Máster en

Química Teórica y Modelización Computacional (TCCM)

Modelling the Binding of Azobenzene to the Human Voltage-Gated Sodium Channel Nav 1.4

Carlos Gómez Rodellar

Trabajo Fin de Máster. Curso 2018-2020

Director: Juan José Nogueria Pérez
Lugar de realización: Facultad de Ciencias
Departamento de Química

This page is left blank on porpouse

Table of Contents

Acknowledgments

Special thanks to Dr. Juan José Nogueira Pérez, director of the thesis, for his help and for the invaluable enthusiasm he shows for his students and the research group. Thanks to Vito Palmisano, TCCM partner, for being always aviable for trouble shooting and doubt sharing. Thanks to Hannah Pollak for setting up the protein model used in this project. Thanks to Aiste Milūtė for making the quarantine better and being an infinite source of support. I would also like to thank family and friends for patience and love during the quarantine.

1 Abstract

1.1 Abstract (ENG)

Ion channels have an important role in biological processes and are nowadays exploited as potential drug targets. Photopharmacology attempts to use photosensitive drugs for precise spatiotemporal control of drug action. These drugs have already been used in ion channels to target channelopathies which cause blindness and chronical pain. In this project the binding of the photoswitch Azobenzene to the α structure of the human Na_v 1.4 ion channel has been theoretically investigated by computational means. Specifically, Gaussian Accelerated Molecular Dynamics has been employed for enhancing the sampling of the conformational space and to compute the free energy surface. This allowed the identification of 4 main binding pockets which are located close to dominium II and inside a hydrophobic cavity between dominium II and dominium III inside the channel. Conventional Molecular Dynamics has been used to assess the sampling efficiency of the enhanced sampling approach. Then, Molecular Mechanics/Generalized Born Surface Area was used to estimate the binding free energy for each pocket and the contribution of each protein residue. Results indicate that the main interactions between ligand and ion channel are of van der Waals nature as well as a significant contribution from non-polar solvation effects. These non-polar solvation effects may be driven by Azobenzene's protection of hydrophobic residues from solvent, which displayed typical "pi-stacking" conformations present in aromatic rings.

1.2 Abstract (ESP)

Los canales iónicos tienen una relevancia importante en procesos biológicos y actualmente están siendo usados como dianas terapéuticas. La fotofarmacología se basa en el uso de medicamentos sensibles a la luz para obtener un control espaciotemporal preciso de su actividad, estos medicamentos ya han sido usados para tratar patologías en canales iónicos las cuales causan ceguera y dolor crónico. Debido a esto, la unión entre Azobenzeno (un ("photoswitch") y la estructura α del canal humano Na_v 1.4 ha sido estudiada teóricamente en este proyecto usando métodos computacionales. Las trayectorias generadas por "Gaussian Accelerated Molecular Dynamics" han sido usada para obtener un mejor muestreo del espacio de fase. La reponderación de estas trayectorias para obtener la superficie de energía libre original ha sido llevada a cabo con series de McLaurin de orden 10, esto a su vez fue usado para obtener los sitios de unión del Azobenzeno con el canal iónico. Estos sitios de unión se encontraban o bien cerca del dominio II o bien entre una cavidad hidrófoba entre los dominios II y III del canal. La técnica "Conventional Molecular Dynamics" fue usada para comparar el incremento de muestreo obtenido y para obtener trayectorias de exploración de los puntos de unión. A continuación, "Mechanics/Genralized Born Surface Area" usó estas trayectorias para calcular incrementos de energía libre de unión, mostrando que las interacciones de van der Waals eran las dominantes y que había una contribución significativa por parte de la energía de solvatación no polar. Esta última contribución probablemente es causada por la protección de aminoácidos hidrofóbicos respecto al disolvente por parte del Azobenzeno el cual tomaba conformaciones de "pi-stacking" típicas entre anillo aromáticos.

2 Introduction

2.1 Introduction and Motivation

Synthetic organic chemicals for modulation of life processes were first used in the 1840s. William Morton, a dentist, made the first true demonstration of ether as an inhalation anaesthetic at what is now called the Ether Dome at Massachusetts General Hospital in 1846. [1] Drug discovery has been driven by chemistry but increasingly guided by pharmacology and the clinical sciences. It can be said that drug research has made an enormous contribution to medicine in the past century. After reaching a degree of maturity with confirmation of Avogadro's atomic hypothesis, acid-base theory, and an ordered periodic table, chemistry began to be applied to problems outside chemistry itself, establishing pharmacology as a well-defined scientific discipline. Kekulé's theory of organic aromatic molecules gave a strong impulse to the development of coal-tar derivatives, particularly dyes. This had a profound impact on medicine due to the development of tissue selective dyes which led anatomists to postulate the existence of chemoreceptors in cells which could be exploited therapeutically.

Biochemistry has influenced drug research by including the concepts of enzymes and receptors which were found to be good drug targets. The importance of molecular biology became such that new drugs are not only generated by chemists but from a dialogue between chemists and biologists. While biologists explain biochemical mechanisms of action, chemists synthesize new chemical structures. The principal promise of molecular biology is to give a potential understanding of processes at the molecular level and determine optimal molecular targets for drug action.[2]

Nowadays, computational (*in silico*) methods are being applied to pharmacology for hypothesis development and testing. These methods include: Quantitative Structure-Activity Relationships (QSAR)[3], Homology Models, Similarity Searching, Molecular Modeling, and Machine Learning among many others. Such methods are frequently used in the discovery and optimization of molecules. At the same time, they offer clarification of distribution, metabolism, absorption, toxicity, and physicochemical characterization of potential drug candidates.[4]

Ion channels are promising therapeutically exploited receptors. These are pore-like structures that regulate the flow of ions across the cell membrane. Their objective of this is to control ion concentrations inside the cell. Chemically, they are transmembrane proteins with a gated, water filled pore and are generally selectively permeable to different ions. They are broadly divided into voltage gated (voltage determines pass of ions), ligand gated (a ligand controls the opening—closure of the channel), and mechanosensitive (gating due to mechanical deformations). These structures are involved in several aspects of physiology like nerve muscle relaxation, cognition, sensory transduction, regulation of blood pressure, and cell proliferation. Several diseases have been linked with them: neurological Indications, kidney failure, cardiac disorders, perception of pain, and blindness. Their potential use as drug targets for therapeutic treatment is known already. In fact, some local anaesthetics such as lidocaine and bupivacaine act on voltage gated sodium channels (Na_v). Some other examples are capsaicin (spicy flavour) or menthol (minty effect) which act on ligand gated ion channels.[5]

Venoms of some snakes, scorpions, spiders and some sea animals have evolved to act with potency and selectivity against voltage gated potassium channels (K_v), voltage gated calcium

channels (Ca_v) and Na_v as well as on some ligand gated ion channels. Research has already been conducted in order to study venoms which could serve as potential drug candidates.[6] There is a close relationship between toxicity and drugs. While toxins can be used in small amounts as therapeutic agents, an overdose of them can be toxic for the organism. This happens because active drug molecules usually lack spatiotemporal control. In other words, once applied they start acting straight away (no temporal control) and in many cases they act with poor target specificity (spatial control) affecting other structures as side effects. However, light can be controlled quite precisely both in space and time. The suggestion is to functionalize drug molecules with a light sensitive part which will allow for light-controlled activity. This approach is denominated photopharmacology.[7] Furthermore, this therapeutic approach has been widely applied to ion channels[8,9], the first attempts dating from the late 1960s.[10]

Light-matter interaction is not always easy to describe by experimental means, but it can be approached by computational ones. Absorption of light depends on the restructuration of charge when a molecule is hit by a photon and the consequent dipole moment generated. This is altered not only by the structure of the molecule itself but also of the environment that the molecule is exposed to. Several Quantum Mechanical (QM) methods such as Complete Active Space Self-Consistent Field (CASSCF) or Time Dependent Density Functional Theory (TD-DFT) are used to explain the underlying physics of these phenomena.[11]

In the present project, the first step of the mode of action of a light-sensitive molecule interacting with an ion channel has been studied with computational methods. Specifically, the ligand binding of a photoswitch (azobenzene) to a human voltage gated sodium channel present in human cardiac tissue and respiratory skeletal muscle (Na_v 1.4)[12] has been theoretically investigated.

During the previous section three main concepts were introduced: Drug design by computational approaches, ion channels, and photopharmacology. Having this in mind, the idea of studying together photo-sensitive drugs inside ion channels for possible control of ion flow using a computational approach, is the motivation of this project. Ideally, the final goal of a long-term research would be to describe the light interaction of a light-sensitive potential drug candidate outside and inside an ion channel, and then, to study the dynamics of that drug as an ion channel control agent. This would require of three phases:

1. Phase 1: Study the dynamics of the drug inside of the protein without light-interaction. This would give an idea of the affinity between the drug and the ion channel and justify if further stages are feasible (work developed in this project).
2. Phase 2: Study the light-matter interaction properties of the drug candidate both inside and outside the channel. In other words, get the absorption spectrum of the drug and study the influences of the ion channel (outlook and future research).
3. Phase 3: Study the excited state dynamics of the molecule inside the channel upon light excitation. In addition to this, study the repercussions on the ion flow both in the fundamental state of the drug and the excited state (outlook and future research).

The full study of the whole process is too large and complex to be handled in a Master Thesis. Therefore, the research conducted in this project is limited to phase 1.

In this project, the binding of Azobenzene, a photoswitch, has been studied using computational *(in silico)* methods to the human voltage gated sodium channel Na_v1.4. In particular, Molecular Dynamics (MD)[13] and Gaussian Accelerated Molecular Dynamics (GaMD)[14] have been employed to determine the binding pockets, Molecular Mechanics energies combined with generalized

Born and surface area continuum solvation (MM/GBSA)[15] have been used for pocket free energy estimation for ligand-protein binding. Reweighting of GaMD energy surface was done using a self-modified version of PyReweighting[16] scripts . Azobenzene probability distributions were obtained by self-developed scripts implemented in Python 3.6 .[17] These methods will be explained in the chapter 3 Methods.

2.2 Ion Channels as Therapeutic Targets

2.2.1 General Concepts About Ion Channels

Ion channels are proteins which form macromolecular pores in cell membranes. These pores are selectively used by ions to move in and out of the cellular cytoplasm (Fig 2.1). This flow of ions is regulated by a wide range of stimuli, such as: the presence other molecules, membrane potential changes, temperature, and mechanical force, among others. Due to the importance of ion channels and their regulatory role in the body, pathologies such as arrythmia and cystic fibrosis arise from mutations in the genes that codify them. As a consequence, these proteins are being exploited nowadays as potential therapeutic targets.[18]

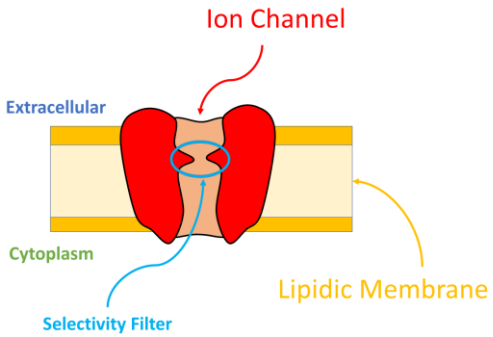

Fig 2.1 Scheme of an Ion Channel: It is a structure which goes through the lipidic cell membrane, which has different permeability depending on the ion. This permeability is achieved through a structure called selectivity filter which is formed by 4 monomers.

Ion Channels are broadly classified into two groups: Voltage gated and ligand Gated depending on the mechanism that governs the flow of ions. Nevertheless, more types can be found, such as phosphorylation gated and stretch or pressure gated channels. All of them are schematically shown in Fig 2.2. Although the picture of a gate opening, and closing can be convenient for some types of channels (Na_v or K_v channels) others may undergo complicated conformational changes like twisting[18].

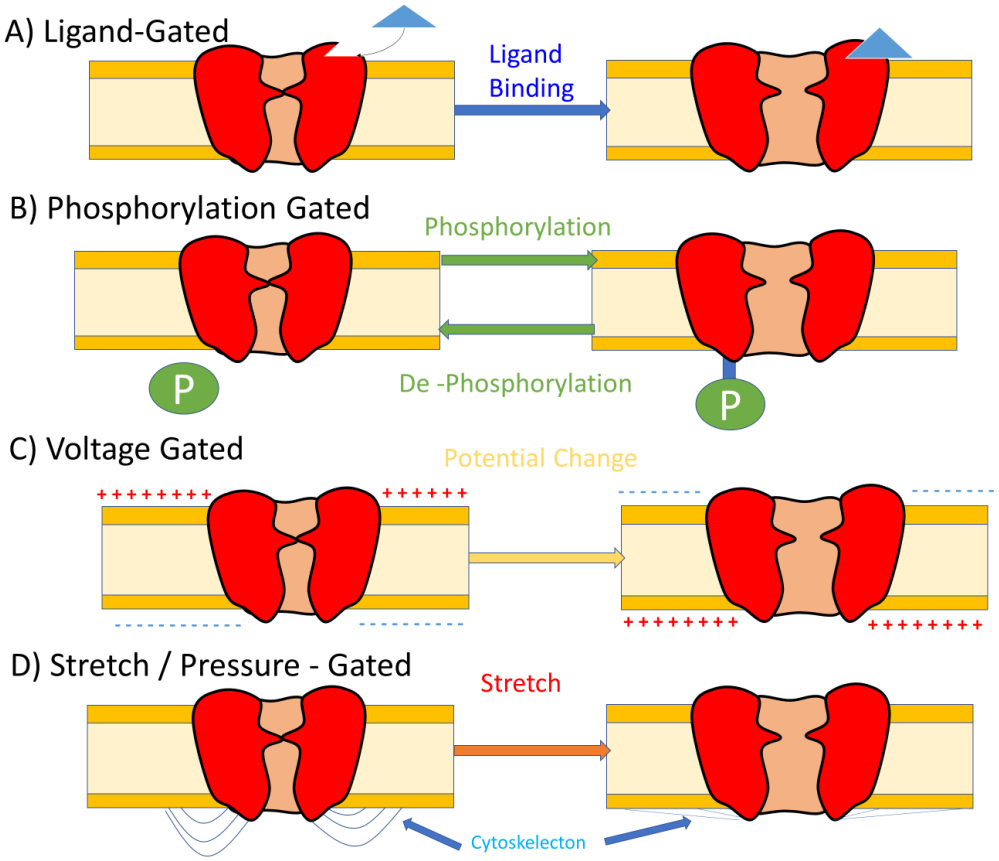

Fig. 2.2 Schematic Representation of Stimuli Control in Ion Channels: A) Ligand-Gated channels open/close when receiving a ligand that binds to a receptor. B) Phosphorylation-Dephosphorilatyon regulates the channel by addition/subtraction of a phosphate. C) Voltage can open and close some channels. Conformational change is induced by protein domains with net charge that can sense potential differences. D) Mechanical action apports energy to change the cytoskeleton and impose conformational changes in the protein.

In this project, the emphasis has been put on the voltage gated type channels. Some examples of this subtype are: K_v, Ca_v and Na_v . Specifically, the focus of this project has been centred in this last type, Na_v. This is not without a reason, due to their physiological importance in processes such as cell signalling and their involvement in neuronal activity.[18]

2.2.2 Importance of Voltage Gated Sodium Channels: Nerve Cell Signalling

The fundamental mechanism that governs electrical impulses in cells is summarized here. When a nerve cell is at rest, ion channels which are selectively permeable to K^+ but considerably less permeable to Na^+ are opened. This causes K^+ ions to leak out from the cell leaving unneutralized negative ions at the inner surface of the membrane, causing a resting potential. An initial stimuli of 10 mV which causes a minor increment of that membrane's potential (from -65 mV to -55mV) suddenly provokes the Na_v channels to open, allowing Na^+ to get inside the cell and neutralize the unbalanced negative charges. This results in a greater potential increment named action potential which induces the opening of other neighbouring Na_v channels and the transport of additional Na^+ ions from the extracellular region to the cytoplasm. The action potential is, therefore, carried across the axons allowing for electrical signalling between nerve cells. After the membrane has been discharged re-equilibration of potassium and sodium ions after each action potential is done thanks to the sodium potassium pump (Na^+-K^+-ATPase).[18] Not all signalling consists of action potentials since these are used for long range communication

between nerve cells, another variation, synaptic potentials, are used for local purposes. If instead of an increment of 10 mV from the resting potential a drop happens (to -75mV), a consequent action potential will be harder to trigger causing an inhibitory effect in the nerve cell.

Because of their importance, drugs that block Na_v channels are used as local anaesthetics and to treat diseases like epilepsy, bipolar disorder, chronic pain, and cardiac arrhythmia. In addition, several psychoactive compounds such as Cocaine are known for acting on Na_v channels.[19,20]

2.2.3 Structure of Voltage Gated Ion Channels

The general structure of voltage gated ion channels regardless of the type (Ca_v, K_v, or Na_v) has some common characteristics. Four Domains (DI , DII , DIII , and DIV) arranged around a central pore form the zone in charge of ion permeation as shown in Fig 2.3. These domains are covalently bonded for Na_v and Ca_v channels, for K_v channels the domains remain as four separate protein chains. Each domain is made up of six transmembrane helices (TM), the first four (TM1-TM4) form *a* "Voltage Sensing Domain (VSD)", while the two remaining (TM5-TM6) form the pore domain of the channel. A "P-Loop" joins the TM5 and TM6 inside each domain, and an aminoacid inside this structure (marked by (X)) forms an important region of the selectivity filter of the channel. This selectivity region is a "ring-shaped" structure which has great importance since it determines the permeability of ions through the channel. For bacteria Na_v channels, the selectivity filter is formed by four GLU (E E E E) while for eukaryotic cells this filter is formed by four different aminoacids : ASP (D), GLU (E), LYS (K), and ALA (A). All of these structures form the alpha (α) subunit of the channel which is sufficient for selective voltage-dependent ion permeation.[21]

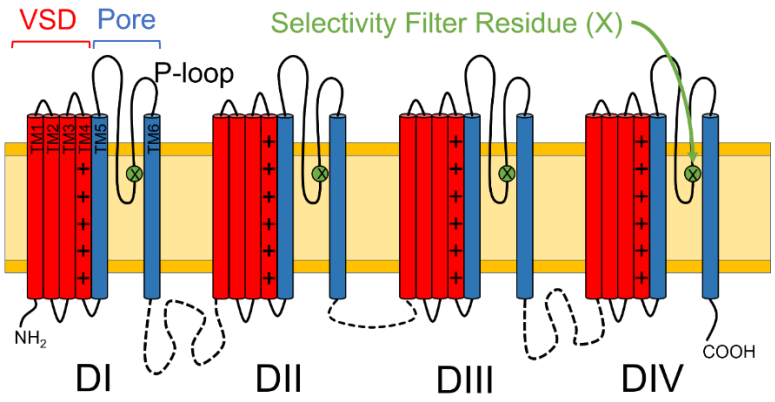

Fig 2.3 Scheme of the voltage gated ion channel general structure: As appreciated each domain contains 6 TM (TM1-6). Between TM5 and TM6 an aminoacid in the P-loop forms the "selectivity filter" in charge of giving different permeabilities to ions that try to pass through the channel. This is a representation of how each domain connects with the next, but not how the channel looks like.

A more accurate picture of how the loop looks is obtained by "folding" the channel, placing each domain around forming a pore as depicted in Fig 2.4.

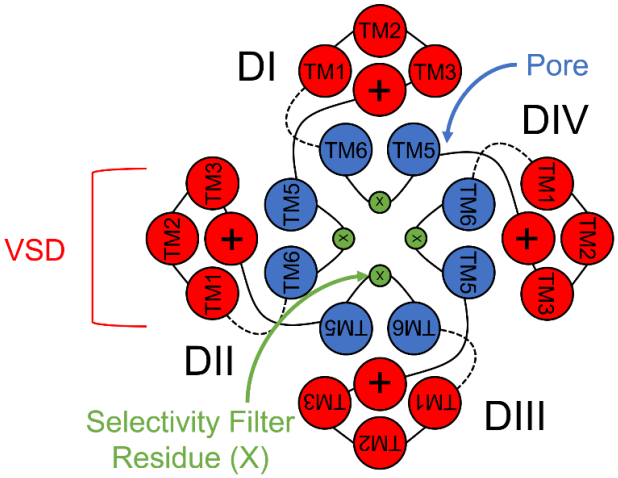

Fig 2.4 Depiction of DI - IV Forming a Pore: All four domains gather around forming a pore with the selectivity filter approximately in the middle of the structure. This visualization corresponds to a top view of the channel.

In addition to the α subunit other beta (β) subunits can appear with the channel. These structures help to regulate membrane trafficking and channel properties.[21]

The regulation mechanism behind voltage-sensing involves positively charged aminoacids (LYS and ARG with NH_3^+ species under physiological pH) in the TM4 of each VSD. Upon changes in the membrane potential these helixes move away from the cytoplasm. Since the VSD are connected to the pore segments, this movement imposes a conformational change in the pore structure allowing the channel to open.[22]

Regarding structural particularities, voltage gated ion channels present some differences depending on which ion they are selective for.[23] In addition, ion channels are different depending on the living organism they belong to. It seems that most mammalian ion channels originate from prokaryote ones. Since ion channels have similar composition but have a percentage of different aminoacids depending on the species and the type of channel , evolutionary studies are conducted in order to allow for better comprehension of the genes that codify the channels and understand the mutations that they undergo.[24–26]

Knowledge about ion channels is quite vast, more than it could be covered here. A recommended book for reading is the first chapter of Principles of Neuroscience. Kandel, E. R., Schwartz, J. H. 1., & Jessell, T. M .[18]

2.2.4 The Human voltage gated Sodium Channel Na_v 1.4

In *Homo sapiens* (humans) there are nine subtypes of Na_v Channels. Na_v1.1, Na_v1.2, Na_v1.3, and Na_v1.6, are primarily present in the central nervous system. Na_v1.4, and Na_v1.5, work in skeletal muscle and heart, and Na_v1.7, Na_v1.8, Na_v1.9, are mainly found in the peripheral nervous system.[12]

In this project the Na_v 1.4 sodium channel has been studied. This structure has been determined recently (2018), and its elucidation opens the gate to study its channelopathies.[12] Since this channel regulates excitations that drive contraction in skeletal respiratory muscles, variants of this structure caused by mutation of the gene that codifies it (SCN4A) can cause diseases like myotonia, periodic paralysis, congenic paralysis, and myasthenic syndrome. It has even been hypothesized that overrepresentations of functionally disruptive variants of SCN4A favour sudden infant death syndrome (SIDS) as well as manifestations of infant life-threatening conditions like apnoea and laryngospasm.[27]

Molecules like Lidocaine and Ranolazine have been researched as therapeutic agents for Na_v 1.4 against prevention and restoration of cardiac arrythmias (heart beating at abnormal rhythm). Nevertheless, drugs aimed for this channel should consider molecular kinetic factors. The slower the ion channel blocker acts, the better the persistent Na^+ late currents are inhibited, stopping the arrythmia and allowing the channel to recharge for the next action potential. However, if the blockers act too fast, the action potential is prevented, increasing adverse effects[28]. Therefore, drug candidates need to be chosen carefully to prevent side effects and undesired toxicity.

The Na_v 1.4 has some structural particularities, as it can be seen in Fig. 2.5. The whole channel is determined as a complex formed by the alpha unit and a β-1 subunit which has two parts: A TM and an external domain (Ig). Voltage sensing involves four to six ARG/LYS residues on the TM4 of the VSD. The appearance of a short linker inserted in a hydrophobic cavity between domains III and IV is used for fast inactivation of the channel.[12]

In Fig. 2.5 a representation of the PDB structure[29] is shown with the same colouring code as for **¡Error! No se encuentra el origen de la referencia.** and **¡Error! No se encuentra el origen de la referencia.**.

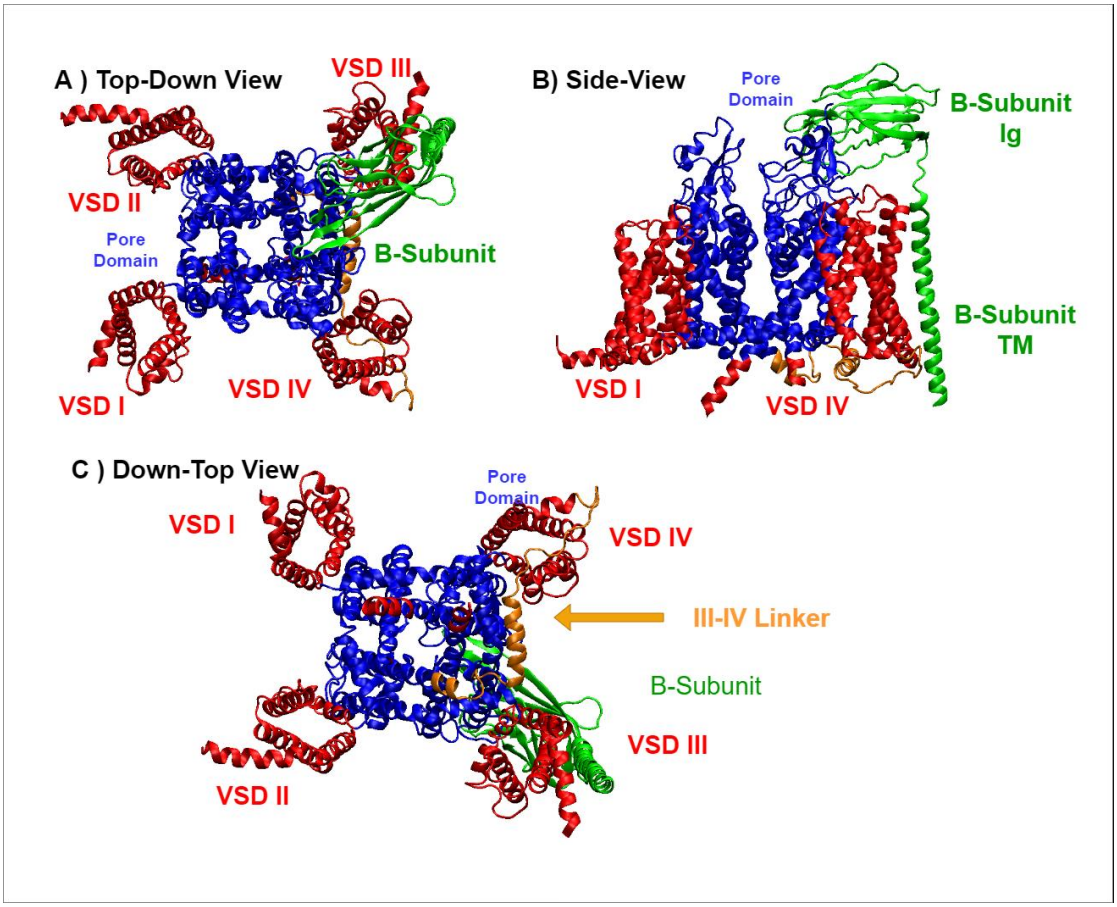

Fig. 2.5 Different Views of the Ion Channel PDB Structure[29] Using VMD[30]: Structure formed by the complex between the α structure (blue and red) and a β subunit (green). A) Top-Down view, that is, looking at the protein from outside the cell membrane side. For clarification it presents the same disposition as **¡Error! No se encuentra el origen de la referencia.**. B) Side-View, only VSDI and VSDIV are indicated since the remaining two are behind. C) Down-Top View, that is, looking from inside the cellular cytoplasm to the protein which is inserted into the lipidic membrane. From this view perspective III-IV linker is visible in the protein.

This is the ion channel investigated in the present project. It was chosen due to its recent structure determination, and its physiological relevance. Only the alpha domain of the whole channel, which is the region in charge of ion transport, was used for the MD simulations. The specifics will be seen in section 4.1.1.

2.3 Photo-sensitive molecules as Ion Channel Blockers

2.3.1 Light-Sensitive Functionalization of Biomolecules

Along the introduction some molecules have been mentioned as drug candidates for ion channels, for example Bupivacaine or Ranolazine[5]. A huge concern for pharmacology is the possible side effects or toxicity associated with a drug. This is mainly caused because drugs act with poor target specificity (no spatial control) and as soon as they reach potential receptors (no temporal control). Light can be controlled with spatiotemporal precision. Therefore, photopharmacology attempts to functionalize biomolecules in order to be selectively activated by photoexcitation.

There are two main approaches for photochemical functionalization of molecules. The first approach consists in protecting the key active functional group of the molecule with a light-sensitive part which "cages" and renders the molecule inactive. Upon excitation, the caging group is removed activating the key functional group of the molecule. This change is often irreversible and hence it is referred as "phototrigger". The second approach consists in causing structural changes, such as ring close/opening and isomerization, in molecules by accessing them using reversible photochemistry. Since the structural changes are reversible, this approach is denominated "photoswitch".

present active/inactive states which can be accessed by reversible photochemistry. This approach is denominated as "photo-switches" and involve structural changes like ring close/opening and isomerization. Fig. 2.6 shows examples of both approaches.

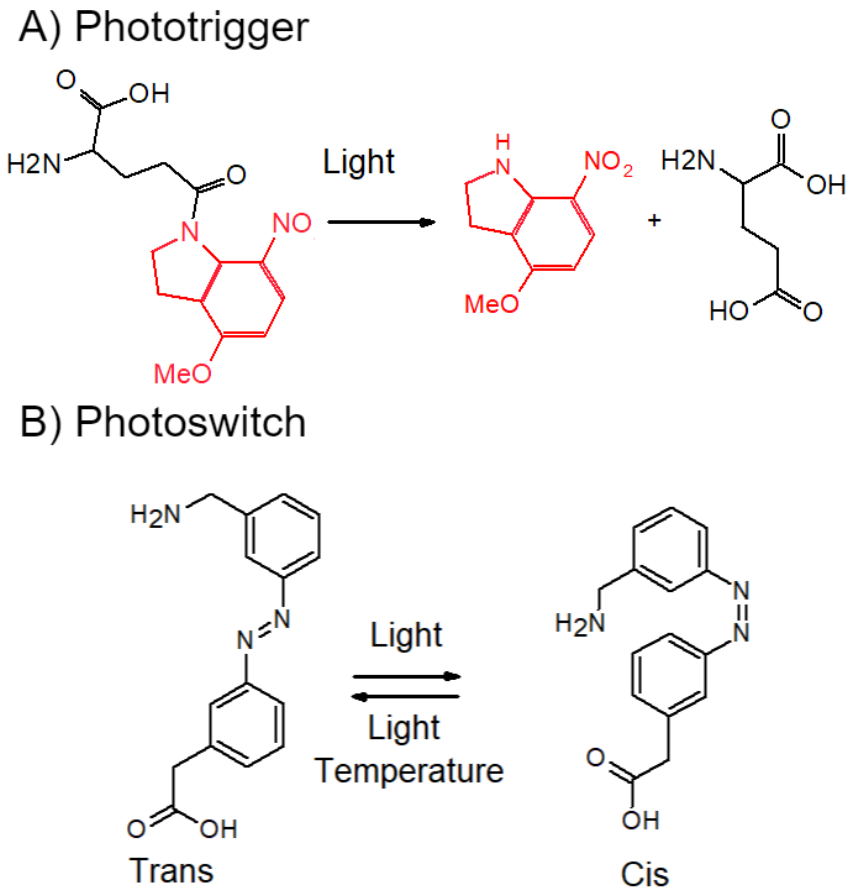

Fig. 2.6 Strategies for Functionalization of Photo-Sensitive Molecules : A) Phototrigger, light breaks the molecule in two pieces yielding an active drug molecule and an innocuous aminoacid (Glutamate).[31] B) Photoswitch, Azobenzene derivative that experiences *cis* – *trans* isomerization when excited by light.[9]

There are some considerations need to be taken into account when using light-sensitive molecules:

1. The molecule should absorb effectively (measured with molar extinction coefficient "ε" or high two-photon cross section) at wavelengths compatible with biological systems, between 340nm and up to 800nm.
2. The quantum yield of the light excitation should be high. In other words, the photochemical change (molecule breaking or isomerization) should happen with high efficiency to reduce light dosage.
3. Light-induced alterations should alter the biological function of the system in a substantial way.
4. Adding a light sensitive part to a molecule should be stable under biological conditions and non-toxic both after and before the excitation.

2.3.2 Azobenzene and Derivatives: Photoswitches as Ion Channel Blockers

Molecular photo-switches have been widely employed for modulation of biomolecules. Examples include : Photocontrol of peptide conformation and protein folding, modulation of ion channels, modulation of G protein-coupled receptors, photocontrol of enzyme activity, and development of photopharmacological drugs among others.[32] Photoswitches have been tested on ion channels , for example, to control neuronal excitability (making them promising therapeutic candidates for treating blindness and pathological pain).[8]

The most common photoswitches are azobenzene and its derivatives. They have been widely employed in several areas: modulation of ion channels in neurons[33],protein modulation[32], *"in vivo"* applications[34], restoration of blindness[35], soft materials[36] and many others[37]. This is not without reason, since azobenzene meets most of the criteria outlined in previous section 2.3.1. Some of the property highlights of azobenzene are:

- The trans to cis isomerization results in a drastic change in geometry (from planar to twisted (*trans-cis)* (Fig. 2.7 A)), and polarity.
- The adsorption spectra of the two isomers are different, allowing for conversion of cis to trans under proper light exposure (Fig. 2.7 B)).
- Azobenzenes are relatively simple and small, making them easily conjugable to various ligands.
- Photoisomerization occurs on a ps time scale, faster than biological processes (like ligand binding which can even take up to μs) allowing for temporal control.
- Azobenzenes are very photostable. Which allows for many photochemical cycles without fatigue of the molecule.
- The energetic barrier between two conformations is high enough to keep them isolated. In dark conditions with absence of light the trans conformation which is thermodynamically more stable predominates the mixture for about 99.99%.

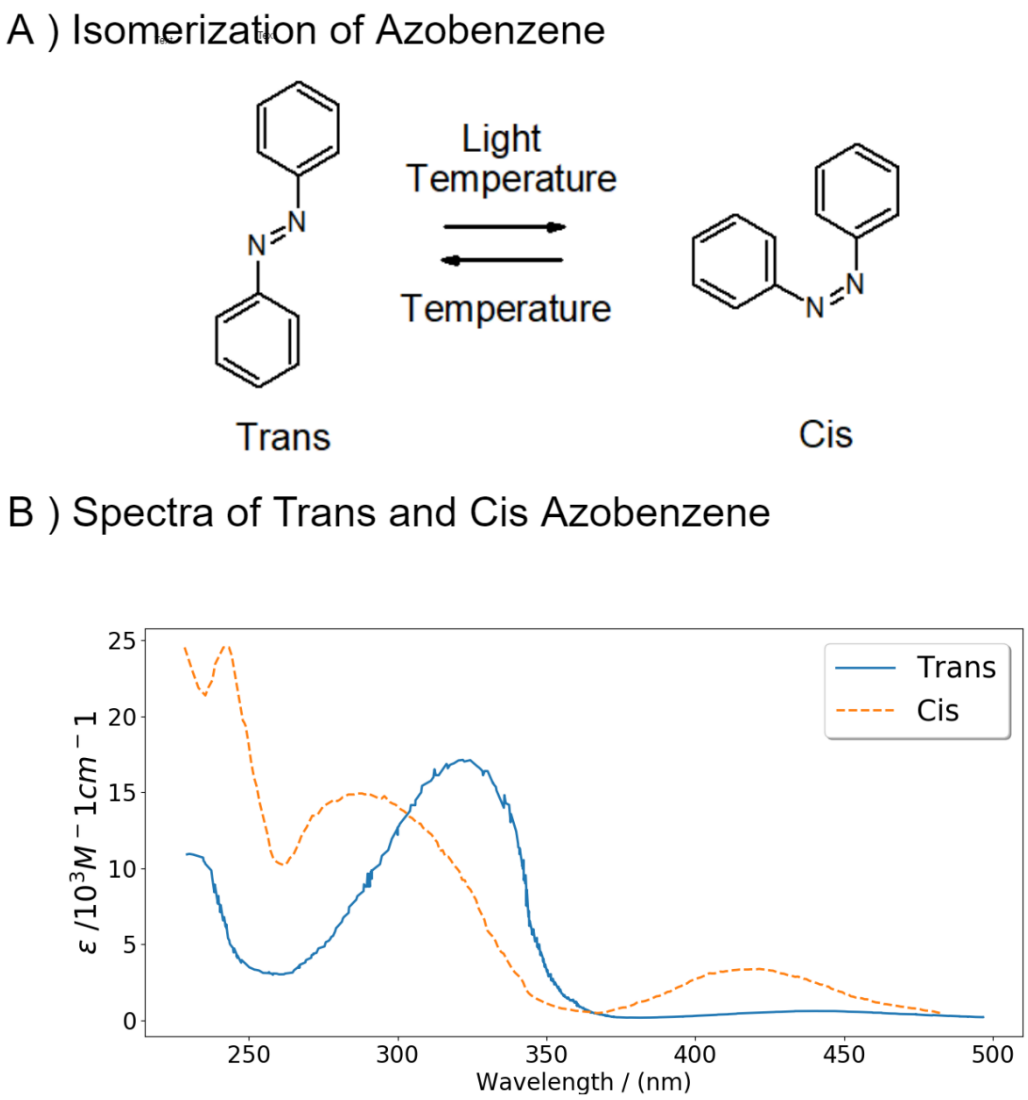

Fig. 2.7. Isomerization of Azobenzene and Spectra : A) Isomerization of Azobenzene and its conversion caused by photo or thermal excitation. B) Absorption spectra for *trans* and *cis* Azobenzene. As it can be appreciated that there is a shift in molar extinction coefficient between both conformations. Data taken from the NIST website both for *trans*-Azobenzene (NIST website[38] and original reference [39]) and for *cis*-Azobenzene (NIST website[40] and original reference[41]).

Azobenzene is found as a structural "scaffold" unit for many photoswitches. In other words, more complicated molecules present Azobenzene as a functional group which undergoes photoisomerization when excited with the right light pulse. Functionalization of Azobenzene with other chemical groups is done for different reasons. First, azobenzene presents mostly non-polar interactions due to its chemical nature. Therefore, the functionalization with polar groups allows for better binding with polar residues in proteins. At the same time, already existent biomolecules like peptides or lipids can be functionalized with azobenzene to change its biological activity.[42] Other functionalizations are used for development of molecular motors[43] as well as for construction of molecular sensors.[44] The second reason is that addition of functional groups to azobenzene causes a change in the electronic structure of the system, altering its photochemical properties and therefore its light induced isomerization process. For example, substitution by amino groups at *ortho* or *para* ring positions causes red shifts of the spectrum. Incorporating a donating group at *para* in one ring and an acceptor group at the *para*

of the other ring generates a "pull-push" situation which can lead to further red shifts in the spectrum.[9] Direct substituents of azobenzene should ,therefore, take into account the possible light absorption changes, examples of this can be found at the work published by Oleg Sadovski, Andrew A. Beharry, Fuzhong Zhang, and G. Andrew Woolley (2009) regarding "Spectral Tuning of Azobenzene Photoswitches for Biological Applications".[45] As a last remark, the mechanism that governs photoisomerization of azobenzene is not trivial and there are several studies that have tried to clarify it.[46,47]

Photoswitch modulation of voltage gated ion channels relies on blocking the channel when the switch is in one state, and unblocking it when accessing the other state of the switch by photo excitation. Depending on the photoswitch and the channel considered, the trans or the cis form will be the one that blocks the flow of ions (Fig. 2.8).

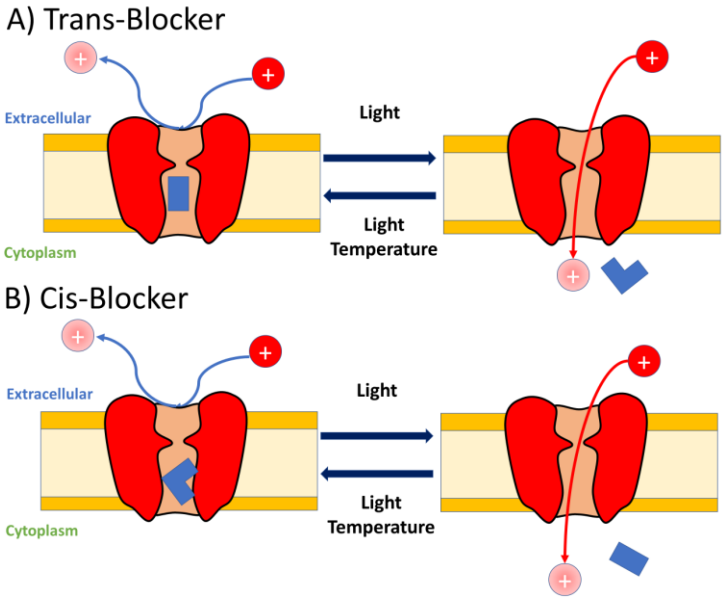

Fig. 2.8 Exemplification of Blocking of a Channel by an Azobenzene Containing Molecule: A) Case where the blocking occurs in trans conformation and unblocking occurs upon excitation. B) Case where the block occurs in cis conformation.

The blocking characteristics of the photoswitch and the binding affinity will be different for each ligand-channel complex.

2.4 Project Objective

The objective of the project is to study the binding mechanism and possible pockets between Azobenzene and the human Na_v 1.4. In order to do this, three steps are followed. First, the construction of a truncated model of the α structure of the ion channel with Azobenzene inside is done followed by an equilibration with classical MD. Second, an exploration of the conformational space with GaMD and determination of the possible binding pockets by reweighting the biased trajectories to obtain a free energy surface is done. Third, estimation of the free binding energy of each pocket and the description of the most relevant interactions using MM/GBSA.

3 Methods

3.1 Molecular Dynamics

In this section general concepts about Molecular Dynamics and its applications are given. The objective is not to deepen into the concepts and fundaments, but rather to give a general view of how these methodologies work. Special thanks to the research group's YouTube channel MoBioChem[48] .

3.1.1 Introduction to Molecular Dynamics (I): Dynamic Description of a Molecular System.

MD currently plays an important role in order to understand and predict properties of molecular systems and for predictive molecular design. The basic idea behind is to build a molecular system using the particles that compose it and propagate the motion of the system over time to investigate its evolution.[13] The first MD paper was published by Alder and Wainright in 1959.[49]

Regarding the description of a molecular system there are several points that need to be addressed:

- What are the fundamental units or particles of the system and how many of them are there?
- What are the mathematical forms of the interactions between particles?
- What is the fundamental law for evolving the system in time?

To describe the evolution of the system along time, the characteristics of mass and velocity need to be known. For velocities close to the speed of light, relativistic effects need to be considered. For lighter masses than 1 amu (atomic mass unit) (Hydrogen's mass, close to 1 proton) quantum effects need to be considered. The two criteria, mass and velocity, leave 4 equations for propagating the dynamics of the system: Dirac (Quantum + Relativistic), Schrödinger (Quantum + Non-Relativistic), Einstein (Classic + Relativistic), and Newton (Classic + Non-Relativistic) as seen in Fig 3.1 . Relativistic effects need to be considered for high velocities, such as the ones experienced by lanthanide electrons due to the strong attraction with the nucleus or in the field of high energy physics. However, this is far away from this project so it will be left aside. Out of the two remaining theories, Schrödinger and Newton, the use of the first one is denominated "Quantum Dynamics" (QD) and allows for the description phenomena like tunnelling or ultrafast processes. Newton's equations of motion are used to describe the dynamics of macroscopic objects.[50]

At the chemical level, the particles described are usually atomic nuclei and electrons. Treating such small entities in a simulation requires of the use of Quantum Mechanics (QM) for their proper description. Under Born-Oppenheimer's approximation, the degrees of freedom for nuclear and electronic motions can be decoupled since electrons move much faster than atomic nuclei. This changes the description of the system into solving Schrödinger's equation for the electrons and moving the nuclei in the consequent field. In order to give a description of electrons methods like "Hartree-Fock" (HF) or "Density Functional Theory" (DFT) can be applied.

Another approach is to treat atoms or groups of atoms, instead of electrons and nuclei, as building blocks of the simulated molecular model. The potential energy relationships between the atoms can be described by simple potential energy functions called force fields which gives the foundation for "Molecular Mechanics" (MM). Here, connectivity between atoms is modelled in a "ball-spring" fashion to simulate molecules. While the computational cost for QM methods scales fast and in general can only describe hundreds of atoms, MM can easily describe hundreds of thousands or even millions of atoms inside a simulation. Propagation of trajectories with QD results extremely useful when studying ultrafast processes[51,52], nevertheless, the computational cost for these methods scales fast with the number of atoms and the time of simulation desired. For studying slow processes like ligand binding (orders of ns or even μs) in a large system (like an ion channel) a more efficient method like MM is required[53].

The path chosen is to use atoms as the fundamental particles of the molecular model of the simulation and propagate them using Newton's equations of motion. This is done by assuming that all atoms, since they have higher mass than 1 amu, behave like classical macroscopic objects. This the foundation of MD, combining it with the MM philosophy (use of force fields), gives the term "Classical MD". This approach does not give an electronic description of atoms and molecules reducing the complexity of calculations. Consequently, classical MD is unable to describe essential chemical phenomena like bond breaking and formation or light-matter interaction. This is solved at the cost of increased computational complexity by introducing QM methods to calculate the forces acting upon the atoms (*ab initio* MD) or by a hybrid approach (QM/MM).

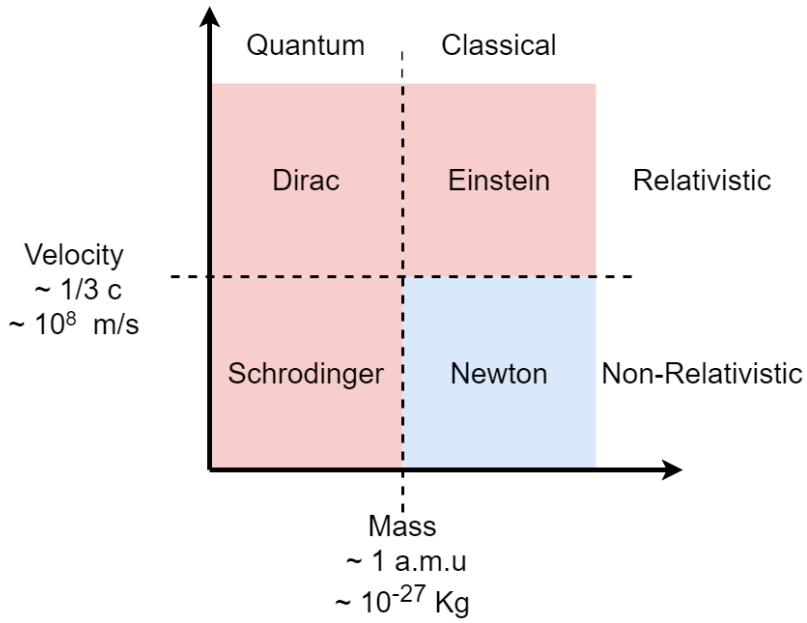

Fig. 3.1 Dynamic Equations*:* Depending on the speed and the mass of the particles Involved in the Simulation. Different physical models need to be applied to describe nature. In blue at the bottom-right Newton's equation and the area where MD simulations are performed.

3.1.2 Introduction to Molecular Dynamics (II): Applications

The term MD refers to a wide branch of methodologies under the approach of using atoms as building blocks and Newton' equations of motion for propagating the evolution of the system. It is largely employed in the fields of material science, structural biochemistry, biophysics, enzymology, molecular biology, pharmaceutical chemistry and biotechnology[54]. Some examples of publications employing MD are: "Structural Properties of Glassy and Liquid Sodium Tetra-silicate : Comparison Between Ab-Initio and Classical Molecular Dynamics"[55], "Anharmonic Force Constants Extracted From First-Principles Molecular Dynamics : Applications to Heat Trasnfer"[56], "Application of Molecular Dynamics Simulations to the Study of Ion-Bombarded Metal Surfaces"[57], and "Molecular Dynamics Simulations of Biomolecules".[58]

MD can study thermodinamical and time-dependent (kinetic) phenomena. This enables an understanding of molecular processes which are inaccessible otherwise. Nevertheless, MD trajectories only provide atomic positions, velocities, and single point energies. The computation of macroscopic properties also requires of the application of statistical mechanics.[54]

3.1.3 Thermodynamics and Statistical Mechanics Concepts

Both Thermodynamics and Statistical Mechanics are deeply interconnected and used to obtain information from MD simulations. While thermodynamics give meaning to macroscopic values like enthalpies, binding free energies, pressure and other quantities, statistical mechanics bridges these properties with the microscopic behaviour of a molecular system.

Statistical mechanics tend to focus on macroscopic thermodynamic properties and microscopic equilibrium behaviour. The cornerstone is Boltzmann's factor, which relates the probability of visiting a determined state with its energy (Eq. 3.1).

$$P(s) = \frac{e^{\left(-\frac{V(s)}{k_b T}\right)}}{Q} \qquad \text{(Eq. 3.1)}$$

Where $V(s)$ is the potential energy when visiting the state s, k_b is Boltzmann's Constant and T is the temperature of the system. The term "Q" is denominnated partition function, which is related to the number of states that are thermally accessible for the system. The thermodynamic properties of the system depend on the partition function.

Jumps from one state to another happen in a stochastic way, random and predictable only in terms of average behaviour. For simple systems, these jumps can be predicted quite easily, and transition rates can be computed. For complex systems with millions of states a large number of local minima in addition to diffuse energy barriers originates a very complex landscape where transitions between states become difficult to predict. The combination of all possible states of the system is denominated as "phase space".[54]

To study changes between states the concept of ensemble is introduced. An ensemble is a thermodinamical system with a specified volume, composition, temperature, and mass permeability, replicated a certain number of times. Different number of replicas of the system will be in different states of energy $V(s)$ and thus the whole ensemble will be in one specific configuration. By applying Boltzmann's factor (Eq. 3.1) it can be guessed that the lowest energy configuration will be most probable to be visited. At the thermodynamic limit (number of

particles of the system tends to infinity) the most probable configuration will dominate the properties of the system. In this fashion, Statistical Mechanics uses concepts like ensemble and partition function to derive and calculate macroscopic thermodynamic quantities from microscopic properties[59]. Different ensembles are characterized by the thermodinamical variables that remain independent (not to change) during their evolution. In Table 3.1 some examples are given.

When a measurement is done experimentally, the particles of the system will be in different states and therefore the result obtained will be a population weighted average of the properties of each stat (the ensemble average is obtained by infinite repetitions of the system). A MD trajectory is a single sample of a process where the system will visit some states during the course of the simulation, obtaining time averages for the properties of the system. If the time of sampling would be infinite, then time average and ensemble average would match, however, simulating infinite times a system or a single system for an infinite amount of time is not possible. Under the ergodic assumption, time averages are taken as ensemble averages even if sampling times are not infinite. In this way using MD trajectories and knowing the ensemble of the simulation, macroscopic thermodynamic properties can be derived such as internal energy, entropy, and free energy increments among others.[13]

Table 3.1 Popular Ensembles: Where N stands for number of particles, V is volume, T is temperature, P is pressure , E is energy, and μ stands for chemical potential. The Function at Equilibrium refers to the function that is minimized in that ensemble to reach equilibrium conditions.

Independent Variables	Name	Function at Equilibrium
NVE	Microcanonical	S, Entropy
NVT	Canonical	F, Helmholtz Free Energy
NPT	Isothermal-Isobaric	G, Gibbs Free Energy
μVT	Grand-Canonical	Ω, Grand Potential

The objective of this project is not to go too deep into Thermodynamics or Statistical Mechanics, but some basic knowledge is required for two concepts that are going to be discussed in the following section.

A last concept to cover is the calculation of free energy which is related with the probability distribution of each one of the states sampled. In other words, count how many times a simulation visits a state and normalize the amount by the total number of visits performed, after this, apply (Eq. 3.2) to obtain the increment of free energy for each state. The evolution of the free energy along a specific pathway is denominated as potential of mean force (PMF).

$$\Delta G\ (s) = k_b T \ln K\ (s) \qquad \text{(Eq. 3.2)}$$

Where K is the value for the probability distribution for the current state s.

3.1.4 Sampling of Phase Space

In the previous section, the concept of energy state was introduced. It was shown that the transitions between configurations in a thermodinamical ensemble are related with the relative probabilities of each state given by Boltzmann's Factor (Eq. *3.1*). The link between ensemble and

time-average was justified with the ergodic assumption. The idea of a link between MD trajectories and thermodynamic ensembles was introduced.

With this in mind, one question arises. Are MD simulations able to visit all relevant states even if there are energy barriers that prevent those states from being accessed? This is where the concept of "sampling" comes to play. How many possible states or configurations is the MD simulation able to visit during the evolution of the system? The more the number of states the simulation can visit, the better the sampling. Even for high probability states, large energy barriers will cause MD simulations to have a low probability of transition between them. Therefore, depending on the starting point of the system in the phase space, some parts will be sampled while other ones will not. In addition to this, low probability states will probably not be visited. Some of these situations can be seen in Fig. 3.2.

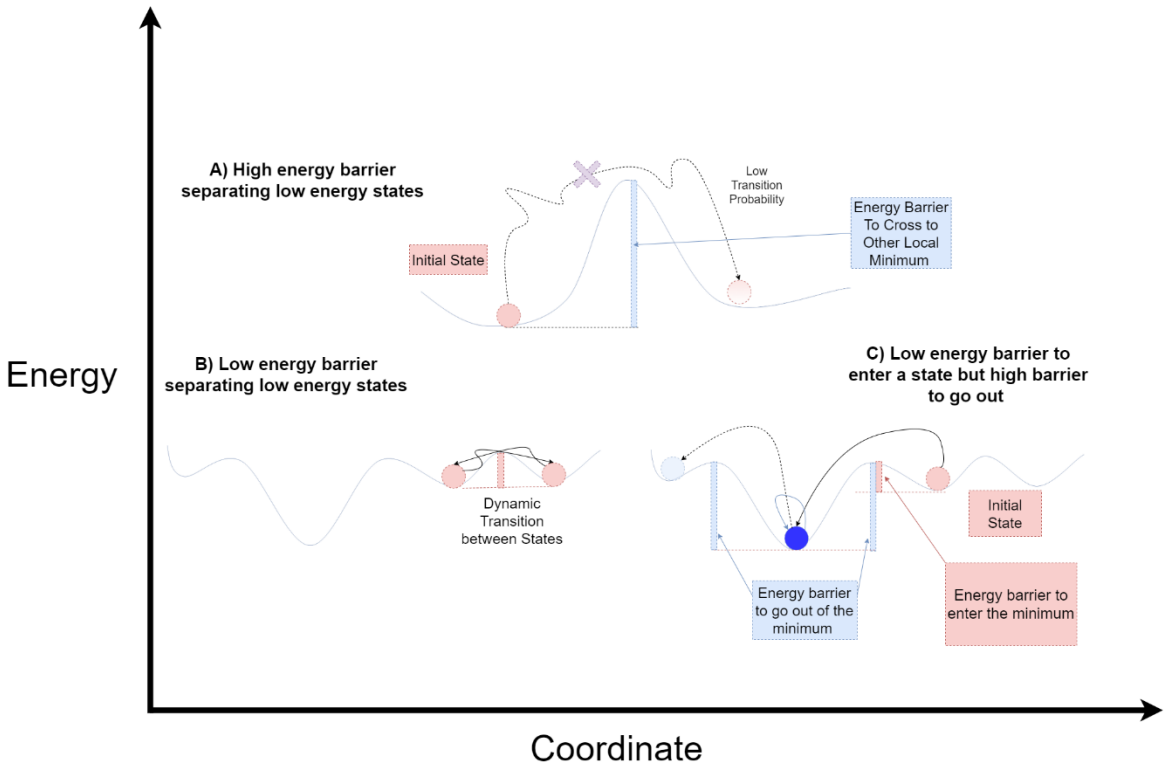

Fig. 3.2 Different Energy State Situations: A) High energy separation between two low energy states preventing sampling across the energy barrier. B) Low energy barrier between similar energy states allowing for dynamical transition between them. C) Initial barrier between states low but high energy barrier to come out. System will be trapped on the local minimum.

As shown in Fig. 3.2 there are different scenarios. If the energy barriers between two states are not very high and the energy of both states is similar, there will be a dynamic transition from one state to another B). Nevertheless, if two states have the same energy but a high separation barrier the transitions will have a low probability of occurring (A)). A system can cross a barrier and fall in a global or a very deep local minimum C). In this case, the energy barrier to come out will be larger than the ones crossed to fall in that minimum. Note that scaping from the minimum is not impossible, but the probability is low.

When performing an MD simulation not all the relevant states are probably visited causing less sampling of the phase space. To avoid this, enhanced sampling techniques are used, some examples are: Replica Exchange Molecular Dynamics, Metadynamics, Umbrella Sampling, and

Simulated Annealing[60]. Different enhanced sampling techniques are divided into two groups. First, the techniques that are applied when the reaction coordinate (RC) to explore is known (Eg: Umbrella Sampling[61]) . The second group is applied when the RC is not known (Eg: GaMD[62]).

In this project GaMD was used to accelerate the simulations and to sample, in principle, a larger fraction of the phase space than when using conventional MD.

3.1.5 Equations of Motion.

As stated before, Molecular Dynamics employs Newton's Equations of Motion to study the evolution of the system along time. Therefore, to understand Molecular Dynamics, the starting point is to describe the Equations of Motion which act as the main cornerstone of this methodology.

Given a system composed of *N* atoms, atom i experiences a force F_i which can be decomposed into a mass m_i and an acceleration a_i. This is the second Newton's law and can be developed further considering that the acceleration of an object is the derivative of the velocity v_i with respect to time. Moreover, the momentum of an object p_i is defined as the mass times the velocity. These relations can be seen in (Eq. 3.3).

$$F_i = m_i * a_i = m_i \frac{dv_i}{dt} = \frac{dp_i}{dt} \qquad \text{(Eq. 3.3)}$$

For a conservative system, one in which the potential energy depends only on the positions of the atoms but not in time, the force exerted on atom i can be described as the negative gradient of the potential energy V with respect to the coordinates $q_i$of the atom as seen in (Eq. 3.4*)*.

$$F_i = -\frac{dV}{dq_i} \qquad \text{(Eq. 3.4)}$$

By combining (Eq. 3.3) and (Eq. 3.4) into (Eq. 3.5) one gets the first Hamilton's equation. This equation accounts for the time evolution of the momenta depending on the gradient of the potential energy.

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = -\frac{dV}{\mathrm{d}q_i} \qquad \text{(Eq. 3.5)}$$

The classical kinetic energy "T_i" of an atom i of a system can be expressed in terms of its mass and velocity, or linear momentum and mass as seen in (Eq. 3.6).

$$T_i = \frac{1}{2} m_i v_i = \frac{1}{2} \frac{m_i^2}{m_i} v_i^2 = \frac{p_i^2}{2m_i} \qquad \text{(Eq. 3.6)}$$

In order to relate the time evolution of the coordinates with respect to time, the derivative of the kinetic energy with respect to the linear momentum of the atom is calculated as seen in (Eq 3.7).

$$\frac{\mathrm{d}T_i}{\mathrm{d}p_i} = \frac{2p_i}{2m_i} = \frac{p\mathrm{i}}{m_i} = \frac{m_i \cdot v_i}{m_i} = v_i = \frac{dq_i}{dt} \qquad \text{(Eq 3.7)}$$

Reorganizing (Eq 3.7) one gets the second Hamilton's equation (Eq 3.8).

$$\frac{\mathrm{d}q_i}{\mathrm{d}t} = \frac{\mathrm{d}T_i}{\mathrm{d}p_i} \qquad \text{(Eq 3.8)}$$

Both equations (Eq. 3.5) and (Eq 3.8) account for the time evolution of the momenta and coordinates of every atom in the system[63].

A fundamental drawback is the fact that both equations cannot be solved analytically. For this purpose, different numerical methods have been used.

3.1.6 Numerical Integration of the Equations of Motion in Molecular Dynamics.

In order to solve the equations of motion in a coupled way, numerical integration is used. This approach is not specific of MD simulations, but it is found in a ubiquitous way in the field of calculus.[64]

The philosophy behind the numerical integration is to solve iteratively the equations of motion for a given time-step. For a given position $q(t)$ and a velocity $v(t)$ an update is done to know the properties of the system after a time increment Δt. In this way, the new set of coordinates $q(t + \Delta t)$ and the new set of velocities $v(t + \Delta t)$ are determined. With the new set of coordinates calculated one can use these solutions as new inputs for the integration algorithm, giving a description of the trajectory of the atoms along time.

Several algorithms have been developed with the objective of numerically integrating the equations of motion. Three of them can be seen in Table 3.2.

Table 3.2 Common Integration Algorithms Used in MD

Beeman:	$q_{n+1} = q_n + v_n \Delta t + (4a_n - a_{n-1})\, {\Delta t^2}/{6}$ $v_{n+1} = v_n + (2a_{n+1} + 5a_n - a_{n-1})\, {\Delta t^2}/{6}$
Leapfrog:	$q_{n+1} = q_n + v_{n+1/2}\Delta t$ $v_{n+1/2} = v_{n-1/2} + a_n \Delta t$
Verlet:	$q_{n+1} = q_n + v_n \Delta t + a_n \Delta t^2/2$ $v_{n+1} = v_n + (a_{n+1} + a_n)\Delta t/2$

Where n is the current time-step iteration (t) , $n + 1$ is the new time step to be calculated ($t + \Delta t$) and $n - 1$ is the previous to the current time-step iteration ($t + \Delta t$). In addition, q stands for atom coordinates, v for atom velocities, and a for atom accelerations. Note that in the Leapfrog algorithm velocities are updated each half time-step iteration.

Apart from the integrators seen in Table 3.2) other ones are used such as the Runge-Kutta or the "Semi-Implicit Algorithm". Each integrator gives different performance and numerical accuracy when used. The comparison between algorithms is beyond the scope of this project, nevertheless there are differences in between the performance and numerical accuracy that they present. The Verlet and the Leapfrog algorithms are easy to implement, computationally cheap, and probably the most popular ones.[65]

Regarding the integration algorithm used in this project, the AMBER18 software with Langevin Dynamics employs a simple Leapfrog algorithm[66]. Nevertheless, as it will be discussed, propagating a dynamic trajectory is not only integrating the equations of motion but also taking into account other factors like a barostat (to control pressure) and a thermostat (control the temperature of the simulation).

In summary, the positions and velocities are iteratively calculated and, therefore, are used as an input-output of the algorithm. However, the acceleration terms of the integrators seen in Table 3.2 are necessary for the calculations but the algorithms themselves do not give a describe these terms. The following question arises: How to calculate the accelerations for each time-step? To answer this question lets re-examine what has been seen previously in this chapter about acceleration. The acceleration is defined as the derivative of velocity with respect to time. It determines the magnitude and the orientation of the variation of the velocity and it is related with the force exerted on the system as seen before in Newton's equation (Eq. 3.3). Moreover, the force exerted on the atoms is the negative gradient of the potential energy with respect to the positions of the atoms, as shown in (Eq. 3.4). Therefore, there is a need to compute the potential energy of the system in order to calculate the forces and therefore the acceleration of them atoms. Different approaches for potential energy calculation will be discussed in section 3.1.7.

In summary, what it is needed for the integration of the equations of motion is, a set of coordinates and velocities for the integration algorithm and a way to calculate the potential energy of the system.

For any point in the middle of the trajectory coordinates and velocities come as a result of the previous iteration of the algorithm. For the initial iteration, coordinates of systems are chosen from theoretical (QM or MM Optimizations) or experimental data. Initial velocities are assigned randomly to each atom from a Maxwell-Boltzmann distribution (Eq. 3.9) for a given temperature. This is done assuming thermal equilibrium between all the regions of the simulation[63].

$$P(v) = \left(\frac{m}{2\pi k_b T}\right)^{3/2} 4\pi v^2 e^{-\frac{mv^2}{2k_b T}} \qquad \text{(Eq. 3.9)}$$

3.1.7 Methods for Computing the Potential Energy of the System

As seen in the previous section there is a need for computing the potential energy of the system to calculate the forces underwent for each atom. There are different ways to do this:

1. Using force fields, MM Approach: Force fields are mathematical expressions that account for the interactions between atoms in the system (MM description). Depends on theoretical/experimental parametrization. This is referred as classical MD.
2. Employing QM Methods: the idea behind is to solve Schrödinger's equation. Applying quantum mechanics, the potential energy of a system can be determined by solving its electronic structure since this will determine both the bonding interactions between atoms and the non-bonding interactions. This approach is referred as ab-initio Molecular Dynamics.
3. Mixed Quantum-Mechanics / Molecular-Mechanics (QM/MM): Employing both a QM description for a part of the system and a force field description for the rest.

Each method has its own advantages and drawbacks.

Force fields are simple to compute but they rely on high parametrization. Consequently, the election of force-field will influence the results of the simulation. For example, force fields optimized for globular proteins fail to properly describe properties of intrinsically disordered proteins[67]. One huge drawback that most of force fields rely on bond and angle description by harmonic potentials and, therefore, are not able to describe bond breaking and formation in molecules, rendering them useful for structural applications but not for simulating chemical processes which involve reorganization of atoms.

QM methods give, in principle, a more accurate and universal description of the potential energy of the system. This statement must be taken with care, since inside the QM field there are a lot of different methods that have their own particularities. Even inside the QM methods the use of empirical or theoretical parameters are widely used. Moreover, the selection of QM methods (Hartree-Fock (HF) , density functional theory (DFT) or semi-empirical methods) to study the same phenomenon like hydrogen bonding or stacking of DNA bases will not yield the same results[68]. Moreover, the computational cost for QM methods is much larger than that of the use of force fields, rendering QM methods not practical when dealing with simulations with thousands or hundreds of thousands of atoms (which, for example, has been the case in this project). Evolving for long times. Nevertheless, they can compute bond breaking in molecules, making them useful for simulations that consider reorganization of atoms inside a system.

QM/MM methods offer an intermediate approach between the ones mentioned above. While one part of the system (usually reduced and specific) is described via the QM methodology, the rest is described via a force field approach. QM/MM result useful when wanting to compute an MD simulation of a system where some part needs to be described with QM methodology such as when studying light matter interaction properties or bond breaking in a specific part of a system. An example of a publication of light interaction is "Study of the Quantum Yield of Tryptophan in Proteins"[69] , while one involving bond breaking "Modelling of the Ras-Gap of Guanosine Triphosphate"[70]. Both examples have to do with proteins, and it is not without a reason, although just a specific part of a protein may be involved with light interaction or bond breaking, the description of the environment is necessary for correct prediction of phenomena.

QM/MM methods are highly employed in biological systems since the size of the simulation is large enough not to be treated with QM methods but require QM description in order to study processes such as biological catalysis[71,72].

In the following section, the Force Field approach is going to be described more in detail since it has been the method used in this project (MM approach, Classical MD).

3.1.8 Force Fields

A force field is a sum of simple mathematical equations that relate the positions of the atoms with the potential energy of the system. Each one of the mathematical equations accounts for a specific interaction inside the system (bond energies, angle energies, etc). Instead of referring to the positions of the atoms, it is usually referred to the internal coordinates of the molecules and atoms that compose the system (bond distance, angles, relative distances between pairs of atoms). The general form of a force field can be seen in (Eq. 3.10).

$$V_{FF} = \sum V_{Bond} + \sum V_{Angle} + \sum V_{Dihedral} + \sum V_{VdW} + \sum V_{El} + \sum V_{Other} \qquad \text{(Eq. 3.10)}$$

Where V_{Bond} accounts for the energy required for stretching the bond connecting two atoms inside a molecule, V_{Angle} represents the energy needed to bend an angle inside a molecule, $V_{Dihedral}$ is the energy required to rotate a dihedral angle formed by 4 atoms covalently connected inside a molecule (torsion energy), E_{VdW} describes the non-bonded Van der Waals interactions between atoms, V_{El} is the electrostatic interaction of atoms with different partial charges, and finally V_{Other} accounts for other interactions which may or may not be present in the description of the force field (improper, coupling between bond-angle contributions, …).[50] A schematic representation of these interactions can be seen in Fig. 3.3.

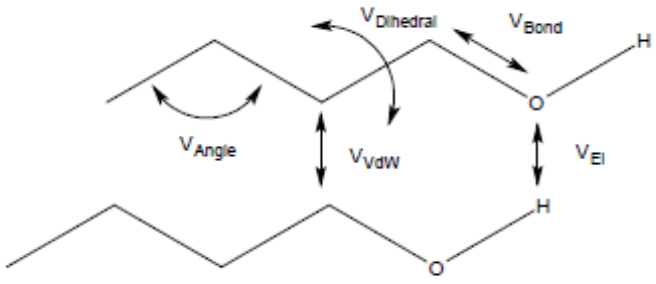

Fig. 3.3 Sketch of the Different Force Field Contributions: Example done with two molecules of butan-1-ol. This figure is representative. Note that while non-bonded interactions are indicated between molecules, they can also appear between parts of the same molecule.

Depending on the force field, each term has a different mathematical expression for each energy contribution and more or less terms are included inside V_{Other}. The most common expressions used in general Force Field are shown below.

The term V_{Bond} accounts for the stretching or contraction of a bond inside the molecule. Indeed, it gives the idea of molecules being balls (atoms) joined by springs (bonds). It is usually expressed in a simple harmonic oscillator way (Eq. 3.11) although it can be a more complicated expression like a Morse oscillator.

$$V_{Bond} = \frac{1}{2} k_{bond} (r - r_0)^2 \qquad \text{(Eq. 3.11)}$$

Where k_{bond} stands for the force constant and r_0 as the equilibrium distance of the bond.

Here, the fact that MM cannot account for bond breaking and formation can be noticed. While the real attractive potential between two atoms decays to zero at infinite distance, i.e., has an anharmonic behaviour, in the harmonic oscillator expression a bond stretched at infinite distance gives infinite potential energy. While this may be good for modelling situations near the equilibrium distance, it fails to model a large distortion of the bonds. Graphically, it has the form of a parabola (quadratic equation).

The term $\boldsymbol{V_{Angle}}$ is modelled again in a harmonic oscillator way (Eq. 3.12).

$$V_{Angle} = \frac{1}{2} k_{angle} (\theta - \theta_0)^2 \qquad \text{(Eq. 3.12)}$$

In this case k_{angle} accounts for force constant of the angle. That is, the resistance of the angle to be displaced from its equilibrium value θ_0.

The term $V_{Dihedral}$ accounts for the energy needed to make a torsion around a bond modifying a dihedral angle composed by 4 consecutive atoms covalently bonded. It is described by a Fourier potential (Eq. 3.13).

$$V_{Dihedral} = k_{torsion}(1 + \cos(n\omega - \gamma) \qquad \text{(Eq. 3.13)}$$

Where $k_{torsion}$ is the force constant of the rotation (height of rotation energy barriers), n is the number of minima in the potential, and finally γ is the phase of the rotation.

Up until here, the terms discussed are denominated “bonded interactions” since they account for energy terms between atoms inside a molecule that are connected by chemical bonds. The following terms are denominated non-bonded and they do not need a chemical bond to appear. They account for attraction/repulsion interactions between atoms, inter molecularly or intra molecularly with a usual separation of by 3 or more bonds.

The term V_{VdW} accounts for inter or intra molecular van Der Waals interactions between atoms. It is computed using a Lennard-Jones 12-6 potential (Eq. 3.14). The Lennard-Jones potential is repulsive for short distances, attractive around an equilibrium value and decays to zero when the separation increases. This behaviour can be seen in Fig. 3.4.

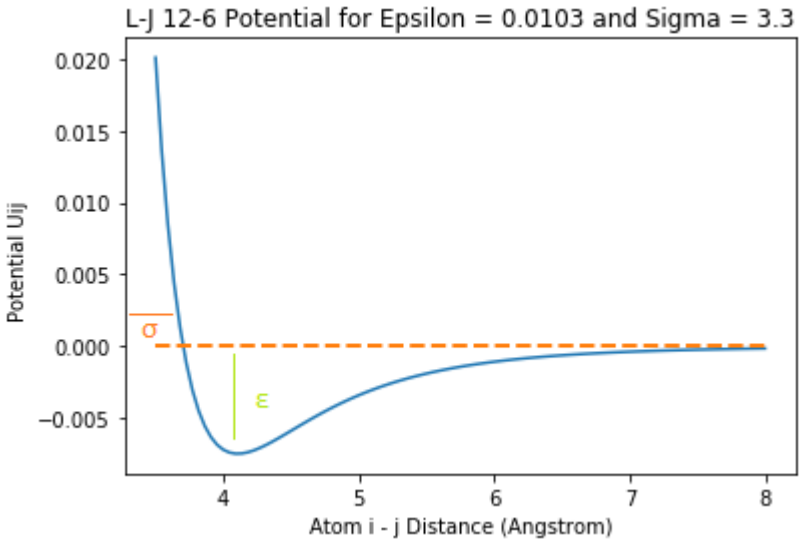

Fig. 3.4 Lennard Jones Potential: Potential Described by (Eq. 3.14*)*. Lennard-Jones equation implemented in Python 3.6[17] and plotted with Matplotlib[73] library.

$$V_{vdW\,ij} = 4\ \varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \qquad \text{(Eq. 3.14)}$$
$$Where\ \ \varepsilon_{ij} = \sqrt{\varepsilon_i\,\varepsilon_j}\ \ and\ \sigma_{ij} = \frac{\sigma_i + \sigma_j}{2}$$

The potential is calculated for the pair of atoms i and j. The term σ is a parameter that accounts for the position at which the potential energy is zero and the term ε accounts for the depth of the well. As it can be seen σ parameters are taken for each atom and then the arithmetical mean is done for both. For ε the geometrical mean is done instead.

The term V_{El} accounts for electrostatic interactions between different atoms. It is modelled with the Coulomb interaction potential (Eq. 3.15).

$$V_{El} = \frac{Q_i Q_j}{4\pi E r_{ij}} \qquad \text{(Eq. 3.15)}$$

Here, Q_i and Q_j refer to the partial charges of the pair of atoms involved, E is the electric permeability of the medium, and r_{ij} the distance between both atoms.

The term V_{other} includes the rest of contributions which are not always present. While the contributions described until this point are general of all force fields, there could be other terms added. Some of these terms consider out of plane bending (improper torsions) and terms which account for couplings between bond stretching, angle stretching and dihedral torsion. The addition of each term and the mathematical expression depends on the specific force field considered.

Although a priori force fields are useful in many cases and easy to implement, some aspects of them need to be considered carefully:

- First, bonds angles and dihedrals may look like a very chemically intuitive picture, nevertheless, this is just a model and they do not have physical meaning. In quantum mechanics, concepts like bonds do not exists as an entity and, therefore, using them imposes sacrificing part of the real behaviour of the system.

- Second, force Fields rely extremely on parametrization. Each atom inside the molecule will need to have a separate set of parameters to fill the force constants, the equilibrium distances/angles, the σ and ε values and partial charges. This distinction is not only done between different elements but for one same element for different kinds of atoms. For example, a C-C bond will not have the same characteristics than a double bond C=C or a C-C bond in a benzene ring. They will have different properties like the orientation of substituents (angles) and bonds strengths. Of course, this will also cause changes in the Lennard Jones parameters and partial charges. This is exemplified in Fig. 3.5.

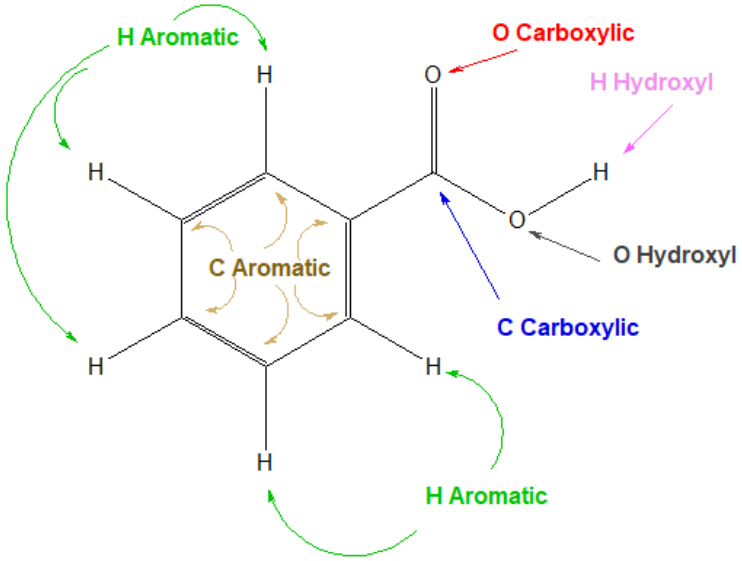

Fig. 3.5. Different Types of Atoms Inside a Molecule: Here, atoms are labelled differently according to their chemical nature inside the molecule of Benzoic Acid.

Because of this parametrisation, different force fields are used for different systems. Most of the force fields treat in the general way described above, harmonic terms for bonds and angles, Fourier series for torsions, Lennard-Jones, and a Coulomb potential for pair interatomic interactions. Main differences between Force Fields originate from the approaches taken to derive the parameters. Minor differences include how the software packages handle technical details in long range electrostatics or interactions between atoms in the same molecule for different force fields. Moreover, parameters for a given atom type in one force field cannot be compared with the values for another force field, generally causing non-transferability of parameters.

Some of the commonly used force fields are: CHARMM, AMBER, GROMOS, OPLS, AIREBO. Each Force Field has different versions according to the year of release and update. Each one of them has its own characteristics making them useful for different kinds of simulations.[74]

The last aspect to consider is the description polarizability. Since atomic charges are usually fitted to meet experimental or theoretical criteria and remaining static, they are not able to describe properly effects of fluctuations in the charge distribution or polarization. For this purpose, two approaches are taken: First, the thought of parametrising force fields so that they

would include implicitly these effects. Second, to include explicitly polarizability in the force field. The latter approach gives the concept of Polarizable Force Fields. These force fields try to implement charge polarization effects using induced dipole moments, fluctuating charge models or Drude oscillator models. Nevertheless, the use of these methods includes more computational effort and some of the underlying physical models need improvement, regarding specially charge transfer, charge penetration and short-range interactions.[75] Some of the polarizable force fields are AMOEBA[76] or the CHARMM Drude Force Field.[77]

Having introduced the equations of motion, their integration, and the potential energy calculation, in the following other concepts which are necessary for MD simulations will be discussed.

3.1.9 Periodic Boundary Conditions

MD can simulate a range of experimental conditions. Nevertheless, the size of the systems that can be simulated is far from real objects. While MD simulations can simulate about an order of 10^5 - 10^6 particles, real objects have an order of moles, where 1 mol is $6.022*10^{23}$. Therefore, to recreate bulk conditions an approximation is done[78]. Periodic boundary conditions duplicate infinite times in all directions the simulated system as seen in Fig. 3.6.

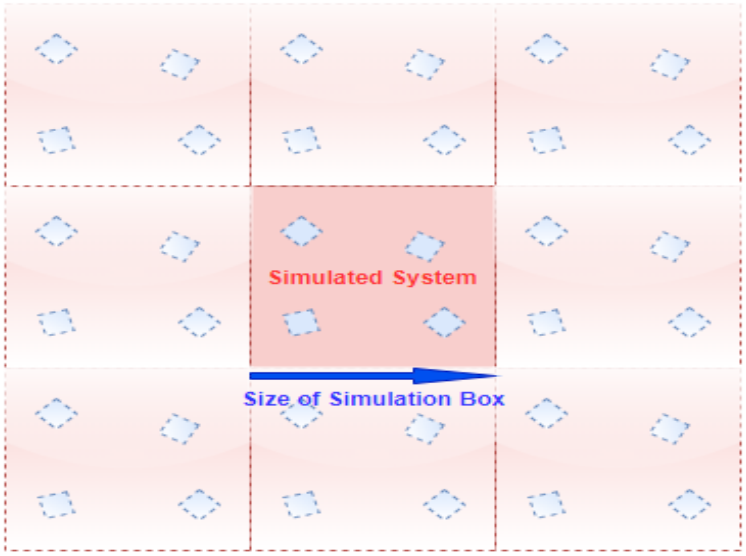

Fig. 3.6. Scheme of Periodic Boundary Conditions

In this way if a particle leaves the main simulation box, it will re-enter from the opposite side from a replicated box. The second achievement is to ensure that the interatomic interactions are satisfied in all directions giving the sense of bulk conditions. Nevertheless, since it is not desirable to count particle-particle interactions multiple times, a cut-off is imposed to the interatomic interactions. This is mainly done for the computation of van der Waals interactions since the decay of the potential is quite fast. In addition to this, before the cut-off distance, a switch function is applied to make the potential reach to zero faster. For the electrostatic interactions a cut-off is not desirable since truncating these interactions would cause worse artifacts than including interactions with multiple copies of the same particle. Instead, what are often termed “cu-toffs” for electrostatic interactions are shifts from short-range to long-range treatments .[13]

The calculation of electrostatic potentials is not trivial and is evaluated via techniques such as Ewald Summation. This methodology uses a Fourier transform in space to simplify and accelerate the convergence electrostatic calculations.[79]

Typically, a cubic lattice or rectangular is used for the replication, but it is not restricted to this geometry and other systems may be used. Other geometries are principally like truncated octahedrons are used in order to minimize the number of particles that interact with each other, saving computational resources.[54]

3.1.10 Temperature Control

MD simulations are used to observe and obtain properties of some system of study. This includes emulating experimental conditions. For this purpose and recalling what was shown in section 3.1.3 about thermodinamical ensembles, entities called thermostats are included in order to control the temperature of the simulation. The physical foundation for these methods is the thermodinamical relationship between the velocity of the particles of a system and the temperature. This is explained via the "Equipartition Theorem" (Eq. 3.16).

$$\frac{3}{2}Nk_bT = \langle \sum_{j=1}^{N} \frac{1}{2} m_j v_j^2 \rangle \qquad \text{(Eq. 3.16)}$$

The brackets indicate time average over the summation. If instead of time-averaging the summation it is taken as an individual value for one snapshot of the simulation, it is called instantaneous temperature. Instantaneous temperatures will not be equal to the time average quantity and at least in the canonical ensemble they will fluctuate around the desired temperature of the simulation.

Thermostats work by altering the Newtonian Equations of motion which belong inherently to the microcanonical ensemble (NVE, constant energy). Therefore, their application drifts away the simulation from Newtonian dynamics. To calculate dynamical properties of systems like such as diffusion coefficients, thermostats should be turned off. While all of them give non-physical dynamics, some of them affect less the calculation of some dynamical properties. Thermostats can be divided into deterministic and stochastic depending on the use of random numbers. They can be also classified as global or local depending on their effect on the whole simulation or only on a specific set inside of it.

Some thermostats operate by rescaling velocities inside the simulation without considering the equations of motion. Not all the velocities are rescaled in each iteration, otherwise Newtonian Dynamics would be deformed so much that the result would be a set of random collisions. Other thermostats use an implicit bath of particles or explicit degrees of freedom in the equations of motion.

A simple Thermostat is the "Simple Velocity Rescaling" where the momenta of the particles are rescaled so that the instant temperature of the simulation would match the desired temperature. It gives highly non-physical dynamics and causes undesired artifacts. A list of popular thermostats are : Gaussian, Berendsen, Andersen, Langevin, and Nosé-Hoover thermostats.[13]

3.1.11 Pressure Control

In MD simulations it is common to calculate the pressure of a system or monitor it to perform simulations under the (NPT) ensemble. The method to calculate the pressure is to sum all the forces on all atoms as described in the virial equation (Eq. 3.17).

$$P = \frac{Nk_bT}{V_{olume}} + \frac{1}{3V}\langle \sum_{i=1}^{N} q_i F_i \rangle \qquad \text{(Eq. 3.17)}$$

Where P is the pressure of the system, k_b is Boltzmann's constant, V_{olume} is the volume of the system, q_i and F_i are the position and the force applied over atom i.

For pair-additive Force Fields (probably most common used form) the virial equation is usually written as shown in (Eq. 3.18).

$$P = \frac{Nk_bT}{V_{olume}} + \frac{1}{6V}\langle \sum_{i=1}^{N} \sum_{j \neq i}^{N} r_{ij} F_{ij} \rangle \qquad \text{(Eq. 3.18)}$$

Where $r_{ij} = q_i - q_j$ and F_{ij} is the force on atom i due to atom j.

Equation (Eq. 3.18) is the most common expression to compute the pressure. When applying periodic boundary conditions there are some precautions to be taken. On the one hand, if the force field used is pair-additive, no problem is found. On the other hand, if the Force Field used is not pair-additive, under periodic boundary conditions, (Eq. 3.18) does not hold. Instead, a solution is to evaluate numerically the thermodinamical definition of pressure (Eq. 3.19).

$$P = -\left(\frac{dE}{dV}\right)_T = \frac{Nk_bT}{V_{olume}} - \left\langle \frac{dV}{dV_{olume}} \right\rangle \qquad \text{(Eq. 3.19)}$$

Where the term $\frac{Nk_bT}{V_{olume}}$ stands for the kinetic energy contribution to the pressure and the $\left\langle \frac{dV}{dV_{olume}} \right\rangle$ term is the potential energy contribution to the pressure. The potential energy is noted as V.

The expressions seen above are the fundament for pressure measurement in an MD simulation[80].

In an analogous way to temperature, pressure is given as a time averaged quantity. Moreover, it is possible to calculate the instantaneous pressure for a snapshot of the simulation in the same fashion. For the NHP and NPT ensembles the instantaneous pressure will oscillate around the desired pressure of simulation.

The general idea of controlling pressure is that barostats work as a fictitious piston on the system. The piston enclosures the simulation and acts in all directions uniformly. In this way, it applies a uniform compression/expansion in the particles of the simulation which will change how they interact with the "enclosure" of the simulation. The impacts of the particles cause a stress in the box of simulation and serve as a barostat.

A simple barostat to implement is the simple volume Rescaling. It works in an analogous way to the velocity rescaling thermostat seen in section 3.1.10. The Volume of the system is rescaled so that the instantaneous pressure of the simulation matches the desired one. Although easy to implement it does not sample properly the system and gives artifacts because of its use. Some popular Barostats are : Berendsen, Andersen, Parrinello-Rahman, Monte Carlo[13].

3.2 GaMD

Accelerated MD (aMD) is an enhanced sampling technique which applies a non-negative bias potential to the potential energy surface of biomolecules which accelerates transitions between conformations. In contrast to other enhanced sampling techniques it does not require a chosen RC which makes it useful for sampling the conformational space of a system without previous knowledge of imposed restrains[14]. Applying aMD alters the potential energy of the system and therefore the system will have a greater number of occurrences in conformations which would not be observed in a conventional MD (cMD). To recover the original free energy surface a reweighting is done by knowing the weight of the potential bias added. [81]

GaMD provides the acceleration and the lack of definition of a RC given by aMD while at the same time adding harmonic potentials which follow a near-gaussian distribution, improving energy reweighting of the free energy surface and reducing noise in the calculations. Since the aim of this project is to study the binding of a small organic molecule (Azobenzene) with a membrane protein (human Na_v 1.4 channel), RC are unknown, justifying the use of GaMD.

The core idea behind GaMD is to smooth the potential energy surface of biomolecules by adding a harmonic boost potential (which should resemble the most a near-gaussian distribution for accurate reweighting). For a system with "N" atoms at positions $\vec{q} = \{q_1 \dots, \vec{q}_N\}$ a boost potential is added for all values lower than a threshold E (Eq. 3.20).

$$\Delta V(\vec{q}) = \frac{1}{2} k_V (E - V(\vec{q}))^2 \quad V(q) < E \qquad \text{(Eq. 3.20)}$$

Where "ΔV" is the potential boost added, k_V is the harmonic force constant for the boost, E is the threshold, and $V(\vec{q})$ is the potential energy that the system has with the atoms in positions $\vec{q}$.

The total potential of the biased system is then $V^*(\vec{q}) = V(q) + \Delta V(\vec{q})$ as seen in (Eq. 3.21).

$$V^*(\vec{r}) = V(q) + \tfrac{1}{2} k_V (E - V(\vec{q}))^2 \qquad \text{(Eq. 3.21)}$$

The potential boost must fulfil three criteria. First, the potential must leave intact the relative disposition of the potentials in the original energy surface. That is, if $V_1(\vec{r}) < V_2(\vec{r})$ then ${V_1}^*(\vec{r}) < {V_2}^*(\vec{r})$. Second, the potential energy difference between two points in the biased surface must be smaller than the energy difference in the original one ${V_1}^*(\vec{r}) - {V_2}^*(\vec{r}) < V_1(\vec{r}) - V_2(\vec{r})$. The third criteria to fulfil is to ensure a narrow distribution ((Eq. 3.22)) to have a near-gaussian harmonic boost potential and accurate reweighting (particularly for cumulant expansion to the second order technique that will be commented further on).

$$\sigma_{\Delta V} = k_V(E - V_{av})\sigma_V \leq \sigma_0 \quad \text{(Eq. 3.22)}$$

Where $\sigma_{\Delta V}$ is the standard deviation of the boost potential, V_{av} is the average of the system potential, and "σ_0" is a user-defined upper limit (e.g. $10k_bT$).

To fulfil these criteria there is a stablished range for the threshold of the bias potential. (Eq. 3.23).

$$V_{max} \leq E \leq V_{min} + \frac{1}{k_V} \quad \text{(Eq. 3.23)}$$

Which to be fulfilled k_V itself must satisfy (Eq. *3.24)*.

$$k_V \leq \frac{1}{V_{max} - V_{min}} \quad \text{(Eq. 3.24)}$$

To complete the inequation, k_V is expressed as (Eq. *3.24)* times a constant k_0 which takes values of zero to one.

$$k_V \equiv k_0\left(\frac{1}{V_{max} - V_{min}}\right) \quad \text{(Eq. 3.25)}$$
$$\text{Where } 0 \leq k_0 \leq 1$$

The value of the force constant (proportionality related with the potential boost applied) is therefore related with the maximum and minimum energies of the system. The variable k_0 controls the magnitude of "k_V" and through it the value of the potential boost. The higher k_0, the higher value of the potential value added and thus the more accelerated the simulation is. Before running GaMD productions, preparation stages are used in order to estimate the maximum acceleration which can be achieved for the simulated system.

The threshold for adding the boost potential can be set to two limits, the "lower bound" ($E = V_{max}$) and the upper bound ($E = V_{min} + 1/k_V$). The first is a more conservative approach which results in filling the potential energy surface while leaving unaltered the high energy barriers (Fig. 3.7). The second approach is recommended if the sampling wants to be enhanced even more with highest acceleration although this was not tested in the original publication.[14]

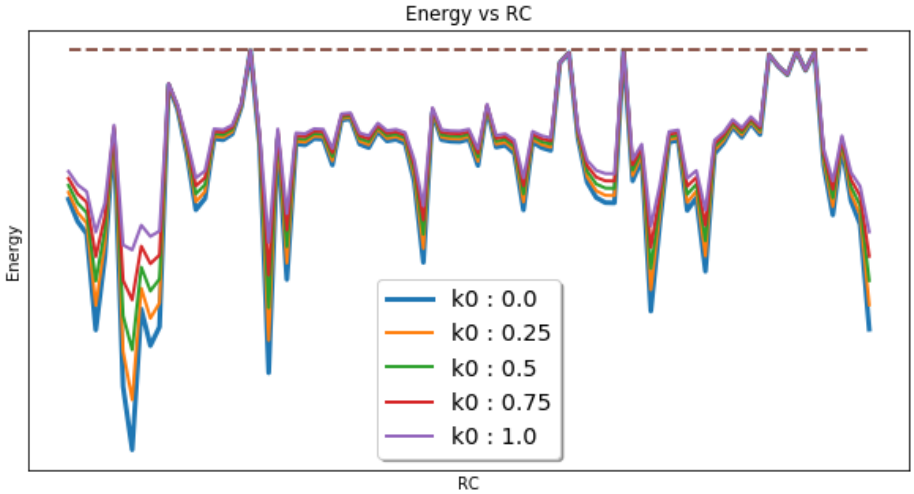

Fig. 3.7 GaMD Potential Energy Smoothing with Increasing k0 and Threshold Set to Lower Bound : For each value of k_0 the surface becomes smoother. It can be noted that while energy peaks remain mostly unaltered energy minima become filled thus allowing for enhanced exploration of states. The dashed line represents the energy threshold when set to its lower bound ($E = V_{max}$). GaMD equations implemented in python 3.6[17] and Ammonia potential energy surface taken from "Harvard Dataverse"[82,83].

Potential boosts can be applied to add only the total potential boost ΔV_P, dihedral potential boost ΔV_D, or a dual potential boost (both at the same time). Generally, the dual-boost potential provides higher acceleration power.

As stated before, reweighting is used to obtain the original free energy surface from a biased GaMD simulation. This is done by approximating Boltzmann's factor to its unbiased value using the weights of the bias potential applied. Three methods are used for this purpose. First, Exponential average, this method suffers from large statistical noise. Second, McLaurin Series provides less noise than the exponential average method. It is based on approximating Boltzmann's factor with a McLaurin series up to the 5th-10th order (Eq. 3.26). Third**,** Cumulant Expansion to the 2nd Order which according to authors gives the best results. However, it relies highly on the gaussianity of the harmonic boost potential (measured though its anharmonicity), moreover, it is only applicable to relative to relatively small systems (100 aminoacids) otherwise the noise becomes too high (stated in in the webpage where authors offer at public disposal the python scripts to carry out the reweighting).[16] Due to the size of the protein in this project (more than 500 aminoacids) cumulant expansion could not be applied due to statistical noise in reweighting.

$$\langle e^{\beta\,\Delta V} \rangle = \sum_{k=0}^{\infty} \frac{\beta^k}{k!} \langle \Delta V^k \rangle \qquad \text{(Eq. 3.26)}$$

3.3 Ligand Binding Affinity Estimation: MM/PBSA and MM/GBSA

To study structure-based drug design the interaction between Ligand (L) and Receptor (R) must be determined. The binding can be described by a simple chemical reaction (Eq. 3.27*)*:

$$L + R \leftrightarrow LR \quad \text{(Eq. 3.27)}$$

This process is governed by an equilibrium constant K_{LR} which can be related with the free energy of the process ΔG by (Eq. 3.28).

$$\Delta G = -RT \ln K_{LR} \quad \text{(Eq. 3.28)}$$

Where "R" is the ideal gas constant with units $\left[\frac{Kj}{mol}\right] or \left[\frac{Kcal}{mol}\right]$, and T is the temperature of the system.

The greater the affinity constant for the LR pair is, the greater the increment of free energy for the binding results to be. Computational methods aim to predict LR binding energies to save time and reduce costly and time-consuming experimental processes.

The most widely common drug design computational methods are docking and scoring which are efficient but not particularly accurate. On the other hand, Alchemical Perturbation (AP) methods make use of statistical mechanics calculations and are in principle highly accurate at the cost of performing Monte Carlo or Molecular Dynamics calculations which require of extensive phase space sampling. At a middle ground, end-point methods offer more accuracy whit reduced computational cost. They act by sampling the end states of the reaction, some examples of this approach are linear response approximation (LRA), linear integration energy (LIE) and Molecular Mechanics / Poisson-Boltzmann (MM/PBSA) methods. MM/PBSA was developed in the late 90's and has been used extensively for protein design, protein-protein design, conformer stability and other settings[15].

In MM/PBSA ΔG is estimated from free energies of reactants and products of the chemical reaction seen in (Eq. *3.27* by (Eq. *3.29*.

$$\Delta G_{Bind} = \langle G_{PL} \rangle - \langle G_P \rangle - \langle G_L \rangle \quad \text{(Eq. 3.29)}$$

The free energy of the Receptor (R) the Ligand (L) and the complex (LR) are calculated with (Eq. 3.30).

$$G = V_{bind} + V_{el} + V_{VdW} + G_{pol} + G_{np} - TS \quad \text{(Eq. 3.30)}$$

Where V_{bind} corresponds to the bonded terms of the force field(V_{Bond}, V_{Angle}, $V_{Dihedral}$), V_{VdW} is the van der Waals interaction and is the V_{el} is the electrostatic interaction from the force field as well (recall force field expression (Eq. 3.10)). G_{pol} is the polar contribution and G_{np} is the non-polar contribution to the solvation free energy, while S is the entropy of the system estimated with a normal-mode analysis of vibrational frequencies.

G_{pol} accounts for the electrostatic interaction between solute and solvent, it is estimated typically by solving Poisson-Boltzmann equation (MM/PBSA approach) or by using Generalized Born (GB) implicit solvation model (giving the MM/GBSA approach). Poisson-Boltzmann's equation results in a complicated expression which is derived from Laplace's equation to study the gradient of the electrostatic potential of a set of particles.[84] Generalized Born model treats the solvation environment as a continuum dielectric model which is easier to compute[85] and allows for per residue free energy decomposition.

G_{np} accounts for cavitation, dispersion and repulsion solvation energies as well as the attractive and repulsive van der Waals interactions between solute and solvent. Although these terms should be included explicitly, in MM/PBSA and GB/PBSA it is given by the linear relation from the solvent accessible surface area (SASA)[15]. Since hydrophobic aminoacids will tend to bury themselves into proteins staying away from the solvent the degree of exposure of the protein is calculated by SASA is used to estimate environment free energies.[86]

The average free energy values from (Eq. 3.29) should be strictly estimated from three different simulations, however, it is common to simulate the complex and create ensemble-averages of the receptor and ligand by removing the atoms from the simulation (Eq. 3.31). This has some advantages like cancelling the force field bonded terms from (Eq. 3.30).

$$\Delta G_{Bind} = \langle G_{PL} - G_P - G_L \rangle_{PL} \quad \text{(Eq. 3.31)}$$

Binding free energies in the presence of solvent are evaluated with the thermodinamical cycle seen in Fig. 3.8.

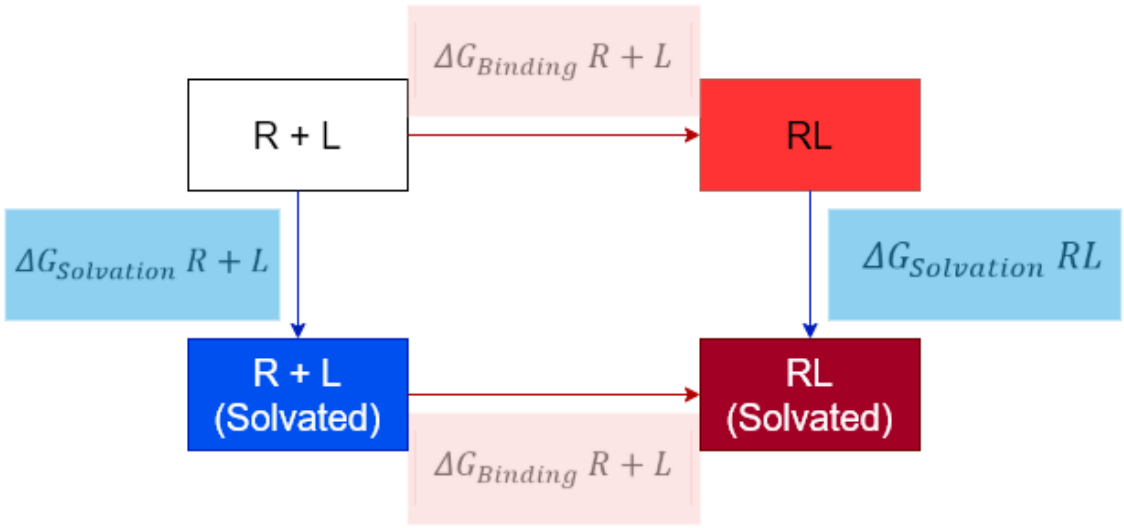

Fig. 3.8 Binding Thermodinamical Cycle: By taking into account free energy increments between solvation and binding the calculation of RL(solvated) can be done from its components.

It should be commented that MM/GBSA allows for energy decomposition per residue. Being able to estimate the interaction energies of the ligand with each aminoacid of the receptor. Regarding the quality of calculations there is not agreement if MM/PBSA is better or not compared with MM/GBSA since the quality of the results depends on the system applied. These methods may be useful to improve results of docking or virtual screening and to grasp hidden tendencies in sets of ligand affinity. Nevertheless, the approximations employed make them highly reliant on the setup chosen for the calculation[15]. The accuracy does not seem sufficient for later states of predictive drug design. Due to their modest computational cost and relative accuracy their use is popular.

4 Results and Discussion

4.1 Computational Details and Work Scheme

Although the work done in this project is going to be mentioned throughout the computational details, in this section a quick summary and a flow chart are given in order to understand better the procedure of this project.

First, a model of the human Na_v 1.4 with Azobenzene inside was constructed. This included a bilayer membrane, water, and solvated ions. Equilibration of this system gave an equilibrated geometry which was used for 4 GaMD simulations and 4 cMD simulations of each 1000ns each (these simulations are also referred as replicas along the results discussion). GaMD was used to enhance sampling of the conformational space and cMD was used to compare the acceleration with unbiased trajectories. From GaMD trajectories, 4 Azobenzene geometries were obtained for each binding pocket. Following this, another 100ns of cMD was used to explore the binding pocket to be used by MM/GBSA to make free energy estimation of binding energies and per residue decomposition. All of this process is schematically shown in Fig. 4.1.

As a result of this process, an idea of the sampling achieved by GaMD identification of binding pockets and free energy calculations were obtained.

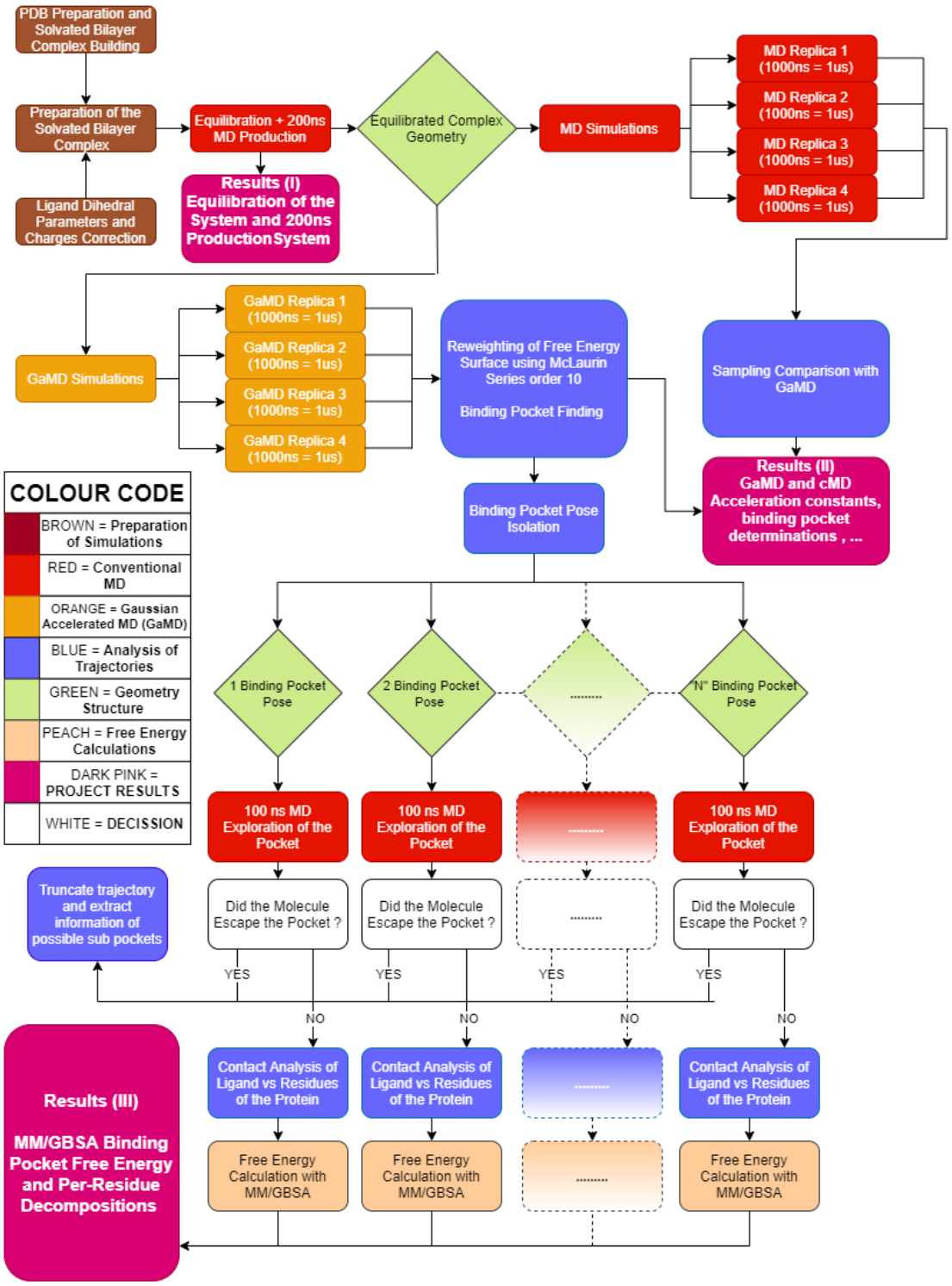

Fig. 4.1 Workflow Carried Out in this Project

4.1.1 Na_v 1.4 and Azobenzene Model Construction.

A truncated model of the α structure of the human Na_v 1.4 voltage gated ion channel[12] using VMD[30] with PDB ID : 6AGF[29]. The VSD and the β_1 were removed, remaining the TM 5 and 6 of the four protein domains of the α structe, selected residues for the model can be seen in Table 4.1.

Table 4.1 Selected residues for Na_v 1.4

Protein Domain	*Chain ID*	*Residues*
I	A ; B	234 – 286 ; 336 – 451
II	C	683 – 805
III	D	1143 – 805
IV	E	1464 - 1601

Azobenzene was constructed using "CHARMM-GUI Ligand Reader and Modeler"[87–89] and parametrized using CHARMM General Force Field (CGenFF) 2.2.0[90]" . Initial penalty scores were too high for nitrogen-containing dihedral angles as well as for atomic charges. Ligand Force constants for the dihedral angles -C-N=N-C- and C-C-N=N- inside azobenzene were taken from McCullagh, Franco, Ratner and Schatz [91] (Appendix A Table 7.1) . A geometry optimization of azobenzene using MP2, atomic charges were calculated using HF with the basis 6-31G(d) implemented in Gaussian 16[92] and fitted using the Restrained Electrostatic Potential (RESP)[93] method using the ANTECHAMBER[94] software package as done by Kingsland , Samai, Yan, Ginger and Maibaum.[95] (Appendix A Table 7.2) . Azobenzene was inserted 10 Å below (z axis direction) from the centre of mass of the channel inside its main vestibule.

Both the truncated model of the Human Na_v 1.4 ion channel and the modified Azobenzene's parameter and charge files, were uploaded to CHARMM-GUI Membrane Builder.[89,96–100] Here, the system was aligned along the z-axis, terminal N- and C- were amidated and acetylated and inserted into a POPC bilayer of 100 by 100 lipids. Water Molecules were added to solvate the system and a concentration of 0.15 M was added. Description of the lipids and protein was done using CHARMM36m[101,102], water was described using the TIP3P[103,104] model. Atom numbers can be seen in Table 4.2. Inputs were generated for AMBER 18[105] which was employed for all MD simulations. Lennard-Jones Parameters for Ions were taken from Joung and Cheatham[106] and modified with the parameter editor ParmED.[107]

Table 4.2 Atom numbers in the simulation system

Protein	Membrane	Water	Ligand	Na^+	Cl^-	Total
9328	26800	62910	24	69	56	99187

4.1.2 Equilibration of the system

An energy minimization of the system was done employing 5000 maximum number of steps minimizations using 100 steepest-descent algorithm steps for using restraints in the system. A Non-bonded cutoff of 12.0 Å was used and a function switch was settled to 10.0 Å. Positional restraints were used for aminoacid residues (10.0 kcal mol^{-1} $Å^{-2}$) and for membrane lipids (2.5 kcal mol^{-1} $Å^{-2}$). 400 distance and angle restraints were used for different atoms inside the lipid membrane.

The system was equilibrated in 6 MD simulations of 125 ps (0.001 ps time-step) where restraints successively decreased Table 4.3. First two simulations were done under the NVT ensemble with Langevin Dynamics (friction coefficient 1.0 ps^{-1}) and a target temperature of 303.15 K. The same cutoff and function switch as the minimization case was employed. The four remaining simulations were done under the NPT ensemble with same parameters for Langevin Dynamics, non-bonded cut-offs, and function switches. A semi-isotropic Monte Carlo (MC) Barostat was used for pressure regulation targeted at 1.0 bar and constant surface tension with 2 interfaces and γ = 0.0 dyne /cm. SHAKE algorithm was present in all simulations to avoid calculating forces of bonds containing Hydrogen. Ewald sum was used for calculation of electrostatic interactions

Table 4.3 Positional restraints for equilibration simulations. In addition to this, all simulations contained 400 distance and angle constraints for lipid atoms.

Equilibration Simulation	Force constant / kcal mol^{-1} $Å^{-2}$	
	Protein	Membrane
1	10	2.5
2	5	2.5
3	2.5	1.0
4	1.0	0.5
5	0.5	0.1
6	0.1	----

4.1.3 cMD Productions

All cMD (conventional, not accelerated) productions in this project were done using the same parameters. NPT ensemble with 0.002 ps time-step with Langevin Dynamics (friction coefficient 1.0 ps^{-1}) with target temperature 303.15 K, non-bonded cutoff of 12.0 Å and function switch of 10.0 Å with SHAKE algorithm constraining bonds with hydrogen. Semi-isotropic MC Barostat targeting to 1.0 bar with constant surface tension γ = 0.0 dyne /cm and two interfaces. Ewald sum was used for calculation of electrostatic interactions.

Four cMD of 1000ns (500000000 steps) were done with these parameters as well as the 100ns (50000000 steps) cMD explorations of the binding pockets after determination with GaMD.

4.1.4 GaMD Productions and Reweighting

The core parameters for GaMD productions were the same as for cMD (previous section, 4.1.3). Both the potential energy and the dihedrals were boosted. Maximum standard deviations for both boosts were kept by default as a value of 6.0. GaMD needs of a preparation phase before running productions where variables from the potential energy of the system are collected. This process lasted for 50ns using a 0.002 ps time-step where time steps were assigned in this way : 1000000 steps for conventional equilibration of the system without collecting variables (ntcmdprep) , 5000000 steps for collecting variables in the conventional simulations (ntcmd), 2500000 steps for equilibrating the system after adding the boost potential (ntebprep), and 20000000 steps for adding boost potential and collecting statistical variables. In addition to this, calculation of potential averages and deviations was done every 5000 steps (ntave). All this step must be included into nstlim (number of total time steps for MD).

For a 250ns GaMD simulation (50ns preparation + 200ns production) 125000000 steps are be used (25000000 for preparation and 100000000 for production) with a 0.002 ps time-step.

Reweighting of GaMD trajectories was done using a self-made modification of the Pyreweighting[16] scripts. Order 10 McLaurin series was employed to recover the unbiased free energy surface.

4.1.5 RC Calculation

To reweight the GaMD free energy surface two RCs must be chosen. In the original GaMD publication $N_{Contacts}$ vs Root Mean Square (RMSD) of the Ligand were chosen to study the binding of benzene to T4-Lysozime.[14] However, these RC do not seem to seem appropriate for describing different binding pockets and they do not seem spatially intuitive. These RC belong to a triangle obtained by three points, the first one is the centre of mass of Azobenzene (CoM Azobenzene), the second and the third points are the centre of mass of the alpha carbons of the DEKA filter (CoM α_c DEKA) and the centre of mass of the alpha carbons of four residues selected near the lower gate of the protein (CoM α_c Gate)(residues seen in Table 4.4).By determining these three points in each snapshot of the simulation and calculating the distances between them, the angles of the triangle formed can be known by the cosine rule. One distance and angle are chosen out of the three in order to serve as RC as seen in Fig. 4.2.

Table 4.4 Residues selected to define RC

Residue	DEKA Filter	Lower Gate
1	ASP 124	VAL 164
2	GLU 248	LEU 287
3	LYS 394	VAL 443
4	ALA 521	ILE 581

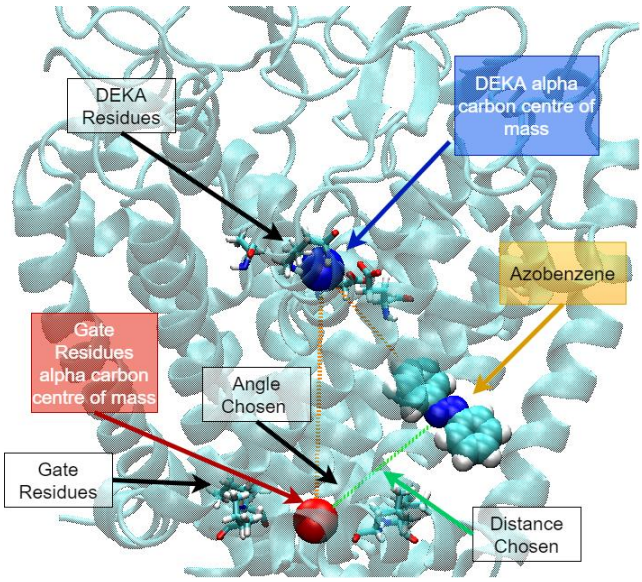

Fig. 4.2 Points Used for RC Definition: Three points are chosen forming a triangle, CoM α_c DEKA, CoM α_c Gate, and CoM Azobenzene. One RC is the distance between CoM α_c DEKA and CoM Azobenzene (RC1) and the other RC is the angle CoM α_c DEKA - CoM α_c Gate - CoM Azobenzene (RC2).

This pseudo-polar coordinate (1 angle and 1 distance, while polar coordinates is 2 angles and 1 distance) allow to follow the trajectory of Azobenzene along the simulations. They will be used

as RC when performing the reweighting of GaMD trajectories and when studying the probability distribution of Azobenzene inside the protein.

For reweighting and probability distribution comparison, 2D histograms will be used with a discretization of RC1 equal to 0.08 Å and RC2 equal to 1º.

4.1.6 MM/GBSA Free Energy Estimation and Per Residue Decomposition

MM/GBSA free energy calculations and per residue decompositions were done using the program MMBSA.py[108] using 500 frames equidistantly taken from pocket exploration cMD. The same salt concentration was used as the used for building the bilayer complex 0.15M of NaCl. Per residue energy decomposition was done for protein residues closer than 5 Å of Azobenzene for more than 2% of times during the pocket exploration cMD. Contact analysis of cMD simulations was done CPPTRAJ[109].

4.2 Results (I). Equilibration of the System and 200ns MD Production

During the protocol described in 4.1.2 the system was minimized performing short equilibrations with successively reduced constrains. Following this, a cMD production of 200ns was done in order to fully equilibrate the system. To check whether the protein is equilibrated the (RMSD) of the alpha carbons of its backbone is inspected to ensure that there is no significant variation at the end of the equilibration (Fig. 4.3).

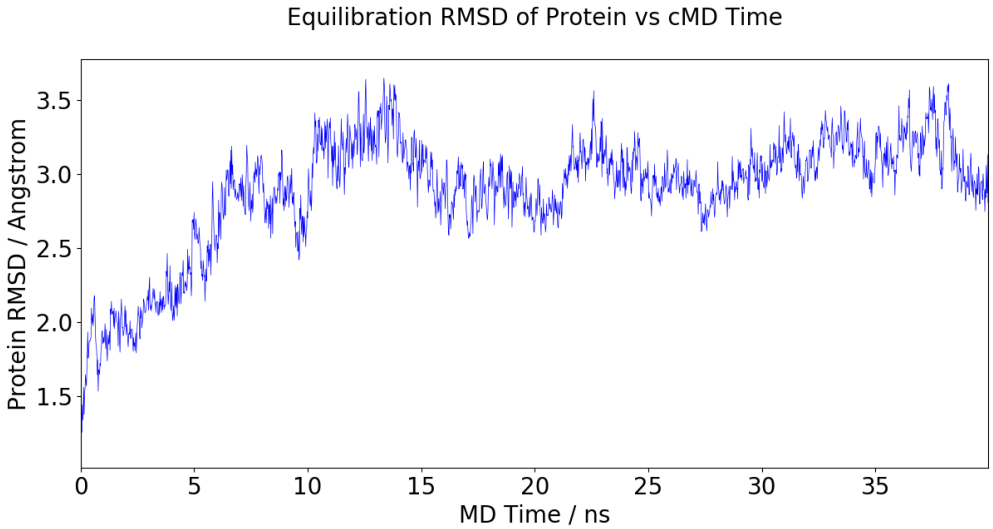

Fig. 4.3 RMSD for 200ns cMD Production After Initial Equilibration: RMSD of α carbons of protein backbone (average = 2.88 Å, standard deviation = 0.39 Å)

Indeed, the RMSD of the protein shows a stabilization around 2.8-3.0 Å with fluctuations around this point. Visualization of the simulation box of the system is shown Fig. 4.4.

The last frame of the 200ns of the cMD simulation after equilibration was used for initial geometries for four GaMD and another four cMD simulations.

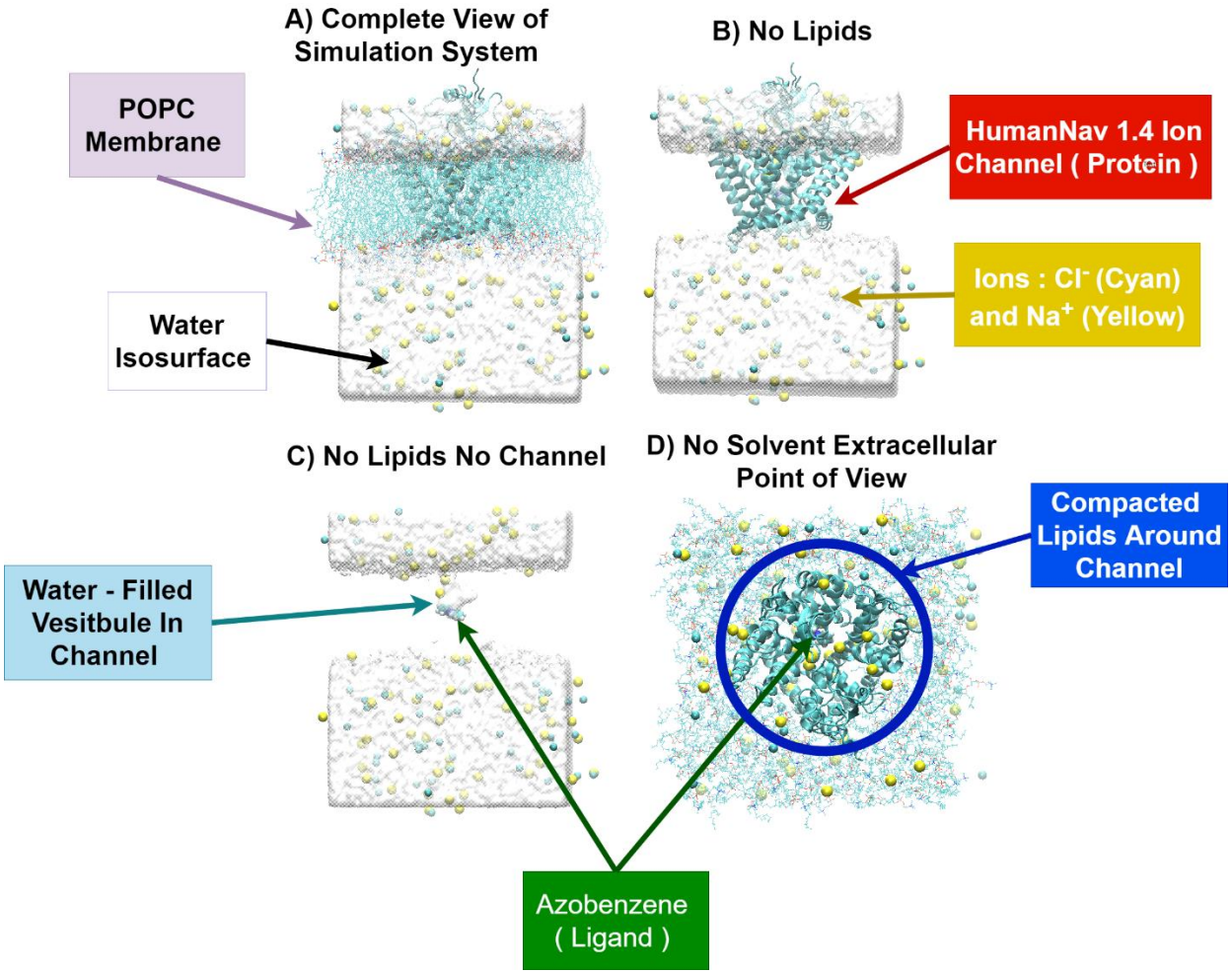

Fig. 4.4 Different Views of the System After Equilibration : A) Complete View of Simulation System, with water represented as an isosurface (using VMD's VolMap tool) and the POPC membrane represented as transparent lines. B) No Lipids. View of the channel without lipids with the channel and the solvated ions highlighted. C) No Lipids and no Channel. Here the Azobenzene inserted inside the Channel can be appreciated as well as existence of water density inside a vestibule, characteristic of ion channels. C) No Solvent Extracellular Point of View. Packing of lipids around the protein is visually inspected to check equilibration after membrane building around the protein.

4.3 Results (II). GaMD and cMD

4.3.1 GaMD Acceleration Constant

As seen in section 3.2 GaMD needs some equilibration steps to collect statistical information (potential maximum, minimum, average and deviation) about the potential energy of the system in order to estimate the potential boost applied. As recalled there, our GaMD simulations were assigned a total of 50ns (25000000 iteration steps) to carry out the preparation phase. Recalling section 3.2 the magnitude of the boost potential is controlled by k_0 which takes values between 0.0 and 1.0 where this second value indicates that the highest acceleration of the simulations is achieved while at the same time maintaining near-gaussianity distribution of the harmonic potential boost applied. The evolution of k_0 for each one of the four independent GaMD

simulations (the term "replica" is used as well as simulation in this project for referring to each one of the 1000ns GaMD or cMD simulations) can be seen in Fig. 4.5.

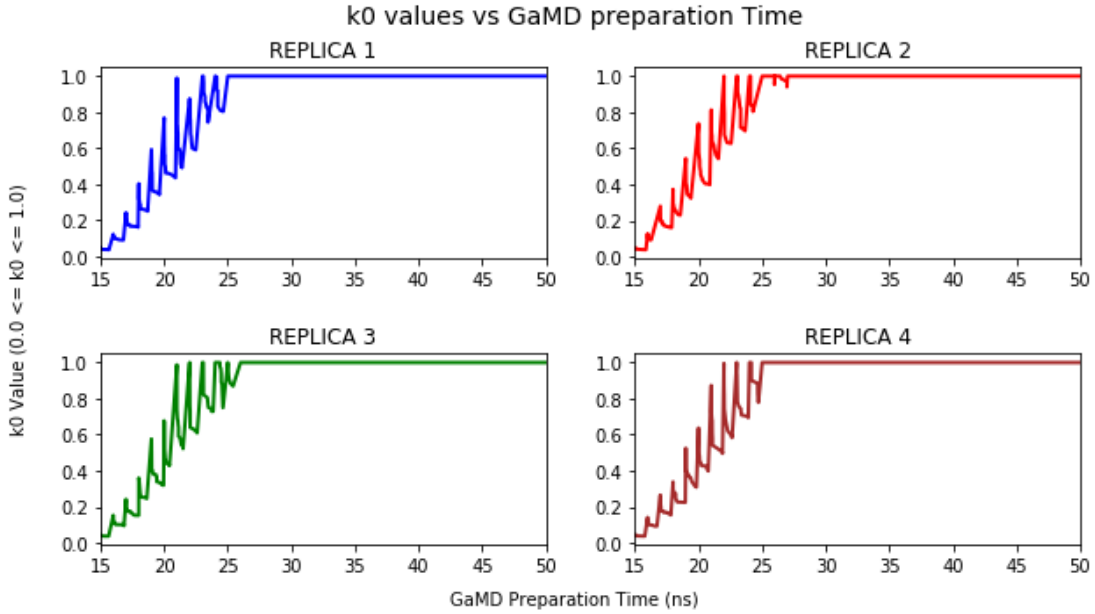

Fig. 4.5 k0 vs GaMD Replicas Preparation Time : As it can be seen, all 4 Simulations reach the value of 1.0 for k_0 before 30ns. This indicates that the boost applied will be maximum conserving a near-gaussian-like shape.

These results suggest that highest acceleration possible using the lower bound for the energy threshold is achieved in all four GaMD replicas and therefore the sampling will be enhanced.

4.3.2 GaMD Sampling Compared With cMD

In previous section 4.3.1 it was seen that under GaMD's own formulation, the acceleration for the four simulations was maximum due to a value of k_0 of 1.0. However, to have a qualitative idea of this acceleration, and of the sampling improvement given by GaMD, four cMD simulations starting from the same geometry and with the same time extension (1000ns) were done. Although starting from the same geometry, all GaMD and cMD simulations are independent since initial velocities for all of them were generated randomly.

To check the phase space sampling, the distribution of the states visited by Azobenzene following the RCs defined in section 4.1.5 is calculated. A probability distribution is calculated by doing a 2D histogram of the RC and normalizing the values. However, the visualization of probability distributions seems to underrepresent less visited conformations. It is better to compare the sampling in terms of reverse probability distributions. This is obtained by taking a probability distribution and setting the state with highest probability as zero, subtracting each state value to this one. Instead of having peaks of probability, valleys of probability will be obtained being the deepest the most visited one. To motivate the use of reverse histograms a comparison is shown in Fig. 4.6.

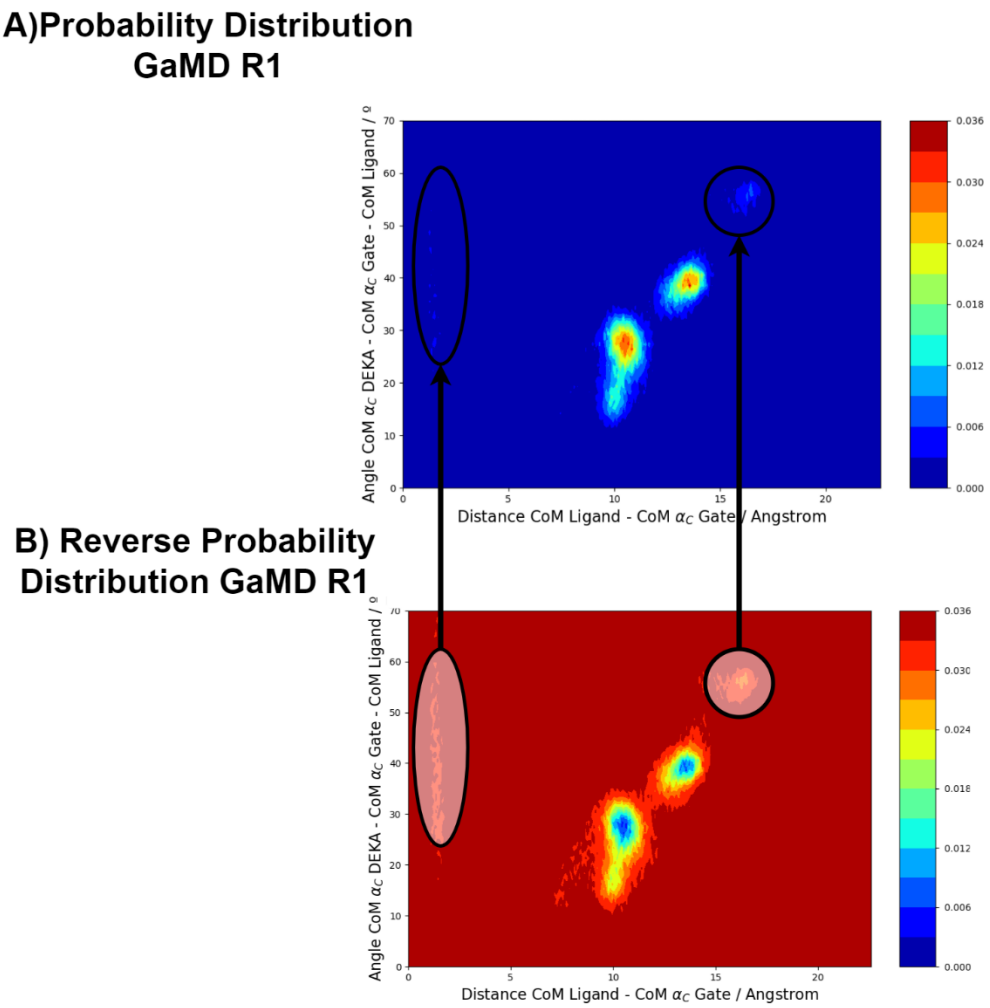

Fig. 4.6 Comparison Between Probability Distribution and Reverse Probability Distribution for GaMD Replica 1 (R1): A) Reverse Probability Distribution GaMD R1 B) Probability Distribution for GaMD R1. While in probability distributions the maximum corresponds to most visited conformations, in the reverse probability distribution is on the contrary, the lower the value the more visits the system does to that point. Two highlight areas show that low probability distributions when visualized in the reverse way they seem stand out more. The contours of visited conformations are "richer" and thus allow for better interpretation of the trajectory results.

Comparison between GaMD and cMD is done using reverse histograms for better visualization of results. All four GaMD and cMD reverse probability distributions can be seen in Fig. 8.1 of Appendix B. For better visualization of results GaMD and cMD results are combined into one plot and compared in Fig. 4.7. In this figure, highly visited are already identified with each binding pocket inside the simulation.

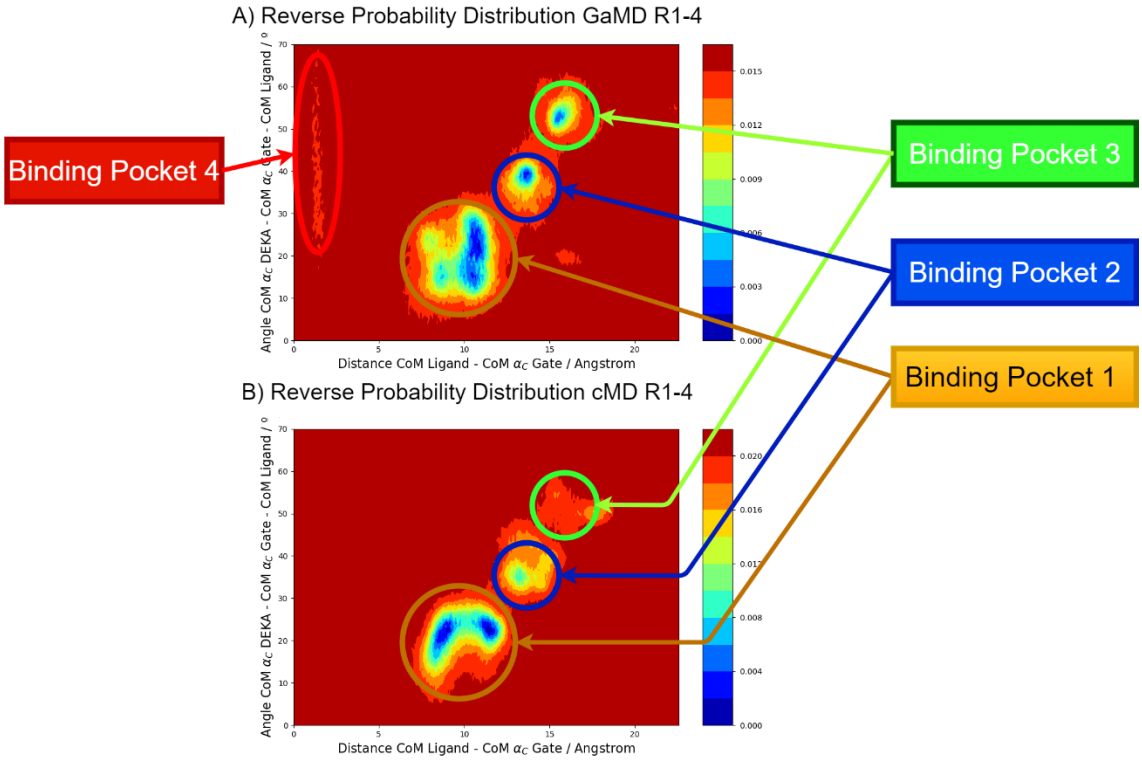

Fig. 4.7 Reverse Probability Distributions Combined for All Four GaMD and All Four cMD Trajectories: A) GaMD Replicas 1 to 4 (4000 ns). B) cMD Replicas 1 to 4 (4000 ns). Both GaMD and cMD seem to predict three highly visited conformations (binding pockets 1, 2 and 3). Nevertheless, GaMD predicts an additional binding pocket which is quite concentrated in terms of x units but spans a long way across y units (binding pocket 4). Binding pockets are clusters of poses that the ligand takes when binding with to the protein. Because of this, the ligand takes several poses and the apparition of sub pockets (different dispositions inside one binding pocket) can happen. This seems to be the prediction for cMD binding pocket 1 where two clear energy minima appear and that will be treated as sub pockets.

At first sight it can be appreciated that GaMD results (Fig. 4.7 A) seems to sample the conformational space better than cMD. There are some observations that confirm this: GaMD contours seem more "diffuse" (more exploration of energy states around minimum values), the low energy conformations seem better defined (GaMD gives more "concentrated" high probability points, see Binding pockets 2 and 3.), and GaMD predicts the apparition of another binding pocket which cMD does sample at all (binding pocket 4). Therefore, it can be said GaMD is indeed enhancing the sampling of the system.

GaMD seems to enhance the sampling of the system. Recalling from section 3.2, the lower bound to the threshold energy (which is the option chosen when performing simulations in this project), enhances the sampling but even more acceleration can be managed by setting the energy threshold to the higher bound, at least this was claimed but not tested in the original publication. The motivation for choosing an approach which gives smaller GaMD boosts in this project has two main reasons. First, a dual potential boost was performed, resulting in both the total potential energy and the dihedral angles of the system being biased, this means, that the structure of the ion channel could be altered if the energy added is too high. Second, GaMD is quite novel (2015) and the size of the simulations carried in the method publications[14,110] are not as large as the one attempted in this project justifying a more conservative approach as a first contact with this technique.

In a recent publication draft (April 2020) Yinglong Miao one of the contributors of GaMD reported the development of a new enhanced sampling technique known as "Ligand GaMD" (LiGaMD). This new methodology promised only to bias non-bonded interactions in order to explore better the possible binding phase space. Furthermore, the authors used the upper bound approach to the threshold energy and indeed reported an increment in sampling activity compared to the lower bound approach[111].

4.3.3 GaMD Reweighting of Free Energy Surface

In previous section the sampling of phase space was compared between GaMD and cMD. It was shown that GaMD offered a better sampling, fulfilling the first reason why GaMD was chosen. The second reason is to offer a better reweighting of the free energy surface due to the near-gaussian shape distribution of the harmonic boost potential applied.

GaMD simulations are biased, that is, that the system will visit low frequency states much more than it should by cMD. The estimation of the unbiased populations from biased populations can be done by 3 methods as recalled in section 3.2. Cumulant expansion should give the best results but it gives too much noise for proteins larger than 100 residues (here the system is about 500 residues). The second method, McLaurin series gives better results[81] than the third method (exponential average) and this is why it is employed in this project. Conceptually, reweighting is a fairly simple concept, calculate a reverse probability distribution for the trajectories as done before (section 4.3.2) but instead of normalizing the histogram using the total number of counts, approximate Boltzmann's factor by knowing the weights of the boost potential and then use this value to generate a weighted histogram of the RC distribution. After, take this weighted reverse probability distribution and calculate the free energy of each point using (Eq. 3.2). This generates a free energy surface which can be used for binding pocket identification and even drug pathway elucidation[110].

The four free energy surfaces for the GaMD replicas can be seen in Appendix B Fig. 8.2. For better interpretation of results, the data of all of them was combined and used to reweight a more complete free energy surface (Fig. 4.8).

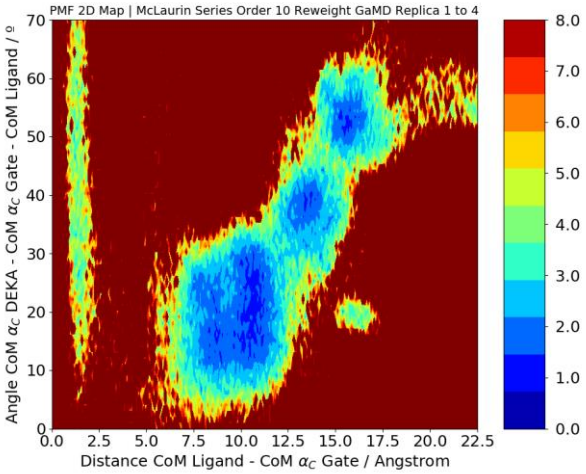

Fig. 4.8 Free Energy Surface for 4 GaMD Simulations Combined (Colour Bar in kcal/mol)

This free energy surface gives a good foundation for binding pocket identification and an idea of the relative stability between them. Although possible binding pockets were identified in Fig.

4.7 the advantage of the figure seen above is that the colour scale is more intuitive since energy is a physical quantity which gives a clearer view of the stability of the pockets.

4.3.4 Determination of Binding Pockets inside the channel

In previous section the free energy surface for the combination of 4 GaMD replicas was shown and its value for binding pocket identification was hinted at. Using Fig. 4.8, visual inspection, and the evolution of the RC along the simulations (Appendix B Fig. 8.3) the identification of a ligand pose belonging to each binding pocket was extracted. This can be seen Fig. 4.9.

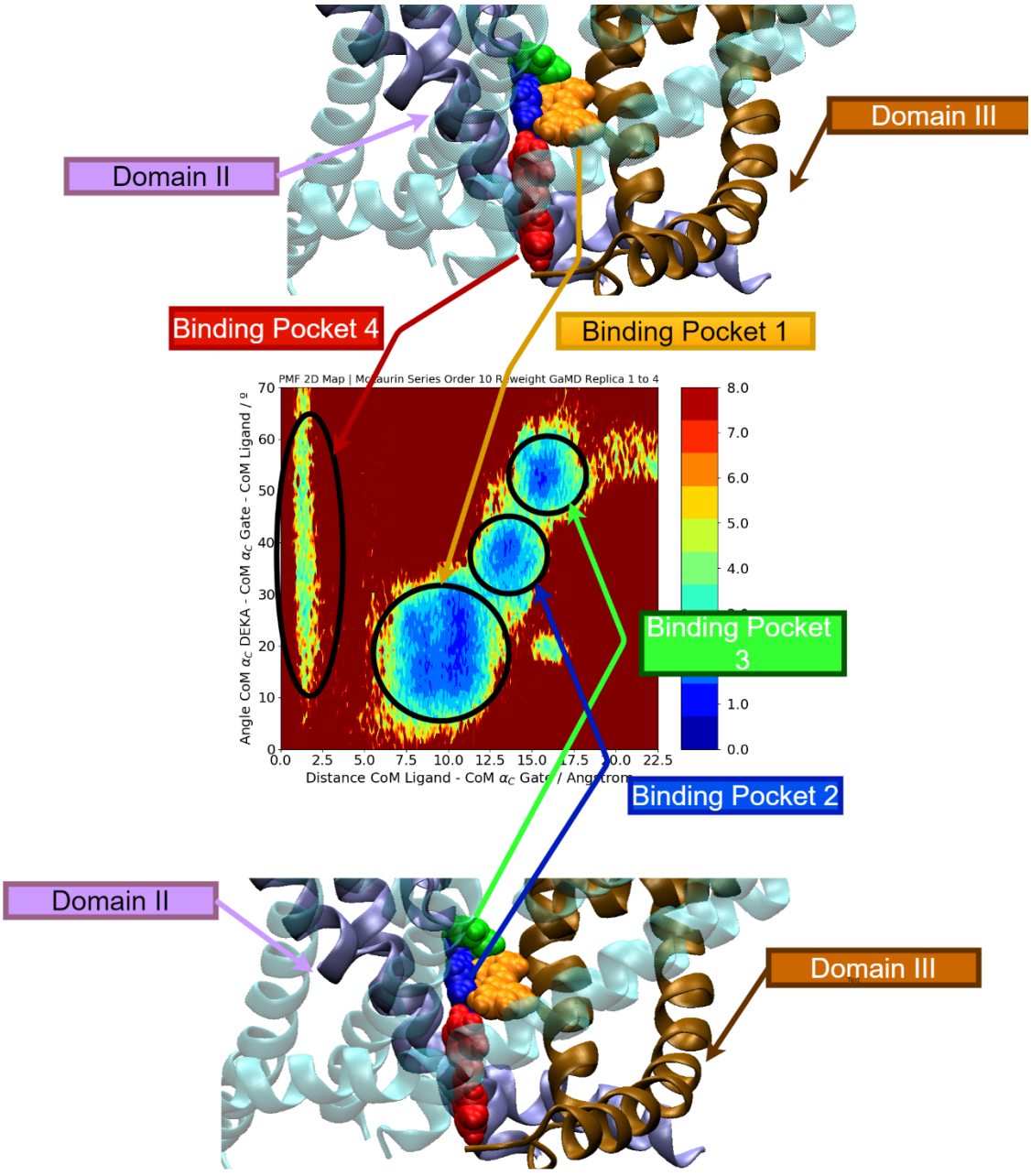

Fig. 4.9 Azobenzene Poses Selected From Each Binding Pocket Present in the Free Energy Surface (Fig. 4.8): As observed, binding pockets seems to be somewhere between DII and DIII inside.

Binging pockets seem to be forming an arch between the lower gate of the channel (binding pocket 4) going through DII and DIII (binding pockets 1,2) and reaching a cavity close to the lipidic

membrane (binding pocket 3). Binding pocket 4 presents a very distinct shape in the free energy surface due to its proximity to the gate of the protein. Since binding pocket 4 is close to the gate point chosen for RC definition (CoM α_c Gate), small variations in the CoM of Azobenzene will cause a great fluctuation in the angle RC2.

Binding pockets seem to be clustered around the same area inside the protein, close by or inserted between DII and DIII. Since azobenzene is a non-polar molecule the expected interactions will probably be with hydrophobic residues inside the protein (PHE, LEU, ILE, …). Indeed, as seen in Fig. 4..10 binding pockets are in close contact with DII. The cavity between D II and DIII has a high amount of non-polar residues which are buried away from the water filled internal vestibule of the ion channel.

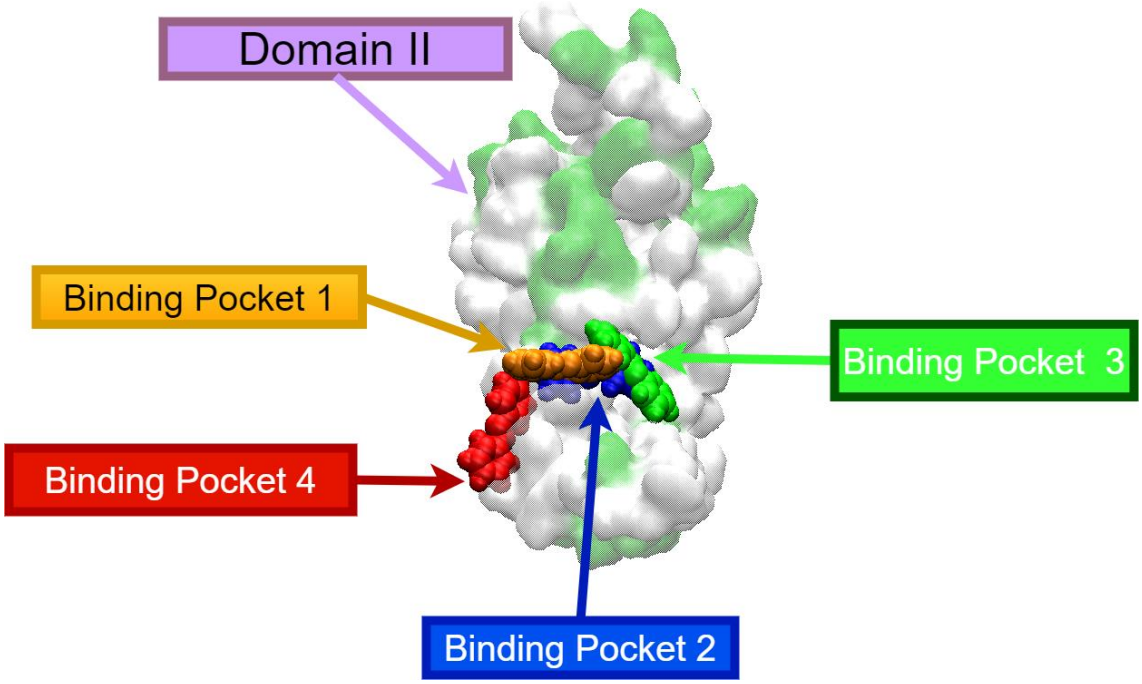

Fig. 4.10 Surface Representation of DII of the Ion Channel with Azobenzene Geometries Belonging to Each Binding Pocket, View from DIII Side: Ion channel colour code by residue type (White = Hydrophobic , Green = Polar). DI-DIII-DIV are not represented.

Therefore, the position of the pockets seems to give information of the kind of interactions that will be present.

4.3.5 100ns cMD on Binding Pockets

In previous section 4 geometries belonging to each binding pocket were extracted from GaMD trajectories. MM/GBSA requires of several simulation snapshots to perform an average estimation of the free energy for each binding pocket, nevertheless, using GaMD trajectories for this purpose could generate artifacts since its trajectories are biased. Because of this, 100ns of cMD were done to have trajectory snapshots which would behave according to the unbiased (original) free energy surface. The Azobenzene geometries seen in Fig. 4.9 were used as starting geometries of the cMD simulations. Another reason to do this, is to check the stability of the pockets, in other words, if the ligand escapes before 100ns from the pocket it suggests that it is not a very robust binding site. The RC for the trajectories of each pocket plotted on top of the GaMD free energy surface are seen in Fig 4.11.

.

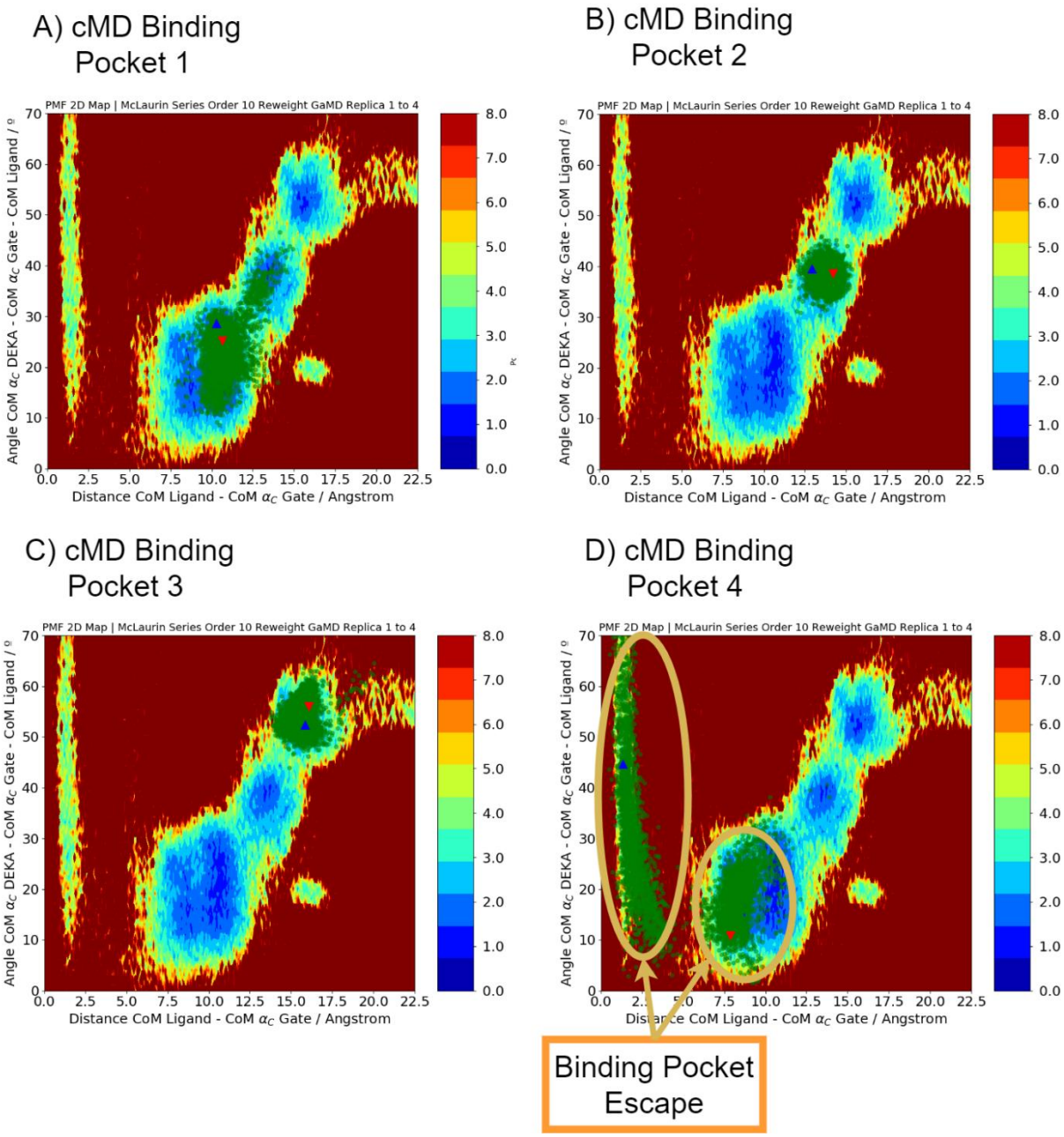

Fig. 4.11 100ns cMD Productions Plotted on Top of the GaMD free Energy Surface (Fig. 4.8) (colour bar in kcal/mol): A) cMD Binding Pocket 1 B) cMD Binding Pocket 2 C) cMD Binding Pocket 3 D) cMD Binding Pocket 4**.** Green dots plotted over free energy surface indicate the trajectory of Azobenzene. Blue triangle pointing up Indicates the starting position and red triangle down indicates the final position of the trajectory. For D) the dynamic which starts in binding pocket 4 escapes and ends up in part of binding pocket 1.

Dynamics seem compliant with GaMD energy surface. In the case of binding pockets 2 and 3 (Fig. 4.11 B) and C)) Azobenzene is trapped for the whole simulation. For binding pocket 1 (Fig. 4.11 A)), Azobenzene seems to sample minimally part of binding pocket 2. One should keep in mind that the free energy surface is just a bidimensional representation of an energy landscape which depends on many degrees of freedom. This is the origin possible mismatches between

the real dynamics of the system and the free energy surface shown. For binding pocket 4 Azobenzene escaped the pocket after 50ns of simulation. To ensure that this was not caused by mere chance, a second cMD simulation was done, again, Azobenzene escaped after 50ns of simulation, confirming that the relative stability of pocket 4 is presumably smaller compared with the rest of binding pockets.

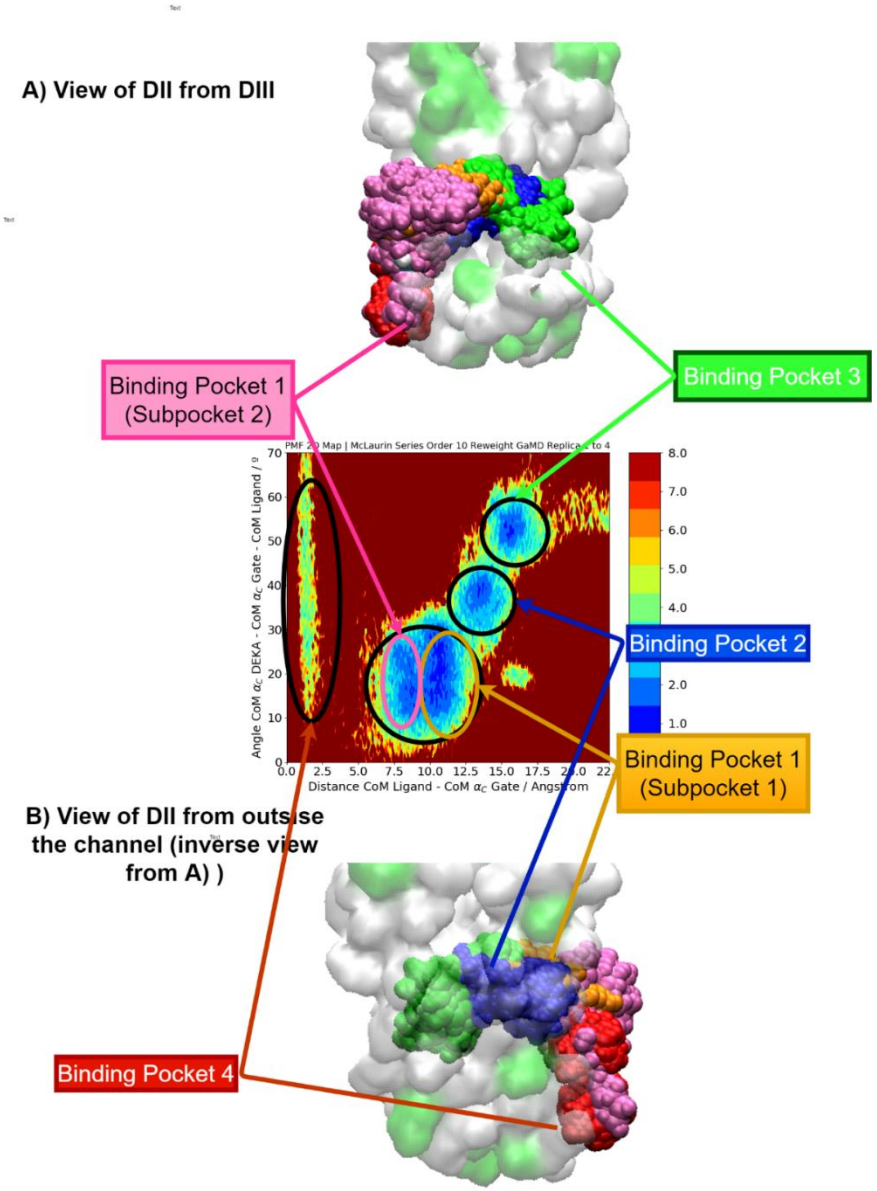

Fig. 4.12 Visualization of 100 Snapshots Belonging to the 100ns cMD trajectories Seen in Fig. 4.11: A) View of DII from DIII and B) View of DII from outside the channel (inverse view from A)). DI to DIII removed, leaving DII remaining. Colour code for ligands is the same as previous figures. Colour code for protein depends on residue type (White = Hydrophobic, Green = Polar).

When Azobenzene escapes from binding pocket 4 to binding pocket 1 (Fig. 4.11 D)) it can be appreciated that it stays in a region of binding pocket 1 which is not explored in Fig. 4.11 A). This suggests that there is the existence of two sub-binding pockets inside binding pocket 1 as it was pointed out by the 4 cMD simulations of 1000ns in Fig. 4.7 B). Therefore, binding pocket 1 is formally divided into 2 sub pockets, one sub pocket represented by the trajectory seen in Fig. 4.11 A) (binding pocket 1 – sub pocket 1) and a second sub pocket represented by the trajectory after Azobenzene escaped from binding pocket for in Fig. 4.11 D) (binding pocket 1 -sub pocket 2).

Binging pockets are not represented by a single ligand pose. Instead, a more accurate sense of the binding pocket dimensions is obtained by plotting several trajectory snapshots along the 100ns cMD simulations. A simultaneous plot of 100 points of each trajectory from Fig. 4.11 A), B), C) and D) gives a sense of the space where the molecule is binded into the protein (Fig. 4.12) by representation of density of Azobenzene poses.

4.4 Results (III). MM/GBSA Binding Pocket Free Energy and Per-Residue Decomposition

500 equidistant frames from the 100 ns cMD trajectories seen in Fig. 4.11 were taken for MM/GBSA to calculate binding free energies. In Table 4.5 free binding energies for each pocket can be seen.

Table 4.5 Free energy (kcal/mol) calculations by MM/GBSA

Energy Contribution	Binding Pocket 1		Binding Pocket 2	Binding Pocket 3	Binding Pocket 4
	Sub pocket 1	Sub pocket 2			
V_{VdW}	-24.4409	-24.7383	-29.7468	-31.1607	-29.8775
V_{EL}	0.0962	-0.8559	-0.7558	-2.2490	-2.7173
G_{GB}	10.5925	11.0854	14.9284	22.4755	12.6864
G_{SURF}	-3.4752	-3.4739	-4.0953	-4.3756	-4.1717
ΔG Gas	-24.3447	-27.5942	-30.5027	-33.4097	-32.5948
ΔG Solv	7.1173	7.6061	10.8331	18.0999	8.5148
ΔG Total	-17.2274	-17.9881	-19.6696	-15.3098	-24.0800

Recall from section 3.3 that the bonded terms MM/GBSA were cancelled due to the extraction of trajectories for complex receptor and ligand. VdW stands for the van der Waals interaction and V_{EL} the electrostatic interaction between complex and ligand given by the force field. G_{GB} is the polar solvation to the free energy calculated by the GB model. G_{SURF} is the free energy contribution due to non-polar solvation calculated by the SASA method. The ΔG Gas and ΔG Solv account for the different free energies depending on the state and are used to determine the total binding free energy as recalled from the thermodinamical cycle in Fig. 3.8.

The first thing that can be observed from Table 4.5 is that binding for all the pockets seems mainly driven by the term VdW. This term accounts both for repulsion and attraction due to van der Waals interactions between ion channel and ligand, since Azobenzene is a non-polar molecule, this is an expected result.

Another result is the abnormal favourable total interaction that binding pocket 4 presents. Although in cMD simulations Azobenzene escaped from pocket 4 before 50ns MM/GBSA predicts it to be the most stable pocket. This is clearly in contradiction with the GaMD free energy surface obtained and the 100ns cMD trajectories starting from each pocket. A possible

explanation is that due to the truncation of the trajectory 100ns cMD, the free energy calculation is done over snapshots that are particularly stable, distorting the real behaviour of the system.

As commented in the computational details for this section, residues which kept a closer contact than 5 Å in the 100ns cMD exploration for at least 2% of the simulation were used to perform a per residue decomposition of the free energies. Of course, these aminoacids will be different for each binding pocket. For binding pocket 1 sub pocket 1 the per residue decomposition can be seen in Table 4.6.

Table 4.6 Per Residue Energy (kcal/mol) Decomposition for Binding Pocket 1 . Three Highest VdW contributors (blue) and Three Highest NPS (green) highlighted.

BP1 – Res Decomposition	Residue	VdW	V_{EL}	Polar Solvation	Non-polar solvation	TOT-Average
0	LEU 245	-0.31199767	-0.05527987	0.02304681	-0.19668453	-0.54091526
1	CYS 246	-0.7205343	0.06651534	-0.08157907	-0.49819894	-1.23379696
2	VAL 280	-0.61202644	0.06468241	-0.13481491	-0.56419464	-1.24635358
3	LEU 281	-0.92712268	0.00170399	0.18595986	-0.86840226	-1.60786109
4	**PHE 284**	**-1.41351973**	0.07307267	0.21387102	**-0.94481052**	**-2.07138655**
5	PHE 393	-0.81380496	-0.12602776	0.21546045	-0.72596289	-1.45033515
6	ILE 429	-0.48105409	-0.06250211	0.07773221	-0.42396891	-0.8897929
7	**PHE 434**	**-1.0995809**	-0.01179869	0.02515851	**-0.82296079**	**-1.90918187**
8	**LEU 437**	**-1.046171**	-0.09032057	0.16590175	**-0.94544469**	**-1.91603452**
9	ILE 441	-0.28609643	-0.01806221	0.0830148	-0.25967539	-0.48081922
10	PHE 571	-0.43072712	-0.07711977	0.13761062	-0.42626747	-0.79650374

In concordance to Table 4.5 the dominating interactions seem to be driven by van der Waals interactions (VdW). On the other hand, the second dominant solvation seems to be non-polar solvation interactions. This means that, the non-polar interactions of ligand and receptor separately is less favourable than when they are binded together. Non-polar solvation is driven by several factors. The first is the van der Waals attraction, repulsion between solvent and solute, this term will become a bit less favourable when the binding occurs since part of the ligand and the receptor will be employed in the binding. The second term to consider is the cavitation which is related with the energy needed to generate a cavity in the solvent to accommodate the solute. The hypothesis laid here is to suppose that when the binding occurs the energy needed to accommodate the complex is reduced. This causes a more favourable interaction between solvent and complex than solvent and non-binded solvent and ligand. Motivation for this comes from the significant of non-polar contributions for each pocket and for the solvation free energies for each residue. The three most contributing aminoacids are shown in Fig. 4.13.

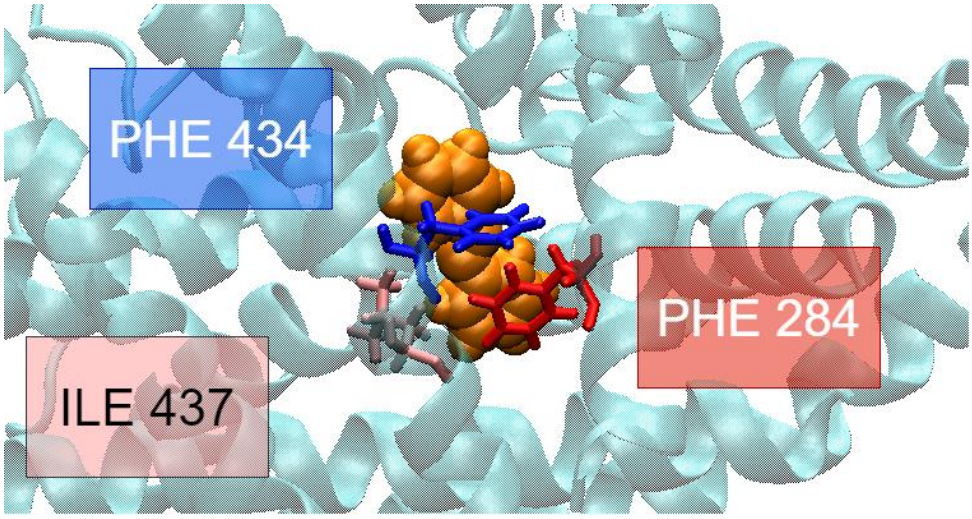

Fig. 4.13 Snapshot of Azobenzene in Binding Pocket 1 with the Three Most Interacting Residues from Table 4.6.

Pi stacking is used in several areas of chemistry to define an important energy contribution between aromatic groups. It is manifested usually by interaction between aromatic rings with three characteristic structures. The first being displaced face to face stacking, the second being edge to face stacking (T-stacking), and the least favourable aligned face to face stacking[112]. The description of this interaction is still a matter of debate[113], in addition to this, the CHARMM36m force field does not have a way to explicitly define this stacking interactions. Nevertheless, as seen above in Fig. 4.13 PHE 284 is stacking on top of the azo (-N=N-) group and PHE 434 is stacking in a T-stacking way with one of the rings of Azobenzene. These structures confirm favourable interaction between Azobenzene and aromatic residues. However, while stacking with PHE gives favourable energies, interactions with non-aromatic residues like ILE still result in favourable interactions as well. Values in Table 4.6 seem to indicate that stacking with aromatic residues has a large contribution to the interaction energy, nevertheless, non-aromatic residues play an important role too. Azobenzene seems to stack in a significant way with PHE and not with other aromatic residues like TRP (only in binding pocket 1 sub pocket 2) and TYR.

Due to the positions of the binding pockets as seen in Fig. 4.12, part of the favourable interaction could be caused by protection of the hydrophobic cavity between D II and III upon binding of Azobenzene. Hydrophobic aminoacids in proteins tend to bury into the protein to stay away from the solvent. When Azobenzene binds, it covers part of the exposed surface of these hydrophobic residues increasing stability inside the protein. Indeed, recalling section 3.3, the SASA method, which calculates the non-polar solvation interaction in MM/GBSA by measuring the exposure of residues to solvent, predicts more favourable interactions when azobenzene binds.

In summary, while van der Waals interactions between Azobenzene and channel seem to be the dominating interactions when stabilizing the complex, non-polar solvation has a significant contribution for Azobenzene residue interaction probably due to the probable reduction of cavitation energy in the solvent combined with surface protection from the solvent of hydrophobic residues.

A point to consider is the particular stacking of PHE 284 with azobenzene is seen in Fig. 4.14. The stacking of PHE occurs on top of the azo group causing a double face displaced stacking with respect to both rings. Azobenzene is a highly conjugated molecule so it would be interesting to carry an investigation regarding whether this situation leads to any extra stabilization.

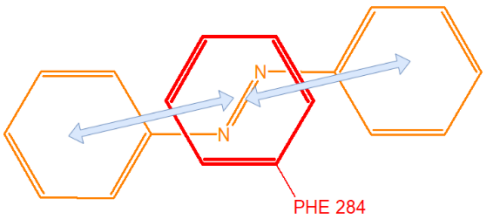

Fig. 4.14 Scheme of PHE 284 Stacking Azobenzene

For binding pocket 1 sub pocket 2 (Appendix B Table 8.1 and Fig. 8.4) the main residue interaction is again with PHE 284, being the most relevant contributions VdW and non-polar solvation. Due to this, and the total energy similarity with binding pocket 1 sub pocket 1, it reaffirms the decision to classify both as sub-pockets inside of the same pocket. Apart from PHE 284 major interactions are with LEU 281 and TYR 578.

Binding pocket 2 seems to be the most stable of all the pockets as seen in Table 4.5. The 1000ns cMD replicas did not seem to sample in an extensive way this pocket (Fig. 4.7 B)). However, once a geometry of this pocket was isolated and exploration was done using 100ns of cMD Azobenzene remained without scaping. This could be interpreted as a point in favour GaMD since it managed to predict in a clear way the lowest energy binding pocket while cMD did not sample it properly. Again, main interactions are with hydrophobic residues (Appendix B Table 8.2 CYS 246, VAL 280 , PHE 393 and PHE 434) where stacking appears with two PHE (Appendix B Fig. 8.5).

Binding pocket 3 is the least stable as seen in Table 4.5 . The same kind of interactions found for other pockets dominate the binding (VdW and non-polar solvation). Main interactions are experienced with hydrophobic aminoacids (VAL 280, ILE 430 and PHE 434). In this case although PHE is present it does not have the highest energy contribution, breaking the trend observed until now (Appendix B Table 8.3 and Fig. 8.6).

Regarding binding pocket 4 as discussed before, cMD and GaMD simulations do not seem to agree with the high stability predicted by MM/GBSA. Therefore, the results obtained regarding this pocket should be taken as orientative. For this pocket no aromatic residues had a significant contribution, binding was driven by VdW and non-polar solvation interactions, and hydrophobic residues were responsible for interactions (LEU 285, ILE 445, ILE 579, ILE 582) (Appendix B Table 8.4 and Fig. 8.7).

Overall, from energy contributions and Azobenzene's reverse probability distribution (Fig. 4.7*)* it seems that binding will mainly take place in binding pocket 1 and binding pocket 2.

A final thought to consider is that the interactions described above would probably change upon photoisomerization of azobenzene due to the apparition of a total dipole moment in the molecule because of the change in symmetry. This would probably alter the interactions that drive binding, probably making Azobenzene to interact with polar residues leaving the hydrophobic cavity exposed and causing important changes in non-polar solvation terms.

5 Conclusions

Voltage gated ion channels are involved in many relevant biological processes, making the elucidation of their atomic resolution structure a way to understand and potentially treat diseases originated from channelopathies. Binding of Azobenzene (a photoswitch) to a potential drug target, the human Na_v 1.4 ion channel present in skeletal respiratory muscle and cardiac tissue whose resolved structure dates from 2018, has been investigated by computational approaches.

The use of GaMD has proven to enhance the sampling of the phase space significantly, improving the identification of binding pockets and even predicting an additional binding pocket which cMD did not visit (this is extracted from of using each technique in four simulations of 1000 ns). One of the advantages that GaMD offers is an improved reweighting of the free energy surface by adding near-gaussian harmonic potentials as a boost to the system. Three methods are offered to perform the reweighting, the first is exponential average, the second is McLaurin series and the third is cumulant expansion to the 2^{nd} order. According to authors, cumulant expansion to the 2^{nd} order provides the best results, however, for systems larger than 100 residues the energetic noise becomes too high to use this method[16]. Since the system simulated in this project was about 550 residues McLaurin series of order 10 was used since it gave best results than exponential average. The free energy surface was obtained and used to identify 4 binding pockets (with the identification of two sub pockets inside binding pocket 1).

Binding of azobenzene inside the α structure of the channel seems to occur mainly close to DII or inserted in a hydrophobic cavity between DII and DIII. Binding pocket 4 was identified near the lower gate of the protein, binding pocket 1 (with its two sub pockets) was identified by the entrance of the hydrophobic cavity, binding pocket 2 was inserted inside the cavity, and binding pocket 3 appeared partly inserted at the end of the cavity. The 4 binding pockets seem to form an “arch” which spans from the lower gate of the channel through the aforementioned hydrophobic cavity. Both cMD and GaMD seem to indicate that main binding will probably occur in binding pockets 1 and 2. The quality of the predicted binding pockets was tested by 100 ns of cMD exploration, showing that Azobenzene escaped the relatively unstable pocket predicted only by GaMD (binding pocket 4).

Regarding the increment in binding free energy made by MM/GBSA, truncation of the cMD pocket exploration trajectory that started in binding pocket 4 seems to generate some artifacts. While Azobenzene escaped before 50ns into the cMD simulation and GaMD’s free energy map predicted less relative stability compared to others, MM/GBSA predicted it to be the most stable. As expected by the non-polar nature of Azobenzene, the main interaction that drove binding in all pockets was van der Waals contribution. However, non-polar solvation effects contributed in a significant way, especially when studying the per residue decomposition of the free energies. Non-polar solvation contributions are stabilizing the ligand channel complex upon binding. This is probably driven by the blocking of the hydrophobic cavity by Azobenzene, protecting it from solvent contact, and reducing the cavitation energy cost in the solvation. Characteristic “pi-stacking” conformations were identified between Azobenzene and aromatic residues of the channel, especially with PHE. These PHE residues seem to be responsible for large part of non-polar solvation effects through the disposition of stacking conformations with Azobenzene.

In summary, GaMD has been proven to enhance the sampling of the phase space, and while not being able to use cumulant's expansion for reduced noise reweighting, 10 Order McLaurin series allowed for free energy surface recovery where 4 main binding pockets were identified and 2 sub pockets were localized inside binding pocket 1. The energy nature behind binding pocket stabilization was done mainly by van der Waals interactions, however, non-polar solvation still had a significant fraction of the per residue decomposition interactions. Therefore, the objective of the project seems fulfilled as the description of the binding between Azobenzene and the human Na_v 1.4 seems to have been described.

As indicated in the project objectives (section 2.4) , the work carried here is just the first phase of a much larger study that attempts to describe the regulation of the flow of ions through the channel by photo excitation of Azobenzene. As an outlook to the project, the next step would be to study the light absorption spectrum of Azobenzene in water and inside the channel to determine possible changes due to environment effects. Furthermore , functionalization of Azobenzene (with groups like -NH_2 or -OH) should alter in an interesting way its binding properties and photochemistry, opening a wide range of possible photoswitches that could be tested if Azobenzene fails to regulate the flow of ions across the channel. Regarding the use of GaMD, setting the energy threshold to the upper bound should provide even more acceleration of the dynamics, which could be an interesting choice for future simulations. In addition to this, Yinglong Miao, one of the contributors of GaMD, described in a recent publication draft (April 2020) LiGaMD[111], a technique that only biases non-bonded interactions and that should be useful for studying binding processes like the one carrier out in this project.

6 References

The last consult for all internet references was done in July 2020

1. Foye's Principles of Medicinal Chemistry, 7e | Pharmacy | Health Library. https://pharmacy.lwwhealthlibrary.com/book.aspx?bookid=758 (June 2020).
2. Drug Discovery: A Historical Perspective | Science. https://science.sciencemag.org/content/287/5460/1960.abstract (June 2020).
3. Lien, E. J., Ren, S., Bui, H.-H. & Wang, R. Quantitative structure-activity relationship analysis of phenolic antioxidants. *Free Radic. Biol. Med.* **26**, 285–294 (1999).
4. Ekins, S., Mestres, J. & Testa, B. In silico pharmacology for drug discovery: applications to targets and beyond. *Br. J. Pharmacol.* **152**, 21–37 (2007).
5. Bagal, S. K. *et al.* Ion channels as therapeutic targets: a drug discovery perspective. *J. Med. Chem.* **56**, 593–624 (2013).
6. Snake Venoms in Drug Discovery: Valuable Therapeutic Tools for Life Saving. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6832721/.
7. Sarma, P. & Medhi, B. Photopharmacology. *Indian J. Pharmacol.* **49**, 221–222 (2017).
8. Mourot, A., Tochitsky, I. & Kramer, R. H. Light at the end of the channel: optical manipulation of intrinsic neuronal excitability with chemical photoswitches. *Front. Mol. Neurosci.* **6**, 5 (2013).
9. A. Beharry, A. & Andrew Woolley, G. Azobenzene photoswitches for biomolecules. *Chem. Soc. Rev.* **40**, 4422–4437 (2011).
10. Deal,W.J.,Erlanger,B.F., and Nachmansohn,D.(1969). Photoregulation of biological activity by photochromic reagents. 3. Photoregulation of bioelectricity by acetylcholine receptor inhibitors. Proc.Natl.Acad.Sci.U.S.A. 64, 1230–1234.
11. Barone, V. & Polimeno, A. Integrated computational strategies for UV/vis spectra of large molecules in solution. *Chem. Soc. Rev.* **36**, 1724–1731 (2007).
12. Structure of the human voltage-gated sodium channel Nav1.4 in complex with β1 | Science. https://science.sciencemag.org/content/362/6412/eaau2486. (June 2020)
13. Braun, E. *et al.* Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **1**, 5957 (2018).
14. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation | Journal of Chemical Theory and Computation. https://pubs.acs.org/doi/abs/10.1021/acs.jctc.5b00436. (June 2020)
15. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10**, 449–461 (2015).
16. PyReweighting: Energetic reweighting of accelerated molecular dynamics Simulations. http://miao.compbio.ku.edu/PyReweighting/.(June 2020)
17. *Python 3 Van Rossum, G. & Drake, F.L., 2009. Python 3 Reference Manual, Scotts Valley, CA: CreateSpace.* (2009).
18. *Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). Principles of neural science. New York: McGraw-Hill, Health Professions Division.*
19. Heard, K., Palmer, R. & Zahniser, N. R. Mechanisms of acute cocaine toxicity. *Open Pharmacol. J.* **2**, 70–78 (2008).
20. Matthews, J. C. & Collins, A. Interactions of cocaine and cocaine congeners with sodium channels. *Biochem. Pharmacol.* **32**, 455–460 (1983).
21. Nogueira, J. J. & Corry, B. Ion Channel Permeation and Selectivity. *The Oxford Handbook of Neuronal Ion Channels* https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190669164.001.0001/oxfordhb-9780190669164-e-22 (2018) doi:10.1093/oxfordhb/9780190669164.013.22. (June 2020)
22. Catterall, W. A. Ion Channel Voltage Sensors: Structure, Function, and Pathophysiology. *Neuron* **67**, 915–928 (2010).
23. Purves, D. *et al.* The Molecular Structure of Ion Channels. *Neurosci. 2nd Ed.* (2001).
24. Jegla, T. J., Zmasek, C. M., Batalov, S. & Nayak, S. K. Evolution of the Human Ion Channel Set. *Comb. Chem. High Throughput Screen.* **12**, 2–23 (2009).
25. Ranganathan, R. Evolutionary origins of ion channels. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 3484–3486 (1994).
26. Strong, M., Chandy, K. G. & Gutman, G. A. Molecular evolution of voltage-sensitive ion channel genes: on the origins of electrical excitability. *Mol. Biol. Evol.* **10**, 221–242 (1993).
27. Männikkö, R. *et al.* Dysfunction of NaV1.4, a skeletal muscle voltage-gated sodium channel, in sudden infant death syndrome: a case-control study. *The Lancet* **391**, 1483–1492 (2018).
28. Wang, G. K., Calderon, J. & Wang, S.-Y. State- and Use-Dependent Block of Muscle Nav1.4 and Neuronal Nav1.7 Voltage-Gated Na+ Channel Isoforms by Ranolazine. *Mol. Pharmacol.* **73**, 940–948 (2008).
29. Bank, R. P. D. RCSB PDB - 6AGF: Structure of the human voltage-gated sodium channel Nav1.4 in complex with beta1. https://www.rcsb.org/structure/6AGF. (June 2020)
30. *Humphrey, W., Dalke, A. and Schulten, K., 'VMD - Visual Molecular Dynamics', J. Molec. Graphics, 1996, vol. 14, pp. 33-38. http://www.ks.uiuc.edu/Research/vmd/.*(June 2020)
31. Ellis-Davies, G. C. R. Caged compounds: photorelease technology for control of cellular chemistry and physiology. *Nat. Methods* **4**, 619–628 (2007).
32. Zhu, M. & Zhou, H. Azobenzene-based small molecular photoswitches for protein modulation. *Org. Biomol. Chem.* **16**, 8434–8445 (2018).
33. Cabré, G. *et al.* Rationally designed azobenzene photoswitches for efficient two-photon neuronal excitation. *Nat. Commun.* **10**, 907 (2019).
34. Dong, M., Babalhavaeji, A., Samanta, S., Beharry, A. A. & Woolley, G. A. Red-Shifting Azobenzene Photoswitches for in Vivo Use. *Acc. Chem. Res.* **48**, 2662–2670 (2015).

35. How Azobenzene Photoswitches Restore Visual Responses to the Blind Retina. *Neuron* **92**, 100–113 (2016).
36. Boelke, J. & Hecht, S. Designing Molecular Photoswitches for Soft Materials Applications. *Adv. Opt. Mater.* **7**, 1900404 (2019).
37. Böckmann, M., Doltsinis, N. L. & Marx, D. Azobenzene photoswitches in bulk materials. *Phys. Rev. E* **78**, 036101 (2008).
38. Nist Data for Trans-Azobeznene https://webbook.nist.gov/cgi/cbook.cgi?ID=C103333&Mask=400. (June 2020)
39. den Hertog, H. J., Henkens, C. H. & van Roon, J. H. Reactivity of 4-nitropyridine-n-oxide (III) Reduction in alkaline medium. *Recl. Trav. Chim. Pays-Bas* **71**, 1145–1151 (1952).
40. NIST data for Cis-Azobenzene https://webbook.nist.gov/cgi/cbook.cgi?ID=C1080166&Units=SI&Mask=400#UV-Vis-Spec.
41. Le Ferve, R.J.W.; Northcott, J., J. Chem. Soc., 1952, 4082.
42. Jafari, M. R., Lakusta, J., Lundgren, R. J. & Derda, R. Allene Functionalized Azobenzene Linker Enables Rapid and Light-Responsive Peptide Macrocyclization. *Bioconjug. Chem.* **27**, 509–514 (2016).
43. Browne, W. R. & Feringa, B. L. Making molecular machines work. in *Nanoscience and Technology* 79–89 (Co-Published with Macmillan Publishers Ltd, UK, 2009). doi:10.1142/9789814287005_0009.
44. Joshi, G. K. *et al.* Ultrasensitive Photoreversible Molecular Sensors of Azobenzene-Functionalized Plasmonic Nanoantennas. *Nano Lett.* **14**, 532–540 (2014).
45. Sadovski, O., Beharry, A. A., Zhang, F. & Woolley, G. A. Spectral Tuning of Azobenzene Photoswitches for Biological Applications. *Angew. Chem. Int. Ed.* **48**, 1484–1486 (2009).
46. Theoretical Study of the Isomerization Mechanism of Azobenzene and Disubstituted Azobenzene Derivatives | The Journal of Physical Chemistry A. https://pubs.acs.org/doi/abs/10.1021/jp057413c.
47. Fujino, T., Arzhantsev, S. Yu. & Tahara, T. Femtosecond/Picosecond Time-Resolved Spectroscopy of trans- Azobenzene: Isomerization Mechanism Following S2(ππ*) ← S0Photoexcitation. *Bull. Chem. Soc. Jpn.* **75**, 1031–1040 (2002).
48. MoBioChem - YouTube. https://www.youtube.com/channel/UC49JN0vDkGzBxHs_8ze20Lg. (June 2020)
49. Alder, B. J. & Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.* **31**, 459–466 (1959).
50. Jensen, F. *Introduction to Computational Chemistry*. (John Wiley & Sons, 2017).
51. Eng, J., Gourlaouen, C., Gindensperger, E. & Daniel, C. Spin-Vibronic Quantum Dynamics for Ultrafast Excited-State Processes. *Acc. Chem. Res.* **48**, 809–817 (2015).
52. Thoss, M. & Wang, H. Quantum dynamical simulation of ultrafast molecular processes in the condensed phase. *Chem. Phys.* **322**, 210–222 (2006).
53. Bottaro, S. & Lindorff-Larsen, K. Biophysical experiments and biomolecular simulations: A perfect match? *Science* **361**, 355–360 (2018).
54. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins | Chemical Reviews. https://pubs.acs.org/doi/10.1021/cr040426m. (June 2020)
55. Ispas, S., Benoit, M., Jund, P. & Jullien, R. Structural properties of glassy and liquid sodium tetrasilicate: comparison between ab initio and classical molecular dynamics simulations. *J. Non-Cryst. Solids* **307–310**, 946–955 (2002).
56. Anharmonic force constants extracted from first-principles molecular dynamics: applications to heat transfer simulations - IOPscience. https://iopscience.iop.org/article/10.1088/0953-8984/26/22/225402/meta.
57. Application of molecular dynamics simulations to the study of ion-bombarded metal surfaces: Critical Reviews in Solid State and Materials Sciences: Vol 14, No sup1. https://www.tandfonline.com/doi/abs/10.1080/10408438808244782. (June 2020)
58. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
59. Atkins. Química Física de Peter Atkins | Editorial Médica Panamericana. https://www.medicapanamericana.com/es/libro/atkins-quimica-fisica. (June 2020)
60. Bernardi, R. C., Melo, M. C. R. & Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1850**, 872–877 (2015).
61. Umbrella sampling - Kästner - 2011 - WIREs Computational Molecular Science - Wiley Online Library. https://onlinelibrary.wiley.com/doi/full/10.1002/wcms.66.
62. Miao, Y. & McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Theory, Implementation, and Applications. *Annu. Rep. Comput. Chem.* **13**, 231–278 (2017).
63. Molecular Dynamics - chapter 1: Equations of Motion - YouTube. https://www.youtube.com/watch?v=8iHER6IP6Ds.
64. Numerical Differential Equation Methods. in *Numerical Methods for Ordinary Differential Equations* 55–142 (John Wiley & Sons, Ltd, 2016). doi:10.1002/9781119121534.ch2.
65. Smith, R. & Harrison, D. E. Algorithms for molecular dynamics simulations of keV particle bombardment. *Comput. Phys.* **3**, 68 (1989).
66. Amber18 Manual.
67. Zapletal, V. *et al.* Choice of Force Field for Proteins Containing Structured and Intrinsically Disordered Regions. *Biophys. J.* **118**, 1621–1633 (2020).
68. Sponer, J., Leszczynski, J. & Hobza, P. Hydrogen bonding and stacking of DNA bases: a review of quantum-chemical ab initio studies. *J. Biomol. Struct. Dyn.* **14**, 117–135 (1996).
69. Callis, P. R. & Vivian, J. T. Understanding the variable fluorescence quantum yield of tryptophan in proteins using QM-MM simulations. Quenching by charge transfer to the peptide backbone. *Chem. Phys. Lett.* **369**, 409–414 (2003).

70. Grigorenko, B. L., Nemukhin, A. V., Topol, I. A., Cachau, R. E. & Burt, S. K. QM/MM modeling the Ras–GAP catalyzed hydrolysis of guanosine triphosphate. *Proteins Struct. Funct. Bioinforma.* **60**, 495–503 (2005).
71. Senn, H. M. & Thiel, W. QM/MM Methods for Biological Systems. in *Atomistic Approaches in Modern Biology: From Quantum Chemistry to Molecular Simulations* (ed. Reiher, M.) 173–290 (Springer, 2007). doi:10.1007/128_2006_084.
72. Reiher, M. Approaches in Modern Biology.
73. *J. D. Hunter, 'Matplotlib: A 2D Graphics Environment', Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.*
74. Harrison, J. A. *et al.* Review of force fields and intermolecular potentials used in atomistic computational materials research. *Appl. Phys. Rev.* **5**, 031104 (2018).
75. Jing, Z. *et al.* Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annu. Rev. Biophys.* **48**, 371–394 (2019).
76. Ponder, J. W. *et al.* Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **114**, 2549–2564 (2010).
77. Patel, D. S., He, X. & MacKerell, A. D. Polarizable Empirical Force Field for Hexopyranose Monosaccharides Based on the Classical Drude Oscillator. *J. Phys. Chem. B* **119**, 637–652 (2015).
78. *Molecular Dynamics - chapter 3: Periodic Boundary Conditions, Temperature and Pressure*.
79. Wells, B. A. & Chaffee, A. L. Ewald Summation for Molecular Simulations. *J. Chem. Theory Comput.* **11**, 3684–3695 (2015).
80. Louwerse, M. J. & Baerends, E. J. Calculation of pressure in case of periodic boundary conditions. *Chem. Phys. Lett.* **421**, 138–141 (2006).
81. Miao, Y. *et al.* Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. *J. Chem. Theory Comput.* **10**, 2677–2689 (2014).
82. Schmitz, G. A. Machine Learning for Potential Energy Surface Construction: A Benchmark Set. (2019) doi:10.7910/DVN/C9ISSX.
83. Schmitz, G. A. ammonia-DZ-F12-ADGA-DEFAULT-1M_PES-DMP2.xyz. (2019) doi:10.7910/DVN/C9ISSX/CRVGEA.
84. Fogolari, F., Brigo, A. & Molinari, H. The Poisson–Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recognit.* **15**, 377–392 (2002).
85. Onufriev, A. V. & Case, D. A. Generalized Born Implicit Solvent Models for Biomolecules. *Annu. Rev. Biophys.* **48**, 275–296 (2019).
86. Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R. & Meiler, J. Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *J. Mol. Model.* **15**, 1093–1108 (2009).
87. CHARMM-GUI. http://www.charmm-gui.org/?doc=input/ligandrm. (June 2020)
88. Kim, S. *et al.* CHARMM-GUI ligand reader and modeler for CHARMM force field generation of small molecules. *J. Comput. Chem.* **38**, 1879–1886 (2017).
89. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).
90. CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2888302/.
91. McCullagh, M., Franco, I., Ratner, M. A. & Schatz, G. C. DNA-Based Optomechanical Molecular Motor. *J. Am. Chem. Soc.* **133**, 3452–3459 (2011).
92. *Gaussian 16, Revision C.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.*
93. Woods, R. J. & Chappelle, R. Restrained electrostatic potential atomic partial charges for condensed-phase simulations of carbohydrates. *Theochem* **527**, 149–156 (2000).
94. Wang, J., Wang, W., Kollman P. A.; Case, D. A. 'Automatic atom type and bond type perception in molecular mechanical calculations'. Journal of Molecular Graphics and Modelling , 25, 2006, 247260.
95. Kingsland, A., Samai, S., Yan, Y., Ginger, D. S. & Maibaum, L. Local Density Fluctuations Predict Photoisomerization Quantum Yield of Azobenzene-Modified DNA. *J. Phys. Chem. Lett.* **7**, 3027–3031 (2016).
96. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).
97. Lee, J. *et al.* CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **12**, 405–413 (2016).
98. Jo, S., Lim, J. B., Klauda, J. B. & Im, W. CHARMM-GUI Membrane Builder for Mixed Bilayers and Its Application to Yeast Membranes. *Biophys. J.* **97**, 50–58 (2009).
99. Jo, S., Kim, T. & Im, W. Automated Builder and Database of Protein/Membrane Complexes for Molecular Dynamics Simulations. *PLoS ONE* **2**, e880 (2007).
100. Lee, J. *et al.* CHARMM-GUI *Membrane Builder* for Complex Biological Membrane Simulations with Glycolipids and Lipoglycans. *J. Chem. Theory Comput.* **15**, 775–786 (2019).
101. CHARMM36m: an improved force field for folded and intrinsically disordered proteins - CHARMM. https://www.charmm.org/charmm/showcase/featured-research/charmm36m-an-improved-force-field-for-folded-and-intrinsically-disordered-proteins/.(June 2020)
102. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).

103. TIP3P Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. J. Chem. Phys.1983, 79, 926.
104. TIP3P Modified Neria, E.; Fischer, S.; Karplus, M. J. Chem. Phys. 1996, 105, 1902.
105. *AMBER 18 D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. SalomonFerrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman (2018), AMBER 2018, University of California, San Francisco.*
106. Joung, I. S. & Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **112**, 9020–9041 (2008).
107. ParmEd — ParmEd documentation. http://parmed.github.io/ParmEd/html/index.html.
108. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations | Journal of Chemical Theory and Computation. https://pubs.acs.org/doi/abs/10.1021/ct300418h.
109. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
110. Bhattarai, A. & Miao, Y. Gaussian Accelerated Molecular Dynamics for Elucidation of Drug Pathways. *Expert Opin. Drug Discov.* **13**, 1055–1065 (2018).
111. Ligand Gaussian accelerated molecular dynamics (LiGaMD): Characterization of ligand binding thermodynamics and kinetics | bioRxiv. https://www.biorxiv.org/content/10.1101/2020.04.20.051979v1.
112. McGaughey, G. B., Gagné, M. & Rappé, A. K. pi-Stacking interactions. Alive and well in proteins. *J. Biol. Chem.* **273**, 15458–15463 (1998).
113. R. Martinez, C. & L. Iverson, B. Rethinking the term “pi-stacking”. *Chem. Sci.* **3**, 2191–2201 (2012).

7 Appendix A: Azobenzene Model Construction Parameters

Table 7.1 Parameters for Azobenzene's azo group dihedrals angles. Adapted from McCullagh, Franco, Ratner and Schatz[91]

Dihedral Angle	Force constant / kcalmol^{-1} Å^{-2}	Equilibrium value (º)
-C-N=N-C	14.0000	180.0
-C-C-N=N-	4.4250	180.0

Table 7.2 Azobenzene Geometry Optimized with MP2 and Partial Charges Calculated with HF 6-31G(d) and fitted with RESP

Atom	Partial Charge
C1	0.044990
C2	-0.180867
C3	-0.171029
C4	-0.180867
C5	-0.044990
C6	0.091007
N1	-0.120885
N2	-0.120885
C7	0.091007
C8	-0.044990
C9	-0.180867
C10	-0.171029
C11	-0.180867
C12	-0.044990
H1	0.156361
H2	0.089272
H3	0.161355
H4	0.156361
H5	0.089272
H6	0.089272
H7	0.156361
H8	0.161355
H9	0.156361
H10	0.089272

8 Appendix B: Results

Figures

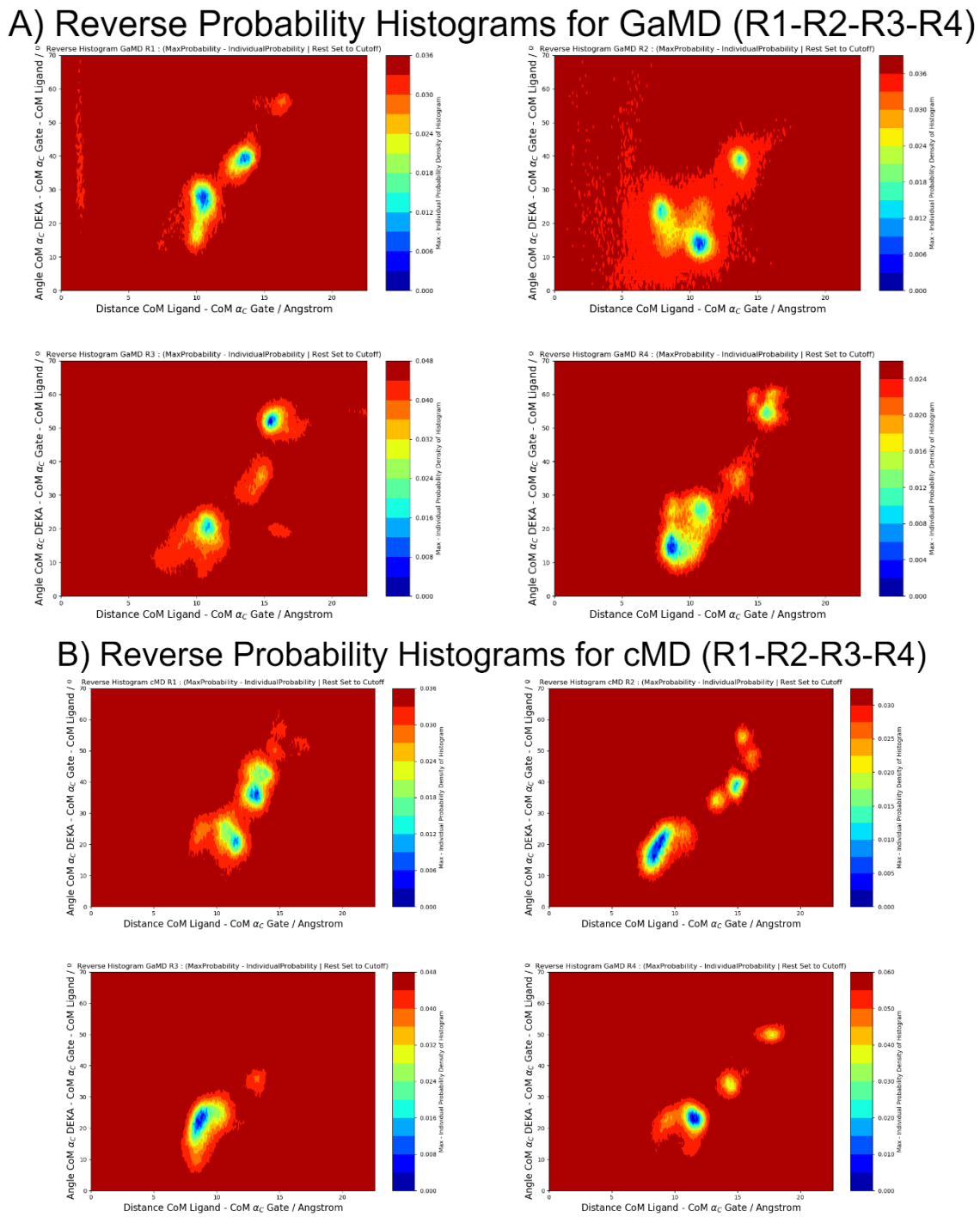

Fig. 8.1 GaMD vs cMD Reverse Histogram Comparison (Each Replica is 1000ns) : A) GaMD Reverse Probability Distributions. B) cMD Reverse Probability Distributions. Each Replica (8 in total 4 GaMD and 4 cMD) is 1000ns. At first Sight It seems that GaMD has more "diffuse" contours, indicating a possible acceleration and biasing of the trajectory. Nevertheless, both techniques seem to predict three highly visited conformations. A better way is to visualize all of them together and then make the comparison

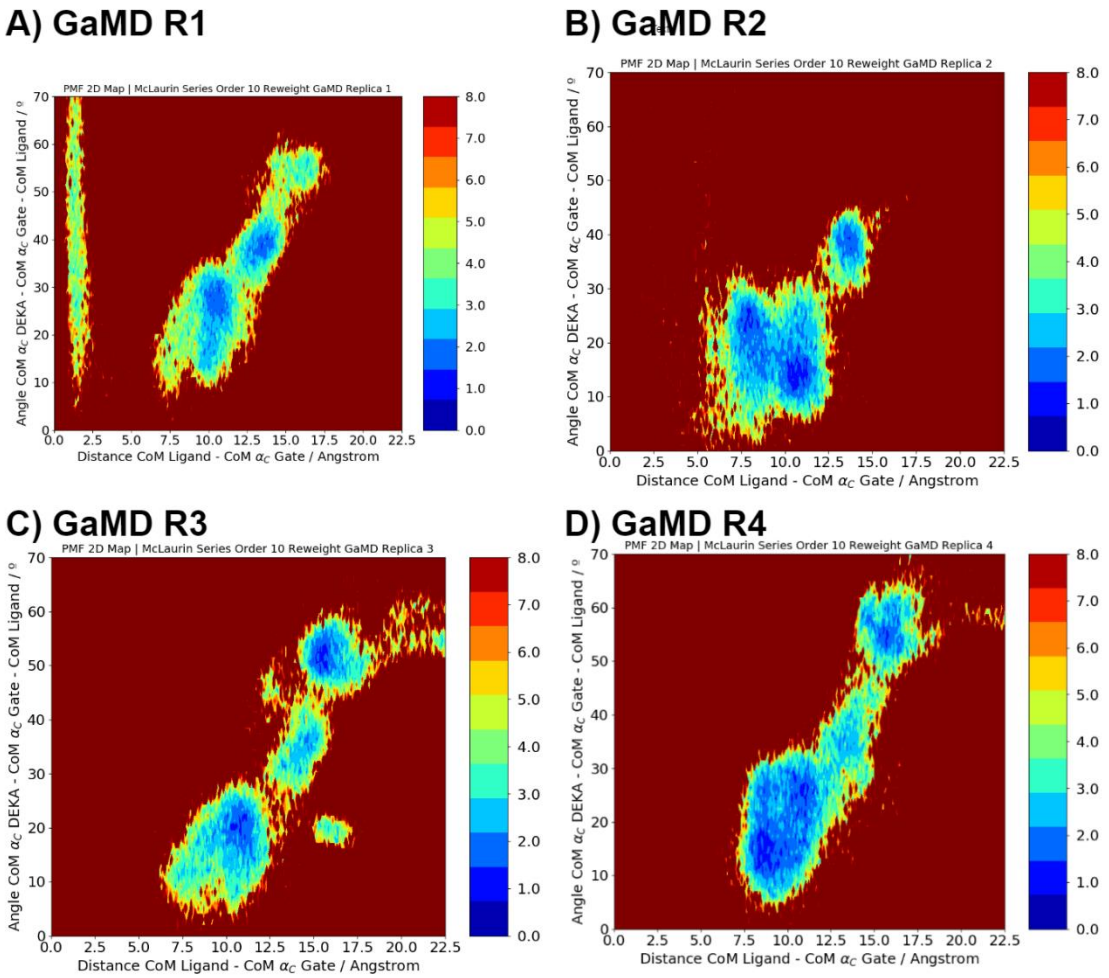

Fig. 8.2 Free energy Surface Obtained by Reweighting the Trajectory of Each GaMD Replica (colour bar in kcal/mol) : A) GaMD R1 B) GaMD R2 C) GaMD R3 D) GaMD R4 . Each replica lasts for 1000ns

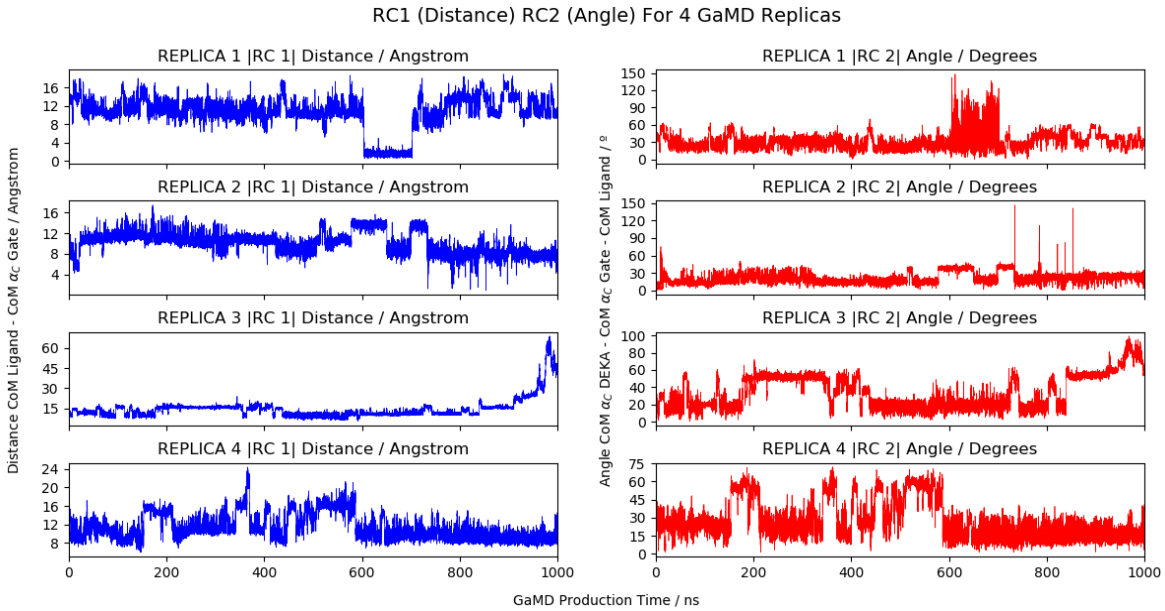

Fig. 8.3 Temporal Evolution of the RC Along Each GaMD Replica

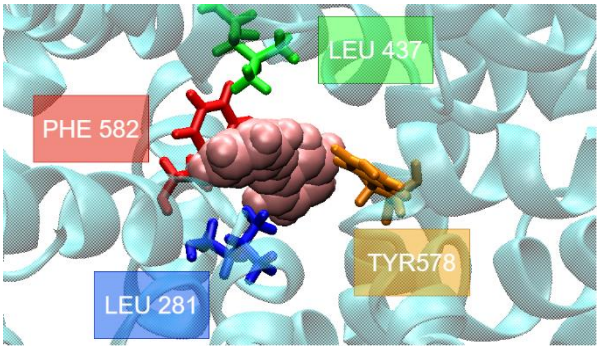

Fig. 8.4 Four Most Interacting Residues with Azobenzene, Binding Pocket 1 Sub Pocket 2 . Partial Stacking with PHE 582 Can Be Appreciated.

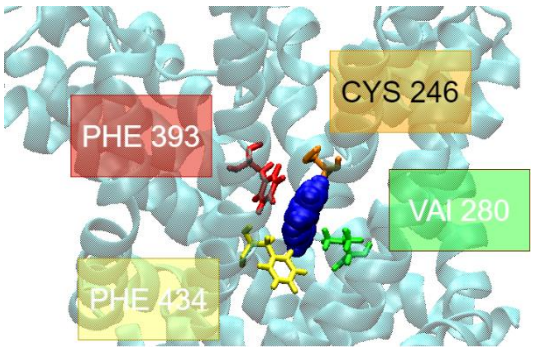

Fig. 8.5 Four Most Interacting Residues with Azobenzene, Binding Pocket 2 .Edge to Face Stacking by PHE434 can be Appreciated

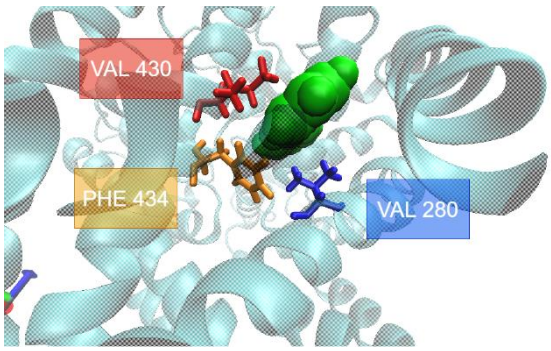

Fig. 8.6 Three Most Interacting Residues with Azobenzene, Binding Pocket 3 .Edge to Face Stacking by PHE434 Can Be Appreciated.

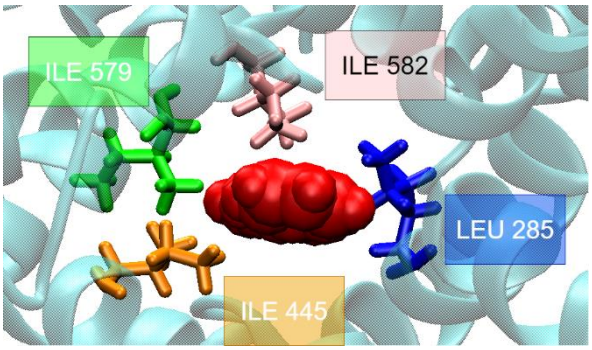

Fig. 8.7 Four most Interacting Residues with Azobenzene, Binding Pocket 4. No Presence of Aromatic Rings.

Tables

Table 8.1 Per residue free energy (kcal/mol) decomposition for binding pocket 1 sub pocket 1.Four Highest VdW contributors (red) and Three Highest NPS (green) highlighted.

BP1 Sub pocket 2 - Res Decomposition	Residue	VdW	V_{EL}	Polar Solvation	Non-polar solvation	TOT-Average
0	LEU 161	-0.35496744	-0.02813886	0.10928401	-0.31944789	-0.59327018
1	CYS 246	-0.41754286	0.0057384	-0.00210357	-0.34048562	-0.75439365
2	VAL 280	-0.38110099	-0.01438209	-0.06932317	-0.32142562	-0.78623187
3	**LEU 281**	**-1.12869778**	-0.13025986	0.26740964	**-1.00682856**	**-1.99837655**
4	**PHE 284**	**-1.77962309**	0.11633217	0.17644529	**-1.2075191**	**-2.69436472**
5	LEU 285	-0.37249038	0.04812204	0.05144527	-0.27637683	-0.5492999
6	PHE 393	-0.71762127	-0.11767989	0.189737	-0.72722501	-1.37278917
7	PHE 434	-0.37676955	0.09885033	-0.03926949	-0.27675731	-0.59394602
8	**LEU 437**	**-0.84071034**	-0.02988195	0.11410846	**-0.7362334**	**-1.49271722**
9	ILE 441	-0.75614339	-0.10551429	0.22638989	-0.62577503	-1.26104282
10	VAL 575	-0.46616191	-0.10934148	0.14975639	-0.34704687	-0.77279387
11	**TYR 578**	**-0.99335446**	-0.01466166	0.07227834	**-0.83388318**	**-1.76962095**
12	ILE 579	-0.46925352	0.02509257	0.03383024	-0.41861261	-0.82894333
13	ILE 582	-0.34417445	0.05339619	-0.00705385	-0.34502773	-0.64285984

Table 8.2 Per residue free energy (kcal/mol) decomposition for binding pocket 2. Four Highest VdW contributors (red) and Three Highest NPS (green) highlighted

BP2 – Res Decomposition	Residue	VdW	V_{EL}	Polar Solvation	Non-polar solvation	TOT-Average
0	PHE 242	-0.74117681	-0.07713025	0.14785893	-0.53629377	-1.2067419
1	LEU 245	-1.02209033	0.27921427	-0.17300846	-0.7164342	-1.63231872
2	CYS 246	**-1.41549378**	0.00644036	0.00393937	-0.94797279	**-2.35308684**
3	VAL 280	**-1.23194775**	-0.22448506	0.11738034	**-1.01432676**	**-2.35337923**
4	LEU 281	-0.7100776	-0.01412657	0.23122847	-0.72711952	-1.22009521
5	PHE 284	-0.81721845	-0.22271644	0.19536452	-0.57997173	-1.42454211
6	PHE 393	**-1.85180173**	0.0556601	0.41076488	**-1.25162963**	**-2.63700638**
7	ILE 429	-0.8370606	-0.03337388	-0.03216797	-0.59677208	-1.49937453
8	ILE 430	-0.65214076	0.10891515	0.0137362	-0.49760051	-1.02708992
9	PHE 434	**-1.40323329**	-0.15119405	0.3094885	**-1.00170202**	**-2.24664086**
10	LEU 437	-0.50779476	-0.06690343	0.13597925	-0.52301499	-0.96173392

Table 8.3 Per residue free energy (kcal/mol) decomposition for binding pocket 3.Three Highest VdW contributors (red) and Three Highest NPS (green) highlighted

BP3 – Res Decomposition	Residue	VdW	V_{EL}	Polar Solvation	Non-polar solvation	TOT-Average
0	SER 183	-0.30643951	-0.05203244	0.03083254	-0.2412944	-0.56893381
1	THR 191	-0.7280037	-0.26919478	0.11383534	-0.71735721	-1.60072035
2	LEU 194	-0.91603023	-0.05261253	0.17289437	-0.67364059	-1.46938898
3	PHE 242	-0.9599987	-0.32198686	0.4645749	-0.63853615	-1.45594681
4	LEU 245	-0.96218383	0.23140211	-0.00935792	-0.7203274	-1.46046704
5	CYS 246	-0.66181724	0.01376475	0.00360789	-0.53222461	-1.17666922
6	TRP 249	-0.00403688	0.00301404	0.00402964	0	0.0030068
7	**VAL 280**	**-1.14425823**	-0.0822427	0.08900345	**-0.91880891**	**-2.05630639**
8	**ILE 430**	**-1.88006202**	-0.28718763	0.37869546	**-1.2806674**	**-3.06922159**
9	**PHE 434**	**-1.57087869**	0.04407102	0.12342283	**-1.19134398**	**-2.59472883**

Table 8.4 Per residue free energy (kcal/mol) decomposition for binding pocket 3.Three Highest VdW contributors (red) and Three Highest NPS (green) highlighted

BP4 – Res Decomposition	Residue	VdW	V_{EL}	Polar Solvation	Non-polar solvation	TOT-Average
0	ASN 158	-0.34865576	-0.0382706	-0.06152889	-0.26577931	-0.71423456
1	LEU 161	-0.70108115	-0.06146728	0.14753583	-0.5765206	-1.19153319
2	LEU 281	-0.66077204	-0.00046006	0.02835074	-0.63088852	-1.26376987
3	**PHE 284**	-1.07673944	-0.02229995	0.12303321	-0.8074136	-1.78341977
4	**LEU 285**	**-1.65531511**	-0.27688431	0.37947781	-1.13926545	**-2.69198706**
5	LEU 288	-1.05267884	0.02577602	0.15586165	-0.77007321	-1.64111437
6	LEU 289	-0.60682013	0.02554241	0.07652928	-0.5872986	-1.09204705
7	ILE 441	-0.77192347	-0.04543476	0.17271456	-0.52484678	-1.16949045
8	**ILE 445**	**-1.06445205**	-0.15413443	0.2584417	-0.91952966	**-1.87967444**
9	TYR 578	-0.86761609	-0.03093078	-0.05910012	-0.67670667	-1.63435366
10	**ILE 579**	**-1.31603776**	-0.19342187	0.29502866	-0.86508429	**-2.07951526**
11	**ILE 582**	**-1.59331299**	-0.07991827	0.19565461	-1.21070652	**-2.68828318**
12	LEU 583	-0.56463927	0.05255342	0.06894681	-0.55828018	-1.00141922
13	PHE 586	-0.48179138	-0.0546554	0.1354989	-0.34978059	-0.75072846