A Computational Study on GPCR Activation Mechanisms: Insights from Adrenaline Binding and G-Protein Dissociation

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Natural Sciences By Research

by

Keshavan Seshadri
20171051
keshavan.seshadri@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500032, India
August 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**A Computational Study on GPCR Activation Mechanisms: Insights from Adrenaline Binding and G-Protein Dissociation**" by **Keshavan Seshadri**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. Marimuthu Krishnan

To my parents, my brother, and a brighter future

Acknowledgments

I want to extend my sincere gratitude and appreciation to the following individuals who have played a significant role in the completion of this thesis, titled "**A Computational Study on GPCR Activation Mechanisms: Insights from Adrenaline Binding and G-Protein Dissociation**."

First and foremost, I am deeply thankful to my advisor, **Prof. Marimuthu Krishnan**. It has been an incredible honor to be his Master's student. His dedication, guidance, and unwavering support have been invaluable throughout my research. I am grateful for the time, ideas, and knowledge he invested in me, which have made this experience productive and enriching. May this thesis stand as evidence of the tremendous patience and steadfast faith he has bestowed upon me! Thank you, Prof. Krishnan, for being an exemplary mentor and showing me the true essence of conducting research.

I am eternally grateful to my parents, Latha Narayanan and Seshadri Raman, who have been my constant pillars of strength. Your unwavering belief in me, unconditional love, and unwavering support have driven my accomplishments. I owe my determination and resilience to your unwavering encouragement. I am also grateful to my brother Madhavan Seshadri for constantly pushing me to pursue my dreams and providing me with the opportunities and resources to succeed.

Furthermore, I extend gratitude to my friends and colleagues whose camaraderie, intellectual discussions, friendship, love, and support have been invaluable. Their presence has provided me with the necessary motivation, laughter, and shared experiences, making this academic endeavor 10x enjoyable and fulfilling.

I am also indebted to the faculty and staff at IIIT Hyderabad for creating a research-centric environment. Their expertise, guidance, and commitment to research and academic excellence have shaped my knowledge in both *Computer Sciences* and *Computational Natural Sciences*. I also thank the HPC center at IIIT Hyderabad for providing the computational resources necessary for my research.

Finally, I want to acknowledge all the scholars, researchers, and pioneers whose work has laid the foundation for my research and whose dedication to pursuing knowledge has inspired me and broadened my horizons. I am profoundly humbled and enthusiastic to be a part of the scientific community as I hope to contribute to the ongoing quest for scientific understanding and discovery, pushing the boundaries of science! In conclusion, I sincerely appreciate everyone who has supported me on this academic and research journey. Your belief in me, guidance, and encouragement have been invaluable. This thesis is a testament to our collective efforts and the transformative power of collaboration and mentorship.

Abstract

GPCRs are the most prominent family of membrane proteins that serve as major targets for one-third of the drugs produced. A detailed understanding of the molecular mechanism of drug-induced activation and inhibition of GPCRs is crucial for the rational design of novel therapeutics. The binding of the neurotransmitter adrenaline to the β_2-adrenergic receptor (β_2AR) is known to induce a *flight or fight* cellular response, but much remains to be understood about binding-induced dynamical changes in β_2AR and adrenaline. In Chapter 3, we examine the potential of mean force (PMF) for the unbinding of adrenaline from the orthosteric binding site of β_2AR and the associated dynamics using umbrella sampling and molecular dynamics (MD) simulations. The calculated PMF reveals a global energy minimum, which corresponds to the crystal structure of β_2AR-adrenaline complex, and a meta-stable state in which the adrenaline is moved slightly deeper into the binding pocket with a different orientation compared to that in the crystal structure. The orientational and conformational changes in adrenaline during the transition between these two states and the underlying driving forces of this transition are also explored. Based on clustering of MD configurations and machine learning-based statistical analyses of time series of relevant collective variables, the structures and stabilizing interactions of these two states of the β_2AR-adrenaline complex are also investigated. In Chapter 4, the PMF of G-Protein dissociation, and GDP dissociation are studied. In order to study the entire GPCR activation pathway, models have been constructed for 3 different states (inactive, intermediate, and active states) of the G-Protein-β_2AR complex. The results suggest that the binding of GDP to G-Protein favours G-Protein dissociation from β_2AR. In summary, the findings enhance our understanding of GPCR activation and the advancement of new therapies aimed at GPCRs by offering comprehensive insights into conformational changes, energetics, and critical residues involved. The methodologies and discoveries presented in this study establish the groundwork for an automated process to explore ligand dynamics and interactions between drugs and GPCRs, thereby facilitating future research endeavors in this domain.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Proteins

Proteins are macro-molecules present in all living organisms and comprise one or more amino acids [1]. They play an essential role in sustaining life, affecting nearly every biochemical process [1, 2]. They have a wide range of functions as they serve as critical biomolecules, including enzymes, hormones, receptors, antibodies, transporters, and catalysts for various reactions occurring within organisms [1, 2]. A protein's nature or distinct function is primarily determined by its amino acid sequence, structure, three-dimensional shape, conformation, and associated interactions within itself and the surrounding environment [1]. Proteins also belong to one of three primary macronutrients required for the human body: carbohydrates, fats, and proteins [3]. These macronutrients provide the body energy and play a vital role in its structure and functioning.

Proteins in the human body are built by 20 different amino acids, of which nine amino acids are considered essential and cannot be produced by the human body [1, 2]. These essential amino acids have to be consumed externally through food and comprise the following amino acids, histidine (HIS), isoleucine (ILE), leucine (LEU), lysine (LYS), methionine (MET), phenylalanine (PHE), threonine (THR), tryptophan (TRP), and valine (VAL). The non-essential amino acids comprise the other 11 amino acids that can be produced by the human body, namely alanine (ALA), arginine (ARG), asparagine (ASN), aspartic acid (ASP),

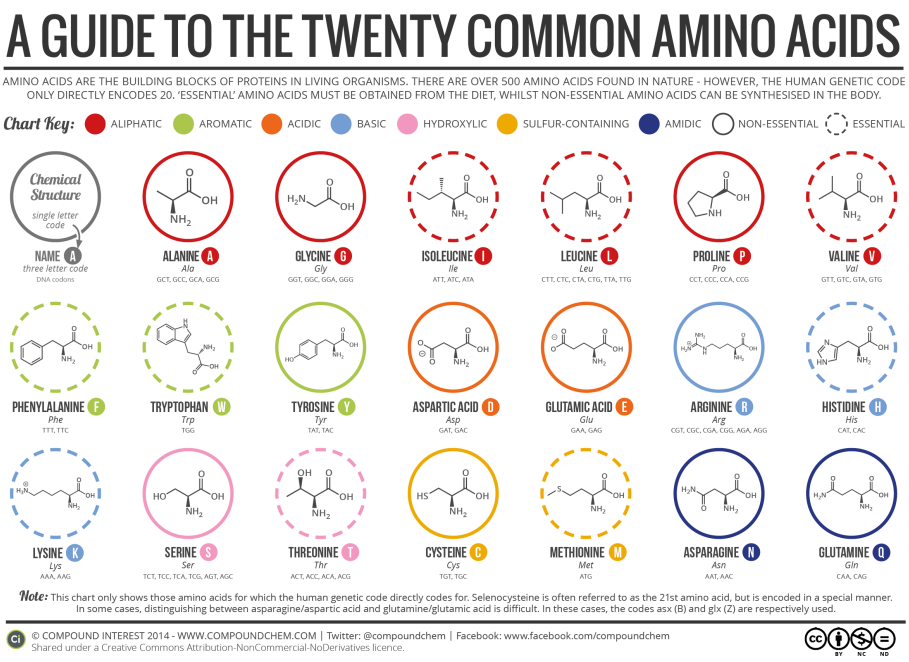


Figure 1.1: Structures of 20 Common Amino Acids. Figure adapted from `https://www.compoundchem.com/2014/09/16/aminoacids/`

cysteine (CYS), glutamic acid (GLU), glutamine (GLN), glycine (GLY), proline (PRO), serine (SER), and tyrosine (TYR). [1, 2]

1.1.1 Central Dogma of Molecular Biology

The genetic code comprised of DNA (double-strand) or RNA (single-strand) consists of the information required to build a protein from the amino acids [1, 2]. The "central dogma of molecular biology" (Figure 1.2) encompasses the processes involving gene expression, the information flow and the formation of proteins inside the cell [4]. It states that the information flows from nucleic acid to nucleic acid, and from nucleic acid to protein. The information cannot be transferred from protein back to either protein or a nucleic acid [4].

1.1.1.1 DNA Replication

During DNA replication, each strand of a DNA molecule is used as a template to create two new DNA molecules, also known as semi-conservative replication [2]. (refer Figure 1.3) This process primarily occurs during cell division, to ensure that each cell receives a copy of the original DNA [2].

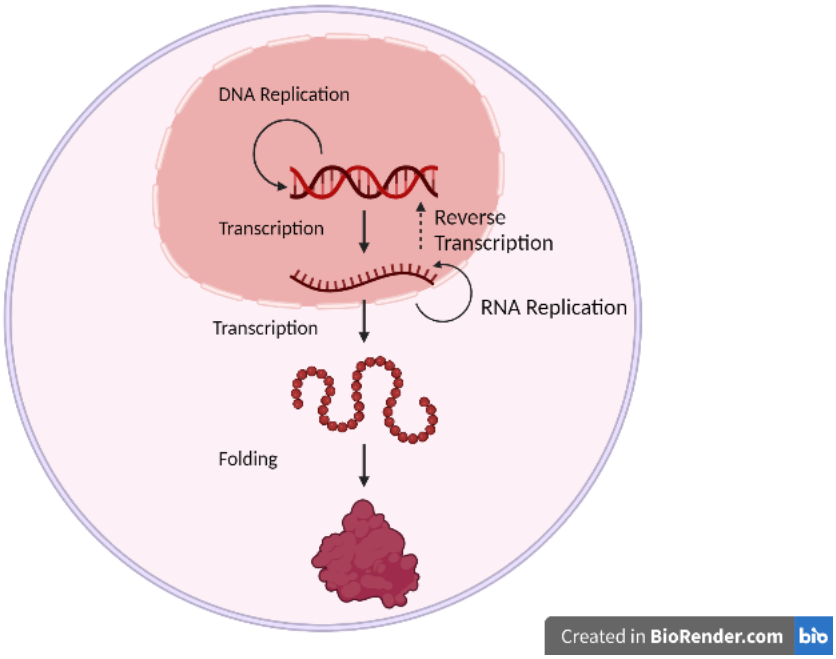


Figure 1.2: Graphical illustration of gene expression and information flow in cells. It also known as "Central Dogma of Molecular Biology" as coined by Francis Crick. [4] Created with `BioRender.com`

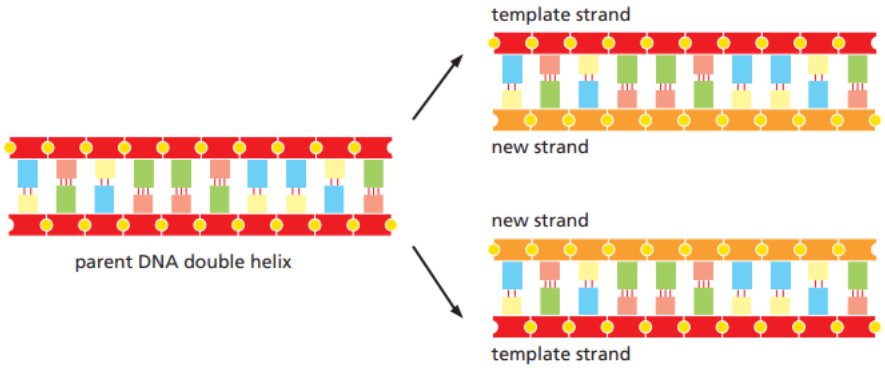


Figure 1.3: In this process, the two strands of a DNA double helix are pulled apart, and each serves as a template for synthesis of a new complementary strand. Figure and Caption adapted from Alberts et al. [2]

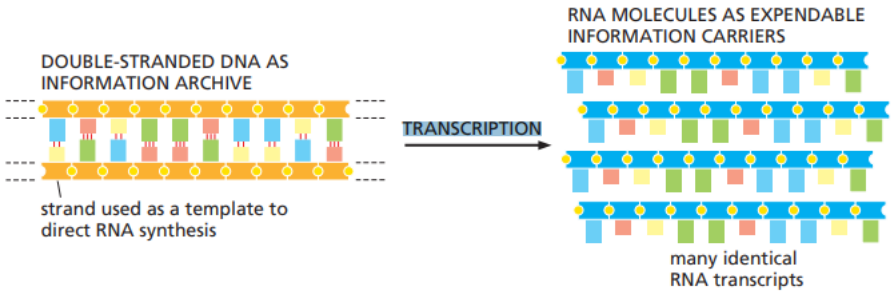


Figure 1.4: A given segment of this DNA guides the synthesis of many identical RNA transcripts, which serve as working copies of the information stored in the archive. Many different sets of RNA molecules can be made by transcribing different parts of a cell's DNA sequences, allowing different types of cells to use the same information store differently. Figure and Caption adapted from [2]

1.1.1.2 Transcription

The process of copying a segment of DNA to RNA is called transcription [2]. (refer Figure 1.4) The hydrogen bonds between the two strands of the DNA double helix are broken, and used as a template to create RNA [2]. The RNA which encodes for the creation of proteins is called messenger RNA (mRNA) [2]. The mRNA generated leave the nucleus through the nuclear pore [2].

1.1.1.3 Translation

Translation is the process of formation of proteins from the mRNA molecule. This process occurs either in the ribosomes or the endoplasmic reticulum. The four different nucleotides present in the mRNA molecule are read in the groups of 3, totaling 4 x 4 x 4 = 64 triplet combinations [2]. Each of these triplet combination is called a codon. However, these codons code for only 20 different amino acids, since multiple codons code for the same amino acid as illustrated in Figure 1.1 [2].

1.1.2 Structure

On the basis of structure proteins can be classified into primary, secondary, tertiary, and quaternary [1, 2]. The structure of a protein is primarily determined by its specific amino acid sequence. The amino-acids form covalent bonds with each other at their C and N terminus. This bond is more commonly referred as a peptide bond [2]. Upon the formation of multiple such peptide bonds, a long unbranched chain of amino acids is formed, which is known as a

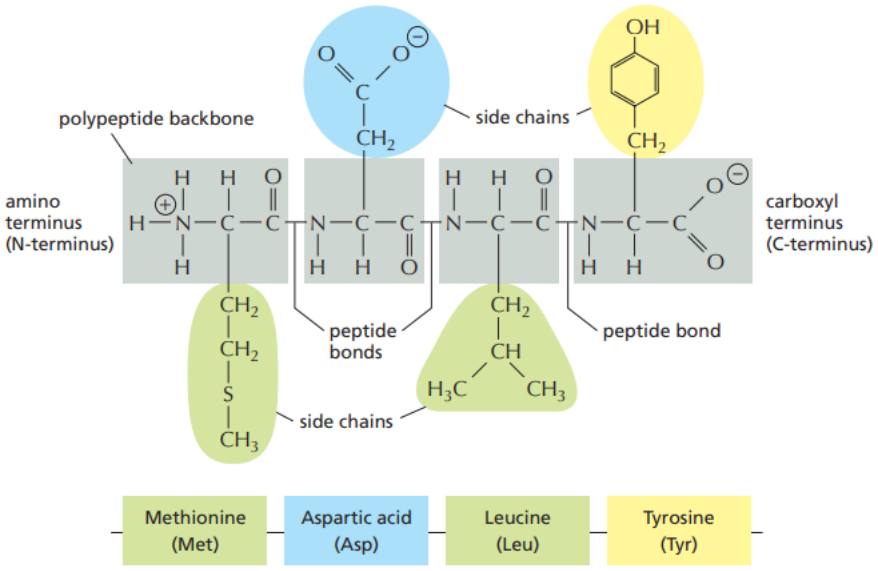


Figure 1.5: Components of a protein. Figure and caption adapted from [2]

polypeptide. (refer Figure 1.5) This polypeptide obtained after translation is a an unfolded conformation of protein, also referred to as the primary structure [2].

Other than peptide bonds, non-covalent interactions also contribute to the determination of the protein structure. Hydrogen bonds, Polar (electrostatic), van der Waals, aromatic stacking, disulphide bridges between Cysteine molecules, hydrophobic interactions are some of the common forces involved in protein structure and conformation determination. These forces contribute to the folding of the protein to make the protein biologically functional. The folding of a protein or a polypeptide primarily occurs in the cytoplasm. Special proteins called chaperones accelerate the protein folding process [2].

1.1.3 Classification

Proteins can be classified primarily based on their function, structure, activity and location within the cell [1, 2]. Each polypeptide chain consists of a N-terminus and a C-terminus [1, 2]. The secondary structure comprises a locally folded structure within a polypeptide. The structure is usually a α-helix or β-sheet (stabilized primarily through the formation of hydrogen bonds), while other structures, including the 3-10 helix and π helix, are rarely observed in naturally formed proteins. The tertiary structure is a folded polypeptide structure formed by multiple secondary structures. The quaternary structure is formed by two or more polypeptide chains with spatially arranged tertiary structures [1, 2]. On the basis of their location within the cell, proteins can be classified as intracellular, extracellular or membrane proteins. The

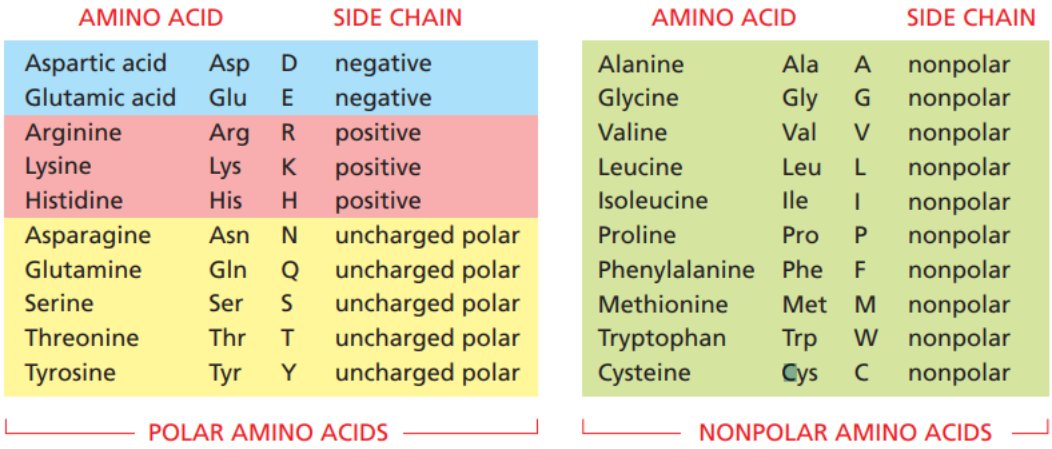


Figure 1.6: The 20 common amino acids along with their 3-letter and 1-letter abbreviation. There are 10 polar and 10 non-polar amino-acids. Figure adapted from [2]

membrane proteins can be either be permanently attached (integral) or temporarily attached (peripheral) to the cell membrane. The permanently- attached integral membrane proteins can further be classified based into (1) integral polytopic proteins (transmembrane proteins), and (2) integral monotropic proteins. Similarly on the basis of their activity they can be classified as active, inactive or partially active.

1.1.4 Receptors

Receptors are bio-chemical structures made up of proteins which receives and converts signals [2, 5]. The signaling molecule (ligand) binds to the receptor protein and produces a physiological changes within the cell. Based on location the receptors can be classified into 2 categories:

1. Trans-membrane Receptor: They are also known as cell-surface receptors, and are embedded in the plasma membrane. The G-Protein Coupled Receptors (GPCR), ligand-gated ion channels, and hormone receptors [5].

2. Intracellular Receptor: They are present inside the cell. They include both the cytoplasmic receptors and the nuclear receptors [5].

The mechanism of signaling through the binding of extracellular signaling molecules to aforementioned trans-membrane receptor and the intracellular receptor is illustrated in Figure 1.8.

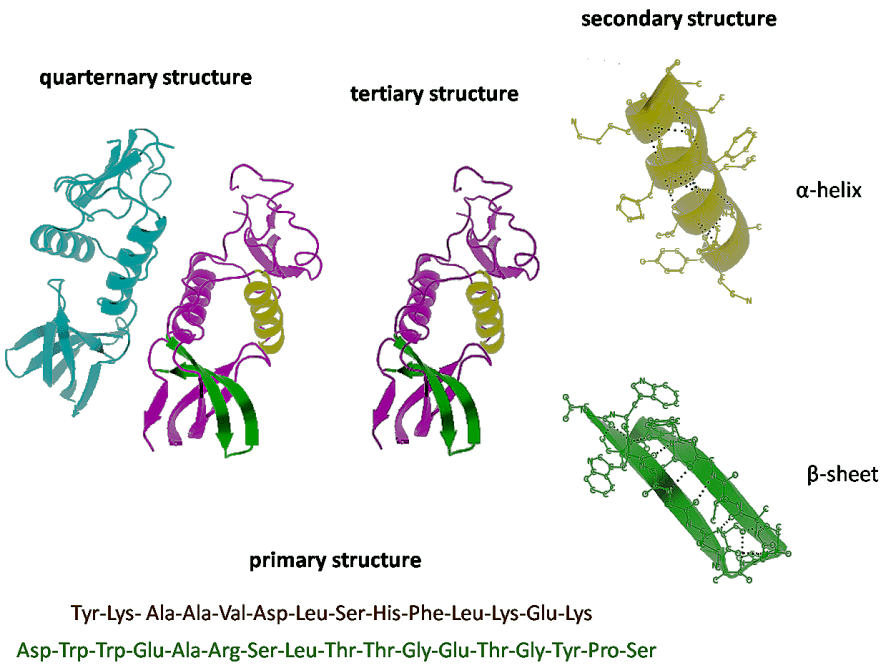


Figure 1.7: Protein Structure Classification. Figure adapted from `https://en.wikipedia.org/wiki/Protein_structure`

1.2 G-Protein Coupled Receptor

The G-protein coupled receptor (GPCR), as its name implies, is a receptor intricately linked to an intracellular heterotrimeric guanosine-nucleotide binding protein, commonly referred to as G-protein [1]. These receptors are distinguished by their seven alpha-helical transmembrane domains, hence known as 7-TM receptors. GPCRs are involved in mediating cellular responses to a diverse range of stimuli, including hormones, neurotransmitters, odorants, tastants, and other ligands, thereby regulating various physiological functions such as vision, smell, taste, homeostasis, behavior, and mood [6–8].

With around 850 GPCRs encoded in the human genome, GPCRs constitute the largest superfamily of cell surface receptors. Among these 850 GPCRs, ~ 350 GPCRs detect hormones, growth factors, and other endogenous ligands, while ~ 500 GPCRs function as olfactory (smell) and gustatory (taste) receptors [1]. The significance of GPCRs in human health becomes evident while considering their involvement in myriad medical conditions and diseases such as allergies, depression, blindness, diabetes, and cardiovascular defects. Furthermore, mutations in GPCRs have been found in ~ 20% of all cancers, cementing their essence in pathology [1]. Unsurprisingly, GPCRs serve as targets for approximately one-third of all manufactured drugs

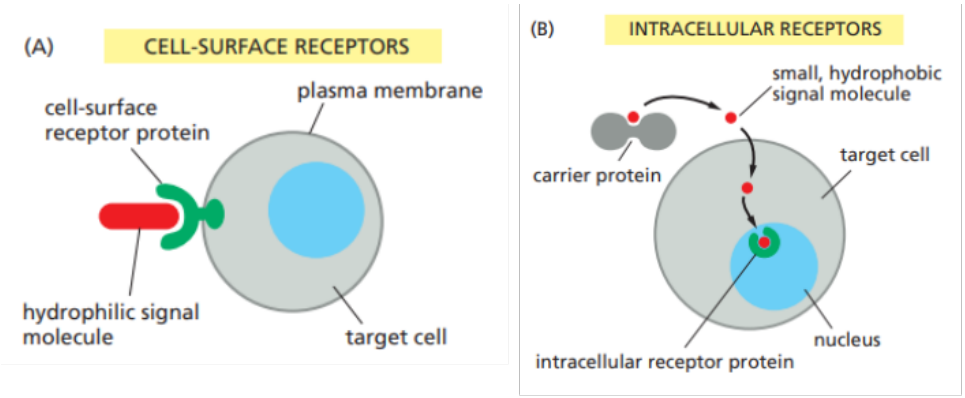


Figure 1.8: Simple mechanism of binding of extracellular signaling molecules to A) Transmembrane Receptor B) Intracellular Receptor. Figure adapted from [2]

[9–11]. Studying GPCRs and their associated mechanisms is paramount, given this substantial therapeutic relevance. The 2012 Nobel Prize in Chemistry was granted to Brian Kobilka and Robert Lefkowitz for their contributions to unraveling the structural and functional aspects of GPCRs, providing insights into their activation and signaling processes [12]. GPCRs remain a focal point in research, with eight other Nobel Prizes awarded for research on GPCR and the associated G protein signaling.

A signaling molecule or a ligand acts as the first chemical messenger that initiates the activation of a G-protein coupled receptor (GPCR) [1]. Upon ligand binding to the GPCR, a cascade of intricate conformational changes is triggered within the GPCR. These conformational changes also compel an allosteric transition to facilitate an interaction between the GPCR and the G-protein [1]. Moreover, the GPCR activation also activates the G-Protein, which in turn promotes the exchange of GDP (guanosine diphosphate) bound to the G-protein with GTP (guanosine triphosphate) [1]. Following this exchange, the G-protein dissociates from the GPCR, leading to the detachment of the $G\alpha$ subunit, which remains bound to GTP, from the $G\beta\gamma$ dimer. Subsequently, the detached $G\alpha$ and $G\beta\gamma$ sub-units interact with effector enzymes/proteins, thereby modulating their activity [1]. This modulation instigates alterations in the cytosolic concentration of small ions or molecules, including the $3', 5' - cyclic$ cyclic adenosine monophosphate (cAMP) or Calcium (Ca^{2+}) ions. These molecules function as secondary messengers, consequently activating or inhibiting various targets, such as protein kinases, and initiating further downstream signaling cascades [1].

1.2.0.1 Classification

The G-Protein Coupled Receptors (GPCRs) can be classified into distinct families based on their structural and functional characteristics. A commonly used classification system is the A-

F classification system which divides GPCRs into classes A, B, C, D, E, and F. However, a more recently developed classification uses the GRAFS system denoting Glutamate (C), Rhodopsin (A), Adhesion (B2), Frizzled/Taste2 (F), Secretin (B1) [13]. The human GPCR classification and the crystal structures found as of 2015 is illustrated in the Figure 1.9.

Class [1]	GRAFS [2] Family	Note	# Human receptors [3]	Crystal Structures (7TM domain)
A	Rhodopsin	Includes ~400 olfactory receptors [4]	282 (+ 400 olfactory)	2000: rhodopsin [5], 2007: β_2-adrenoceptor; and today: 23 receptors in total [6]
B	Secretin	Also referred to as B1	15	2013: Corticotropin-releasing factor receptor 1 [7] and Glucagon receptor [8]
	Adhesion	Also referred to as B2	33	–
C	Glutamate	Includes Taste type 1 receptors	22	2014: Metabotropic glutamate receptor 1 [76] and 5 [77]
F	Frizzled		11	2013: Smoothened [78]
O (Other)	Taste2	Re-classified as own family [79], although originally grouped with Frizzled [2]	24	–

Figure 1.9: Human GPCR classification and crystal structures (as of 2015). Classes D and E do not exist in human, and are fungal mating pheromone receptors and cAMP receptors, respectively. Figure and caption adapted from [14]

Each class exhibits unique structural features and signaling mechanisms, contributing to the functional diversity of GPCRs across various physiological processes. The roles of each class in classes A-F and class O/T:

1. Class A, also known as rhodopsin-like receptors, represents the largest group (85% of all GPCRs) and includes receptors for neurotransmitters, hormones, and sensory stimuli.

2. Class B receptors are involved in peptide hormone signaling important for vasodilation, Ca^{2+} homeostasis and blood glucose regulation [15, 16].

3. Class C receptors are characterized by their role in sensing amino acids and metabotropic glutamate receptors.

4. Class D receptors consist of fungal mating pheromone receptors.

5. Class E receptors consist of cyclic AMP receptors present in Dictyostelium discoideum (commonly referred as slime molds) which control its development [17–19].

6. Class F consists of the Frizzled GPCRs associate with the Wnt signaling pathway.

7. Class O/T originally grouped with Frizzled consists of Taste2 receptors associated with Taste sensation.

1.2.0.2 Numbering System

The most commonly used numbering systems [14] for GPCR classes are as follows: class A utilizes the Ballesteros-Weinstein (BW) [20] numbering system, class B uses the Wootten [21] numbering system, class C employs the Pin [22] numbering system, and class F utilizes the Wang [23] numbering system. These systems are all based on the Ballesteros-Weinsteing system which primarily has a number "$a.b$" for each residue in the TM helices, wherein a corresponds to the trans-membrane helix number ranging from 1-7. The number b denotes the residue position relative to the most conserved position at number 50. Alternative numbering systems proposed for class A were given by Oliveira [24] and Baldwin/Schwartz [25, 26].

1.2.0.3 β_2-Adrenergic Receptor

The β_2-Adrenergic Receptor (β_2AR) is a class A GPCR whose natural agonists are the hormones and neurotansmiters, adrenaline and noradrenaline. It plays a crucial role in promoting relaxation of smooth muscles found in the gastrointestinal tract, bronchi, urinary bladder, uterus, and seminal tract. Additionally, it facilitates improved blood flow and vasodilation in arteries while enhancing muscle contraction and motility in striated muscles. Moreover, it is responsible for regulating the secretion of insulin and glucagon in the pancreas.

Upon binding of adrenaline, the β_2AR is activated and the allosteric changes in the receptor-Gs complex allows the GDP to be exchanged with GTP. The activated $Gs\alpha$ is separated from the $Gs\beta\gamma$ dimer and binds to the adenylyl cyclase enzyme. The adenylyl cyclase catalyzes the formation of adenosine 3', 5' cyclic monophosphate (cAMP) and pyrophosphate from adenosine triphosphate (ATP). (refer Figure 1.10) The cAMP is a secondary messenger whose increase in concentration may lead to the activation of protein kinase A [27], Popdc (popeye domain containing proteins) [28], exchange proteins like EPAC [29], and cyclic nucleotide-gated ion channels [30]. PKA i a cAMP-dependen protein, when activate PKA phosphorylates eznymes promoting muscle contraction in the heart, enzymes catalyzing glycogen to glucose conversion, and transcription factors [31].

The agonists binding to the β_2AR can be classified as

(i) SABA (Short-acting β_2 agonists): Spasmolytics used for the treatment of asthma and COPD (salbutamol, procaterol, isoprenaline) and tocolytic agents used to suppress premature labour (fenoterol, salbutamol, isoxsuprine).

(ii) LABA (Long-acting β_2 agonists): Spasmolytics used for the treatment of asthma and COPD (salmeterol, formoterol, clenbuterol)

(iii) ultra-LABA: Spasmolytics used for the treatment of asthma and COPD (carmoterol, indacaterol, olodaterol)

Antagonists include beta-blockers (butoxamine, propranolol, ICI-118,551) which are used in the management of abnormal heart rhythms, or prevention of immediate second heart attack after an initial heart attack.

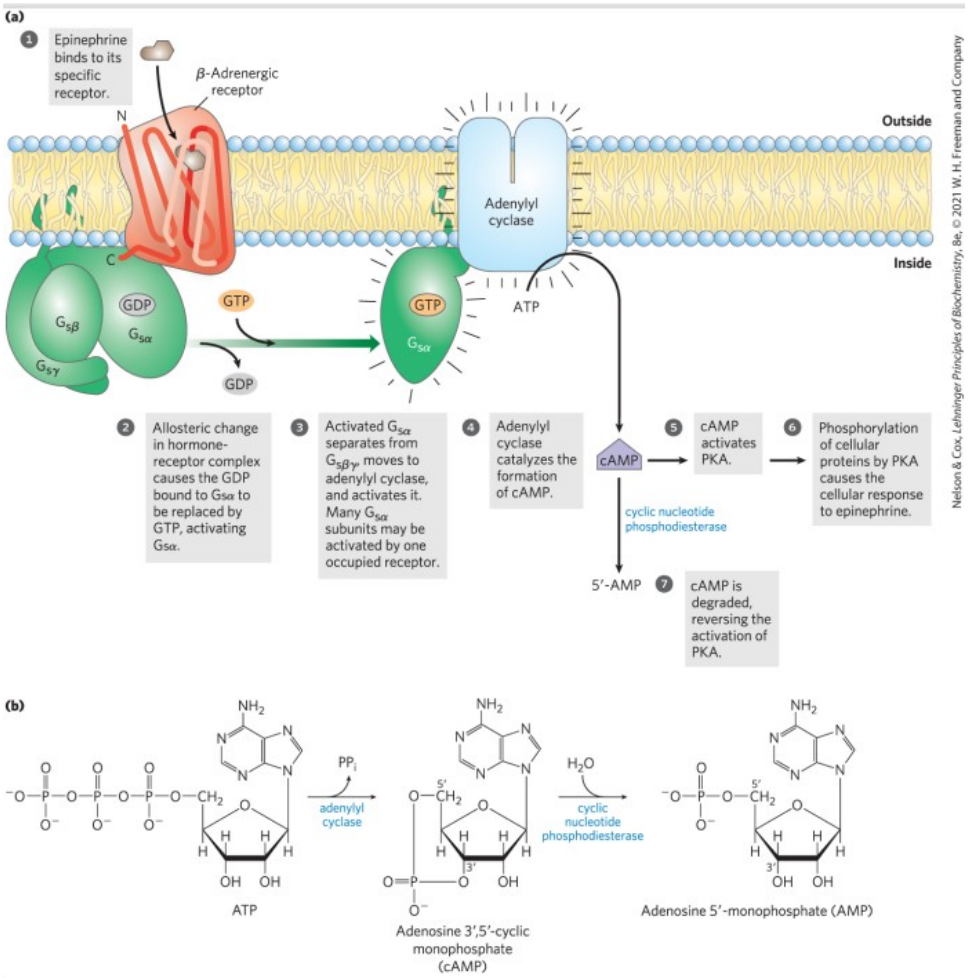


Figure 1.10: Transduction of the epinephrine signal: the β-adrenergic pathway. (a) The mechanism that couples binding of epinephrine to its receptor with activation of adenylyl cyclase; the seven steps are discussed in the text. The same adenylyl cyclase molecule in the plasma membrane may be regulated by a stimulatory G protein (Gs), as shown, or by an inhibitory G protein (Gi, not shown). Gs and Gi are under the influence of different hormones. Hormones that induce GTP binding to Gi cause inhibition of adenylyl cyclase, resulting in lower cellular [cAMP]. (b) The combined action of the enzymes that catalyze steps and , synthesis and hydrolysis of cAMP by adenylyl cyclase and cAMP phosphodiesterase, respectively. Figure and caption adapted from [1]

1.3 G-Protein

Guanine nucleotide-binding proteins (G-Proteins), are molecular switches present within the cells, and play an important role in signal transduction through the GPCRs. As the name suggests, G-Proteins are bound to either the guanosine triphosphate (GTP) or the guanosine diphosphate (GDP). The G-Proteins are considered active when the GTP is bound, and are considered inactive when GDP is bound. G-proteins also belong to the broader group of enzymes known as GTPases.

G-proteins can be divided into 2 classes:

(i) small G-proteins: Consists of monomeric Small GTPases, which are hydrolyzes GTP molecule. They are homologous to the $G\alpha$ sub-unit of a heterotrimeric G-Protein, but can function individually. They contain the Ras superfamily which further contains the 7 sub-families: Ras, Ral, Rit, Rap, Rheb, Rad, and Miro. Each sub-family has specific funtions, cell proliferation (Ras), cell morphology (Rho), nuclear transport (Ran) and vesicle transport (Rab).

(ii) Heterotrimeric G proteins: Coupled to the GPCRs, they are composed of three subunits: $G\alpha$, $G\beta$, and $G\gamma$. Upon activation of the GPCR, the G_α sub-unit binds and hydrolyzes GTP, while the $G\beta$ and $G\gamma$ sub-units remain associated as a stable dimeric complex. The G_α and the $G\beta\gamma$ dimer bind to effector proteins/enzymes and regulate their activity. During termination of the transduced signal, special proteins called RGS (Regulator of G protein signalling) proteins help in the hydrolysis of the GTP to GDP, which in turn helps in ther reunification of $G\alpha$ with the $G\beta\gamma$ dimer. There are four significant G-protein families: Gi/Go, Gq, Gs, and G12/13 [32]. The roles of some of the heterotrimeric G-Protein families are illustrated in Figure 1.11.

S.No	G-Protein Family	Role
1.	Gs	Stimulates adenylate cyclase through Beta-adrenergic receptors. It increases the calcium conductance in the Heart.
2.	Gi	Has multiple forms of inhibiting the adenylate cyclase through Alpha2-adrenergic receptors.
3.	Gt	It is also known as Transducin. It binds to the photoreceptor(rhodopsin) in the retina.
4.	Go	It can regulate the calcium channels.
5.	Gk	It can regulate the potassium channels.
6.	Gp	It can regulate phospholipases.

Figure 1.11: The role of different G-Protein families.

1.4 GPCR Specific Ligands

The signaling molecule or ligand that binds to the GPCR can be classified into Agonists, Antagonists, Inverse Agonists, Allosteric Modulator or Biased ligand. The role of each ligand is further discussed in Figure 1.12.

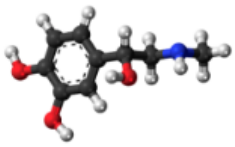
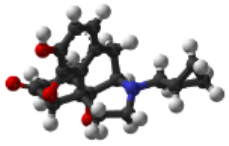
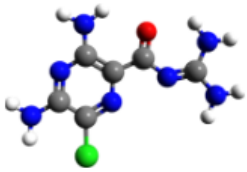
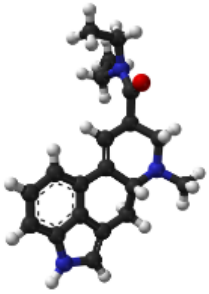
S.No	Type	Role	Structure
1	Agonists	A chemical that binds to the receptor to initiate a biological response(Activation). Ex: Epinephrine(Adrenaline)	
2	Antagonists	A chemical that blocks the receptor instead of activating it. Ex: Caffeine	
3	Inverse Agonists	A chemical that binds to a receptor and produces an opposite response compared to that of an agonist. Ex: Naltrexone	
4	Allosteric Modulator	A chemical that changes the response of a receptor to external stimulus by binding at the allosteric site. Ex: Amiloride	
5	Biased Ligand	The biased GPCR ligands preferentially activates either the G-protein or the β-arrestin mediated downstream signaling pathway. Ex: Lysergic acid diethylamide (LSD)	

Figure 1.12: The role of different ligands binding to GPCRs.

1.5 Research Focus

In Chapter 3, we model and study the dynamics of adrenaline in the orthosteric binding pocket of $\beta_2 AR$ using molecular dynamics, umbrella sampling, and machine learning. We obtain the PMF plot of adrenaline dissociation from the orthosteric binding pocket to study the energetics of the system. Using a clustering algorithm, we model the structure of these two sta-

ble states, a meta-stable and most-stable, and study the factors responsible for state transition. Furthermore, we identify the interacting residues, orientational and conformational dynamics of adrenaline in the binding pocket. To find the residues responsible for the inward motion of adrenaline, we calculated pairwise interaction energies and used statistical and machine learning-based regression methods.

In Chapter 4, we model the $\beta_2 AR$-Gs system and study both the dissociation of Gs upon the activation of $\beta_2 AR$ and the dissociation of GDP from $G\alpha$ using umbrella sampling and molecular dynamics. Furthermore, we create base models of adrenaline-$\beta_2 AR$-Gs-GDP/GTP in their inactive, intermediate, and active states, respectively.

In the next chapter, we discuss the computational methods and tools to model and study complex GPCR-G-protein systems. We also discuss key concepts, including molecular dynamics, umbrella sampling, and machine learning methods.

Chapter 2

Computational Methods

Contents

2.1 Overview

The use of computational methods to simulate multi-particle systems predates back to the 1950s. One of the first computational algorithms developed was the Monte Carlo algorithm, conceptualized in the eighteenth century uses randomness to solve deterministic problems. The development of such computer-based methods for studying the interactions between proteins and ligands has been a significant area of focus for computational biophysicists [33] and has been a continuing area of interest even in the 21st century, as evidenced by a review by Shoichet and Kobilka in 2011 [34]. In the 1970s and 1980s, molecular dynamics (MD) simulations materialized as an effective method for studying the motion and dynamics of proteins and other macromolecules [33]. MD simulations obey the laws of classical mechanics to solve the equations of motion for a system of interacting atoms/molecules. One of the earliest successful MD simulations for biophysical systems was in predicting the structure of the protein BPTI in 1977 [35]. The development of computational methods for studying protein-ligand interactions and discovering new drugs/ligands has continued to gain momentum since the 1990s. In particular, using virtual screening methods such as molecular docking for drug discovery became increasingly popular [36]. This method involves screening large databases of potential

drug molecules against a target protein to identify molecules that bind with high affinity. Using machine learning algorithms to perform big-data analysis in the 21st century has revolutionized computational biophysical research, enabling the exploration of complex biological systems at a remarkable level of detail [37, 38]. Today, computational biophysics has become an integral part of modern life sciences research, contributing significantly to the understanding and explaining of the molecular mechanisms behind various biological processes [39]. As supercomputers become faster every year and the field continues to evolve, computational methods will play an even more significant role in advancing our understanding of biophysical systems.

This chapter introduces key computational methods involved in the study, including Molecular Dynamics, Enhanced Sampling Methods, Statistical Methods and Machine Learning.

2.2 Molecular Dynamics Simulations

Molecular dynamics (MD) simulations are a powerful computational tool used to study the dynamics and behavior of atoms/molecules over time [40]. MD simulations use classical mechanics to simulate the motion of particles in a system, where a potential energy function is used to describe the forces between particles [40]. The potential energy function used in MD simulations is commonly represented by a force field, a mathematical model describing the interactions between atoms/molecules [40]. The force field is parameterized using experimental data or quantum mechanical calculations. The CHARMM, AMBER, and GROMOS force fields are the most commonly used force fields [40]. The equations of motion for MD simulations is based on the Newton's second law of motion, which states that the change of motion of an object is proportional to the force impressed; and is made in the direction of the straight line in which the force is impressed [40–42].

The modern form of Newton's second law states that the time derivative of the momentum (for constant mass m, becomes $m\boldsymbol{a}$) is the net force [42].

$$\boldsymbol{F} = m\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = m\boldsymbol{a} \tag{2.1}$$

where m is the mass of the object, $\boldsymbol{v}$ its velocity and $\boldsymbol{a}$ its acceleration. In an N particle system, molecular dynamics mandates that the net force be calculated on all the N particles. Therefore, for any given particle i, the equation becomes,

$$\boldsymbol{F_i} = m_i\boldsymbol{a_i} = m\dot{\boldsymbol{v}}_{\boldsymbol{i}} = m\ddot{\boldsymbol{r}} \tag{2.2}$$

where $\boldsymbol{F_i}$ is its net force vector, m_i is its mass, and $\boldsymbol{a_i}$ is its acceleration, $\boldsymbol{v}$ is its velocity and $\boldsymbol{r}$ its position.

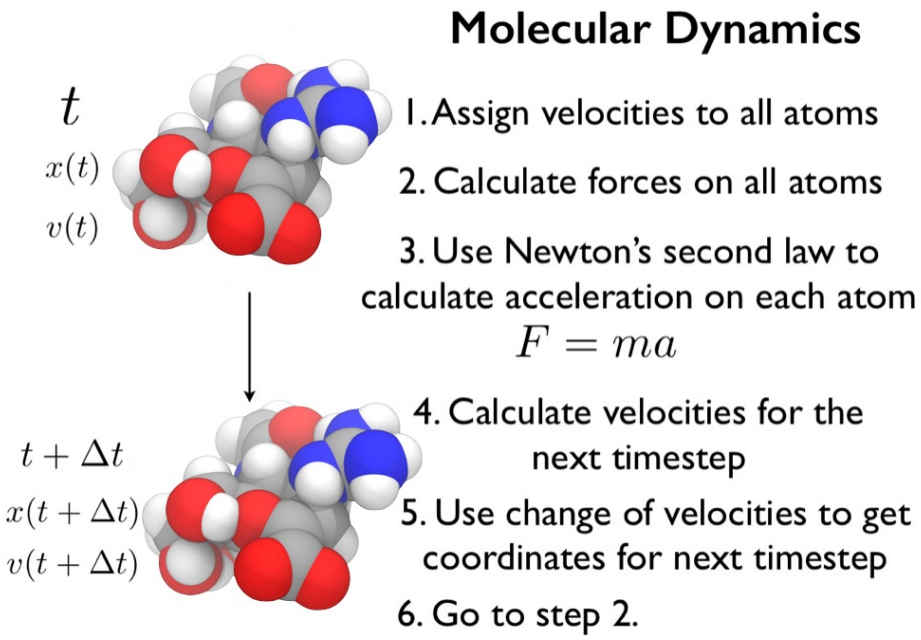


Figure 2.1: Steps involved in Molecular Dynamics Simulations. Figure adapted from `https://chryswoods.com/dynamics/dynamics/theory.html`

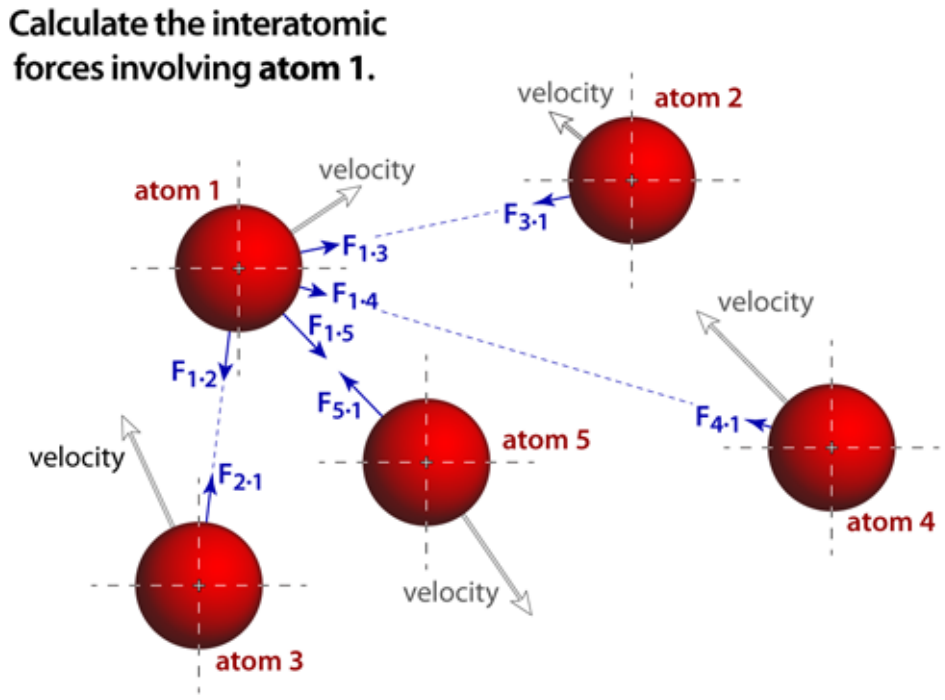


Figure 2.2: An illustration of the inter-atomic forces (only non-bonded interactions). Figure adapted from `http://www.atomsinmotion.com/book/chapter5/md`

To calculate the net force on each atom, all inter-atomic forces need to be summed. For example, in figure 2.2, the net force on atom 1, $\boldsymbol{F_1}$ is given by the equation 2.3.

$$\boldsymbol{F_1} = \boldsymbol{F_{12}} + \boldsymbol{F_{13}} + \boldsymbol{F_{14}} + \boldsymbol{F_{15}} \tag{2.3}$$

These inter-atomic forces can be written as a derivative of their inter-atomic potentials since $F(r) = -\frac{\mathrm{d}U}{\mathrm{d}r}$. A more generalized equation for the net force on a particle i, at a given time t, is given by the equation 2.4.

$$F_i(r) = -\Delta_i U_{total} \tag{2.4}$$

The number of times, each of these inter-atomic forces is computed depends on the Sum(N), where N is the total number of particles. Assuming each pair is computed, Sum(N)=N(N+1)/2. However, distance based pair-wise cutoffs between 2 particles are implemented using a *switching function* in molecular dynamics softwares to increase efficiency [40]. A pairwise cutoff of 10Å or 12Å is commonly used, since distances greater than the cutoff distance amount to negligible inter-atomic forces. The next sub-section comprises the computational determination of the inter-atomic potential, U_{total}, through force-fields.

2.2.1 Force Field

A force field which consists of the functional form and parameter sets of atoms/particles, provides a computational method to calculate the potential energy of the system and the force between any two atoms. These parameters are determined from physics and chemistry experiments, and quantum mechanics. The force fields describe the energy landscape. The force on a particle is derived from the gradient of the potential energy to the atom/particle's coordinates. The potential energy of the system can be represented as the sum of the "bonded" (covalent bond) terms, and the "non-bonded" (non-covalent interactions) terms. The *bonded* terms include the bonds, angles, and torsions (dihedrals + inversions). The *non-bonded* terms primarily include the electrostatic and van der Waals energy. Although additional terms could be added to the potential energy equation to account for improper torsions, cross terms, and explicit hydrogen bonds, a majority of the mainly used force fields (CHARMM, AMBER, GROMOS) have the terms mentioned in the equations 2.6 and 2.7 common [40].

$$U_{total} = U_{bonded} + U_{non-bonded} \tag{2.5}$$

$$U_{bonded} = U_{bond} + U_{angle} + U_{torsion} \tag{2.6}$$

$$U_{non-bonded} = U_{electrostatic} + U_{vdW} \tag{2.7}$$

The bonded terms bonds and angles can be approximated with a spring and therefore, the forces (for bond stretch, angle bend) can be written using the simple Hooke's law. The Hooke's law is given by the equation 2.8.

$$F_s = kx \tag{2.8}$$

where F_s is the force required to extend or compress the spring, x is the distance by which the spring is extended or compressed, and k is the spring constant. Now, the potential energy of the spring is given by the equation 2.9.

$$U_{spring} = \frac{1}{2}kx^2 \tag{2.9}$$

Similarly, the potential energy of the bonded terms can be written as follows:

$$U_{bond} = \sum_{All\ bonds} \frac{1}{2}k_b(b - b_o)^2 \tag{2.10}$$

$$U_{angle} = \sum_{All\ angles} \frac{1}{2}k_\theta(\theta - \theta_o)^2 \tag{2.11}$$

A majority of the force fields use the cosine series expansion to model the torsional term of the potential energy. They usually include both the dihedral angle between the planes defined by four atoms/particles and the out of plane inversions.

$$U_{torsion} = \sum_{All\ dihedrals} k_{dihedral}(1 - cos(n\omega - \gamma_{dihedral}) + \sum_{All\ inversions} k_{inversion}(1 - cos(n\omega - \gamma_{inversion}) \tag{2.12}$$

The electrostatic forces between any two atoms/particles is given by the Coulomb's law, equation 2.13. The electrostatic potential energy can be derived from the Coulomb's law is given in equation 2.14.

$$|F| = k_e \frac{|q_a||q_b|}{r_{ab}^2} \tag{2.13}$$

$$U_{electrostatic} = \sum_{coulomb} k_e \frac{|q_a||q_b|}{r_{ab}} \tag{2.14}$$

The van der Waal's interaction is usually modelled with the Lennard-Jones potential which is given in equation 2.15.

$$U_{vdW} = \sum_{vdW} 4\epsilon_{ab}\left[\left(\frac{\sigma_{ab}}{r_{ab}}\right)^{12} - \left(\frac{\sigma_{ab}}{r_{ab}}\right)^{6}\right] \tag{2.15}$$

Once the inter-atomic potentials are computed through force-fields, the inter-atomic forces can be determined (refer equation 2.4). The next step in molecular dynamics involves the computation of velocities and position, evolving through time with numerical integration.

2.2.2 Numerical Integration

In order to calculate the position and velocity of a particle at the next time step, a *Taylor series* expansion can be used. The velocity and position at next time-step (Δt) is given by the following equations 2.16 and 2.17.

$$v(t+\Delta t) = v(t) + a(t)\Delta t + \frac{1}{2}\frac{\mathrm{d}^2 v}{\mathrm{d}t^2}\Delta t^2 + \frac{1}{3!}\frac{\mathrm{d}^3 v}{\mathrm{d}t^3}\Delta t^3 + O(\Delta t^4) + \dots \tag{2.16}$$

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 + \frac{1}{3!}\frac{\mathrm{d}^3 r}{\mathrm{d}t^3}\Delta t^3 + O(\Delta t^4) + \dots \tag{2.17}$$

However, the taylor series expansion is computationally expensive to caculate. Therefore, numerical integration methods such as Verlet and Velocity Verlet algorithms are required.

2.2.3 Verlet

In the verlet algorithm, the *Taylor series* expansion of position is truncated to the first four terms. Similarly, the position backward in time is written in the following equation 2.18.

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 - \frac{1}{3!}\frac{\mathrm{d}^3 r}{\mathrm{d}t^3}\Delta t^3 + O(\Delta t^4) + \dots \tag{2.18}$$

Adding the first four terms in equation 2.17 and 2.18, gives, equation 2.19

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + a(t)\Delta t^2 + O(\Delta t^4) \tag{2.19}$$

In the Verlet algorithm, the velocities are not required to determine the next position. However, the calculation of velocity is required for calculating the kinetic energy. One can compute velocity using the mean value theorem and position, equation 2.20

$$v(t) = \frac{r(t+\Delta t) - r(t-\Delta t)}{2\Delta t} + O(\Delta t^2) \tag{2.20}$$

The above equation 2.20 compromises accuracy and has an error of $O(\Delta t^2)$ which is a lot worse than the four-term truncated Taylor series equation 2.16.

2.2.4 Velocity Verlet

The *Velocity Verlet* algorithm is a numerical integration method. It has a unique property that it calculates both the position and velocity at the same time. The following are the equations of the *Velocity Verlet algorithm*

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 \tag{2.21}$$

$$v(t+\Delta t) = v(t) + \frac{1}{2}[a(t+\Delta t) + a(t)]\Delta t + O(\Delta t^3) \tag{2.22}$$

It is possible to derive the *Velocity Verlet* algorithm from the basic verlet algorithm as follows. Ignoring higher-order error terms, from equation 2.20, we can write $r(t - \Delta t)$ as,

$$r(t - \Delta t) = r(t + \Delta t) - 2v(t)\Delta t \tag{2.23}$$

Now, we substitute, equation 2.23, in equation 2.19, we get the truncated *Taylor series* expansion of position which is used in the Velocity Verlet algorithm.

$$r(t + \Delta t) = 2r(t) - r(t + \Delta t) + 2v(t)\Delta t + a(t)\Delta t^2 \tag{2.24}$$

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 \tag{2.25}$$

To compute velocities, we write $v(t + \Delta t)$ from equation 2.20 as,

$$v(t + \Delta t) = \frac{r(t + 2\Delta t) - r(t)}{2\Delta t} \tag{2.26}$$

Similarly, we write $r(t + 2\Delta t)$, from equation 2.19 as,

$$r(t + 2\Delta t) = 2r(t + \Delta t) - r(t) + a(t + \Delta t)\Delta t^2 \tag{2.27}$$

Substituting equation 2.27 in equation 2.26, we get,

$$v(t + \Delta t) = \frac{r(t + \Delta t) - r(t)}{\Delta t} + \frac{1}{2}a(t + \Delta t)\Delta t \tag{2.28}$$

Now, we eliminate $r(t + \Delta t) - r(t)$ from equation 2.25, to get the velocity, as

$$v(t + \Delta t) = v(t) + \frac{1}{2}[a(t + \Delta t) + a(t)]\Delta t \tag{2.29}$$

Standard implementations of the velocity Verlet algorithm use a "half-timestep" velocity computation. The equations 2.25 and 2.29 can be modified as follows

$$r(t + \Delta t) = r(t) + v(t + \frac{1}{2}\Delta t)\Delta t \tag{2.30}$$

$$v(t + \Delta t) = v(t + \frac{1}{2}\Delta t) + \frac{1}{2}a(t + \Delta t)\Delta t \tag{2.31}$$

2.3 Energy Minimization

The spatial arrangement of atoms in the initial structure may not be energetically favorable. This issue can be addressed through energy minimization, a process that mitigates inter-atomic forces and guides the system towards a local (point A) or global minimum (point C) on the potential energy surface (PES). (refer Figure 2.3) When the system reaches its most stable state, it adopts a conformation corresponding to the global minimum on the PES. In the next sub-sections, we will discuss in brief the two significant energy minimization algorithms that have been employed in the thesis.

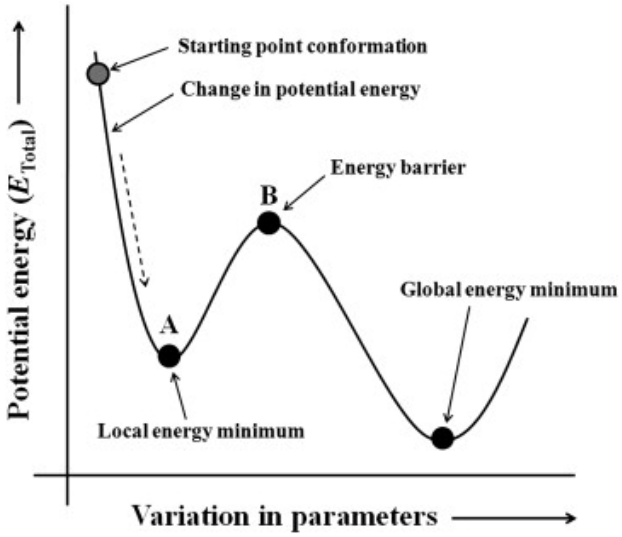


Figure 2.3: A figure depicting the different phases during energy minimization. Figure adapted from `https://lammpstube.com/2020/08/05/minimization-of-energy/`

2.3.1 Steepest Descent

The *steepest descent* algorithm is a simple and efficient algorithm primarily used in preliminary energy minimization. In each iteration, a consistent step (η) is taken in the direction opposite to the local gradient of the potential energy (U(r)). A new configuration (arrangement of atoms) denoted by r_{n+1} is generated from current configuration, r_n as follows:

$$r_{n+1} = r_n - \eta \Delta U(r_n) \tag{2.32}$$

2.3.2 Conjugate Gradient

The conjugate gradient method involves the selection or computation of successive orthogonal (conjugate) direction vectors. In each iteration, a step is taken along the chosen direction vector, and the step size for the next iteration is determined. Notably, the algorithm ensures that previously visited directions are not revisited in the future.

Compared to the steepest descent method, the conjugate gradient method may initially exhibit slower convergence during energy minimization. However, it becomes increasingly efficient as the system approaches the energy minimum. As a result, it is commonly employed after initially utilizing the steepest descent algorithm to accelerate the optimization process.

2.4 Ensembles

An ensemble is a collection or group of individual members, where each member represents a potential state or outcome of the actual system. In a majority of the real-world situations, randomness or uncertainty is present [43]. Therefore, rather than focusing on a single outcome or specific configuration, an ensemble takes into account all possible outcomes or configurations and assigns probabilities or weights to each of them [43].

In statistical mechanics, an ensemble represents a collection of systems with the same macroscopic conditions, such as number of particles (N), temperature (T), total energy (E), pressure (P), or volume (V). Each individual system within the ensemble corresponds to a different micro-configuration or micro-state of the particles in the system [43]. On studying the ensemble's properties and/or behavior as a whole, statistical mechanics provides insights into the average/expected behavior of the individual systems. (refer Figure 2.4) The four commonly used statistical ensembles are

1. Microcanonical Ensemble: Constant number of particles (N), volume (V), and total energy (E)

2. Canonical Ensemble: Constant number of particles (N), volume (V), and temperature (T). To keep the temperature in a canonical ensemble constant, a thermostat is often used. Various techniques to control temperature involve velocity-rescaling, Langevin thermostat, Berendensen thermostat, Noose-Hoover thermostat, and Andersen thermostat [43].

3. Grand Canonical Ensemble: Constant chemical potential (μ), volume (V), and temperature (T)

4. Isobaric-Isothermal Ensemble: Constant number of partices (N), Pressure (P), and Temperature (T). To maintain constant pressure, a barostat is often used. Techniques used to control temperature include volume-rescaling, Berendensen barostat, and Andersen barostat [43].

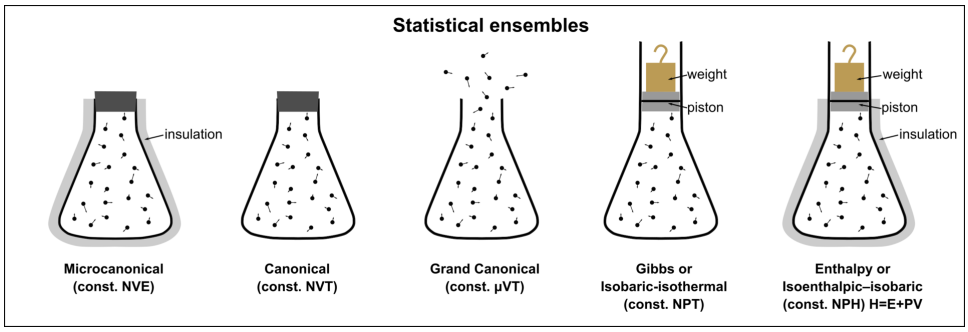


Figure 2.4: Commonly used statistical ensembles. Figure adapted from `https://en.wikipedia.org/wiki/Ensemble_%28mathematical_physics%29#/media/File:Statistical_Ensembles.png`

2.5 Free Energy

The term "free energy" was coined by *Hermann von Helmholtz* in the year 1882 [44]. The free energy presents the amount of energy available to do thermodynamic work at a constant temperature [44]. It is usually represented by the Helmholtz free energy, A, or the Gibbs free energy, G [40]. At constant volume and temperature (canonical ensemble) the free energy is given by the Helmholtz free energy equation 2.33.

$$A = U - TS = -\frac{1}{\beta} \ln Q_{NVT} \tag{2.33}$$

where A is the Helmholtz free energy, U is the potential energy, T is the temperature, S is the entropy of the system, and Q_{NVT} is the canonical partition function. Similarly, the free energy used under isobaric-isothermal (constant pressure (P) and temperature (T)) conditions is the given by the Gibbs Free Energy equation 2.34.

$$G = U - TS + PV = -\frac{1}{\beta} \ln Q_{NPT} \tag{2.34}$$

where G is the Gibbs free energy, U is the potential energy, T is the temperature, S is the enthalpy of the system, and Q_{NPT} is the isobaric-isothermal partition function.

The difference in free energy is essential for any chemical reaction to occur and is crucial to determine thermodynamic factors such as ligand binding rate constants [45]. However, the accurate determination of free energies or change in free energies through molecular dynamics or monte carlo simulations is not possible due to inadequate sampling of high energy regions [40, 45]. Therefore, the use of enhanced sampling techniques such as umbrella sampling is required to calculate the free energy profile of the system [45].

2.5.1 Collective Variable

A collective variable (CV) or a reaction coordinate, ε, is a continuous parameter used to distinguish two different thermodynamic states [45]. It can either be single or multi-dimensional. Broadly, there are two different types of collective variables commonly used:

1. Geometric CV: Distance, angle, dihedral, RMSD. They are also known as physics-intuitive/physics-based collective variable [46, 47]. Normally, around 10-100 physics-based collective variables are required to efficiently sample a biological process due to the number of conformation changes involved [46]. One can study only a limited number of conformational changes by using one CV [46, 47].

2. Abstract CV: A linear/non-linear combination of one or more geometric CVs. It is difficult to interpret the physical meaning of abstract CVs. Various dimensionality reduction methods, such as Principal Component Analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), local linear embedding (LLE), isomap, and diffusion map, are commonly employed to extract high-variance CVs from the data [46–48]. Likewise, for the exploration of slow CVs using lengthy unbiased MD trajectories, Time-lagged Independent Component Analysis (TICA), Variational Approach to Markov Processes (VAMP), Markov State Models (MSMs), and Extended Dynamical Mode Decomposition (EDMD) have been instrumental [46–48]. Deep learning-based approaches encompass techniques such as Variational Dynamics Encoders (VDEs), Variational Approach to Markov Processes Networks (VAMPnets), Deep Canonical Correlation Analysis (Deep CCA), and Time-lagged Autoencoders (TAEs) [46, 48].

In this thesis, the main objective is to design tailored geometric collective variables (CVs) for comprehending dynamic and conformational transformations in GPCR systems. By employing well-suited collective variables, we can investigate the conformational changes and construct potential of mean force (PMF) plots associated with them.

2.5.2 Potential of mean force

The potential of mean force (PMF) represents the free energy surface along the reaction coordinate(s) or chosen collective variable(s) [40, 45]. It is given by the equation

$$A(\varepsilon) = -\frac{1}{\beta} \ln Q(\varepsilon) + C \tag{2.35}$$

where A(ε) denotes the free energy along the collective variable, ε and Q(ε) is the corresponding partition function.

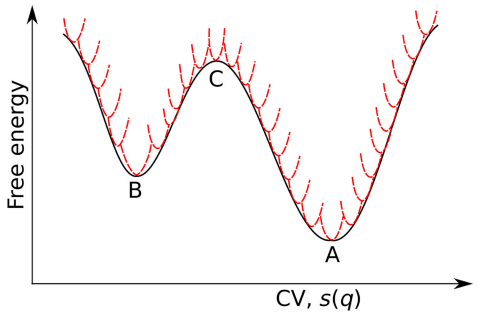


Figure 2.5: Schematic illustration of the umbrella sampling method. The red dash lines represent the harmonic bias potentials that are added to the system Hamiltonian at different CV points (windows) along the CV space. Figure and caption adapted from [50]

2.5.3 Umbrella Sampling

Umbrella Sampling, proposed by Torrie and Valleau in 1977 [49], is an enhanced-sampling technique widely employed to surmount energy barriers along a chosen collective variable, ε. The umbrella sampling method involves dividing the sampling space into multiple windows, each associated with an initial structure positioned along the collective variable. In order to confine the system within a narrow phase space along the collective variable while ensuring adequate overlap between adjacent windows, a biasing potential, typically a harmonic bias, is applied to each window.

The overall potential energy of the system, $U_i(\varepsilon)$, including the biasing potential, in the window i, is given by the equation 2.36

$$U_{\text{i, biased}}(\varepsilon) = U_{\text{i, unbiased}}(\mathbf{r}) + V_i(\varepsilon) \tag{2.36}$$

where $U(\mathbf{r})$ represents the original unbiased potential energy function of the system, dependent on the atom coordinates $\mathbf{r}$. The harmonic biasing potential $V(\varepsilon)$ for each window i is given by the equation:

$$V_i(\varepsilon) = \frac{1}{2} k_i(\varepsilon - \varepsilon_i)^2 \tag{2.37}$$

where k_i is the spring constant and ε_i is the target value of the collective variable for window i. The PMF of each window in umbrella sampling is given by

$$A_i(\varepsilon) = -K_b T ln(P_i^b) - V_i(\varepsilon) + C_i \tag{2.38}$$

Here, P_i^b is obtained through sufficient MD simulation, and C_i is a constant independent of ε, and is equal to $-k_b T ln < (exp\{-\beta V_i(\varepsilon)\} >$. Molecular dynamics simulations of the biased system are then performed, and the potential of mean force (PMF) is subsequently

calculated using the weighted histogram analysis method (WHAM) [51] or umbrella integration [49]. To compute the global probability distribution, WHAM assigns weights to each window combines individual window to minimize the error of $P^u(\varepsilon)$ [51]. The following are the WHAM equations [50, 51]:

$$P^u(\varepsilon) = \frac{\Sigma_i^N h_i(\varepsilon)}{\Sigma_j^N n_j[\beta exp(F_j - V(\varepsilon)]} \tag{2.39}$$

$$F_j = -k_b T ln \Sigma_q P_u(\varepsilon) exp(-\beta V(\varepsilon)) \tag{2.40}$$

where β is the inverse temperature($\frac{1}{k_b T}$), i and j are indexes of the windows, F_j is the factor determined for window j, n_j is the total number of data-points in window j, $P^b(\varepsilon)$ is the probability distribution function of the collective variable obtained from the biased simulations, and $P^u(\varepsilon)$ is the reference probability distribution function in the absence of bias.

In summary, *Umbrella Sampling* is an effective enhanced sampling technique that utilizes biased molecular dynamics simulations with harmonic potentials to sample conformational states along a chosen collective variable, ε, to compute the potential of mean force (PMF) and provide valuable insights on the underlying free energy landscape. In this thesis, we use Umbrella Sampling to compute the PMFs for different GPCR systems to understand the free energy landscape.

2.6 Statistical Methods and Machine Learning

2.6.1 Correlation

Correlation is a measure of the statistical relationship (or dependence) between variables, often columns in the a dataset. Correlation coefficient is a value that ranges between -1 and +1. Here +1 corresponds to positive correlation, which means when x always changes in the direction of y. Conversely, -1 implies x and y change in the opposite direction to each other, and 0 implies they are not related to each other (independent). Some of the correlation coefficients used are Pearson's r, Spearman's ρ or r_s, Cramer's ϕ, Point-biserial, and Kendall's τ.

2.6.1.1 Pearson's Correlation Coefficient

The Pearson's correlation coefficient, denoted by r, measures the linear relationship between two variables. It is given by the equation,

$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{2.41}$$

2.6.1.2 Spearman's Correlation Coefficient

The Spearman's correlation between two variables is the same as the Pearson's correlation between the rank of the two variables. Spearman's correlation, denoted by r_s or ρ, assesses the monotonic relationships between the two variables regardless of it being linear or non-linear. It is given by the equation,

$$r_s = \frac{Cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \tag{2.42}$$

2.6.2 Machine Learning

Machine learning (ML) is a sub-set of Artificial Intelligence (AI), where models are built and trained on data to learn and make predictions and/or decisions without explicit programming. Machine learning, often used as a "black-box", is a powerful and useful tool used in protein-structure prediction, drug discovery, bio-informatics, collective variable discovery, and analysis of large data amongst others. These include classification, regression, clustering, dimensionality reduction, In the next subsections, we explore some of the ML methods used in feature selection in Chapter 3.

2.6.2.1 Linear Regression

Linear Regression is a part of supervised machine learning algorithms which provides a linear approach in modeling the relationship between the dependent variable (y, also known as the target variable or response variable) and independent variables ($[x_1, x_2, \ldots x_n]$, also known as predictor variables or features) [52]. It is given by the equation 2.43,

$$y = \beta_0 + \Sigma_{i=1}^{n} \beta_i * x_i \tag{2.43}$$

where β_i are the regression coefficients or weights. Though linear regression is usually used for prediction or inference of the dependent variable, it can also be used or feature selection. In the case of linear regression, the regression coefficients/weights (β_i) are the feature importance scores. They give valuable insights on the strength (magnitude) as well as direction (sign) on the relationship between the dependent variable (y) and the independent variables ($[x_1, x_2, \ldots x_n]$).

2.6.2.2 Decision Trees

A decision tree is a tree-like hierarchical model wherein each node is a test on a condition with each branch representing the possible outcome on the test [52]. (refer Figure 2.6) In decision tree, the feature importance is determined upon calculating the probability of an observation falling into a particular node. The Breiman feature importance [52, 53] is calculated as follows,

$$I_l^2(t) = \Sigma_{t=1}^{N} \hat{i}_t^2 I(v(t) = l) \tag{2.44}$$

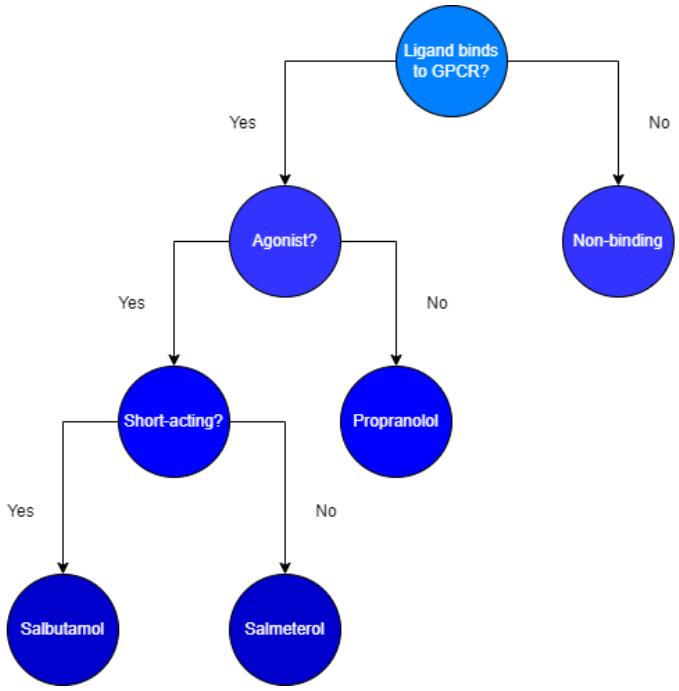


Figure 2.6: A simple decision tree classifying GPCR ligands.

where T is the decision tree, J is the number of internal does, l is the feature for which importance is calculated, v(t) is the feature for which node t is split, $\hat{i}^2$ is the reduction in the error metric used for splitting (usually Mean Squared Error).

2.6.2.3 Random Forest

Random forest is an ensemble of decision trees. Decision trees are combined with the bagging method with an idea that combining multiple decision trees betters the overall result. In random forests, the feature importance can be calculated as the average feature importance across decision trees [52]. (refer equation 2.45) In sci-kit, the feature importance is calculated using the *gini* index.

$$I_l^2(t) = \frac{1}{M} \Sigma_{m=1}^{M} I_l^2(T) \tag{2.45}$$

where M is the decision trees.

2.6.2.4 XGBoost

XGBoost, also known as extreme gradient boosting, is a recent boosting-based ensemble machine learning algorithm. It combines multiple weak learning models (like decision trees) to reduce the overall bias and errors by weighted sampling and merging [54]. The feature importance in XGBoost can be calculated based on weight (the number of times a feature

occurs in a tree), gain (average gain of splits using the feature), and cover (average coverage of splits using the feature).(https://xgboost.readthedocs.io/en/latest/python/python_api.html#xgboost.Booster.get_score)

2.6.2.5 K Nearest Neighbours

The K Nearest Neighbours (KNN) is a non-parametric supervised machine learning algorithm [52, 55]. In KNN regressor, the output is predicted based on the local interpolation of targets based on the nearest neighbours of the training set. (https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html) SInce the feature importance for KNN is not defined directly, one can calculate the feature importance through permutation importance. It involves permuting the values of a single feature while keeping the others constant and observing the effect on the model's performance. By comparing the model's performance before and after the permutation, we can assess the impact of each feature on the prediction accuracy. Consequently, features causing a significant drop in performance when permuted are considered more important, as they contribute crucial information for the model's predictions.

Chapter 3

Molecular Dynamics and Machine Learning Study of Adrenaline Dynamics in the Binding Pocket of GPCR

Contents

3.1 Abstract

GPCRs are the most prominent family of membrane proteins that serve as major targets for one-third of the drugs produced. A detailed understanding of the molecular mechanism of drug-induced activation and inhibition of GPCRs is crucial for the rational design of novel therapeutics. The binding of the neurotransmitter adrenaline to the β_2-adrenergic receptor (β_2AR) is known to induce a *flight or fight* cellular response, but much remains to be understood about binding-induced dynamical changes in β_2AR and adrenaline. In this article, we examine the potential of mean force (PMF) for the unbinding of adrenaline from the orthosteric binding site of β_2AR and the associated dynamics using umbrella sampling and molecular dynamics (MD) simulations. The calculated PMF reveals a global energy minimum, which corresponds to the crystal structure of β_2AR-adrenaline complex, and a meta-stable state in which the adrenaline is moved slightly deeper into the binding pocket with a different orientation compared to that in the crystal structure. The orientational and conformational changes in adrenaline during the transition between these two states and the underlying driving forces of this transition are

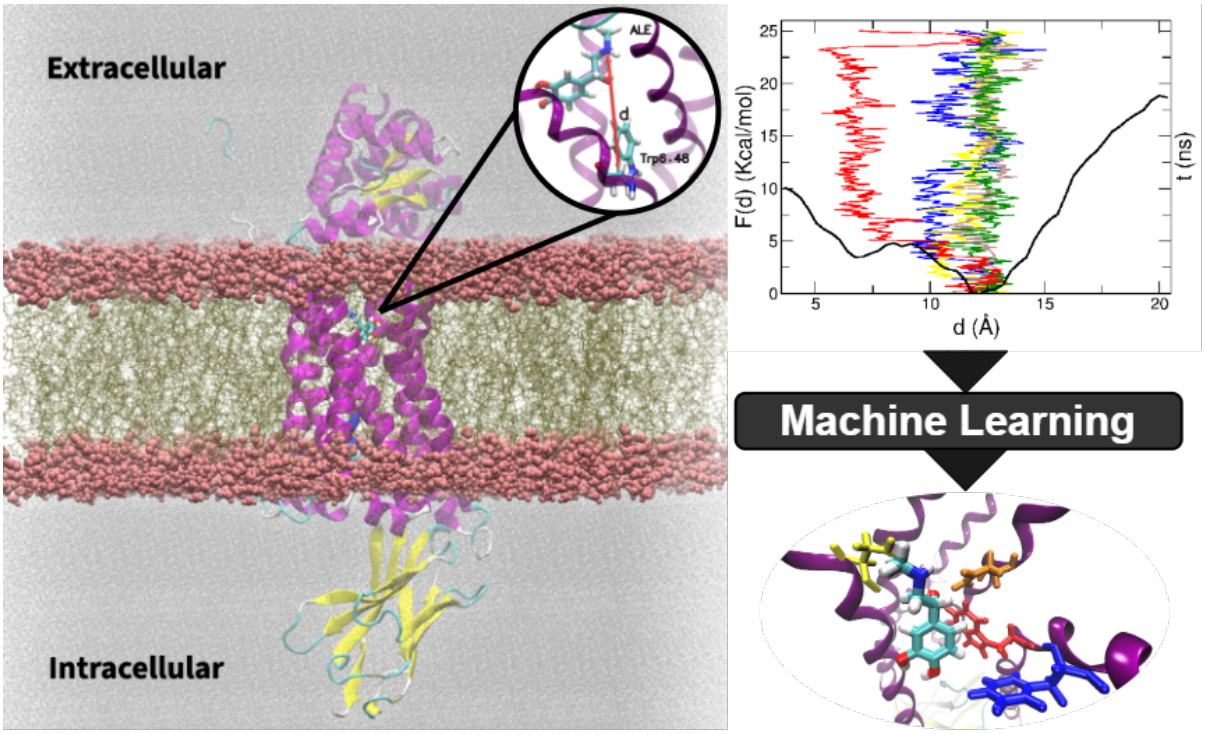


Figure 3.1: TOC image

also explored. Based on clustering of MD configurations and machine learning-based statistical analyses of time series of relevant collective variables, the structures and stabilizing interactions of these two states of the β_2AR-adrenaline complex are also investigated.

3.2 Introduction

G-Protein Coupled Receptors (GPCRs) are the prime trans-membrane protein targets for approximately one-third of all the approved drugs [9–11]. They mediate cells' response to hormones and neurotransmitters and play a vital role in regulating sensory functions such as vision, smell, and taste, and physiological functions such as homeostasis, behavior, and mood regulation [6–8]. A diverse range of molecular stimuli, including odorants, tastants, vitamins, neurotransmitters, hormones, pheromones, nucleotides, lipids, and ions, bind to the GPCRs, and signals are transduced to activate various regulatory processes [56]. The ligands that bind to GPCRs can be classified based on their functions into agonists, inverse agonists, antagonists, and allosteric modulators [57]. The ligand-induced conformational changes in GPCRs are critical for activating and regulating different signaling pathways [58–60].

The binding of an agonist to the orthosteric site close to the extracellular end of the receptor activates the GPCR [57]. The ligand binding induces conformational changes in GPCR that are

relayed through the trans-membrane (TM) regions to the G-protein bound at the intracellular end of the receptor [61]. In the absence of the ligand, the G-protein exists in the inactive state, in which the G_β and G_γ sub-units are attached to the GDP-bound G_α sub-unit [62]. However, upon ligand binding to the receptor, the G-protein changes to the active state, in which the G_α sub-unit detaches from both the $G_{\beta\gamma}$ dimer and the receptor and GDP is exchanged for GTP [62]. The agonist binding induced activation of the GPCR and G-protein ultimately leads to the initiation of various downstream signaling events responsible for cellular functions.

The human β_2-adrenergic receptor (β_2AR) is a cell surface receptor primarily located in nerve endings, vascular and bronchial smooth muscle cells, epithelial cells, and immune cells, thereby playing an essential role in regulating the vascular, pulmonary, cardiac, and metabolic functions [63].

Adrenaline and noradrenaline are the natural agonists of β_2AR, while other adrenergic drugs commonly used for treating asthma and chronic obstructive pulmonary disease (COPD) mimic the functions of these natural agonists [64]. One of the major molecular dynamics simulations study on β_2AR by Dror et al. used the special-purpose Anton [65] supercomputer to study the deactivation process and gain subsequent insights on the activation process [66]. The β_2AR shifted from a crystallographic-obtained agonist-bound active state conformation to an inactive conformation in a time-scale of 2-50 μs in the absence of a G-Protein or G-Protein mimetic nanobody [66, 67]. Since then several studies have identified key events in the activation of β_2AR including the outward movement of the trans-membrane helix 6 (TM6), the opening of the hydrophobic lock, the regulation of different molecular switches, and the collapse of Na^+ binding pocket [68–76]. The ligands initially bind to the extracellular vestibule containing the residues $Tyr^{7\cdot35}$, $Phe193^{ECL2}$, $Ala^{5\cdot39}$, $His^{6\cdot58}$, and $Val^{6\cdot59}$ before entering the binding pocket [77].The residues $Leu^{3\cdot43}$, $Leu^{6\cdot40}$ and $Ile^{6\cdot41}$ form the hydrophobic lock [68–71], which upon opening loosens the TM3-TM6 packing and allows the rotation of the TM6 helix during activation [72]. Here, the Ballesteros-Weinstein (BW) numbering scheme [20] is used to denote residues, where $Leu^{6\cdot40}$ refers to the leucine residue of the sixth trans-membrane helix of β_2AR that is at the 40^{th} position relative to the most conserved residue ($Pro^{6\cdot50}$ residue) at the 50^{th} position. [78] The aromatic residues in the extracellular ligand binding region are often conserved, and the rearrangement of their aromatic stacking interactions occurs during activation [73].The conserved core triad formed by the residues present below the binding pocket, $Phe^{6\cdot44}$, $Pro^{5\cdot50}$, $Ile^{3\cdot40}$ undergo rearrangement upon agonist binding and may initiate β_2AR signaling cascade [79, 80]. The activation process involves four molecular switches: (i) $Trp^{6\cdot48}$ toggle/transmission switch, ii) $Tyr^{7\cdot53}$ toggle switch, (iii) the ionic lock between TM3 and TM6, and (iv) 3-7 lock between TM3 and TM7 [74]. The activation of β_2AR also involves the collapse of Na^+ pocket located around the $Asp^{2\cdot50}$ residue [75]. The $Trp^{6\cdot48}$ residue is probably the most

common switch amongst all GPCRs, and it acts as a rotamer toggle switch and a transmission switch necessary for the activation of β_2AR [81, 82].

Molecular dynamics and enhanced sampling studies suggested that the residues $Trp^{3\cdot28}$, $Asp^{3\cdot32}$, $Val^{3\cdot33}$, $Val^{3\cdot36}$, $Phe193^{ECL2}$, $Tyr^{5\cdot38}$, $Ser^{5\cdot42}$, $Ser^{5\cdot43}$, $Ser^{5\cdot46}$, $Trp^{6\cdot48}$, $Phe^{6\cdot51}$, $Phe^{6\cdot52}$, $Asn^{6\cdot55}$, $Tyr^{7\cdot35}$, $Ile^{7\cdot36}$ and $Asn^{7\cdot39}$ were found to collectively interact with different β_2AR ligands [46, 83–85]. Moreover, the residues $Asp^{3\cdot32}$ and $Asn^{7\cdot39}$ frequently formed ionic interactions and hydrogen bonds respectively with the ligands and were therefore suggested to be key residues for ligand binding [83, 85]. The GPCRmd server hosts unbiased MD trajectories across various GPCR classes and receptor-ligand complexes, including the β_2AR-adrenaline complex, and offers a protein-ligand interaction network visualization tool [86]. Despite considerable progress, much remains to be understood about the ligand dynamics during and after the activation of β_2AR.

Previous mutagenesis experiments on the human β_2AR receptor revealed the importance of hydrogen bonds formed by residues $Asn^{7\cdot39}$ and $Asn^{6\cdot55}$ with L-adrenaline for effective binding [87, 88]. Similarly, a point mutation involving $Asp^{2\cdot50}$ with $Asn^{2\cdot50}$ reduced L-adrenaline binding affinity by ~ 140-fold [89], while replacing $Asp^{3\cdot49}$ with $Asn^{3\cdot49}$ increased binding affinity by ~ 13-fold [90]. The GPCRDB database [14, 91] provides a comprehensive summary of structural and mutagenesis data related to crucial residues involved in the binding and recognition of adrenaline and other agonists with the β_2AR [92]. This database has proven instrumental in conducting a thorough comparison between the residues identified through our method and the data available in GPCRDB [14, 91].

In this paper, we examine the dynamics of adrenaline (ALE) within the binding pocket of β_2AR using umbrella sampling and molecular dynamics simulations of the active state of adrenaline-bound β_2AR. The $Trp^{6\cdot48}$ residue, which is a highly conserved residue located in the orthosteric-binding site of β_2AR, is used as a reference point, relative to which the translational, orientational and conformational dynamics of the adrenaline are investigated. Using umbrella sampling simulations with a suitable distance-based collective variable, we have calculated the potential of mean force (PMF) for the unbinding adrenaline from the orthosteric binding pocket of β_2AR. The results reveal the presence of a meta-stable state in addition to the most-stable global energy minimum state for the ligand within the binding pocket of β_2AR. Clustering is used to identify structures representing the aforementioned states. We also used various machine learning algorithms to identify the most important residues responsible for state transition based on the pairwise residue interaction energy values between adrenaline and residues in the binding pocket.

3.3 Methods

3.3.1 Models

The crystal structures of the adrenaline-free inactive (PDB ID: 2RH1) [93] and the adrenaline-bound active (PDB ID: 4LDO) [94] states of β_2AR were chosen as the starting configurations for the molecular dynamics simulation of these systems. It is to be noted that the active-state crystal structure (PDB ID: 4LDO) specifically modelled only the pharmacologically active and endogenous neurotransmitter L-adrenaline and not its enantiomer (D-adrenaline) [94–96]. These receptors were inserted into the pre-equilibrated phosphatidylcholine (POPC) lipid bilayer using the OPM database [97] and CHARMM-GUI software with default parameters [98–102]. The total numbers of POPC molecules in the lipid bilayers of the active and inactive β_2AR models are 383 and 319, respectively. The Amber Lipid17 [103] and AMBER19SB [104] force fields were used for the lipid bilayer and the receptors, respectively. The Generalized Amber Force Field (GAFF) [105, 106] with the AM1-BCC [107] partial charges for the adrenaline and the TIP3P [108] model for water were used. A suitable number of sodium and chloride ions were added to the systems to maintain a salt concentration of 0.15 M. The box dimensions for the resulting models were 110 Å x 110 Å x 140 Å with 36331 water molecules for the inactive state and 120 Å x 120 Å x 160 Å with 52134 water molecules for the active state. The GPCRDB database [14, 91] was used to identify residues interacting with adrenaline and was also used to number the residues in the BW numbering scheme. The final adrenaline-free inactive β_2AR model contained the T4-lysozyme and the adrenaline-bound active β_2AR model contained both the T4-lysozyme and G-protein mimicking Camelid antibody fragment. These structures stabilize the conformation of β_2AR in their inactive [93] and active state [94]. Moreover, the presence of G-protein or a G-protein mimicking nanobody increases the lifetime of the structure in the active state[66, 67, 109]. However, caution must be exercised while interpreting the results, as the obtained free energy profile could differ for different models based on the presence or absence of the intracellular binding partner[110]. In this study, the adrenaline-free inactive β_2AR model is only used to study any changes in the dihedral angle of $\mathrm{Trp}^{6.48}$.

3.3.2 Molecular Dynamics Simulation

The potential energies of these membrane models were minimized using the steepest gradient descent for 10000 cycles and conjugate gradient descent for another 10000 cycles [111, 112]. The energy-minimized systems were gradually heated to 303.15 K at 1 atm pressure and equilibrated there for 5 ns, followed by a production run of 50 ns under constant temperature and constant pressure conditions. The temperature was controlled using the Langevin ther-

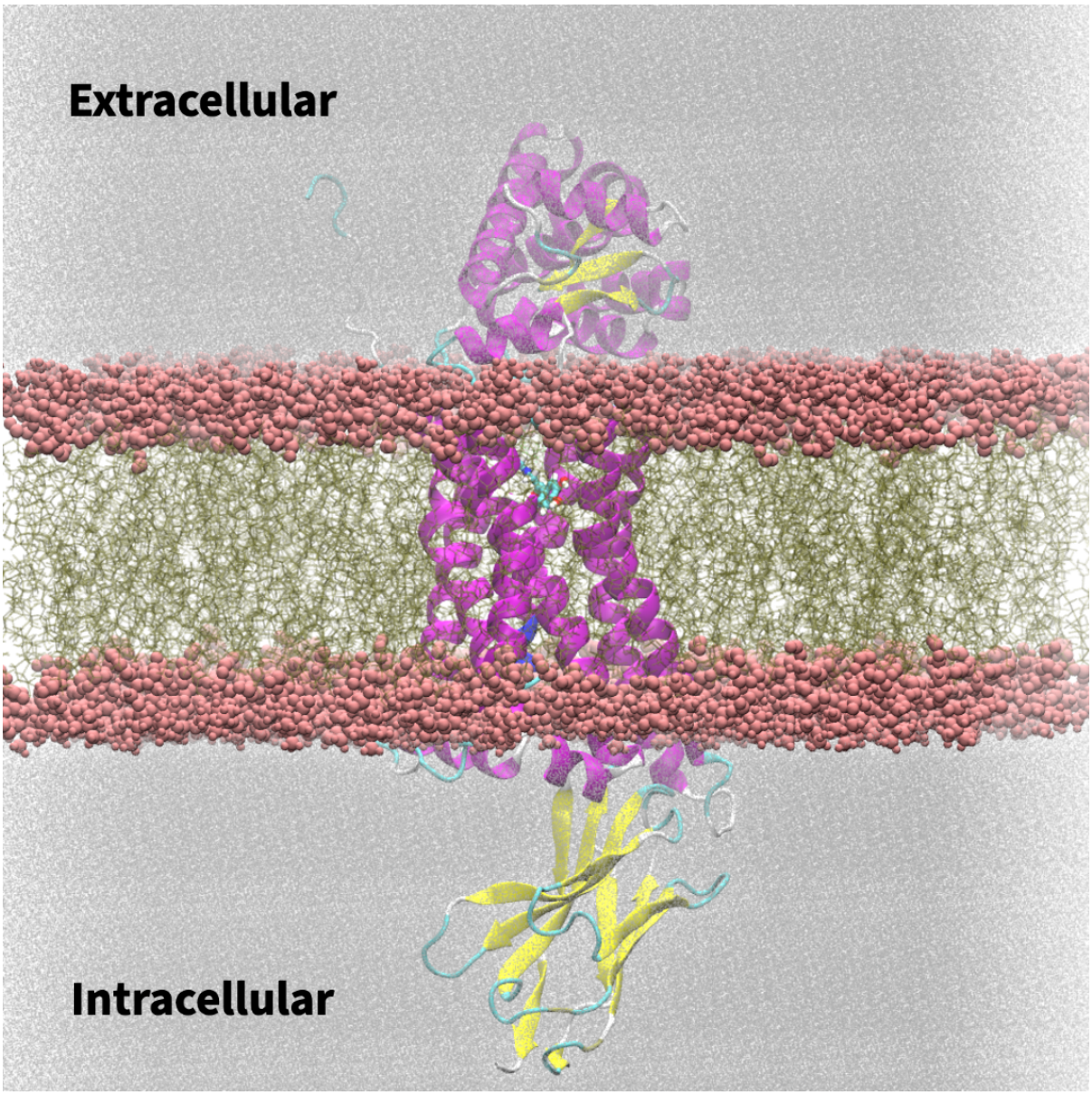


Figure 3.2: A snapshot of T4L-adrenaline-bound β_2AR with camelid antibody fragment embedded in a hydrated POPC lipid bilayer membrane (head-group (red) and tail groups (green)). Water molecules are shown in grey color.

mostat [113], and the pressure was controlled using the anisotropic Berendsen barostat [114]. SHAKE algorithm [115] was used to constrain bonds involving hydrogen atoms. The particle mesh Ewald [116] approach was used for treating the electrostatic interactions with a real-space cutoff distance of 10 Å. The equations of motion were integrated using the velocity Verlet algorithm [117] with a time step of 2 fs, with data written every 100 ps. All the simulations, including the enhanced sampling simulations, were done using the AMBER18 software [118]. Starting from different initial velocities, we performed four additional MD simulations, each for 25 ns; all these simulations were further extended to 500 ns, resulting in a total simulation time of 2.5 μs for the adrenaline-bound active β_2AR model. All the trajectory analyses were carried out using the Amber CPPTRAJ [119] and VMD package [120].

3.3.3 Umbrella Sampling

Umbrella sampling (US) simulations were performed on the adrenaline-bound β_2AR to examine the binding energetics of the system. The distance (denoted by d) between the $C\alpha$ of $\text{Trp}^{6.48}$ and the nitrogen of adrenaline was chosen as the collective variable (CV) for the umbrella sampling simulations (Figure 3.3) based on a previously established protocol [121]. Since the position of $\text{Trp}^{6.48}$ is more or less unchanged and remains conserved during activation, it serves as a valuable reference for understanding the adrenaline dynamics within the binding pocket of β_2AR. The CV was varied from 4 Å to 20 Å in steps of 0.5 Å, resulting in a total of 33 windows. The US simulation for each window was carried out in the NPT ensemble for 6 ns under the same conditions as unbiased MD runs. The spring constant used for the harmonic biasing potential was 10 kcal/mol/Å² for production and 200 kcal/mol/Å² for minimization. The Weighted Histogram Analysis Method (WHAM) (`membrane.urmc.rochester.edu/?page_id=126`) was used to determine the potential of mean force (PMF) from CV trajectories obtained from the US simulations [51]. The last 5-ns long umbrella sampling trajectory per window was used for the WHAM.

3.3.4 Adrenaline-Receptor Interaction Energy

In order to characterize the interaction between adrenaline and β_2AR, we computed the pairwise interaction energy between adrenaline and the binding pocket residues of the receptor using the Linear Interaction Energy [122] (LIE) module from Amber CPPTRAJ package. This interaction energy is a sum of the van der Waals and electrostatic energies between adrenaline and a given binding pocket residue of β_2AR.

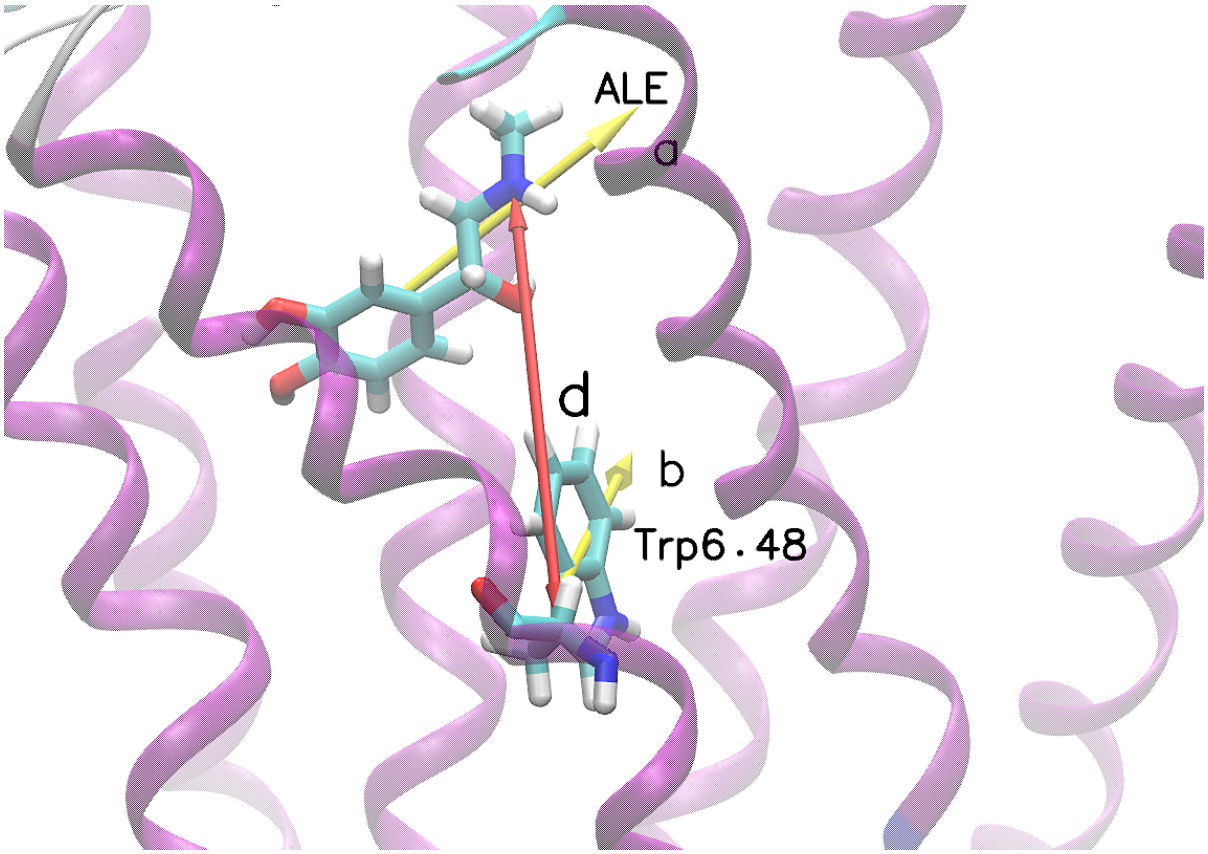


Figure 3.3: Collective variables: d is the distance between the $C\alpha$ atom (cyan) of $\text{Trp}^{6.48}$ and the nitrogen atom (blue) of adrenaline (ALE). $\boldsymbol{a}$ and $\boldsymbol{b}$ (yellow arrows) are the vectors corresponding to the molecular long axes of adrenaline and $\text{Trp}^{6.48}$, respectively.

3.3.5 Correlation Analysis

Given MD-derived time series of different collective variables and energies of the system, we examined the correlation between different variables using the Pearson's correlation (r) [123] and Spearman's correlation coefficient (r_s) [124], which are defined as follows,

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.1}$$

$$r_s = \frac{Cov(R(X),R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}} \tag{3.2}$$

where, $Cov(X,Y)$ is the covariance of X and Y, $Cov(R(X),R(Y))$ denotes the covariance of rank variables $R(X)$ and $R(Y)$ of X and Y, respectively, σ_X, σ_Y, $\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are the standard deviation of X, Y, $R(X)$, and $R(Y)$, respectively.

3.4 Results and Discussion

3.4.0.1 Potential of Mean Force for Drug Dissociation

The umbrella sampling derived potential of mean force (PMF), $F(d)$, for unbinding adrenaline from the orthosteric binding pocket of β_2AR shows a global energy minimum at $d = 12.1$ Å, which is close to the value of d calculated from the experimental crystal structure of adrenaline-bound β_2AR (11.84 Å), and a meta-stable minimum at $d = 6.8$ Å (Figure 3.4). The meta-stable state is found to be ~ 3.5 kcal/mol less stable than the state at $d = 12.1$ Å. The activation barrier for the transition from the most-stable state to the meta-stable state is ~ 4.8 kcal/mol, whereas it is ~ 1.3 kcal/mol for the reverse transition. The presence of two energy minima on the PMF indicates that adrenaline can adopt at least two distinct poses at two different sites within the binding pocket of β_2AR. Since the value of d of the meta-stable state is less than the global energy minimum state, it appears that adrenaline can move deeper inside the binding pocket of β_2AR.

An activation barrier of ~ 4.8 kcal/mol suggests that the transition from the most-stable state to the meta-stable state is an activated rare event. The rarity of this transition is evident in the time series of d obtained from five independent short 25-ns unbiased MD trajectories (Figure 3.4). Four trajectories (labeled as TJ1-TJ4) sampled only the global energy minimum basin. In contrast, the other trajectory (labeled as TJ5) exhibited barrier-crossing transitions between the states within the timescale of the MD simulation. The observed transitions in TJ5 substantiate the existence of the meta-stable state for the adrenaline-β_2AR complex. The presence of a "shoulder state" ($9.5 < d < 11.5$) is observed as the system samples around the local maxima, but is unable to cross the barrier in TJ1 and TJ2. To further examine the relative stability of the meta-stable state and the structural changes accompanying this

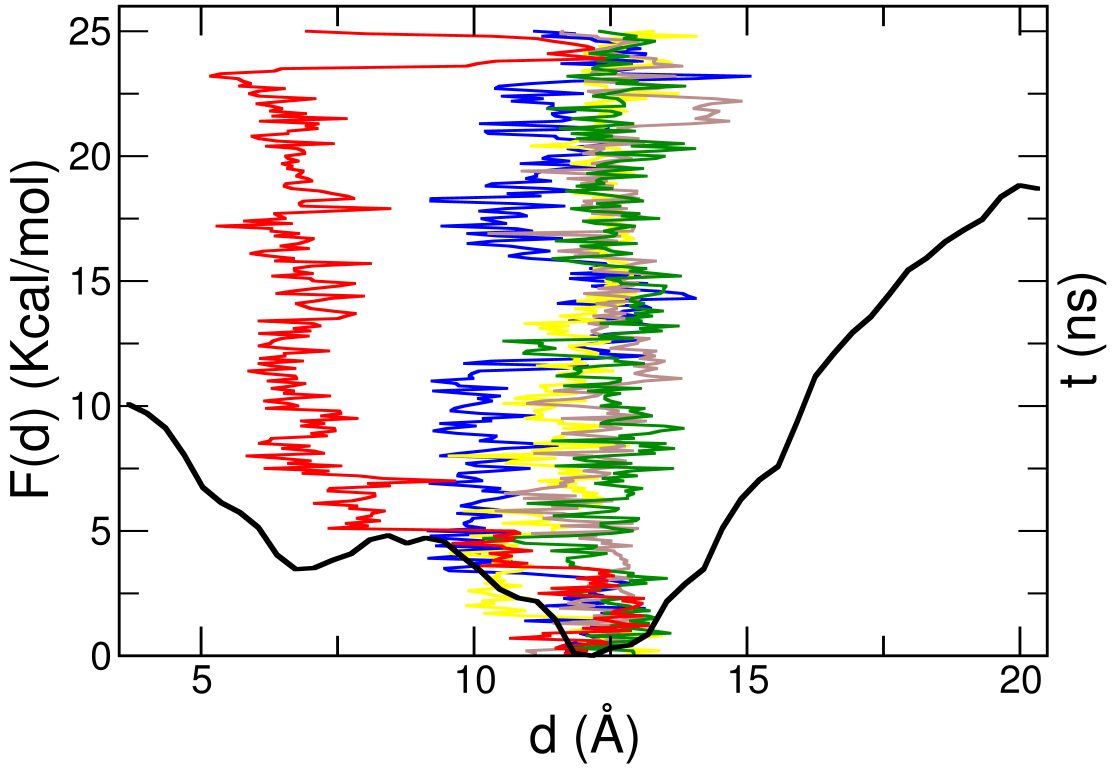
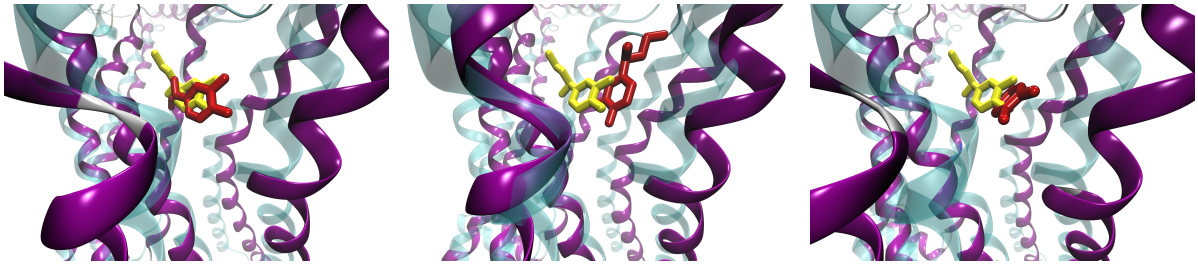


Figure 3.4: Computed PMF as a function of d (black) is shown along with the time series of d obtained from the five independent 25 ns unbiased MD trajectories.

transition, the trajectories TJ1-5 were further extended to 500 ns. The time series of d (refer Figure S1a) shows that given sufficient simulation time, all the systems in trajectories TJ1-5 exhibited barrier crossing transitions at least once during the course of the simulation. Most notably, the system in trajectory TJ5 stayed in the meta-stable state for most of the simulation time. However, amongst trajectories TJ1-4, only the system in TJ3 managed to stay for a long period of time ($\sim$ 103 ns) in the meta-stable state after its state transition at $\sim$ 397 ns.

To characterize and examine the structures of the β_2AR-adrenaline complex in the aforementioned states, the K-means clustering algorithm based on RMSD was employed on the combined simulation trajectory (TJ1-5) totaling 2.5 μs with K=10 iterated over 10000 times into 10 clusters, namely $c0 - c9$ with $c0$ being the most populated cluster, and $c9$ being the least populated cluster. Table 3.1 summarizes the results obtained from the clustering analysis. The rationale behind choosing 10 clusters was to strike a balance between capturing the relevant conformational space of the system (based on the collective variable, d) and avoiding excessive fragmentation or aggregation of states. It is important to note that alternative clustering algorithms or different numbers of clusters could potentially yield different insights or interpretations of the data. The clusters that fall near the global-minimum basin, including the "shoulder state" clusters ($9.5 < d < 11.5$), are highlighted in bold font in table 3.1, with their population adding to 55.4% of the total population. Among these clusters, $c4$ had its mean d values closest to the most-stable state configuration. Similarly, cluster $c2$ was closest to the meta-stable state configuration. The cluster $c1$ has a mean value of d = 9.14 Å is closest to the local maximum at around $d \approx 8.8$ Å. Therefore, we consider the cluster centroid structures (Figure 3.5) of $c1$, $c4$, and $c2$ likely correspond to local maximum, most-stable, and meta-stable states of the adrenaline-β_2AR complex, respectively.



(a) (b) (c)

Figure 3.5: Representative median structures of the (a) local energy maximum (cluster c1) (b) most-stable (cluster c4) and (c) meta-stable (cluster c2) states of the β_2AR-adrenaline complex (magenta) obtained from the K-means clustering algorithm aligned on the β_2AR-adrenaline crystal structure (PDB ID: 4LDO, cyan). Adrenaline is colored in red in the clusters and its aligned crystallographic poses are colored in yellow.

Cluster	Population (%)	σ	d (Å)	θ ($^\circ$)	δ_3($^\circ$)
c0	**22.1**	**3.54**	**11.10**	**127.59**	**80.17**
c1	18.7	4.62	9.14	99.54	81.87
c2	13.2	5.57	6.74	67.61	80.26
c3	12.6	4.41	7.89	74.66	91.27
c4	**8.8**	**4.24**	**12.32**	**106.09**	**18.34**
c5	**8.1**	**4.06**	**10.26**	**78.78**	**53.03**
c6	**5.9**	**3.76**	**9.75**	**81.24**	**105.34**
c7	**5.0**	**5.43**	**9.73**	**63.14**	**96.54**
c8	**4.1**	**4.59**	**10.63**	**161.56**	**82.11**
c9	**1.4**	**2.79**	**10.27**	**146.53**	**96.69**

Table 3.1: Clusters c0 to c9 in the decreasing order of population with their corresponding standard deviation (σ) from their respective centroids, and their mean d, θ and δ_3 values.

To further understand the transition between the meta-stable and most-stable states, we examined the changes in adrenaline-β_2AR contacts as the system undergoes this transition. For a given configuration in the trajectory, a residue is said to be in contact with adrenaline if any of its atoms is within 4.5Å from any of the atoms of adrenaline. A total of 47 contact residues were identified during the 500 ns molecular dynamics simulation (TJ5). Upon comparing the cluster-average structures of the most-stable (c4) and meta-stable (c2) states, the following contact residues Cys$^{6\cdot47}$, Trp$^{6\cdot48}$ were removed on state transition from meta-stable state to the most-stable state. Consequently, the following new contacts with residues His$^{6\cdot58}$, Val$^{6\cdot59}$, Thr195^{ECL2}, Tyr$^{7\cdot43}$, Tyr$^{5\cdot38}$, Ala$^{5\cdot39}$ were formed. The contact with residues Ser$^{5\cdot42}$, Ser$^{5\cdot43}$, Ser$^{5\cdot46}$, Phe$^{45\cdot52}$, Val$^{3\cdot33}$, Phe$^{6\cdot51}$, Phe$^{6\cdot52}$, Val$^{3\cdot36}$, Asn$^{6\cdot55}$, Thr$^{3\cdot37}$ remained for both the stable and meta-stable cluster averaged states. However, studying the formation and loss of contacts during state transitions only provides information on directly interacting residues within a particularly short radius (4.5 Å) and may fail to include other critical residues responsible for state transition outside the given radius. Moreover, it becomes increasingly complex to track and understand the role of each and every contact residue during the course of a long MD simulation.

In TJ5, the state transition of the adrenaline-β_2AR complex from meta-stable to most-stable state occurs at $\sim$ 23-ns. It is anticipated that the system continues to stay in the most-stable state as its free energy is in the global energy minimum. However, this state transition is

very short-lived as we observe that the system comes back to the meta-stable state by ~ 26-ns instead of staying in the most-stable state (refer Figure S1a). We postulate that this return to the meta-stable state is probably due to other degrees of freedom that are coupled to d that are yet to relax to the most-stable state. From the trajectory TJ5, we also observed a relative orientation motion of adrenaline with respect to $Trp^{6.48}$ and a conformational change in adrenaline during the state transition. The orientational and conformational changes exhibited by adrenaline during the course of the MD simulation are discussed in the following sections.

3.4.0.2 Orientational Dynamics

To investigate the orientational dynamics of adrenaline in the binding pocket, we calculated the time series of the angle (θ) between the long molecular axes of adrenaline and $Trp^{6.48}$ from unbiased MD trajectories. The axis of the least mean square distance of atoms was considered as the long axis of the molecule in the present study and is denoted by vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ for adrenaline and $Trp^{6.48}$, respectively. Subsequently, θ was calculated as the angle between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ (shown in Figure 3.3).

The time evolution of θ obtained from MD trajectories (refer Figure S1b) reveals that θ fluctuates primarily around three different values (~ 130°, ~ 90°, and ~ 70°) and exhibits abrupt transitions between them at different time intervals. The most populated orientational state with $\theta \approx 130°$ corresponds to the most-stable state of the system on the PMF profile. In contrast, the other orientational states ~ 90°, and ~ 70° corresponds to the local maximum and the meta-stable local energy minimum state on the PMF profile respectively. In trajectory TJ3, as the system transitions from most-stable state to meta-stable state θ changes from ~ 130° to ~ 40° and fluctuates erratically between ~ 5° and ~ 110° for a short time span (470 $ns < t <$ 486 ns) during the simulation.

The time series of d unambiguously showed that adrenaline and $Trp^{6.48}$ came closer to each other during the transition from the most-stable state to the meta-stable state of the system. The time series of θ indicates that this transition also accompanies a significant reorientation of adrenaline in the binding pocket of β_2AR. A substantial degree of similarity and synchrony observed between the transitions in d and θ suggests a strong correlation between the orientational and translational dynamics of adrenaline in the binding pocket of β_2AR. This correlation between d and θ is evident in the d-versus-θ scatter plot shown in Figure 3.6. A sinusoidal wave is fit on the data points which suggests that the adrenaline appears to rotate relative to $Trp^{6.48}$ as it moves deep inside the binding pocket. As the state transitions from meta-stable state (~ 70°) to the most-stable state (~ 130°), the sine equation reduces to a linear correlation. The cluster centroids of $c1$, $c4$, and $c2$ observed at ($d = 9.14$Å, $\theta = 99.54°$) and ($d = 12.32$Å,

$\theta = 106.09°$), ($d = 6.74$Å, $\theta = 67.61°$) in the scatter plot correspond to the local-maximum, most-stable, and the meta-stable states of the system, respectively. (refer Table 3.1)

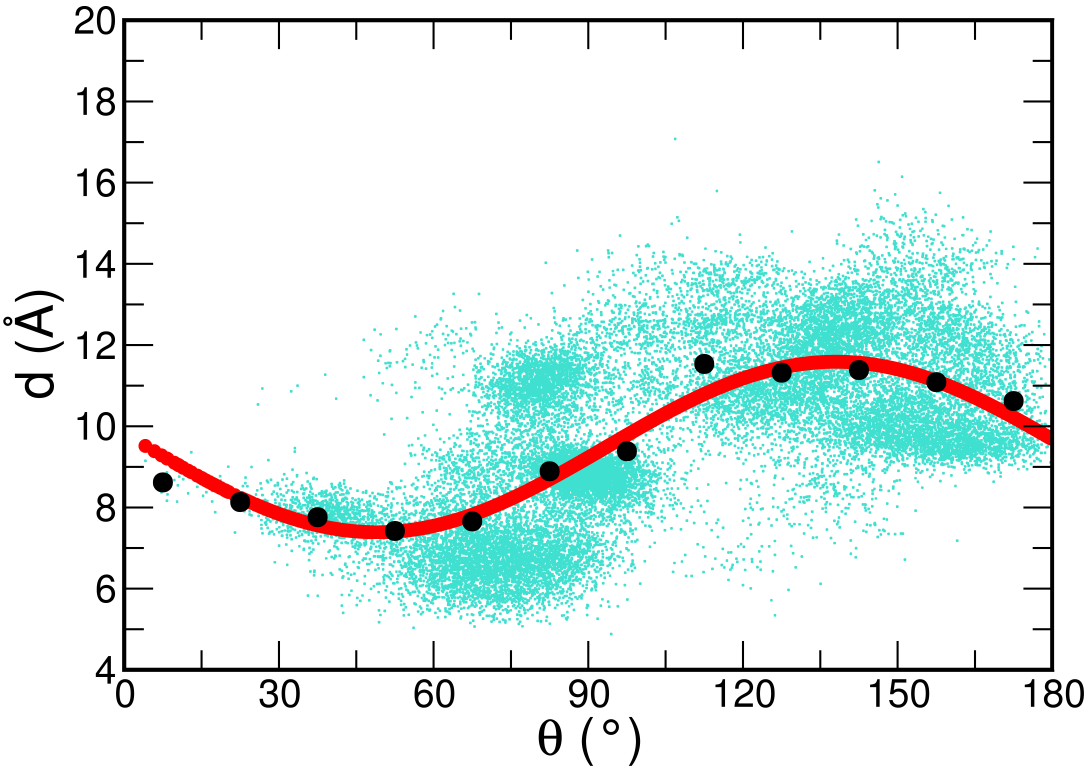
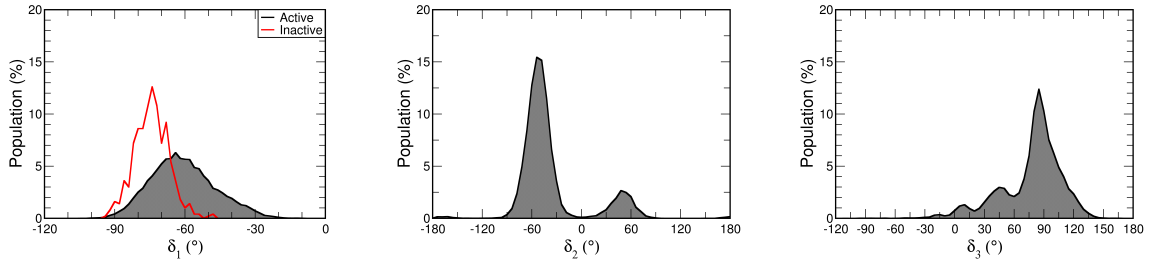


Figure 3.6: The scatter plot of d versus θ (in turquoise) obtained from all the MD trajectories generated. The black points are the bin mean values of d with bins from 0° to 180°, with bin size 15°. The red line $y = 9.49 + 2.10 * sin(2.02 * (x * \pi/180) + 2.98)$ is the best sinusoidal fit to the data points.

3.4.0.3 Conformational Dynamics

In order to assess the conformational dynamics of adrenaline and Trp$^{6.48}$ residue in the binding pocket of β_2AR, we calculated the $N - C\alpha - C\beta - C\gamma$ and $O - C - C - N$ dihedral angles of Trp$^{6.48}$ and adrenaline, respectively. Also, the pseudo-dihedral angle (corresponding to $C\gamma_{ALE} - N_{ALE} - C\alpha_{Trp^{6.48}} - C\gamma_{Trp^{6.48}}$) formed by the molecular planes of adrenaline and Trp$^{6.48}$ was also calculated to probe changes in their relative orientations. In what follows, the $N - C\alpha - C\beta - C\gamma$, $O - C - C - N$, and $C\gamma_{ALE} - N_{ALE} - C\alpha_{Trp^{6.48}} - C\gamma_{Trp^{6.48}}$ dihedral angles are denoted by δ_1, δ_2, and δ_3, respectively.



(a) (b) (c)

Figure 3.7: Histograms of dihedral angles (a) δ_1 (b) δ_2 (c) δ_3 for 5 x 500 ns adrenaline bound active state simulations (TJ1-5) shown in black. In (a), the histogram of δ_1 for adrenaline-free inactive state simulation is shown in red.

The crystal structure values for δ_1 for inactive (PDB ID: 2RH1) and active state β_2AR (PDB ID: 4LDO) are $-73.48°$ and $-83.79°$ respectively. The histogram of δ_1 (refer Figure 3.7a) indicates that $Trp^{6.48}$ prefers the *gauche(-)* conformation with an average dihedral angle of $-75.29°$ and $-61.78°$ in the inactive state and active state MD simulations respectively. No noticeable conformational transitions are observed in the time series of δ_1 even when adrenaline came closer to $Trp^{6.48}$ (refer Figure S2a). This restriction of $Trp^{6.48}$ to stay in the *gauche(-)* conformation was earlier attributed to its steric hindrance from $Phe^{6.52}$ residue [125].

Unlike $Trp^{6.48}$, adrenaline exhibits significant conformational changes during the course of the simulation. The histogram of the dihedral angle, δ_2 (refer Figure 3.7b) reveals two major conformations adopted by adrenaline, $\sim -60°$ (*gauche (-)*) and $\sim +60°$ (*gauche (+)*). Occasionally, adrenaline also adopts a trans conformation, corresponding to a δ_2 value of around $\sim \pm 180°$ (refer Figure S2b). Note that in the experimental crystal structure of the adrenaline-β_2AR complex, adrenaline adopts the *gauche(-)* conformation with $\delta_2 \sim -60°$ [126]. The histogram of δ_3 which denotes the angle between the molecular planes of $Trp^{6.48}$ and adrenaline about the collective variable d showed 3 distinct peaks at $\sim 90°$, $\sim 45°$, and $\sim 15°$(refer Figure 3.7c).

3.4.1 Adrenaline-Receptor Interaction Energy

To identify the key residues that contribute to the state transition, we calculate the residue-wise interaction energy between the binding site residues and adrenaline. A total of 71 binding pocket residues that are located within 10Å from the binding pocket were chosen for this interaction energy calculation. The time series of the total non-bonded energy (a sum of the electrostatic and van der Waals interactions, denoted by E) between a given residue and adrenaline was obtained from the unbiased MD trajectories for each chosen binding pocket

residue. As expected, the highest mean residue-wise interaction energy values belonged to those residues with strong polar, hydrophobic, hydrogen bonds and van der Waals interactions.

Residues	$Asp^{3\cdot32}$	$Val^{3\cdot33}$	$Asn^{6\cdot55}$	$Ser^{5\cdot42}$	$Phe^{45\cdot52}$	$Phe^{6\cdot52}$
$\bar{E}$ (kcal/mol)	-7.59	-3.73	-3.59	-2.88	-2.52	-2.16

Table 3.2: Mean residue-wise interaction energy (in kcal/mol) of key binding pocket residues that interact strongly with adrenaline across all unbiased MD trajectories (TJ1-5).

The recognition of meta-stable states and associated collective variables could be accelerated and automated using methods such as Principal Component Analysis (PCA), Time-Independent Component Analysis (TICA), and Variational Approach for Markov Processes (VAMP) [46, 48]. However, it is necessary to note that such methods require large amount of MD data and the resulting collective variables generated from these methods may be abstract and challenging to interpret directly [46, 48]. In this study, the collective variable, d is a bio-physics aware feature as it gives insight on the depth of the ligand within the binding pocket, and also distinguishes between the meta-stable and most-stable states. Therefore, it can be used for as a target variable for correlation and machine learning-based studies concerning ligand dynamics [127].

Given the time series of θ, $\delta_1, \delta_2, \delta_3$, and the residue-wise interaction energy, we examine which of these time series correlate reasonably well with the time series of d. The calculated Pearson's and Spearman's correlation coefficients for CVs that showed significant correlation with d are presented in Table 3.3. As observed before, θ shows high correlation with d and the interaction energies of a few binding pocket residues ($Phe^{6\cdot52}$, $Asn^{6\cdot55}$, $Trp^{6\cdot48}$, $Ser^{5\cdot43}$ and $Ser^{5\cdot42}$) also showed reasonable correlation with d. Although these residues appear to be important for the state transition, a more detailed machine learning-based non-linear correlation analysis between the interaction energy of individual residues and d is necessary.

Correlation	θ	$Phe^{6\cdot52}$	$Asn^{6\cdot55}$	$Trp^{6\cdot48}$	$Ser^{5\cdot43}$	$Ser^{5\cdot42}$
Pearson	0.62	0.64	0.44	0.40	0.38	0.17
Spearman	0.63	0.56	0.41	0.44	0.37	0.23

Table 3.3: Correlation coefficient values of θ and the residues $Phe^{6\cdot44}$, $Asn^{6\cdot55}$, $Trp^{6\cdot48}$, $Ser^{5\cdot43}$, and $Ser^{5\cdot42}$ with d.

We calculated the feature importance coefficients by fitting the interaction energy versus d data into the following machine learning regression models: (1) Linear Regression, (2) Decision

Trees, (3) Random Forest, (4) XGBoost, (5) K Neighbours Regressor [54, 128]. A 10-fold cross-validation was performed whilst training each of the aforementioned models, and their average feature importance scores were used to identify and rank the residues that are most likely responsible for predicting the state of the system.

ML Methods	Residues Identified
Linear Regression	$Phe^{6\cdot52}$, $Trp^{6\cdot48}$, $Tyr^{5\cdot38}$, $Phe^{6\cdot51}$, $Val^{3\cdot33}$
Decision Trees	$Phe^{6\cdot52}$, $Trp^{6\cdot48}$, $Trp^{3\cdot28}$, $Asp^{3\cdot32}$, $Asn^{7\cdot39}$
Random Forest	$Phe^{6\cdot52}$, $Trp^{6\cdot48}$, $Trp^{3\cdot28}$, $Asp^{3\cdot32}$, $Asn^{7\cdot39}$
XGBoost	$Phe^{6\cdot52}$, $Asn^{7\cdot39}$, $Trp^{6\cdot48}$, $Trp^{3\cdot28}$, $Asp^{3\cdot32}$
K Neighbours	$Asp^{3\cdot32}$, $Asn^{7\cdot39}$, $Asn^{6\cdot55}$, $Phe^{6\cdot52}$, $Ser^{5\cdot42}$

Table 3.4: The five most important residues identified by the ML methods in the decreasing order of feature importance scores.

All the residues identified in Table 3.4 are also identified in the structural and mutagenesis data available in the GPCRDB database [14, 91]. Amongst these residues, $Phe^{6\cdot52}$, $Trp^{6\cdot48}$, $Asn^{7\cdot39}$, $Asp^{3\cdot32}$ were the four most important residues according to the above ML algorithms.

$Phe^{6\cdot52}$ participates in edge-to-face aromatic interaction with the ligands and is also coupled to the rotation of $Trp^{6\cdot48}$ [125]. $Trp^{6\cdot48}$ apart from being a transmission switch [81, 82], it is also a part of the conserved polar network with $Asp^{3\cdot32}$ and plays a major role in GPCR activation [79, 127]. $Asp^{3\cdot32}$ is a part of the 3-7 lock switch and the opening of 3-7 lock is integral to ligand binding and is one of the first molecular switch to be activated in GPCRs [82]. $Asn^{7\cdot39}$ forms a hydrogen bond with adrenaline, stabilizing the receptor conformation [94], which is consistent with prior mutagenesis experiments [87]. Similarly, the K Neighbours algorithm identified $Asn^{6\cdot55}$ as another important residue which is known to hydrogen bond with adrenaline, corroborated by mutagenesis experiments [88]. Although the identified residues are used to determine the movement of adrenaline within the binding pocket, they also play an significant role in GPCR signaling and activation. Remarkably, all these residues have been previously identified as important residues in prior studies, solidifying the significance of our method.

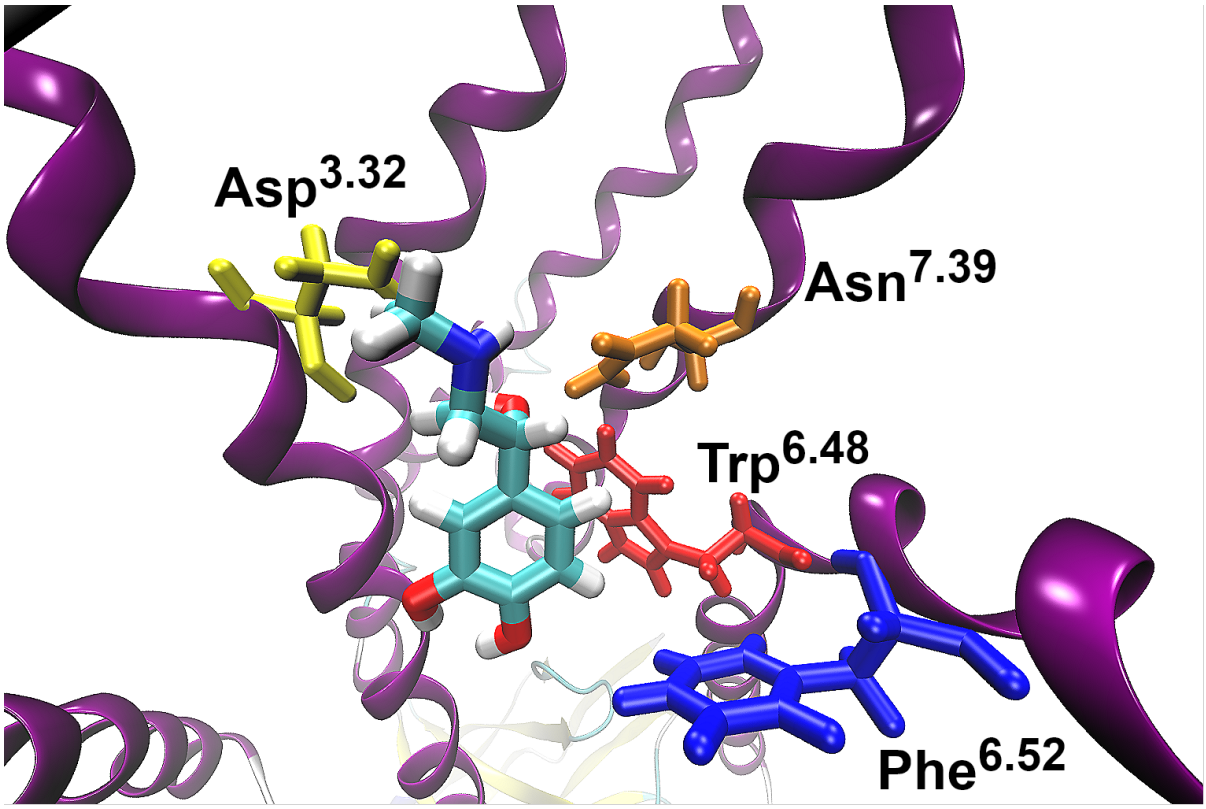


Figure 3.8: Residues Phe$^{6.52}$ (blue), Trp$^{6.48}$ (red), Asn$^{7.39}$ (orange), and Asp$^{3.32}$ (yellow) in top-view were identified as the most important residues by ML methods. Adrenaline (colored by atom name) is also shown in the Licorice representation.

3.5 Conclusion

As GPCRs are major drug targets, the medicinal and pharmaceutical importance of their interactions with drugs are widely recognized and researched. However, a detailed elucidation of the molecular mechanism of drug-induced activation or inhibition of GPCRs is non-trivial and challenging to probe experimentally. Molecular dynamics simulations have emerged as a powerful tool to investigate intricate conformational changes that underlie the activation of these receptors, thereby allowing us to gain detailed insight into the underlying molecular mechanisms and associated energetics.

In this study, we used umbrella sampling and molecular dynamics simulations to explore the dynamics of adrenaline in the orthosteric binding pocket of β_2AR. Our results revealed the existence of two distinct stable states for the adrenaline-β_2AR complex: a most-stable state and a meta-stable state that are separated by an energy barrier. Using the K-means clustering algorithm on multiple unbiased MD trajectories that sampled the energy basins corresponding to these two states, the most-probable structures corresponding to these states are determined.

The potential of mean force computed using the umbrella sampling method enabled us to quantify the relative stability of these states and the activation barriers to transit between them. We examined the conformational and orientational changes in the binding partners that occur during the transition between these two stable states.

To better understand the key interactions that govern the transitions between these states, we computed the pairwise interaction energy between adrenaline and each of the binding pocket residues of the receptor. Using statistical methods and machine learning-based regression models, the correlation between the interaction energy of individual binding pocket residues with adrenaline and the collective variables that describe the transition was examined. Correlation analysis showed that the residues Phe$^{6.52}$, Asn$^{6.55}$, Trp$^{6.48}$, Ser$^{5.43}$ and, Ser$^{5.42}$ showed reasonable correlation with movement of adrenaline within the binding pocket. The results from machine learning-based analysis revealed that the residues Phe$^{6.52}$, Trp$^{6.48}$, Asn$^{7.39}$, and Asp$^{3.32}$ contribute to the motion of adrenaline inside the binding pocket and also govern state transition. Interestingly, all these residues are known to play a critical role in GPCR signaling and activation.

In summary, the present work showcases the effectiveness of combining molecular dynamics and enhanced sampling simulations with machine learning-based statistical methods to characterize the orientational and conformational dynamics of drugs in the binding pockets of GPCRs. The results and the methodology reported here can provide a basis for an automated pipeline for molecular-level investigation of ligand dynamics and drug-GPCR interactions and a deeper understanding of the complex and intricate mechanisms involved in GPCR activation.

3.6 Data and Software Availability

The AMBER software used to perform MD simulations is available at `https://ambermd.org/GetAmber.php`. The CHARMM-GUI software used to build the membrane models of the receptor for input to MD simulations is available at `https://www.charmm-gui.org/?doc=input`. The AMBER input files: coordinates, and parameter files used for the current study is provided in the Supporting Information. The VMD software used for visualization is available open-source at `https://www.ks.uiuc.edu/Research/vmd/`. The protein structures of the adrenaline-free (PDB ID: 2RH1) and adrenaline-bound (PDB ID: 4LDO) β_2AR are retrievable free of charge from the Protein Data Bank (`https://www.rcsb.org/`). The GPCRDB database used for BW numbering and identifying interacting residues is available at `https://gpcrdb.org/`. The dataset and code used for correlation and machine learning-based regression analysis is made publicly available at `https://github.com/K7S3/GPCR_ML_Residue_Importance`.

3.7 Supporting Information Available

- Time-series of d and θ (Figure S1), Time-series of δ_1, δ_2, and δ_3 (Figure S2), histogram demonstrating overlap between umbrella sampling windows (Figure S3), comparison with trajectories in GPCRmd database (Figure S4), centroid structures of all 10 clusters obtained from K-means clustering (Figure S5), and brief overview on the AMBER input and parameter files included (PDF)
- AMBER coordinates, structures, parameters, restart, and input files, cluster average and centroid structures (`b2ar_ale_dynamics.zip`)

This material is available in Appendix: Supporting Information Chapter and is also available free of charge at `https://pubs.acs.org/doi/10.1021/acs.jcim.3c00401`.

Chapter 4

Unraveling G-Protein Activation and GDP binding using Molecular Dynamics Simulations and Umbrella Sampling

Contents

4.1 Introduction

As we had previously discussed in Chapter 2, the G-Protein Coupled Receptors (GPCRs) are seven trans-membrane (7 TM) that play a crucial role in cellular signaling and are responsible for vital physiological functions, including taste, vision, olfaction, and neuro-transmission. They are coupled to the intracellular heterotrimeric G-Protein and also serve as primary drug targets for around one-third of all the drugs produced. In this study we focus on agonist induced dynamic and energetic changes with G-Protein. The binding of an agonist induces conformational changes in the GPCR, and activates the GPCR. Once the GPCR is activated, allosteric transition is forced wherein the intracellular G-Protein gets activated and is forced to leave the GPCR. The intracellular G-Protein consists of three sub-units $G\alpha$, $G\beta$, and $G\gamma$. A guanosine di-phosphate molecule (GDP) is bound to the α sub-unit in the inactive state of G-Protein. When the G-Protein gets activated, conformational changes occur in the G-Protein, favoring the dissociation of the GDP molecule in exchange for guanosine tri-phosphate (GTP). The cell typically maintains a GTP:GDP ratio of 10:1 in the cytoplasm, ensuring a GTP molecule is readily available for this exchange. Moreover, the binding of the GTP molecule enables conformational changes that allow the dissociation of the active Gα sub-unit (bound to GTP) from

the G$\beta\gamma$ dimer and the GPCR. Upon dissociation, the Gα and the G$\beta\gamma$ dimer bind to different effector proteins for the downstream signaling cascade.

In this study, we model the active state $\beta_2 AR$ bound to adrenaline and G_s protein. Further, we employ molecular dynamics, and umbrella sampling to study the dynamical and energetic changes involved in the adrenaline-induced $\beta_2 AR$ activation and the subsequent G_s activation. Additionally, we also study the dissociation of G_s protein from the $\beta_2 AR$ and the dissociation of GDP from $G\alpha$. We also build end state models from completely inactive end states of $\beta_2 AR$-G_s complex, to their completely active end states in the presence of adrenaline and GDP/GTP molecules.

4.2 Methods

4.2.1 Models

4.2.1.1 GDP Unbound Model

The crystal structures of the unbound inactive (PDB ID: 2RH1)[93] and the G-Protein bound (PDB ID: 3SN6)[129] states of the β_2 adrenergic receptor were used as initial configurations for molecular dynamics simulations. These receptors were inserted into the pre-equilibrated phosphatidylcholine (POPC) lipid bilayer using the OPM database[97] and CHARMM-GUI software[98–102]. The total numbers of POPC molecules in the lipid bilayers of the G-Protein bound and unbound $\beta_2 AR$ models are 383 and 319, respectively. The Amber Lipid17[103] and AMBER19SB[104] force fields were used for the lipid bilayer and the receptors, respectively. The Generalized Amber Force Field (GAFF)[105, 106] with the AM1-BCC[107] partial charges for the adrenaline and the TIP3P[108] model for water were used. A suitable number of sodium and chloride ions were added to the systems to maintain a salt concentration of 0.15 M. The box-dimensions for the resulting models were 110 Å x 110 Å x 140 Å with 36331 water molecules for the inactive state and 120 Å x 120 Å x 160 Å with 53622 water molecules for the active state.

4.2.1.2 GDP Bound Model

Two different methods were used to build the GDP bound models: 1) Docking GDP to G-Protein at its binding site. 2) Use structure alignment with existing PDB structures.

4.2.1.3 Docking Based Model

The GDP binding region on the $G\alpha$ of G-Protein is determined by guanine recognition motif, a P-loop that envelops the a and b phosphates of GTP, and two dynamic structural elements

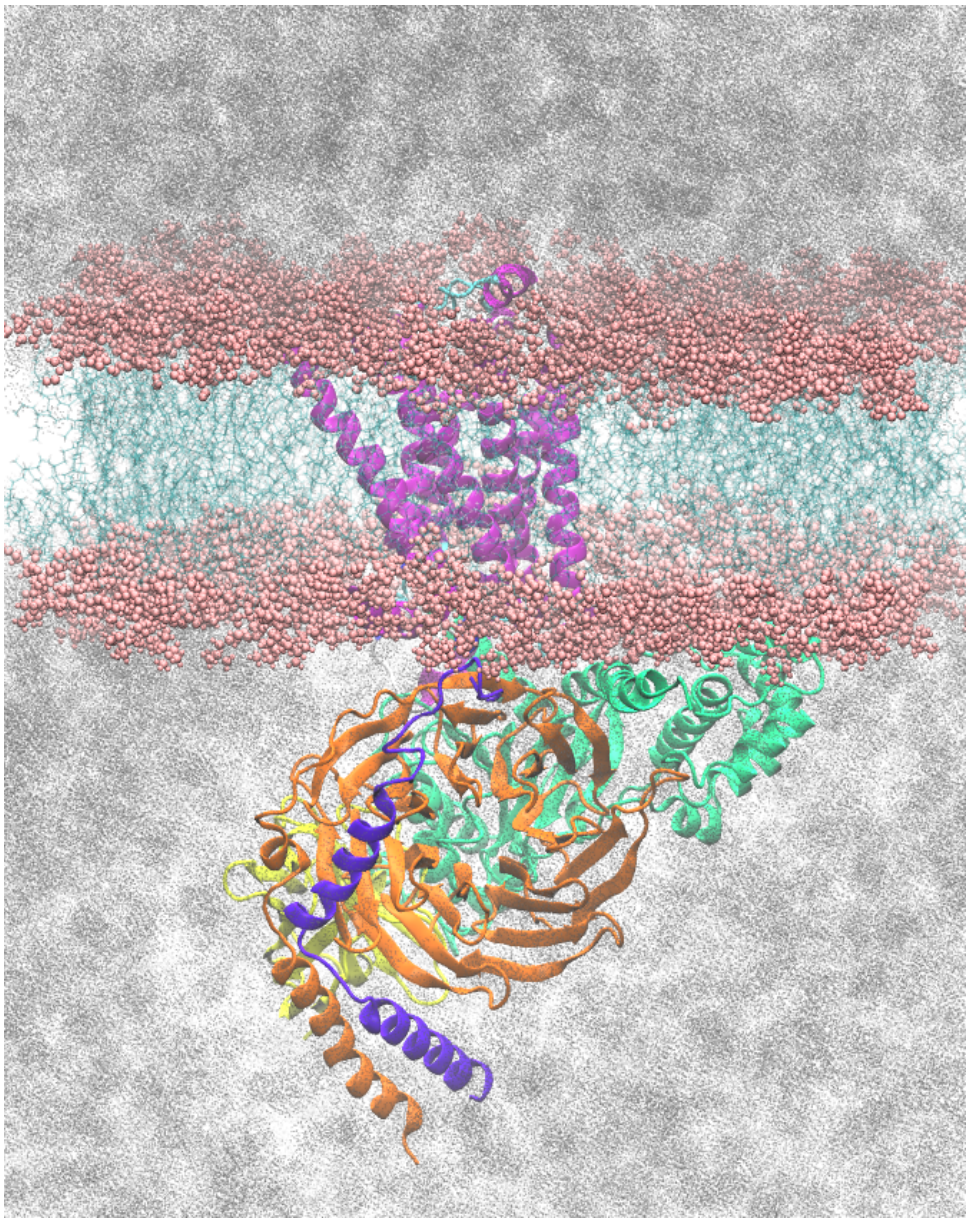


Figure 4.1: A snapshot of β_2 adrenergic receptor bound to G-Protein (based on PDB ID: 3SN6) embedded in a hydrated POPC lipid bilayer membrane with one head-group PC shown in red and two tail-groups OL and PA shown in green. Water molecules are shown in grey color.

called switch I and switch II [130]. Following which, three residues were identified to estimate the docking site for the GDP on the $G\alpha$ as shown in figure 4.2.

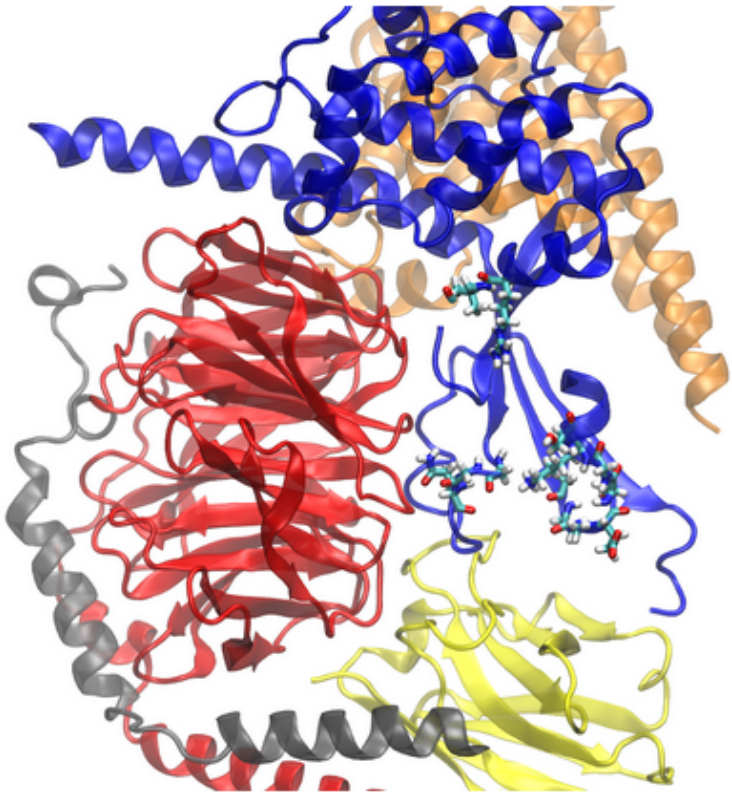


Figure 4.2: Binding Site Determination for GDP

The geometric center of the three residues was calculated and a bounding box of 20x20x20Å was used for docking with the SMINA software package[131]. Subsequently, nine docked poses for the GDP in the binding pocket were produced and the one with the best score to be used as the starting structure for further simulations. However, docking based method for determining ligand pose is used only when the crystal structure for the bound state of the ligand is unknown. In our case, we have both the GDP and GTP bound states available, and therefore we used structural alignment to model the states.

4.2.1.4 Structure Alignment Based Model

To model the combined adrenaline-β_2AR-G-Protein-GDP complex, we combined three different protein structures with PDB IDs: 4LDO, 3SN6, and 6EG8 based on structure alignment of the common β_2AR in the above files. The VMD tool, MultiSeq was used for structure alignment. The combined structure was modified to include adrenaline, β_2AR, G-Protein, and

GDP. The combined model was then embedded in POPC bilayer and was used as input for CHARMM-GUI.

4.2.2 Molecular Dynamics Simulation

The potential energies of these membrane models were minimized using steepest gradient descent for 500000 cycles and conjugate gradient descent for another 500000 cycles[111, 112]. The energy minimized systems were gradually heated to 303.15 K at 1 atm pressure and equilibrated there for 5 ns, followed by a production run of 50 ns under constant temperature and constant pressure conditions. The temperature was controlled using the Langevin thermostat[113], and the pressure was controlled using the anisotropic Berendsen barostat[114]. SHAKE algorithm[115] was used to constrain bonds involving hydrogen atoms. The particle mesh Ewald[116] approach was used for treating the electrostatic interactions with a real-space cutoff distance of 10 Å. The equations of motion were integrated using the velocity Verlet algorithm[117] with a time step of 2 fs. All the simulations, including the enhanced sampling simulations, were done using the AMBER18 software[118]. All the trajectory analyses were carried out using the Amber CPPTRAJ[119] and VMD package[120].

4.2.3 Umbrella Sampling

Umbrella sampling simulations were performed on the G-Protein bound β_2AR systems in the absence of GDP with different collective variables to study G-Protein dissociation from GPCR. Similarly, US simulations were performed on G-Protein bound β_2AR system in the presence of GDP to study GDP dissociation.

4.2.3.1 Collective Variables

1. C_{d1}: Distance (d_1) between the $C\alpha$ of residue Y219 of β_2AR and the Y391 $C\alpha$ of residue of G-protein.

2. C_{d2}: Distance (d_2) between the GDP's phosphate groups and the backbone of Lys52G.H1.1-Thr54G.H1.3 in the Gαq subunit.

The collective variables, d_1 was varied from 4.0Å to 30.0Å in steps of 0.5Å resulting in a total of 53 windows and d_2 was varied from 1.0Å to 30.0Å in step of 0.5Å resulting in a total of 59 windows.

Each US simulation (6 ns trajectory per window) was performed in the NPT ensemble, under the same conditions as those of unbiased MD runs. The spring constant used for the harmonic biasing potential was 10 $Kcal/mol/Å^2$ for production and 200 $Kcal/mol/Å^2$ for minimization.

The Weighted Histogram Analysis Method (WHAM) (`membrane.urmc.rochester.edu/?page_id=126`) was used to determine the potential of mean force (PMF) from CV trajectories obtained from the US simulations. The last 5 ns of the umbrella sampling trajectory per window was used for the WHAM.

4.2.3.2 Modeling End States

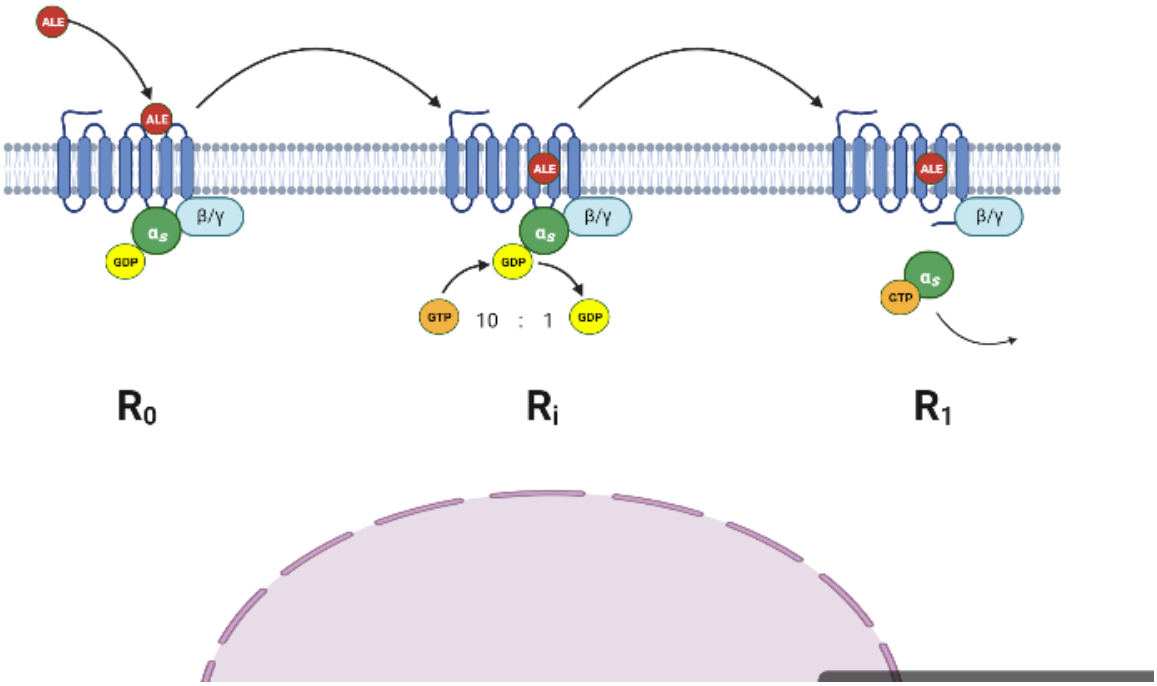


Figure 4.3: The figure shows the three states, R_0, R_i, and R_1, depicting the GPCR and G-Protein activation pathway. Created with `BioRender.com`

Model R_0 was prepared by structural alignment of 4 PDB files with PDB IDs: 2RH1, 4LDO, 3SN6, and 6EG8 to include inactive state β_2AR bound to GDP-bound G_s. Adrenaline was translated 25Å in the z-axis from its orthosteric binding site, placing it directly above the extracellular loop 2 (ECL2), one of the first point of contacts for adrenaline during GPCR activation. Model R_i was prepared by structural alignment of 3 PDB files with PDB IDs: 4LDO, 3SN6, 6EG8 to include active state β_2AR with adrenaline bound at its orthosteric binding pocket and GDP-bound G_s.

To prepare the models R_0 and R_i for NEB, the inactive and active β_2AR were sequence-aligned to ensure 100% similarity. The missing residues in the sequence-aligned models of inactive and active β_2AR were modelled to a target sequence obtained from the gene encoding data of β_2AR using the online SWISS-MODEL software. Furthermore, the models obtained were subjected to 500000 cycles of steepest gradient minimization and 500000 cycles of conjugate gradient minimization.

Model R_1 was prepared by structural alignment of 4 PDB files with PDB IDs: 4LDO, 3SN6, 1AZT, 1XHM to include adrenaline-bound active β_2AR, GTP-bound $G\alpha$ and $G\beta\gamma$ sub-units.

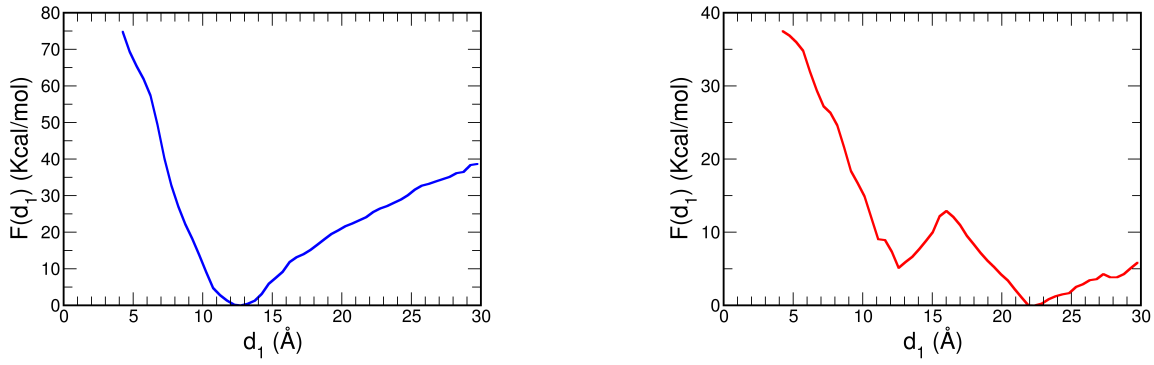
The final models R0, Ri, and R1 contained a total of 8459 protein and ligand atoms. These receptors were inserted into the pre-equilibrated phosphatidylcholine (POPC) lipid bilayer using the OPM database[97] and CHARMM-GUI software[98–102]. The total numbers of POPC molecules in the models were 536 molecules. The 79487 water molecules and 536 POPC molecules. The Amber Lipid17[103] and AMBER19SB[104] force fields were used for the lipid bilayer and the receptors, respectively. The Generalized Amber Force Field (GAFF)[105, 106] with the AM1-BCC[107] partial charges for the adrenaline and the TIP3P[108] model for water were used. A suitable number of sodium and chloride ions were added to the systems to maintain a salt concentration of 0.15 M. The box-dimensions for the resulting models were 150 Å x 150 Å x 176 Å with 79487 water molecules. The total number of atoms were same across R0, Ri and R1, and amounted to 317346 atoms.

4.3 Results and Discussion

4.3.0.1 G-Protein Dissociation

The potential of mean force (PMF) was derived from the umbrella sampling simulations of dissociating G-Protein from β_2AR on the GDP-free (PDB ID: 3SN6) and GDP-bound active-state β_2AR-G_s models. The PMF on the GDP-free model indicates a global energy minimum of $d_1 = 12.6$Å and the PMF on the GDP-bound model indicates a local energy minimum at $d_1 = 12.6$Å and a global energy minimum at $d_1 = 22.2$Å as shown in figure 4.4. The value of d calculated from the GDP-free crystal structure of the G-Protein bound β_2AR (PDB ID: 3SN6) is 12.2Å. There is a steep increase in the free energy F(d) as the G-Protein moves closer in the binding pocket. However, the same pattern is not observed as the value of d increases. The free energy increases somewhat similarly as the value of d_1 increases till $d_1 = 16.3$Å for the GDP-free model. Afterwards, the rate of increase in the free energy decreases and the graph plateaus at higher values of d at around $d_1 \sim 30$Å for the GDP-free model.

In the GDP-bound model, we can observe two stable states of the system, a meta-stable state at $d_1 = 12.6$Å and the most-stable state at $d = 22.2$Å. There is a local maximum



(a) (b)

Figure 4.4: Free Energy Profile of G-Protein Unbinding (a) in the absence of GDP (b) in the presence of GDP

at $d = 16.0$Å and the energy barrier to cross from the meta-stable state to the most-stable state is around 8 kcal/mol and the free energy required for the reverse transition is around 13 kcal/mol. Moreover, the rate of increase in the free energy F(d) for the GDP-bound model as the G-Protein moves closer is almost half that of GDP-free model for the same. The binding of GDP appears to have reduced the overall energy of the system, and also has created a more stable state further away from the G-Protein's original binding site. As the value of d increases greater than $d = 22.2$Å, there is very minimal increase in the free energy. This indicates that the binding of GDP to the G-Protein, favours the dissociation of G-Protein from $\beta_2 AR$.

4.3.0.2 GDP Dissociation

Similarly, the potential of mean force derived from the umbrella sampling simulation of the GDP bound active $\beta_2 AR$-G_s model (R_i) for dissociating the GDP from the binding pocket of Gα indicates a global energy minimum at d_2=5.2Åas shown in figure 4.5. Experimentally, the free energy required to dissociate GDP from it's binding pocket is around 10 kcal/mol. This can also be seen in the PMF plot, the free energy increases gradually till 10 kcal/mol and reaches a maximum of around 30 kcal/mol at $d_2 \sim 20$Å. Moreover, a plateau is observed in the free energy as d_2 is increased beyond 20Å. A very steep increase in free energy F(d_2) is observed when GDP is brought closer within the $G\alpha$ sub-unit.

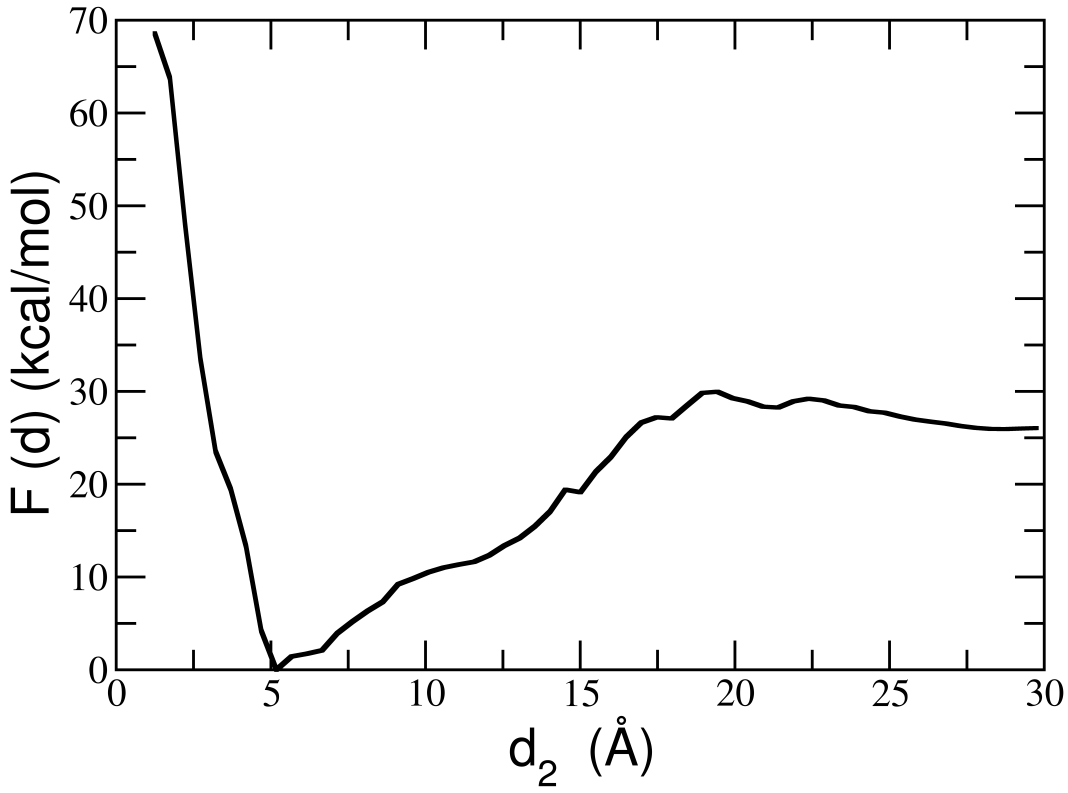


Figure 4.5: Free Energy Profile of GDP unbinding.

4.4 Conclusion

In conclusion, the potential of mean force (PMF) calculations obtained from umbrella sampling simulations provide valuable insights into the dissociation process of G-Protein from the $\beta_2 AR$ receptor in both GDP-free and GDP-bound states. In the GDP-free model, the PMF shows a steep increase in free energy as the G-Protein approaches the binding pocket, with a plateau observed at higher distances. On the other hand, in the GDP-bound model, two stable states are identified, with a local energy maximum between them. The binding of GDP appears to reduce the overall energy and stabilizes the G-Protein in a more distant position. The PMF analysis suggests that the binding of GDP favors the dissociation of G-Protein from the $\beta_2 AR$ receptor. Additionally, the PMF calculations for GDP dissociation from the Gα sub-unit indicate a global energy minimum at a certain distance, with a steep increase in free energy as GDP approaches the binding pocket. These findings enhance our understanding of

the complex dynamics involved in G-Protein dissociation and provide important information for drug design targeting the $\beta_2 AR$ receptor.

Furthermore, three models (R0, Ri, R1) with equal number of atoms were built by combining crystal structures of $\beta_2 AR$, G-Protein, and ligands to model inactive, intermediate and active state of $\beta_2 AR$-G_s systems. These models were further inserted into lipid bilayers, and provide end points which can be used for further Nudged Elastic Band (NEB) calculations. A future study using the NEB simulations could provide insights into GPCR and G-Protein signaling by analyzing the energy landscape and transition pathways, aiding our understanding of both the GPCR and G-Protein activation and signaling dynamics.

Chapter 5

Conclusion

In this thesis, we have employed a combination of molecular dynamics (MD) simulations, enhanced sampling techniques, statistical analysis, and machine learning-based regression models to gain insights into the intricate mechanisms underlying drug-G protein-coupled receptor (GPCR) interactions and G-protein dissociation from the $\beta_2 AR$ receptor. Our findings shed light on the conformational changes, energetics, and key residues involved in these processes, providing a deeper understanding of GPCR activation and signaling dynamics.

In Chapter 3, we focused on investigating the dynamics of adrenaline in the orthosteric binding pocket of the $\beta_2 AR$ receptor using umbrella sampling and MD simulations. Our results revealed the existence of two distinct stable states for the adrenaline-$\beta_2 AR$ complex, separated by an energy barrier. Through K-means clustering and analysis of multiple unbiased MD trajectories, we identified the most-probable structures corresponding to these states. By quantifying the relative stability and activation barriers between these states using the potential of mean force, we gained valuable insights into the energetics of the transition. Moreover, we examined the conformational and orientational changes occurring in the binding pocket during this transition.

To better understand the key interactions governing the transition between the stable states, we computed the pairwise interaction energy between adrenaline and each binding pocket residue of the receptor. Employing statistical methods and machine learning-based regression models, we established correlations between the interaction energy of individual residues and the collective variables describing the transition. Notably, residues Phe$^{6.52}$, Trp$^{6.48}$, Asn$^{7.39}$, and Asp$^{3.32}$ emerged as crucial contributors to the motion of adrenaline within the binding pocket and the overall state transition. Significantly, these residues are known to play critical roles in GPCR signaling and activation, highlighting their importance in the dynamics of ligand binding and receptor activation.

In Chapter 4, our focus shifted to the dissociation process of the G-protein from the $\beta_2 AR$ receptor in both GDP-free and GDP-bound states. Through potential of mean force (PMF)

calculations obtained from umbrella sampling simulations, we gained valuable insights into the energy landscape and stability of the G-protein dissociation process. In the GDP-free model, the PMF exhibited a steep increase in free energy as the G-protein approached the binding pocket, with a plateau observed at higher distances. Conversely, in the GDP-bound model, two stable states were identified, separated by a local energy maximum. Interestingly, the binding of GDP appeared to reduce the overall energy and stabilize the G-protein in a more distant position, favoring its dissociation from the $\beta_2 AR$ receptor. Furthermore, PMF calculations for GDP dissociation from the Gα subunit revealed a global energy minimum at a certain distance, followed by a steep increase in free energy as GDP approached the binding pocket.

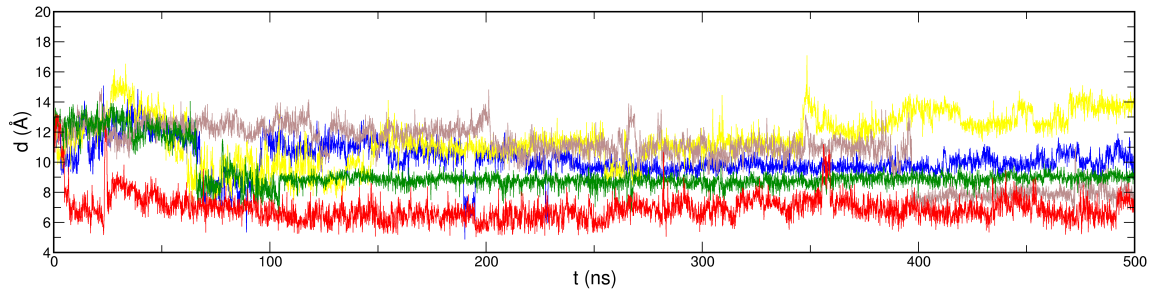
To further enhance our understanding of GPCR and G-protein signaling, we constructed three models representing the inactive, intermediate, and active states of the $\beta_2 AR$-G_s system. These models, built by combining crystal structures of the $\beta_2 AR$, G-protein, and ligands, were inserted into lipid bilayers and can serve as endpoints for future Nudged Elastic Band (NEB) calculations. We note that further NEB simulations based on the above states may hold the potential to provide detailed insights into the energy landscape and transition pathways involved in GPCR and G-protein signaling, offering a comprehensive understanding of their activation and signaling dynamics.

Overall, this thesis demonstrates the effectiveness of combining MD simulations, enhanced sampling techniques, statistical analysis, and machine learning approaches to unravel the complex dynamics of ligand-receptor interactions and GPCR signaling. By providing detailed insights into the conformational changes, energetics, and key residues involved, our study contributes to the broader understanding of GPCR activation and the development of novel therapeutics targeting GPCRs. The methodologies and findings presented here lay the foundation for an automated pipeline for investigating ligand dynamics and drug-GPCR interactions, fostering future research in this field.

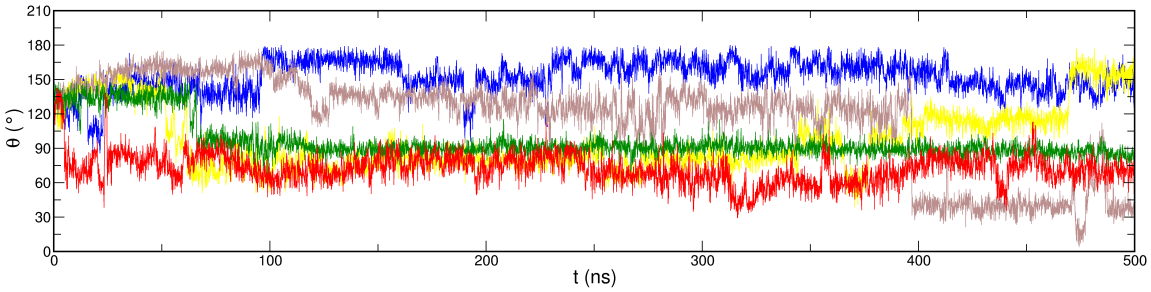
Appendix A

Supporting Information for Chapter 3

A.1 Timeseries of d and θ



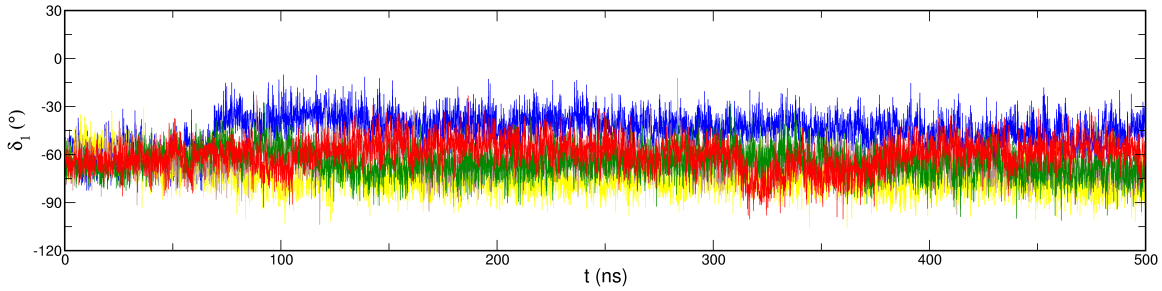
(a)



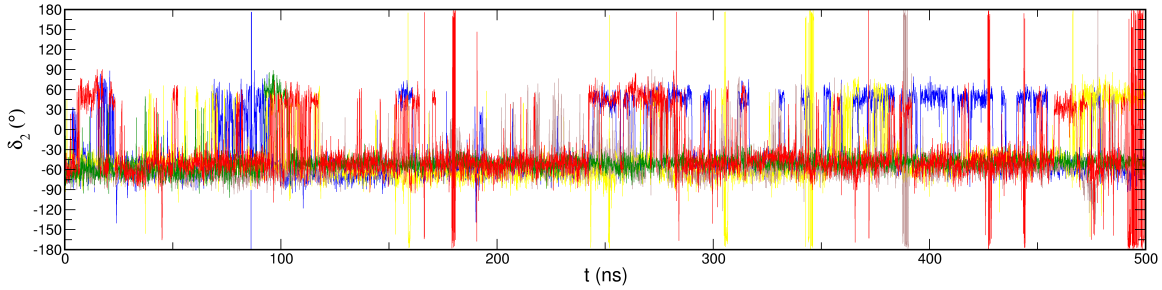
(b)

Figure A.1: Time series of (a) d (b) θ obtained from the 500 ns unbiased MD trajectories (TJ1-5) represented by colors blue, yellow, brown, green and red respectively

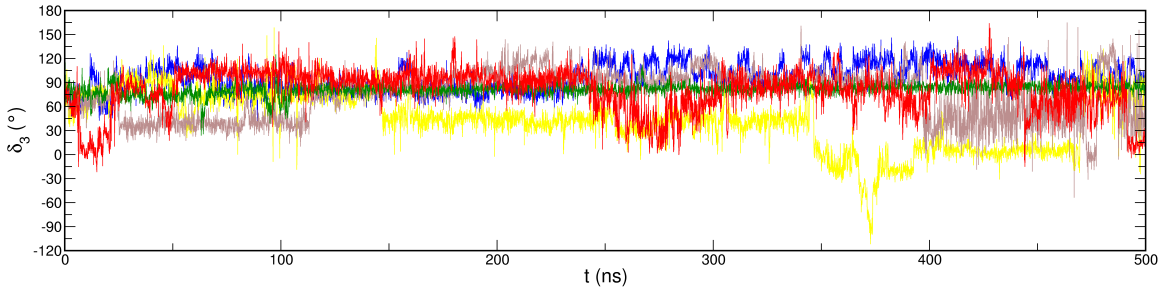
A.2 Time series of δ_1, δ_2, and δ_3



(a)



(b)



(c)

Figure A.2: Time series of (a) δ_1 (b) δ_2 (c) δ_3 obtained from the 500 ns unbiased MD trajectories (TJ1-5) represented by colors blue, yellow, brown, green and red respectively

A.3 Sufficiency of Umbrella Sampling Trajectories

The histogram A.3 shows that there is sufficient overlap between any two consecutive windows.

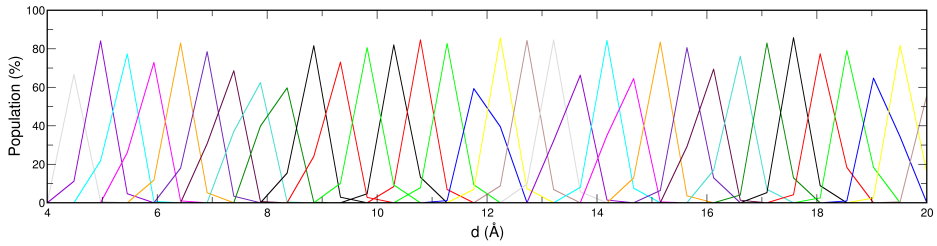


Figure A.3: Histograms of 33 Umbrella Sampling windows of collective variable d varying from 4 Å to 20 Å.

A.4 Comparison with GPCRmd Trajectories

Here, we compared our unbiased MD trajectories of adrenaline-bound active state MD trajectories with the simulations in the GPCRmd database. However, it is important to acknowledge the distinctions between our model and the one employed in GPCRmd. Our adrenaline-bound active β_2AR model contains additional components, namely T4-lysozyme and a G-Protein mimetic nanobody, which allows for the stabilization of the receptor in its active state. Furthermore, our simulations employ a 2 fs timestep and utilize the AMBER force field in conjunction with the AMBER/18 software, while the trajectories in GPCRmd (`https://submission.gpcrmd.org/view/117/`) employ a 4 fs timestep and the CHARMM 36M force field with ACEMD software.

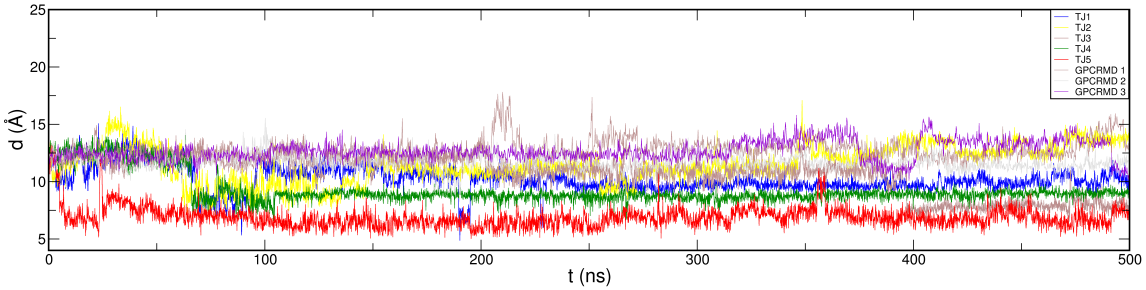
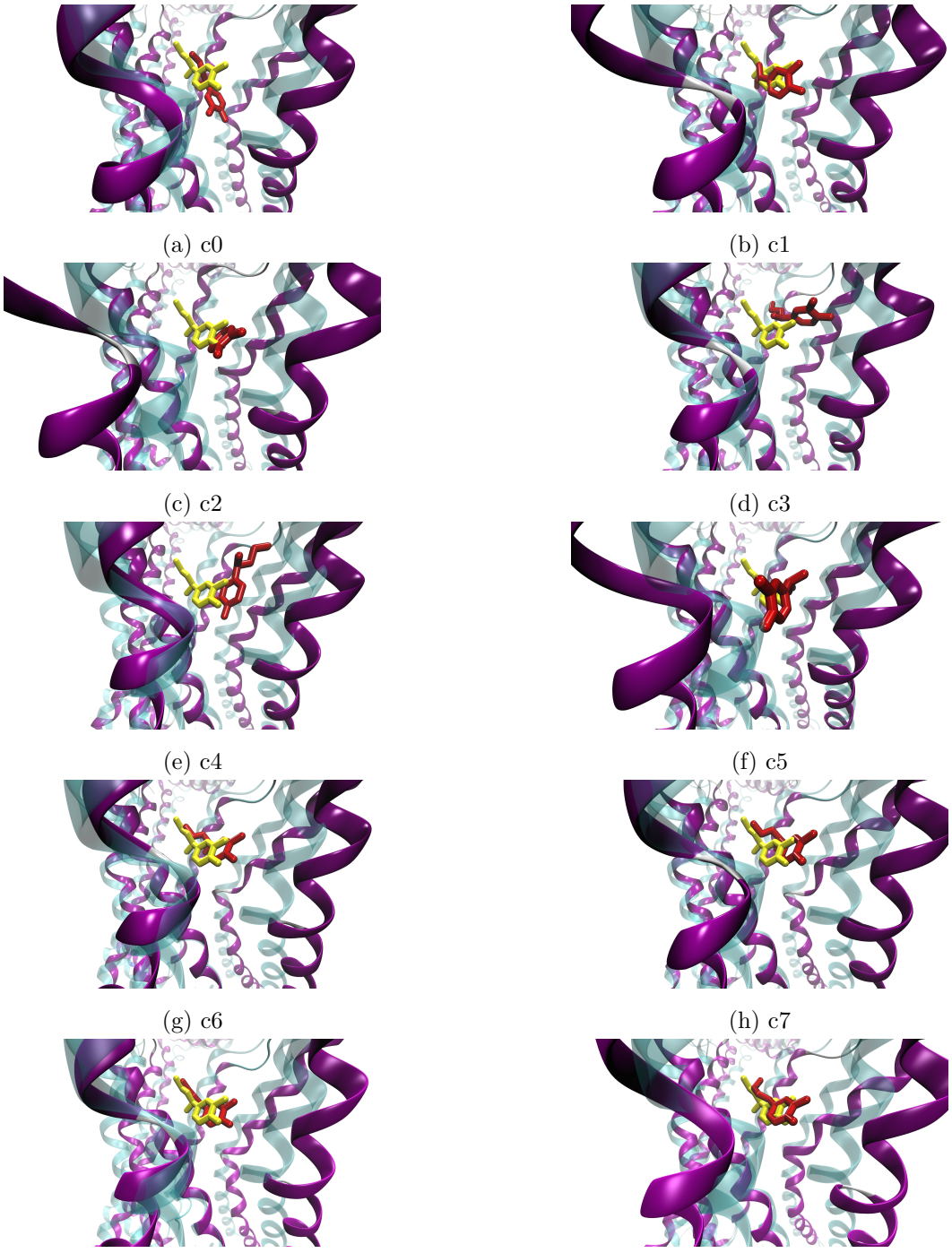


Figure A.4: Timeseries of d of adrenaline-bound active β_2AR simulations (TJ1-5) and the 3 trajectories available in GPCRmd database.

Given these differences, a direct comparison between our results and those from GPCRmd is limited. Nevertheless, we have compared the time-series of d in both our trajectories and the GPCRmd and found that the minimum value of d was 8.27Å (refer Figure A.4). This value is higher than the meta-stable d value, 6.8 Å. However, it is to be noted that the PMF plots for the same collective variable d could be different based on the nature of the intracellular binding partner [110].

A.5 K-Means Clustering Analysis

Structures obtained through K-Means Clustering Analysis with K=10 clusters.



(a) c0 (b) c1

(c) c2 (d) c3

(e) c4 (f) c5

(g) c6 (h) c7

(i) c8 (j) c9

Figure A.5: Figures a-j represent the representative median structures of clusters c0-c9 obtained from the K-means clustering algorithm (magenta) aligned on the β_2AR-adrenaline crystal structure (PDB ID: 4LDO, cyan). Adrenaline is colored in red in the clusters and its aligned crystallographic poses are colored in yellow.

A.6 Files Included

The following files related to the study of adrenaline dynamics in $\beta_2 AR$ are included in the `b2ar_ale_dynamics.zip` file available at `https://pubs.acs.org/doi/10.1021/acs.jcim.3c00401`. The `b2ar_ale_dynamics.zip` contains three directories `Models`, `AMBER_Input_Files`, and `Clusters`.

A.7 Models

The `Models` directory contains the coordinates, structures, parameters and restart files obtained after embedding the $\beta_2 AR$ in POPC bi-layer obtained using CHARMM-GUI. It contains two sub-directories `2rh1` (adrenaline-free model, corresponding to PDB ID: 2RH1) and `4ldo` (adrenaline-bound model, corresponding to PDB ID: 4LDO).

In the sub-directory, `2rh1`, the following files are included:

1. `2rh1_membrane_model.pdb` (Coordinate File for unbiased MD)
2. `2rh1_membrane_model.psf` (Structure File for unbiased MD)
3. `2rh1_membrane_model.rst7` (AMBER Restart File for further simulations)
4. `2rh1_membrane_model.parm7` (AMBER Parameter File)

In the sub-directory, `4ldo`, the following files are included:

1. `4ldo_membrane_model.pdb` (Coordinate File for unbiased MD)
2. `4ldo_membrane_model.psf` (Structure File for unbiased MD)
3. `4ldo_membrane_model.rst7` (AMBER Restart File for further simulations)
4. `4ldo_membrane_model.parm7` (AMBER Parameter File)
5. `complex.pdb` (Coordinate File for Umbrella Sampling Runs)

A.8 AMBER Input Files

The AMBER minimization, heating, equilibration and production input files for unbiased MD simulations is included. Similar protocol was followed for Umbrella Sampling (US) simulations with harmonic restraints whose restraint files are included.

1. `01_Min.in` (Input File for Minimization)

2. `02_Heat.in` (Input File for Heating till 100K)

3. `03_Heat.in` (Input File for Heating from 100K to 303K)

4. `04_Hold.in` (Input File for 5 ns Equilibration run)

5. `05_Prod.in` (Input File for 50 ns Production run)

6. `rest_min` (Restraint File for US Minimization)

7. `rest_prod` (Restraint File for US Equilibration and Production)

A.9 Clusters

1. `rep.`c_{0-9}`.pdb` (Contains the representative structures (cluster centroid) for clusters co to c9)

2. `avg.`c_{0-9}`.pdb` (Contains the average structures for clusters co to c9)

Related Publications

- Keshavan Seshadri, Marimuthu Krishnan, Molecular Dynamics and Machine Learning Study of Adrenaline Dynamics in the Binding Pocket of GPCR *Journal of Chemical Information and Modeling*, Article ASAP, Published July 6th, 2023, doi: 10.1021/acs.jcim.3c00401
  URL: `https://doi.org/10.1021/acs.jcim.3c00401`

- Keshavan Seshadri, Peng Liu, and David Ryan Koes, The 3Dmol.js Learning Environment: A Classroom Response System for 3D Chemical Structures, Journal of Chemical Education 2020 97 (10), 3872-3876, doi: 10.1021/acs.jchemed.0c00579
  URL: `https://doi.org/10.1021/acs.jchemed.0c00579`

- Keshavan Seshadri, Marimuthu Krishnan, Unraveling G-Protein Activation and GDP binding using Molecular Dynamics Simulations and Umbrella Sampling, July 2023 (Manuscript in Preparation)

Bibliography

[1] D.L. Nelson and M. Cox. *Lehninger Principles of Biochemistry: International Edition*. Macmillan Learning, 2021. ISBN 9781319381493. URL https://books.google.co.in/books?id=NYoJzgEACAAJ.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. W.W. Norton & Company, August 2017. doi: 10.1201/9781315735368. URL https://doi.org/10.1201/9781315735368.

[3] Bernard J. Venn. Macronutrients and human health for the 21st century. *Nutrients*, 12(8):2363, August 2020. doi: 10.3390/nu12082363. URL https://doi.org/10.3390/nu12082363.

[4] FRANCIS CRICK. Central dogma of molecular biology. *Nature*, 227(5258):561–563, August 1970. doi: 10.1038/227561a0. URL https://doi.org/10.1038/227561a0.

[5] J.E. Hall. *Guyton and Hall Textbook of Medical Physiology*. Guyton Physiology. Elsevier Health Sciences, 2015. ISBN 9781455770052. URL https://books.google.co.in/books?id=3sWNCgAAQBAJ.

[6] Daniel M. Rosenbaum, Søren G. F. Rasmussen, and Brian K. Kobilka. The structure and function of g-protein-coupled receptors. *Nature*, 459(7245):356–363, May 2009. doi: 10.1038/nature08144. URL https://doi.org/10.1038/nature08144.

[7] Lisa A. Catapano and Husseini K. Manji. G protein-coupled receptors in major psychiatric disorders. *Biochim. Biophys. Acta, Biomembr.*, 1768(4):976–993, April 2007. doi: 10.1016/j.bbamem.2006.09.025. URL https://doi.org/10.1016/j.bbamem.2006.09.025.

[8] Otavio Cabral-Marques, Alexandre Marques, Lasse Melvær Giil, Roberta De Vito, Judith Rademacher, Jeannine Günther, Tanja Lange, Jens Y. Humrich, Sebastian Klapa, Susanne Schinke, Lena F. Schimke, Gabriele Marschner, Silke Pitann, Sabine Adler, Ralf Dechend, Dominik N. Müller, Ioana Braicu, Jalid Sehouli, Kai Schulze-Forster, Tobias

Trippel, Carmen Scheibenbogen, Annetine Staff, Peter R. Mertens, Madlen Löbel, Justin Mastroianni, Corinna Plattfaut, Frank Gieseler, Duska Dragun, Barbara Elizabeth Engelhardt, Maria J. Fernandez-Cabezudo, Hans D. Ochs, Basel K. al Ramadi, Peter Lamprecht, Antje Mueller, Harald Heidecke, and Gabriela Riemekasten. GPCR-specific autoantibody signatures are associated with physiological and pathological immune homeostasis. *Nat. Commun.*, 9(1), December 2018. doi: 10.1038/s41467-018-07598-9. URL `https://doi.org/10.1038/s41467-018-07598-9`.

[9] Andrew L. Hopkins and Colin R. Groom. The druggable genome. *Nat. Rev. Drug Discovery*, 1(9):727–730, September 2002. doi: 10.1038/nrd892. URL `https://doi.org/10.1038/nrd892`.

[10] Krishna Sriram and Paul A. Insel. G protein-coupled receptors as targets for approved drugs: How many targets and how many drugs? *Mol. Pharmacol.*, 93(4):251–258, January 2018. doi: 10.1124/mol.117.111062. URL `https://doi.org/10.1124/mol.117.111062`.

[11] Alexander S. Hauser, Misty M. Attwood, Mathias Rask-Andersen, Helgi B. Schiöth, and David E. Gloriam. Trends in GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discovery*, 16(12):829–842, October 2017. doi: 10.1038/nrd.2017.178. URL `https://doi.org/10.1038/nrd.2017.178`.

[12] The Nobel Prize in Chemistry 2012 — nobelprize.org. `https://www.nobelprize.org/prizes/chemistry/2012/press-release/`, 2012. [Accessed 02-Jun-2023].

[13] Helgi B. Schiöth and Robert Fredriksson. The GRAFS classification system of g-protein coupled receptors in comparative perspective. *General and Comparative Endocrinology*, 142(1-2):94–101, May 2005. doi: 10.1016/j.ygcen.2004.12.018. URL `https://doi.org/10.1016/j.ygcen.2004.12.018`.

[14] Vignir Isberg, Chris de Graaf, Andrea Bortolato, Vadim Cherezov, Vsevolod Katritch, Fiona H. Marshall, Stefan Mordalski, Jean-Philippe Pin, Raymond C. Stevens, Gerrit Vriend, and David E. Gloriam. Generic GPCR residue numbers – aligning topology maps while minding the gaps. *Trends in Pharmacological Sciences*, 36(1):22–31, January 2015. doi: 10.1016/j.tips.2014.11.001. URL `https://doi.org/10.1016/j.tips.2014.11.001`.

[15] Julia K. Archbold, Jack U. Flanagan, Harriet A. Watkins, Joseph J. Gingell, and Debbie L. Hay. Structural insights into RAMP modification of secretin family g protein-coupled receptors: implications for drug development. *Trends in Pharmacological Sciences*, 32(10):591–600, October 2011. doi: 10.1016/j.tips.2011.05.007. URL `https://doi.org/10.1016/j.tips.2011.05.007`.

[16] Christoph Parthier, Steffen Reedtz-Runge, Rainer Rudolph, and Milton T. Stubbs. Passing the baton in class b GPCRs: peptide hormone activation via helix induction? *Trends in Biochemical Sciences*, 34(6):303–310, June 2009. doi: 10.1016/j.tibs.2009.02.004. URL https://doi.org/10.1016/j.tibs.2009.02.004.

[17] Peter S. Klein, Tzeli J. Sun, Charles L. Saxe, Alan R. Kimmel, Ronald L. Johnson, and Peter N. Devreotes. A chemoattractant receptor controls development in dictyostelium discoideum. *Science*, 241(4872):1467–1472, September 1988. doi: 10.1126/science.3047871. URL https://doi.org/10.1126/science.3047871.

[18] C L Saxe, G T Ginsburg, J M Louis, R Johnson, P N Devreotes, and A R Kimmel. CAR2, a prestalk cAMP receptor required for normal tip formation and late development of dictyostelium discoideum. *Genes &amp Development*, 7(2):262–272, February 1993. doi: 10.1101/gad.7.2.262. URL https://doi.org/10.1101/gad.7.2.262.

[19] R L Johnson, C L Saxe, R Gollop, A R Kimmel, and P N Devreotes. Identification and targeted gene disruption of cAR3, a cAMP receptor subtype expressed during multicellular stages of dictyostelium development. *Genes &amp Development*, 7(2):273–282, February 1993. doi: 10.1101/gad.7.2.273. URL https://doi.org/10.1101/gad.7.2.273.

[20] Juan A. Ballesteros and Harel Weinstein. [19] integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in g protein-coupled receptors. In *Methods Neurosci.*, pages 366–428. Elsevier, 1995. doi: 10.1016/s1043-9471(05)80049-7. URL https://doi.org/10.1016/s1043-9471(05)80049-7.

[21] Denise Wootten, John Simms, Laurence J. Miller, Arthur Christopoulos, and Patrick M. Sexton. Polar transmembrane interactions drive formation of ligand-specific and signal pathway-biased family b g protein-coupled receptor conformations. *PNAS*, 110(13):5211–5216, March 2013. doi: 10.1073/pnas.1221585110. URL https://doi.org/10.1073/pnas.1221585110.

[22] Jean-Philippe Pin, Thierry Galvez, and Laurent Prézeau. Evolution, structure, and activation mechanism of family 3/c g-protein-coupled receptors. *Pharmacology &amp Therapeutics*, 98(3):325–354, June 2003. doi: 10.1016/s0163-7258(03)00038-x. URL https://doi.org/10.1016/s0163-7258(03)00038-x.

[23] Chong Wang, Huixian Wu, Tama Evron, Eyal Vardy, Gye Won Han, Xi-Ping Huang, Sandy J. Hufeisen, Thomas J. Mangano, Dan J. Urban, Vsevolod Katritch, Vadim Cherezov, Marc G. Caron, Bryan L. Roth, and Raymond C. Stevens. Structural basis for

smoothened receptor modulation and chemoresistance to anticancer drugs. *Nature Communications*, 5(1), July 2014. doi: 10.1038/ncomms5355. URL https://doi.org/10.1038/ncomms5355.

[24] L. Oliveira, A. C. M. Paiva, and G. Vriend. A common motif in g-protein-coupled seven transmembrane helix receptors. *Journal of Computer-Aided Molecular Design*, 7(6): 649–658, December 1993. doi: 10.1007/bf00125323. URL https://doi.org/10.1007/bf00125323.

[25] J.M. Baldwin. The probable arrangement of the helices in g protein-coupled receptors. *The EMBO Journal*, 12(4):1693–1703, April 1993. doi: 10.1002/j.1460-2075.1993.tb05814.x. URL https://doi.org/10.1002/j.1460-2075.1993.tb05814.x.

[26] Thue W Schwartz. Locating ligand-binding sites in 7tm receptors by protein engineering. *Current Opinion in Biotechnology*, 5(4):434–444, August 1994. doi: 10.1016/0958-1669(94)90054-x. URL https://doi.org/10.1016/0958-1669(94)90054-x.

[27] Judy L. Meinkoth, Arthur S. Alberts, Wei Went, Dominic Fantozzi, Susan S. Taylor, Masatoshi Hagiwara, Marc Montminy, and James R. Feramisco. Signal transduction through the cAMP-dependent protein kinase. *Molecular and Cellular Biochemistry*, 127-128(1):179–186, November 1993. doi: 10.1007/bf01076769. URL https://doi.org/10.1007/bf01076769.

[28] Subreena Simrick, Roland F. Schindler, Kar-Lai Poon, and Thomas Brand. Popeye domain-containing proteins and stress-mediated modulation of cardiac pacemaking. *Trends in Cardiovascular Medicine*, 23(7):257–263, October 2013. doi: 10.1016/j.tcm.2013.02.002. URL https://doi.org/10.1016/j.tcm.2013.02.002.

[29] Johannes L. Bos. Epac proteins: multi-purpose cAMP targets. *Trends in Biochemical Sciences*, 31(12):680–686, December 2006. doi: 10.1016/j.tibs.2006.10.002. URL https://doi.org/10.1016/j.tibs.2006.10.002.

[30] U. Benjamin Kaupp and Reinhard Seifert. Cyclic nucleotide-gated ion channels. *Physiological Reviews*, 82(3):769–824, January 2002. doi: 10.1152/physrev.00008.2002. URL https://doi.org/10.1152/physrev.00008.2002.

[31] Donal A. Walsh and Scott M. Van Patten. Multiple pathway signal tran sduction by the cAMP-dependent protein kinase. *The FASEB Journal*, 8(15):1227–1236, December 1994. doi: 10.1096/fasebj.8.15.8001734. URL https://doi.org/10.1096/fasebj.8.15.8001734.

[32] Clare Ellis and. The state of GPCR research in 2004. *Nature Reviews Drug Discovery*, 3(7):577–626, July 2004. doi: 10.1038/nrd1458. URL https://doi.org/10.1038/nrd1458.

[33] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253 (5489):694–698, 1975.

[34] Brian K Shoichet and Brian K Kobilka. Structure-based drug screening for g-protein-coupled receptors. *Trends in biotechnology*, 29(7):382–390, 2011.

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

[36] Douglas B Kitchen, Henry Decornez, John R Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949, 2004.

[37] Mohammed alQuraishi. Proteins as machines for learning. *Nature*, 568(7752):299–301, 2019.

[38] Andrew W Senior, R James Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustinus Zidek, Andrew W Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577 (7792):706–710, 2020.

[39] Jochen S Hub. *Biophysics: a practical approach*. John Wiley & Sons, 2017.

[40] A.R. Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall, 2001. ISBN 9780582382107. URL https://books.google.co.in/books?id=kB7jsbV-uhkC.

[41] Steven C Frautschi, Richard P Olenick, Tom M Apostol, and David L Goodstein. *The mechanical universe*. Cambridge University Press, Cambridge, England, January 2008.

[42] Newton Isaac, N W Chittenden, Motte Andrew, and Hill Theodore Preston. *Newton's Principia: The Mathematical Principles of Natural Philosophy*. University of California Libraries. Daniel Adee, New York, 1846.

[43] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Computational science. Elsevier Science, 2001. ISBN 9780080519982. URL https://books.google.co.in/books?id=5qTzldS9ROIC.

[44] R. Baierlein. *Thermal Physics*. Cambridge University Press, 1999. ISBN 9780521658386. URL https://books.google.co.in/books?id=fqUU71spbZYC.

[45] Johannes Kästner. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6):932–942, May 2011. doi: 10.1002/wcms.66. URL https://doi.org/10.1002/wcms.66.

[46] Ming Chen. Collective variable-based enhanced sampling and machine learning. *The European Physical Journal B*, 94(10):5175–5192, October 2021. doi: 10.1140/epjb/s10051-021-00220-w. URL https://doi.org/10.1140/epjb/s10051-021-00220-w.

[47] Soumendranath Bhakat. Collective variable discovery in the age of machine learning: reality, hype and everything in between. *RSC Advances*, 12(38):25010–25024, 2022. doi: 10.1039/d2ra03660f. URL https://doi.org/10.1039/d2ra03660f.

[48] Hythem Sidky, Wei Chen, and Andrew L. Ferguson. Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. *Molecular Physics*, 118(5), March 2020. doi: 10.1080/00268976.2020.1737742. URL https://doi.org/10.1080/00268976.2020.1737742.

[49] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, February 1977. doi: 10.1016/0021-9991(77)90121-8. URL https://doi.org/10.1016/0021-9991(77)90121-8.

[50] Qinghua Liao. Enhanced sampling and free energy calculations for protein simulations. In *Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly*, pages 177–213. Elsevier, 2020. doi: 10.1016/bs.pmbts.2020.01.006. URL https://doi.org/10.1016/bs.pmbts.2020.01.006.

[51] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. THE weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, October 1992. doi: 10.1002/jcc.540130812. URL https://doi.org/10.1002/jcc.540130812.

[52] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2009. doi: 10.1007/978-0-387-84858-7. URL https://doi.org/10.1007/978-0-387-84858-7.

[53] A. D. Gordon, L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. *Biometrics*, 40(3):874, September 1984. doi: 10.2307/2530946. URL https://doi.org/10.2307/2530946.

[54] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, August 2016. ACM. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

[55] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, January 1967. doi: 10.1109/tit.1967.1053964. URL https://doi.org/10.1109/tit.1967.1053964.

[56] John A. Allen and Bryan L. Roth. Strategies to discover unexpected targets for drugs active at g Protein–Coupled receptors. *Annu. Rev. Pharmacol. Toxicol.*, 51(1):117–144, February 2011. doi: 10.1146/annurev-pharmtox-010510-100553. URL https://doi.org/10.1146/annurev-pharmtox-010510-100553.

[57] Daniel Wacker, Raymond C. Stevens, and Bryan L. Roth. How ligands illuminate GPCR mol. pharmacol. *Cell*, 170(3):414–427, July 2017. doi: 10.1016/j.cell.2017.07.009. URL https://doi.org/10.1016/j.cell.2017.07.009.

[58] Dianne M. Perez and Sadashiva S. Karnik. Multiple signaling states of g-protein-coupled receptors. *Pharmacol. Rev.*, 57(2):147–161, May 2005. doi: 10.1124/pr.57.2.2. URL https://doi.org/10.1124/pr.57.2.2.

[59] G. Vauquelin and I. Van Liefde. G protein-coupled receptors: a count of 1001 conformations. *Fundam. Clin. Pharmacol.*, 19(1):45–56, February 2005. doi: 10.1111/j.1472-8206.2005.00319.x. URL https://doi.org/10.1111/j.1472-8206.2005.00319.x.

[60] Jonathan D. Urban, William P. Clarke, Mark von Zastrow, David E. Nichols, Brian Kobilka, Harel Weinstein, Jonathan A. Javitch, Bryan L. Roth, Arthur Christopoulos, Patrick M. Sexton, Keith J. Miller, Michael Spedding, and Richard B. Mailman. Functional selectivity and classical concepts of quantitative pharmacology. *J. Pharmacol. Exp. Ther.*, 320(1):1–13, June 2006. doi: 10.1124/jpet.106.104463. URL https://doi.org/10.1124/jpet.106.104463.

[61] Narendra Tuteja. Signaling through g protein coupled receptors. *Plant Signaling Behav.*, 4(10):942–947, October 2009. doi: 10.4161/psb.4.10.9530. URL https://doi.org/10.4161/psb.4.10.9530.

[62] Lubert Stryer and Henry R. Bourne. G proteins: A family of signal transducers. *Annu. Rev. Cell Biol.*, 2(1):391–419, November 1986. doi: 10.1146/annurev.cb.02.110186.002135. URL https://doi.org/10.1146/annurev.cb.02.110186.002135.

[63] Taylor J. Maxwell, Margaret-Mary Ameyaw, Stuart Pritchard, Nadia Thornton, Gbolahan Folayan, Jessie Githang'a, Anne Indalo, Mohammed Tariq, Abeer Mobarek, David A. Evans, David Ofori-Adjei, Alan R. Templeton, and Howard L. McLeod. Beta-2 adrenergic receptor genotypes and haplotypes in different ethnic groups. *Int. J. Mol. Med.*, 16(4): 573–580, October 2005. ISSN 1107-3756.

[64] Charlotte K. Billington, Raymond B. Penn, and Ian P. Hall. Agonists. In *Handbook of Experimental Pharmacology*, pages 23–40. Springer International Publishing, 2016. doi: 10.1007/164_2016_64. URL `https://doi.org/10.1007/164_2016_64`.

[65] David E. Shaw, Ron O. Dror, John K. Salmon, J. P. Grossman, Kenneth M. Mackenzie, Joseph A. Bank, Cliff Young, Martin M. Deneroff, Brannon Batson, Kevin J. Bowers, Edmond Chow, Michael P. Eastwood, Douglas J. Ierardi, John L. Klepeis, Jeffrey S. Kuskin, Richard H. Larson, Kresten Lindorff-Larsen, Paul Maragakis, Mark A. Moraes, Stefano Piana, Yibing Shan, and Brian Towles. Millisecond-scale molecular dynamics simulations on anton. In *Proc. Conf. High Performance Computing Networking, Storage and Analysis*, New York, NY, USA, November 2009. ACM. doi: 10.1145/1654059.1654126. URL `https://doi.org/10.1145/1654059.1654126`.

[66] Ron O. Dror, Daniel H. Arlow, Paul Maragakis, Thomas J. Mildorf, Albert C. Pan, Huafeng Xu, David W. Borhani, and David E. Shaw. Activation mechanism of the β_2-adrenergic receptor. *PNAS*, 108(46):18684–18689, October 2011. doi: 10.1073/pnas.1110499108. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1110499108`.

[67] Daniel M. Rosenbaum, Cheng Zhang, Joseph A. Lyons, Ralph Holl, David Aragao, Daniel H. Arlow, Søren G. F. Rasmussen, Hee-Jung Choi, Brian T. DeVree, Roger K. Sunahara, Pil Seok Chae, Samuel H. Gellman, Ron O. Dror, David E. Shaw, William I. Weis, Martin Caffrey, Peter Gmeiner, and Brian K. Kobilka. Structure and function of an irreversible agonist- adrenoceptor complex. *Nature*, 469(7329):236–240, January 2011. doi: 10.1038/nature09665. URL `https://doi.org/10.1038/nature09665`.

[68] Melanie P. Wescott, Irina Kufareva, Cheryl Paes, Jason R. Goodman, Yana Thaker, Bridget A. Puffer, Eli Berdougo, Joseph B. Rucker, Tracy M. Handel, and Benjamin J. Doranz. Signal transmission through the CXC chemokine receptor 4 (CXCR4) transmembrane helices. *PNAS*, 113(35):9928–9933, August 2016. doi: 10.1073/pnas.1601278113. URL `https://doi.org/10.1073/pnas.1601278113`.

[69] Benjamin G. Tehan, Andrea Bortolato, Frank E. Blaney, Malcolm P. Weir, and Jonathan S. Mason. Unifying family a GPCR theories of activation. *Pharmacol.*

Ther., 143(1):51–60, July 2014. doi: 10.1016/j.pharmthera.2014.02.004. URL https://doi.org/10.1016/j.pharmthera.2014.02.004.

[70] Xinbing Han, Souvenir D. Tachado, Henry Koziel, and William A. Boisvert. Leu1283.43 (L128) and val2476.40 (V247) of CXCR1 are critical amino acid residues for g protein coupling and receptor activation. *PLoS ONE*, 7(8):e42765, August 2012. doi: 10.1371/journal.pone.0042765. URL https://doi.org/10.1371/journal.pone.0042765.

[71] Maria Martí-Solano, Ferran Sanz, Manuel Pastor, and Jana Selent. A dynamic view of molecular switch behavior at serotonin receptors: Implications for functional selectivity. *PLoS ONE*, 9(10):e109312, October 2014. doi: 10.1371/journal.pone.0109312. URL https://doi.org/10.1371/journal.pone.0109312.

[72] Qingtong Zhou, Dehua Yang, Meng Wu, Yu Guo, Wanjing Guo, Li Zhong, Xiaoqing Cai, Antao Dai, Wonjo Jang, Eugene I Shakhnovich, Zhi-Jie Liu, Raymond C Stevens, Nevin A Lambert, M Madan Babu, Ming-Wei Wang, and Suwen Zhao. Common activation mechanism of class a GPCRs. *eLife*, 8, December 2019. doi: 10.7554/elife.50279. URL https://doi.org/10.7554/elife.50279.

[73] Steven O. Smith. Deconstructing the transmembrane core of class a g protein–coupled receptors. *Trends Biochem. Sci*, 46(12):1017–1029, December 2021. doi: 10.1016/j.tibs.2021.08.006. URL https://doi.org/10.1016/j.tibs.2021.08.006.

[74] Slawomir Filipek. Molecular switches in GPCRs. *Curr. Opin. Struct. Biol.*, 55:114–120, April 2019. doi: 10.1016/j.sbi.2019.03.017. URL https://doi.org/10.1016/j.sbi.2019.03.017.

[75] Yinglong Miao, Alisha D. Caliman, and J. Andrew McCammon. Allosteric effects of sodium ion binding on activation of the M3 muscarinic g-protein-coupled receptor. *Biophys. J.*, 108(7):1796–1806, April 2015. doi: 10.1016/j.bpj.2015.03.003. URL https://doi.org/10.1016/j.bpj.2015.03.003.

[76] Yinglong Miao and J Andrew McCammon. G-protein coupled receptors: advances in simulation and drug discovery. *Curr. Opin. Struct. Biol.*, 41:83–89, December 2016. doi: 10.1016/j.sbi.2016.06.008. URL https://doi.org/10.1016/j.sbi.2016.06.008.

[77] Ron O. Dror, Albert C. Pan, Daniel H. Arlow, David W. Borhani, Paul Maragakis, Yibing Shan, Huafeng Xu, and David E. Shaw. Pathway and mechanism of drug binding to g-protein-coupled receptors. *PNAS*, 108(32):13118–13123, July 2011. doi: 10.1073/pnas.1104614108. URL https://doi.org/10.1073/pnas.1104614108.

[78] Vignir Isberg, Bas Vroling, Rob van der Kant, Kang Li, Gert Vriend, and David Gloriam. GPCRDB: an information system for g protein-coupled receptors. *Nucleic Acids Res.*, 42 (D1):D422–D425, December 2013. doi: 10.1093/nar/gkt1255. URL https://doi.org/10.1093/nar/gkt1255.

[79] Weijiao Huang, Aashish Manglik, A. J. Venkatakrishnan, Toon Laeremans, Evan N. Feinberg, Adrian L. Sanborn, Hideaki E. Kato, Kathryn E. Livingston, Thor S. Thorsen, Ralf C. Kling, Sébastien Granier, Peter Gmeiner, Stephen M. Husbands, John R. Traynor, William I. Weis, Jan Steyaert, Ron O. Dror, and Brian K. Kobilka. Structural insights into μ-opioid receptor activation. *Nature*, 524(7565):315–321, August 2015. doi: 10.1038/nature14886. URL https://doi.org/10.1038/nature14886.

[80] Xinyu Xu, Jonas Kaindl, Mary J. Clark, Harald Hübner, Kunio Hirata, Roger K. Sunahara, Peter Gmeiner, Brian K. Kobilka, and Xiangyu Liu. Binding pathway determines norepinephrine selectivity for the human ar over . *Cell Res.*, 31(5):569–579, October 2020. doi: 10.1038/s41422-020-00424-2. URL https://doi.org/10.1038/s41422-020-00424-2.

[81] Xavier Deupi and Jörg Standfuss. Structural insights into agonist-induced activation of g-protein-coupled receptors. *Curr. Opin. Struct. Biol.*, 21(4):541–551, August 2011. doi: 10.1016/j.sbi.2011.06.002. URL https://doi.org/10.1016/j.sbi.2011.06.002.

[82] B. Trzaskowski, D. Latek, S. Yuan, U. Ghoshdastider, A. Debinski, and S. Filipek. Action of molecular switches in GPCRs - theoretical and experimental studies. *Curr. Med. Chem.*, 19(8):1090–1109, March 2012. doi: 10.2174/092986712799320556. URL https://doi.org/10.2174/092986712799320556.

[83] H. C. Stephen Chan, Slawomir Filipek, and Shuguang Yuan. The principles of ligand specificity on beta-2-adrenergic receptor. *Sci. Rep.*, 6(1), October 2016. doi: 10.1038/srep34736. URL https://doi.org/10.1038/srep34736.

[84] Yanxiao Han, John R. D. Dawson, Kevin R. DeMarco, Kyle C. Rouen, Slava Bekker, Vladimir Yarov-Yarovoy, Colleen E. Clancy, Yang K. Xiang, and Igor Vorobyov. Elucidation of a dynamic interplay between a beta-2 adrenergic receptor, its agonist, and stimulatory g protein. *PNAS*, 120(10), February 2023. doi: 10.1073/pnas.2215916120. URL https://doi.org/10.1073/pnas.2215916120.

[85] Argha Mitra, Arijit Sarkar, and Attila Borics. Universal properties and specificities of the -adrenergic receptor-gs protein complex activation mechanism revealed by all-atom

molecular dynamics simulations. *Int. J. Mol. Sci.*, 22(19):10423, September 2021. doi: 10.3390/ijms221910423. URL https://doi.org/10.3390/ijms221910423.

[86] Ismael Rodríguez-Espigares, Mariona Torrens-Fontanals, Johanna K. S. Tiemann, David Aranda-García, Juan Manuel Ramírez-Anguita, Tomasz Maciej Stepniewski, Nathalie Worp, Alejandro Varela-Rial, Adrián Morales-Pastor, Brian Medel-Lacruz, Gáspár Pándy-Szekeres, Eduardo Mayol, Toni Giorgino, Jens Carlsson, Xavier Deupi, Slawomir Filipek, Marta Filizola, José Carlos Gómez-Tamayo, Angel Gonzalez, Hugo Gutiérrez de Terán, Mireia Jiménez-Rosés, Willem Jespers, Jon Kapla, George Khelashvili, Peter Kolb, Dorota Latek, Maria Marti-Solano, Pierre Matricon, Minos-Timotheos Matsoukas, Przemyslaw Miszta, Mireia Olivella, Laura Perez-Benito, Davide Provasi, Santiago Ríos, Iván R. Torrecillas, Jessica Sallander, Agnieszka Sztyler, Silvana Vasile, Harel Weinstein, Ulrich Zachariae, Peter W. Hildebrand, Gianni De Fabritiis, Ferran Sanz, David E. Gloriam, Arnau Cordomi, Ramon Guixà-González, and Jana Selent. GPCRmd uncovers the dynamics of the 3d-GPCRome. *Nat. Methods*, 17(8):777–787, July 2020. doi: 10.1038/s41592-020-0884-y. URL https://doi.org/10.1038/s41592-020-0884-y.

[87] S Suryanarayana and B K Kobilka. Amino acid substitutions at position 312 in the seventh hydrophobic segment of the beta 2-adrenergic receptor modify ligand-binding specificity. *Mol. Pharmacol.*, 44(1):111–114, July 1993. URL https://pubmed.ncbi.nlm.nih.gov/8101966/.

[88] K Wieland, H M Zuurmond, C Krasel, A P Ijzerman, and M J Lohse. Involvement of asn-293 in stereospecific agonist recognition and in activation of the beta 2-adrenergic receptor. *PNAS*, 93(17):9276–9281, August 1996. doi: 10.1073/pnas.93.17.9276. URL https://doi.org/10.1073/pnas.93.17.9276.

[89] Fu Zon Chung, Cheng-Dian Wang, Paul C. Potter, J. Craig Venter, and Claire M. Fraser. Site-directed mutagenesis and continuous expression of human beta-adrenergic receptors. identification of a conserved aspartate residue involved in agonist binding and receptor activation. *J. Biol. Chem.*, 263(9):4052–4055, March 1988. URL https://pubmed.ncbi.nlm.nih.gov/2831218/.

[90] C M Fraser, F Z Chung, C D Wang, and J C Venter. Site-directed mutagenesis of human beta-adrenergic receptors: substitution of aspartic acid-130 by asparagine produces a receptor with high-affinity agonist binding that is uncoupled from adenylate cyclase. *PNAS*, 85(15):5478–5482, August 1988. doi: 10.1073/pnas.85.15.5478. URL https://doi.org/10.1073/pnas.85.15.5478.

[91] Gáspár Pándy-Szekeres, Christian Munk, Tsonko M Tsonkov, Stefan Mordalski, Kasper Harpsøe, Alexander S Hauser, Andrzej J Bojarski, and David E Gloriam. GPCRdb in 2018: adding GPCR structure models and ligands. *Nucleic Acids Res.*, 46(D1):D440–D446, November 2017. doi: 10.1093/nar/gkx1109. URL https://doi.org/10.1093/nar/gkx1109.

[92] Christian Munk, Kasper Harpsøe, Alexander S Hauser, Vignir Isberg, and David E Gloriam. Integrating structural and mutagenesis data to elucidate GPCR ligand binding. *Curr. Opin. Pharmacol.*, 30:51–58, October 2016. doi: 10.1016/j.coph.2016.07.003. URL https://doi.org/10.1016/j.coph.2016.07.003.

[93] V. Cherezov, D. M. Rosenbaum, M. A. Hanson, S. G. F. Rasmussen, F. S. Thian, T. S. Kobilka, H.-J. Choi, P. Kuhn, W. I. Weis, B. K. Kobilka, and R. C. Stevens. High-resolution crystal structure of an engineered human 2-adrenergic g protein-coupled receptor. *Science*, 318(5854):1258–1265, October 2007. doi: 10.1126/science.1150577. URL https://doi.org/10.1126/science.1150577.

[94] Aaron M. Ring, Aashish Manglik, Andrew C. Kruse, Michael D. Enos, William I. Weis, K. Christopher Garcia, and Brian K. Kobilka. Adrenaline-activated structure of -adrenoceptor stabilized by an engineered nanobody. *Nature*, 502(7472):575–579, September 2013. doi: 10.1038/nature12572. URL https://doi.org/10.1038/nature12572.

[95] Veronica Ianno', Philippe Négrier, and Philippe Espeau. Adrenaline system: another rare case of conglomerate with partial solid solutions. *J. Therm. Anal. Calorim.*, 138(2):997–1002, May 2019. doi: 10.1007/s10973-019-08301-w. URL https://doi.org/10.1007/s10973-019-08301-w.

[96] FAJGA MANDELBAUM, OLGA B. HENRIQUES, and S. B. HENRIQUES. Comparative hyperfibrinogenæmic action of d- and l-adrenaline. *Nature*, 178(4529):363–364, August 1956. doi: 10.1038/178363b0. URL https://doi.org/10.1038/178363b0.

[97] Mikhail A. Lomize, Irina D. Pogozheva, Hyeon Joo, Henry I. Mosberg, and Andrei L. Lomize. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.*, 40(D1):D370–D376, September 2011. doi: 10.1093/nar/gkr703. URL https://doi.org/10.1093/nar/gkr703.

[98] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. Charmm-gui: A web-based graphical user interface for charmm. *J. Comput. Chem.*, 29(11):1859–1865, March 2008. doi: 10.1002/jcc.20945. URL https://doi.org/10.1002/jcc.20945.

[99] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, 30(10):1545–1614, May 2009. doi: 10.1002/jcc.21287. URL `https://doi.org/10.1002/jcc.21287`.

[100] Jumin Lee, Xi Cheng, Jason M. Swails, Min Sun Yeom, Peter K. Eastman, Justin A. Lemkul, Shuai Wei, Joshua Buckner, Jong Cheol Jeong, Yifei Qi, Sunhwan Jo, Vijay S. Pande, David A. Case, Charles L. Brooks, Alexander D. MacKerell, Jeffery B. Klauda, and Wonpil Im. CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J. Chem. Theory Comput.*, 12(1):405–413, December 2015. doi: 10.1021/acs.jctc.5b00935. URL `https://doi.org/10.1021/acs.jctc.5b00935`.

[101] Emilia L. Wu, Xi Cheng, Sunhwan Jo, Huan Rui, Kevin C. Song, Eder M. Dávila-Contreras, Yifei Qi, Jumin Lee, Viviana Monje-Galvan, Richard M. Venable, Jeffery B. Klauda, and Wonpil Im. CHARMM-GUIMembrane buildertoward realistic biological membrane simulations. *J. Comput. Chem.*, 35(27):1997–2004, August 2014. doi: 10.1002/jcc.23702. URL `https://doi.org/10.1002/jcc.23702`.

[102] Sunhwan Jo, Taehoon Kim, and Wonpil Im. Automated builder and database of Protein/Membrane complexes for molecular dynamics simulations. *PLoS ONE*, 2(9):e880, September 2007. doi: 10.1371/journal.pone.0000880. URL `https://doi.org/10.1371/journal.pone.0000880`.

[103] IR Gould, AA Skjevik, CJ Dickson, BD Madej, and RC Walker. Lipid17: A comprehensive AMBER force field for the simulation of zwitterionic and anionic lipids. *in prep.*, 2018.

[104] Chuan Tian, Koushik Kasavajhala, Kellon A. A. Belfon, Lauren Raguette, He Huang, Angela N. Migues, John Bickel, Yuzhang Wang, Jorge Pincay, Qin Wu, and Carlos Simmerling. ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *J. Chem. Theory Comput.*, 16(1):528–552, November 2019. doi: 10.1021/acs.jctc.9b00591. URL `https://doi.org/10.1021/acs.jctc.9b00591`.

[105] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, 2004. doi: 10.1002/jcc.20035. URL `https://doi.org/10.1002/jcc.20035`.

[106] Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.*, 25(2):247–260, October 2006. doi: 10.1016/j.jmgm.2005.12.005. URL https://doi.org/10.1016/j.jmgm.2005.12.005.

[107] Araz Jakalian, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. parameterization and validation. *J. Comput. Chem.*, 23(16):1623–1641, October 2002. doi: 10.1002/jcc.10128. URL https://doi.org/10.1002/jcc.10128.

[108] Pekka Mark and Lennart Nilsson. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 k. *J. Phys. Chem. A*, 105(43):9954–9960, November 2001. doi: 10.1021/jp003020w. URL https://doi.org/10.1021/jp003020w.

[109] Aashish Manglik, Tae Hun Kim, Matthieu Masureel, Christian Altenbach, Zhongyu Yang, Daniel Hilger, Michael T. Lerch, Tong Sun Kobilka, Foon Sun Thian, Wayne L. Hubbell, R. Scott Prosser, and Brian K. Kobilka. Structural insights into the dynamic process of β_2-adrenergic receptor signaling. *Cell*, 161(5):1101–1111, May 2015. doi: 10.1016/j.cell.2015.04.043. URL https://doi.org/10.1016/j.cell.2015.04.043.

[110] Noureldin Saleh, Passainte Ibrahim, and Timothy Clark. Differences between g-protein-stabilized agonist-GPCR complexes and their nanobody-stabilized equivalents. *Angew. Chem. Int. Ed.*, 56(31):9008–9012, July 2017. doi: 10.1002/anie.201702468. URL https://doi.org/10.1002/anie.201702468.

[111] Donald W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.*, 11(2):431–441, June 1963. doi: 10.1137/0111030. URL https://doi.org/10.1137/0111030.

[112] Hestenes Magnus Rudolph and Stiefel Eduard. *Methods of conjugate gradients for solving linear systems*, volume 49. NBS Washington, DC, 1952.

[113] Richard J. Loncharich, Bernard R. Brooks, and Richard W. Pastor. Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers*, 32(5):523–535, May 1992. doi: 10.1002/bip.360320508. URL https://doi.org/10.1002/bip.360320508.

[114] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, October 1984. doi: 10.1063/1.448118. URL https://doi.org/10.1063/1.448118.

[115] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23(3):327–341, March 1977. doi: 10.1016/0021-9991(77)90098-5. URL https://doi.org/10.1016/0021-9991(77)90098-5.

[116] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: AnN·log(N) method for ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, June 1993. doi: 10.1063/1.464397. URL https://doi.org/10.1063/1.464397.

[117] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76(1):637–649, January 1982. doi: 10.1063/1.442716. URL https://doi.org/10.1063/1.442716.

[118] D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, V.W.D. Cruzeiro III, R.E. Duke T.A. Darden, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, S. Izadi Y. Huang, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, K.M. Merz D.J. Mermelstein, 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. Salomon-Ferrer, R.C. Walker J. Swails, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York, and P.A. Kollman. Amber 2018, university of california, san francisco, 2018.

[119] Daniel R. Roe and Thomas E. Cheatham. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.*, 9(7):3084–3095, June 2013. doi: 10.1021/ct400341p. URL https://doi.org/10.1021/ct400341p.

[120] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *J. Mol. Graphics*, 14:33–38, 1996.

[121] Noureldin Saleh, Passainte Ibrahim, Giorgio Saladino, Francesco Luigi Gervasio, and Timothy Clark. An efficient metadynamics-based protocol to model the binding affinity and the transition state ensemble of g-protein-coupled receptor ligands. *J. Chem. Inf. Model.*, 57(5):1210–1217, May 2017. doi: 10.1021/acs.jcim.6b00772. URL https://doi.org/10.1021/acs.jcim.6b00772.

[122] Johan Åqvist, Carmen Medina, and Jan-Erik Samuelsson. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng., Des. Sel.*, 7(3):385–391, 1994. doi: 10.1093/protein/7.3.385. URL https://doi.org/10.1093/protein/7.3.385.

[123] David A Freedman. *Statistical models*. Cambridge University Press, Cambridge, England, 2 edition, April 2009.

[124] Patrick Schober, Christa Boer, and Lothar A. Schwarte. Correlation coefficients. *Anesth Analg.*, 126(5):1763–1768, May 2018. doi: 10.1213/ane.0000000000002864. URL https://doi.org/10.1213/ane.0000000000002864.

[125] Aditi Tandale, Manali Joshi, and Durba Sengupta. Structural insights and functional implications of inter-individual variability in -adrenergic receptor. *Sci. Rep.*, 6(1), April 2016. doi: 10.1038/srep24379. URL https://doi.org/10.1038/srep24379.

[126] Daniela R. Silva, Josué M. Silla, Lucas A. Santos, Elaine F. F. da Cunha, and Matheus P. Freitas. The role of intramolecular interactions on the bioactive conformation of epinephrine. *Mol. Inf.*, 38(6):1800167, April 2019. doi: 10.1002/minf.201800167. URL https://doi.org/10.1002/minf.201800167.

[127] Parisa Mollaei and Amir Barati Farimani. Activity map and transition pathways of g protein-coupled receptor revealed by machine learning. *J. Chem. Inf. Model.*, 63(8):2296–2304, April 2023. doi: 10.1021/acs.jcim.3c00032. URL https://doi.org/10.1021/acs.jcim.3c00032.

[128] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus F. M. Ng, B. Liu, Philip S. Yu, Zhi-Hua Zhou, Michael S. Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14:1–37, 2007.

[129] S.G.F. Rasmussen, B.T. DeVree, Y. Zou, A.C. Kruse, K.Y. Chung, T.S. Kobilka, F.S. Thian, P.S. Chae, E. Pardon, D. Calinski, J.M. Mathiesen, S.T.A. Shah, J.A. Lyons, M. Caffrey, S.H. Gellman, J. Steyaert, G. Skiniotis, W.I. Weis, R.K. Sunahara, and B.K. Kobilka. Crystal structure of the beta2 adrenergic receptor-gs protein complex, July 2011. URL https://doi.org/10.2210/pdb3sn6/pdb.

[130] Stephen R. Sprang. Invited review: Activation of g proteins by GTP and the mechanism of g-catalyzed GTP hydrolysis. *Biopolymers*, 105(8):449–462, May 2016. doi: 10.1002/bip.22836. URL https://doi.org/10.1002/bip.22836.

[131] David Ryan Koes, Matthew P. Baumgartner, and Carlos J. Camacho. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, February 2013. doi: 10.1021/ci300604z. URL https://doi.org/10.1021/ci300604z.