretical developments regarding the probability of exchanges in the context of an infinite reservoir with the desired mole fraction for all lipid components are presented in the next section. The neMD/MC reservoir algorithm is then examined and tested for few illustrative systems with exchanges of dilauroyl-phosphatidylcholine (DLPC) and anionic dilauroyl-phosphatidylglycerol (DLPG) lipids.

5.2 Theoretical developments

We consider an extended combined system corresponding to a membrane bilayer comprising a simulation box (s) together with a very large external reservoir (r). There are two type of lipids in the extended system, a and b. We want to swap the lipids of type a and b between the simulation box and the large external reservoir, as depicted in Figure 5.1.

In the extended combined system, the total number of lipids of type a is N_a, and the total number of lipids of type b is N_b. The number of lipids of type a and b in the simulation system is n_a and b_b, and the total number of lipids in the simulation system, $N_s = n_a + n_b$, is fixed. The number of lipids of type a and b in the external reservoir is $N_a - n_a$ and $N_b - n_b$.

For clarity, the formal development is pursued with a finite number of lipids in the extended system. At the final stage, we will take the limit that the external reservoir is much larger than the simulated system, with $N_a \gg n_a$ and $N_b \gg n_b$.

We write the constrained partition function of the extended system as,

$$\begin{aligned}\Xi = \sum_{n_a \geq 0} \sum_{n_b \geq 0} \delta_{n_a+n_b,N_s} & \frac{N_a!}{n_a!(N_a-n_a)!} \frac{N_b!}{n_b!(N_b-n_b)!} \\ & \int_s d\mathbf{R}_s \, e^{-\beta U_s(n_a,n_b)} \int_r d\mathbf{R}_r \, e^{-\beta U_r(N_a-n_a,N_b-n_b)}\end{aligned} \tag{5.1}$$

where N_a and N_b are the total number of molecules. Introducing the scaled coordinates $\mathbf{X} \equiv \{\mathbf{x}_a, \ldots, \mathbf{x}_n\}$ such that the coordinates become dimensionless and varies between 0 and 1, we re-write this as,

$$\begin{aligned}\Xi = \sum_{n_a \geq 0} \sum_{n_b \geq 0} \delta_{n_a+n_b,N_s} & \frac{N_a!}{n_a!(N_a-n_a)!} \frac{N_b!}{n_b!(N_b-n_b)!} \\ & (V_s)^{n_a+n_b} \int_s d\mathbf{X}_s e^{-\beta U_s(n_a,n_b)} \\ & (V_r)^{N_b-n_b+Na-n_a} \int_r d\mathbf{X}_r \; e^{-\beta U_r(N_a-n_a,N_b-n_b)}\end{aligned} \tag{5.2}$$

The probability of a given configuration with n_a and n_b particles in the inner region is

$$\begin{aligned}\mathcal{P}(n_a,n_b) = & \frac{1}{\Xi} \frac{N_a!}{n_a!(N_a-n_a)!} \frac{N_b!}{n_b!(N_b-n_b)!} \\ & (V_s)^{n_a+n_b} (V_r)^{N_b-n_b+N_a-n_a} \\ & e^{-\beta[U_s(n_a,n_b)+U_r(N_a-n_a,N_b-n_b)]}\end{aligned} \tag{5.3}$$

An important prerequisite in constructing a valid non-equilibrium simulation algorithm is that the system relaxes to the correct statistical properties when the channel is submitted to equilibrium boundary conditions. Let us construct a Markov chain for the system in which the fraction of particles 1 and 2 can vary by +1 (creation) or −1 (destruction) via random transitions. This random walk in the number of particles can be indicated schematically as,

$$\dots \leftrightarrow (n_a - 1, n_b + 1) \leftrightarrow (n_a, n_b) \leftrightarrow (n_a + 1, n_b - 1) \leftrightarrow \dots$$

Stepping toward the right replaces a lipid b by a lipid a, Stepping toward the left replaces a lipid a by a lipid b. There is an infinite number of Markov chains with transition probabilities $k_{n_a,n_b \to n_a+1,n_b-1}$ and $k_{n_a+1,n_b-1 \to n_a,n_b}$ converging towards the equilibrium probabilities given by Eq. (5.3). A sufficient condition to insure that the transition probabilities will yield the correct equilibrium probabilities is to impose the condition of detailed balance, i.e.,

$$\mathcal{P}(n_a, n_b)\ k_{n_a,n_b \to n_a+1,n_b-1} = \mathcal{P}(n_a + 1, n_b - 1)\ k_{n_a+1,n_b-1 \to n_a,n_b} \tag{5.4}$$

To proceed further, we need to determine the ratio of the equilibrium probabilities, $\mathcal{P}(n_a + 1, n_b - 1)/\mathcal{P}(n_a, n_b)$. If we add one molecule of type a and remove a molecule of type b,

$$\begin{aligned} \frac{\mathcal{P}(n_a + 1, n_b - 1)}{\mathcal{P}(n_a, n_b)} &= \frac{n_a!(N_a - n_a)!}{(n_a + 1)!(N_a - n_a - 1)!} \frac{n_b!(N_b - n_b)!}{(n_b - 1)!(N_b - n_b + 1)!} e^{-\beta \Delta W} \\ &= \frac{(N_a - n_a)}{(n_a + 1)} \frac{n_b}{(N_b - n_b + 1)} e^{-\beta \Delta W} \end{aligned} \tag{5.5}$$

where ΔW is the neMD work to go from the potential energy $[U_s(n_a, n_b) + U_r(N_a - n_a, N_b - n_b)]$ to the potential energy $[U_s(n_a + 1, n_b - 1) + U_r(N_a - n_a - 1, N_b - n_b + 1)]$. Therefore, we have that

$$\frac{k_{n_a,n_b \to n_a+1,n_b-1}}{k_{n_a+1,n_b-1 \to n_a,n_b}} = \frac{\mathcal{P}(n_a + 1, n_b - 1)}{\mathcal{P}(n_a, n_b)} \tag{5.6}$$

$$= \frac{(N_a - n_a)}{(n_a + 1)} \frac{n_b}{(N_b - n_b + 1)} e^{-\beta \Delta W} \tag{5.7}$$

Detailed balance provides only a constraint on the relative magnitude of the transition probabilities. We write

$$k_{n_a,n_b\to n_a+1,n_b-1} \quad = \quad \left(\frac{(N_a - n_a)\,n_b}{C}\right)\min\Big\{1, e^{-\beta\Delta W}\Big\} \tag{5.8}$$

and

$$k_{n_a+1,n_b-1\to n_a,n_b} \quad = \quad \left(\frac{(n_a+1)\,(N_b - n_b + 1)}{C}\right)\min\Big\{1, e^{-\beta\Delta W}\Big\} \tag{5.9}$$

Setting the constant, $C = (N_a - n_a)\,n_b + (n_a+1)\,(N_b - n_b + 1)$, we get,

$$\begin{aligned} k_{n_a,n_b\to n_a+1,n_b-1} = &\left(\frac{(N_a - n_a)\,n_b}{(N_a - n_a)\,n_b + (n_a+1)\,(N_b - n_b + 1)}\right) \\ &\min\Big\{1, e^{-\beta\Delta W}\Big\} \end{aligned} \tag{5.10}$$

and

$$\begin{aligned} k_{n_a+1,n_b-1\to n_a,n_b} = &\left(\frac{(n_a+1)\,(N_b - n_b + 1)}{(N_a - n_a)\,n_b + (n_a+1)\,(N_b - n_b + 1)}\right) \\ &\min\Big\{1, e^{-\beta\Delta W}\Big\} \end{aligned} \tag{5.11}$$

which can be shifted to the initial state n_a, n_b by decreasing n_a by 1 and increasing n_b by 1,

$$\begin{aligned} k_{n_a,n_b\to n_a-1,n_b+1} = &\left(\frac{n_a\,(N_b - n_b)}{(N_a - n_a + 1)\,(n_b+1) + n_a\,(N_b - n_b)}\right) \\ &\min\Big\{1, e^{-\beta\Delta W}\Big\} \end{aligned} \tag{5.12}$$

to obtain an expression for the removal of a lipid of type a, Now, we take the limit that the external reservoir is much larger than the simulated system, with $N_a \gg n_a$ and $N_b \gg n_b$, and that it can essentially be treated as an infinite reservoir with fixed mole fractions $f_a =$

$N_a/(N_a + N_b)$ and $f_b = N_b/(N_a + N_b)$,

$$\lim_{N_a,N_b\to\infty} k_{n_a,n_b\to n_a+1,n_b-1} \;=\; P_a^{\text{incr}} \times \min\Big\{1, e^{-\beta\Delta W}\Big\} \tag{5.13}$$

and

$$\lim_{N_a,N_b\to\infty} k_{n_a,n_b\to n_a-1,n_b+1} \;=\; P_a^{\text{decr}} \times \min\Big\{1, e^{-\beta\Delta W}\Big\} \tag{5.14}$$

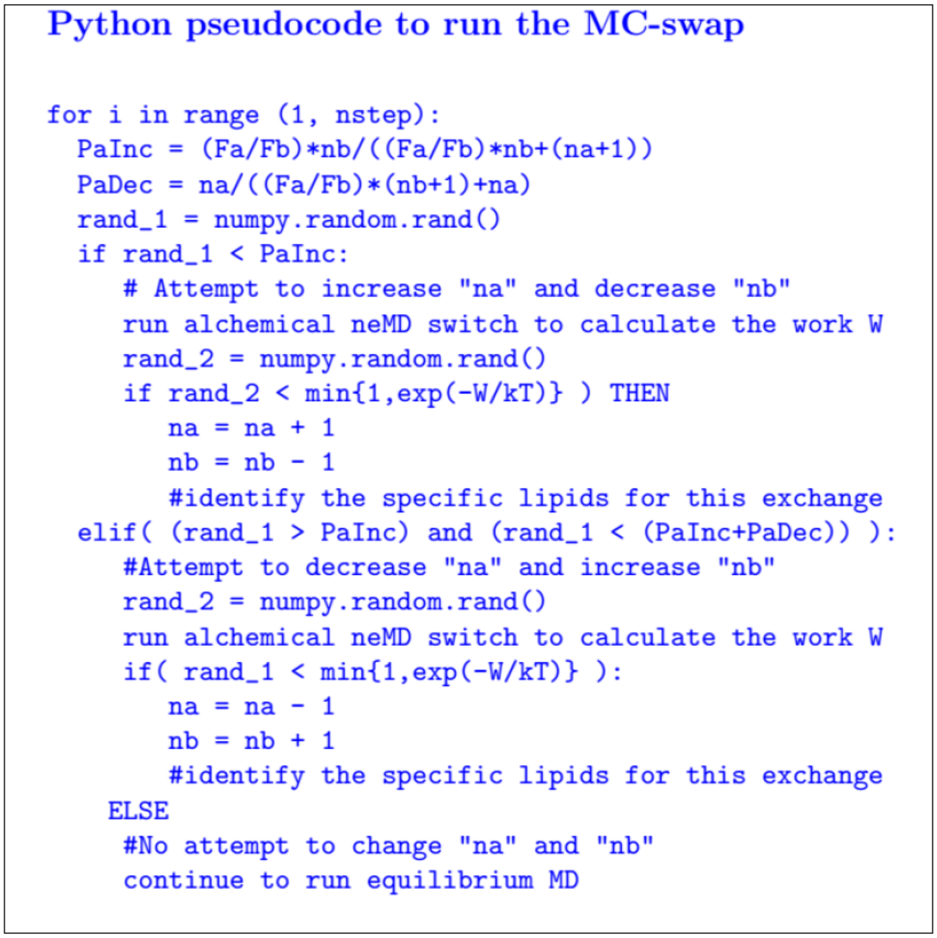
where

$$\begin{aligned} P_a^{\text{incr}} &= \left(\frac{(N_a/N_b)\, n_b}{(N_a/N_b)\, n_b + (n_a+1)}\right) \\ &= \left(\frac{(f_a/f_b)\, n_b}{(f_a/f_b)\, n_b + (n_a+1)}\right) \end{aligned} \tag{5.15}$$

and

$$\begin{aligned} P_a^{\text{decr}} &= \left(\frac{n_a}{(N_a/N_b)\,(n_b+1) + n_a}\right) \\ &= \left(\frac{n_a}{(f_a/f_b)\,(n_b+1) + n_a}\right) \end{aligned} \tag{5.16}$$

are the probability for attempting to increase or decrease the number of lipid of type a, respectively. It is noted that the probabilities for increasing or decreasing are symmetric with respect to a and b. These expressions were derived assuming that only lipid a and b are exchanged. However, if the lipid of type a is anionic, it may be necessary to simultaneously swap a cation (c) with a water molecules (w) to maintain charge neutrality in the simulated system. Further analysis shows that accounting for this additional exchange introduces a factor of $((n_c+1)/n_w)\,(f_w/f_c)$ multiplying (n_a+1) in the expression for P_a^{incr}, or the n_a in the expression for P_a^{decr}, where $f_c = N_c/(N_c + N_w)$ and $f_w = N_w/(N_c + N_w)$ are the mole fraction of cations and water molecules in the reservoir, respectively. This analysis shows



Python pseudocode to run the MC-swap

```
for i in range (1, nstep):
  PaInc = (Fa/Fb)*nb/((Fa/Fb)*nb+(na+1))
  PaDec = na/((Fa/Fb)*(nb+1)+na)
  rand_1 = numpy.random.rand()
  if rand_1 < PaInc:
     # Attempt to increase "na" and decrease "nb"
     run alchemical neMD switch to calculate the work W
     rand_2 = numpy.random.rand()
     if rand_2 < min{1,exp(-W/kT)} ) THEN
        na = na + 1
        nb = nb - 1
        #identify the specific lipids for this exchange
  elif( (rand_1 > PaInc) and (rand_1 < (PaInc+PaDec)) ):
     #Attempt to decrease "na" and increase "nb"
     rand_2 = numpy.random.rand()
     run alchemical neMD switch to calculate the work W
     if( rand_1 < min{1,exp(-W/kT)} ):
        na = na - 1
        nb = nb + 1
        #identify the specific lipids for this exchange
    ELSE
     #No attempt to change "na" and "nb"
     continue to run equilibrium MD
```

Figure 5.2: Pseudocodo for the MC-swap

that the multiplicative factor should be very close to unity, $((n_c+1)/n_w) \approx (f_c/f_w)$, as long as the salt solution in the simulated system remains stably at the same concentration as the reservoir. For the sake of simplicity, this additional factor was not included in the present simulations. The basic steps of the algorithm are given in Figure 5.2.

5.3 Computational methodology

5.3.1 Swapping protocol

Following Figure 5.2, the character of the attempted swapped is randomly chosen based on the probabilities P_a^{incr} and P_a^{decr} based on Eqs. (5.15) and (5.16). Figure 5.3 shows a

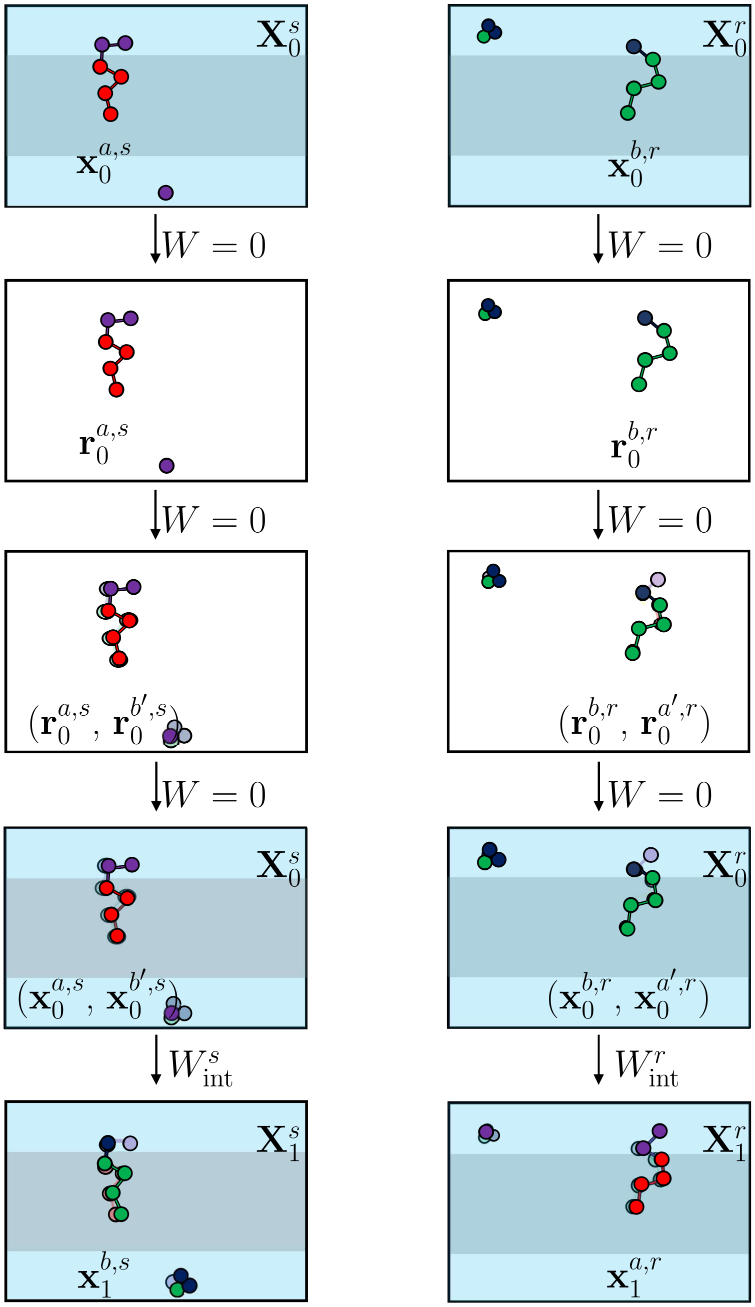


Figure 5.3: Schematic representation of the exchange methodology. s and r represent the simulation system (s) and the reservoir (r), $\mathbf{x}_0^{a,s}$ and $\mathbf{x}_0^{b,r}$ are the coordinates and velocities of the molecules a and b in their initial systems. $\mathbf{r}_0^{a,s}$ and $\mathbf{r}_0^{b,r}$ are the coordinates of the molecules in the vacuum. $\mathbf{r}_0^{a',r}$, $\mathbf{r}_0^{b',s}$, $\mathbf{x}_0^{a',r}$ and $\mathbf{x}_0^{b',s}$ are the coordinates (and velocities if written as $\mathbf{x}$) of the molecules in their new conformation in the other molecular system before the exchange. $\mathbf{x}_1^{a,r}$ and $\mathbf{x}_1^{b,s}$ are the positions and velocities of a and b in their final state after the exchange. $\mathbf{X}^s$ and $\mathbf{X}^r$ are the coordinates and velocities of the rest of the systems, indexed as 0 for the initial values, and 1 for the final values.

detailed representation of the swapping procedure. If the attempt is to increase (decrease) the number of lipid of type a, then a lipid of type a (type b) is randomly chosen in the simulation system (called s in Figure 5.3). Simultaneously, a lipid of type b (type a) is randomly chosen in the reservoir (called r in Figure 5.3). To maintain charge neutrality, a negative lipid is associated with a random cation, and the neutral lipid is associated with a random water molecule. The lipids are swapped in the membrane, and the water molecule and ion are exchanged in the bulk. The exchanges follow the two steps procedure developed by the authors (Szczepaniak et al. (2024)). For the molecular system s, the lipid a (b) is chosen, then copied in vacuum. The position of all the atoms is restraints. The lipid b (a) chosen in r is then copied in this box of vacuum, such that the carbon tails, common between the two lipids, are aligned, the polar heads are approximately superposed, and the oxygen of the water is aligned on the ion. A weak restraint is added between the polar heads and another one between the ion and water. A simulation in vacuum is run to equilibrate the conformation of the molecule b (a). Then, an alchemical simulation is run in the molecular system s, with the common atoms of a and b restraints, and the conservation of the weak restraint between the polar heads, and the water and ion. During this alchemical exchange, the molecule a (b) is decoupled, and the molecule b (a) are coupled. From this simulation, the nonequilibrium work associated with the exchange carried out over the finite switching time t_{sw} is computed using a Thermodynamic Integration (TI) procedure,

$$W_{\text{int}}^{s} = \int_{0}^{t_{\text{sw}}} \left(\frac{\partial U_{\text{int}}}{\partial \lambda} \right) \dot{\lambda}(t) \, dt \tag{5.17}$$

The same method is used for the molecular system r. During the alchemical simulation r, the molecule b (a) is decoupled and the molecule a (b) are coupled, and the work associated with the exchange W_{int}^{r} is computed using the same equation 5.17. Then, a Metropolis test is run on the nonequilibrium work to accept or reject the exchange:

$$\begin{aligned} &T^{(\mathrm{a})}(\mathbf{x}_0^{a,s}, \mathbf{X}_0^s, \mathbf{x}_0^{b,r}, \mathbf{X}_0^r \rightarrow \mathbf{x}_1^{a,r}, \mathbf{X}_1^s, \mathbf{x}_1^{b,s}, \mathbf{X}_1^r) \\ &= \min\left[1, e^{-\Delta W/k_B T}\right] \end{aligned} \tag{5.18}$$

where $\Delta W = W_{\mathrm{int}}^s + W_{\mathrm{int}}^r$. This Metropolis test ensures the microscopic detailed balance, and thus the convergence toward the equilibrium Boltzmann distribution (Chen and Roux (2015c)). Symmetric momentum reversal conditions are applied before and after the attempted exchange (Chen and Roux (2014)).

Note that the whole exchange and this Metropolis test happen only if the probability to attempt such an exchange is high enough. There is a first test using the probabilities (5.15) and (5.16) to know if an exchange is possible, and which lipids would be exchanged.

5.3.2 Simulation details

In order to prove the accuracy and the efficiency of the methodology, four examples are being studied. The lipids considered are DLPG (1,2-dilauroyl-sn-glycero-3-phosphoglycerol) and DLPC (1,2-Dilauroyl-sn-glycero-3-phosphocholine). Two patches are being equilibrated, each of them with 50 DLPC per layer. One is a pure membrane, the other one has a peptide in the middle of the membrane. The peptide is a helix of 22 leucines in the membrane, with 9 arginines at both ends. The arginines of this peptide should have a high affinity with negatively charged lipids. These two patches are equilibrated with two reservoirs: one is a homogeneous mixture of DLPG and DLPC, with 50 lipids of each kind per layer and the other one is a membrane with 99 DLPC and 1 DLPG per layer. The two patches and the two reservoirs are shown Figure 5.4. The four examples studied are the equilibration of each patch with the two reservoirs.

All the systems are created using the CHARMM-GUI membrane builder (Lee et al.

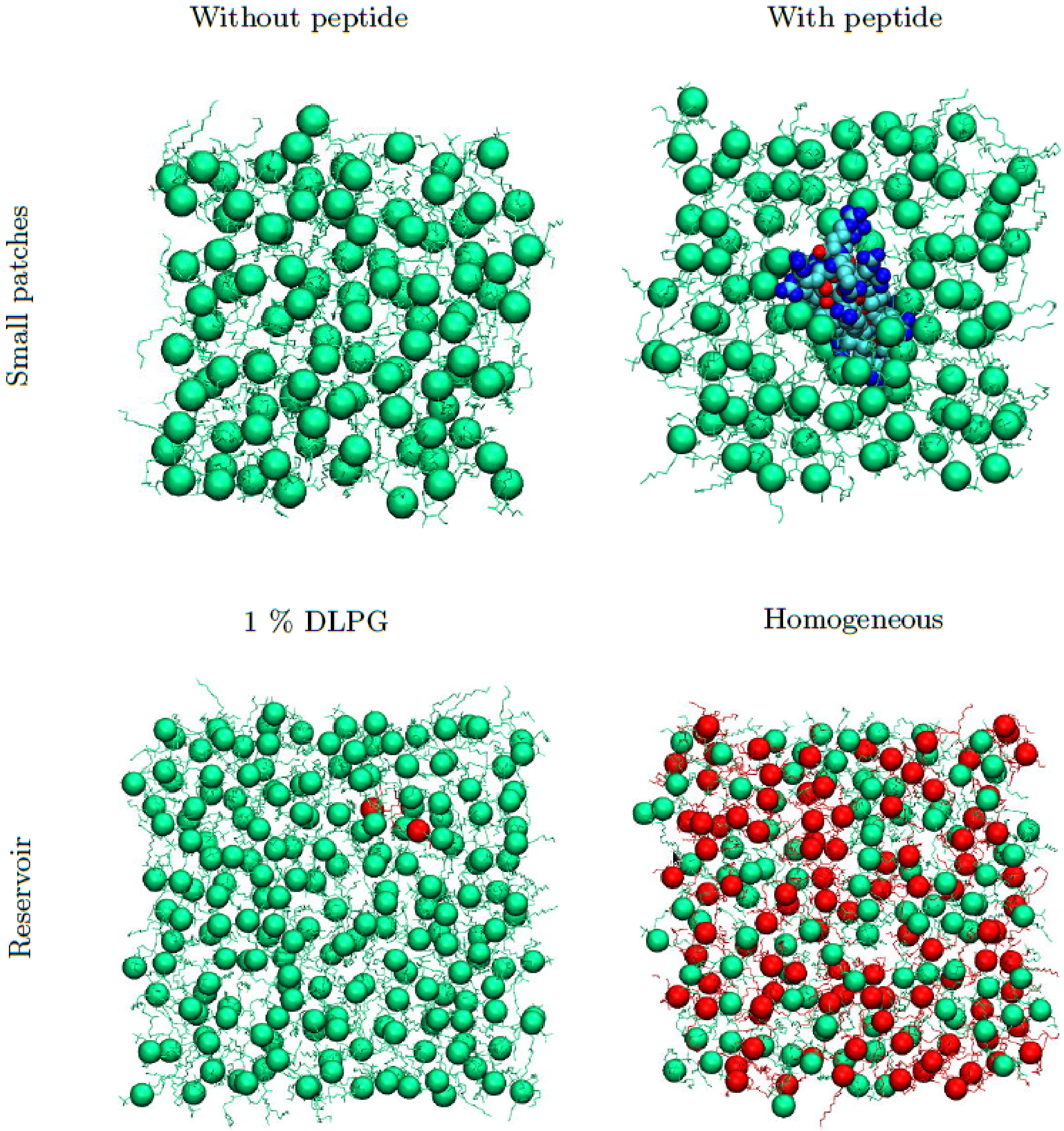


Figure 5.4: Systems used for the study. On the first row, the two patches that need to be equilibrated. On the second row, the two reservoirs. In green: DLPC, in red: DLPG, in blue and cyan: the peptide. The beads correspond to the position of the phosphate group of each lipid.

(2016)). The two reservoirs are approximately 80 Å × 80 Å × 80 Å, with about 50,000 atoms. The small patch without peptide is 55 Å × 55 Å × 80 Å,with about 25,000 atoms, and the one with a peptide is 50 Å × 60 Å × 94 Å with about 35,000. In each system, the concentration of NaCl is set up to be at 0.1 M. All simulations are performed using NAMD (Phillips et al. (2005)), in the NPT ensemble. The water is described using the TIP3P model (Jorgensen et al. (1983)), the lipids, ions and peptide are described using the nonpolarizable force field CHARMM36 (Best et al. (2012)). To keep the temperature fixed at 315 K and the pressure fixed at 1.0315 bar, Langevin thermostat and piston are used (Feller et al. (1995)). To handle long-range interactions, the Particle-Mesh Ewald algorithm is used (Procacci et al. (1996)), and the van der Waals and electrostatic interactions are truncated above 12 Å at a switching distance of 14 Å. To constrain covalent bonds, the SHAKE/RATTLE (Ryckaert et al. (1977); Andersen (1983)) algorithm is used, and the SETTLE algorithm (Miyamoto and Kollman (1992)) is used for the water. The Hydrogen Mass Repartitioning (HMR) scheme (Hopkins et al. (2015)) is used, and all equilibrium MD and neMD/MC simulations were carried out with a time-step of 2 fs.

The alchemical exchanges are performed over 50 ps, and the equilibrium MD simulations are 100 ps long. The length of the MD is independent of the switches, so can be made longer if desired. The main focus of this project being to exchange lipids between two systems, it was not necessary to make it long in this context. The alchemical exchange is done without soft-core potential (Zacharias et al. (1994); Beutler et al. (1994)), with a linear modification of the Van der Waals and electrostatic interactions for λ varying from 0 to 1. The alchemical TI code developed by Radak (Suh et al. (2018)) is used. Radak and Roux (Radak and Roux (2016)) showed that the efficiency in the neMD/MC would decrease if the switch was too long, even though it would increase the probability to accept and exchange. Therefore, the length of the alchemical switch has been chosen to be as short as possible. The two alchemical simulations (the one in the small patch and the one in the reservoir) are done using the same parameters. As shown in the theoretical developments, it can happen that the probability for attempting an exchange is too low to run an alchemical simulation. A first test is done to know if it is possible to try an exchange, and if it is, if it is to increase or decrease the number of DLPG. If this first test allows the attempt of an exchange, a second Metropolis test is done after the exchange to accept or reject it (equation 5.18). Because some times, the first test will not even allow an exchange, and because the switch is so short that it does not have a high probability to be accepted, the equilibrium MD is performed after 65 attempts. Even if the first test does not allow to run the exchanges, it is considered in these 65 attempts. The concentration of the lipids in the reservoir is supposed to be constant. Therefore, it is not necessary to keep the lipid distribution after the exchange. To limit the computational cost associated with the reservoir, a unique trajectory of the membrane is generated using brute force MD. At each exchange, a frame is randomly selected in the trajectory, and from this frame, the coordinates of the molecules as well as their velocities are extracted. These are used during the alchemical exchange, but, after the exchange, once the simulation is done and the work is computed, the newly generated lipid distribution is not conserved. The next attempt will be done using another frame of the initial trajectory.

5.4 Results

Figure 5.5 shows the work associated with the exchanges of lipids between the patch with peptide and the reservoir at 1 % of DLPG. The histograms for the other simulations are all similar (not shown).

Figure 5.6 shows the evolution of the mole fraction of DLPG in each membrane equilibrated with the reservoir at 1 % of DLPG. The two systems are not converged yet. The expected mole ratio in each membrane is about 1 %, but the mole ratio presented is about 0.5 %. To verify the validity of the equations, the membrane with peptide is equilibrated with the reservoir at 1 % of DLPG. But for this test only, the exchanges in the membranes are

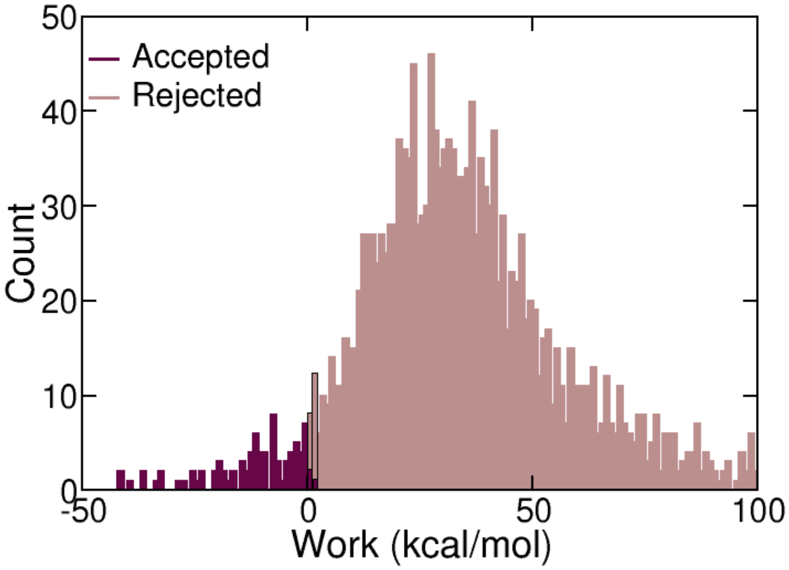


Figure 5.5: Histogram of the work associated with the exchanges of lipids between the patch with peptide and the reservoir at 1 % of DLPG. Brown and indigo stand for rejected and accepted exchanges, respectively.

always accepted, so only the probabilities of increasing or decreasing the number of DLPG (equations 5.15 and 5.16) control the mole ratio of DLPG in the membrane. This test shows that the mean value of the mole ratio of DLPG is converging to 1 (Figure 5.7 top). So the probabilities do not justify the difficulty to observe the convergence of the systems in Figure 5.6. Another hypothesis is that after insertion of a lipid, the membrane needs to relax. When comparing the work associated to the attempts to increase or to decrease the number of DLPG, it shows that decreasing the number of DLPG is associated with a lower value of the work (Figure 5.7 bottom). In the algorithm, the MD simulation to equilibrate the system is done after 65 exchanges. The membranes probably need more time to relax and stabilize the insertion of the lipids.

Figure 5.8 shows the evolution of the mole fraction of DLPG in each membrane equilibrated with the reservoir at 50 % of DLPG. Both membranes show fluctuations around 50 % of DLPG. The membrane with the peptide converges faster to this value than the system

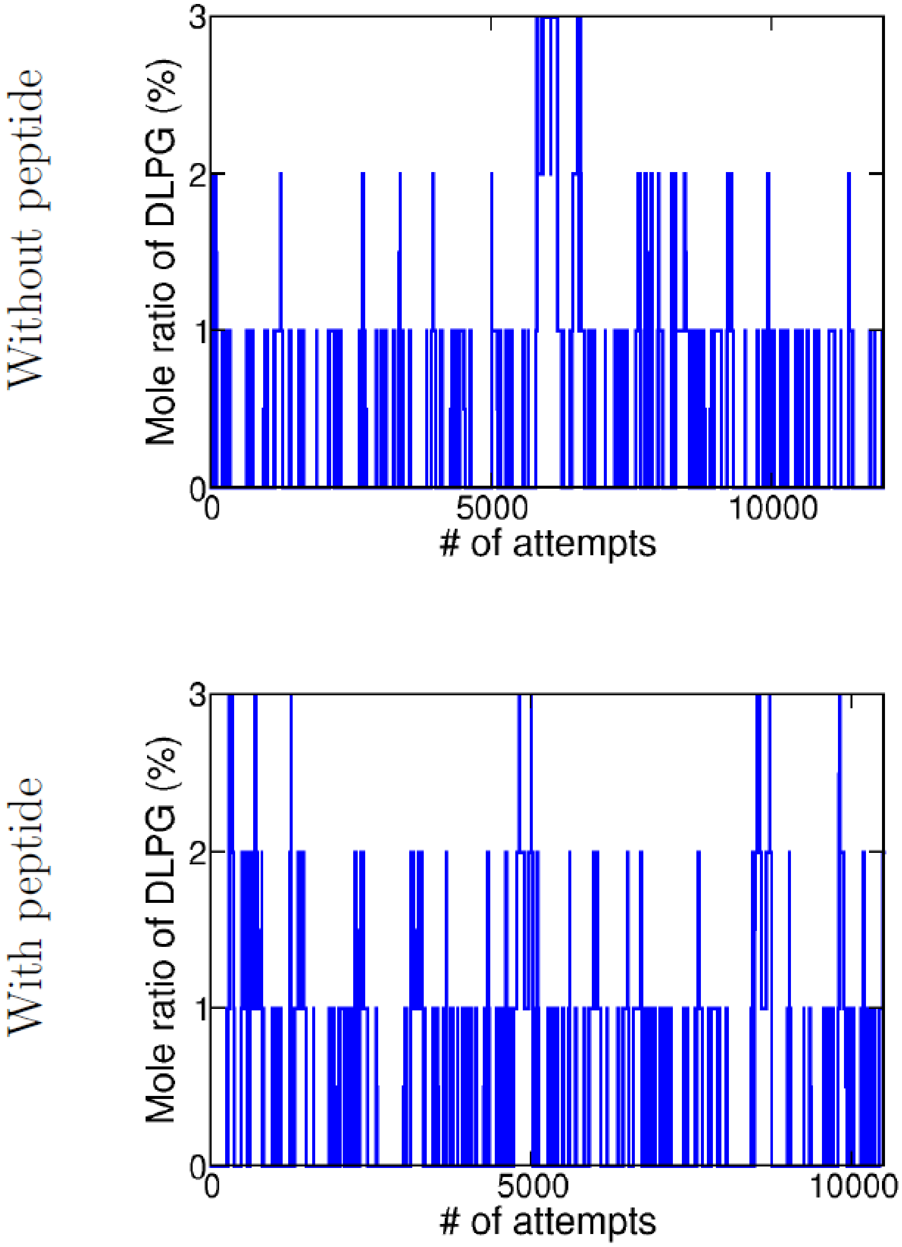


Figure 5.6: Evolution of the number of DLPG in the simulated systems. The systems are equilibrated with a reservoir at 1 % of DLPG (top) membrane without a peptide (bottom) membrane with a peptide. The average without peptide is 0.40 %. The average with peptide is 0.44 %.

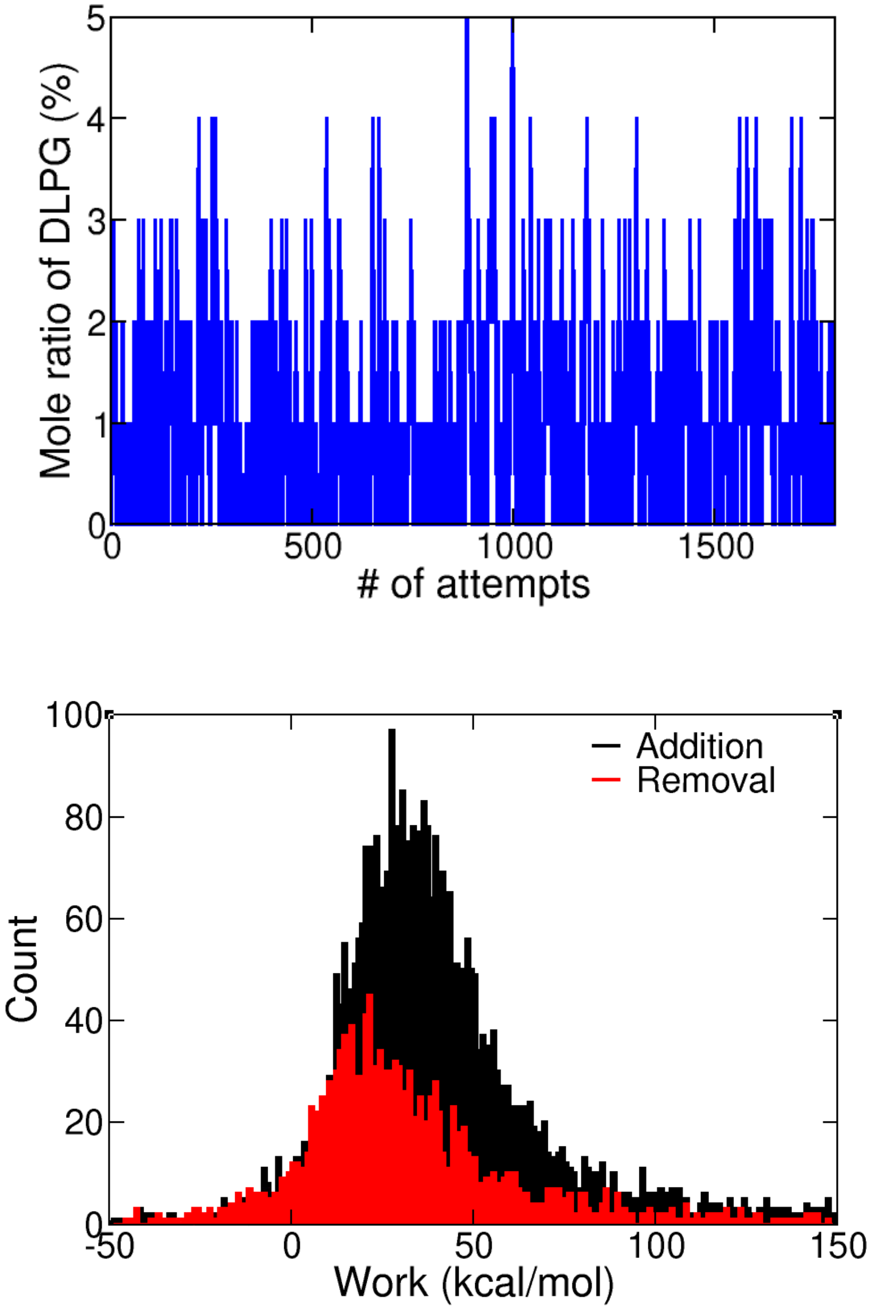


Figure 5.7: Top: Evolution of the number of DLPG in the simulated systems. The system is equilibrated with a reservoir at 1 % of DLPG membrane with a peptide. The average without peptide is 1 %. Bottom: Work association with the simulations to add (black) or remove (red) a DLPG.

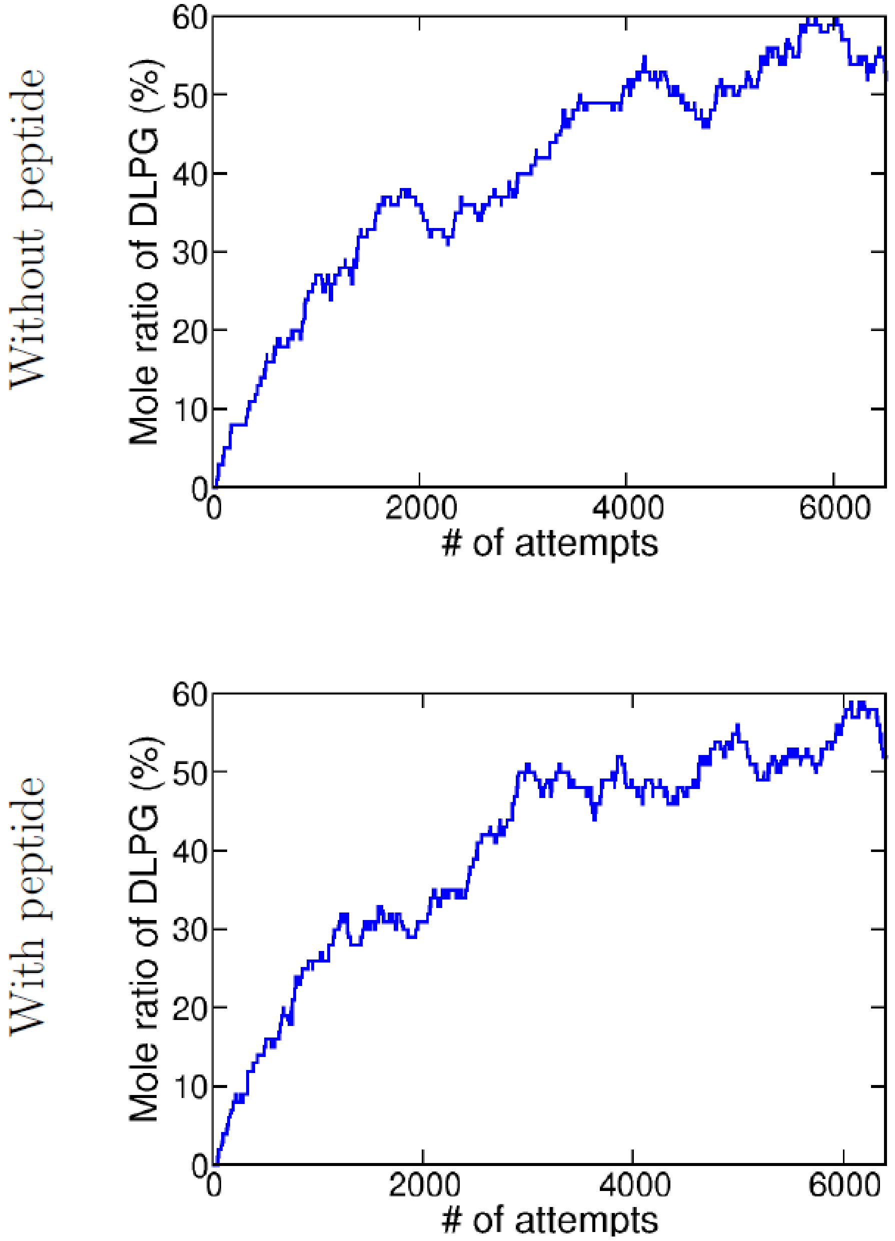


Figure 5.8: Evolution of the number of DLPG per systems. The systems are equilibrated with a reservoir at 50:50 of DLPG and DLPC (top) membrane without a peptide (bottom) membrane with a peptide. The average without peptide is ???. The average with peptide is ???.

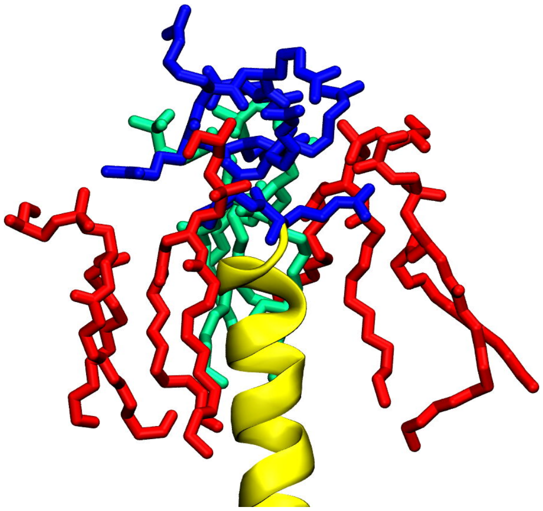


Figure 5.9: Local enrichment of DLPG in the neighborhood of the peptide resulting from the neMD/MC simulation. The lipid molecules within 2 Å of the peptide (in yellow) in the membrane equilibrated with the homogeneous reservoir at 50% DLPG mole ratio. The blue sticks represent the arginine, the DLPG are in red and the DLPC are in green.

without the peptide. It indicates that the peptide drives and stabilizes the composition of membrane.

Figure 5.9 shows the lipid distribution within 2 Å of the peptide. When the membrane is equilibrated with the homogeneous reservoir, the attempts to increase the number of DLPG are the majority, at least until a homogeneous composition of the membrane is reached. The DLPG inserted in the membrane have more time to relax before there is an attempt to get removed, so they diffuse to bind to the peptide. The specificity of protein-lipid binding is reproduced by the neMD/MC method.

5.5 Conclusion

Adopting a finite model that is representative of a complex biological lipid membrane to carry out MD simulations is often challenging. The difficulties are further heightened when considering an inhomogeneous system in which some components are present at low abundance. In practice, simulation of membranes with a fixed composition are inherently unable to adapt in response to a local perturbation. For example, the mean number of negatively charged lipids may increase in the neighborhood of a positively charged protein, but this cannot occur in a finite system if the number of those lipids is too small. The mean response to a local perturbation is associated with number fluctuations, which are ignored in a closed system with fixed composition.

To enforce equilibrium between a simulated system and an infinite surrounding bath, we designed a novel hybrid nonequilibrium molecular dynamics - Monte Carlo (neMD/MC) algorithm, in which a randomly chosen lipid molecule in the simulated system is swapped with a lipid picked in a separate system standing as a thermodynamic "reservoir" with the desired mole fraction for all lipid components. In essence, the algorithm is akin to standard experimental procedures, where the lipid composition of a sample in terms of the mole fraction is chosen deliberately.

The neMD/MC exchange algorithm is tested with a few illustrative systems with a DLPC:DLPG lipid mixture. In practice, the exchange algorithm attempts to swap the PC and PG polar head groups while retaining the conformation of the identical hydrocarbon chains. A cation picked randomly was associated with the anionic DLPG molecule to preserve charge neutrality.

The tests show that the algorithm is able to populate the simulation system in a manner consistent with the mole fraction present in the thermodynamic reservoir, and enable number fluctuations consistent with the finite size. A particular advantage of a neMD/MC formulation based on exchange between a simulated system and a reservoir is that it bypasses the need to determine the excess chemical potential of the lipid of type a and b that would be required in a Grand Canonical Monte Carlo algorithm. It is our hope that the algorithm will provide a useful methodology to generate realistic simulations of complex multi-component membranes.

In the immediate future, the neMD/MC exchange algorithm will be expanded to treat multi-component systems with variations in polar head groups and hydrocarbon chains

(Harayama and Riezman (2018); Kopitz (2017); Lorent et al. (2020); Symons et al. (2021)). One specific focus will be the phosphatidylinositol bisphosphates (PIP2) molecule, which makes up around 1% of the plasma membrane composition but is found in higher concentrations near intrinsic proteins and separate membrane domains (Van Den Bogaart et al. (2011); Mandal (2020)).

CHAPTER 6

PENTAMERIC LIGAND-GATED ION CHANNELS

The function of membrane proteins can be modulated by the properties of the surrounding lipids, either by the bulk properties of the lipid bilayer (thickness, rigidity, etc.) or by the association of specific lipids to dedicated sites at the membrane-protein interface. A typical example of membrane proteins exhibiting such a dependency is the family of Pentameric Ligang-Gated Ion Channels (pLGICs). I studied the dynamics of nicotinic acetylcholine receptors (nAChRs). I specifically focused on agonist-bound structures of different nAChRs. Using MD simulations and computational electrophysiollogy, I measured their conductances and discuss the structrual organization of nAChRs in a desensitized state. Using Alchemical transformations, I further studied the specific association of charged lipids to binding sites suggested by recent CryoEM structures and coarse-grained simulations.

6.1 Dynamics of nAChRs

6.1.1 Introduction

Transport of molecules into and out of the cells is a key physiological process. Ion channels are a specific class of membrane proteins, including Voltage Gated Ion Channel (VGIC), conveying ions in response to a transmembrane voltage (Yellen (2002)), or the Pentameric Ligang-Gated Ion Channels (pLGICs). PLGICs are neurotransmitter receptors found in a large variety of organisms. In the human being, they are found in the nervous system or at neuromuscular junctions. Upon agonist binding, they convey either cations when the neurotransmitter is the acetylcholine or serotonin, and anions when the neurotransmitter is the γ-aminobutyric acid or glycine (Salari et al. (2014)).

PLGICs are either homo- or heteropentamers formed by 5 subunits. The different stoichiometries of each subunit dictate the affinity toward a given neurotransmitter, binding

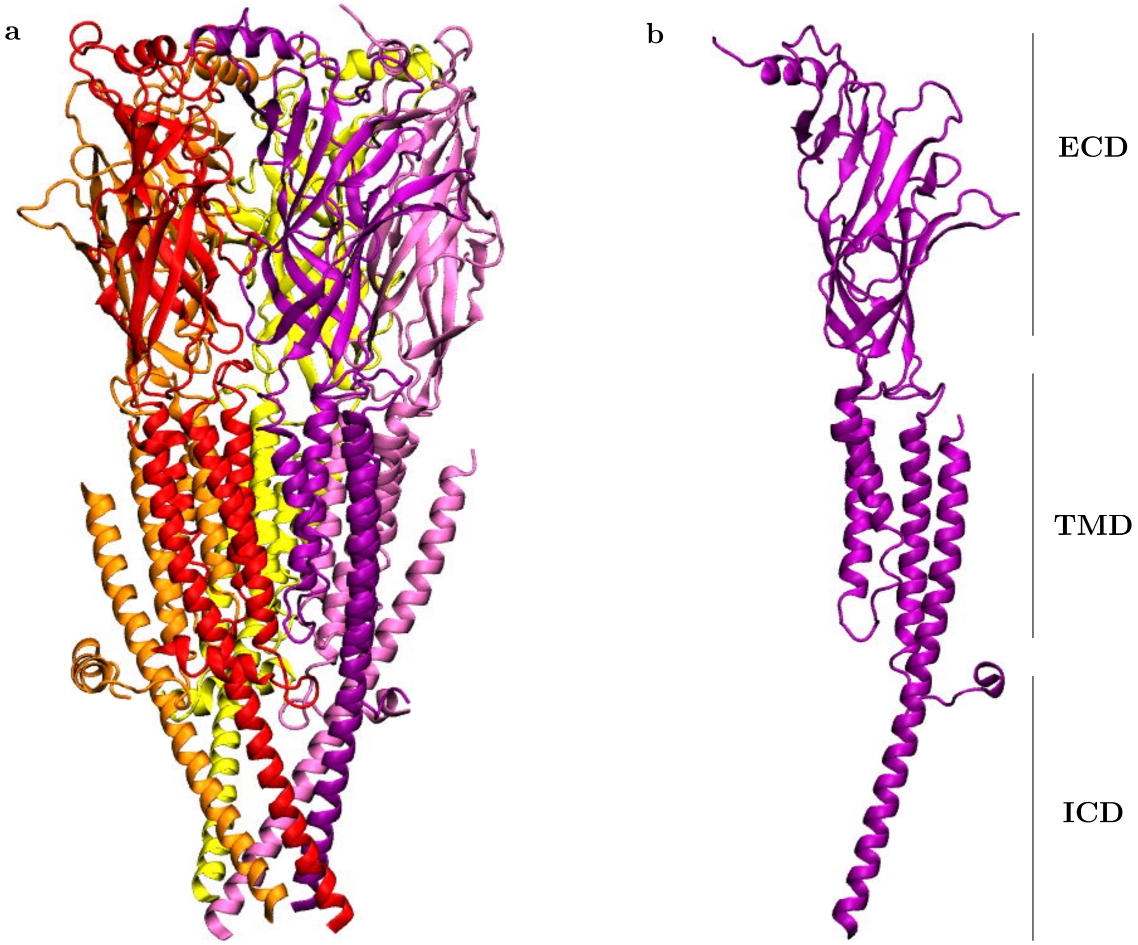


Figure 6.1: Protein $\alpha 3\beta 4$. **a**, the 5 subunits forming the channel, **b**, one subunit and the different cellular domaines.

at the interface between adjacent subunits (Albuquerque et al. (2009)). Figure 6.1 depicts a typycal pGLIC, the $\alpha 3\beta 4$ neuronal acetylcholine receptor. It is a heteropentamer, made of α and β subunits. The binding site of neurotransmitters are located in the Extra-Cellular Domain (ECD). On the other side of the membrane, the Intra-Cellular Domain (ICD) regulates the activity of the protein. The pore of the protein is located in the transmembrane domain (TMB) (Changeux and Taly (2008); Pless and Sivilotti (2018)). Electrophysiology studies show that upon agonist binding, the channel goes from a closed state to an open conformation and after a few milliseconds reaches a desensitized state (Sakmann et al. (1980); daCosta and Baenziger (2013)). Figure 6.2 shows a schematic

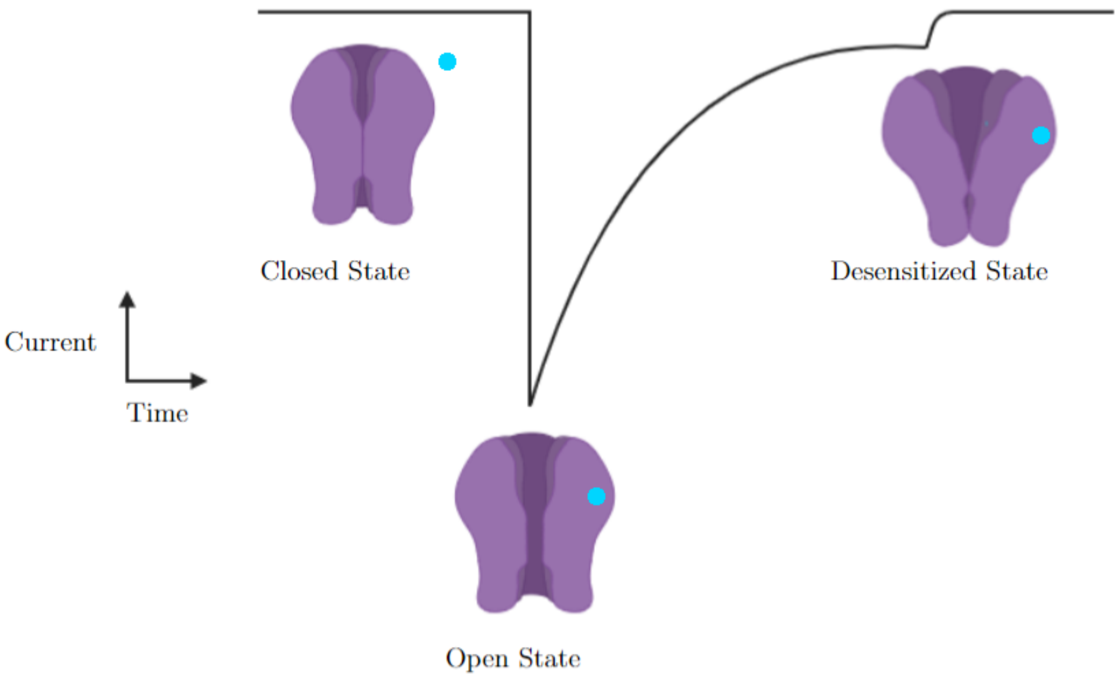


Figure 6.2: Schematic representation of the three configurations adopted by the pLGICs and the current measured by electrophysiology. The black line represents the evolution of the current induced in response to the binding of a neurotransmitter (in blue).

representation of the electrophysiology spectrum of the current depending on the three states of the channel. In the closed state, the ions do not cross the pore and therefore no current is measurable. When bound to a neurotransmitter, the pore of the channel opens and a current is established through the protein. Then, the protein reaches a desensitized state, in which the neurotransmitters are still bound to the protein but the latter does conduct ions anymore. Characterizing desensitized states as well as deciphering the overall transition cycle remain highly challenging (Basak et al. (2017)).

Morales-Perez et al. determine the structure of the desensitized human $\alpha4\beta2$ nicotinic receptor. In the latter, the pore region is non-conducting. By comparing their structure to that of of a GABA receptor obtained in desensitizing conditions, they conclude that desensitization is probably operating differently accross the pGLICs family (Morales-Perez et al. (2016)). Walsh et al. study several stoechiometries of the $\alpha4\beta2$ receptor. They show that the different interfaces between subunits change the binding affinity with the nicotine and the organisation of cholesterol around the TMD (Walsh et al. (2018)). Gharpure et al. determine the structure of $\alpha3\beta4$, another nicotinic receptor, and perform electrophysiology experiments and MD simulations on the protein. The binding site is less specific to the one of $\alpha4\beta2$, but the structure of the pore between the two receptors is similar. Yet, MD simulations on $\alpha4\beta2$ show that the pore becomes dehydrated because of a reduction of the radius, while the pore of $\alpha3\beta4$ remains hydrated (Gharpure et al. (2019)). This study shows that desensitized states have different structural properties and behaviors that the open and closed state. This is consistent with previous MD simulations on nicotinic receptors, showing that the radius of the desensitized pores are different from those of the open and closed states (Yu et al. (2019); Oliveira et al. (2019)). In another study, Zarkadas et al. run simulations on the *Torpedo* nicotinic receptor, and show that the desensitized state of the pore collapses through the simulations - the radius decreases and the pore is not hydrated anymore (Zarkadas et al. (2022)). Yu et al. compare the positional rearrangements of the domains and the radius of the pore between different structures of different receptors (nicotinic, GABA, etc). They show that not only the pore but also the conformations of the proteins (including the ECD) are different between the closed and desensitized states (Yu et al. (2019)). The reorganisation of the protein in each state is supported by a study by Nury et al. In a 1 μs long simulation of a nicotinic receptor homologue, they show that the closing mechanism of the channel is a propagating process, modifying the conformation of the whole protein (Nury et al. (2010)).

Gharpure et al. also emphasize the importance of the lipids surrounding the protein (Gharpure et al. (2019)). It has been shown that there is a higher affinity of the protein with cholesterol or anionic lipids (Thompson and Baenziger (2020); Petroff et al. (2022)). DaCosta et al. show that with specific membranes compositions, the agonist binding is uncoupled from the desensitisation (daCosta et al. (2009)). Cerdan et al. also show that the environment influences the behavior of the pores of pLGICs, and that the native environment should be used when determining the structures, especially for open states (Cerdan et al. (2018)). Zarkadas et al. showed there could also be POPC inside the pore (Zarkadas et al. (2022)). Ananchenko et al. resort to coarse Grained simulations to identify binding sites to a nAChr protein with POPC, POPA (anionic) and cholesterol. They identified binding sites for lipids at the surface of the pore, with higher affinity to POPA and cholesterol compare to POPC (Ananchenko et al. (2024)).

All of these studies show how complex it is to firmly establish what are the essential structural determinants of a desensitized state. Here, I performed a series of MD simulations of different nAChr receptors. Focusing on the dynamics of the conduction pore and its conductance, I reconcile the heterogeneity of structural data underlying desensitization.

6.1.2 Methodology

Molecular assays:

Simulations were performded for with the structures of $\alpha3\beta4$, code PDB 6PV7 (Gharpure et al. (2019)), and of $\alpha7$, codes PDB 7KOQ (Noviello et al. (2021)) and 7EKP (Zhao et al. (2021)). The structure 7EKP is elucidated in detergent, 7KOQ in a nanodisc of lipids, and they have a different resolution (2.85 Å and 3.60 Å, respectively). Using the CHARMM-GUI input generator (Lee et al. (2016)), all proteins were embbeded in a POPC:POPA: Cholesterol mixture at a ratio of 3:1:1. All systems were fully hydrated by a 22.5 Å water layer. The concentration of ions is set to 1 M, which is higher than physiological concentration in order to increase the permeation probability.

Simulations parameters:

All simulations are performed in the NPT ensemble using the program NAMD (Phillips et al. (2005)). The nonpolarizable CHARMM36 force field (Best et al. (2012)) is used for the protein, ions, and lipids. Water is described with TIP3P (Jorgensen et al. (1983)). A

Langevin thermostat and a Langevin piston are employed to maintain a temperature of 300K and a pressure of 1.0315 bar, respectively (Feller et al. (1995)). The Particle–Mesh Ewald algorithm is used to handle long–range interactions (Procacci et al. (1996)). Above 12 Å, electrostatic and Van der Waals interactions are truncated at a switching distance of 14 Å. The SHAKE/RATTLE (Ryckaert et al. (1977); Andersen (1983)) algorithm is used to constrain covalent bonds involving hydrogen atoms and the SETTLE algorithm (Miyamoto and Kollman (1992)) is utilized for water molecules. In all simulations, the Hydrogen Mass Repartitioning (HMR) scheme (Hopkins et al. (2015)) is used. The structure generated by CHARMM-GUI is initially equilibrated with restraints on the backbone for 40 ns. The restraints are progressively removed during 44 ns. Finally, a MD simulation is run without any restraints for 200 to 500 ns depending on the systems.

Computational electrophysiology:

An electric field is applied to measure the conductance of the channels. Two methods have been implemented to apply transmembrane potentials. The first one is a "charge imbalance". Two membranes are stacked, forming two solvent reservoirs. In each of them, excess charges ($+Q$ and $-Q$) are added to for the charge imbalance. The second one applies an additional force $F = q_i\mathbf{E}$, where $(q_i)_{i \in N}$ is the partial charge of the atom i. In this case, the transmembrane potential depends on the electric field $\mathbf{E}$ and the size of the box. Despite being non-periodic by construction, it is compatible with periodic boundary conditions (Gumbart et al. (2012); Kasparyan and Hub (2023)). The method used for this section is the second one. For the simulation, a electric field of 600 mV is applied. Experimental electrophysiolology usually requires a lower potential to measure the conductance of the protein, but using a higher potential increases the permeation probability in a given amount of time. Knowing the potential applied to the membrane (600 mV), and knowing the flow of ions through the pore (which is analogous to the current) gives information on the conductance, as the conductance is equal to the current divided by the

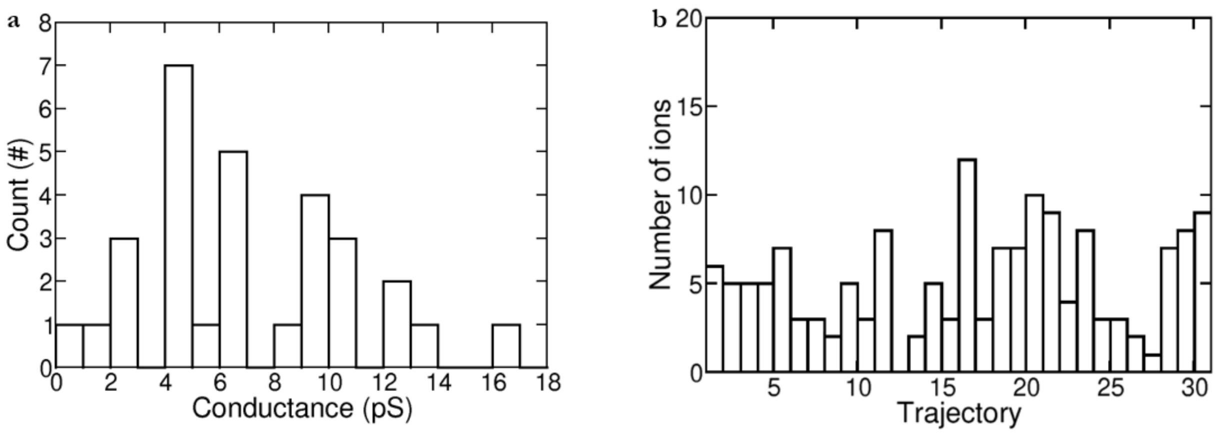


Figure 6.3: **a**: Conductance of the protein $\alpha3\beta4$ (6PV7) in presence of 600 mV electric Field. **b**: Number of ions crossing the pore in each of the 30 trajectories of 200 ns.

potential.

6.1.3 Results

Protein $\alpha3\beta4$ - PDB 6PV7:

30 simulations of the $\alpha3\beta4$ are calculated to provide a statistically meaningful estimate of the conductance of the pore. Figure 6.3 shows the histogram of the different conductance measured through the 30 trajectories of 200 ns (left) and the number of ions crossing the pore for each trajectory (right). The hydrophobic pore of $\alpha3\beta4$ is conductive in the majority of the simulations, with a maximum value of $\approx$ 17 pS and a mean value of $\approx$ 7 pS.

Figure 6.4 shows the evolution of the pore of the radius on the trajectory 1 (the numbering of the trajectories is consistent with the Figure 6.3 **b**). The figure 6.4 **a** depicts the pore hydration of the protein in its structure determined by CryoEM. Figure 6.4 **b** shows the evolution of the density of the water inside the apolar pore. Figure 6.4 **c** depicts the pore hydration of the protein after 200 ns of trajectory. Figure 6.4 **d** shows the evolution of the radius of the pore measured with the HOLE program (Smart et al. (1996)). The radius of the pore remains constant, and the pore remains hydrated through the whole simulations.

Figure 6.3 shows that even if most trajectories allow ions to cross the pore, some trajec-

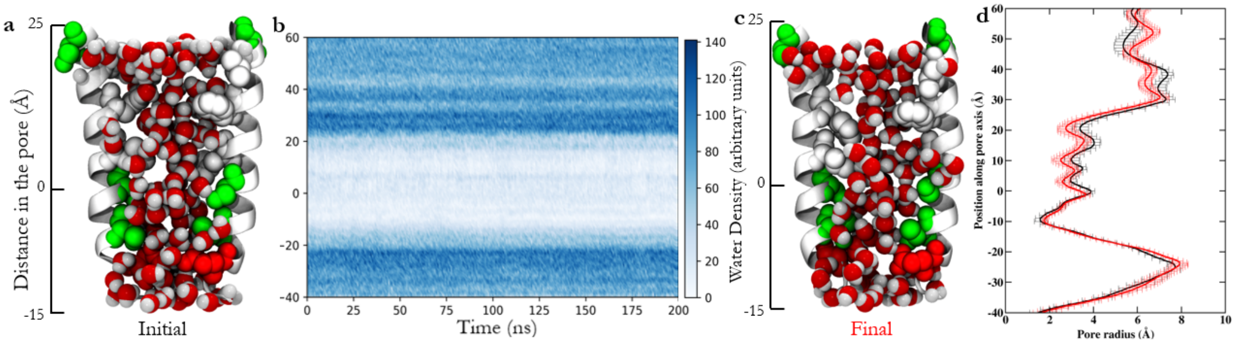


Figure 6.4: Structure of the apolar pore of $\alpha3\beta4$ (6PV7) at the initial (**a**) and final (**c**) frames of the first 200 ns long trajectory. In Van der Waals spheres representation, in green, the polar residues, in red, the acidic residues, in white the apolar residues and in the middle, in red and white, the water crossing the pore. **b**: evolution of the Water density in the pore. **d**: Evolution of the radius of the pore. In black, radius in the 20 first ns of the trajectory, in red, the radius in the last 20 ns.

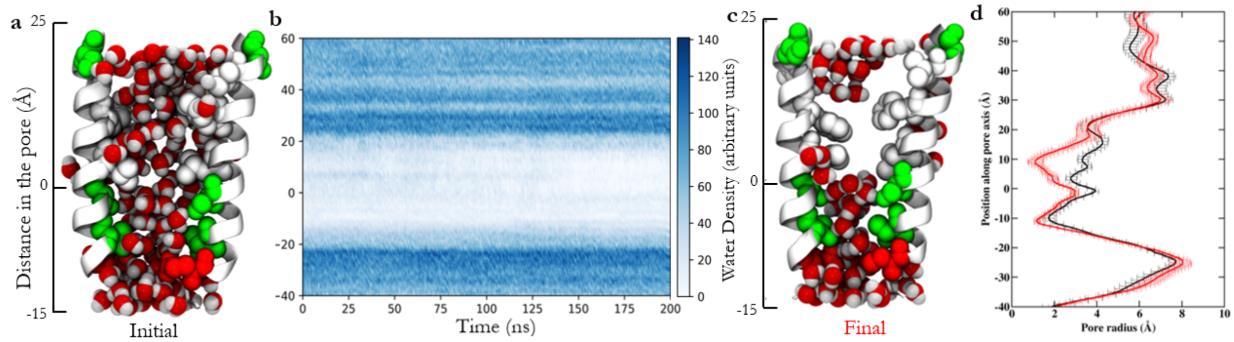


Figure 6.5: Structure of the apolar pore of $\alpha3\beta4$ (6PV7) at the initial (**a**) and final (**c**) frames of the trajectory 14. In Van der Waals spheres representation, in green, the polar residues, in red, the acidic residues, in white the apolar residues and in the middle, in red and white, the water crossing the pore. **b**: evolution of the Water density in the pore. **d**: Evolution of the radius of the pore. In black, radius in the 20 first ns of the trajectory, in red, the radius in the last 20 ns.

tories present a different behavior. For example, no permeation event is observed in the trajectory 14. Figure 6.5 shows the structure and analysis of the trajectory 14. It appears that, in this trajectory, the pore collapses – it becomes dehydrated after a decrease of the pore radius (see Figures 6.5 **b** and **d**).

Potential of Mean Force:

Potential of Mean Force (PMF) simulations have been computed by Chris Chipot to study the crossing of the pore of $\alpha3\beta4$ by an ion NA^{+}. Using Colvars (Fiorin et al. (2013)), a planar harmonic restraint ensures that the ion stays inside the pore. The PMF is computed

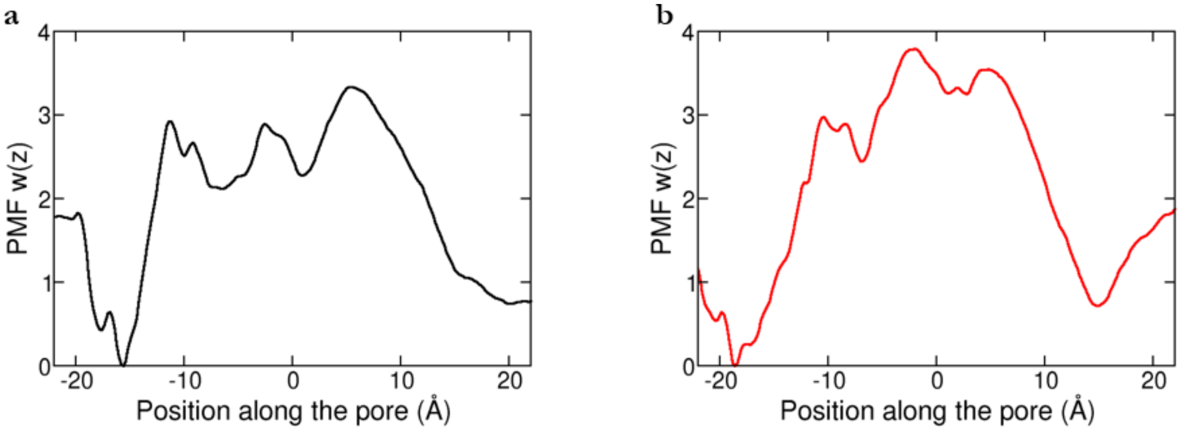


Figure 6.6: Free-energy landscape of an ion in the pore of the $\alpha3\beta4$ (**a**) and in the pore of the *Torpedo* driven into the structure of $\alpha3\beta4$ (**b**).

along z for 1.6 μs using WTM-eABF. Figure 6.6 **a** shows the free energy landscape of the ion in the pore. Even if the structure of $\alpha3\beta4$ is desensitized, the energy profile is similar to the PMF of an ion in the open conformation of the serotonin receptor 5-HT3 (Polovinkin et al. (2018)). In addition, Chris drove the structure of the protein *Torpedo* fish, previously simulated as a collapsing pore (Zarkadas et al. (2022)), into the structure of the $\alpha3\beta4$ using Targeted Molecular Dynamics. With the colvars module, the RMSD of the pore of *Torpedo* is modified to match the RMSD of $\alpha3\beta4$ with a force constant that changes through the simulation (Fiorin et al. (2013)). A PMF of the ion Na^+ crossing the pore of *Torpedo* driven in the conformation of $\alpha3\beta4$ has then been computed (see Figure 6.6 **b**). The *Torpedo* pore does not collapse anymore, and the PMF shows similarities with the PMF of the ion crossing $\alpha3\beta4$.

Protein $\alpha7$ - PDB 7EKP:

5 trajectories are generated with the 7EKP structure for 500 ns (Zhao et al. (2021)). The protein $\alpha7$ with the structure of 7EKP shows a very narrow pore in all trajectories (Figure 6.7). In the Figure 6.7 **a**, the structure of the pore is still restraint in the conformation defined by cryo-EM. In the latter conformation, the pore is narrow enough to prevent the entry of water molecules. Figures 6.7 **c** and **d** show that the pore collapses concomitantly to

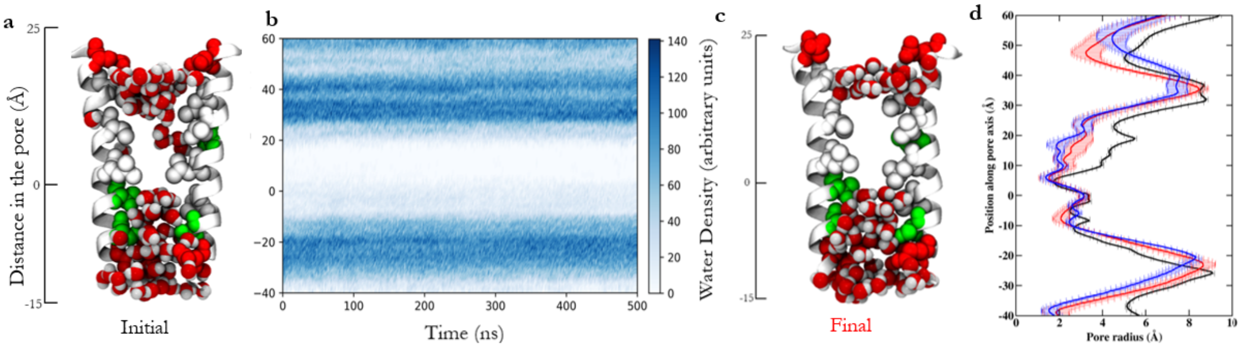


Figure 6.7: Structure of the apolar pore of $\alpha 7$ (7EKP) at the initial (**a**) and final (**c**) frames of a 500 ns long trajectory. In Van der Waals spheres representation, in green, the polar residues, in red, the acidic residues, in white the apolar residues and in the middle, in red and white, the water crossing the pore. **b**: evolution of the Water density in the pore. **d**: Evolution of the radius of the pore. In black, radius in the 20 first ns of the trajectory with the restraints, in red, the radius in the first 20 ns after the release of the structure, in blue the radius in the last 20 ns.

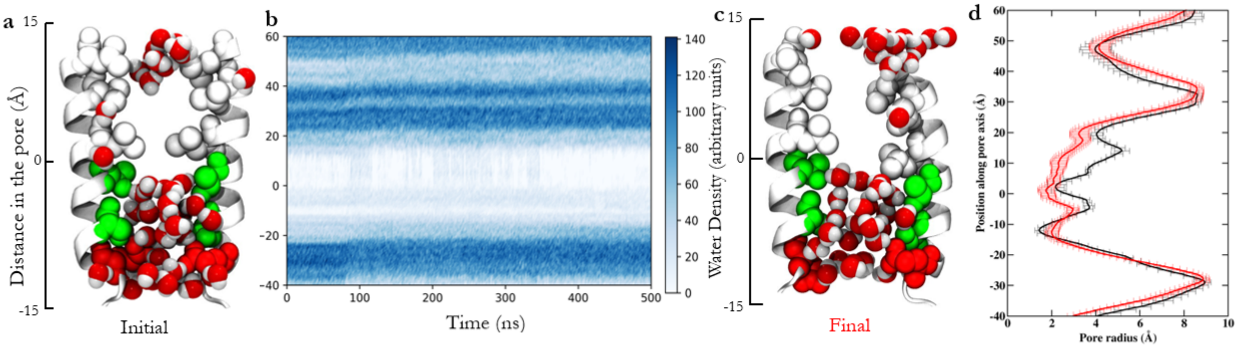


Figure 6.8: Structure of the apolar pore of $\alpha 7$ (7KOQ) at the initial (**a**) and final (**c**) frames of the trajectory 1. In Van der Waals spheres representation, in green, the polar residues, in red, the acidic residues, in white the apolar residues and in the middle, in red and white, the water crossing the pore. **b**: evolution of the Water density in the pore. **d**: Evolution of the radius of the pore. In black, radius in the 20 first ns of the trajectory, in red, the radius in the last 20 ns.

the release of the restraints.

Protein $\alpha 7$ - PDB 7KOQ:

5 trajectories are generated with the structure 7KOQ for 500 ns (Noviello et al. (2021)). Two kinds of behaviour of the pore have been observed, as shown on Figures 6.8 and 6.9.

Figure 6.8 shows a collapse of the pore, with a low water density in the pore (Figure 6.8 **b**), and a decrease of the radius of the pore through the trajectory (Figure 6.8 **d**). Figure 6.9, on the other side, shows a hydrated pore. The initial structures (Figures 6.8 **a** and 6.9

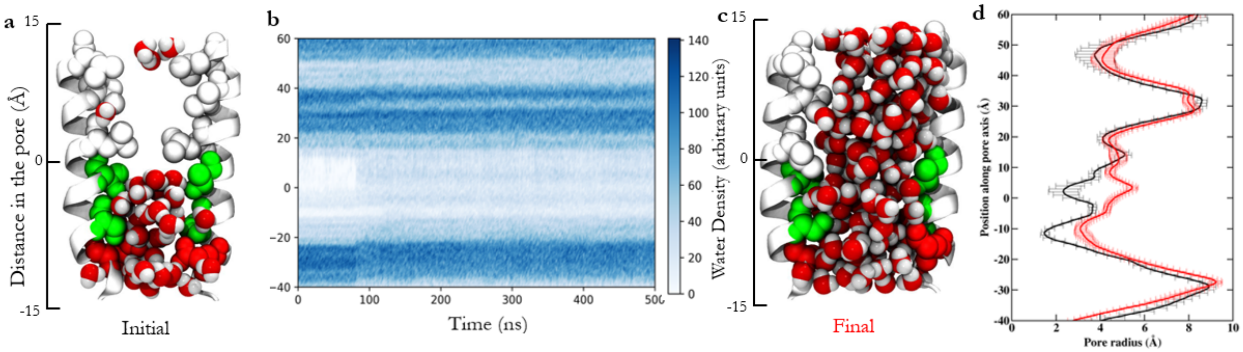


Figure 6.9: Structure of the apolar pore of $\alpha7$ (7KOQ) at the initial (**a**) and final (**c**) frames of the trajectory 2. In Van der Waals spheres representation, in green, the polar residues, in red, the acidic residues, in white the apolar residues and in the middle, in red and white, the water crossing the pore. **b**: evolution of the Water density in the pore. **d**: Evolution of the radius of the pore. In black, radius in the 20 first ns of the trajectory with the restraints, in red, the radius in the first 20 ns after the release of the structure, in blue the radius in the last 20 ns.

a) are the cryo-EM structure, with a low pore hydration. Yet, in the Figure 6.9 **b**, the water density in the pore increases after releasing the restraints, emphasized in the Figure 6.9 **c** and in the Figure 6.9 **d**, which shows that the radius is larger at the the simulation.

Figure 6.10 shows a superimposition of the transmembrane helices lining the hydrophobic pore in two structures 7EKP and 7KOQ. The structures are close, but not exactly the same. The RMSD of the backbone of the pore of 7EKP compared to 7KOQ is 0.8. This small displacement may explain the difference of dynamics. It has been previously observed in sodium channel pores that a small difference in the position of the residues of two similar proteins may propagate to induce major changes in the dynamics of the pore (McCusker et al. (2012)).

6.1.4 Conclusion and discussion

The dynamics and the stability of the pore of several desensitized structures of nicotinic receptors are being studied. The pore of the $\alpha3\beta4$ protein remains hydrated in the majority of the simulations, consistently with previous simulations on this structure (Gharpure et al. (2019)). Trajectories of the protein $\alpha7$ with the initial structure corresponding to the PDB

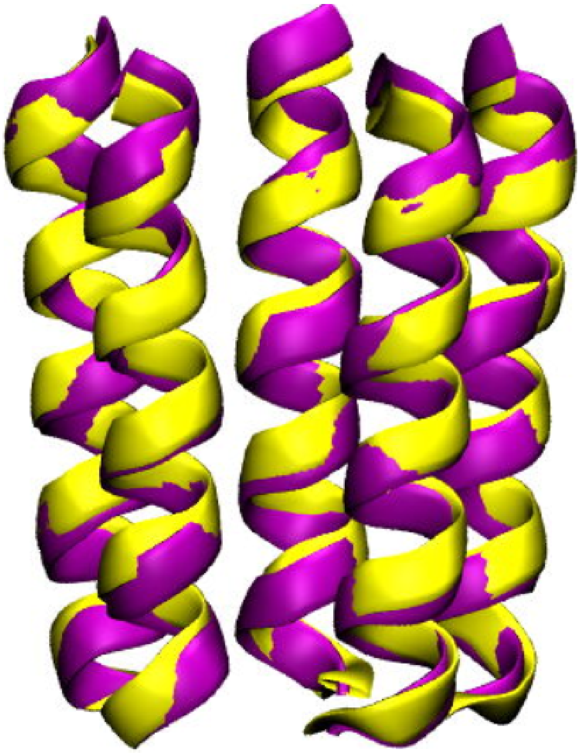


Figure 6.10: Comparison of the secondary structure apolar core of the structure 7EKP (yellow) and 7KOQ (purple).

ID 7EKP indicate that the pore gets dehydrated as its radius decreases. The collapsing of desensitized pLGICs pores have also been previously observed in several studies (Morales-Perez et al. (2016); Zarkadas et al. (2022)). Another structure of this protein, (code PDB 7KOQ) shows both behavior: the pore remains hydrated or collapses, depending on the simulations. The two structures have a small difference in the definition of the structure of the pore, and it seems to induce major changes in the dynamics. The PMF calculations, performed by Chris Chipot, support that structural variations induce behavior differences. Driving the structure of the pore of the *Torpedo* fish (collapsing in its Cryo-EM structure) into the structure of $\alpha3\beta4$ seems to induce a stability and a conductivity to the *Torpedo* pore.

Overall, it seems that the desensitized state covers an ensemble of different states, with close structures, but very different dynamics and behaviors. The experimental methods to obtain the structures of the desensitized structures are likely to affect the conformations of the protein. Most Cryo-EM structures are obtained from proteins embedded in a nanodisc. Cryo-EM maps also show some diffuse density in the pore. Zarkadas et al. showed that adding lipids inside the pore matches the density and prevents the pore from collapsing (Zarkadas et al. (2022)).

6.2 Specificity of the protein-lipid binding

In this section, I present a collaborative work with Anna Ananchenko (University of Ottawa) aimed at quantifying the relative affinity of a nicotinic receptor for different types of lipids.

6.2.1 Introduction

The previous section presented study of the dynamics of desensitized states of nicotinic receptor. This section focuses on another aspect of the pLGICs: the protein-lipid binding specificity. The lipid distribution surrounding the TMD influences the dynamics of the protein, by stabilizing a state over another. DaCosta and Baenziger showed that in a pure membrane of PC lipids (zwiterrionic), the opening mechanism of the nAChR is not responsive to agonist binding, and that anionic lipids stabilize a resting configuration proned to open upon binding. The difference in the size of polar heads also creates a curvature in the membrane that could favor protein coupling (daCosta et al. (2009)).

DaCosta et al, and Thompson and Baenzinger show the pLGIs have a higher affinity for anionic lipids and cholesterol, than for zwiterrionic lipids (daCosta et al. (2009); Thompson and Baenziger (2020)). Walsh et al. show that the binding affinity the cholesterol depends on the subunits composing the pentamers (Walsh et al. (2018)). Ananchenko et al. study the interactions between the *Torpedo* nAChr and a lipid bilayer composed of charged, zwiterrionic and cholesterol. After performing 30 μs long CG trajectories, they analyze the lipid density around the protein and identify dedicated binding sites. Figure 6.11 shows the lipid density for different membrane composition. In a pure membrane of POPC, lipids bind at specific site located between adjacent subunits. In a membrane with POPC and POPA, the binding sites are mostly occupied mostly by POPA. In a membrane with POPC and cholesterol, the cholesterol also bind favorably, but at different binding sites than POPC. Finally, in a membrane with the three lipids POPC, POPA and cholesterol, the binding sites are mostly populated either by POPA or cholesterol. The POPA have three favored binding sites (at

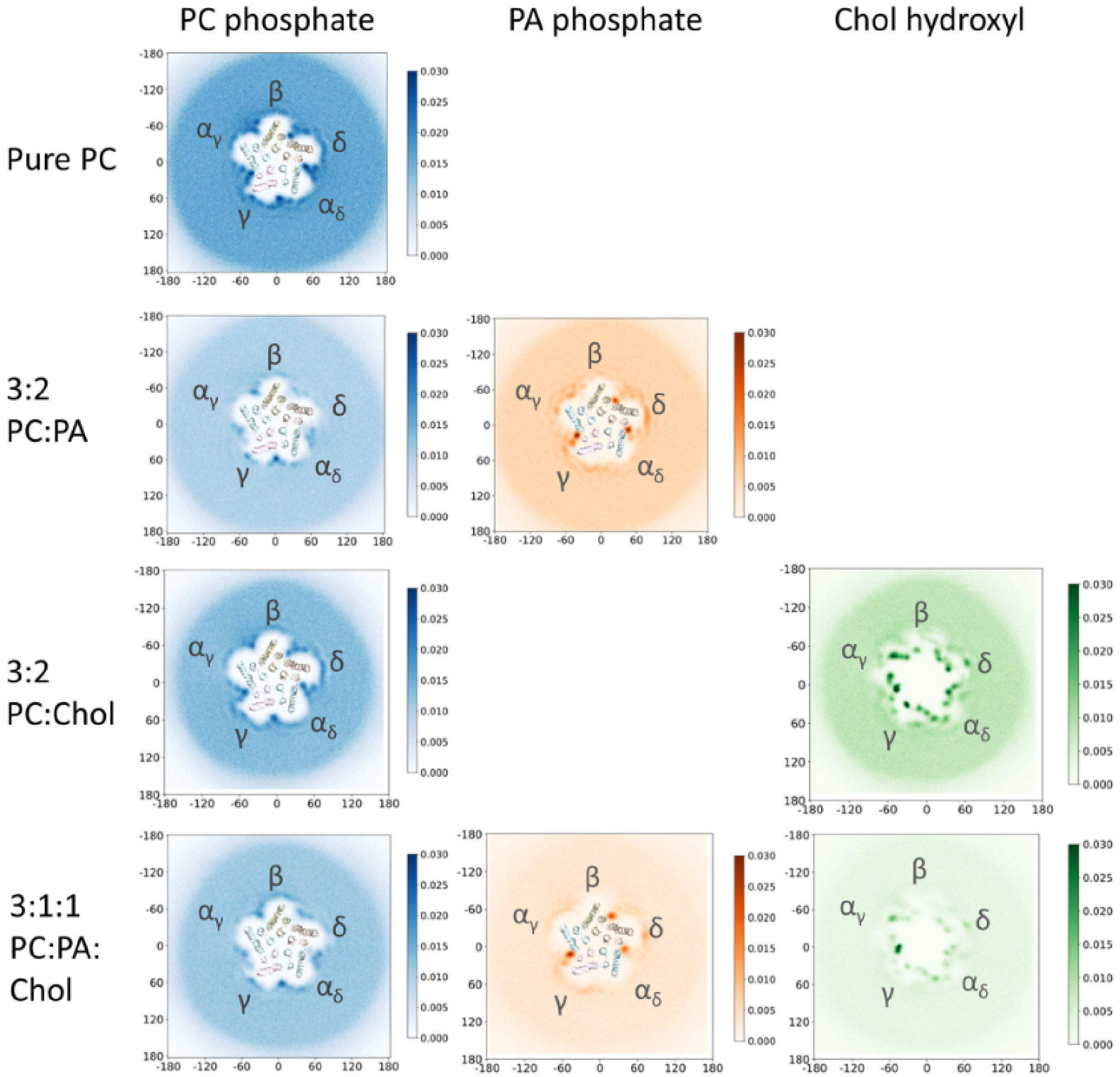


Figure 6.11: Top-down 2D headgroup density plots for PC, PA and Chol in the inner leaflet from simulations of the apo nAChR. Each density plot represents the lipid headgroup densities averaged over 3×30 μs CG-MD trajectories (Ananchenko et al. (2024)).

the interfaces of $\alpha_\gamma - \gamma$, $\alpha_\delta - \delta$ and $\beta - \delta$). The POPC density at the $\gamma - \alpha_\delta$ interface is higher even when there are other lipids in the membrane (Ananchenko et al. (2024)).

Petroff et al. quantified the different affinities between a channel protein ELIC and PC, PE and PS lipids, resorting to Streamlined Alchemical Free Energy Perturbation (SAFEP). To ensure the reversibility and enhance the convergence of the binding free energy simulations, restraints on the ligand are added. Instead of using the set of restraints previoulsy defined in the Introduction and the Chapter 2 (on the RMSD and on the five angles translating the relative orientation), the SAFEP method resorts to a single Distance-to-Bound Configuration (DBC) coordinate (Petroff et al. (2022); Santiago McRae et al. (2023)). Using SAFEP, they computed the variations in the binding affinity when lipids are mutated in the membrane or at the protein-lipid interface. The $\Delta\Delta G$ calculated for POPE to POPC (both zwiterrionic lipids, but POPE has a smaller polar head) is around 4 kcal/mol. The $\Delta\Delta G$ calculated for POPE to POPG (zwiterrionic to anionic lipid) is around -2 kcal/mol, consistently with the higher affinity of the protein for anionic lipids (Petroff et al. (2022)).

In this section, we aimed at quantifying the affinitiy of the *Torpedo* nACHR for POPC and POPA at binding sites identified by Ananchenko et al. using alchemical transformations (Ananchenko et al. (2024)). This project is still ongoing.

6.2.2 Methodology

Protein:

The protein-lipid interactions are studied for the *Torpedo* nAChr (code PDB: 7QL5 (Zarkadas et al. (2022))), with a membrane composed of 3:2 POPC:POPA. Here we only focus on the binding site at the $\alpha_\gamma - \gamma$ subunits interface, mostly populated by POPA (see Figure 6.11) (Ananchenko et al. (2024)).

The software CG2AT2 is used (Vickery and Stansfeld (2021)) to generate the initial conformation of the alchemical simulation. It generates the corresponding all-atoms represen-

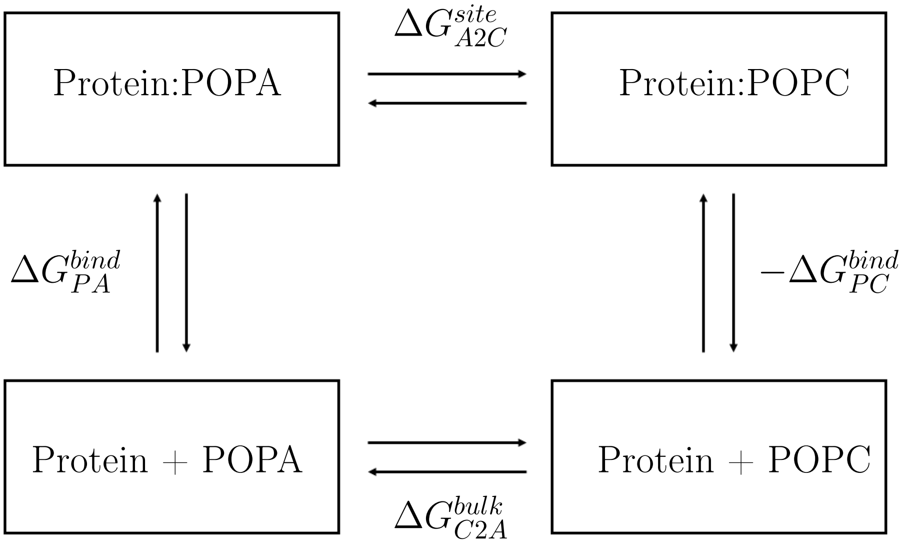


Figure 6.12: Thermodynamic cycle for the mutation. In the first line, the lipid is in the binding site. In the second line, the lipid is in the bulk.

tation from a CG structure taken from one of the 30 μs trajectory, which is then thermalized for 250 ns.

Alchemical simulations:

The thermodynamic cycle depicted Figure 6.12 is used to compute the difference of binding free energy between a POPA and a POPC. The upper line in the cycle represents an alchemical transformation of a POPA to a POPC at the binding site and the lower line stands for the reverse transformation in the bulk. The difference of binding free energy is:

$$\Delta\Delta G = \Delta G^{site}_{A2C} + \Delta G^{bluk}_{C2A} = \Delta G^{bind}_{PC} - \Delta G^{bind}_{PA} \tag{6.1}$$

When restraining the lipid in the binding site to lower the configurational entropy during the transformation, the difference in free energy associated with the mutation ΔG^{site}_{A2C} (first line of Figure 6.12) is computed in three steps (see Figure 6.13). First, a restraint on the RMSD of the phosphate group is added to the POPA. Next, the restrained POPA in the binding site is mutated into the POPC. Finally, the restraint is removed. The contribution

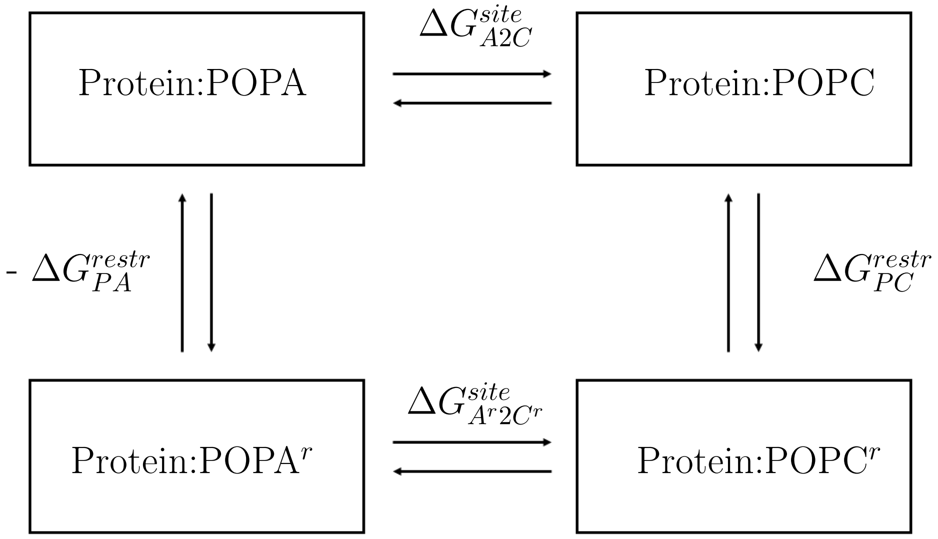


Figure 6.13: Thermodynamic cycle for the addition of a restraint during the mutation. The exponent r indicates if the lipid is restraint.

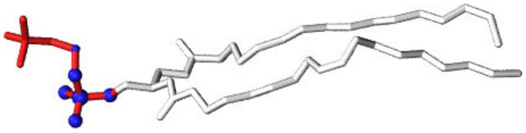


Figure 6.14: Dual topology for the mutation. In white, the carbon tail, common between the two structures. In red, the POPC polar head. In blue, the POPA polar head.

of the restraint to the free energy is evaluated using Thermodynamic Integration (TI). According to the cycle, the value of ΔG^{site}_{A2C} corresponds to:

$$\Delta G^{site}_{A2C} = \Delta G^{restr}_{PA} + \Delta G^{site}_{A^r2C^r} - \Delta G^{restr}_{PC} \tag{6.2}$$

To mutate one POPC into a POPA, a dual topology is used (Figure 6.14). The common carbon tail remains unchanged during the simulation and the polar heads (including the phosphate group) are mutated.

Computational details:

Two simulations are computed separately: a mutation of POPA in POPC in the binding site, and a mutation of POPC in POPA in a bulk membrane. The molecular system for the mutation in the binding site is 317,147 atoms, with 78,756 water molecules. The total box is 121 x 121 x 214 Å. For the mutation in the bulk membrane, a membrane with the 3:2 ratio of POPC:POPA is created. It consists in a 52,301 atoms patch, with 8,767 water molecules. The total box is 82 x 82 x 75 Å.

Both trajectories are generated using 50 windows, with a $\Delta\lambda = 0.02$ between each window. Each trajectory is 100 ns long. The timestep is 4 fs. The force field used is a CHARMM36 (Best et al. (2012)), the simulations are run using the NAMD program (Phillips et al. (2005)). A Langevin thermostat and a Langevin piston are employed to maintain a temperature of 300K and a pressure of 1.0315 bar, respectively (Feller et al. (1995)).

Restraints:

Different sets of restraints have been considered. Based on a restraint similar to the single Distance-to-Bound Configuration coordinate introduced in the SAFEP method (Santiago McRae et al. (2023)), a harmonic restraint is applied to maintain the lipid in the binding site and ensuring that no other lipid can access the site during the transformation process. The restraint is applied on the distance between the lipid and the site in a way that does not affect the dynamics of the end-states.

Another strategy consists in restraining the RMSD of the phosphate group of the mutated lipid. The contribution to the free energy of the latter restraint is evaluated using Thermodynamic Integration (TI).

6.2.3 Results

Mutations with a restraint to maintain the lipid in the binding site:

In this test, the lipid is maintained in the binding site by a harmonic restraint.

Figure 6.15 shows the free energy variation associated with the mutations in the bulk

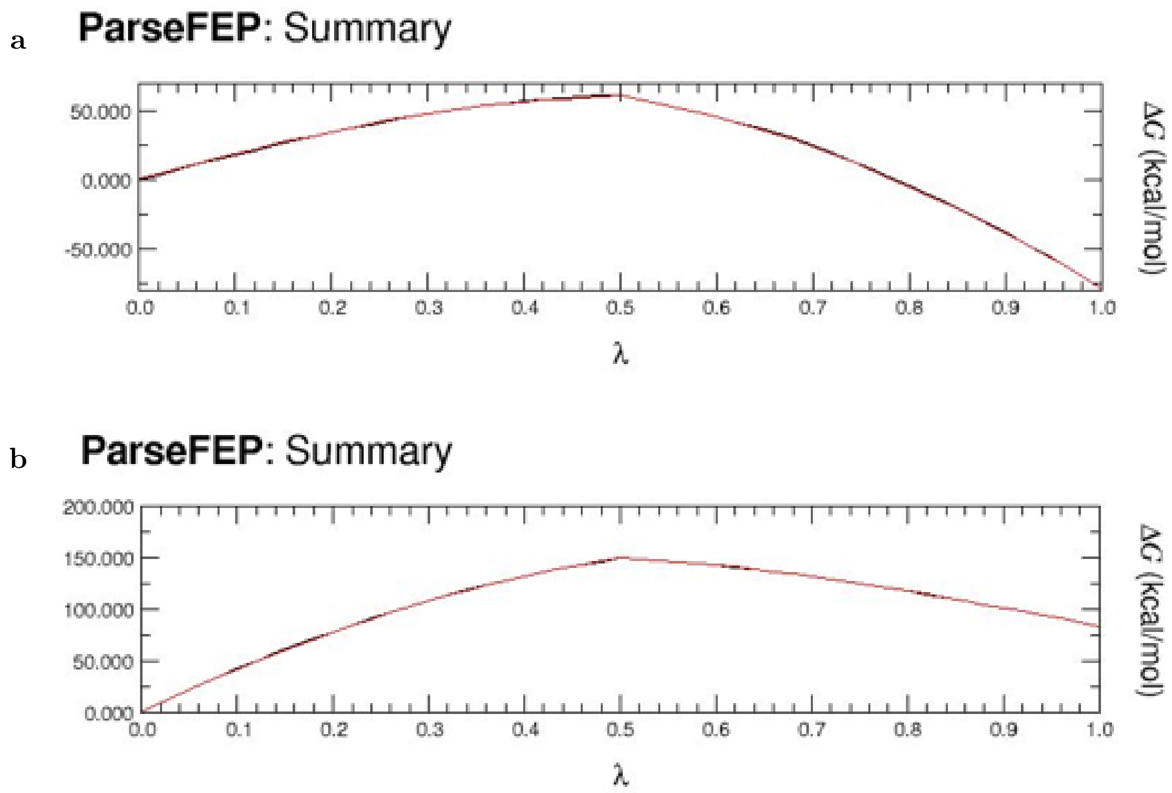


Figure 6.15: Free energy change associated with the mutations in the bulk (**a**) and in the binding site (**b**). These curves have been generated using the Plugin ParseFEP (Liu et al. (2012))

(Figure 6.15 **a**) and in the binding site (Figure 6.15 **b**). The hysteresis between the forward and the backward trajectories are in a ±1 kcal/mol range. The perturbation associated with the mutation in the binding site is $\Delta G^{site}_{A2C} = 82.75$ kcal/mol, and the perturbation associated with the mutation in the bulk is $\Delta G^{bulk}_{C2A} = -80.33$ kcal/mol, leading for a relative affinity to:

$$\Delta\Delta G = \Delta G^{site}_{A2C} + \Delta G^{bluk}_{C2A} = 2.4\text{kcal/mol} \tag{6.3}$$

Mutations with a restraint on the RMSD of the phosphate group:

A restraint is added on the RMSD of the phosphate group of the lipid in the binding site. The cycle for the mutation in the binding site is the one described Figure 6.13. These simulations are not converged as the thesis is being written, so the results are not going to be discussed.

6.2.4 Conclusion and discussion

The binding specificity of nicotinic receptors with lipids is being explored. Alchemical transformations are computed to compare the affinity of the receptor for an anionic lipid with the affinity for a zwitterionic lipid. A bound POPA is mutated into a POPC, and a POPC in the bulk membrane is mutated into a POPA. Several strategies have been explored. All the transformations are not converged, so no conclusion on the methodology can be drawn yet.

The transformation with a restraint to maintain the lipid in the binding site indicates that the difference of affinity between an anionic and a zwitterionic lipid in the binding site of nAChR is $\Delta\Delta G = 2.4$ kcal/mol. This value is positive, which is consistent with a preferential binding with POPA.

The alchemical transformation consists in mutating a zwitterionic lipid into an anionic one (or the other way around). The net charge of the system is not conserved. When the charge of the system is not kept neutral, artefacts in the calculations of energies are observed (Hub et al. (2014)). To correct these artefacts, several routes have been tested (Papadourakis et al. (2023)). Doing the two mutations (zwitterionic to anionic and anionic to zwitterionic) in an unique simulation keeps the charge constant, but requires a large patch (Rashid et al. (2013)). Additional terms can *a posteriori* be computed to correct the binding free energy (Rocklin et al. (2013)). These terms are yet not so trivial to evaluate in membrane (Wu and Biggin (2022)). Adding or removing a counterion is also a strategy to keep the net charge constant and limit the computational errors (Buslaev et al. (2022)). Finally, it is also possible to neglect these artefacts, as the induced errors decrease with the size of the box (Simonson and Roux (2016); Radak et al. (2017); Papadourakis et al. (2023)). We are currently testing a strategy where an ion Cl^- is inserted when POPC is coupled and removed when POPA is coupled to keep the charge constant. The goal of this test is to ensure that no error is being induced by the non conservation of the net charge.

CHAPTER 7

CONCLUSION AND DISCUSSION

This PhD thesis focuses on the study of biomolecular complexes by means of theoretical approaches. I resorted to equilibrium Molecular Dynamics simulations to compute binding free energies and to study the dynamics of proteins. I contributed to the development of a protocol for the Binding Free Energy Estimator 2 plugin (Fu et al. (2022)). This plugin is an automatized tool to setup and analyze binding free energy calculations for protein-ligand and protein-protein complexes. I essentially focused on setting up part of the protocol aimed at computing free energy of associations of ligands binding buried in membrane protein. The plugin was adapted to account for a semi-istotropic environment such as a lipid bilayer. Special care was given to the alchemical transformation strategy to ensure the reversibility of the calculations by using a proper set of restraints, and by carefully sampling the hydration of the buried binding site.

I further employed alchemical transformations to quantify the effect of point mutations on the stability of complexes involving two proteins. I specifically studied two families of proteins involved in morphogenesis. Proteins of each family interact specifically with each others to guide the formation of the synaptic network. I focused on mutations at the interface of two complexes, the cognate Dpr6-DIPα and the non-cognate Dpr6-DIPγ. With coworkers, we also compared several computational methods to see how fast and accurate each of them are to investigate the binding selectivity of a large family of complexes. Alchemical transformations are accurate when performing mutations, but are too slow to be used in high-thoughput calculations. The mutations at the interface of a cognate complex (Dpr6-DIPα) and of a non-cognate complex (Dpr6-DIPγ) pointed out a few residues that are important in the selectivity mechanism. For example, the perturbations in binding free energy associated with the mutation of the residues I114 or Y123 of Dpr6 in alanine are much higher in the cognate complex, indicating a more favorable interaction with the corresponding DIP protein. Poisson-Boltzmann calculations are much cheaper and faster, but are not as precise and require additional correction terms to produce results consistent with experiments. LDA-AIMS method is fast, computationally cheap and yield to a distinguishability between cognate and non-cognate of nearly 0.8, but does not not accurately estimate the perturbation associated with mutations at the interface. Goulard Coderc de Lacam et al. developed Machine Learning models to distinguish between cognate and non-cognate complexes. They also identified key residues evolved in the binding selectivity, assessing the efficiency of Machine Learning to study specific protein:protein complexes (Goulard Coderc De Lacam et al. (2024)).

MD simulations do not extensively sample all configurations of complex biomolecular systems when the associated free energy landscape is rugged. For example, the reorganisation of lipid distribution in an inhomogeneous membrane is limited by the slow lateral diffusion of the lipid molecules. The use of nonequilibrium simulations can therefore be an alternative to sample the large configurational changes. The work associated with a nonequilibrium transformation can be used in a hybrid Monte Caro method, to sample the space of configurations differently than in MD. I developed a hybrid nonequilibrium Molecular Dynamics-Monte Carlo (neMD/MC) methodology to enhance the sampling of lipid distribution in the membrane. Lipids are exchanged two by two in a membrane using alchemical simulations. Based on a Metropolis Monte Carlo criteria, exchanges are done until the thermodynamic equilibrium distribution is reached (Nilmeier et al. (2011); Chen and Roux (2014, 2015c); Radak and Roux (2016)). I focused on sampling exchanges between lipid bearing different charges, a challenge never addressed in the literature. The developed method successfully reproduces lipid distribution of lipid mixture and in the surrounding of a model transmembrane peptide (Szczepaniak et al. (2024)). Several modifications to the method were developed to make it more efficient. In the context of sampling lipids around a membrane protein, the choice of lipids proposed for exchange can be biased, such that at least one lipid is in the neighborhood of the protein. I extended the neMD/MC algorthm to lipid exchanges between two bilayer systems evolving separetly. For example, the lipids phosphatidylinositol bisphosphates (PIP2) represent approximatively 1 % of the plasma membrane composition, but are found at higher concentration around membrane proteins. A membrane of several hundreds lipids would be necessary to reproduce the low concentration and the local enrichment of PIP2.

The neMD/MC algorithm has been employed for sampling a lipid mixture composed of two lipids differing solely through their headgroups. Extension to any other headgroup is straightforward and requires only the construction of the corresponding dual-topologies. The neMD/MC method is theoretically extensible to exchange all kind of lipids, but in practice, the work associated with the transformation of very different lipids has to be negative enough to accept exchanges. Cholesterol, phosphatidylinositol phosphate lipids (PIP) or cardiolipins are are known to interact with membrane proteins (Van Den Bogaart et al. (2011); Mandal (2020); Petroff et al. (2022); Ananchenko et al. (2024)), but they are structurally very different from a phospholipid. The efficiency of neMC/MC is closely related to the constraints used on the common atoms between the similar lipids. Exchanging phospholipids with cholesterol, PIP2 or cardiolipins would require a drastic modification of our strategy. For example, it would be interesting to try to exchange one cardiolipin with two phospholipids to decrease the sterical perturbation.

I studied the dynamics of nicotinic acetylcholine receptors (nAChRs), proteins of the family of pentameric Ligand-Gated Ion Channels (pGLICs). PLGICs are neurotransmitter receptors found in the nervous system or at neuromuscular junctions. Without neurotransmitter, pLCGIC are in a closed conformation and do not convey molecules into or out of the cell. Upon agonist binding, the conformation of the protein changes, leading to a pore opening. After a long exposure to agonist, the protein reaches a desensitized state, which is no longer conductive. Studies on the desensitized states of nicotinic receptors show that different structure have very different dynamics. The pore can either be hydrated (Gharpure et al. (2019)), or collapse. MD studies also suggested that diffuse densities at the level of the conductance pore observed in Cryo-EM could associated to lipids (Zarkadas et al. (2022)). MD simulations have been conducted for three structures of desensitized nicotinic receptors: one of $\alpha3\beta4$ and two of $\alpha7$, obtained in detergent and in lipid nanodisc. 30 trajectories of 200 ns have been generated for $\alpha3\beta4$. The majority of the simulations of $\alpha3\beta4$ show a stable pore, that remains hydrated, consistently with previous simulations (Gharpure et al. (2019)). 5 trajectories of 500 ns have been generated on each structure of $\alpha7$. The pore of one structure collapses in every simulations, the other one collapses in 3 simulations or remains open and hydrated in the 2 others. The two PDB used for $\alpha7$ have small structural variations, inducing differences in the dynamics and behavior of the pore. PMF calculations, performed by Chris Chipot, support that structural variations induce behavior differences. Driving the structure of the pore of the nicotinic receptor of the *Torpedo* Fish (collapsing in its Cryo-EM structure) into the structure of the pore of $\alpha3\beta4$ induces the stability of the radius of the pore and the capacity to conduct ions.

Previous studies showed that the function of PLGICs is modulated by the composition of the membrane. For example, pure POPC membrane decouples the agonist binding from the opening mechanism (daCosta et al. (2009)). Anachenko et al. showed using CG simulations of the desensitized state of the nAChR of the *Torpedo* fish in different membrane compositions that lipids bind at very specific site on the protein surface. The observed sites can bind with POPC in a pure POPC membrane but mostly recognize the charged POPA lipid in POPC:POPA and POPC:POPA:cholesterol mixtures (Ananchenko et al. (2024)). In collaboration with Anna Ananchenko, I contributed to alchemical transformations to quantify the difference of binding affinity of the nAChR of the *Torpedo* fish for POPA compared to POPC. A bound POPA is mutated into a POPC, and a POPC in a bulk membrane is mutated into a POPA. The difference in affinity for the binding site locate at the $\alpha_\gamma - \gamma$ subunits interface is $\Delta\Delta G = \Delta G^{site}_{A2C} + \Delta G^{bluk}_{C2A} = 2.4\mathrm{kcal/mol}$, which is consistent with the higher affinity of the protein to POPA. Additional strategies are being explored, to evaluate the efficiency and the accuracy of the method. The objective is to establish a rigorous method to determine protein:lipid binding affinity. Preliminary results indicate that our method enables the quantification of the binding free energy perturbation associated with the mutation of POPA in POPC with a high accuracy. It would be interesting to study the difference of affinity for POPA and POPC in the other binding sites, or in structures of the open and closed states of the protein.

The optimization of the route to perform alchemical simulations was a major aspect of my Ph.D. In equilibrium and nonequilibirium simulations, adding restraints lower the number of degrees of freedom, reducing thereby the configurational entropy to be sampled along the transformation. Use of restraints was also key for exchanging lipids in membrane with our neMD/MC algorithm. In the future, it would be interesting to apply the methodologies developed in this thesis to explore the modulation of the function of membrane proteins by the membrane. The neMD/MC method can generate a more complex and realistic configuration of the lipids surrounding the protein. The method to compute lipid:protein affinity can refine the model by precisely describing the lipids in the specific binding sites. The response of membrane proteins to ligands, for example in the context of drug design, could be more accurately predicted by computational methods, as they would be more efficient to model realistic biological phenomena.

REFERENCES

Abascal, J. L. F. and Vega, C. (2005). A general purpose model for the condensed phases of water: TIP4P/2005. *The Journal of Chemical Physics*, 123(23):234505.

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

Abrams, C. and Bussi, G. (2013). Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy*, 16(1):163–199.

Adachi, K., Oiwa, K., Yoshida, M., Nishizaka, T., and Kinosita, K. (2012). Controlled rotation of the F1-ATPase reveals differential and continuous binding changes for ATP synthesis. *Nature Communications*, 3(1):1022.

Adelusi, T. I., Bolaji, O. Q., Ojo, T. O., Adegun, I. P., and Adebodun, S. (2023). Molecular Mechanics with Generalized Born Surface Area (MMGBSA) Calculations and Docking Studies Unravel some Antimalarial Compounds Using Heme O Synthase as Therapeutic Target. *ChemistrySelect*, 8(48):e202303686.

Akash, S., Bibi, S., Yousafi, Q., Ihsan, A., Mustafa, R., Farooq, U., Kabra, A., Alanazi, M. M., Alanazi, A. S., and Al Kamaly, O. (2023). Ligand-based drug design of Pinocembrin derivatives against Monkey-Pox disease. *Arabian Journal of Chemistry*, 16(11):105241.

Alberts, B., Johnson, A., and Lewis, J. (2002). *Molecular Biology of the Cell. 4th edition.* Garland Science.

Albuquerque, E. X., Pereira, E. F. R., Alkondon, M., and Rogers, S. W. (2009). Mammalian Nicotinic Acetylcholine Receptors: From Structure to Function. *Physiological Reviews*, 89(1):73–120.

Alder, B. J. and Wainwright, T. E. (1957). Phase Transition for a Hard Sphere System. *The Journal of Chemical Physics*, 27(5):1208–1209.

Ananchenko, A., Gao, R. Y., Dehez, F., and Baenziger, J. E. (2024). State-dependent binding of cholesterol and an anionic lipid to the muscle-type Torpedo nicotinic acetylcholine receptor. *Communications Biology*, 7(1):437.

Andersen, H. C. (1980). Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393.

Andersen, H. C. (1983). Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics*, 52(1):24–34.

Aqvist, J. and Marelius, J. (2001). The Linear Interaction Energy Method for Predicting Ligand Binding Free Energies. *Combinatorial Chemistry & High Throughput Screening*, 4(8):613–626.

Ashley, J., Sorrentino, V., Lobb-Rabe, M., Nagarkar-Jaiswal, S., Tan, L., Xu, S., Xiao, Q., Zinn, K., and Carrillo, R. A. (2019). Transsynaptic interactions between IgSF proteins DIP-α and Dpr10 are required for motor neuron targeting specificity. *eLife*, 8:e42690.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., Van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876.

Bai, F., Morcos, F., Cheng, R. R., Jiang, H., and Onuchic, J. N. (2016). Elucidating the druggable interface of protein-protein interactions using fragment docking and coevolutionary analysis. *Proceedings of the National Academy of Sciences U.S.A.*, 113(50).

Barducci, A., Bonomi, M., and Parrinello, M. (2011). Metadynamics. *WIREs Computational Molecular Science*, 1(5):826–843.

Barish, S., Nuss, S., Strunilin, I., Bao, S., Mukherjee, S., Jones, C. D., and Volkan, P. C. (2018). Combinations of DIPs and Dprs control organization of olfactory receptor neuron terminals in Drosophila. *PLOS Genetics*, 14(8):e1007560.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173.

Basak, S., Schmandt, N., Gicheru, Y., and Chakrapani, S. (2017). Crystal structure and dynamics of a lipid-induced potential desensitized-state of a pentameric ligand-gated channel. *eLife*, 6:e23886.

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2):245–268.

Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690.

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., and MacKerell, A. D. (2012). Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone Φ, Ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273.

Beutler, T. C., Mark, A. E., van Schaik, R. C., Gerber, P. R., and van Gunsteren, W. F. (1994). Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters*, 222(6):529–539.

Bingham, R. J., Findlay, J. B. C., Hsieh, S.-Y., Kalverda, A. P., Kjellberg, A., Perazzolo, C., Phillips, S. E. V., Seshadri, K., Trinh, C. H., Turnbull, W. B., Bodenhausen, G., and Homans, S. W. (2004). Thermodynamics of Binding of 2-Methoxy-3-isopropylpyrazine and 2-Methoxy-3-isobutylpyrazine to the Major Urinary Protein. *Journal of the American Chemical Society*, 126(6):1675–1681.

Birnbaum, M. E., Dong, S., and Garcia, K. C. (2012). Diversity-oriented approaches for interrogating T-cell receptor repertoire, ligand recognition, and function. *Immunological Reviews*, 250(1):82–101.

Blazhynska, M., Gumbart, J. C., Chen, H., Tajkhorshid, E., Roux, B., and Chipot, C. (2023). A Rigorous Framework for Calculating Protein–Protein Binding Affinities in Membranes. *Journal of Chemical Theory and Computation*, 19(24):9077–9092.

Blount, Z. D. (2015). The unexhausted potential of E. coli. *eLife*, 4:e05826.

Bochicchio, D., Panizon, E., Ferrando, R., Monticelli, L., and Rossi, G. (2015). Calculating the free energy of transfer of small solutes into a model lipid membrane: Comparison between metadynamics and umbrella sampling. *The Journal of Chemical Physics*, 143(14):144108.

Bonora, M., Patergnani, S., Rimessi, A., De Marchi, E., Suski, J. M., Bononi, A., Giorgi, C., Marchi, S., Missiroli, S., Poletti, F., Wieckowski, M. R., and Pinton, P. (2012). ATP synthesis and storage. *Purinergic Signalling*, 8(3):343–357.

Born, M. and Oppenheimer, R. (1927). Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389(20):457–484.

Boughter, C. T., Borowska, M. T., Guthmiller, J. J., Bendelac, A., Wilson, P. C., Roux, B., and Adams, E. J. (2020). Biochemical patterns of antibody polyreactivity revealed through a bioinformatics-based analysis of CDR loops. *eLife*, 9:e61393.

Braun, E., Gilmer, J., Mayes, H. B., Mobley, D. L., Monroe, J. I., Prasad, S., and Zuckerman, D. M. (2019). Best Practices for Foundations in Molecular Simulations [Article v1.0]. *LiveCoMS*, 1(1).

Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614.

Brown, S. P. and Muchmore, S. W. (2009). Large-Scale Application of High-Throughput Molecular Mechanics with Poisson-Boltzmann Surface Area for Routine Physics-Based Scoring of Protein-Ligand Complexes. *Journal of Medicinal Chemistry*, 52(10):3159–3165.

Bryant, P., Pozzati, G., and Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, 13(1):1265.

Brünger, A. T. (1997). X-ray crystallography and NMR reveal complementary views of structure and dynamics. *Nature Structural and Molecular Biology*, 4 Suppl:862–865.

Buslaev, P., Aho, N., Jansen, A., Bauer, P., Hess, B., and Groenhof, G. (2022). Best Practices in Constant pH MD Simulations: Accuracy and Sampling. *Journal of Chemical Theory and Computation*, 18(10):6134–6147. Publisher: American Chemical Society.

Bussi, G. and Laio, A. (2020). Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2(4):200–212.

Buß, O., Rudat, J., and Ochsenreither, K. (2018). FoldX as Protein Engineering Tool: Better Than Random Based Approaches? *Computational and Structural Biotechnology Journal*, 16:25–33.

Carrillo, R. A., Özkan, E., Menon, K. P., Nagarkar-Jaiswal, S., Lee, P.-T., Jeon, M., Birnbaum, M. E., Bellen, H. J., Garcia, K. C., and Zinn, K. (2015). Control of Synaptic Connectivity by a Network of Drosophila IgSF Cell Surface Proteins. *Cell*, 163(7):1770–1782.

Cerdan, A. H., Martin, N. E., and Cecchini, M. (2018). An Ion-Permeable State of the Glycine Receptor Captured by Molecular Dynamics. *Structure*, 26(11):1555–1562.e4.

Changeux, J.-P. and Taly, A. (2008). Nicotinic receptors, allosteric proteins and medicine. *Trends in Molecular Medicine*, 14(3):93–102.

Chen, Y., Kale, S., Weare, J., Dinner, A. R., and Roux, B. (2016). Multiple Time-Step Dual-Hamiltonian Hybrid Molecular Dynamics – Monte Carlo Canonical Propagation Algorithm. *Journal of Chemical Theory and Computation*, 12(4):1449–1458.

Chen, Y. and Roux, B. (2014). Efficient hybrid non-equilibrium molecular dynamics - Monte Carlo simulations with symmetric momentum reversal. *The Journal of Chemical Physics*, 141(11):114107.

Chen, Y. and Roux, B. (2015a). Constant-pH Hybrid Nonequilibrium Molecular Dynamics–Monte Carlo Simulation Method. *Journal of Chemical Theory and Computation*, 11(8):3919–3931.

Chen, Y. and Roux, B. (2015b). Enhanced Sampling of an Atomic Model with Hybrid Nonequilibrium Molecular Dynamics—Monte Carlo Simulations Guided by a Coarse-Grained Model. *Journal of Chemical Theory and Computation*, 11(8):3572–3583.

Chen, Y. and Roux, B. (2015c). Generalized Metropolis acceptance criterion for hybrid non-equilibrium molecular dynamics—Monte Carlo simulations. *The Journal of Chemical Physics*, 142(2):024101.

Cheng, R. R., Morcos, F., Levine, H., and Onuchic, J. N. (2014). Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proceedings of the National Academy of SciencesU.S.A.*, 111(5).

Cheng, S., Ashley, J., Kurleto, J. D., Lobb-Rabe, M., Park, Y. J., Carrillo, R. A., and Özkan, E. (2019). Molecular basis of synaptic specificity by immunoglobulin superfamily receptors in Drosophila. *eLife*, 8.

Cherniavskyi, Y. K., Fathizadeh, A., Elber, R., and Tieleman, D. P. (2020). Computer simulations of a heterogeneous membrane with enhanced sampling techniques. *The Journal of Chemical Physics*, 153(14):144110.

Chipot, C. and Pohorille, A., editors (2007). *Free energy calculations. Theory and applications in chemistry and biology.* Springer Verlag, Berlin, Heidelberg, New York.

Christopher, J. A., Brown, J., Doré, A. S., Errey, J. C., Koglin, M., Marshall, F. H., Myszka, D. G., Rich, R. L., Tate, C. G., Tehan, B., Warne, T., and Congreve, M. (2013). Biophysical Fragment Screening of the beta $_1$ -Adrenergic Receptor: Identification of High Affinity Arylpiperazine Leads Using Structure-Based Drug Design. *Journal of Medicinal Chemistry.*, 56(9):3446–3455.

Comer, J., Gumbart, J. C., Hénin, J., Lelièvre, T., Pohorille, A., and Chipot, C. (2015). The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To Ask. *Journal of Physical Chemistry B*, 119(3):1129–1151.

Cooper, G. M. (2000). *The cell: a molecular approach.* ASM Press [u.a.], Washington, DC, 2. ed edition.

Coppock, P. S. and Kindt, J. T. (2009). Atomistic Simulations of Mixed-Lipid Bilayers in Gel and Fluid Phases. *Langmuir*, 25(1):352–359.

Corradi, V., Sejdiu, B. I., Mesa-Galloso, H., Abdizadeh, H., Noskov, S. Y., Marrink, S. J., and Tieleman, D. P. (2019). Emerging Diversity in Lipid–Protein Interactions. *Chemical Reviews*, 119(9):5775–5848.

Cosmanescu, F., Katsamba, P. S., Sergeeva, A. P., Ahlsen, G., Patel, S. D., Brewer, J. J., Tan, L., Xu, S., Xiao, Q., Nagarkar-Jaiswal, S., Nern, A., Bellen, H. J., Zipursky, S. L., Honig, B., and Shapiro, L. (2018). Neuron-Subtype-Specific Expression, Interaction Affinities, and Specificity Determinants of DIP/Dpr Cell Recognition Proteins. *Neuron*, 100(6):1385–1400.e6.

Courgeon, M. and Desplan, C. (2019). Coordination between stochastic and deterministic specification in the *Drosophila* visual system. *Science*, 366(6463):eaay6727.

Crooks, G. E. (1999). Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3):2721–2726.

daCosta, C. and Baenziger, J. (2013). Gating of Pentameric Ligand-Gated Ion Channels: Structural Insights and Ambiguities. *Structure*, 21(8):1271–1283.

daCosta, C. J., Medaglia, S. A., Lavigne, N., Wang, S., Carswell, C. L., and Baenziger, J. E. (2009). Anionic Lipids Allosterically Modulate Multiple Nicotinic Acetylcholine Receptor Conformational Equilibria. *Journal of Biological Chemistry*, 284(49):33841–33849.

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092.

Das, R. and Baker, D. (2008). Macromolecular Modeling with Rosetta. *Annual Review of Biochemistry*, 77(1):363–382.

de Joannis, J., Jiang, Y., Yin, F., and Kindt, J. T. (2006). Equilibrium Distributions of Dipalmitoyl Phosphatidylcholine and Dilauroyl Phosphatidylcholine in a Mixed Lipid Bilayer: Atomistic Semigrand Canonical Ensemble Simulations. *Journal of Physical Chemistry B*, 110(51):25875–25882.

Deng, Y. and Roux, B. (2008). Computation of binding free energy with molecular dynamics and grand canonical Monte Carlo simulations. *The Journal of Chemical Physics*, 128(11):115103.

Dowhan, W. (1997). MOLECULAR BASIS FOR MEMBRANE PHOSPHOLIPID DIVERSITY: Why Are There So Many Lipids? *Annual Review of Biochemistry*, 66(1):199–232.

Egberts, E. and Berendsen, H. J. C. (1988). Molecular dynamics simulation of a smectic liquid crystal with atomic detail. *The Journal of Chemical Physics*, 89(6):3718–3732.

Eriksson, M. A. and Roux, B. (2002). Modeling the Structure of Agitoxin in Complex with the Shaker K+ Channel: A Computational Approach Based on Experimental Distance Restraints Extracted from Thermodynamic Mutant Cycles. *Biophysical Journal*, 83(5):2595–2609.

Ewald, P. P. (1921). Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, 369(3):253–287.

Faraggi, E., Dunker, A. K., Sussman, J. L., and Kloczkowski, A. (2018). Comparing NMR and X-ray protein structure: Lindemann-like parameters and NMR disorder. *Journal of Biomolecular Structure and Dynamics*, 36(9):2331–2341.

Faruk, N. F., Peng, X., Freed, K. F., Roux, B., and Sosnick, T. R. (2022). Challenges and Advantages of Accounting for Backbone Flexibility in Prediction of Protein–Protein Complexes. *Journal of Chemical Theory and Computation*, 18(3):2016–2032.

Fathizadeh, A. and Elber, R. (2018). A mixed alchemical and equilibrium dynamics to simulate heterogeneous dense fluids: Illustrations for Lennard-Jones mixtures and phospholipid membranes. *The Journal of Chemical Physics*, 149(7):072325.

Fathizadeh, A., Valentine, M., Baiz, C. R., and Elber, R. (2020). Phase Transition in a Heterogeneous Membrane: Atomically Detailed Picture. *Journal of Physical Chemistry Lett.*, 11(13):5263–5267.

Feller, S. E., Zhang, Y., Pastor, R. W., and Brooks, B. R. (1995). Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of Chemical Physics*, 103(11):4613–4621.

Fiorin, G., Klein, M. L., and Hénin, J. (2013). Using collective variables to drive molecular dynamics simulations. *Molecular Physics*, 111(22-23):3345–3362.

Frenkel, D. and Smit, B. (2002). *Understanding Molecular Simulation*. Elsevier.

Fu, H., Cai, W., Hénin, J., Roux, B., and Chipot, C. (2017). New Coarse Variables for the Accurate Determination of Standard Binding Free Energies. *Journal of Chemical Theory and Computation*, 13(11):5173–5178.

Fu, H., Chen, H., Blazhynska, M., Goulard Coderc de Lacam, E., Szczepaniak, F., Pavlova, A., Shao, X., Gumbart, J. C., Dehez, F., Roux, B., Cai, W., and Chipot, C. (2022). Accurate determination of protein:ligand standard binding free energies from molecular dynamics simulations. *Nature Protocols*, 17(4):1114–1141.

Fu, H., Chen, H., Cai, W., Shao, X., and Chipot, C. (2021). BFEE2: Automated, Streamlined, and Accurate Absolute Binding Free-Energy Calculations. *Journal of Chemical Information and Modeling*, 61(5):2116–2123.

Fu, H., Shao, X., Cai, W., and Chipot, C. (2019). Taming Rugged Free Energy Landscapes Using an Average Force. *Accounts of Chemical Research*, 52(11):3254–3264.

Fu, H., Zhang, H., Chen, H., Shao, X., Chipot, C., and Cai, W. (2018). Zooming across the Free-Energy Landscape: Shaving Barriers, and Flooding Valleys. *Journal of Physical Chemistry Lett.*, 9(16):4738–4745.

Genheden, S. and Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5):449–461.

Gharpure, A., Teng, J., Zhuang, Y., Noviello, C. M., Walsh, R. M., Cabuco, R., Howard, R. J., Zaveri, N. T., Lindahl, E., and Hibbs, R. E. (2019). Agonist Selectivity and Ion Permeation in the $\alpha3\beta4$ Ganglionic Nicotinic Receptor. *Neuron*, 104(3):501–511.e6.

Ghysels, A., Krämer, A., Venable, R. M., Teague, W. E., Lyman, E., Gawrisch, K., and Pastor, R. W. (2019). Permeability of membranes in the liquid ordered and liquid disordered phases. *Nature Communications*, 10(1):5616.

Gill, S. C., Lim, N. M., Grinaway, P. B., Rustenburg, A. S., Fass, J., Ross, G. A., Chodera, J. D., and Mobley, D. L. (2018). Binding Modes of Ligands Using Enhanced Sampling (BLUES): Rapid Decorrelation of Ligand Binding Modes via Nonequilibrium Candidate Monte Carlo. *Journal of Physical Chemistry B*, 122(21):5579–5598.

Goossens, K. and De Winter, H. (2018). Molecular Dynamics Simulations of Membrane Proteins: An Overview. *Journal of Chemical Information and Modeling*, 58(11):2193–2202.

Goulard Coderc De Lacam, E., Roux, B., and Chipot, C. (2024). Classifying Protein–Protein Binding Affinity with Free-Energy Calculations and Machine Learning Approaches. *Journal of Chemical Information and Modeling*, 64(3):1081–1091.

Gras, S., Burrows, S. R., Turner, S. J., Sewell, A. K., McCluskey, J., and Rossjohn, J. (2012). A structural voyage toward an understanding of the MHC-I-restricted immune response: lessons learned and much to be learned. *Immunological Reviews*, 250(1):61–81.

Grewer, C., Gameiro, A., Mager, T., and Fendler, K. (2013). Electrophysiological Characterization of Membrane Transport Proteins. *Annual Review of Biophysics*, 42(1):95–120.

Grubmüller, H. (1995). Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Physical Review E*, 52(3):2893–2906.

Gumbart, J., Khalili-Araghi, F., Sotomayor, M., and Roux, B. (2012). Constant electric field simulations of the membrane potential illustrated with simple systems. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1818(2):294–302.

Gumbart, J. C., Roux, B., and Chipot, C. (2013a). Efficient Determination of Protein–Protein Standard Binding Free Energies from First Principles. *Journal of Chemical Theory and Computation*, 9(8):3789–3798.

Gumbart, J. C., Roux, B., and Chipot, C. (2013b). Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *Journal of Chemical Theory and Computation*, 9(1):794–802.

Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Research*, 36(9):3025–3030.

Harayama, T. and Riezman, H. (2018). Understanding the diversity of membrane lipid composition. *Nature Reviews Molecular Cell Biology*, 19(5):281–296.

Hassaine, G., Deluz, C., Grasso, L., Wyss, R., Tol, M. B., Hovius, R., Graff, A., Stahlberg, H., Tomizaki, T., Desmyter, A., Moreau, C., Li, X.-D., Poitevin, F., Vogel, H., and Nury, H. (2014). X-ray structure of the mouse serotonin 5-HT3 receptor. *Nature*, 512(7514):276–281.

Hollingsworth, S. A. and Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron*, 99(6):1129–1143.

Holzmann, N., Chipot, C., Penin, F., and Dehez, F. (2016). Assessing the physiological relevance of alternate architectures of the p7 protein of hepatitis C virus in different environments. *Bioorganic & Medicinal Chemistry*, 24(20):4920–4927.

Hopkins, C. W., Le Grand, S., Walker, R. C., and Roitberg, A. E. (2015). Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *Journal of Chemical Theory and Computation*, 11(4):1864–1874.

Huang, K. and García, A. E. (2014). Acceleration of Lateral Equilibration in Mixed Lipid Bilayers Using Replica Exchange with Solute Tempering. *Journal of Chemical Theory and Computation*, 10(10):4264–4272.

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H., and MacKerell, A. D. (2017). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1):71–73.

Hub, J. S., De Groot, B. L., Grubmüller, H., and Groenhof, G. (2014). Quantifying Artifacts in Ewald Simulations of Inhomogeneous Systems with a Net Charge. *Journal of Chemical Theory and Computation*, 10(1):381–390.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38.

Hénault, C. M., Govaerts, C., Spurny, R., Brams, M., Estrada-Mondragon, A., Lynch, J., Bertrand, D., Pardon, E., Evans, G. L., Woods, K., Elberson, B. W., Cuello, L. G., Brannigan, G., Nury, H., Steyaert, J., Baenziger, J. E., and Ulens, C. (2019). A lipid site shapes the agonist response of a pentameric ligand-gated ion channel. *Nature Chemical Biology*, 15(12):1156–1164.

Hénin, J., Lelièvre, T., Shirts, M. R., Valsson, O., and Delemotte, L. (2022). Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]. *LiveCoMS*, 4(1).

Hünenberger, P. H. (2005). Thermostat Algorithms for Molecular Dynamics Simulations. In Abe, A., Joanny, J.-F., Albertsson, A.-C., Duncan, R., Kausch, H.-H., Kobayashi, S., Dušek, K., Lee, K.-S., De Jeu, W., Leibler, L., Nuyken, O., Long, T. E., Terentjev, E., Voit, B., Manners, I., Wegner, G., Möller, M., Dr. Holm, C., and Prof. Dr. Kremer, K., editors, *Advanced Computer Simulation*, volume 173, pages 105–149. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Advances in Polymer Science.

Im, W., Beglov, D., and Roux, B. (1998). Continuum solvation model: Computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. *Computer Physics Communications*, 111(1-3):59–75.

Ingólfsson, H. I., Melo, M. N., van Eerden, F. J., Arnarez, C., Lopez, C. A., Wassenaar, T. A., Periole, X., de Vries, A. H., Tieleman, D. P., and Marrink, S. J. (2014). Lipid Organization of the Plasma Membrane. *Journal of the American Chemical Society*, 136(41):14554–14559.

Jarzynski, C. (1997). Nonequilibrium Equality for Free Energy Differences. *Physical Review Letter*, 78(14):2690–2693.

Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584.

Jo, S., Cheng, X., Lee, J., Kim, S., Park, S.-J., Patel, D. S., Beaven, A. H., Lee, K., Rui, H., Park, S., Lee, H. S., Roux, B., MacKerell, A. D., Klauda, J. B., Qi, Y., and Im, W. (2017). CHARMM-GUI 10 years for biomolecular modeling and simulation: Biomolecular Modeling and Simulation. *Journal of Computational Chemistry*, 38(15):1114–1124.

Jo, S., Kim, T., Iyer, V. G., and Im, W. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11):1859–1865.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Karplus, M. and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural and Molecular Biology*, 9(9):646–652.

Kasparyan, G. and Hub, J. S. (2023). Equivalence of Charge Imbalance and External Electric Fields during Free Energy Calculations of Membrane Electroporation. *Journal of Chemical Theory and Computation*, 19(9):2676–2683.

Kindt, J. T. (2011). Atomistic simulation of mixed-lipid bilayers: mixed methods for mixed membranes. *Molecular Simulation*, 37(7):516–524.

Kirkwood, J. G. (1935). Statistical Mechanics of Fluid Mixtures. *The Journal of Chemical Physics*, 3(5):300–313.

Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., and Cheatham, T. E. (2000). Calculating Structures and Free Energies of Complex Molecules: Combining

Molecular Mechanics and Continuum Models. *Accounts of Chemical Research*, 33(12):889–897.

Kopitz, J. (2017). Lipid glycosylation: a primer for histochemists and cell biologists. *Histochemistry and cell biology*, 147(2):175–198.

Kortemme, T., Kim, D. E., and Baker, D. (2004). Computational Alanine Scanning of Protein-Protein Interfaces. *Science STKE*, 2004(219).

Kroese, D. P., Brereton, T., Taimre, T., and Botev, Z. I. (2014). Why the Monte Carlo method is so important today. *WIREs Computational Statistics*, 6(6):386–392.

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992). THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021.

Kästner, J. (2011). Umbrella sampling. *WIREs Computational Molecular Science*, 1(6):932–942.

Laganowsky, A., Reading, E., Allison, T. M., Ulmschneider, M. B., Degiacomi, M. T., Baldwin, A. J., and Robinson, C. V. (2014). Membrane proteins bind lipids selectively to modulate their structure and function. *Nature*, 510(7503):172–175.

Leach, A. R. (2009). *Molecular modelling: principles and applications.* Pearson/Prentice Hall, Harlow, 2. ed., 12. [dr.] edition.

Lee, J., Cheng, X., Swails, J. M., Yeom, M. S., Eastman, P. K., Lemkul, J. A., Wei, S., Buckner, J., Jeong, J. C., Qi, Y., Jo, S., Pande, V. S., Case, D. A., Brooks, C. L., MacKerell, A. D., Klauda, J. B., and Im, W. (2016). CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *Journal of Chemical Theory and Computation*, 12(1):405–413.

Levine, B. G., Stone, J. E., and Kohlmeyer, A. (2011). Fast analysis of molecular dynamics trajectories with graphics processing units—Radial distribution function histogramming. *Journal of Computational Physics*, 230(9):3556–3569.

Li, S., Wu, S., Wang, L., Li, F., Jiang, H., and Bai, F. (2022). Recent advances in predicting protein–protein interactions with the aid of artificial intelligence algorithms. *Current Opinion in Structural Biology*, 73:102344.

Liu, P., Dehez, F., Cai, W., and Chipot, C. (2012). A Toolkit for the Analysis of Free-Energy Perturbation Calculations. *Journal of Chemical Theory and Computation*, 8(8):2606–2616.

Lockless, S. W. and Ranganathan, R. (1999). Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science*, 286(5438):295–299.

Lorent, J. H., Levental, K. R., Ganesan, L., Rivera-Longsworth, G., Sezgin, E., Doktorova, M., Lyman, E., and Levental, I. (2020). Plasma membranes are asymmetric in lipid unsaturation, packing and protein shape. *Nature Chemical Biology*, 16(6):644–652.

Mackay, D., Berens, P., Wilson, K., and Hagler, A. (1984). Structure and dynamics of ion transport through gramicidin A. *Biophysical Journal*, 46(2):229–248.

Mandal, K. (2020). Review of PIP2 in Cellular Signaling, Functions and Diseases. *International Journal of Molecular Sciences*, 21(21):8342.

Mark, P. and Nilsson, L. (2001). Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *Journal of Physical Chemistry A*, 105(43):9954–9960.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W., and Huber, R. (1983). The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 39(4):480–490.

Marrink, S. J., Corradi, V., Souza, P. C., Ingólfsson, H. I., Tieleman, D. P., and Sansom, M. S. (2019). Computational Modeling of Realistic Cell Membranes. *Chemical Reviews*, 119(9):6184–6226.

Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and de Vries, A. H. (2007). The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *Journal of Physical Chemistry B*, 111(27):7812–7824.

Martins De Oliveira, V., Liu, R., and Shen, J. (2022). Constant pH molecular dynamics simulations: Current status and recent applications. *Current Opinion in Structural Biology*, 77:102498.

McCammon, J. A., Gelin, B. R., and Karplus, M. (1977). Dynamics of folded proteins. *Nature*, 267(5612):585–590.

McCusker, E. C., Bagnéris, C., Naylor, C. E., Cole, A. R., D'Avanzo, N., Nichols, C. G., and Wallace, B. (2012). Structure of a bacterial voltage-gated sodium channel pore reveals mechanisms of opening and closing. *Nature Communications*, 3(1):1102.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Mey, A. S., Allen, B. K., Bruce Macdonald, H. E., Chodera, J. D., Hahn, D. F., Kuhn, M., Michel, J., Mobley, D. L., Naden, L. N., Prasad, S., Rizzi, A., Scheen, J., Shirts, M. R., Tresadern, G., and Xu, H. (2020). Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *LiveCoMS*, 2(1).

Milne, J. L. S., Borgnia, M. J., Bartesaghi, A., Tran, E. E. H., Earl, L. A., Schauder, D. M., Lengyel, J., Pierson, J., Patwardhan, A., and Subramaniam, S. (2013). Cryo-electron microscopy – a primer for the non-microscopist. *The FEBS Journal*, 280(1):28–45.

Minh, D. D., Bui, J. M., Chang, C.-e., Jain, T., Swanson, J. M., and McCammon, J. A. (2005). The Entropic Cost of Protein-Protein Association: A Case Study on Acetylcholinesterase Binding to Fasciculin-2. *Biophysical Journal*, 89(4):L25–L27.

Miyamoto, S. and Kollman, P. A. (1992). Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):952–962.

Morales-Perez, C. L., Noviello, C. M., and Hibbs, R. E. (2016). X-ray structure of the human $\alpha4\beta2$ nicotinic receptor. *Nature*, 538(7625):411–415.

Morcos, F. and Onuchic, J. N. (2019). The role of coevolutionary signatures in protein interaction dynamics, complex inference, molecular recognition, and mutational landscapes. *Current Opinion in Structural Biology*, 56:179–186.

Mori, T., Jung, J., and Sugita, Y. (2013). Surface-Tension Replica-Exchange Molecular Dynamics Method for Enhanced Sampling of Biological Membrane Systems. *Journal of Chemical Theory and Computation*, 9(12):5629–5640.

Mori, T., Miyashita, N., Im, W., Feig, M., and Sugita, Y. (2016). Molecular dynamics simulations of biological membranes and membrane proteins using enhanced conformational sampling algorithms. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1858(7):1635–1651.

Morton, A. and Matthews, B. W. (1995). Specificity of ligand binding in a buried nonpolar cavity of T4 lysozyme: Linkage of dynamics and structural plasticity. *Biochemistry*, 34(27):8576–8588.

Muegge, I. and Hu, Y. (2023). Recent Advances in Alchemical Binding Free Energy Calculations for Drug Discovery. *ACS Medicinal Chemistry Letters Journal*, 14(3):244–250.

Muller, M. P., Jiang, T., Sun, C., Lihan, M., Pant, S., Mahinthichaichan, P., Trifan, A., and Tajkhorshid, E. (2019). Characterization of Lipid–Protein Interactions and Lipid-Mediated Modulation of Membrane Protein Function through Molecular Simulation. *Chemical Reviews*, 119(9):6086–6161.

Nakamura, M., Baldwin, D., Hannaford, S., Palka, J., and Montell, C. (2002). Defective Proboscis Extension Response (DPR), a Member of the Ig Superfamily Required for the Gustatory Response to Salt. *The Journal of Neuroscience*, 22(9):3463–3472.

Nandigrami, P., Szczepaniak, F., Boughter, C. T., Dehez, F., Chipot, C., and Roux, B. (2022). Computational Assessment of Protein–Protein Binding Specificity within a Family of Synaptic Surface Receptors. *Journal of Physical Chemistry B*, 126(39):7510–7527.

Nilmeier, J. P., Crooks, G. E., Minh, D. D. L., and Chodera, J. D. (2011). Nonequilibrium candidate Monte Carlo is an efficient tool for equilibrium simulation. *Proceedings of the National Academy of Sciences*, 108(45):E1009–E1018.

Nina, M., Beglov, D., and Roux, B. (1997). Atomic Radii for Continuum Electrostatics Calculations Based on Molecular Dynamics Free Energy Simulations. *Journal of Physical Chemistry B*, 101(26):5239–5248.

Noviello, C. M., Gharpure, A., Mukhtasimova, N., Cabuco, R., Baxter, L., Borek, D., Sine, S. M., and Hibbs, R. E. (2021). Structure and gating mechanism of the $\alpha7$ nicotinic acetylcholine receptor. *Cell*, 184(8):2121–2134.e13.

Nury, H., Poitevin, F., Van Renterghem, C., Changeux, J.-P., Corringer, P.-J., Delarue, M., and Baaden, M. (2010). One-microsecond molecular dynamics simulation of channel gating in a nicotinic receptor homologue. *Proceedings of the National Academy of SciencesU.S.A.*, 107(14):6275–6280.

Oliveira, A. S. F., Shoemark, D. K., Campello, H. R., Wonnacott, S., Gallagher, T., Sessions, R. B., and Mulholland, A. J. (2019). Identification of the Initial Steps in Signal Transduction in the $\alpha4\beta2$ Nicotinic Receptor: Insights from Equilibrium and Nonequilibrium Simulations. *Structure*, 27(7):1171–1183.e3.

Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK_a Predictions. *Journal of Chemical Theory and Computation*, 7(2):525–537.

Pan, A. C., Jacobson, D., Yatsenko, K., Sritharan, D., Weinreich, T. M., and Shaw, D. E. (2019). Atomic-level characterization of protein–protein association. *Proceedings of the National Academy of SciencesU.S.A.*, 116(10):4244–4249.

Papadourakis, M., Sinenka, H., Matricon, P., Hénin, J., Brannigan, G., Pérez-Benito, L., Pande, V., Van Vlijmen, H., De Graaf, C., Deflorian, F., Tresadern, G., Cecchini, M., and Cournia, Z. (2023). Alchemical Free Energy Calculations on Membrane-Associated Proteins. *Journal of Chemical Theory and Computation*, 19(21):7437–7458.

Park, S., Levental, I., Pastor, R. W., and Im, W. (2023). Unsaturated Lipids Facilitate Partitioning of Transmembrane Peptides into the Liquid Ordered Phase. *Journal of Chemical Theory and Computation*, 19(15):5303–5314.

Petroff, J. T., Dietzen, N. M., Santiago-McRae, E., Deng, B., Washington, M. S., Chen, L. J., Trent Moreland, K., Deng, Z., Rau, M., Fitzpatrick, J. A. J., Yuan, P., Joseph, T. T., Hénin, J., Brannigan, G., and Cheng, W. W. L. (2022). Open-channel structure of a pentameric ligand-gated ion channel reveals a mechanism of leaflet-specific phospholipid modulation. *Nature Communications*, 13(1):7017.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802.

Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., McGreevy, R., Melo, M. C. R., Radak, B. K., Skeel, R. D., Singharoy, A., Wang, Y., Roux, B., Aksimentiev, A., Luthey-Schulten, Z., Kalé, L. V., Schulten, K., Chipot, C., and Tajkhorshid, E. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics*, 153(4):044130.

Pless, S. A. and Sivilotti, L. G. (2018). A tale of ligands big and small: An update on how pentameric ligand-gated ion channels interact with agonists and proteins. *Current Opinion in Physiology*, 2:19–26.

Pohorille, A., Jarzynski, C., and Chipot, C. (2010). Good Practices in Free-Energy Calculations. *Journal of Physical Chemistry B*, 114(32):10235–10253.

Polovinkin, L., Hassaine, G., Perot, J., Neumann, E., Jensen, A. A., Lefebvre, S. N., Corringer, P.-J., Neyton, J., Chipot, C., Dehez, F., Schoehn, G., and Nury, H. (2018). Conformational transitions of the serotonin 5-HT3 receptor. *Nature*, 563(7730):275–279.

Polêto, M. D. and Lemkul, J. A. (2022). Integration of experimental data and use of automated fitting methods in developing protein force fields. *Communications Chemistry*, 5(1):38.

Ponomarenko, E. A., Poverennaya, E. V., Ilgisonis, E. V., Pyatnitskiy, M. A., Kopylov, A. T., Zgoda, V. G., Lisitsa, A. V., and Archakov, A. I. (2016). The Size of the Human Proteome: The Width and Depth. *International Journal of Analytical Chemistry*, 2016:1–6.

Procacci, P., Darden, T., and Marchi, M. (1996). A Very Fast Molecular Dynamics Method To Simulate Biomolecular Systems with Realistic Electrostatic Interactions. *Journal of Physical Chemistry*, 100(24):10464–10468.

Radak, B. K., Chipot, C., Suh, D., Jo, S., Jiang, W., Phillips, J. C., Schulten, K., and Roux, B. (2017). Constant-pH Molecular Dynamics Simulations for Large Biomolecular Systems. *Journal of Chemical Theory and Computation*, 13(12):5933–5944.

Radak, B. K. and Roux, B. (2016). Efficiency in nonequilibrium molecular dynamics Monte Carlo simulations. *The Journal of Chemical Physics*, 145(13):134109.

Rahman, A. (1964). Correlations in the Motion of Atoms in Liquid Argon. *Physical Review*, 136(2A):A405–A411.

Rahman, A. and Stillinger, F. H. (1971). Molecular Dynamics Study of Liquid Water. *The Journal of Chemical Physics*, 55(7):3336–3359.

Ramadoss, V., Dehez, F., and Chipot, C. (2016). AlaScan: A Graphical User Interface for Alanine Scanning Free-Energy Calculations. *Journal of Chemical Information and Modeling*, 56(6):1122–1126.

Rao, S., Klesse, G., Lynch, C. I., Tucker, S. J., and Sansom, M. S. P. (2021). Molecular Simulations of Hydrophobic Gating of Pentameric Ligand Gated Ion Channels: Insights into Water and Ions. *Journal of Physical Chemistry B*, 125(4):981–994.

Rashid, M. H., Heinzelmann, G., Huq, R., Tajhya, R. B., Chang, S. C., Chhabra, S., Pennington, M. W., Beeton, C., Norton, R. S., and Kuyucak, S. (2013). A Potent and Selective Peptide Blocker of the Kv1.3 Channel: Prediction from Free-Energy Simulations and Experimental Confirmation. *PLOS ONE*, 8(11):e78712.

Rocklin, G. J., Mobley, D. L., Dill, K. A., and Hünenberger, P. H. (2013). Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *The Journal of Chemical Physics*, 139(18):184103.

Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. (2004). Protein Structure Prediction Using Rosetta. In *Methods in Enzymology*, volume 383, pages 66–93. Elsevier.

Rose, M., Hirmiz, N., Moran-Mirabal, J., and Fradin, C. (2015). Lipid Diffusion in Supported Lipid Bilayers: A Comparison between Line-Scanning Fluorescence Correlation Spectroscopy and Single-Particle Tracking. *Membranes*, 5(4):702–721.

Roux, B. (2011). *Molecular Machines*. WORLD SCIENTIFIC.

Roux, B. (2021). *Computational Modeling And Simulations Of Biomolecular Systems*. World Scientific Publishing Company, Incorporated, world scientific publishing company, incorporated edition.

Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341.

Sakmann, B., Patlak, J., and Neher, E. (1980). Single acetylcholine-activated channels show burst-kinetics in presence of desensitizing concentrations of agonist. *Nature*, 286(5768):71–73.

Salari, R., Murlidaran, S., and Brannigan, G. (2014). Pentameric ligand-gated ion channels: insights from computation. *Molecular Simulation*, 40(10-11):821–829.

Salinas, V. H. and Ranganathan, R. (2018). Coevolution-based inference of amino acid interactions underlying protein function. *eLife*, 7:e34300.

Santiago McRae, E., Ebrahimi, M., Sandberg, J. W., Brannigan, G., and Hénin, J. (2023). Computing absolute binding affinities by Streamlined Alchemical Free Energy Perturbation [Article v1.0]. *LiveCoMS*, 5(1).

Schneider, T. and Stoll, E. (1978). Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Physical Review B*, 17(3):1302–1322.

Schwarzl, S. M., Tschopp, T. B., Smith, J. C., and Fischer, S. (2002). Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas entropy correction? *Journal of Computational Chemistry*, 23(12):1143–1149.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research*, 33(Web Server):W382–W388.

Schärer, K., Morgenthaler, M., Paulini, R., Obst-Sander, U., Banner, D. W., Schlatter, D., Benz, J., Stihle, M., and Diederich, F. (2005). Quantification of Cation–π Interactions in Protein–Ligand Complexes: Crystal-Structure Analysis of Factor Xa Bound to a Quaternary Ammonium Ion Ligand. *Angewandte Chemie International Edition*, 44(28):4400–4404.

Sergeeva, A. P., Katsamba, P. S., Cosmanescu, F., Brewer, J. J., Ahlsen, G., Mannepalli, S., Shapiro, L., and Honig, B. (2020). DIP/Dpr interactions and the evolutionary design of specificity in protein families. *Nature Communications*, 11(1):2125.

Sezgin, E., Levental, I., Mayor, S., and Eggeling, C. (2017). The mystery of membrane organization: composition, regulation and roles of lipid rafts. *Nature Reviews Molecular Cell Biology*, 18(6):361–374.

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

Shoup, D. and Szabo, A. (1982). Role of diffusion in ligand binding to macromolecules and cell-bound receptors. *Biophysical Journal*, 40(1):33–39.

Simonson, T. and Roux, B. (2016). Concepts and protocols for electrostatic free energies. *Molecular Simulation*, 42(13):1090–1101.

Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B., and Sansom, M. S. (1996). HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *Journal of Molecular Graphics*, 14(6):354–360.

Smyth, M. S. (2000). x Ray crystallography. *Molecular Pathology*, 53(1):8–14.

Sodt, A., Pastor, R., and Lyman, E. (2015). Hexagonal Substructure and Hydrogen Bonding in Liquid-Ordered Phases Containing Palmitoyl Sphingomyelin. *Biophysical Journal*, 109(5):948–955.

Song, Y., DiMaio, F., Wang, R.-R., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013). High-Resolution Comparative Modeling with RosettaCM. *Structure*, 21(10):1735–1742.

Souaille, M. and Roux, B. (2001). Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer Physics Communications*, 135(1):40–57.

Sperry, R. W. (1963). Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proceedings of the National Academy of Sciences*, 50(4):703–710.

Stern, H. A. (2007). Molecular simulation with variable protonation states at constant pH. *The Journal of Chemical Physics*, 126(16):164112.

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

Suh, D., Jo, S., Jiang, W., Chipot, C., and Roux, B. (2019). String Method for Protein–Protein Binding Free-Energy Calculations. *Journal of Chemical Theory and Computation*, 15(11):5829–5844.

Suh, D., Radak, B. K., Chipot, C., and Roux, B. (2018). Enhanced configurational sampling with hybrid non-equilibrium molecular dynamics–Monte Carlo propagator. *The Journal of Chemical Physics*, 148(1):014101.

Swanson, J. M., Henchman, R. H., and McCammon, J. A. (2004). Revisiting Free Energy Calculations: A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy. *Biophysical Journal*, 86(1):67–74.

Symons, J. L., Cho, K.-J., Chang, J. T., Du, G., Waxham, M. N., Hancock, J. F., Levental, I., and Levental, K. R. (2021). Lipidomic atlas of mammalian cell membranes reveals hierarchical variation induced by culture conditions, subcellular membranes, and cell lineages. *Soft Matter*, 17(2):288–297.

Szczepaniak, F., Dehez, F., and Roux, B. (2024). Configurational Sampling of All-Atom Solvated Membranes Using Hybrid Nonequilibrium Molecular Dynamics Monte Carlo Simulations. *Journal of Physical Chemistry Lett.*, 15(14):3796–3804.

Szlasa, W., Zendran, I., Zalesińska, A., Tarek, M., and Kulbacka, J. (2020). Lipid composition of the cancer cell membrane. *Journal of Bioenergetics and Biomembranes*, 52(5):321–342.

Tan, L., Zhang, K., Pecot, M., Nagarkar-Jaiswal, S., Lee, P.-T., Takemura, S.-y., McEwen, J., Nern, A., Xu, S., Tadros, W., Chen, Z., Zinn, K., Bellen, H., Morey, M., and Zipursky, S. (2015). Ig Superfamily Ligand and Receptor Pairs Expressed in Synaptic Partners in Drosophila. *Cell*, 163(7):1756–1769.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680.

Thompson, M. J. and Baenziger, J. E. (2020). Structural basis for the modulation of pentameric ligand-gated ion channel function by lipids. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1862(9):183304.

Timm, D. E., Baker, L., Mueller, H., Zidek, L., and Novotny, M. V. (2001). Structural basis of pheromone binding to mouse major urinary protein (MUP-I). *Protein Science*, 10(5):997–1004.

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

Trifonov, E. N. (2011). Vocabulary of Definitions of Life Suggests a Definition. *Journal of Biomolecular Structure and Dynamics*, 29(2):259–266.

Tuckerman, M., Berne, B. J., and Martyna, G. J. (1992). Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics*, 97(3):1990–2001.

Van Den Bogaart, G., Meyenberg, K., Risselada, H. J., Amin, H., Willig, K. I., Hubrich, B. E., Dier, M., Hell, S. W., Grubmüller, H., Diederichsen, U., and Jahn, R. (2011). Membrane protein sequestering by ionic protein–lipid interactions. *Nature*, 479(7374):552–555.

Van Meer, G., Voelker, D. R., and Feigenson, G. W. (2008). Membrane lipids: where they are and how they behave. *Nature Reviews Molecular Cell Biology*, 9(2):112–124.

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., and Mackerell, A. D. (2010). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4):671–690.

Vickery, O. N. and Stansfeld, P. J. (2021). CG2AT2: an Enhanced Fragment-Based Approach for Serial Multi-scale Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*, 17(10):6472–6482.

Wallin, E. and Heijne, G. V. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms: Membrane protein topology. *Protein Science*, 7(4):1029–1038.

Walsh, R. M., Roh, S.-H., Gharpure, A., Morales-Perez, C. L., Teng, J., and Hibbs, R. E. (2018). Structural principles of distinct assemblies of the human $\alpha 4\beta 2$ nicotinic receptor. *Nature*, 557(7704):261–265.

Wang, E., Sun, H., Wang, J., Wang, Z., Liu, H., Zhang, J. Z. H., and Hou, T. (2019). End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews*, 119(16):9478–9508.

Wang, H., Zhang, P., and Schütte, C. (2012). On the Numerical Accuracy of Ewald, Smooth Particle Mesh Ewald, and Staggered Mesh Ewald Methods for Correlated Molecular Systems. *Journal of Chemical Theory and Computation*, 8(9):3243–3256.

Watson, H. (2015). Biological membranes. *Essays in Biochemistry*, 59:43–69.

Woo, H.-J., Dinner, A. R., and Roux, B. (2004). Grand canonical Monte Carlo simulations of water in protein environments. *The Journal of Chemical Physics*, 121(13):6392–6400.

Woo, H.-J. and Roux, B. (2005). Calculation of absolute protein–ligand binding free energy from computer simulations. *Proceedings of the National Academy of SciencesU.S.A.*, 102(19):6825–6830.

Woolf, T. B. and Roux, B. (1994). Molecular dynamics simulation of the gramicidin channel in a phospholipid bilayer. *Proceedings of the National Academy of SciencesU.S.A.*, 91(24):11631–11635.

Wu, Z. and Biggin, P. C. (2022). Correction Schemes for Absolute Binding Free Energies Involving Lipid Bilayers. *Journal of Chemical Theory and Computation*, 18(4):2657–2672.

Xu, L., Sun, H., Li, Y., Wang, J., and Hou, T. (2013). Assessing the Performance of MM/PBSA and MM/GBSA Methods. 3. The Impact of Force Fields and Ligand Charge Models. *Journal of Physical Chemistry B*, 117(28):8408–8421.

Yeh, I.-C. and Wallqvist, A. (2011). On the proper calculation of electrostatic interactions in solid-supported bilayer systems. *The Journal of Chemical Physics*, 134(5):055109.

Yellen, G. (2002). The voltage-gated potassium channels and their relatives. *Nature*, 419(6902):35–42.

Yin, H. and Flynn, A. D. (2016). Drugging Membrane Protein Interactions. *Annual Review of Biomedical Engineering*, 18(1):51–76.

Yin, L., Scott-Browne, J., Kappler, J. W., Gapin, L., and Marrack, P. (2012). T cells and their eons-old obsession with MHC. *Immunological Reviews*, 250(1):49–60.

Yu, R., Tae, H., Xu, Q., Craik, D. J., Adams, D. J., Jiang, T., and Kaas, Q. (2019). Molecular dynamics simulations of dihydro-beta-erythroidine bound to the human $\alpha4\beta2$ nicotinic acetylcholine receptor. *British Journal of Pharmacology*, 176(15):2750–2763.

Yu, W., He, X., Vanommeslaeghe, K., and MacKerell, A. D. (2012). Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of Computational Chemistry*, 33(31):2451–2468.

Zacharias, M., Straatsma, T. P., and McCammon, J. A. (1994). Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *The Journal of Chemical Physics*, 100(12):9025–9031.

Zarkadas, E., Pebay-Peyroula, E., Thompson, M. J., Schoehn, G., Uchański, T., Steyaert, J., Chipot, C., Dehez, F., Baenziger, J. E., and Nury, H. (2022). Conformational transitions and ligand-binding to a muscle-type nicotinic acetylcholine receptor. *Neuron*, 110(8):1358–1370.e5.

Zhao, Y., Liu, S., Zhou, Y., Zhang, M., Chen, H., Eric Xu, H., Sun, D., Liu, L., and Tian, C. (2021). Structural basis of human $\alpha 7$ nicotinic acetylcholine receptor activation. *Cell Research*, 31(6):713–716.

Zhou, X., Song, H., and Li, J. (2022). Residue-Frustration-Based Prediction of Protein–Protein Interactions Using Machine Learning. *Journal of Physical Chemistry B*, 126(8):1719–1727.

Zhu, F. and Hummer, G. (2012). Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry*, 33(4):453–465.

Zinn, K. and Özkan, E. (2017). Neural immunoglobulin superfamily interaction networks. *Current Opinion in Neurobiology*, 45:99–105.

Zipursky, S. L. and Sanes, J. R. (2010). Chemoaffinity Revisited: Dscams, Protocadherins, and Neural Circuit Assembly. *Cell*, 143(3):343–353.

Zwanzig, R. W. (1954). High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics*, 22(8):1420–1426.

Åqvist, J., Medina, C., and Samuelsson, J.-E. (1994). A new method for predicting binding affinity in computer-aided drug design. *Protein Engineering, Design and Selection*, 7(3):385–391.

Özkan, E., Carrillo, R. A., Eastman, C. L., Weiszmann, R., Waghray, D., Johnson, K. G., Zinn, K., Celniker, S. E., and Garcia, K. C. (2013). An Extracellular Interactome of Immunoglobulin and LRR Proteins Reveals Receptor-Ligand Networks. *Cell*, 154(1):228–239.

Özkan, E., Chia, P., Wang, R., Goriatcheva, N., Borek, D., Otwinowski, Z., Walz, T., Shen, K., and Garcia, K. (2014). Extracellular Architecture of the SYG-1/SYG-2 Adhesion Complex Instructs Synaptogenesis. *Cell*, 156(3):482–494.

Šali, A. and Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, 234(3):779–815.

Šali, A., Potterton, L., Yuan, F., Van Vlijmen, H., and Karplus, M. (1995). Evaluation of comparative protein modeling by MODELLER. *Proteins*, 23(3):318–326.