EPFL

Thèse n° 10 182

From Sequence to Dynamics to Function: Computational Design of Allostery and Ligand Selectivity in G-Protein Coupled Receptors

Présentée le 1er mars 2024

Faculté des sciences de la vie
Unité du Prof. Barth
Programme doctoral en science et génie des matériaux

pour l'obtention du grade de Docteur ès Sciences

par

Mahdi HIJAZI

Acceptée sur proposition du jury

Prof. K. Scrivener, présidente du jury
Prof. P. D. Barth, directeur de thèse
Prof. P. W. Hildebrand, rapporteur
Prof. S. Grzesiek, rapporteur
Prof. M. Dal Peraro, rapporteur

École polytechnique fédérale de Lausanne

2024

One day the sun admitted,
I am just a shadow.
I wish I could show you The Infinite Incandescence
that has cast my brilliant image!
I wish I could show,
when you are lonely or in darkness,
the Astonishing Light
of your own Being!
— Hafiz

To Ibrahim…

Acknowledgements

First and foremost, I would like to thank my family, who have planted the seed of the love of knowledge deep inside since I was a young one. Thank you mom for all the discussions that we had throughout, for reminding me of the bigger picture, and that there is much more to life than research and academic work. Thank you dad for your unwavering support and belief in me, and for your positivity whenever we talk. Thank you Ali for the jokes and the fun, for trying hard to make my fashion on point. Finally, thank you Ibrahim for always being there (albeit remotely), and for all the great books you gifted me since childhood.

I would also express my thanks to the members of the Barth lab for being the fantastic support system that they are, and making the PhD journey a bit more tolerable.

I am grateful to Patrick for his guidance, supervision, and support throughout my PhD journey. We have had many insightful discussions and he never fails to propose new and exciting ideas.

I would also like to thank Matteo, Stephan, and Peter for their careful reading of this manuscript, and for the stimulating discussion during my oral exam. Thank you Karen for moderating the oral exam and making sure it was a smooth experience.

Thank you Lucas and Matthieu for taking the time to review the manuscript.
During my time in the Barth lab, I had the pleasure of meeting and working with:

- Rob, the dedicated DM who actually managed to finish a D&D campaign. I already miss your dry humor here in the lab.
- Louis, who was supportive and a pleasure to work with.
- Daniel, from whom I learned a lot during my first days here.
- Lucas, with whom I spent so much time in the lab, and the best board game commentator.
- Aysima, all I am going to say is: noice!
- Aurélien, best shirts AND cleanest dose response curves.
- Matthieu, I have GAINed a lot from our friendship

- Liyan, who always pushed me to participate in social life and activities in SV. Reminder to you, Liyan, that the factory must grow.
- Gabriele, Gabriele, Gabriele, thanks for the socks man.
- Ana, I enjoyed our discussions about both work and life.
- The penguins, it was great to have you around.

To my students, Tom, Simon, Amina, and Mariia, it was fantastic to work with every one of you.
With immense gratitude, I acknowledge Fabrizio for his unwavering presence all throughout my PhD.
I am profoundly grateful for the discord squad, especially Ali and Ahmad, for making life abroad, especially during a pandemic, much more bearable.
Thank you Hussein for all your logistical help during my first PhD days, and for the frequent visits to Lausanne.
I deeply appreciate Mustafa for all the incredibly insightful and useful discussions throughout my PhD.
I wish to thank my Lebanese group of friends in Lausanne (and Switzerland generally), Hamza and Walaa (the fam!), Baqer and Ghofran, Taha and Mira, Hala, Zeinab, Ali, Ghewa, Hadi (both of them), Ahmad, Hassan, and many others!
My heartfelt thanks go to the Dubrovnik coffee club, to Jarla and her adventures, and to Anna and our (almost) weekly coffee breaks.
And finally, a huge thank you to Taenaz for making the last year of my PhD much more bearable.

Lausanne, January 11, 2024 M. H.

Abstract

The phenomenon of allostery, a general property in proteins that has been heralded as "the second secret of life" remains elusive to our understanding and even more challenging to incorporate into protein design. One example of allosteric proteins with great therapeutic potential are G-Protein coupled receptors (GPCRs). GPCRs play a crucial role in regulating numerous physiological reactions triggered by neurotransmitters, hormones, and various environmental stimuli. As a result, GPCRs are targets for nearly one-third of all licensed pharmaceutical drugs. In this thesis, we present a framework to elucidate allosteric signaling applied to GPCRs as model systems. The framework is based on analyzing dynamics in proteins modeled via molecular simulations to (1) extract potential allosteric pathways in the protein and (2) quantify protein response to a perturbation. We employ computational protein design coupled with the aforementioned dynamic analysis to explore the allosteric functions of GPCRs, unveiling mechanistic relationships between agonist ligand chemistry, receptor sequence, structure, dynamics, and allosteric signaling across the dopamine receptor family. The framework is also applied to designed signaling complexes between conformationally dynamic proteins and peptides in chemokine receptors to shed light on the change of allosteric pathways in response to the designs. This work is a step forward toward mechanistic understanding of sequence polymorphism on receptor function and pharmacology, providing valuable insights for selective drug design and rational receptor engineering for both fundamental research and therapeutic applications.

Keywords: protein design, allostery, G-Protein Coupled Receptors, molecular dynamics, bioengineering, protein dynamics.

0.1 List of Publications

1. Yin, J., Chen, K. Y. M., Clark, M. J., **Hijazi, M.**, Kumari, P., Bai, X. C., ..., Barth, P., & Rosenbaum, D. M. (2020). Structure of a D2 dopamine receptor–G-protein complex in a lipid membrane. Nature, 584(7819), 125-129.

2. Keri, D.*, **Hijazi, M.***, Oggier, A., & Barth, P. (2022). Computational rewiring of allosteric pathways reprograms GPCR selective responses to ligands. bioRxiv, 2022-03. *: Co-first

authorship

3. Rudden, L. S., **Hijazi, M.**, & Barth, P. (2022). Deep learning approaches for conformational flexibility and switching properties in protein design. Frontiers in Molecular Biosciences, 9, 928534.

4. Dumas, L., Marfoglia, M., Yang, B., **Hijazi, M.**, Larabi, A., Lau, K., ... & Barth, P. (2023). Uncovering and engineering the mechanical properties of the adhesion GPCR ADGRG1 GAIN domain. bioRxiv, 2023-04.

5. Jefferson, R. E., Oggier, A., Füglistaler, A., Camviel, N., **Hijazi, M.**, Villarreal, A. R., ... & Barth, P. (2023). Computational design of dynamic receptor—peptide signaling complexes applied to chemotaxis. Nature Communications, 14(1), 2875.

Résumé

Le phénomène de l'allostérie, une propriété générale des protéines qui a été qualifiée de "second secret de la vie", demeure aujourd'hui une propriété insaisissable et difficile à intégrer dans la conception que nous avons des protéines. Un exemple de protéines allostériques ayant un grand potentiel thérapeutique est celui des récepteurs couplés aux protéines G (RCPGs). Les RCPGs jouent un rôle crucial dans la régulation de nombreuses réactions physiologiques déclenchées par les neurotransmetteurs, les hormones et autres stimuli environnementaux. En conséquence, les RCPGs sont la cible d'environ un tiers des médicaments pharmaceutiques autorisés à ce jour. Dans cette thèse, nous présentons une méthode pour comprendre et caractériser la signalisation allostérique que nous avons appliqué au système modèle des RCPGs. Cette méthode repose sur l'analyse de la dynamique des protéines modélisées par des simulations moléculaires afin (1) de caractériser les voies allostériques potentielles dans la protéine et (2) de quantifier la réponse de la protéine à diverses perturbations. En combinant de méthodes computationnelles permettant le design de protéines avec des outils d'analyse de leurs dynamiques précédemment citées, nous avons explorer les fonctions allostériques des RCPGs, révélant les relations mécanistiques entre la chimie des ligands agonistes, la séquence, la structure et la dynamique des récepteurs pour caractériser la signalisation allostérique de la famille des récepteurs à la dopamine. Cette nouvelle méthode est appliquée à des complexes de signalisation composés de protéines conformationnellement dynamiques et des peptides dans les récepteurs aux chimiokines afin de comprendre les modifications des voies allostériques en réponse aux mutations introduites dans ces systèmes. Ce travail constitue une avancée vers la compréhension mécanistique de la polymorphie de la séquence sur la fonction ainsi que sur la pharmacologie des récepteurs, fournissant des informations précieuses pour la conception sélective de médicaments et l'ingénierie rationnelle des récepteurs, à la fois pour la recherche fondamentale et les applications thérapeutiques.

Mots clés : conception de protéines, allostérie, récepteurs couplés aux protéines G, dynamique moléculaire, bio-ingénierie, dynamique des protéines.

Glossary

β1AR β_1 adrenergic receptor, a class A GPCR from the aminergic family
β2AR β_2 adrenergic receptor, a class A GPCR from the aminergic family

Agonist a molecule (small or otherwise) that activates a receptor to produce a biological response
Apo the state in which a protein is not bound to a ligand
ATSM allosteric two state model, a framework to describe the behavior of allosteric proteins

BW Ballesteros-Weinstein, generic numbering scheme for class A GPCRs (1)

cAMP cyclic adenosine monophosphate, a second messenger in many cellular signaling pathways.
CCR5 chemokine receptor type 5, a class A GPCR
C-term or **C-terminus** carboxyl terminus
CV collective variable, a structural parameter (or combination of) that can be measured during a simulation
CXCR2 chemokine receptor type 2, a class A GPCR
CXCR4 chemokine receptor type 4, a class A GPCR

DCCM dynamic cross-correlation map, a matrix of cross-correlations extracted from dynamical simulations
DD1R dopamine D1 receptor, a class A GPCR from the aminergic family
DD2R dopamine D2 receptor, a class A GPCR from the aminergic family

EC extracellular
ECL extracellular loop

GDP guanine diphosphate,
Gi G-alpha inhibitory subunit
GPCR G-protein coupled receptor, integral membrane proteins with seven membrane-spanning domains, or helices
Gp G-protein, a family of enzymes that hydrolyze GTP to GDP
Gq G-alpha q subunit

GRK G-protein coupled receptor kinases
Gs G-alpha stimulatory subunit
Gt Transducin, the G-protein that interacts with rhodopsin and is present in vertebrate retina rods and cones
GTP guanine triphosphate,

H5 G-protein carboxy-terminal helix 5

IC intracellular
ICL intracellular loop

KL_1 Local Kullback-Leibler divergence for one body
KLdiv Kullback-Leibler divergence, a measure of distance between two probability distributions

M_2 mutual divergence, the mutual information equivalent for KLdiv
MCMC Markov-chain Monte Carlo
MD molecular dynamics
MI mutual information, a statistical measure that quantifies the degree of dependency or information shared between two random variables
MSA multiple sequence alignment
MWC Monod-Wyman-Changeux, one of the first models of allostery (2)

NMA normal mode analysis, used to describe protein fluctuations about an equilibrium position
NMR nuclear magnetic resonance
N-term or **N-terminus** Amino terminus, protein sequences start at the N-terminus and end at the C-terminus

PAM positive allosteric modulator
PCA principal cmponent analysis, a linear dimensionality reduction method
pdf Probability density function
POPC 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine, a phospholipid naturally found in eukaryotic cell membranes
PTM post-translational modifiction

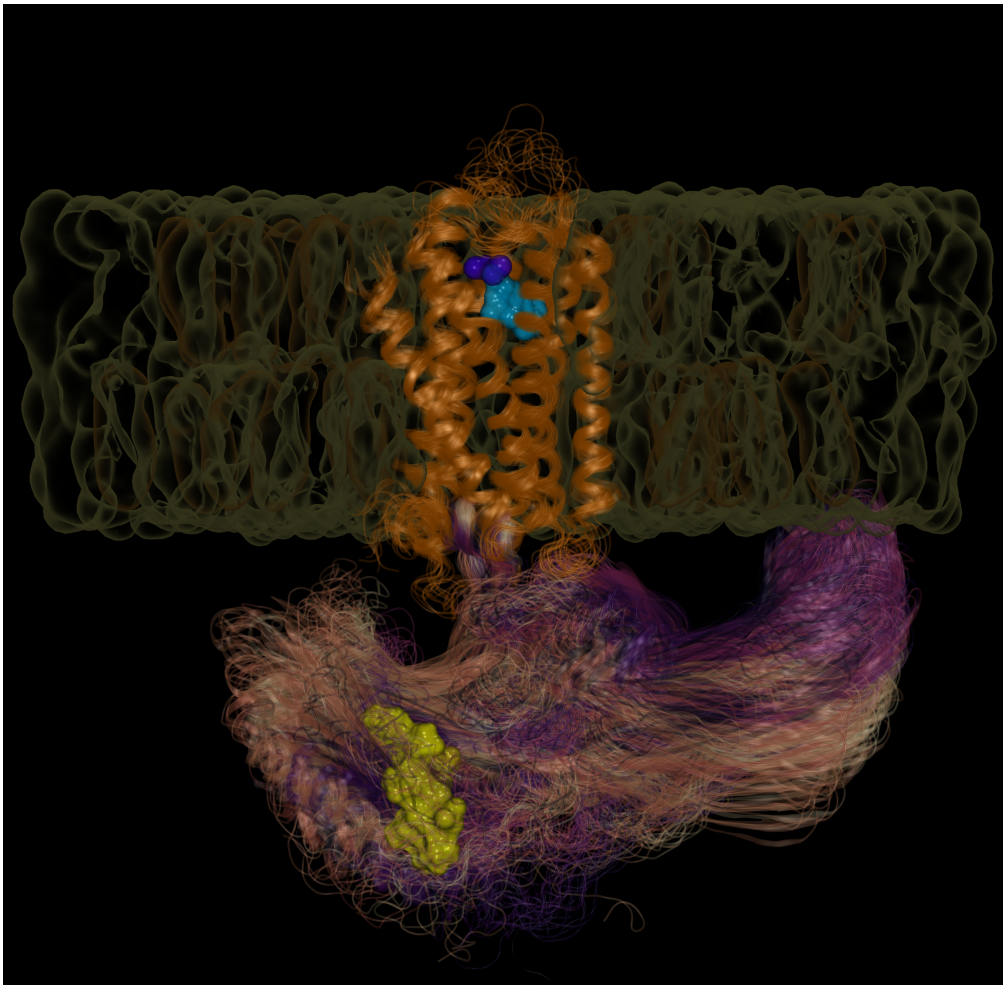
RMSD root mean square deviation
RMSF root mean square fluctuation
RU Rosetta units, energy units for the empirical Rosetta forcefield

TM trans-membrane
TMH trans-membrane helix

TRP transient receptor potential

WT wild type

Contents



Artwork 1: *Serotonin 2B (orange) receptor bound to LSD (cyan/purple) and interacting with β-arrestin, an intracellular partner (beige thru purple). Phosphorylated tail of the receptor that plays an important role in β-arrestin recruitment is shown in yellow surface representation.*

Introduction

"To know the secrets of life, we must first
become aware of their existence"
— Albert Einstein, The World As I See It

0.2 A brief history of allostery

Heralded as "the second secret of life" (3), allostery is the biological phenomenon where two or more sites on a single biological molecule are dynamically coupled despite being out of direct physical interaction range. It regulates a plethora of biomolecular functions, for example, some of the first discovered allosteric proteins that responded to ligand binding were hemoglobin (4) and L-threonine deaminase (5) (in addition to many many others (6; 7; 8; 9; 10; 11; 12; 13; 14; 15; 16)). Even though this phenomenon has been known for a long time, allostery remains rather elusive and its understanding on a molecular level leaves a lot to be desired.

One example that highlights the importance of allosteric regulation is the disease-causing mutation in hemoglobin that leads to sickle cell anemia. Hemoglobin has four subunits, each containing a heme group where oxygen can bind. Allosteric regulation in hemoglobin involves cooperative binding, meaning that when one oxygen molecule binds to a heme group, it increases the affinity of the remaining heme groups for oxygen. This enhances the overall efficiency of oxygen transport in the blood. However, in sickle cell anemia, a single amino acid substitution occurs in one of the hemoglobin subunits. This mutation replaces a glutamic acid residue with valine in the β-globin subunit, resulting in the formation of hemoglobin S (HbS, as opposed to normal adult hemoglobin HbA). HbS has disrupted allosteric regulation, and when oxygen levels are low, HbS molecules tend to polymerize and form long, rigid structures, forming the sickle shape and leading to various complications (17). The presence of diseases caused by allosteric dysfunction opens the possibility of designing allosteric drugs that bind distant regulatory sites in proteins as opposed to classical orthosteric binding (18).

The first model of allostery was proposed by Monod, Wyman, and Changeux (MWC) who proposed a statistical mechanical model of allostery where the protein switches conformational state upon ligand binding (2). The model depends on two main tenets: (1) the protein is an

oligomer involving several subunits, and (2) the existence of a conformational equilibrium (between at least two states, usually called relaxed (R), and tense (T) states) that is shifted (or stabilized) by ligands. G-protein coupled receptors (GPCRs, the model system used in this work) are one of the cases that show allosteric function as monomers (19). In this case, the different trans-membrane helices are thought of as a "mini-oligomer". In short, the MWC model was able to capture the essence of allosteric regulation in a multitude of proteins, and has such found widespread use (20; 21; 22). The MWC model was so influential that it is still used and reviewed 50 years after its inception (19).

An alternative hypothesis to understand allostery is the sequential model put forward by Koshland, Nemethy, and Filmer (23). The model is sequential in that the substrate binds to the molecule in an induced fit fashion, which "induces" the switch from the tense to the relaxed state.

The next jump in perspective was the dynamics based view of allostery, where the relaxed and tense states of proteins can be structurally similar or even (almost) identical. The difference between the two can be discerned by fluctuations of the system upon ligand binding. Cooper and Dryden (24) performed revolutionary work to advance this view, driving the understanding of allostery away from a purely structural view to an ensemble view. Under this view, a change in the population (or the *average* structure) is not required for allostery to occur; it is rather the change in the *distribution* (in terms of frequencies and amplitudes of atomic motions) around the average structure that drives allostery, affecting the binding affinity at the allosteric (distant) site (24).

The ensemble view of allostery has been the focus of a great deal of research since the recent rise of interest in controlling proteins using allostery (25; 26; 27; 28; 29). Other scientists, such as Cui and Karplus, have argued that the "new view" of allostery, emphasizing “population shifts,” is, in fact, an “old view” (30), arguing that the idea of a “population shift” was the basis of the original model of MWC, and that what the "new view" offers is an emphasis on the intrinsic dynamic nature of proteins.

Coupled with the study of allostery is the concept of frustration (to the extent that there was a published perspective titled: "Allostery Frustrates the Experimentalist" (31)) we would like to tell the authors of this perspective that allostery also frustrates the computationalist. A physical system, such as a protein, is frustrated "when it is impossible to simultaneously optimize all the possible interactions" (31). Frustration, like allostery, is seen as a general property of proteins that is crucial for function (32). In essence, the authors reached the conclusion that measurement of allostery is method dependent, especially allostery due to entropic shifts without conformational change in single domain proteins (31). In another review that seeks allosteric networks in PDZ domains, the authors conclude that despite two decades of both computational and experimental studies on PDZ domains (PDZ domains being a classical system to study allostery), "our analysis highlights the contradiction between the different methods and calls for additional work to better understand these allosteric phenomena"

(33). This sentiment was shared in a recent CECAM (Centre européen de calcul atomique et moléculaire) workshop in Oct. 2023 concerning allostery, where during a roundtable discussion, the attendees could not agree on how to store dynamical data that could be used to learn about allostery in different systems.

0.3 Thesis objectives

Given the challenging nature of studying allostery, we will investigate allostery using an integrated approach that combines computational methods with protein design and experimental validation in G-protein coupled receptors (GPCRs).

The first step to achieving this objective is finding a model to describe allostery in GPCRs, which would encapsulate developing an approach to describe this allostery on a residue level, which would then enable design. The underlying assumption is that allostery takes the form of pathways that connect the extra-cellular (EC) and intra-cellular (IC) domains of GPCRs (16).

Due to the variability and method dependence of allosteric residue determination (33; 31), this necessitates developing an multi-faceted approach to tackle the problem from several angles: (1) allosteric transmission through pathways, (2) quantifying allosteric interactions via perturbation response far from the perturbation site, and (3) in-silico mutagenesis to assess stability of the mutant receptor before experimental validation.

To validate the allosteric pathways, we design amino acids along the pathway and observe the response to the mutation in cell-based assays. The assumption is that given the enormous design space of a protein the size of a GPCR, the ability to allosterically modify GPCR signaling supports the validity of the pathway description of allostery in GPCRs.

To bring the work a step closer to real-life application and further validate the underlying methods. we study ligand specific allosteric pathways with the aim of achieving ligand selective designs.

The developed methods are used for dynamic characterization of designed receptors and peptide-receptor complexes, and for quantifying the effects of allosteric modulators on pathways, dynamics, and correlations.

Finally, we attempt to connect the proposed allosteric description to genetic variation and disease causing mutations.

0.4 Thesis structure

In the following chapter (Ch. 1), we review G-protein coupled receptor (GPCR) structure, activation, and function. We also present the underlying methods used in this work. We finally review the literature surrounding allostery and methods of studying it, both experimental and

computational.

The following chapters, which form the bulk of the thesis, are structured by dividing the work into method development (**Part I**) and applications of the developed methods (**Part II**).

Part I presents methods developed as part of my PhD to study allostery. The main contribution is the software (AlloDy), which quantifies allosteric signals in proteins using molecular dynamics simulations. **Part I** presents the architecture of the code, theory behind the method, and benchmarks. **Part I** also contains comparisons between metrics extracted using AlloDy and experimental observables, such as NMR chemical shifts and ligand activation.

Part II presents three major applications to the methods developed in **Part I**. The first application (Ch. 3) uses the concept of allosteric pathways to engineer ligand specific responses in dopamine receptors and explains the observed experimental effects using a combination of perturbation response and pathway models. The second application (Ch. 4) uses the concept of allosteric pathways to explain how peptide-ligand receptor designs attain potent signaling responses through dynamic conformational ensembles in chemokine receptors. The third application (Ch. 5) compares genetic variation and single nucleotide variants (SNVs) with allosteric and evolutionary scores.

Finally, in the last chapter, we summarize the major findings as well as the contributions of this thesis, discuss the limitations of the study, and propose directions for future research.

1 Literature Review

"Everything that living things do can be understood in terms of
the jigglings and wigglings of atoms."
— Richard Feynmann

1.1 G-protein coupled receptors activation and function

G protein-coupled receptors (GPCRs) represent one of the most diverse and ubiquitous superfamilies of cell surface receptors in eukaryotic organisms. These receptors play a pivotal role in cellular communication by transducing extracellular signals into intracellular responses, thus serving as key molecular switches in numerous physiological processes. The discovery and characterization of GPCRs have not only revolutionized our understanding of signal transduction but have also unlocked exciting prospects for drug discovery and therapeutic intervention.

The sheer diversity of GPCRs is striking, with over 800 different GPCR genes identified in the human genome (around 3% of protein encoding genes). This diversity is reflected not only in their ligand specificity but also in their tissue distribution and functional roles. GPCRs serve a wide variety of physiological functions, from sensing the environment through rhodopsin and olfactory receptors to mood regulation through dopamine and serotonin receptors to immune system regulation via chemokine receptors (Fig. 1.1). Given the scope of the processes they regulate, they are involved in many diseases and thus are very attractive drug targets, being the targets of 34% of FDA approved drugs (34).

1.1.1 *In-vivo* function

As integral membrane proteins, GPCRs serve at the frontier of the cell. They sense external cues about the environment and then activate cellular responses. These external stimuli can be light (photons), small molecules, peptides, or even other proteins, making GPCRs highly diverse receptors in terms of both stimulus and function (35) (Fig. 1.1a).

Canonical signaling of GPCRs, as is suggested by their name, happens through coupling and activation of G-proteins (also known as guanine nucleotide-binding proteins). The G-protein binds to a GPCR, and then the GPCR acts as a guanine nucleotide exchange factor, exchanging guanosine diphosphate (GDP, inactive state G-protein) to a guanosine triphosphate (GTP, active state G-protein). G-protein are heterotrimers containing an α, a β, and a γ subunit, and upon nucleotide exchange, the α subunit dissociates from the $\beta\gamma$ subunit and each elicits further downstream signaling (Fig. 1.2). Given that GPCRs regulate many different physiological processes, this variability is visible in G-proteins as well. There are 16 different Gα subtypes, divided into 4 general classes: Gα_s, G$\alpha_{i/o}$, G$\alpha_{q/11}$, and G$\alpha_{12/13}$. Gα_s is the stimulatory G-protein which activates various adenylyl cyclases that then stimulate cyclic adenosine monophosphate (cyclic AMP, or cAMP) production from adenosine triphosphate (ATP). Gα_i, on the other hand, is the inhibitory G-protein that inhibits adenylyl cyclase activation (36)(and thus cAMP production). Gα_q stimulates phospholipase C (PLC) leading to downstream calcium release from the endoplasmic reticulum into the cytoplasm via second messenger signaling (37). Finally, $G\alpha_{12/13}$ are involved in regulation of the actin cytoskeleton through Rho family GTPase signaling (38). Traditionally, GPCRs are seen to have "cognate" G-protein pairing, where a GPCR is selective to a specific G-protein subtype. Recent evidence has shown that many GPCRs have a level of promiscuity, coupling to and activating more than one G-protein subtype (39; 40; 41), as seen in Fig. 1.1b. This complicates attempts to elucidate physiological effects of every G-protein pathway. In effect, GPCR activation can lead either to stimulation of inhibition of cAMP production, calcium ion release into the cytoplasm, or cytoskeleton regulation, all of which have a wide range of effects in cells.

G$\beta\gamma$ acts functionally as a monomer as it is a tightly bound heterodimer where the subunits have not been shown to function separately (42). The G$\beta\gamma$ complex plays two main roles: when it is bound in the heterotrimeric complex, it is a negative regulator that increase Gα's affinity to GDP (43) (and thus favoring the inactive state of Gα). After GDP-GTP exchange, G$\beta\gamma$ separates from Gα and signals on its own as a dimer (44) (Fig. 1.2). G$\beta\gamma$ has been shown to regulate a diverse array of downstream effectors such as G protein-gated inward rectifier channels (GIRKs) (45), adenylyl cyclase (46), phospholipase C (47; 48), calcium channels (49), and gene transcription (50).

Another important intracellular binding partner of GPCRs is the arrestin family. Contrary to the diversity of G-proteins, the human genome encodes only four arrestin genes, two of which are visual arrestins (arrestin-1 and arrestin-4) specific for rhodopsin and cone opsins, and the other two (named β-arrestin 1 and 2) interact with the majority of GPCRs (51). Arrestins function mainly in receptor desensitization and internalization. After heterotrimeric G-protein activation, GPCRs can be desensitized as a form of adaptation to a persistent stimulus for example. The first step is phosphorylation of the active receptor by G protein coupled receptor kinases (GRKs), which increases arrestin affinity for GPCR binding. Upon binding the receptor, arrestin impedes G protein signaling by occluding the intracellular binding site and targets GPCRs for internalization by linking the receptor to internalization machinery, such as clathrin (52; 53). This would lead either to receptor degradation or recycling back to the membrane

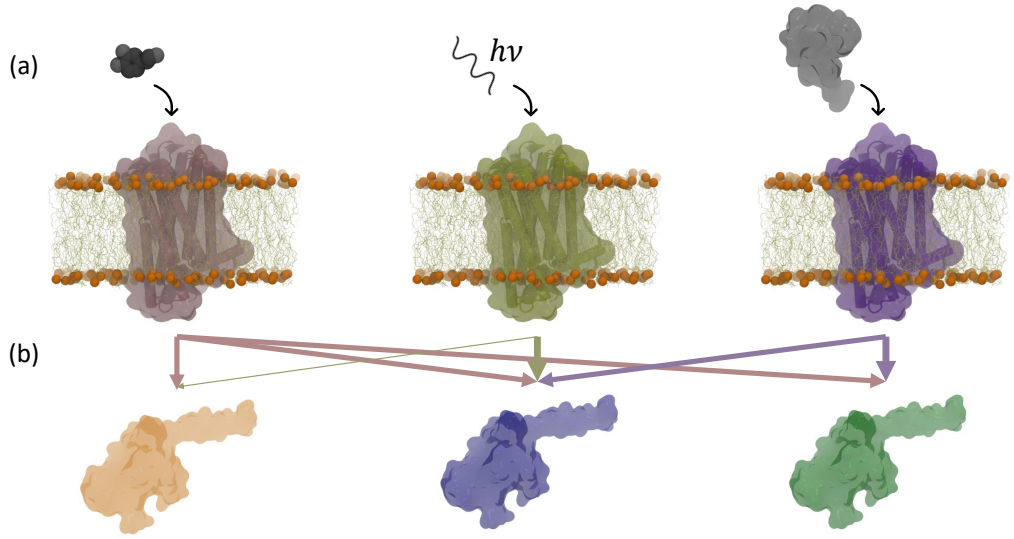


Figure 1.1: GPCR ligand diversity and promiscuity of intracellular binding partners. **(a)** GPCRs can bind/sense a wide array of ligands such as small molecules (pink), light (tan), or peptides/proteins (purple). **(b)** GPCRs can signal through various types of G-proteins, where some receptors have higher G-protein specificity, while other are more promiscuous and can bind to different G-protein subtypes. The arrows are colored in reference to the different receptors, where the arrow thickness represents strength of signaling of a receptor with a G-protein subtype.

(54). Arrestins have also been shown to play a role as regulators of multiple signaling pathways (55)

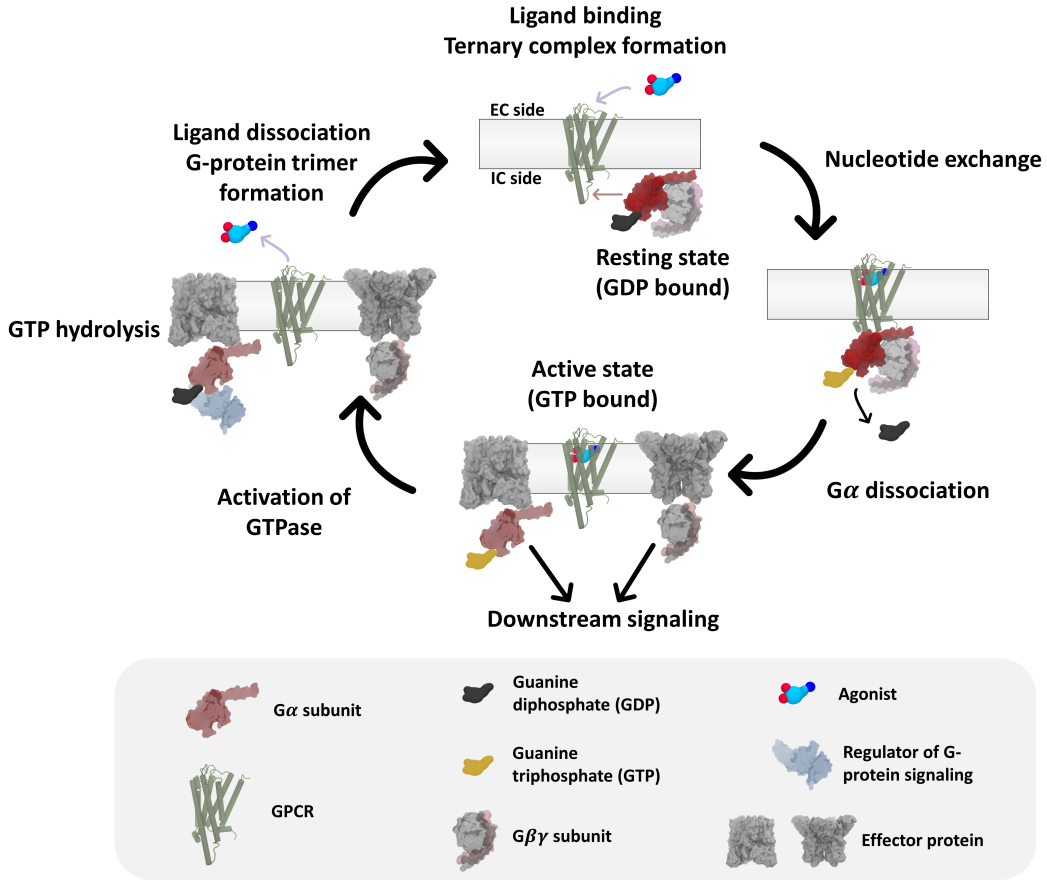


Figure 1.2: **GPCR signaling through the G-protein pathway: Top** formation of agonist and GDP bound ternary complex. The resting state refers to the GDP-bound G-protein. **Right** nucleotide exchange step, where a GDP is exchanged for a GTP in the G-protein α subunit. **Bottom** dissociation of G-protein subunits from the receptor and from each other, yielding a Gα subunit and a tightly bound G$\beta\gamma$ dimer. **Left** Activation of GTPase followed by GTP hydrolysis, allowing for the regeneration of the active form of the Gα subunit, which will re-associate with a G$\beta\gamma$ dimer, forming the resting trimer which could bind to GPCRs again. A ligand bound receptor can activate more than a single G protein before shutting down.

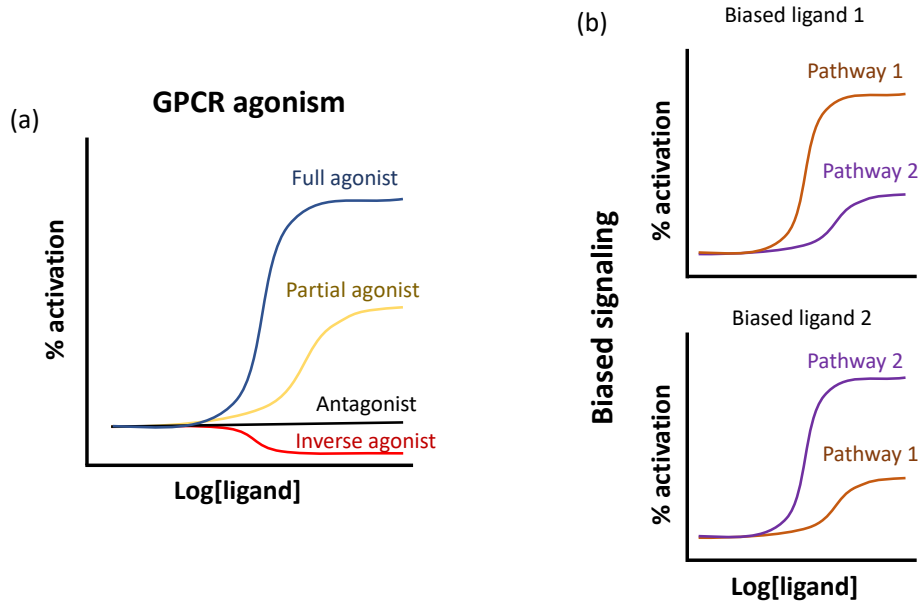


Figure 1.3: **(a) GPCR agonism:** classes of drugs are classified according to their effect compared to a high efficacy endogenous ligand. Full agonist have the highest signal seen with an experimental assay. Partial agonists have positive signals that are lower than full agonists. Antagonists bind the reeptor but have no effect on activity. Inverse agonists lower the basal activity of the receptor. **(b) GPCR biased signaling:** An example of biased signaling where both ligands 1 and 2 activate the same receptors, but ligand 1 prefers downstream effector pathway 1 (for example, a G-protein) while ligand 2 activates pathway 2 more strongly (for example, a β-arrestin pathway)

1.1.2 GPCR activation

In recent years, we have gone a long way in understanding the molecular underpinnings of GPCR activation due to the number of active GPCR structures that have been discovered (56; 57; 58; 59; 60; 61; 62) coupled with NMR (63; 64; 65) and computational studies (especially MD simulations (66)). One of the simplest models of GPCR activation is that agonist binding stabilizes the receptor in its active state, and this conformation favors interaction with either a G-protein or β-arrestin. Deeper investigation shows that GPCRs are not simple on/off switches, however, as most GPCRs exhibit some basal activity even in the absence of a ligand. This indicates that there is an equilibrium between active and inactive populations. In this context, an agonist is a ligand that favors the active state and increases the downstream activation beyond the basal level, and depending on how much it increases the activity, it could be categorized as a full or partial agonist (Fig. 1.3a).

An inverse agonist does the exact opposite, reducing the activity below the basal level, thus favoring an inactive conformation of the GPCR. Antagonists, while not affecting the basal activity, compete with agonists and inverse agonists for the ligand binding site, and thus favor the "neutral" state of the GPCR. Since GPCRs can bind to either a G-protein (or several G-proteins,

depending on the GPCR (39; 40)) or β-arrestin down-streaming pathways, agonists could possibly stimulate both pathways equally or preferably toward one pathway in a phenomenon known as biased agonism (67; 68).

GPCR activation as a function of ligand concentration is traditionally represented as a dose response curve (more details are given in Sec. 1.3.1). The response of the ligand resembles a sigmoid curve that can be essentially described with three quantities, the basal activity (response at zero concentration of any ligand), the maximum response (compared to an endogonous ligand), also known as efficacy, and the half response ligand concentration EC_{50}, also known as potency.

1.1.3 Biased signaling

As previously mentioned, GPCRs can bind more than one type of intracellular binding partners, and their signaling can be discriminatory toward one downstream effector (Fig. 1.3b). Biased signaling can be the result of several factors: the receptor itself can be biased (69; 70); alternatively, the agonists could stabilize a certain conformation of the receptor that favors a specific binding partner (71). The system itself could be biased, with the "system" being defined by non-ligand molecules involved in the signalling process. Examples of system bias are different transducer, effector, or modulatory protein concentrations across cell types or different tissues (72).

Specifically, biased agonism is a hot subject of study since it promises therapeutic applications with reduced side effects. For example, mouse studies of μ-opioid in mice lacking β-arrestin 2 showed enhanced analgesia (73). This has lead to an increased interest in computational studies for elucidating the molecular mechanism (74) that would lead to design of novel therapeutics (75).

1.1.4 Structural features

GPCRs are also known as seven-transmembrane domain receptors because of their characteristic seven-transmembrane alpha helical structure (TM1-TM7) with the N-terminus toward the extracellular (EC) side and the C-terminus toward the intracellular (IC) side. The seven helices form a cavity in the plasma membrane where a ligand binds the EC cavity, and an IC binding partner binds at the IC side. The helices are connected via three intra-cellular (ICL1-3) and three extra-cellular (ECL1-3) loops which vary in length between receptor types, even those belonging to the same subfamily (Take ECL2 and ICL3 of dopamine D1 and D2 receptors, for example). After TM7, there is a shorter intracellular helix that is parallel to the membrane (commonly known as helix 8). H8 is followed by a flexible C-terminus that is variable in length, and contains several post-translational modification (PTM) sites. A general overview is shown in Fig. 1.4.

Generic GPCR numbering: Due to the conserved seven-transmembrane helical structure of

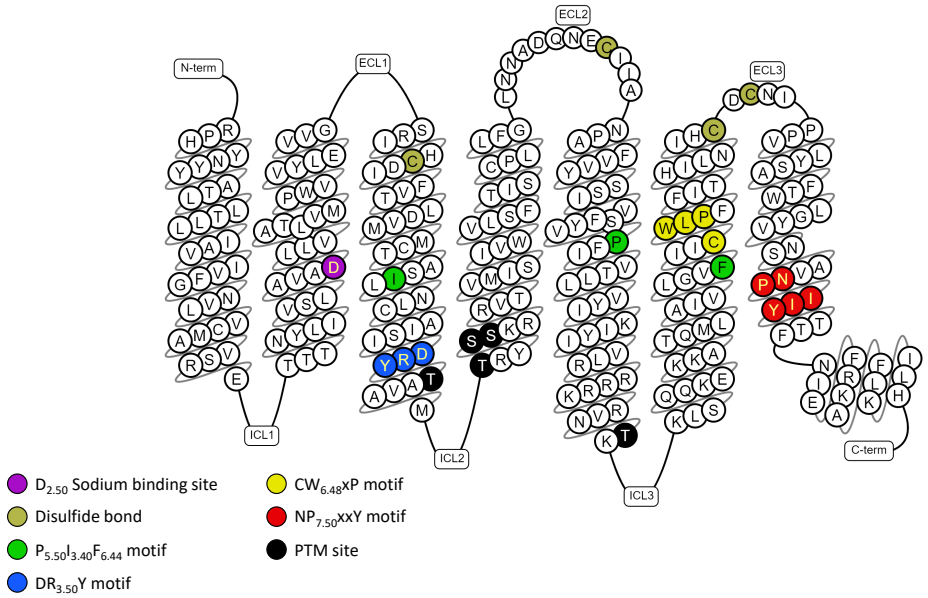


Figure 1.4: GPCR snake plot featuring structure of dopamine D2 receptor, conserved motifs in class A GPCRs, conserved sodium ion binding site, and common post-translational modification (PTM) sites. Residue numbers follow the Ballesteros-Weinstein designation, where the first number is the TM helix and the second the position of the residue compared to the most conserved residue in the helix, designated 50 (1). Generated with GPCRdb (76) (https://gpcrdb.org/protein/drd2_human/)

GPCRs, generic numbering schemes that reference the position on the helix rather than the sequence residue number are very helpful when comparing residue positions across receptors. The most common numbering scheme that is used in this work is the Ballesteros-Weinstein (BW) scheme (1), which exploits the presence of highly conserved residues in each of the seven TM helices in class A GPCRs. The residue number consists of two numbers, the first referring to the helix 1-7, and the second being the position of the residue compared to the highest conserved residue in the helix, denoted 50. For example, 7.50 would be the most conserved residue on helix 7, and 7.53 would be the residue 3 positions away from 7.50 toward the C-terminus. The positions of the most conserved motifs in class A GPCRs are shown in Fig. 1.4.

The transmembrane helices: The seven TM helices form a tight interacting fold that serves as a communication channel between the ligand binding and G-protein (or IC binding partner more generally) binding site. A consensus analysis of various GPCRs unveils a conserved network of 24 interactions between TM helices that are mediated by 36 topologically equivalent amino acids (77) (generic numbering is helpful in this case to show this equivalence, keep in mind that this analysis was from 2013, when a much smaller number of GPCR structures was available). Included in the list of the mentioned amino acids are highly conserved residues such as $N^{1.50}$, $D^{2.50}$, $W^{4.50}$, and $P^{7.50}$ (Fig. 1.4), while other topologically equivalent positions

are more variable between GPCRs. The consensus inter-TM contacts are localized toward the middle to cytoplasmic side of the GPCR. At the center of the TM bundle is TM3, which acts as structural and functional hub that maintains consensus contacts with TM6,5,4, and 2 in addition to non-consensus contacts with TM7 and TM1. This role is aided by the tilt angle of TM3 with respect to the perpendicular to the membrane plane (~ 35°) (77). Another key feature in GPCR TM helices is the presence of conserved proline residues in TM5, 6 and 7 at positions $P^{5.50}$, $P^{6.50}$, and $P^{7.50}$ that introduce kinks by interrupting the intrahelical hydrogen bonding pattern. Other kinks are found at non-conserved positions too. These features are essential for facilitating functional state formation as a response to ligand binding by lowering the energy barriers of conformational changes in the TM helices.

GPCR post-translational modifications (PTMs): Almost all GPCRs are post-translationally modified, although their presence and number varies by receptor. PTMs range from glycosylation sites on ECLs to phosphorylation sites on ICL3 and C-terminus to palmitoylation sites on H8. Glycosylation contributes to GPCR trafficking to the cell membrane and receptor stability (78; 79). Phosphorylation sites affect receptor internalization and desensitization after activation, as we saw previously with arrestin signaling (80). Recent research has shown evidence for different patterns of phosphorylation (the so called phosphorylation barcode) which would modulate arrestin conformation, and thus result in unique functional outcomes (such as desensitization, internalization, or signaling) (81). Palmitoylation plays an important role in membrane targeting and anchoring (82; 83; 84), and mutation of a palmitoylation site in CB1 receptor also impaired its signaling properties (84). Finally, it is important to note that the world of PTMs is vast, and a lot about them is still unknown. There is interplay between PTMs that produces unique effects (79), but that it out of scope for this review.

Extra-cellular loops (ECLs): ECLs play important roles in ligand binding and specificity, and are also potential binding sites for allosteric modulators (85). ECL1 has a highly conserved length (but divergent sequence) in class A and class C GPCRs, and variable length and sequence in class B. Despite its short length in class A, it has been reported to influence the shape of the binding pocket (86; 87; 88) and to affect signalling efficacy of an allosteric agonist (89). ECL1 is also the home of the WxFG motif in almost 90% of class A GPCRs (88). Mutations in the WxFG motif were reported to be capable of ligand binding, but lack signaling responses (86), and mutation of the W impairs receptor trafficking (88). ECL2, on the other hand, shows surprising structural diversity in class A GPCRs, varying from unstructured short loops (DD2R) to alpha helices (β2AR) to beta hairpins (CXCR4, rhodopsin). ECL2 contributes to ligand selectivity (which may explain its structural and sequence diversity, (90; 91)), receptor function (92), ligand binding (93), allosteric modulation (94; 89; 95), and biased signaling (96; 97), and it has been described as a "gatekeeper" of receptor activation. A conserved disulfide bond between ECL2 and TM3 forms a structural constraint across the GPCR superfamily (98). In some GPCRs, the presence of the disulfide bond is essential for signaling, such as C5aR (99) and chemokine CCR8 (93), while in others, such as adenosine A2A receptor, its presence seems to be functionally redundant (100). The final extracellular loop is ECL3, which is small in class A GPCRs. ECL3 forms an intra-loop disulfide bond in cases such as dopamine (101), serotonin

(102), and melanocortin receptor families (103), while it forms a disulfide bond with the N-terminus in other receptors, such as the chemokine family (104) and Angiotensin AT1 (105) (instead of bombarding the reader with a river of references for the structures of every single receptor mentioned, we would invite the reader to check the structures themselves by looking them up on GPCRdb https://gpcrdb.org/structure/) Through those structural constraints, ECL3 plays an important role in stabilizing the ligand binding pocket for aiding N-terminus contacts with the ligand in case of chemokine and AT1 receptors. ECL3 is the least studied between the ECLs, but there is evidence of its involvement in allosteric ligand binding (106) and interactions with ECL2 that are critical for proper folding and signaling (107).

Intra-cellular loops (ICLs): Intra-cellular loops (ICLs) play a wide range of roles in GPCRs, such as IC binding partner specificity, autoregulation, and PTMs. GPCRs contains three ICLs (owing to the presence of seven TM helices), with different loops playing separate roles. ICL1 forms interactions with H8 with a few exceptions (AT2-R for example). Chimeras of AT1 with ICL1 of AT2 and vice versa showed a loss of function of AT1 toward G-q and toward β-arr recruitment, which implies a role for ICL1 toward those pathways in AT receptors (108). ICL1 has also been shown to be used in an "alternative" conformation of β-arrestin binding (109). The second ICL plays a major role in both G-protein and arrestin binding, with the "major" arrestin pose extracted from various structural studies (Fig. 1.8) showing ICL2 in a helical conformation resting in a cleft in the arrestin structure. ICL2 is also involved in GDP release, where mutations in ICL2 in rhodopsin maintain Gt coupling but impair GDP release (110). ICL2 has also been suggested to be a fine-tuning switch in 5HT2AR that can distinguish modes of receptor activation in response to hallucinogenic (such as LSD, shown in Fig. 0.1) and nonhallucinogenic ligands (such as serotonin, also known as 5-HT), where ICL2 conformations depend on the ligand bound in MD simulations (111). The third ICL is the largest and most variable in GPCRs, ranging between 10 and 240 amino acids in length. A recent study has found a relationship between ICL3 length and receptor–G protein binding site conservation. In short, shorter ICL3s tend to have a broad distribution of interface conservation, while longer ICL3s tend to have narrower (and lower on average) interface conservation. The length threshold is reported to be around 46 AAs. The length of ICL3 is also related to how GPCRs achieve specificity, either via the G-protein interface for shorter ICL3s, or via ICL3 gating for the longer ones, thus making ICL3 a determinant of G-protein selectivity (112).

The N and C-termini: Amino-terminus (N-terminus) is a region that is variable in length and is classically involved in receptor trafficking (113) and ligand binding in class B (secretin-like) (114) and class C (glutamate-like) (115) GPCRs. It also binds to peptide and protein ligands in some class A GPCRs and is stabilized with a disulfide bond to ECL3 (70; 116; 117). Furthermore, the N-terminus affects GPCR signaling through proteolysis and other proteolysis-independent modalities summarized here (118). A very interesting and particular case is the N-terminal region in adhesion GPCRs (class B2), which has a large extra-cellular region that contains a conserved G-protein-coupled receptor (GPCR) autoproteolysis-inducing (GAIN) domain and a variable adhesion ligand binding domain (119; 120). The GAIN domain contains a tethered peptide agonist that is often auto-cleaved (121), and is sensitive to mechanical

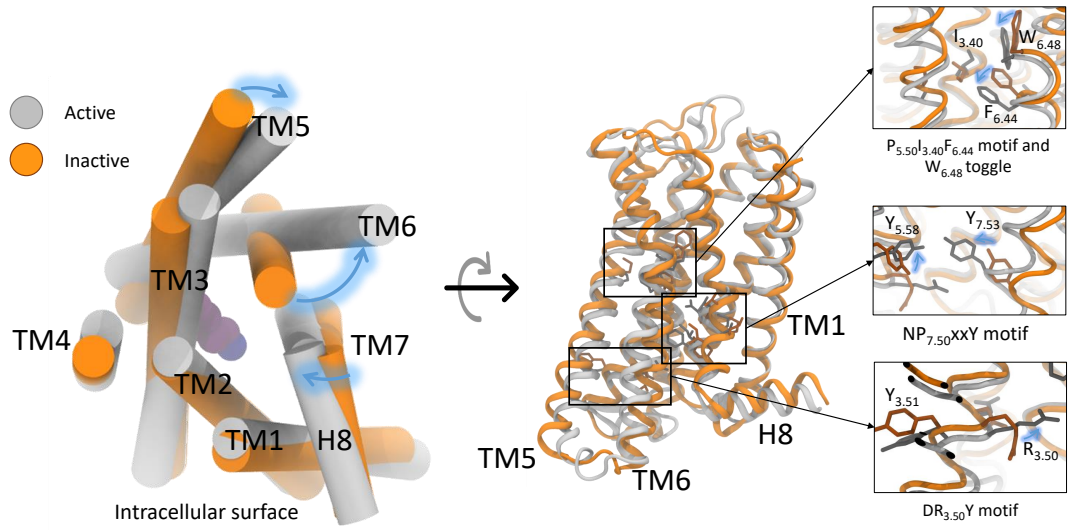


Figure 1.5: **(left) GPCR activation in a snapshot:** Intracellular view of active (grey, PDB code: 6VMS) and inactive (orange, PDB code: 6CM4) GPCR (dopamine D2R) structures. Main motions of TM5, TM6, and TM7 are highlighted. **(right) Conserved microswitches:** motions of highly conserved microswitches involved in class A GPCR activation

signals (122; 123; 124) but can also activate the receptor via spontaneous tethered agonist exposure(125).

The disordered carboxyl terminus (C-terminus) is a substrate for GPCR kinases and also binds β-arrestins (126). Structures of GPCRs bound to β-arrestins clearly show the phosphorylated C-terminal peptide interacting with β-arrestin1, rendered in Fig. 1.8c (127; 128; 129; 130). Recent studies on β2AR have shown that C-terminus inhibits basal activity and agonist signaling in cells lacking β-arrestin. It limits interactions between the GPCR and G-protein by being negatively charged and interacting with the positively charged cytoplasmic side of the receptor, thus acting as an autoinhibitory factor (131).

Structural features of GPCR activation

As mentioned previously, the advent of high resolution cryo-EM has opened the door to elucidation of structures of active tertiary (ligand - GPCR - IC binding partner) complexes, which provides great insight to the structural features that define GPCR activation when compared to inactive or intermediate state structures (mostly solved using X-ray crystallography). When these features are combined with spectroscopy data and computer simulations, we arrive at a molecular understanding of GPCR activation.

The most notable feature of GPCR active state structures is the opening of the cytoplasmic side

to accommodate binding of its IC binding partner (a G-α subunit or an arrestin, Fig. 1.5:left). The staple of this opening is an increase in the distance between TM6 and the stationary TM3, where the exact distance varies depending on the binding partner (larger for G-αs (57) and smaller for G-αi (132), as shown in Fig. 1.7). TM7 also kinks to form interactions with the G-protein, moving closer toward the binding cavity along with H8. At the ligand binding site, we observe a contraction in response to presence of an agonist compared to an apo state, although the differences in the ligand binding site between an agonist and an antagonist/inverse agonist bound state are subtle (133). Interestingly, structural and computational data have revealed that ligand and G-protein binding sites of GPCRs can each sample active or inactive conformations with a degree of independence from each other. This can be seen structurally in agonist bound structures of 5HT1B (102), where cytoplasmic part of the receptor is inactive like despite binding an agonist, and can be observed in the conformations of agonist bound β2-AR MD simulations (66).

In addition to large scale motion of TM6 and TM7 in GPCR activation, smaller scale movements at the amino acid scale, termed microswitches, are critical to GPCR activation and function. Many of these microswitches display correlated motion across sparse networks that connect the EC and IC domains of a GPCR. Moreover, since these microswitches form a unique set of contacts between the inactive and active states, they are ideal targets for protein design (16).

Conserved motifs in class A GPCRs: Of the aforementioned microswitches, some are highly conserved across class A GPCRs and form conserved motifs that are outlined here and shown in Fig. 1.5:right:

Toward the G-protein binding site, one finds the D(E)$^{3.49}$-R$^{3.50}$-Y$^{3.51}$ motif. In the inactive state, R$^{3.50}$ interacts with E$^{6.30}$, either forming a salt bridge (called an ionic lock), such as the case of dark-state rhodopsin (134) or simply interacting at a distance, such as β_1AR inactive state. MD simulations of β_2AR have shown that ionic lock may form temporarily and then dissociate (66; 135), which indicates this lock forms transiently in GPCRs with basal activity. Nonetheless, it plays an important role in stabilizing the inactive state. That is why TM3-6 distance is considered to be one of the order parameters of GPCR activation. There is also an extensive role that the DRY motif plays in forming stabilizing interactions with the G-protein in the active state, as seen in β2-AR active state structure (57). R$^{3.50}$ has been observed to form hydrogen bonds with Y$^{5.58}$ in active state structures of rhodopsin and μ-opioid receptors, and in β2-AR, Y$^{5.58}$A variant does not activate G-s and does not display basal activity (136).

Another conserved motif in GPCRs is the NP$^{7.50}$xxY motif on helix 7. While it does not interact directly with the G-protein, it is critical for forming the active state. Due to the proline at position 3.50 that acts as a helix breaker, TM7 can rotate in the active state, moving Y$^{7.53}$ to a position previously occupied by TM6 in the inactive state. Y$^{7.53}$ then forms a hydrogen bond with Y$^{5.58}$. Due to this motion, the NPxxY RMSD to the inactive state could be used as an order parameter to measure GPCR activation, as will be seen in later sections of this work.

Toward the core of the receptor, the P$^{5.50}$ I$^{3.40}$ F$^{6.44}$ motif, whose side chains act as toggle

rotamers that respond to agonist binding and help transmit the "activating" signal across the receptor. An inward movement of TM5 toward the orthosteric binding site requires $I^{3.40}$ to adopt a different rotamer to avoid steric clashes (see Fig. 1.5). This leads to a shift in $F^{6.44}$ to maintain packing while TM7 moves inward. The conserved $W^{6.48}$, known as Trp toggle, also responds with a concerted motion with the aforementioned residues. These motions are facilitated by the absence of backbone hydrogen bonds caused by the presence of prolines at conserved positions in TM5, TM6, and TM7.

Ligand binding site

The orthosteric ligand binding sites vary between GPCRs given the wide array of diverse ligands that they sense. We will focus on aminergic receptors as those are the class A GPCR families that we mainly deal with in this work.

Aminergic receptors mainly bind their ligands through hydrogen bonding primarily with serines or threonines on TM5 in positions 5.42, 5.43, and 5.46, a salt bridge interaction between the charged nitrogen and $D^{3.32}$, which is fully conserved in human class A aminergic receptors, and $Y^{7.43}$, which forms a hydrogen bond with the aforementioned $D^{3.32}$ and the ligand (Fig. 1.6a). Residue $D^{3.32}$ is essential for ligand binding and could abrogate ligand binding if mutated (137; 138). Contacts with TM5 and TM6 in dopaminergic and serotonergic receptors has been shown to control full agonist vs biased agonist response, where contacts with TM5 only gave a G-protein biased response, while ligand interactions with both TM5 and TM6 (specifically position 6.55) leads to full agonist response. The authors concluded that sequence variation in position 6.55 is nature's way of fine-tuning β-arrestin recruitment (139). Furthermore, the ligand binding site is rich with aromatic residues that could form π-π interactions with the aromatic rings of aminergic ligands. Sequence conservation of the ligand binding site in human aminergic GPCRs is shown in Tab. 1.1.

Table 1.1: Sequence conservation of ligand binding residues in human class A aminergic GPCRs. Residues with highest conservation level are shown in sequence consensus row, with the percentage conservation shown. Additional designations are shown for positions with lower conservation (< 70). Legend: TM = transmembrane helix, BW = Ballesteros-Weinstein numbering scheme. Generated with GPCRdb (76) (https://gpcrdb.org/alignment/render)

TM	3	3	3	5	5	5	5	5	6	6	6	7	7	7
Residue (BW)	32	33	36	42	43	46	47	48	51	52	55	35	39	43
Seq. consensus	D	V	C	S	S	S	F	Y	F	F	N	F	F	Y
	100	58	56	58	50	44	100	72	75	81	28	36	22	89
Aromatic											22	69	36	
H-bonding		22	36	83	69	58					67	31	44	
Small		0	89	75	78	86					11	8	3	
Hydrophobic		100	64	8	28	42					42	81	69	

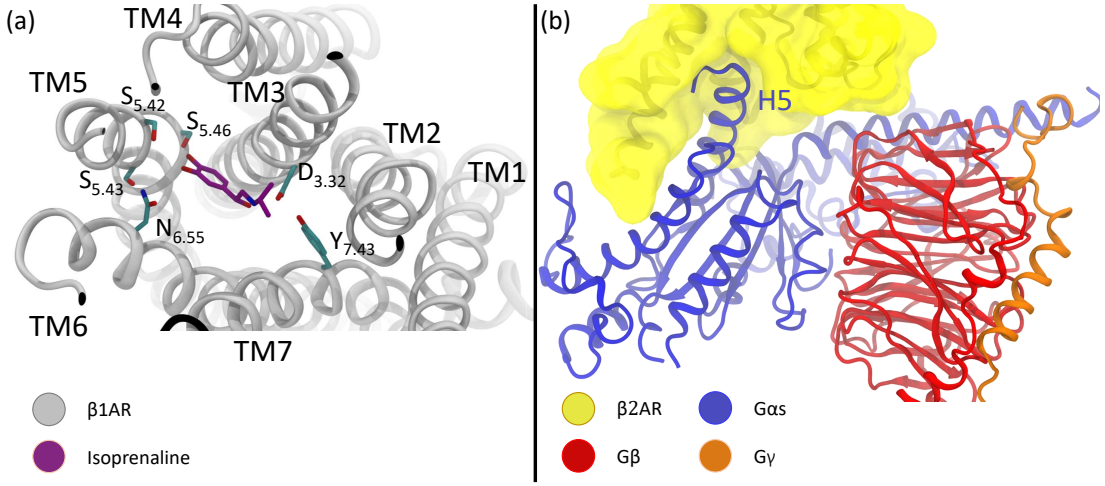


Figure 1.6: **(a) Ligand binding site of β1AR:** extracellular view of β1AR (grey, PDB code: 6H7J (140)) bound to Isoprenaline. Conserved ligand binding residues in class A aminergic GPCRs are highlighted with BW numbering (1). **(a) G-protein binding site in β2AR:** binding of β2AR (yellow, PDB code: 3SN6 (57)) to Gαs-$\beta\gamma$ heterotrimer. C-terminal helix (H5) of the Gα is highlighted.

GPCR binding to different intracellular (IC) binding partners

Given the wide variety of GPCR IC binding partners and G-protein subtypes, studies have attempted to uncover the determinants of G-protein selectivity through functional assays and molecular dynamics simulations (40; 39; 141). Dynamical simulations have shown that when a GPCR binds to its cognate G-protein, the C-terminal helix (also known as Helix 5, Fig. 1.6b) of the G-α domain (which is inserted into the cytoplasmic cavity of the GPCR) assumes a dynamic ensemble of unique orientations. Non-cognate G-proteins, on the other hand, interact weakly and dynamically with latent intracellular GPCR cavities (40). Another study built G-protein chimeras by exchanging the C termini and between Gα-subunits. They find that Gq and Gs coupled receptors display promiscuity by binding to Gi1 to some extent while Gi coupled receptors are more selective (39). One possible explanation for this difference are structural features that are unique to every G-protein variant. For example, Gs-bound active state structures have a larger opening of TM6 than Gi and Gq complexes, while Gq structures have variable TM6 opening with low TM3-7 distance (measured between $R^{3.50}$ and $Y^{7.53}$) as seen in Fig. 1.7. Gi and Go generally have a small opening of TM6 and a slightly larger TM3-7 distance compared to Gs structures. At the time of writing, there is only one class A GPCR structure bound to G13, and it sits in an intermediate position in TM3-6, TM3-7 space between all the other studied structures. The wealth of structural and annotation data now allows construction of what is called the "receptor-G protein couplome" for determining selectivity/promiscuity of GPCRs and G proteins (142). The authors find that half of GPCRs are selective for a unqiue G-protein while 5% promiscuously activate all G-protein families,

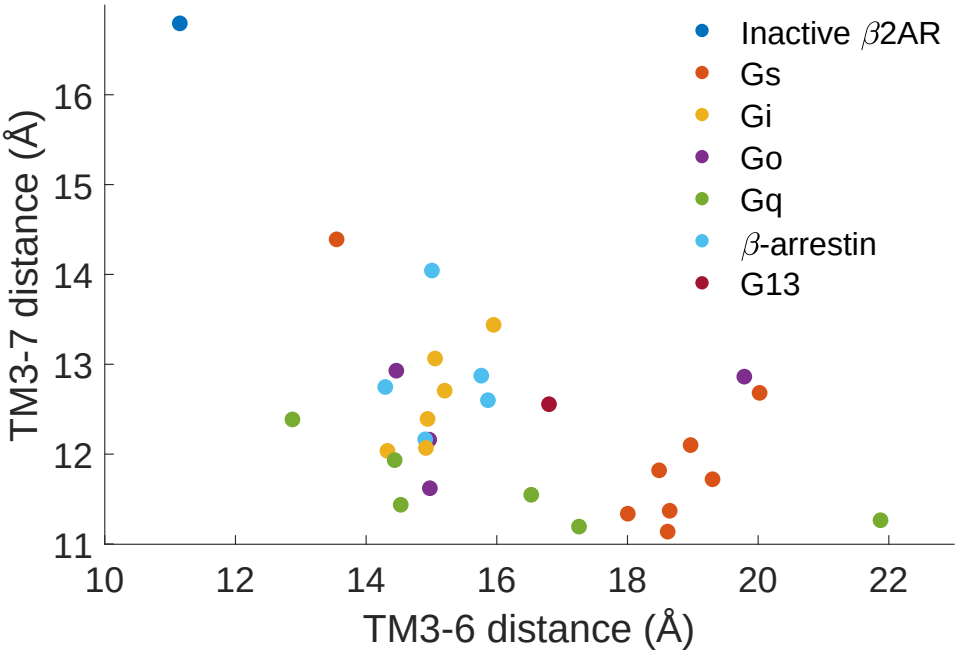


Figure 1.7: **Survey of class A ternary complexes for different IC binding partners:** TM3-6 distances are calculated between $R^{3.50}$ and $X^{6.30}$ (where X is a variable amino acid in the sampled structures), while TM3-7 distances are calculated between $R^{3.50}$ and $Y^{7.53}$. Inactive structure of β2AR (PDB code: 2RH1) is added for reference. PDB codes of ternary complexes used in this figure are: **Gs:** 3SN6, 5G53, 6GDG, 7RMH, 7PIU, 7CKZ, 7XT9, and 7XTB. **Gi:** 6CMO, 6N4B, 6OS9, 8F7Q, 8F7W, and 7JVR. **Go:** 6OIK, 6WWZ, 7EJ0, and 7W2Z. **Gq:** 6OIJ, 7F6H, 7SRR, 7DFL, 7SR8, and 8E9Z. **β-arrestin:** 6PWC, 6U1N, 6TKO, 7R0C, and 7SRS. **G13:** 7T6B. The outliers are 60IJ (X:21.9, Y:11.3, **Gq**), 7EJ0 (X:19.8, Y:12.9, **Go**), and 7RMH (X:13.5, Y:14.4, **Gs**). Note that 7RMH is bound to an engineered miniGs399 that does not signal, so the structure probably does not resemble a functional state

and that almost three quarters of GPCRs activate all G-proteins belonging to the same family. In addition, there is evidence of binding of G-proteins to the membrane through several lipid anchors (143).

Going beyond static structural features, a recent study derived a spatio-temporal code from locations and durations of GPCR-Gα contacts that are critical for G-protein selectivity (141). These contacts were divided into specific and common categories, and it was observed that promiscuous GPCRs tend to sample the common contacts more than G-protein specific ones. Between the studied G-proteins, Gs had the largest number of specific contacts, followed by Gq and then Gi .

As for β-arrestin, it binds the same IC cavity as the G-protein in the activated receptor via insertion of a finger loop, stabilizing a distinct receptor conformation. As seen in Fig. 1.7, opening of the IC cavity is comparable to binding Gi/o, although this does not tell the full story. Arrestin

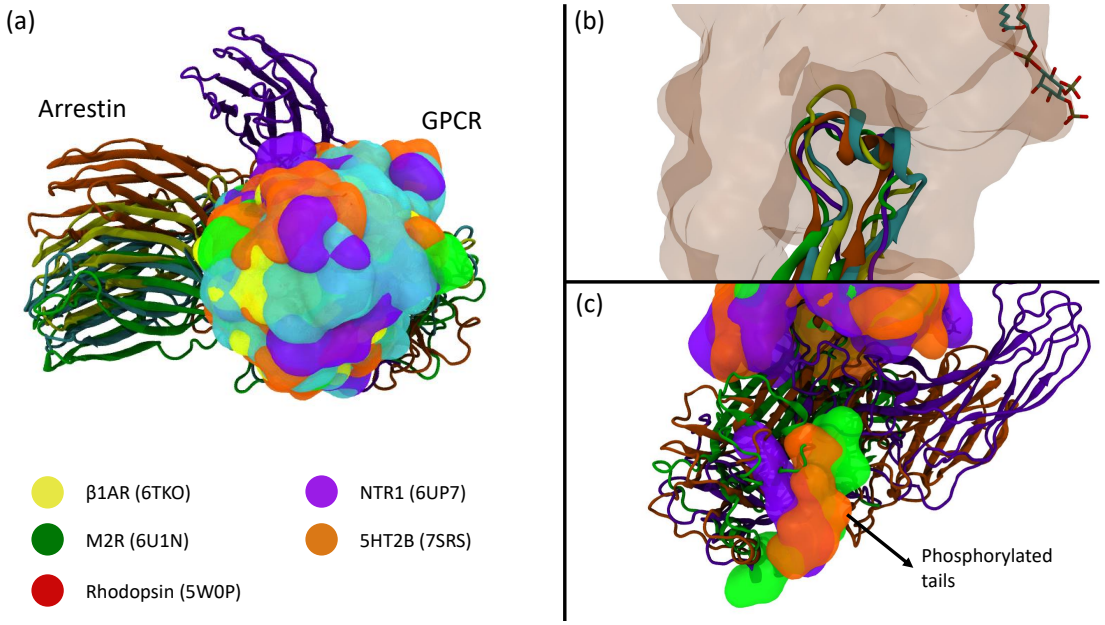


Figure 1.8: **Structural diversity of arrestin binding from available arrestin bound complexes:** arrestins are shown as cartoon, while GPCRs are represented as a translucent surface. **(a) range of arrestin orientations:** after receptor Cα alignment, arrestins span a range of orientations with respect to the receptors. **(b) Diversity of finger loop conformations:** finger loops can either be helical or unstructured. Only 5HT2B (7SRS) receptor is shown in this panel for clarity. **(c) Phosphorylated tails binding:** despite diversity of arrestin conformations, phosphorylated C-terminal tail similarly bind the positively charged N-terminus of arrestin. The tail is shown in surface representation.

binding positions ICL2 in a hydrophobic cleft (with the exception of NTSR1, which binds β-arrestin via ICL1, leading to a largely rotated arrestin structure, shown in Fig. 1.8a, (109)), and mutation I34.51A (ICL2) in 5HT2B reduced agonist induced β-arrestin-1 recruitment by 70% (130). β-arrestin forms contacts with the phosphorylated C-terminus, as indicated in β-arrestin bound structures, and hinted at by mutagensis of possible phosphorylated serines in 5HT2B C-terminus, which reduced β-arrestin-1 recruitment maximum response by 30%–50% in response to LSD. GPCRs lacking a sizeable C-terminus (such as dopamine D2-4 or 5HT1A/B) compensate with a long ICL3 containing several phosphorylation sites that binds β-arrestin. Disulfide crosslinking experiments of 5HT1A and 5HT1B provide evidence for ICL3 binding to the N-terminal domain of β-arrestin1 as well as ICL1 binding (as opposed to ICL2), which suggests an NTSR1 similar "alternate" binding conformation. In this conformation, the positively charged N-terminal domain is accessible to ICL3 (109). It is also possible for a GPCR to partially bind β-arrestin via the phosphorylated tail only (known as the 'tail' conformation) (144; 145). Such a conformation can mediate receptor endocytosis and activation, although G protein signaling desensitization requires a fully engaged GPCR–arrestin complex (146). Binding of either ICL3 or C-terminal peptide occurs at a conserved P-X-P-P motif (where P is a phosphorylation site) which interacts with a K-K-R-R-K-K sequence in the N-domain

of β-arrestins (147). A comparison between available arrestin bound structures and their G-protein counterparts highlights a downward shift of R3.50 side chain to accommodate the arrestin finger loop and a counterclockwise rotation of N7.49 away from D2.50. In addition, while the arrestin finger loop has conserved negatively charged residues across β and visual arrestins (109), the loop could assume various conformations, which in combination with arrestin rigid body rotation is hypothesized to allow arrestin adaptation to various cytoplasmic cavities across the GPCR superfamily (130). Furthermore, there is evidence that β-arrestin preassociates with the plasma membrane, driving coupling to receptors and subsequent activation (148).

Conclusion

As seen in this section, GPCRs play a diverse and pivotal role *in vivo* through the variety of stimuli that they sense and diversity of IC binding partners and downstream responses. While having a conserved 7-TM helix topology, GPCRs show a wealth of variety in the loop, terminal, and ligand binding regions. GPCRs also access a rich activation landscape, exhibiting inactive, intermediate, and multiple active states depending on the bound partners.

In the context of this thesis, we are interested in class A GPCRs as model systems for complex allosteric proteins in which allosteric behavior depends on both structural and dynamic features.

1.2 Underlying methods

1.2.1 Molecular dynamics (MD) simulations

Molecular dynamics (MD) simulations are a powerful and versatile way to study allosteric signaling in a plethora of systems and in different forms of computational experiments. MD simulations are based on solving Newton's equations of motion for particles of a system where the forces are calculated through a molecular mechanics force field in the classical case, and where particles could be individual atoms in all-atom MD, atom-hydrogen pairs in united-atom MD, or a merge of several atoms in coarse-grained MD. A molecular mechanics force field is divided into bonded and non-bonded terms, where bonded terms describe the bonds, angles, and dihedrals formed in a molecule, while non-bonded terms contain an electrostatic and a van der Waals component. Such force fields are typically derived from experimental data fitting (149) or quantum mechanical calculations (150), and more recently machine learning potentials (151).

Without modification, molecular dynamics integrators sample a constant number of particles/volume/energy surface (known as the microcanonical ensemble, or NVE). A visualization of such a trajectory and its sensitivity to inital conditions are shown in Fig. 1.9. An MD simulation can be modified to sample constant temperature via coupling to a thermostat, which

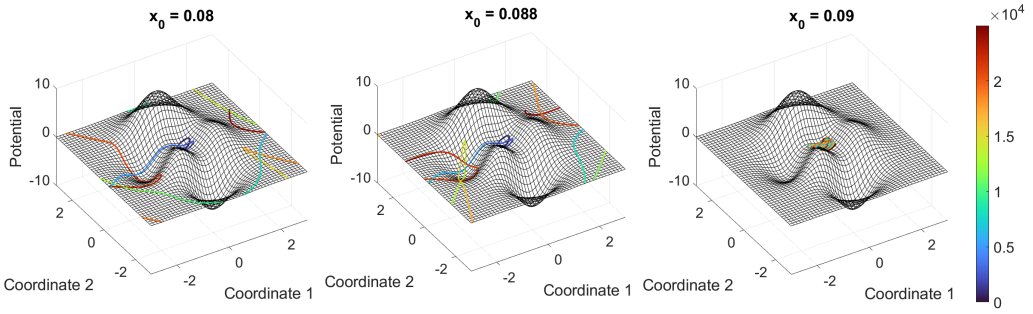


Figure 1.9: Sensitivity of molecular dynamics (MD) simulations to initial conditions. While NVE MD simulations are deterministic, they are chaotic, meaning that a small change in initial conditions can lead to drastically different trajectories. Simulations start at $y_0 = 0$ and x_0 indicated by the title of each figure. A slight change of $x_0 = 0.88$ to $x_0 = 0.9$ leads to the MD simulation getting stuck in the small potential well at the origin. Simulation was performed for 25000 steps with a time step of $dt = 0.001$ with periodic boundary conditions

will add or remove energy to a system, approximating a canonical ensemble (also known as NVT). Popular thermostats include the Nose-Hoover (152; 153), velocity rescaling (154), and Langevin (155; 156) thermostats. Additionally, constant pressure simulations can be achieved by coupling to a barostat, and with the aforementioned modifications, MD can mimic an experiment in the lab at constant temperature and pressure conditions.

In this work, we use all-atom classical unbiased MD to sample conformations of soluble and membrane proteins, and then analyze the trajectories to extract allosteric communication pathways or quantify responses to perturbations in a given system. We also used steered molecular dynamics to study the unfolding/mechanical response of a mechanosensor (122).

For an overview of how MD is used to study allostery in proteins, we refer the reader to Sec. 1.4.2.

1.2.2 Rosetta

Rosetta is a software suite containing an ensemble of algorithms and protocols for computational modeling, design, and analysis of proteins. It employs a combination of physics and knowledge-based terms to construct a forcefield for scoring protein structures, with the terms shown in Fig. 1.10 (157), and works on the assumption that the "native" structure for a given sequence corresponds to the lowest energy structure that can be formed (158; 159)(also known as the thermodynamic hypothesis or Anfinsen's dogma (160)).

Two major algorithms that are the workhorses of Rosetta are gradient-based optimization and Monte Carlo. A stochastic gradient descent algorithm minimizes the score of a protein structure by iteratively adjusting the atom positions and minimizing the energy using the aforementioned energy function. The Markov chain Monte Carlo (MCMC) algorithm, on the

Energy Term	Description	Weight	Units
fa_atr	Attractive energy between two atoms on different residues separated by a distance d	1	kcal/mol
fa_rep	Repulsive energy between two atoms on different residues separated by a distance d	0.55	kcal/mol
fa_intra_rep	Repulsive energy between two atoms on the same residue separated by a distance d	0.005	kcal/mol
fa_sol	Gaussian exclusion implicit solvation energy between protein atoms in different residues	1	kcal/mol
lk_ball_wtd	Orientation-dependent solvation of polar atoms assuming ideal water geometry	1	kcal/mol
fa_intra_sol	Gaussian exclusion implicit solvation energy between protein atoms in the same residue	1	kcal/mol
fa_elec	Energy of interaction between two nonbonded charged atoms separated by a distance d	1	kcal/mol
hbond_lr_bb	Energy of short-range hydrogen bonds	1	kcal/mol
hbond_sr_bb	Energy of long-range hydrogen bonds	1	kcal/mol
hbond_bb_sc	Energy of backbone–side-chain hydrogen bonds	1	kcal/mol
hbond_sc	Energy of side-chain–side-chain hydrogen bonds	1	kcal/mol
dslf_fa13	Energy of disulfide bridges	1.25	kcal/mol
rama_prepro	Probability of backbone φ, ψ angles given the amino acid type	(0.45 kcal/mol)/kT	kT
p_aa_pp	Probability of amino acid identity given backbone φ, ψ angles	(0.4 kcal/mol)/kT	kT
fa_dun	Probability that a chosen rotamer is native-like given backbone φ, ψ angles	(0.7 kcal/mol)/kT	kT
omega	Backbone-dependent penalty for cis ω dihedrals that deviate from 0° and trans ω dihedrals that deviate from 180°	(0.6 kcal/mol)/AU	AU
pro_close	Penalty for an open proline ring and proline ω bonding energy	(1.25 kcal/mol)/AU	AU
yhh_planarity	Sinusoidal penalty for nonplanar tyrosine χ3 dihedral angle	(0.625 kcal/mol)/AU	AU
ref	Reference energies for amino acid types	(1.0 kcal/mol)/AU	AU

Figure 1.10: Energy terms of the Rosetta energy function known as ref2015. Descriptions and weights are from (157)

other hand, aims to simulate protein folding by sampling different conformations and then evaluating the energy of each conformation. For every MCMC step, the energy of the current conformation is compared to the previous one; if its energy is lower, the step is accepted. If the current conformation's energy is higher than the previous, then the Metropolis criterion is applied to accept or reject the step with probability:

$$P(\Delta r) = \exp\left(\frac{-\Delta E}{k_b T}\right), \tag{1.1}$$

where Δr is the conformation change, ΔE is the change in energy due to the MCMC step, k_b is Boltzmann's constant, and T is a temperature factor that modulates the permissiveness of the protocol. Due to the stochastic nature of the mentioned algorithms, multiple replicas are run in parallel to increase the chance of sampling and then converging to the global energetic minimum.

The application of these fundamental algorithms, coupled with others methods (such as fragment-based assembly and homology modeling) has led to many successes in employing Rosetta to design *de novo* proteins (161), membrane proteins (162), enzymes (163), and therapeutic peptides (164).

Rosetta has a wide array of specialized protocols, and choosing which protocol to use is key for using Rosetta to the best of its capability. In addition, Rosetta developers have been recently focusing on machine learning and deep learning methods for structure prediction with RoseTTAFold (165), structure generation (unconditional or topology-constrained, or

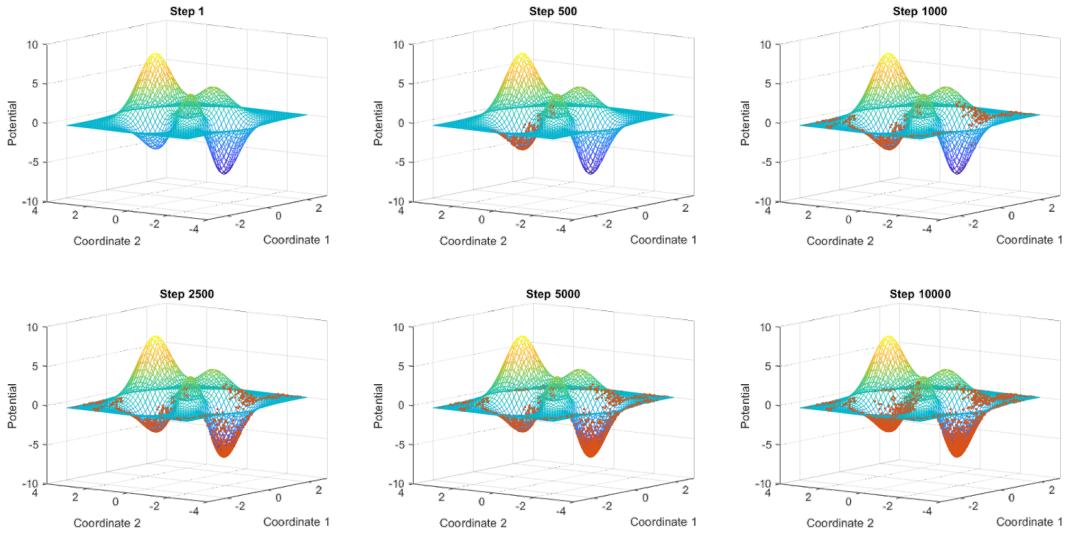


Figure 1.11: Different steps of a Markov Chain Monte Carlo simulation with a Metropolis criterion for one particle in a toy potential. The particle starts at the origin, as shown in the top left frame. The particle eventually populates the two minima found in the energy landscape. Simulation was run for 10000 steps, with a step size of 0.2 and $k_b T = 1$. Periodic boundary conditions were applied.

even symmetric oligomeric complexes) with RFdiffusion (166), and sequence design with ProteinMPNN (167).

In this work, we do not aim to radically change protein structure or sequence or to do *de novo* design. We rather introduce point or double mutations that modify protein function in some quantifiable way. The mainly used protocols are the relax (168) and score (157; 169) protocols for structure preparation and comparison, and RosettaRemodel for rebuilding missing loop regions in the studied structures (170). For performing accurate mutations in a membrane environment, RosettaMembrane is employed, which will be explained in the next section.

1.2.3 RosettaMembrane

Our laboratory created RosettaMembrane specifically for simulating the membrane protein environment via intraprotein and protein-solvent interactions (171). In contrast to Rosetta, RosettaMembrane has additional terms for calculating van der Waals and hydrogen bonding scores in the membrane. The membrane is modeled implicitly by three continuous phases: water, lipidic, and phospholipid head-group. The contribution of each phase to the solvation term depends on its location within the membrane. With these changes, RosettaMembrane has been able to more accurately reproduce native rotamer conformations within the membrane during benchmarking.

In this work, RosettaMembrane is used for in-silico mutagenesis of GPCRs. We use the output

scores from RosettaMembrane coupled with normal mode analysis (next section) to evaluate the quality of mutations and decide on which mutations to go to the next phase of the study.

1.2.4 Normal mode analysis and dynamic cross-correlations

As described in Sec. 1.4.2, NMA methods provide a fast and reliable way of probing protein fluctuations and cross-correlations at a coarse grained level. A normal mode is a pattern of motion where all parts of a given system move sinusoidally with the same frequency and a fixed phase. Normal modes can be derived from very simple mechanical systems (coupled oscillators with 2 masses for example) and can be readily generalized to larger systems in three dimensions with the use of Hessian matrices of the potential (called generalized force matrix in (172)). The general motion of a system can be seen as a superposition of these normal modes. Low frequency modes describe large concerted motions of the system that are energetically favorable (energy of a mode is proportional to square of its frequency)

For the aim of our study, we use NMA as a fast coarse grained approximation to GPCR dynamics. Anisotropic network model (ANM) has been shown to be able to replicate the transition between opsin and rhodopsin by comparing the motions described by the lowest 20 (accessible) modes to rhodopsin to the first principal component of a PCA constructed using the available rhodopsin and opsin structures (173). Furthermore, lowest ANM modes in rhodopsin were coupled with energy minimization with Amber94 forcefield to construct a model for the active Meta II state of rhodopsin that highlighted residues (global hinge sites, peaks in high frequency modes, and sites related to retinal isomerization) with observed experimental effects using decay rate and misfolding data (174).

Given the (relative) success of NMA methods to GPCRs, we use NMA to approximate dynamic cross-correlations which we then relate to allosteric signaling across the GPCR. To do that, we sum dynamic cross-correlations over pairs of allosteric hubs (residues that are important for transmitting the signal from the ligand binding to the intracellular binding interface which are extracted from molecular dynamics (MD) simulations). We hypothesize that this will remove superfluous correlations that are not functionally relevant (16). Dynamic cross-correlation matrices (DCCM) are calculated from the lowest 20 modes (similar to rhodopsin) and then compared between a reference state and a perturbed state, for example: a WT and a mutant receptor, or an inactive and active state receptor. One could push this idea further and define a difference of differences $\Delta\Delta\sum_i\sum_j DCCM(i,j)$, between (active and inactive) and (mutant and WT). Chen et al. have related this quantity to the ligand-effector structural coupling G^c under certain assumptions (16)

$$\Delta\Delta G^c \sim \Delta\Delta\sum_i\sum_j DCCM(i,j), \tag{1.2}$$

where i and j are pairs of allosteric hubs and $DCCM$ is the dynamic cross-correlation matrix

for the given state of the system.

Based on the derivation in (16), eq. 1.2 holds as long as the ligand and receptor contributions to the energy between WT and mutant are similar. This can be achieved by: 1- Avoiding mutations that destabilize the system 2- Avoiding mutations in the ligand binding site/G-protein binding site.

1.3 Models of allostery

1.3.1 Allosteric two state model (ATSM)

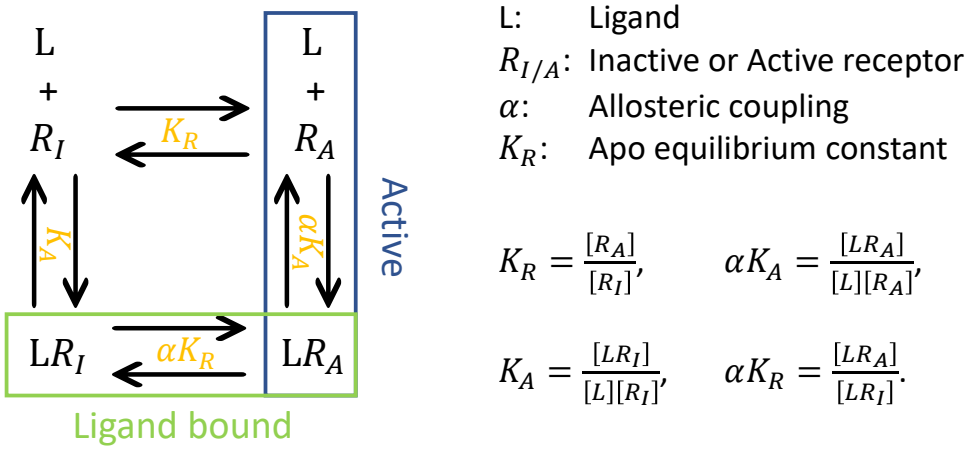


Figure 1.12: Allosteric two state model, where a ligand (L) binds to a two-state receptor (R). The system is considered as a two component system where each component can access two states, as shown in the schematic **(left)**. The terms in the schematic, as well as the equilibrium constants are presented on the **(right)**. An expansion of this explanation to include a thermodynamic aspect and more than 1 ligand and/or active state can be found here (175).

Of significant importance for this study is allostery in receptor activation, specifically GPCRs, as formulated in the allosteric two state model (ATSM) (176). In this model, the system studied is a ligand-receptor pair, where the receptor can occupy one of two states, a resting inactive state (R_I) or an active state (R_A). The receptor has a "basal" equilibrium between these two states in the absence of a ligand described by K_R, and then ligand binding shifts this equilibrium either toward the active state (the ligand being an agonist), toward the inactive state (inverse agonist), or not affecting the equilibrium (antagonist). The ligand has a different association constant (K_A) for the active or the inactive state, with the the allosteric coupling α describing this difference, as seen in Fig. 1.12.

From the cycle in Fig. 1.12:left, we can determine the ratio of active receptors, f_R, to be

$$f_R = \frac{[R_A] + [LR_A]}{[R_I] + [R_A] + [LR_I] + [LR_A]}, \tag{1.3}$$

where the concentrations are defined as in Fig. 1.12:right. Manipulation of the equation using the equilibrium equations in Fig. 1.12:right leads to the fraction of active receptors as a function of ligand concentration $[L]$ and as a function of the parameters of ATSM (K_A, K_R, and α)

$$f_R([L]) = \frac{K_R + \alpha K_R K_A [L]}{1 + K_R + K_A[L] + \alpha K_R K_A [L]}. \tag{1.4}$$

We can use Eq. 1.4 to extract three quantities that describe the dose-response curve in ATSM, the basal activity as $[L] \to 0$, the max response as $[L] \to \infty$, and the middle points of the transition, EC_{50} as seen in Fig. 1.13.

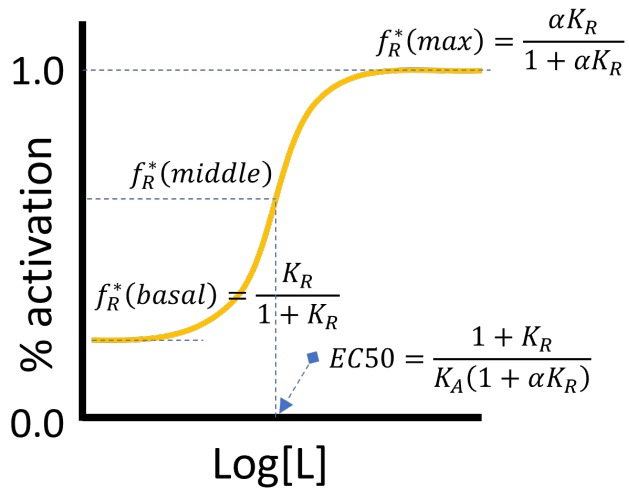


Figure 1.13: A typical dose-response sigmoid curve in ATSM. Under ATSM, an equation describing the fraction of active receptors as a function of ligand concentration can be derived (Eq. 1.4), and then we can extract different limits from the equation as the ligand concentration tends to zero, as the ligand concentration tends to infinity, and at the middle of the sigmoid curve (175).

finally, we can solve for the allosteric efficacy α in terms of K_R, K_A, and EC_{50}:

$$\alpha = \left(\frac{1 + K_R}{K_A EC_{50}} - 1\right) * \frac{1}{K_R}. \tag{1.5}$$

This represents the simplest version of the allosteric two-state model as it pertains to GPCRs. The model could be expanded to incorporate additional states, which would be necessary in the case of biased signaling, allosteric modulation, or presence of multiple ligands (175). Despite its simplicity, the allosteric two-state model remains a valuable instrument for predicting

and elucidating GPCR signaling, albeit primarily at a phenomenological level, rather than a structural or dynamical aspect.

1.3.2 Macroscopic mechanisms of allostery

As described in the introduction, many models have been proposed to describe allostery (Sec. 0.2). Out of these models, three main macroscopic mechanisms emerge: induced fit (KNF sequential model (23)), population shift (MWC model (2)), and entropy driven allostery (Cooper and Dryden (24)).

Weinkam et al. proposed an order parameter to connect microscopic structural from molecular simulations to macroscopic allosteric mechanisms. The order parameter is related to how ligand binding induces motion or cooperativity in the protein. Little to no change in structure in response to the ligand signifies an entropy-driven mechanism, while significant amount of conformational change triggered by ligand binding is a sign of an induced fit mechanism. Between these two extremes lies the population shift (177). Each of these mechanisms seem to better describe one of the proteins mentioned in the study. From this, we can define allosteric mechanism as a function of an order parameter (Fig. 1.14).

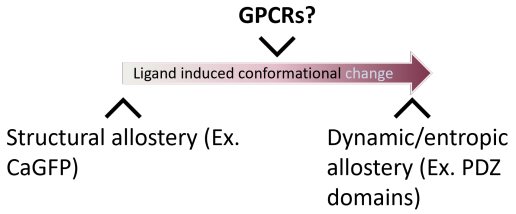


Figure 1.14: Description of allosteric mechanisms as a function of an order parameter (177). GPCRs fit into neither extreme, and thus we need descriptions of both entropic contributions and conformational change to describe their allosteric mechanisms.

1.4 Studying allostery

This section outlines the major methods of studying allostery that are relevant to this work. While the methods mentioned in this section are divided into experimental and computational, and then divided into sub-methods; studying allostery is rarely ever done using only one technique. It is usually studied by combining the strengths of different experimental and computational techniques to achieve an understanding of allostery.

1.4.1 Experimental methods

Any rigorous study of allostery needs to have a basis in experimental data whether for fitting/-training or validation. Allostery has been studied in different and imaginative ways, including but not limited to:

Mutagenesis: one of the more "brute force" ways of studying allostery is mutagenesis. The question is simple: what effect does mutating an amino acid have on protein function? (Of course the interpretation of the specific experimental read-out and relating it to any kind of allosteric mechanism is not trivial, as is discussed in this perspective (178)). Alanine scanning of an area of interest in KCNQ1, a potassium channel, has been used to study its gating process (179). Other studies have quantified the prevalence of allosteric sites in human liver pyruvate kinase by whole-protein alanine scanning (180) and mapped the allosteric mechanism of a GTPase via deep mutational scanning (181; 182). The conclusion from these studies is that allosteric sites are prevalent in the structure, and that an allosteric mechanism that involves a few defined positions is too simplistic. Deep mutational scanning has been applied to GPCRs on the model beta-2 adrenergic receptor by testing 7800 of 7828 possible single amino acid substitutions at four concentrations of the agonist isoproterenol. Mutations were divided into six clusters with varying sensitivity and tolerance (183).

A different approach is targeting mutations to specific allosteric sites with mutation effects predicted from a computational pipeline, also known as rational design. The process would include predicting allosteric sites computationally (check Sec. 1.4.2) and then quantifying the effect of specific substitutions. This has been demonstrated on dopamine D2 receptor by repurposing it to a serotonin biosensor as well as modulating ligand efficacy and basal activity (16). (One could argue that this rational design is an integrated approach, but we added it here in this section for completion)

Structural studies: Structural data provides an invaluable resource to probe into allosteric mechanisms, especially in cases where allostery triggers a conformational change in the studied protein. These structures can be commonly determined by X-ray crystallography, cryogenic electron microscopy (cryo-EM), or NMR spectroscopy.

X-ray crystallography has been used to probe conformational changes (by determining structures of different states/complexes), such as opening or closing of an active site, as seen in phosphoglycerate dehydrogenase (PGDH) (184), minor conformational changes to active site, with an example being DAHP synthase (DAHPS) (185), control of complex formation/oligomerization allosterically, as in ATP phosphoribosyltransferase and epidermal growth factor receptor (EGFR) (186; 187), as well as other cases described in the literature (188). Of special note are the crystal structures of GPCRs, starting from the first structure of rhodopsin (189) to the tertiary GPCR-G protein complex (57) passing through agonist-bound intermediate states (69). (Note that crystallization of the tertiary complex is significantly difficult due to flexibility of the complex; there has been one other X-ray structure of a GPCR-G protein complex since 2011 (190))

Combining structural insight from crystal structures of GPCRs with the recent explosion of GPCR cryo-EM structures gives us a unique opportunity to study allostery in GPCRs (191). Cryo-EM has seen a recent surge with the famous "resolution revolution" (192), making it a popular tool for examining the atomic details of protein structures (193), especially in GPCRs

(Only 15 out of 261 GPCR structures deposited on GPCRdb in the last year (76)(March 2022 - March 2023, https://gpcrdb.org/structure/statistics) were determined using X-ray crystallography, all the others were cryo-EM!). As GPCRs are fundamentally allosteric proteins that populate many different states, it is critical to discern the structures of inactive, intermediate, and active states to understand their function at an atomistic level. Cryo-EM has opened the door to solving active state complexes of GPCRs with G-proteins (194), arrestins (109), and a G protein-coupled receptor kinase (195). Structures of GPCRs bound to allosteric modulators pave the way to studying allosteric modulation at sites distant from the orthosteric binding site, adding another dimension to the study of allostery in these systems (196)(check Sec. 1.4.2 for more details). Furthermore, time-resolved cryo-EM can be used to investigate intermediates between two states, as has been shown on GTP bound Gαs with β2-AR (197). Such intermediate state structures will prove invaluable in understanding protein transitions when combined with dynamics simulations.

Dynamical studies: There are other cases, however, where structural studies are not enough to identify the allosteric mechanism because the active and inactive states are almost identical structurally. One famous example are PDZ domains, very common protein domain binding modules that exhibit "dynamic" allostery. Fuentes et al. employed spin relaxation backbone (^{15}N) and side-chain (^{2}H-methyl) to compare the differences in dynamic parameters upon binding of a peptide target, detecting changes in the domain's ps–ns dynamics (198). More recent studies have employed advances in NMR-based methods to provide an atomic-resolution multi-state allosteric mechanism for PDZ domains, starting with a conformationally selective ligand binding followed by conformational change covering about 25% of the protein, which in turn reduces the allosteric netowrk of the *apo* form. (199). Similar advances allowed for pinning of two different allosteric modes of action of two ligands of the Pin1 enzyme, with one ligand following a dynamic allosteric mechanism and the other a population shift mechanism (200).

NMR studies have also been invaluable in underpinning allostery in GPCRs (201). $^{1}H^{15}N$ NMR has been used to quantify the response of the thermostabilized mutant of the turkey β1-adrenergic receptor (β1AR) to six different ligands, ranging from full agonists to antagonists, as well as correlating NMR chemical shifts to functional data (G-protein efficacy) (202). This study was later complemented by quantifying dynamic equilibria between three major receptor states upon binding of a G protein mimicking nanobody (63). In addition. ^{19}F NMR spectroscopy has showed the importance of cholesterol and cations in signaling of the adenosine A2A receptor (A2AR), where cholesterol acted as an allosteric modulator that increased the active state population and precoupled G-protein-bound state (64). Cations played either an inactive ensemble reinforcing (Na^{+}) or active state stabilizing (Ca^{2+} and Mg^{2+}) role (203).

GPCRs have also been studied with double electron-electron resonance (DEER). In one study, pressure resolved DEER revealed a minor population of unliganded β2AR with signature TM6 movement (that signifies class A GPCR activation), which provides evidence for a conformational selection mechanism in the activation of β2AR (204). Another study elucidated

structural and dynamical responses of μOR to functionally distinct ligands. The authors were able to resolve two active conformations of TM6, which would translate to different G-protein activties (205).

1.4.2 Computational methods

Given the recent development in computational resources and simulation techniques, computational methods have become indispensable in the study of biological systems, and allostery is no exception. Technical details of the relevant methods are left to Sec. 1.2, while this section reviews how the methods were used to study allostery in the literature.

Normal mode analysis (NMA): NMA is based on approximating the potential energy surface of a protein harmonically and then representing protein motions as linear combinations of normal mode vectors (206; 207). It neatly provides a method that resembles the ensemble view of allostery, with the main difference being that conformational sampling is close to an energy minimum. NMA has been used to study dynamics at an all-atom level (208), but coarse graining, as in the elastic network model (ENM), removes the need for structure minimization (so it could be applied directly to crystal coordinates) and is capable of replicating slow modes of motion in sufficient detail using a single parameter model that connects particles within a cutoff with springs described by Hooke's law (172; 209). Low frequency modes are the ones most relevant to allostery since they're cooperative and can capture overall movement of the structure (210). A further development was the Gaussian network model (GNM) by Erman et al. (211; 212), which provides information on individual modes, whether in the form of cross-correlation or residue mean square fluctuation, by assuming that residue fluctuations obey a Gaussian distribution. It falls shorts of providing information on three dimensional motion, however. An extension of the GNM, the anisotropic network model (ANM), was developed to include, as its name suggests, anisotropic fluctuations (213; 214). NMA has been successful in detecting allosteric sites (215; 216), as well as mapping allosteric pathways (210; 217). Flavors of NMA, such as normal mode perturbation analysis, which quantifies the change in dynamics of a protein in the presence of a modulator, can be used to predict allosteric pockets (218). Given that NMA assumes that equilibrium systems fluctuate around a single (well-defined) potential well and that the well takes a harmonic form, this leads to a number of limitations: NMA is only valid in close proximity to equilibrium and uncertainty grows as you move away from it. A direct consequence of this is that normal modes that describe motions of the protein will violate internal constraints (bond lengths, angles, etc ...) in all but the smallest motions unless explicit measures are taken to circumvent them (219).

Molecular dynamics (MD): Since MD simulations provide detailed information about the system dynamics at an atomic level without the simplifications/assumptions of NMA, it is considered the computational gold standard in studying biomolecule dynamics which can enable the elucidation of allosteric mechanisms. MD has two features that make it attractive for the study of allostery: the first is its spatial (sub-angstrom) and temporal (femtosecond)

resolution. The second is the versatility of simulation setups, where controlled perturbations could be introduced to the system to study the exact effect of a mutation, absence or presence of ligand, or application of an external force. Allostery can take many forms and "disguises" (220), making design of the "computational experiment" with MD is of utmost importance as outlined in the case studies presented in Hertig et al (221).

The most straightforward way of studying allostery is to simulate a protein structure and look for correlated motions between sites of interest, or use correlated residues to detect potential allosteric sites (MD simulations produce huge amounts of data which are difficult to discern with direct visualization). Analysis methods based on metrics such as covariance/cross-correlation (222; 223), mutual information (224; 225), and transfer entropy (226) among others (227; 228) have been developed to separate the wheat from the chaff in simulations.

A popular way to understand allosteric networks are graphs (25; 229), and metrics from MD simulations have been combined with graph representations of proteins to elucidate allosteric communication. Co-variation analysis was proposed by Proctor et al. (230), who treated the correlation map as a weighted graph connectivity matrix, and then used Dijkstra's algorithm (231) to find the optimal path of propagation of correlations throughout the network. From this idea of pathways within protein graphs, one could define bottleneck residues that can disconnect a sub-graph, blocking communications between binding and allosteric sites, and thus these residues could serve as targets for design and allosteric modulation. Sethi et al. introduced the idea of "community network analysis", where the protein graph is partitioned into sub-graphs called communities. These communities are loosely coupled to one another, but residues within one community are strongly coupled (232). Bhattacharya and Vaidehi (233) proposed an alternative approach to co-variation analysis that uses time-averaged pairwise mutual information (MI) computed from torsional angles (rather than Cartesian coordinates) and applied this method to GPCRs. This form of studying allostery in proteins will be expanded upon in Sec. 1.2.

MD is also popular to study allosteric regulation and allosteric modulation in GPCRs (234). A study of allosteric modulation of M2 receptor with unbiased MD provided valuable information on the binding modes and binding pathways of the allosteric ligands as well as their cooperativity with the orthosteric ligands. All studied modulators formed cation-pi interactions with extracellular ligands, which needed fine-tuning of cation-pi interactions in the forcefield (235). Allosteric modulation could occur in more subtle ways than binding of allosteric ligands, such as binding of ions or presence of cholesterol in the lipid membrane. Presence of Na^+ ion was shown to affect binding of orthosteric ligands as well as the action of negative allosteric modulators (NAMs) in dopamine D2 receptor(236). Another study suggested a conserved mechanism of Na^+ binding to class A and B GPCRs by calculating free energy profiles from combined MD simulations and Markov state models (237). An unbiased MD-based study found a correlation between the presence of cholesterol in the membrane and the flexibility of the serotonin 5-hydroxytryptamine 2A receptor (5HT2A). This increase was due to decrease of hydrogen bonding between the receptor and the first layer of the lipid

membrane in the presence of cholesterol (238). Another study used MD to discern cholesterol entry and exit pathways from the membrane into the A2AR binding pocket. The authors observed that cholesterol modulates ligand-binding properties in A2AR both orthosterically and allosterically, with a possibility of cholesterol "invading" the orthosteric binding pocket (239).

In other studies, enhanced sampling MD simulations were utilized to better mimic experimental conditions and/or to overcome sampling difficulties. As an example, steered molecular dynamics (SMD) could be used to computationally mimic single molecule pulling experiments (such as those done with atomic force microscopy, although it is vital to note the limits of similarity between both (240; 241)). SMD was combined with equilibrium MD and principal component and correlation analyses by Amaro et al. to find allosteric pathways in glutamine amidotransferase (242). Schoeler et al. employed SMD combined with AFM to find force propagation pathways passing through a mechanically stable multidomain cellulosome protein complex under force (243).

Other enhanced sampling techniques that have proven popular in the study of allostery and activation of GPCRs include (Gaussian) accelerated MD (244; 245; 246), replica exchange (especially variants with solute tempering, such as REST2) (247; 248; 249), and meta-dynamics (250; 251; 252).

Accelerated MD (244) adds a bias potential below a certain energy threshold but leaves the area above the threshold unaltered. The bias could be added either to the dihedral potential or the total potential (or both, called dual-boost aMD (253)) While aMD does not need prior knowledge about any collective variables (CVs), it does need an initial unbiased simulation to collect potential statistics for determining the energy threshold. However, finding appropriate CVs that well describe the system is still required for any free energy calculations, as shown in the free energy study of the M2 muscarinic receptor (254), and a study of the effects of select mutations on the activation landscape of CCR5 (245).

Replica exchange methods (255) attempt to overcome the sampling problem by simulating different replicas of the system over a range of temperature and then exchanging coordinates between replicas periodically using a metropolis criterion (256). The standard temperatures replica exchange method suffers from poor scaling with system size, where the number of required replicas grows as $f^{1/2}$, where f is the number of degrees of freedom in the studied system. The idea of solute tempering (REST) has been implemented to bypass this poor scaling by heating up the solute while leaving the solvent cold in higher temperature replicas, thus reducing the number of required replicas. Further modifications to the scaling of the hamiltonian have been suggested to optimize the solute tempering method and the REST2 version (247) has been successfully applied to study sodium ion allosteric modulation of CXCR4 with and without constitutively active mutants (248). REST2 was combined with NMR and in-cell assays to study biased signaling in μ-opioid receptor, where the authors found distinct binding conformations for biased, unbiased, or partial agonists. The biased ligands

induced a change in conformation around intracellular loop 1 and helix 8, which they used to propose an activation mechanism for unbiased vs biased ligands (249).

Meta-dynamics (257) attempts to escape free energy minima by depositing a history dependent bias potential on the free energy landscape defined by a few CVs. This bias will (given enough time) fill the free energy wells, allowing for an accurate determination of the free energy surface (FES) as a function of the selected CVs. Meta-dynamics proved to be a popular method for FES exploration, but it suffered from a few drawbacks. The first is the question of when to end a run, since deposition of bias peaks leads to oscillating around the correct value. The second is that the simulation could go to regions of the configurational space that are unphysical if the simulation goes on for long enough. Finally, the method is dependent on the choice of CVs, and choosing an appropriate CV for a given system is not trivial (we will later see how machine learning approaches can help with that!). To address the first two issues, an adaptive bias was later introduced that decreases in amplitude as the simulation goes on, converging to the correct FES (well-tempered metadynamics (250)). MetaD methods have been used to study β-arrestin and G-protein complex formation of β2-adrenergic receptor in apo form or in the presence of four ligands (251). The authors found that the arrestin-receptor bound conformation depends heavily on the bound small molecule, and were able to quantify changes in binding free energies of small molecule ligands in presence of either intracellular binding partner.

Machine and deep learning

Over the last decade, machine learning approaches have been developed to assist in studying protein dynamics and allosteric mechanisms (258), and have become a valuable asset that augments both computational and experimental studies. This subchapter will focus on the computational part that is relevant to this work.

Machine learning has helped conformational sampling in MD simulations through choice of reaction coordinates (259; 246), learning of slow modes of motion (260), and equilibrium ensemble generation without the need of explicitly performing simulations (261). For example, Gaussian accelerated MD simulations were combined with deep learning approaches to study the effect of allosteric modulators on GPCR free energy landscape. Attention maps extracted from 2D convolutional neural networks trained on labeled residue contact maps were mixed with flexibility analysis to find reaction coordinates for which free energies would be calculated. They found that allosteric modulators (whether positive or negative) confine GPCRs to mostly one specific conformation for signaling (across class A and B) and that modulators are specific to receptor subtypes (246).

The scope of machine learning methods goes beyond sampling to extraction of allosteric residues and networks. Zhu et al. have applied a neural relational inference (NRI) model based on graph neural networks that reconstructs MD trajectories, thus learning long-range interactions that mediate allosteric communication between distant sites. It has been applied

to elucidate allosteric pathways in Pin1, SOD1, and MEK1 (262).

Conclusion

All of these studies of protein allostery culminate in the ultimate question, can we build a model that can design (or at least aid in designing) allosteric proteins[1]? As design of allosteric proteins is still in its infancy due to a lack of any unifying theoretical framework (despite having "unified" high level views (175); maybe one day we can hope for a "grand unification" of allosteric mechanisms?). Attempts at design include classification of mutational effects of a LOV2 domain by studying over a 100 mutations with a Random Forests algorithm (263), but the bulk of allosteric design (so far) happens classically. Allosteric databases, such as ASD (264) or AlloMAPS 2 (265), will prove very useful for future developments of allosteric residue/site prediction and design. Finally, the advent of generative methods holds great promise for the future of protein design in general (266).

1.4.3 Theoretical considerations

Bidirectionality and microscopic reversibility

A useful principle when studying allostery using computational methods is the concept of bidirectionality of allostery. Assume an allosteric protein binds two ligands A and B. If the binding of ligand A increases the binding affinity of ligand B (a phenomenon known as positive cooperativity), then the binding of ligand B increases the affinity to ligand A by an equal amount (if measured as changes in the free energy of binding, Fig. 1.15a). This could be useful since studying the binding of the effect of one ligand may end up being easier than studying the binding of another, and thus we end up with much shorter simulation times. One example could be seen in this study of allosteric modulators in GPCRs (235). Simulations were run with the presence and absence of the orthosteric ligand, and then allosteric modulators would bind or dissociate during the simulation. Running simulations in this fashion is much more practical since binding of allosteric modulators is typically much faster than that of orthosteric ligands, since they bind to exposed parts of the receptor. A discussion of this point with several case studies is presented here (221).

Another useful principle, microscopic reversibility, states that, at equilibrium, a system (such as a protein) that goes through a transition between two states, will follow the same path in both the forward and reverse directions. This is especially useful when we have two experimentally determined structures of a protein (active and inactive state, for example), and we want to sample the transition between the two states, as sampling one direction may be much more feasible than the other direction (Fig. 1.15b). To know whether microscopic reversibility

[1]What do we mean by protein design? We can envision a targeted design starting from natural scaffolds and modifying allosteric signaling in them, or a more radical *de novo* design of allosteric proteins from the ground up. Protein design is usually treated as a binding problem, while we are treating it in this work as a function problem.

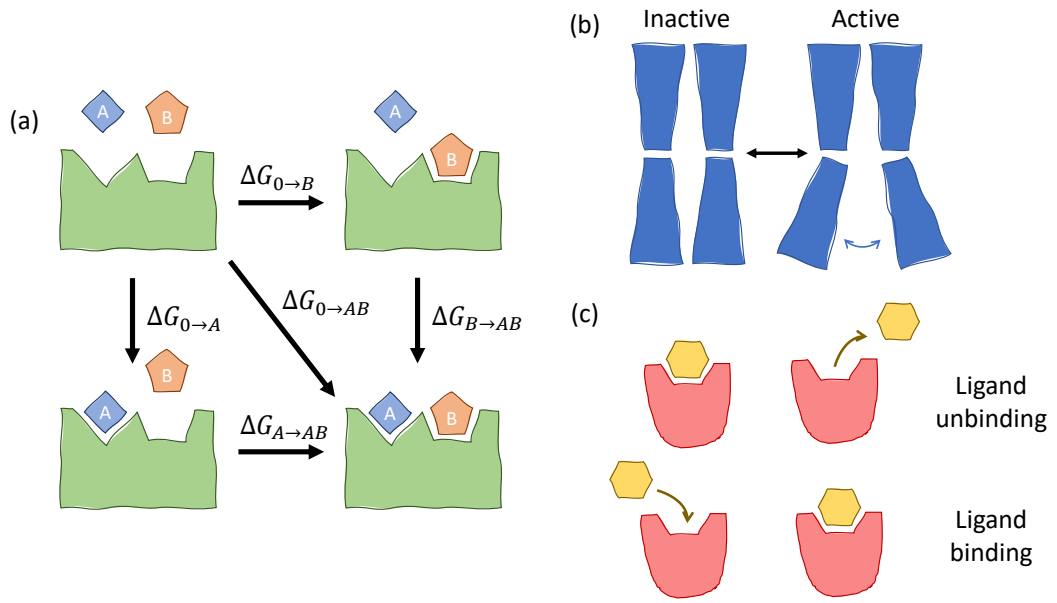
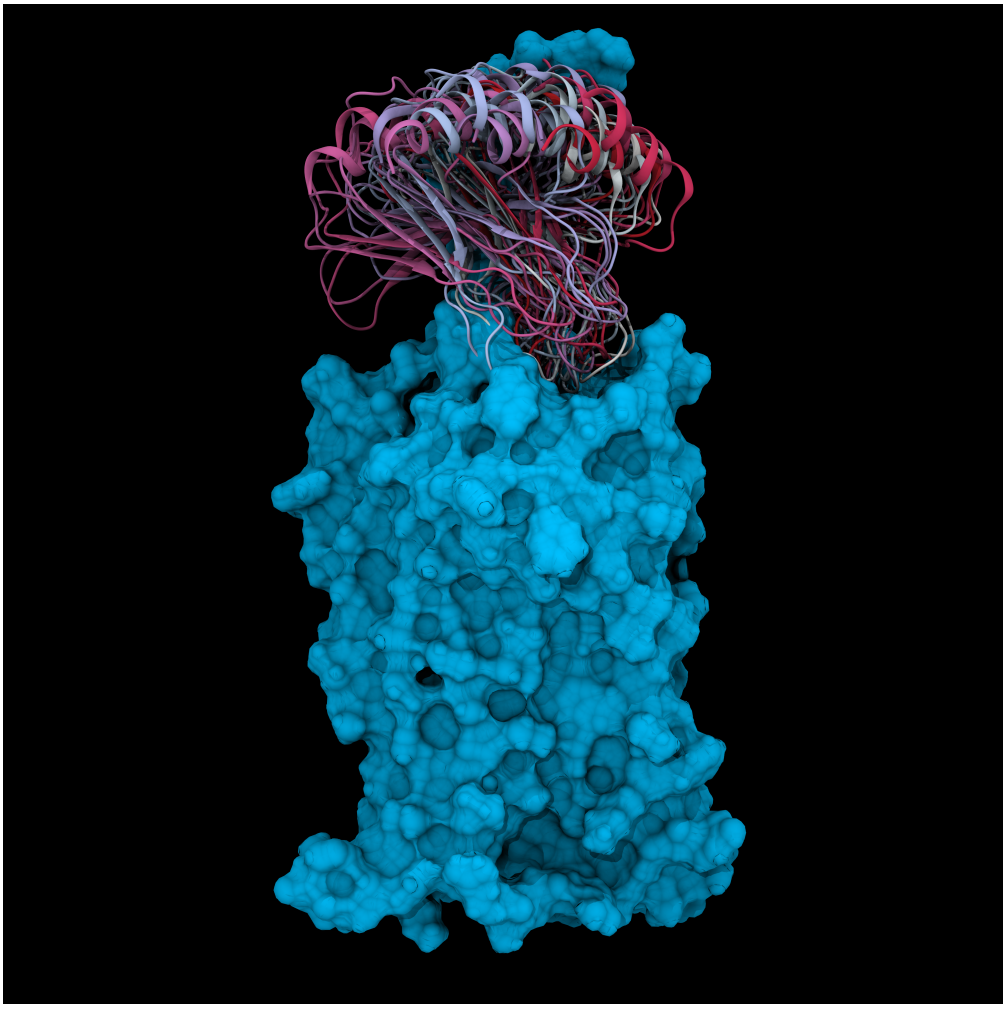


Figure 1.15: **(a) Bidirectionality of allostery:** Assume an allosteric protein binds two ligands A and B, then the free energy difference $\Delta G_{0\to AB}$ is independent of the pathway taken to arrive at state AB. Thus $\Delta G_{0\to AB} = \Delta G_{0\to A} + \Delta G_{A\to AB} = \Delta G_{0\to B} + \Delta G_{B\to AB}$. This implies that if the binding of A is more favorable in the presence of B, then the binding of B is more favorable in the presence of A (and vice versa). **(b) and (c) Microscopic reversibility:** under the assumption of microscopic reversibility, simulating the transition in one direction will give us information about the transition in the other direction. Practical examples are simulating an active to inactive transition rather than the other way around **(b)** and simulating ligand unbinding rather than binding **(c)**

applies to process one is studying, here are a few guidelines (as outlined in the case studies in this paper (267)). Time scales within a state (a state could have substates, with transitions between different substates) should be much shorter than transition times between separate states, which leads to a quasi-Markovian behavior. In addition, for an energy basin that is deep enough, a trajectory's exit point is uncorrelated with its point of entry, which entails approximate symmetry in the system. In essence, one should be careful when choosing a simulation's initial conditions, in addition to initiating the simulation from well-defined states. Quoting Bhatt and Zuckermann: "It appears to be an open question, however, whether biomolecular systems of interest tend to exhibit suitably well-defined states (267)".



Artwork 2: *CXCR2 (cyan) bound to CXCL8 (light grey to hot pink). The conformational ensemble sampled by CXCL8 during a set of MD simulations is shown.*

Method Development Part I

2 Development: AlloDy

GROMACS reminds you: "It seems likely that significant software contributions to existing scientific software projects are not likely to be rewarded through the traditional reputation economy of science. Together these factors provide a reason to expect the over-production of independent scientific software packages, and the underproduction of collaborative projects in which later academics build on the work of earlier ones."
— Howison & Herbsleb

2.1 Main idea and objectives

AlloDy is a package built to analyze molecular dynamics (MD) trajectories of proteins, with a focus on allosteric pathways and ensemble differences in GPCRs.

Objectives of the development process are the following:

- Automate the analysis of MD simulations in an accessible and customizable fashion
- Assemble an ensemble of useful metrics for allosteric analysis in a single package
- Build a system that can integrate multiple simulations and reference states organically into the code
- Integrate GPCR specific features (structural elements, generic numbering, activation states, etc ...) into the analysis

2.1.1 Overview of the approach

Since allostery is an elusive, and at times, ill-defined problem, studying it necessitates using an integrated approach that combines structural features, correlation metrics, and formal comparisons with reference states in order to pick up allosteric signals from the simulations.

As such, the workflow of AlloDy can be divided into three main parts: (1) analysis of a set of simulations of an individual system in the *md2path* module, (2) ensemble comparison between sets of simulations of a test and a reference system using the *kldiv* module, and (3) meta-analysis of simulations of many systems using the *meta* modules.

The *md2path* module, which aims to extract general useful metrics and calculate allosteric pathways from the simulations, could be summarized in the following steps:

- Fetch reference PDBs from database, and then align sequences and input PDBs to reference PDBs
- Load and align trajectories
- Calculate Root-Mean-Square Deviation (RMSD) and Root-Mean-Square Fluctuation (RMSF) of receptor, ligand, and effector (if present)
- Calculate contact map of receptor with ligand and/or effector protein (if they exist)
- Calculate GPCR order parameters for activation states (if protein is a GPCR)
- PCA of ligand binding poses and receptor conformations followed by clustering
- Calculate dihedral time series from trajectories (ϕ, ψ and χ_x)
- Calculate 1st and 2nd order entropies from dihedrals, followed by mutual information (MI)
- Assess convergence of entropies
- Run allosteric pathway calculation:
  - Represent protein as a graph with edge weights being some function of MI
  - Find shortest paths between nodes with significant pair MI
  - Cluster individual paths into super-structures of pathways

As for the *kldiv* module, which compares distributions of dihedrals between test and reference system using Kullback-Leibler (KL) divergences, the steps are:

- Fetch reference PDBs, and then align sequences and input PDBs to reference PDBs
- Load and align trajectories of test and reference ensembles
- Calculate dihedrals from trajectories
- Reconcile dihedrals between test and reference systems (in case of mutations or difference in number of residues)

- Calculate 1st order KL divergences
- Calculate 2nd order KL divergences and mutual divergence (optional step that is time consuming and not very informative)
- Plot dihedral distributions with highest divergences
- Visualize KL divergences residue-wise

The *meta* modules will not be discussed here, and explanation will be left to the documentation on AlloDy's Github page.

2.1.2 Choice of states to simulate

The choice of states of a given system is critical in ensuring extraction of useful information from the simulations. Given that AlloDy compares distributions of dihedrals or distances, it gives best results when simulations sample a well defined functional state of the receptor (ex.: antagonist bound inactive state, agonist bound intermediate state, or agonist-receptor-G-protein tertiary complex).

To make the study of a given system meaningful, at least two states need to be simulated, with one acting as a reference state (such as an inactive state), and the other is the test or target state (such as a ligand bound active state). More target states, such as multiple ligands can be incorporated into the analysis, what matters is the consistency in the choice of a reference state.

2.2 Implementation

AlloDy is available for download at https://github.com/barth-lab/AlloDy under an MPL-2.0 license including instructions and a demo.

AlloDy is written in Matlab (268) and requires Matlab Bioinformatics toolbox (269), MDprot (free to download at https://github.com/barth-lab/mdprot), and VMD (270). It has been tested using Matlab 2021b on Windows and Linux systems.

2.3 Architecture of the code

The general architecture of the objects defined in AlloDy are shown in Fig. 2.1. AlloDy follows trajectory input and reading rules as defined in MDToolbox (271), and it uses a modified version of MDToolbox as a base (MDprot) and builds upon it. The classes defined in this section were initially written by Simon Lietar, a masters project student at the Barth lab from February till July 2022.

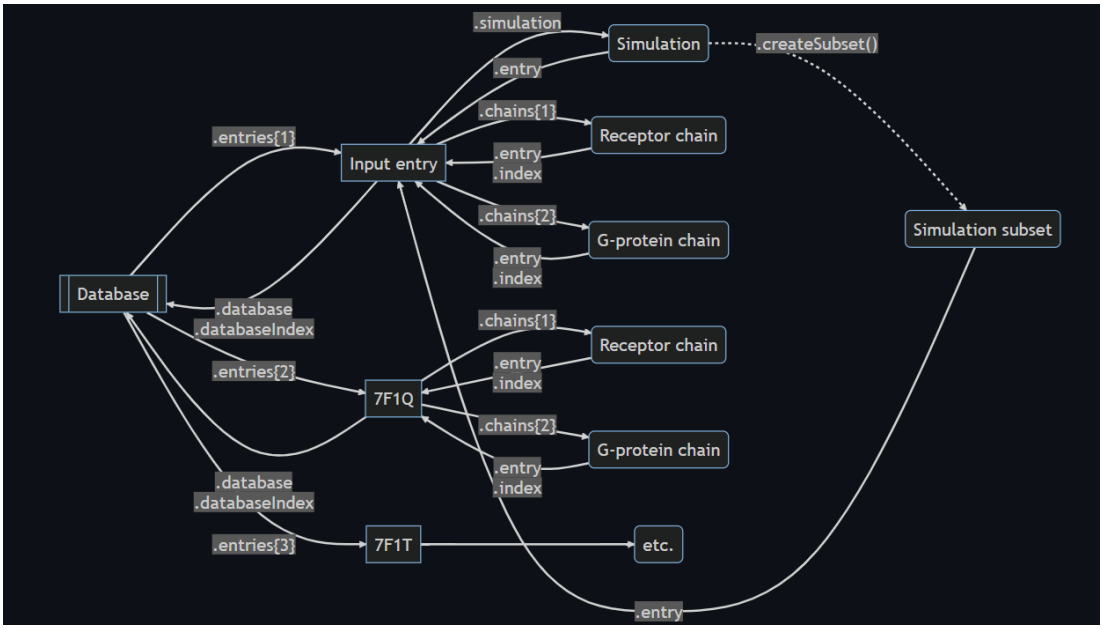


Figure 2.1: Flowchart of AlloDy classes

The code is based on four classes, *database, entry, simulation,* and *chain.*

A *database* object is the main class that contains different entries, where every *entry* contains a PDB file that is either fetched from the RCSB protein database or local directories. The *database* also contains functions that align sequences and structures between different chains of member entries, as well as label residues using generic GPCR numbering. This generic numbering will be useful for fetching conserved residues, calculating activation states of GPCRs, and labeling residues for ease of use of output files.

An *entry* is the object containing the PDB file with reference to *chain* objects, sequence of every chain, protein coordinates, and most importantly, it reads simulation data and creates a corresponding *simulation* object. An *entry* is also very handy to make any kind of atom selections (such as selecting backbone atoms, Cαs, ligand atoms, and so on)

A *simulation* object contains the trajectory from a molecular dynamics simulation and different functions to manipulate and analyze the trajectory. Examples of trajectory manipulation are superimposing frames to a reference frame and concatenating frames from different trajectories using an atom selection and an equilibration cutoff. Trajectory analysis ranges from calculating RMSD and RMSF of atom selections to calculating dihedral time series. From the dihedrals, 1st and 2nd order entropies could be calculated, from which MI is inferred. If more than one simulation set are present, KL divergences between distributions of equivalent dihedrals (according to the sequence alignment) can also be calculated.

Finally, a *chain* object helps the code deal with PDBs containing multiple chains, where every *chain* will contain its own atom and residue indices, secondary structure, and sequence.

For a more detailed description of the code, the objects and the functions that they contain, see AlloDy GitHub page reference.

2.4 *Md2path* module: calculating allosteric pathways from MD simulations

The *md2path* module analyzes a set of simulations for an individual system with the final aim of calculating allosteric pathways, going through the steps mentioned in Sec. 2.1.1.

PDB fetching, sequence alignment, trajectory alignment, and RMSD/RMSF calculation steps are trivial steps. We provide explanation for the rest of the steps in the following sections.

2.4.1 Contact map calculation

Contacts are calculated between non-hydrogen atoms using a dual cutoff scheme. Two atoms get into contact when they are within r_1 of each other and stay in contact until they are further than r_2 apart. This scheme is used to remove high frequency noise at the distance cutoff. Default values for r_1 and r_2 are 3.5 Å and 5 Å. respectively.

Contacts are calculated between the receptor chain and ligand chain (if present) and between the receptor chain and effector chain (if present). In case the effector chain is not present in the simulation, contacts are calculated from the reference active state PDB specified in the input settings file.

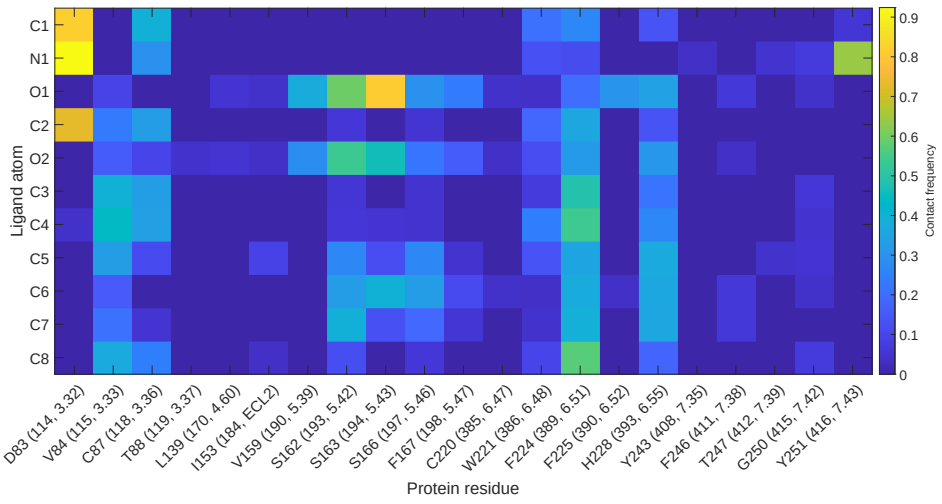


Figure 2.2: Contact map of ligand (dopamine) with receptor (DD2R) with default parameters. Y-axis shows atoms of the ligand, while X-axis shows residues of the receptor, where the first number is the residue number in the simulation, the second is the PDB sequence number, and the third is the Ballesteros-Weinstein designation (1).

2.4.2 Principal component analysis (PCA) of ligand binding poses

Principal component analysis (PCA) is used for clustering ligand conformations accessed during the simulation and finding representative samples from those clusters. Depending on whether the ligand is a peptide or small molecule, either Cα or heavy atom coordinate time series are used as input for the PCA by default. Other options for input coordinates are: vectors from closest receptor residue to the ligand atom concatenated with ligand coordinates or distance between closest receptor residue to the ligand atom concatenated with ligand coordinates. Comparisons between the three sets of input are shown in Fig. 2.4. Adding the aforementioned vectors from the receptor residues to the ligand atoms captures changes in the receptor ligand binding region in addition to variations in the ligand binding pose, alleviating one of the main drawbacks of taking ligand coordinates only.

PCA space is then clustered using a k-means clustering algorithm, with the optimal number of clusters being evaluated by the Calinski-Harabasz criterion (272).

To find representative frames from every cluster, the PCA space is histogrammed in 2 dimensions and frames closest to the center of the highest density bin are chosen as the highest density frame in the cluster. This is done for all clusters, as seen in Fig. 2.3.

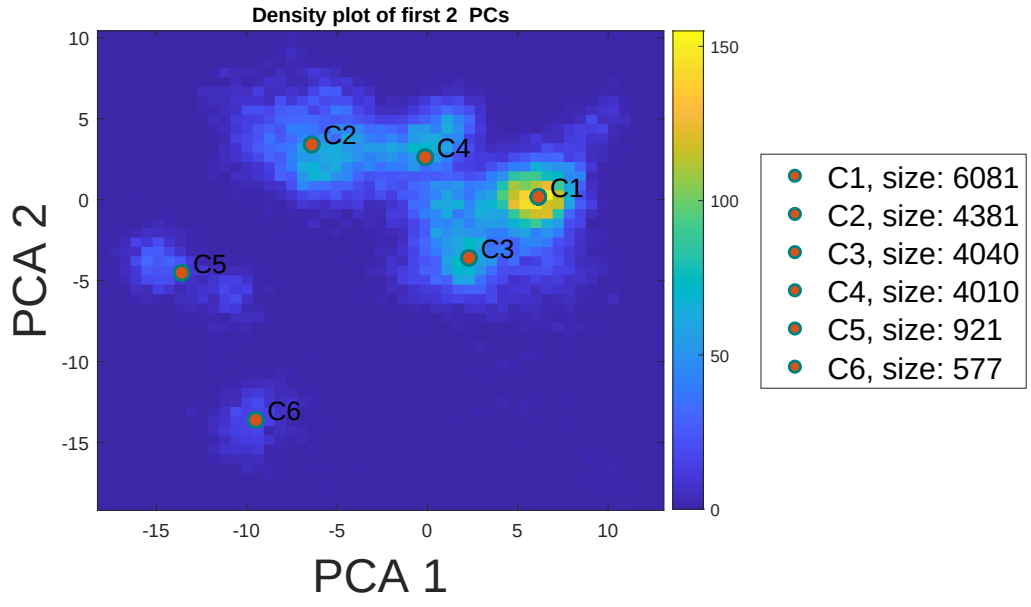


Figure 2.3: Density map of PC space of the first 2 PCs for dopamine bound to DD2R. Cluster centers are highlighted as C1, C2, etc. The size of every cluster (number of frames) is mentioned in the legend. PCA was performed on heavy atom coordinates of dopamine.

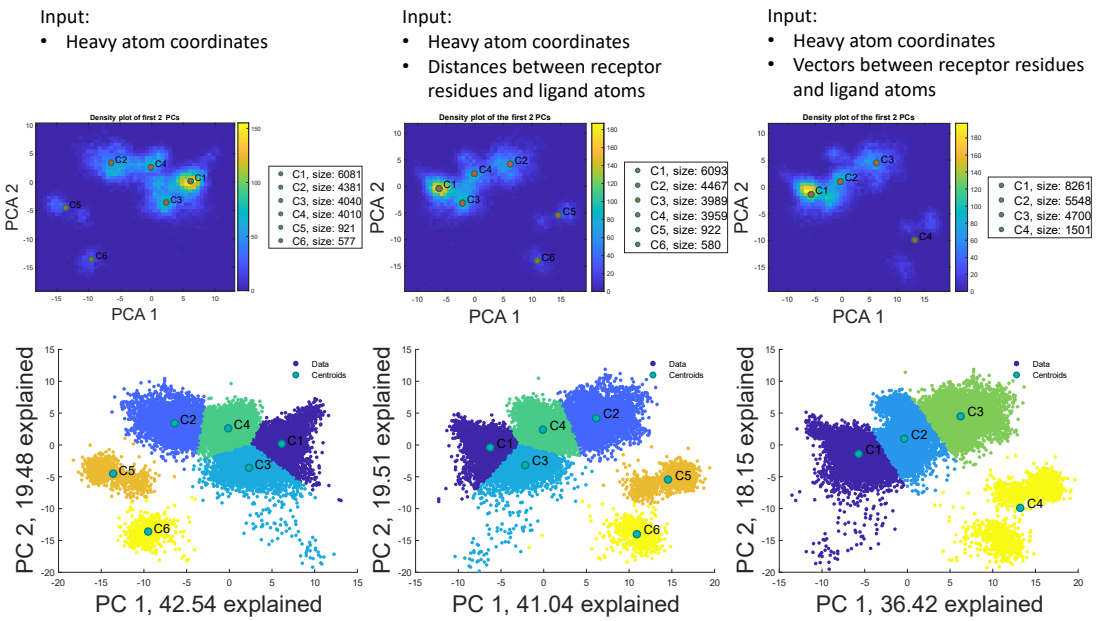


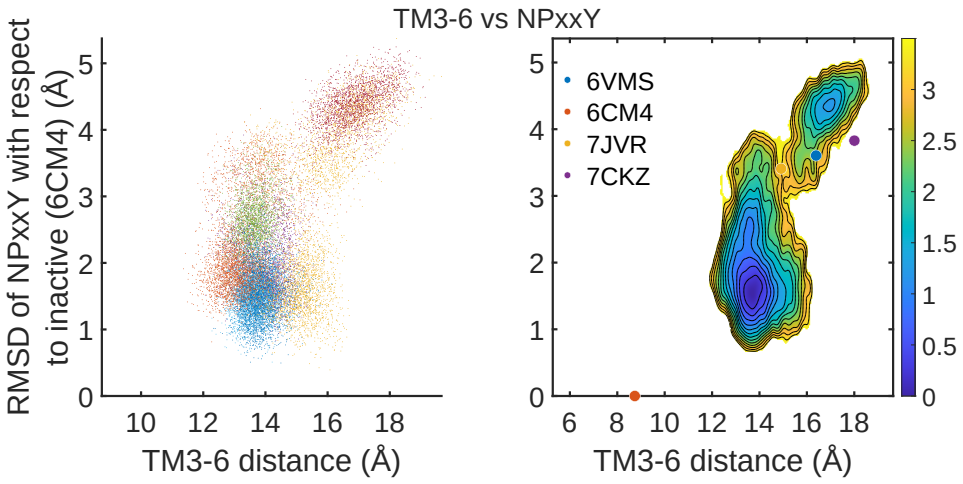
Figure 2.4: Ligand (dopamine) density maps (top) and scatter plots (bottom) of PC space of the first 2 PCs for dopamine bound to DD2R for three sets of input data to the PCA. Cluster centers are highlighted as C1, C2, etc. The size of every cluster (number of frames) is mentioned in the legend.

2.4.3 Principal component analysis (PCA) of and receptor conformations

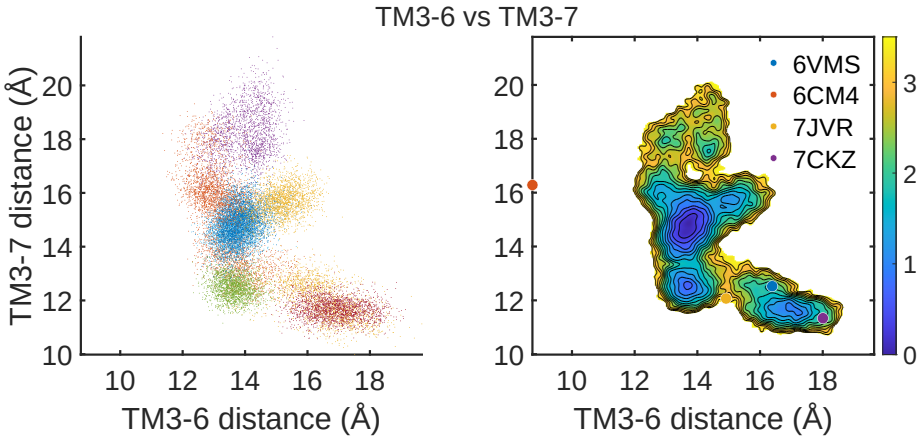
Similarly to the previous section, PCA is performed on receptor conformations. The process is the same with the exception of the choice of input to the PCA. By default, Cα coordinate time series are used, but other options include dihedral time series (backbone and sidechain), backbone dihedral time series, and distances between Cα pairs. Since distances are two body terms, the calculation may become memory intensive. To counteract this, two approximations are made: (1) coarse graining of the receptor, where every N residue Cαs are taken (defaults to 4), and (2) distances with highest variance (as a function of time) are considered (defaults to top 1000 distances).

2.4.4 GPCR activation states (GPCR specific option)

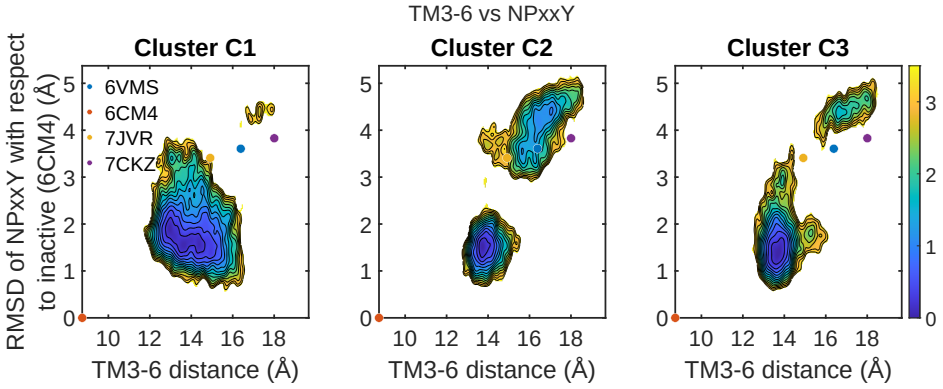
To specify the state of a GPCR during the simulation, we use hallmarks of class A GPCR activation to define the state of the frame in the simulation compared to a reference active or inactive structure. The used order parameters are: distance between TM helix 3 (residue 3.50) and TM helix 6 (residue 6.30), distance between TM helix 3 (residue 3.50) and TM helix 7 (residue 7.53), and RMSD of the NPxxY(7.53) motif to the inactive state reference. Note that these order parameters are not the same for other classes of GPCRs and would require customization. This module also couples with the ligand PCA module to find the activation states of every ligand binding pose cluster (Fig. 2.5c).



(a) Plot obtained by plotting TM3-6 distance vs RMSD of NPxxY motif with respect to inactive reference. (left) scatter colored by replica (10 replicas total). (right) compiled data from all runs.



(b) Plot obtained by plotting TM3-6 distance vs distance. (left) scatter colored by replica (10 replicas total). (right) compiled data from all runs.



(c) Plot of order parameters for the top three ligand binding clusters identified via PCA.

Figure 2.5: Class A GPCR activation states for dopamine-bound D2 receptor simulations.

2.4.5 Mutual information (MI) calculation

Mutual Information (MI) is a statistical measure that quantifies the degree of dependency or information shared between two random variables. It is used in various fields, including information theory, statistics, and bioinformatics, to assess the relationship between two variables. MI measures how much knowing the value of one variable (let's call it A) tells us about the value of another variable (let's call it B). In other words, it quantifies the reduction in uncertainty about variable B when we know the value of variable A.

In this application, MI is calculated from correlated motions extracted from MD simulations. The approach uses internal coordinates with focus on dihedral angles, which are the slowest degree of freedom in the bond-angle-dihedral coordinate system. The advantage of using dihedrals is two-fold, it avoids high frequency noise in the data, and it captures correlated changes in side chain rotamers (225). MI is calculated via the mutual information expansion (MIE), as formulated by Killian et al. (273). In short, a list of all backbone (φ and ψ) and side chain (χ_1 up to χ_5 where applicable) torsion angles is built from the initial structure (similar to a topology file, with a dihedral index, dihedral type (backbone or sidechain), and atom numbers that make up the four dihedral atoms), dihedrals are then extracted every 100 ps after removing the first 50 to 150 ns of every replica of the simulation (depending on the equilibration of the RMSD of Cα coordinates). The dihedral time series are then histogrammed in radian space using 50 bins (50 x 50 bins for 2-dimensional histograms), and the marginal entropy is calculated using the following :

$$S_{\phi_i} = -R \sum_{n=1}^{B_i} P_{\phi_i}(n) \ln\Big(\frac{P_{\phi_i}(n)}{h_\phi}\Big), \tag{2.1}$$

where ϕ_i is the dihedral sampled for residue i, R is the gas constant, and B_i is the number of bins. $P_{\phi_i}(n)$ is the probability of finding ϕ_i in bin n defined as: $P_{\phi_i}(n) = \frac{N_i(n)}{N}$, with $N_i(n)$ being the number of snapshots/datapoints where ϕ_i falls in bin n and N the total number of snapshots/datapoints. h_ϕ is the width of each bin in the histogram defined by the dihedral ϕ_i. For two dihedrals ϕ_i and ψ_i belonging to residues i and j, the joint entropy is written as:

$$S_{\phi_i\psi_j} = -R \sum_{n=1}^{B_i} \sum_{m=1}^{B_j} P_{\phi_i\psi_j}(n,m) \ln\Big(\frac{P_{\phi_i\psi_j}(n,m)}{h_\phi h_\psi}\Big), \tag{2.2}$$

where $P_{\phi_i\psi_j}(n,m)$ is the joint probability of finding ϕ_i in bin n and ψ_i in bin m. We then get the corresponding mutual information term $I_{\phi_i\psi_j}$:

$$I_{\phi_i\psi_j} = S_{\phi_i} + S_{\psi_j} - S_{\phi_i\psi_j}. \tag{2.3}$$

Since probability distributions are calculated from finite length simulations with a finite

number of snapshots, correction for finite size effects is also added:

$$\left\langle S^{observed} \right\rangle = S - \frac{M-1}{2N}, \tag{2.4}$$

where $\left\langle S^{observed} \right\rangle$ is the estimated entropy using N datapoints and M is the number of histogram bins with non-zero probability (274; 275).

Practically, we consider MI between residue pairs i and j, which is simply the sum of MI between the pair of residues' dihedrals:

$$I(i,j) = \sum_{k=\varphi,\psi,\chi's} \sum_{l=\varphi,\psi,\chi's} I_{k,l}. \tag{2.5}$$

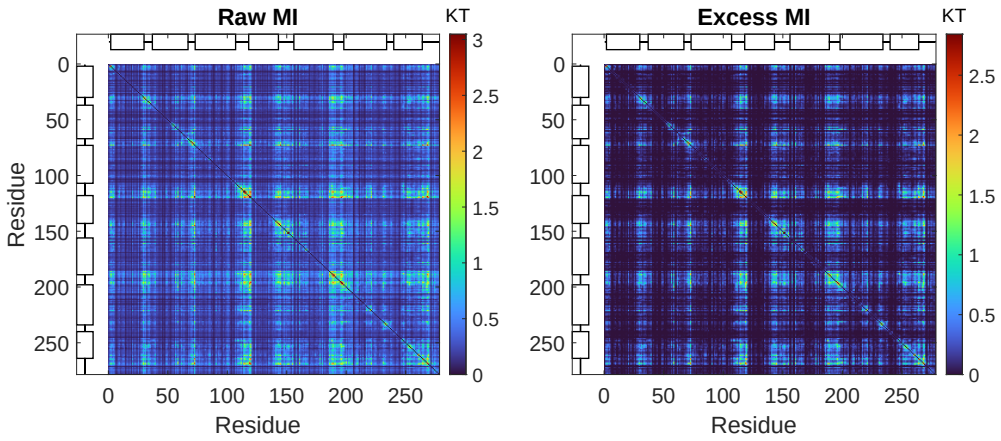


Figure 2.6: **MI maps before and after filtering:** Raw MI (left) and excess MI (right) calculated from dopamine-bound DD2R simulations. 7-TM Helices are shown for reference.

2.4.6 Statistical filtering and significance testing of MI

Another side effect of finite sample sizes is nonzero mutual information in independent datasets. To correct for this effect, we divide the observed MI space into 100 bins. In each MI bin, we randomly pick 5 dihedral pairs (or less if the bin has <5 samples) to represent the bin, and then for every dihedral pair ϕ_i and ψ_j, we shuffle the time series of one of the observed dihedrals and recalculate MI with the shuffled dihedral. This process is repeated until the shuffled dihedral MI converges, and then the average of the resulting MI over the chosen dihedral pairs approximates the nonzero independent MI for a given MI bin. This value of independent MI is then subtracted from all MI values belonging to the bin to get the "excess" MI (Figs. 2.6 and 2.8). These permutations can also be used as a test of significance for MI, and the percentage of MI values from the various permutations that are larger than the observed MI approximates a p-value (225). We used an MI significance level of $p < 0.01$ in our analysis.

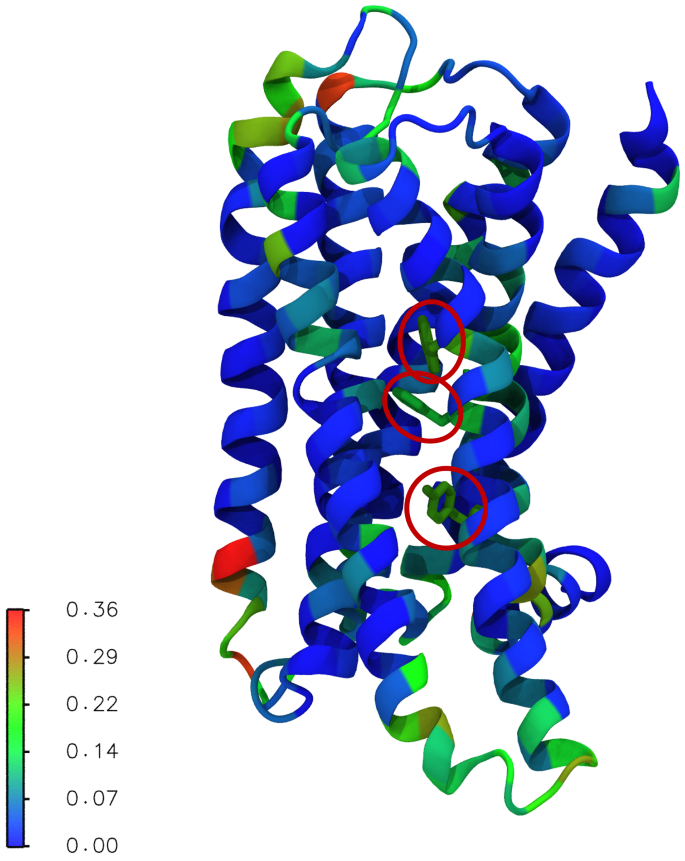


Figure 2.7: **MI mapped on structure:** Average summed MI per residue mapped on the structure of dopamine-bound DD2R. This data is obtained by summing the rows of Fig. 2.6 (right) and then dividing by the number of residues. Backbone is shown in cartoon representation, and conserved residue F6.44, W6.48, and Y7.53 are highlighted in licorice. Note that these conserved residues have high MI in the protein core.

The independent MI versus "raw" MI is shown in Fig. 2.8. The plateau level of independent MI depends on the number of datapoints and number of bins.

2.4.7 Convergence of entropies

In order to acquire precise entropy and MI measurements up to the second order, it is necessary to run simulations for significantly longer durations compared to the time it takes for the slowest autocorrelation and pair correlation processes to unfold. Moreover, the data should ideally be comprised of independent observations. However, due to computational limitations, achieving these ideal conditions is seldom feasible. Consequently, molecules within simulations tend to retain some memory of their initial states. MI filtering and corrections for finite sampling help with this issue, but do not counteract it fully. To assess the convergence of entropies (and thus mutual information), we plot the the dihedral contributions to the

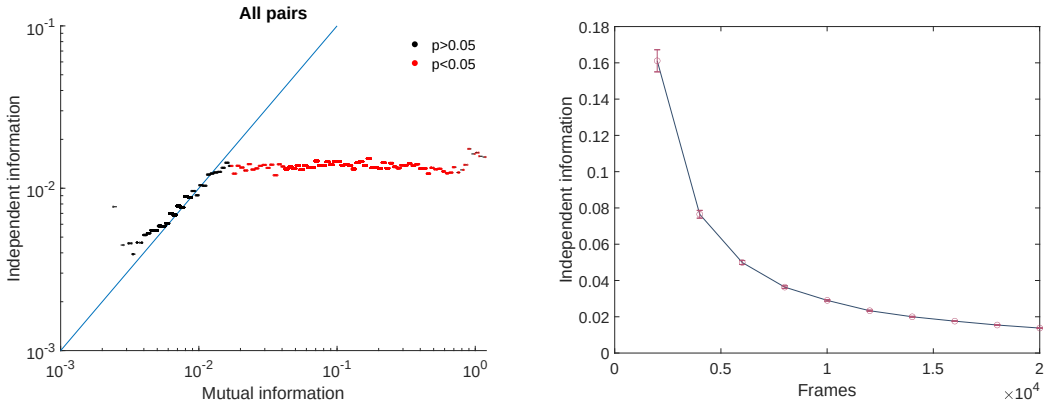


Figure 2.8: **MI filtering and significance testing:** (left) Independent information vs mutual information for dopamine bound DD2R simulations for total simulation time of $2\mu s$. The blue solid line shows the $x = y$ diagonal. Every "cluster" of dots shows a MI bin (100 bins), where the color represents significance (black are not significant, while red are significant, $\alpha = 0.05$). (right) Level of independent information (II) as a function of number of frames considered for the calculation. II level is calculated by averaging over the bins with significant II (red, left panel). Error bars are the standard error of the mean. Frames total to $2\mu s$ of simulation time.

first and second-order approximations to the entropy as a function of number of frames and double check their convergence in the simulations (Fig. 2.9). The first and second-order approximations to S are given by the following equations:

$$S^{(1)} \equiv \sum_{\phi} S_{\phi}, \tag{2.6}$$

and

$$S^{(2)} \equiv S^{(1)} - \sum_{\phi} \sum_{\psi>\phi} I_{\phi\psi}. \tag{2.7}$$

Due to limitations on computation time, we consider all dihedrals for the first-order entropy but only a subset of dihedrals with highest MI contributions for the second-order entropy. In this work we take contributions from the top 100 dihedrals with highest MI. Note that this method of assessing convergence has the downside of looking at the global state of the system rather than individual degrees of freedom.

2.4.8 Allosteric pathway and pipeline calculation

Building on advancements and methods of calculating allosteric networks in the field (233; 15) (Sec. 1.4.2), we studied allostery in GPCRs by representing the protein as a graph and constructing shortest paths (paths of "maximum information") in the graph followed by clustering of said paths. The goals of this process are the following:

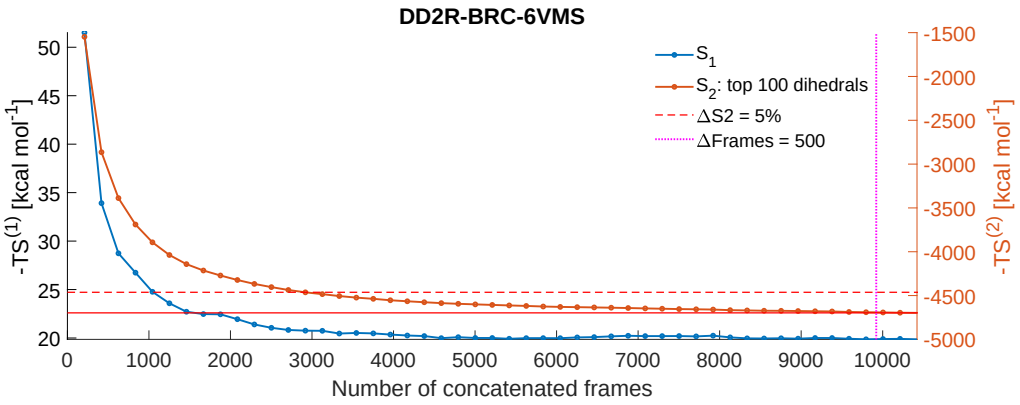


Figure 2.9: **Convergence of entropies:** Convergence of first-order entropy ($S^{(1)}$, blue) and second-order entropy ($S^{(2)}$, red) approximated from dihedral angles. Solid red line shows the final value of $S^{(2)}$ using all frames. Dashed red line shows 5% difference in $S^{(2)}$ for reference. Dotted purple line shows the last 500 frames, equivalent to 50 ns.

1. Get a score for how important a given residue is for allosteric signaling in the receptor under the simulation conditions. This score could be used to target sites for mutagenesis.

2. Establish connectivities between different parts of the receptor, for example between the ligand and G-protein binding sites.

Graph construction

An undirected weighted graph is constructed using residue coordinates and MI. The graph is intentionally undirected due to the utilization of a specific correlation measure in this research, namely mutual information. This choice aligns with the inherent symmetry of MI ($I(a,b) = I(b,a)$), a key characteristic that reflects the mutual dependence between two variables without regard to their order. In contrast, other measures such as transfer entropy exhibit directional properties (226; 276). The construction of such a graph involves three pivotal decisions that significantly influence its structure and utility. These decisions pertain to the selection of nodes, the establishment of edges, and the assignment of edge weights. Each choice in this triad plays a crucial role in shaping the graph and, consequently, in determining its suitability for capturing the intricate relationships within the simulation being analyzed.

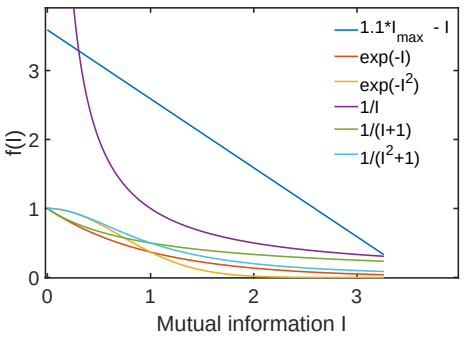


Figure 2.10: Examples of monotone decreasing functions of MI that could be used as edge weights over the range of MI values observed in my simulations.

We consider residues to be nodes centered at the Cα position, where edges are defined between nodes that are less than r_c apart and have significant MI between them. A typical value for r_c is 7.5 Å, which is meant to approximate Cα-Cα cutoff distances used in the literature (277).

Regarding edge weights, the desired property is a monotonic decreasing function of MI. Since MI is non-negative, it suffices that the functions are monotone decreasing over the interval where MI is defined ($[0,\infty]$). Fig. 2.10 shows a variety of functions that fit the aforementioned criteria (with the exception of $1/I$, which is ill defined at $I = 0$. This is not a problem since edges are also not defined when $I = 0$). After visualization of edge weight maps for different functions (Fig. 2.11), we settled on $\exp(-I)$ for the edge weights, but other functions would also serve a similar purpose. For example, a recent preprint uses the negative logarithm of a generalized correlation constructed from mutual information as edge weights in a similar context (278).

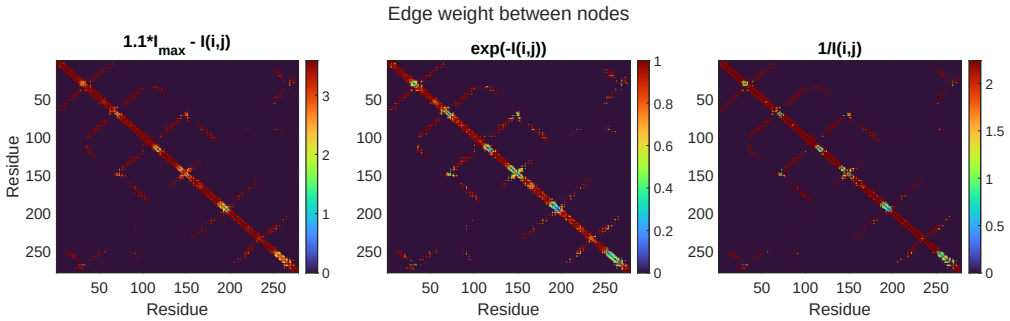


Figure 2.11: **Edge weight maps for various edge weight functions:** Examples of edge weight maps for different functions of mutual information. Functions are shown in the subpanel titles. The colorbar for $1/I(i,j)$ goes beyond the range $[0,2]$, but this range is shown for clarity.

Now that the graph is constructed, the question of *how* do we use the graph to reach the aforementioned goals. Some ideas that could come to mind are centrality metrics (such as betweenness centrality), which would rank the nodes by their importance for the connectivity of the graph. Another operation to perform on the graph is clustering, which would give us communities within the graph that are more connected within each other than with the rest of the graph. Since we are interested in constructing pathways within the protein and ranking residues by their importance, we first construct shortest paths and then cluster them to get both allosteric signaling pathways and residue scores.

Shortest path calculation

After the construction of a graph from protein structure and MI, we calculate a set of shortest paths connecting every pair of nodes that have significant MI and that are further than 10 Å apart using Dijkstra's shortest path algorithm (231) as implemented in Matlab (268).

The distance of 10 Å is chosen by investigating the average MI vs distance plot (Fig. 2.12 and Fig. A.1 for more examples.) and picking a distance where the direct communication peak dies out. After construction of the set of shortest paths between all eligible residue pairs, the pathways are ranked according to the MI of their end nodes. After that, top percentage of pathways that cover I_c% of total MI are chosen for clustering. I_c = 85% has been used in this work. For example, in dopamine-bound DD2R simulations, I_c = 85% would represent taking the top 18% pathways with highest pair MI (Fig. 2.13).

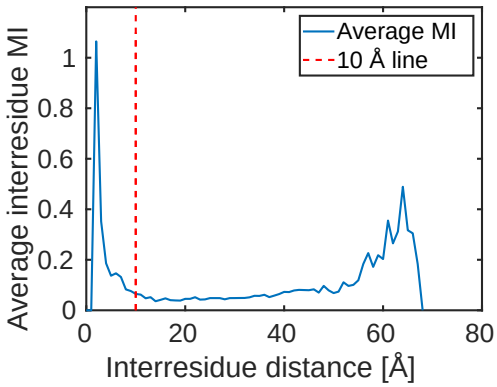


Figure 2.12: Average interresidue MI vs interresidue distance for dopamine-bound DD2R simulation in the active state. The 10 Å line is shown for clarity. The first peak at 2 Å represents "direct communication" between residues in close proximity, while the further peak at 64 Å represents "allosteric communication". Fig. A.1 shows the curve for more class A GPCRs.

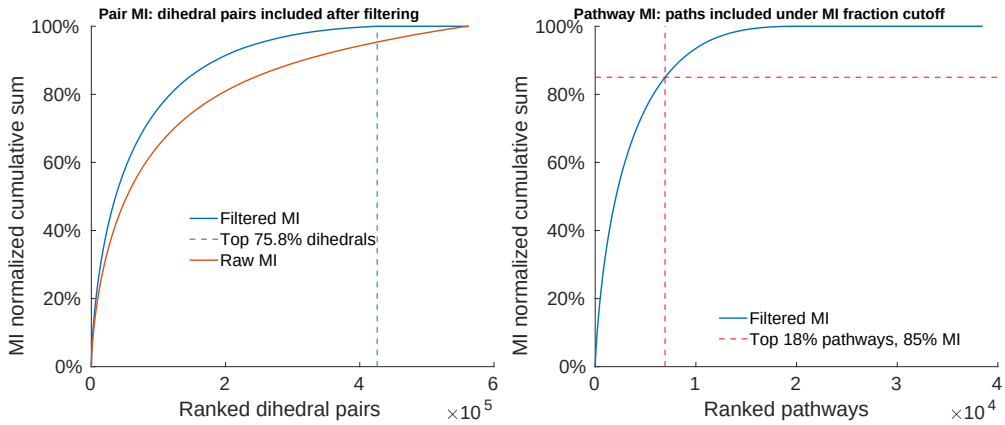


Figure 2.13: **Dihedral pair and pathway MI statistics:** (Left) MI normalized cumulative sum plotted versus dihedral pairs ranked from highest to lowest MI. 75.8% of dihedral pairs contribute significant MI after filtering. (Right) MI normalized cumulative sum plotted versus pathways ranked from highest to lowest end node pair MI. The top 18% pathways with highest pair MI cover 85% of total MI, and are thus considered for clustering.

Path clustering

The next step in the analysis is clustering the constructed paths into larger pathway structures. The first ingredient for clustering is a similarity matrix. A simple measure of similarity is the Jaccard index, which measures the overlap between two sets as the size of the intersection of the two sets divided by the size of their union $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. The Jaccard index as defined here is not suitable since it does not give leeway to consider interactions between residues. For example, two pathways containing residues in close proximity that are physically interacting with one another ought to be clustered together. Instead, we could measure similarity between pathways according to their proximity in the space. Given two pathways k and l of lengths N_k and N_l, denote the number of nodes in k that are within a cutoff distance r_c (typically $r_c = 7.5$ Å) from l as ν_1 and denote the number of nodes in l that are within r_c from k as ν_2. Inspired from similar works in the literature (233), we define an overlap parameter Φ between two pathways k and l:

$$\Phi(k,l) = \max\left(\frac{\nu_1}{N_k}, \frac{\nu_2}{N_l}\right) - \alpha * \frac{|N_k - N_l|}{\rho}, \tag{2.8}$$

where the first term represents the "soft" overlap between k and l and the second term is a penalty for the difference in length between the two pathways. α is a weighting factor for the penalty term, and ρ is the maximum possible path length difference in all considered pathways. ρ includes information about the width of the distribution of path lengths into the overlap metric as a kind of normalization (Fig. 2.14). An overlap cutoff Φ_c is used where any $\Phi < \Phi_c$ is set to zero. We typically use $\Phi_c = 0.75$ in this work. The weighting factor α controls whether smaller paths can integrate with larger overlapping paths in the cluster, if $\alpha \geq 1 - \Phi_c$, then smaller paths cannot cluster with larger overlapping paths, while if $\alpha < 1 - \Phi_c$, then smaller paths can cluster with larger overlapping paths. We generally allow for smaller paths to merge with larger ones to get larger clusters.

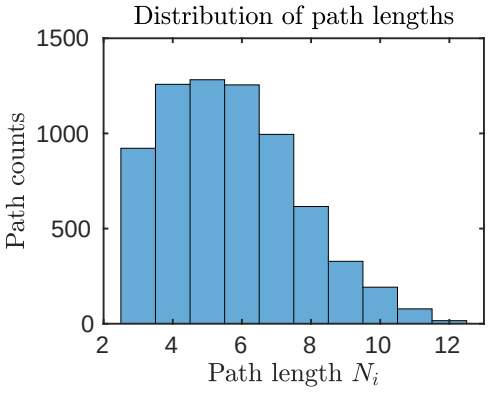


Figure 2.14: Distribution of shortest path lengths considered for clustering in dopamine-bound DD2R simulations.

Now that a similarity matrix between all considered pathways is obtained from $\Phi(k,l)$s, a linkage matrix is constructed from the similarity matrix by pairing paths into binary clusters, then the newly formed pair clusters are grouped into larger clusters until a hierarchical tree is obtained. Under the assumption that the set of shortest paths merge into larger pathway structures that describe allosteric communication, hierarchical clustering is fitting for this application. Note that the shortest paths in graph space represent paths of highest mutual

information. To find the optimal number of clusters, we use a metric that maximizes intra-cluster overlap while minimizing inter-cluster overlap as described previously in the literature (233).

To reduce the number of clusters further and thus improve interpretability, a second clustering step is performed by combining clusters that are too similar. Consider two clusters that resulted from the aforementioned clustering ζ_1 and ζ_2. For every residue m belonging to ζ_1 or ζ_2, we construct a score $\lambda_m(\zeta_x)$ by counting the number of pathways belonging to cluster ζ_x containing residue m. Similarity among clusters is then defined as:

$$\frac{\sum_{m\in(\zeta_1\cap\zeta_2)}(\lambda_m(\zeta_1)+\lambda_m(\zeta_2))}{\sum_{m\in\zeta_1}\lambda_m(\zeta_1)+\sum_{m\in\zeta_2}\lambda_m(\zeta_2)}.$$

Like the first clustering step, a linkage matrix is defined from the similarity matrix, which is followed by hierarchical clustering and determination of number of clusters by maximizing intra-cluster overlap while minimizing inter-cluster overlap.

After both clustering steps, we end up with more populated clusters as shown in Fig. 2.15. The allosteric strength of a cluster η is the MI weighted sum of pathways belonging to the cluster, where every pathway is weighted by the MI value of the pathway's end residues (for distribution differences between mean path MI and path end pair MI, check Fig. A.2). Thus the allosteric strength of cluster η can be written as:

$$\sigma_\eta = \sum_{l\subset\eta} I(i,j), \tag{2.9}$$

where l is a pathway belonging to cluster η, i and j are the end nodes of pathway l, $I(i,j)$ is the MI between residues i and j.

A similar metric can be used to measure the allosteric strength of individual residues by calculating the end node MI-weighted sum of all pathways going through a given residue m:

$$\sigma_m = \sum_{l\forall m\in l} I(i,j). \tag{2.10}$$

Allosteric residues are defined as the residues with the largest number of MI-weighted allosteric pathways passing through them, i.e., largest σ_m (Fig. 2.18a). This metric is versatile as the set of pathways to sum over could be filtered according: belonging to a cluster η, pathways containing ligand binding residues, pathways containing G-protein binding residues, etc ...

Robustness of clusters

To check the robustness of the clustering algorithm, we reran the *md2path* pipeline using 75% of the dopamine-bound DD2R simulations while permuting the sampled frames three times. We report the mean MI vs cluster size, number of clusters, and topology of top clusters in this section. The number of clusters changes slightly between permutations (Fig. 2.16), but the trends remain very similar between different tests.

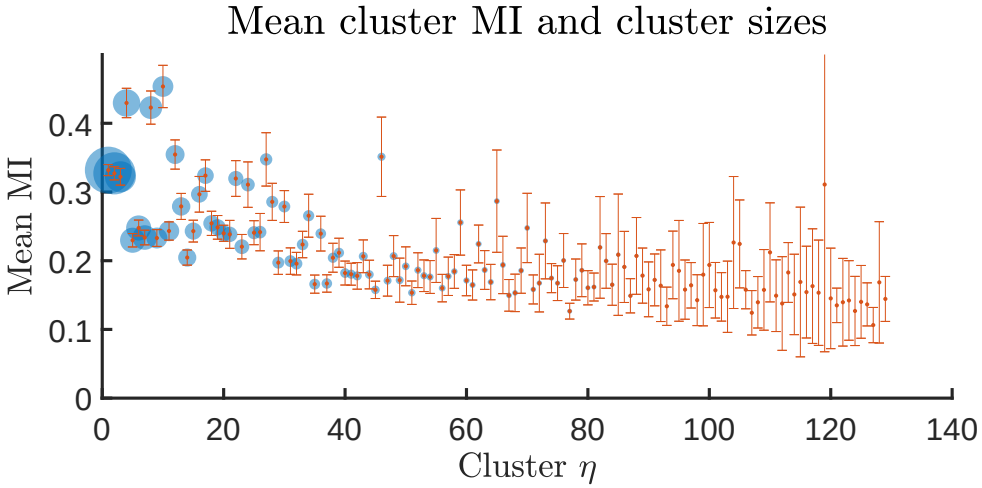


Figure 2.15: **Mean cluster MI and cluster sizes:** mean cluster MI is calculated as the mean of the end node path MI for every path belonging to the cluster. Error bars show standard error of the MI within a given cluster. Cluster sizes are represented by the size of the blue dots (for reference, the largest 3 clusters contain 894, 722, and 392 pathways respectively). The total number of clusters is 129, with 81 clusters containing > 10 pathways.

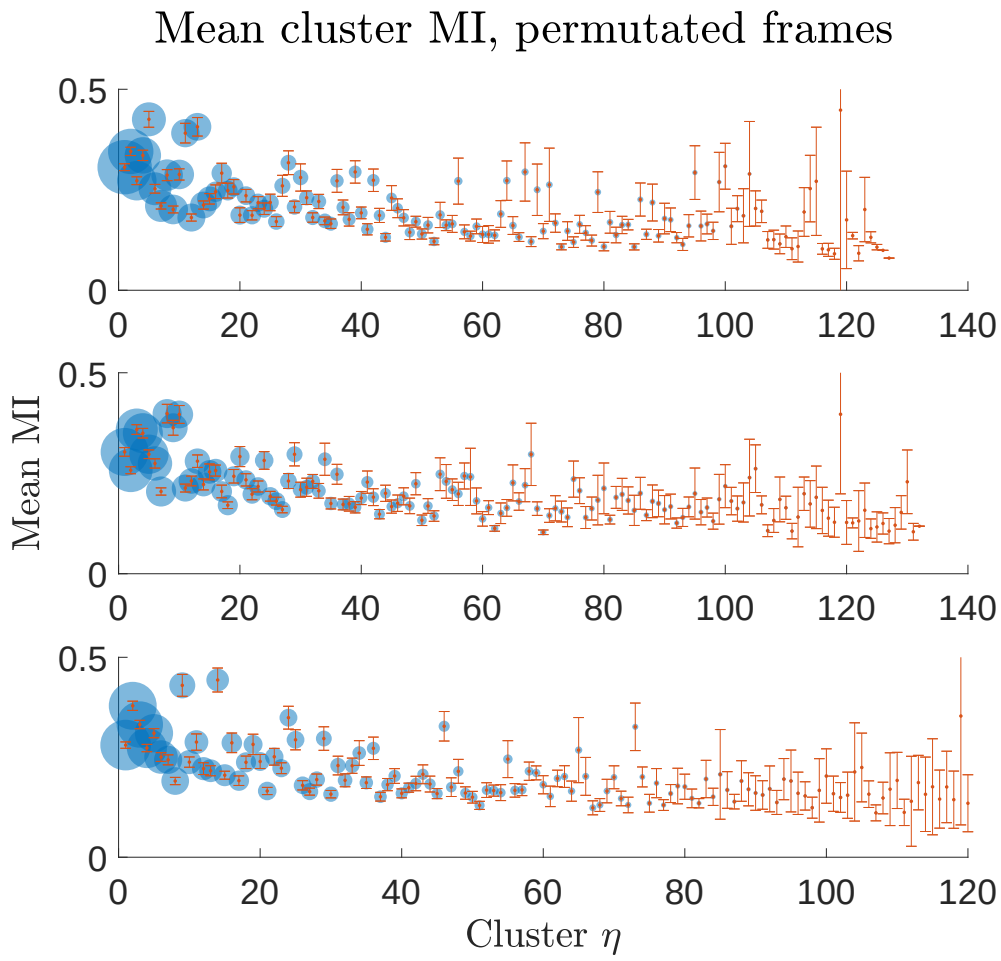


Figure 2.16: **Mean cluster MI and cluster sizes for resampled data:** Data is resampled by permuting sampled frames for 75% of simulated time (translating to 1.5μs). This process is repeated three times. Mean cluster MI is calculated as the mean of the end node path MI for every path belonging to the cluster. Error bars show standard error of the MI within a given cluster. Cluster sizes are represented by the size of the blue dots. Number of clusters from top to bottom panels are 127, 132, and 120 respectively.

Robustness of residue allosteric strength scores σ_m

Similar to the last section, we reran the md2path pipeline using 75% of the dopamine-bound DD2R simulations while permuting the sampled frames three times and extracted the residue allosteric strength scores σ_m and compared them within each other and with the full simulation. The results are reported in Fig. 2.17. The distribution and ranking of the residues is largely similar between the tested sets. One observation is that the full simulation has consistently lower allosteric strengths than the randomized sets that cover 75% of the simulation. A following test where we ran another $5 \times 1\mu s$ simulations of dopamine-bound DD2R confirmed this, showing a relatively similar distribution but yet lower mean σ_m (Fig. A.3). There are two sources for this effect, the first is the MI filtering, where the level of independent MI depends on the sampled frames (Fig. 2.8:left), and thus less sampling leads to generally lower MI due to the independent MI level being higher. The second source is the number of pathways considered for analysis (Fig. 2.13:right), where less sampling leads to taking higher percentage of pathways to reach the desired MI cutoff.

The takeaway from this section is that while the distribution of σ_m is robust to changes in sampling, the absolute values depend on the sampled frames, and thus care must be taken when comparing values of σ_m coming from simulations with vastly different number of frames.

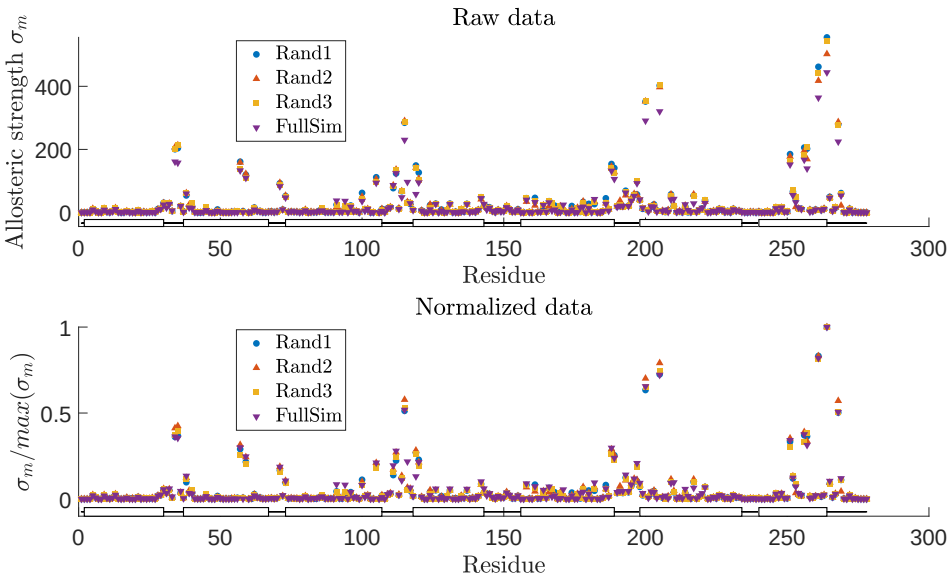


Figure 2.17: **Robustness of allosteric strength scores:** Data is resampled by permuting sampled frames for 75% of simulated time (translating to $1.5\mu s$). This process is repeated three times. (top) Raw allosteric strength scores σ_m are plotted. (bottom) Allosteric strength scores normalized by maximum score for every analyzed set of frames.

Alternatives to path clustering

It is possible to get allosteric scoring and pathway-like structures without going through a clustering step. One option is to filter pathways going through a desired set of residues (ligand binding or effector binding sites for example), and then calculate the density of pathways going through every residue.

A similar alternative that is also density based is to divide the Cartesian space occupied by the protein into voxels and calculate path densities at every voxel. This defines volumes of high density within the protein structure that would be most important for allosteric communication (Fig. 2.18b).

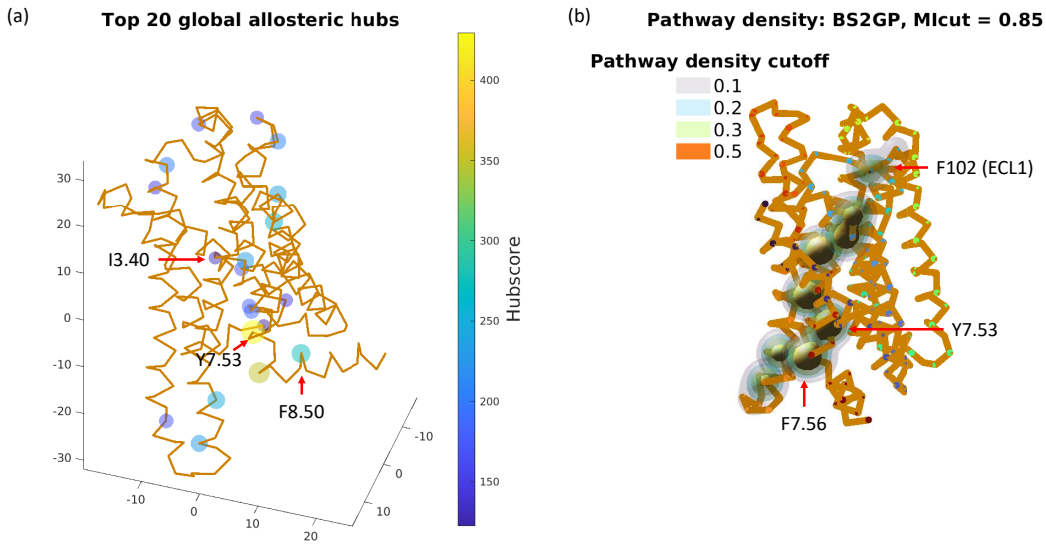


Figure 2.18: **Path density from dopamine-bound DD2R simulations shown on protein structure:** (a) Top 20 global allosteric hubs (without filtering by specific residues) with highest hubscores shown on DD2R structure. (b) Path density for paths filtered by ligand binding and G-protein binding residues using 85% of MI content. Path densities are colored according to their intensity compared to maximum path density. Selected residues are highlighted to help guide the reader.

These approaches are mentioned here for completion since they were considered during development, but they are not used further.

Limitations to the approach

While the path calculation approach has proven to be a powerful method to understand and design allostery in GPCRs, it is not without its limitations:

1. Number of post analysis steps lead to a certain difficulty in interpretability.
2. Complexity of structures that emerge from clustering (clusters η) make comparisons between clusters from different systems difficult.
3. Lack of directionality in the correlation metrics and subsequently in the graph description of the proteins.

While it is possible to deal with more than one ensemble (and formally compare them, as we will see in Sec. 2.5), this approach finds difficulty to deal with the transitions and transition state simulations, since it needs converged statistics describing a somehow well defined state of the system.

2.5 *KLdiv* module: perturbation response by ensemble comparison using Kullback–Leibler divergences

In contrast to the *md2path* module that analyzes a set of simulations for an individual system, *kldiv* module performs an ensemble comparison between sets of simulations of a test and a reference system, going through the steps mentioned in Sec. 2.1.1. KL highlights degrees of freedom that are different between the test and reference sets of simulations by calculating the distance between distributions of a given d.o.f. in the two systems, thus quantifying the perturbation response of the system in question. This perturbation is defined as the difference between the test (or target) and reference systems. Typical perturbations are amino acid substitutions, ligand changes, or protein state differences.

2.5.1 Dihedral reconciliation

Dihedral time series are constructed in a similar fashion to the last section (Sec. 2.4). After constructing the dihedral list for every system, differences in dihedral types are resolved to construct a common topology that applies to both systems. For example, if the test system contains a mutation from Arginine (φ, ψ, and χ_1 to χ_5) to Glycine (φ and ψ), only the common dihedrals present in both simulations (φ and ψ) contribute to the KL divergence.

2.5.2 Kullback-Leibler (KL) divergence calculation

KL calculation is performed as described in Mcclendon et al. (279). From the KL-divergence expansion, we consider the first order term, denoted the "local" KL divergence. To perform this calculation, dihedrals are extracted every 100 ps after removing the first 50 to 150 ns of every replica of the simulation (depending on the equilibration of the RMSD of Cα coordinates). The dihedrals are then histogrammed using 50 bins, and the KL is calculated using the following equation:

$$KL_{\phi} = \sum_{n=1}^{B} P_{\phi}(n) \ln\Big(\frac{P_{\phi}(n)}{P^{*}_{\phi}(n)}\Big), \tag{2.11}$$

where ϕ is the dihedral being sampled, B is the number of bins, $P_{\phi}(n)$ is the probability of finding ϕ in bin n, in the target ensemble, defined as mentioned in the MI calculation section, and $P^{*}_{\phi}(n)$ is the probability of finding ϕ in bin n, in the reference ensemble.

The local KL for a single residue is simply the sum of the KL between reference and target ensembles for each of the applicable dihedrals for a given residue (φ, ψ, and $\chi's$):

$$KL_{res_i} = \sum_{\phi=\varphi,\psi,\chi's} \sum_{n=1}^{B} P_{\phi}(n) \ln\Big(\frac{P_{\phi}(n)}{P^{*}_{\phi}(n)}\Big). \tag{2.12}$$

Despite the similarity of this expression to that of the marginal entropy (S_{ϕ_i}) mentioned before, it is a distinct thermodynamic quantity that measures the dissimilarity of two probability density functions (pdfs), while the marginal entropy S_{ϕ_i} measures the level of disorder of a particular pdf. This makes local KL a suitable criterion for quantifying degree of freedom specific responses to a perturbation to a reference system.

2.5.3 Statistical filtering and significance testing of KL

Corrections to the local KL are applied to counteract the effects of sample variability due to finite sampling (279). To do that, KL is calculated from the reference ensemble using a statistical bootstrapping approach, which is then used for significance testing and for correcting the calculated KL values (which bears similarity to mutual information expansion explained before). Reference sample KL is calculated by dividing the reference simulation into n_b blocks, and then using half the blocks as a proxy "reference" ensemble and the other half as a proxy "target" ensemble. Any non-zero KL between these proxy sets will approximate the bias to KL coming from intra-ensemble variability. The average bias becomes thus:

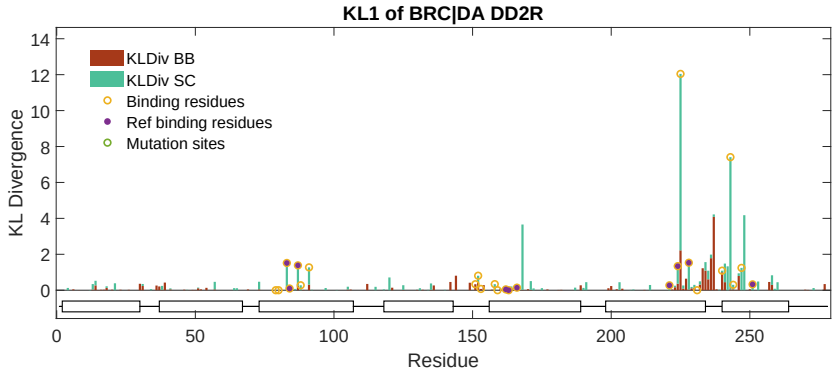
$$KL_{\phi}^{H_0} = \binom{n_b}{n_b/2}^{-1} \sum_{blocks}^{n_b/2} \sum_{n=1}^{B} P_{\phi}(n) \ln\Big(\frac{P_{\phi}^{S}(n)}{P_{\phi}^{S^C}(n)}\Big), \tag{2.13}$$

where S are the subsamples constructed from half the blocks and S^C are the complementary subsamples from the other half of the blocks. The distribution of proxy KL values can be used to derive a p-value for the null hypothesis that the mean KL does not exceed what one would expect based on the variability observed in the reference ensemble. We then define a significance level α, which is used to zero out non-significant KL divergences, and subtract "excess" divergences from the significant ones as follows:

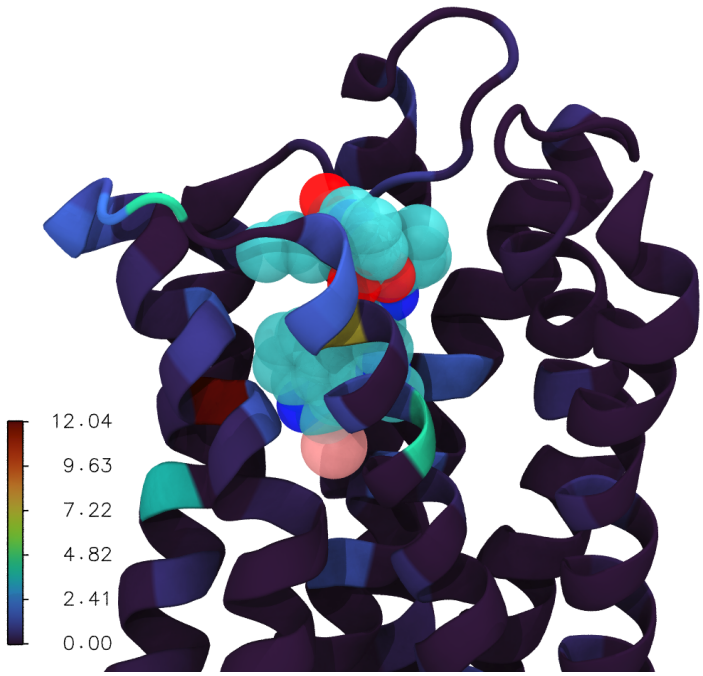
$$\hat{KL}_1 = KL_1 - KL_1^{H_0}.$$

2.5.4 Interpreting the divergences

Since KL divergence is measuring the distance between two distributions, it is reducing these two probability distribution functions (pdfs) into a scalar value. This could lead to difficulties in interpreting the exact values that are output by KL. Despite global KL divergence having connections to the free energy (279), it would be useful to visualize how the exact values of the local KL translate into pdfs. The general trend is that the larger the KL, the more different the two pdfs are, as can be seen in Fig. 2.21 and Fig. 2.22. Note that the KL_1 values reported in the aforementioned figures can arise from a variety of distribution shapes, and thus care must be taken not to overgeneralize from the reported distributions.

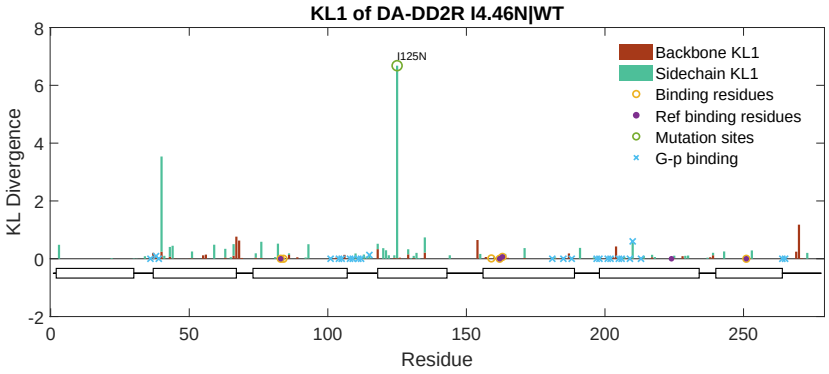


(a) KL per residue for DD2R system with DA-bound simulation as reference and BRC-bound simulation as test. KL contributions are color-coded according to backbone (brown) and sidechain (green). Ligand binding residues for DA and BRC are highlighted. Note that most of the changes are close to ligand binding residues, as is expected when the perturbation is a ligand change. The 7TM helices of GPCRs are highlighted.

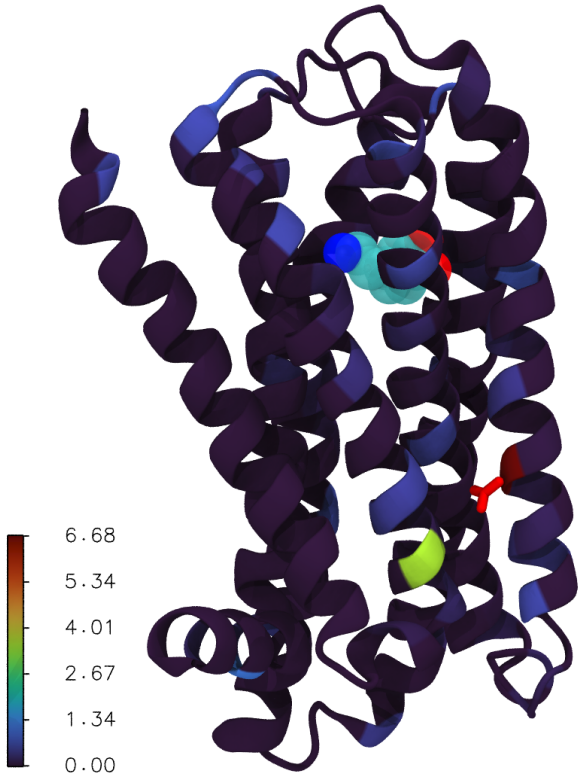


(b) Same KL shown above mapped on the DD2R structure, where the backbone is shown in cartoon representation and colored according to the KL value. Ligand (BRC) is shown in transparent sphere representation.

Figure 2.19: KL per residue for DD2R system with DA-bound simulation as reference and BRC-bound simulation as test.



(a) KL per residue for DA-bound DD2R system with the WT as reference and I4.46N mutated simulation as test. KL contributions are color-coded according to backbone (brown) and sidechain (green). Ligand binding residues, G-protein binding residues, and mutation site are marked. The two residues with highest KL are the mutation site (I4.46N) and Y2.41, which forms a polar interaction with the mutation site. The 7TM helices of GPCRs are highlighted.



(b) Same KL shown above mapped on the DD2R structure, where the backbone is shown in cartoon representation and colored according to the KL value. Ligand (DA) is shown in transparent sphere representation, and mutation site (I4.46N) is shown as licorice.

Figure 2.20: KL per residue for DA-bound DD2R system with WT simulation as reference and I4.46N simulation as target.

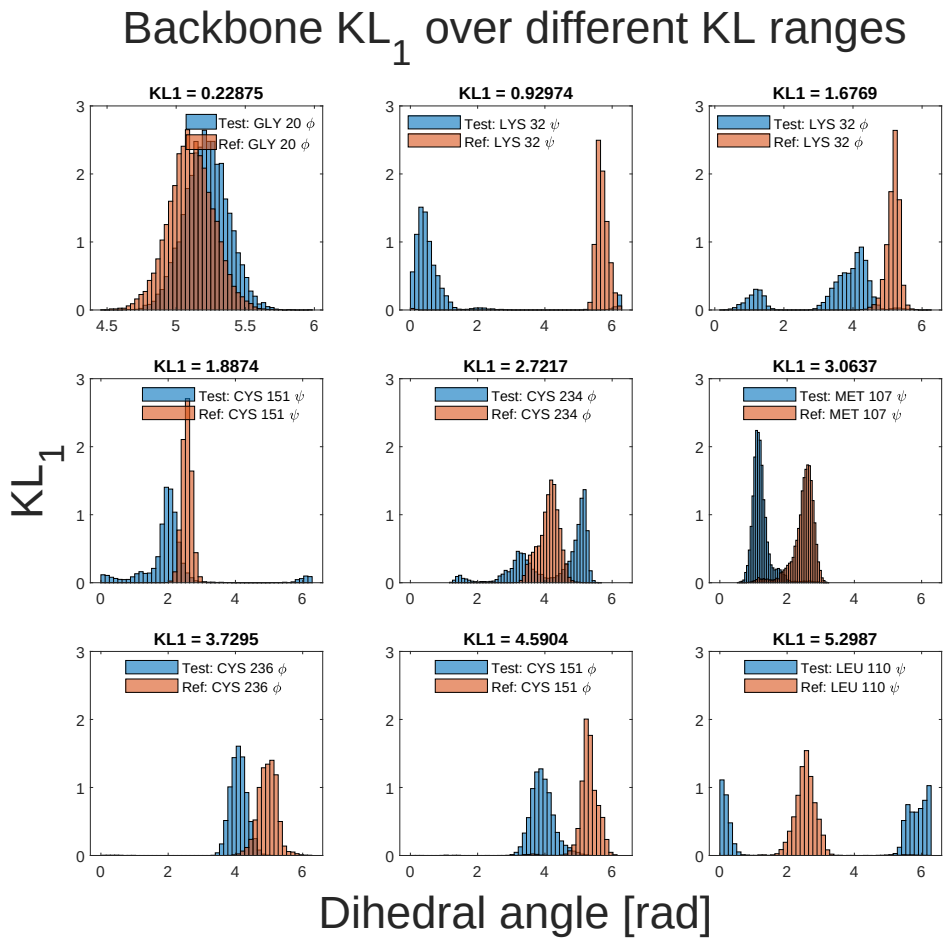


Figure 2.21: KL for a selection of backbone dihedrals starting from smallest to largest KL for dopamine-bound DD2R (active state, blue) simulations as test system and risperidone-bound DD2R (inactive state, brown) simulations as reference system.

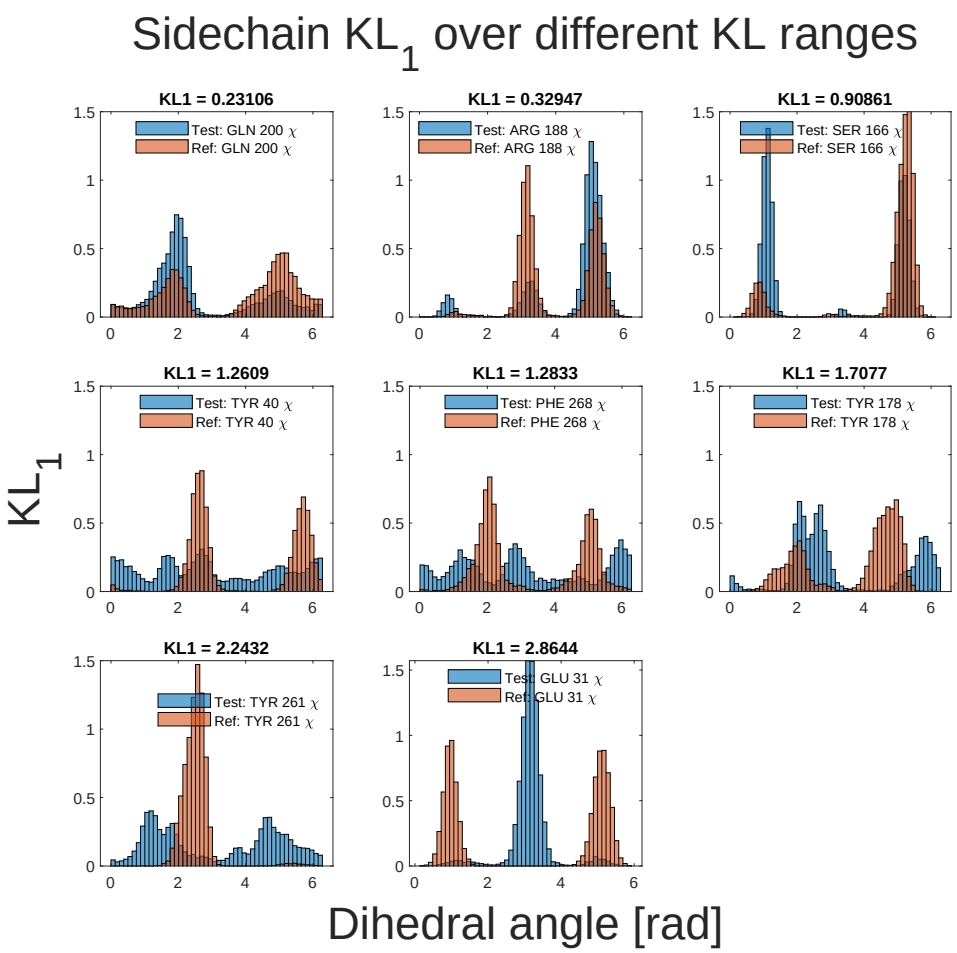


Figure 2.22: KL for a selection of sidechain dihedrals starting from smallest to largest KL for dopamine-bound DD2R (active state, blue) simulations as test system and risperidone-bound DD2R (inactive state, brown) simulations as reference system.

2.5.5 Higher order KL terms

Similar to the mutual information expansion, the Kullback-Leibler expansion also includes higher order terms that contain two or more body interactions. In such an expansion, a term similar to the mutual information can be defined, and is called the mutual divergence M (279). M is defined between two degrees of freedom with pdfs P_i and P_j and joint distribution P_{ij} in the test ensemble, and two degrees of freedom with pdfs P_i^* and P_j^* and joint distribution P_{ij}^* in the reference ensemble:

$$M_2 = \sum_i P_i \ln\left(\frac{P_i}{P_i^*}\right) + \sum_j P_j \ln\left(\frac{P_j}{P_j^*}\right) - \sum_i \sum_j P_{ij} \ln\left(\frac{P_{ij}}{P_{ij}^*}\right). \tag{2.14}$$

The derivation of higher order terms is left to the reference (279). While the mutual information has a simple intuitive interpretation, the mutual divergence between degrees of freedom (d.o.f.) i and j can be interpreted in the following manner: if we know how divergent d.o.f. i is in the test ensemble from the reference ensemble, what do we know about divergence of d.o.f. j between the two ensembles?

To assess the usefulness of the mutual divergence term, M_2 is calculated for systems with different types of perturbations (ligand change, amino acid substitution, activation state change) and observe the additional insight that it may give.

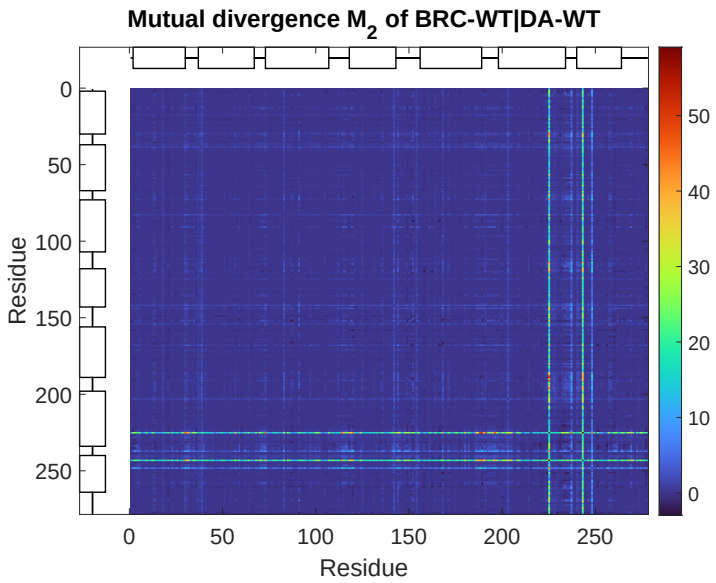
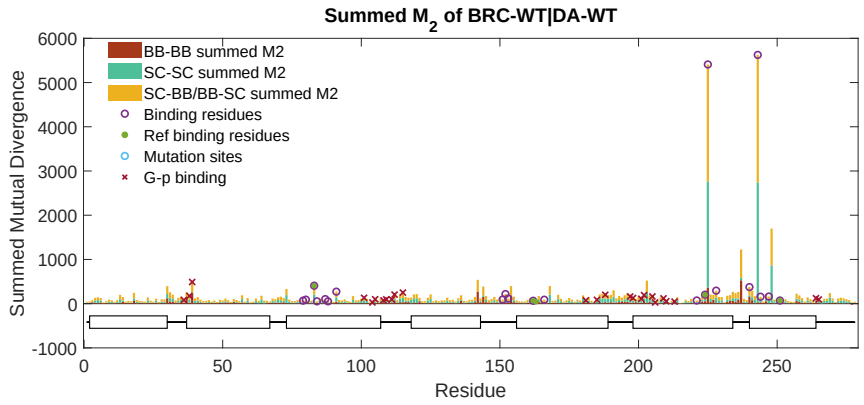


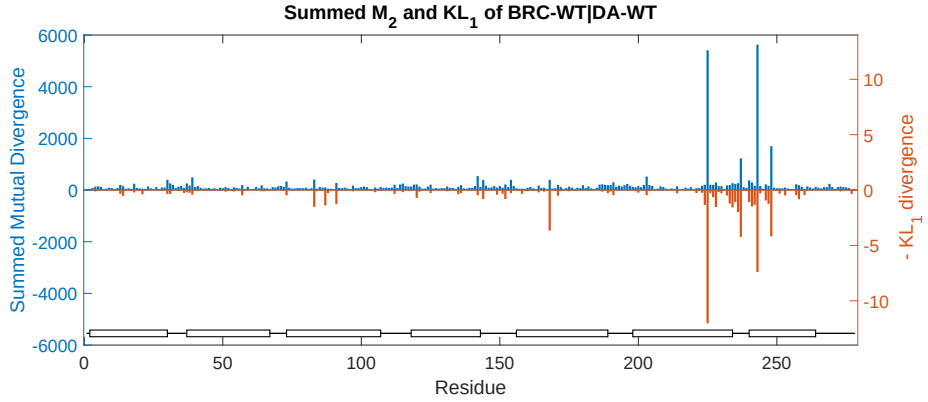
Figure 2.23: Mutual divergence M_2 for DD2R bound to BRC (test ensemble) and DA (reference ensemble).

For the ligand change test, the results are reported for DD2R active state bound to BRC (test ensemble) and DA (reference ensemble). We observe that the mutual divergence is not adding new information to the one term KL divergence, which is signified in the horizontal/vertical

high M_2 lines in Fig. 2.23 and in the very similar peaks when plotting residue summed M_2 and KL_1 on the same axis (Fig. 2.24).



(a) Mutual divergence M_2 summed over residues for DD2R system with the DA-bound as reference and BRC-bound as test ensembles. M_2 contributions are color-coded according to backbone-backbone (brown), sidechain-sidechain (green), and backbone-sidechain (yellow). Ligand binding residues and G-protein binding residues are marked. The largest divergences are in the ligand binding region. The 7TM helices of GPCRs are highlighted.



(b) Same residue summed M_2 (blue, left y-axis) plotted above with KL_1 plotted in the negative direction (red, right y-axis).

Figure 2.24: M_2 and KL_1 comparison of DD2R BRC and DA bound.

Similar results are reported for the two other tested perturbations (amino acid substitution and activation state change), and the figures are found in the appendix (Sec. A.1).

2.6 Relationship of KL-divergences to experimental observables

KL-divergences have been reported to highlight regions showing NMR chemical shifts (IL-2 in response to binding and pH regulation of Talin (279)). In this section, we study the relationship between KL and experimental observables, namely NMR chemical shifts.

When simulating two states of a system, a perturbed and an unperturbed state, the KL will quantify two types of effects, a direct effect in close proximity to the perturbation site and an allosteric effect distant from the perturbation. NMR spectroscopy that reports on structure and dynamics of a protein in response to a ligand or to mutations would make an ideal reference for validation, since differences of chemical shifts between states can identify local changes as well connections in allosteric pathways (202).

While chemical shift calculations from simulations require quantum mechanical methods for geometry optimization and electronic structure calculations, my aim with this comparison is not to re-calculate the chemical shifts *ab-initio*, but to find correlations between what the chemical shifts represent in a certain experiment and the KL-divergences. To this end, we compare KL-divergences calculated from MD simulations using AlloDy to NMR chemical shifts measured for beta-1 adrenergic receptor (β1AR) from the group of Prof. Stephan Grzesiek from the University of Basel (202; 63). What makes the data ideal for comparison is that some of the chemical shifts are interpreted as functional readouts of ligand efficacy or some property of the ligand (affinity, tail volume, etc ...).

2.6.1 Studied system description

In Isogai et al. and Grahl et al., two variants of wild turkey β1AR are studied, a thermostabilized variant (TS) and a variant doubly mutated from TS (YY). The thermostabilized variant was derived from previous turkey β1AR used in crystallography studies by adding three thermostabilizing mutations, the final TS sequence is as follows:

MGAELLSQQWEAGMSLLMALVVLLIVAGNVLVIAAIGSTQRLQTLTNLFITSLACADLVVGLL
VVPFGATLVVRGTWLWGSFLCELWTSLDVLCVTASVETLCVIAIDRYLAITSPFRYQSLMTRA
RAKVIICTVWAISALVSFLPIMMHWWRDEDPQALKCYQDPGCCEFVTNRAYAIASSIISFYIP
LLIMIFVALRVYREAKEQIRKIDRASKRKTSRVMLMREHKALKTLGIIMGVFTLCWLPFFLVN
IVNVFNRDLVPKWLFVAFNWLGYANSAMNPIILCRSPDFRKAFKRLLAFPRKADRRL
HHHHHH.

The YY variant reverts positions 5.58 and 7.53 back to their WT sequences (A5.58Y and L7.53Y from TS). The tyrosines present in these two positions are essential for function and for nanobody binding in the intracellular side. The receptors were studied in an apo state and bound to six different ligands: isoprenaline, dobutamine, alprenolol, carvedilol, cyanopindolol, and atenolol, ranging from full agonist to antagonists (Fig. 2.28).

On the computational side, we prepared the systems for MD simulations by in-silico mutagensis of respective PDB files (check Tab. 2.1 for the full list) using RosettaMembrane (171), building missing loops using RosettaRemodel (170), and ligand protonation using Protoss (280). We then embedded the complex (or receptor in the case of apo simulations) in a POPC lipid membrane followed by solvation and ion addition using CHARMM-GUI (281; 282; 283), and running simulations using Gromacs (284) (See Methods, Sec. 2.6.4).

2.6.2 Results

KL divergences of different perturbations

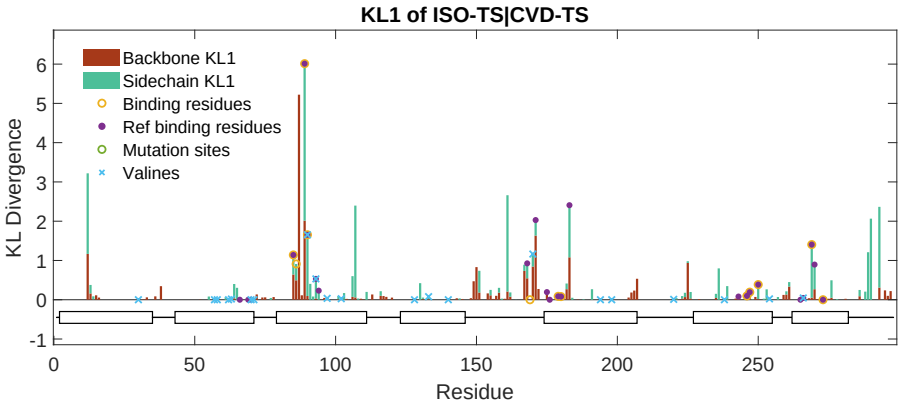
From the simulated systems, we first try to understand the effect of different perturbations introduced in the study before directly comparing with experimental observables.

Effect of agonist vs antagonist: To assess the effect of binding to agonist versus an antagonist, we calculate KL_1 using isoprenalol bound β1AR-TS as target ensemble and carvedilol bound β1AR-TS as reference ensemble (Fig. 2.25). The general differences are highlighted by KL_1 in the ligand binding site as expected, including ECL2. Another pair of significant residues are R3.50 and E6.30, which form an ionic lock that stabilizes the inactive state.

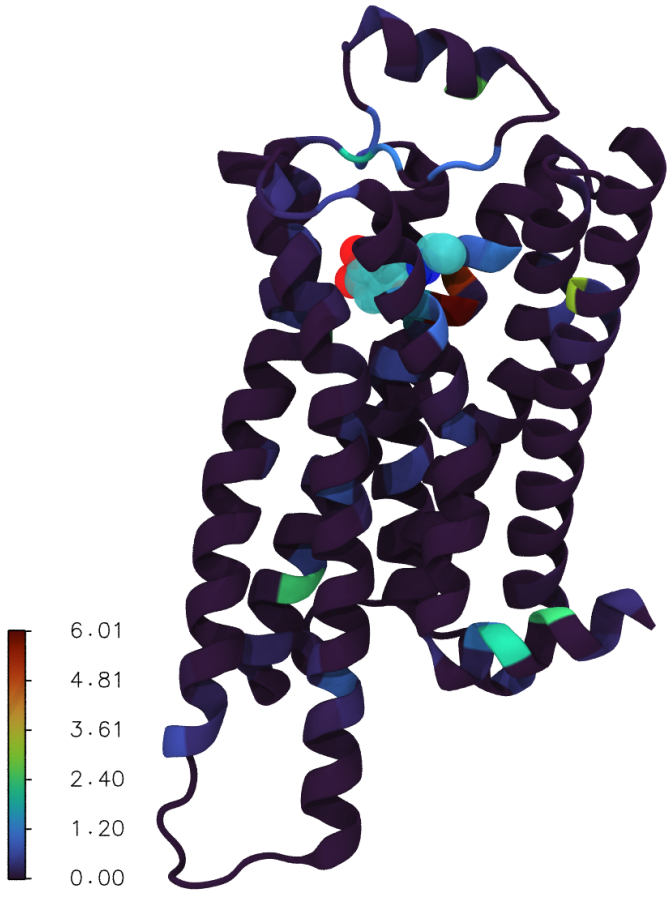
Effect of A5.58Y-L7.53Y double mutation: The double mutation A5.58Y-L7.53Y is crucial for activity of the receptor, for without the two critical tyrosines, the receptor cannot bind its intracellular binding partners. We calculate KL_1 using isoprenalol bound β1AR-YY as target ensemble and isoprenalol bound β1AR-TS as reference ensemble (Fig. 2.26 to assess effect of mutations during simulations starting from an inactive state (PDB code: 2Y03). Large KL signals can be seen in the vicinity of the mutation sites (TM5 and NPxxY motif). Note that KL_1 of A5.58Y residue is small at the site itself due to alanine lacking any sidechain dihedrals, and therefore having only backbone contributions to the divergence. The effect of the mutation is clear on the neighborhood in TM5, however.

Effect of nanobody binding: To capture the effect of nanobody binding, we compare ensembles from simulations of β1AR-YY bound to ISO with and without the nanobody. The YY mutations are necessary for nanobody binding (202), and thus it makes no physical sense to simulate nanobody bound β1AR-TS. Nanobody binding shows the highest KL_1 signal among the studied perturbations, with sidechain signals observed at the core of the receptor (residues V3.40, M5.54, M6.41, and F6.44) and a mixed backbone sidechain signal from the YY mutation sites (Y5.58 and Y7.53) and their neighborhoods. The largest signals, however, come from mostly backbone contributions from ICL3, which is interfacing with the nanobody.

Direct comparison with experimental observables In Isogai et al. (202), the authors correlated NMR chemical shifts at different valine residues with ligand efficacy to Gs (V226(5.57)),

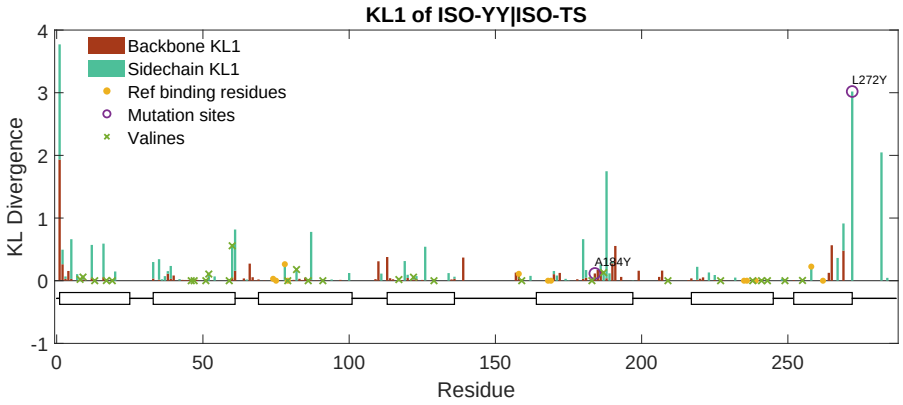


(a) KL per residue where target ensemble is isoprenalol bound β1AR-TS and reference ensemble is carvedilol bound β1AR-TS. Most differences are present at ligand binding sites (as expected), including ECL2. In addition, positions R3.50 and E6.30 also show a KL signal. Helix 8 also shows a significant sidechain KL signal.

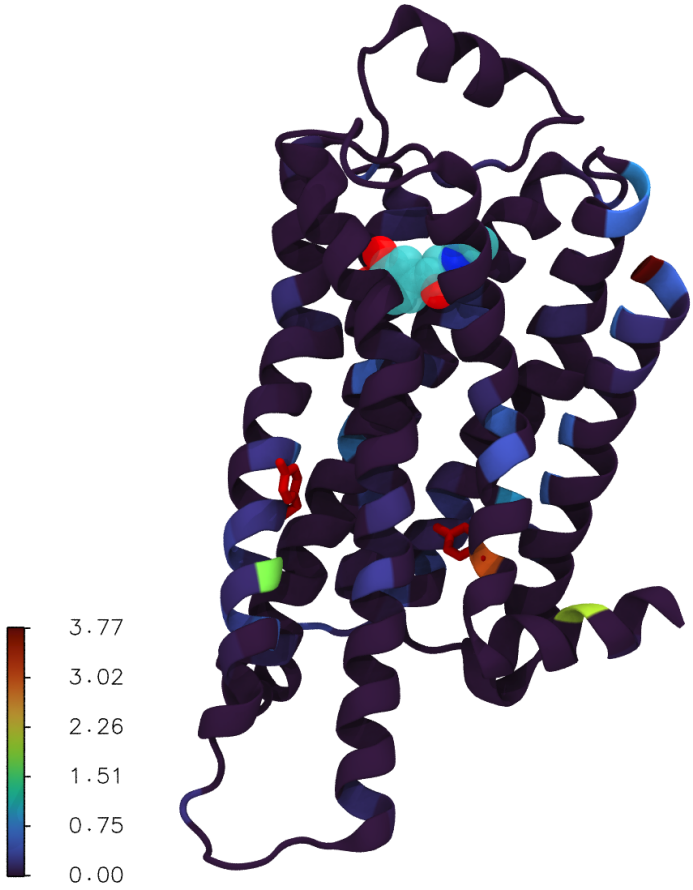


(b) Same KL shown above mapped on the ISO-bound β1AR-TS structure, where the backbone is shown in cartoon representation and colored according to the KL value. Ligand (ISO) is shown in transparent sphere representation.

Figure 2.25: KL_1 **of β1AR-TS to quantify differences between agonist and antagonist binding.**

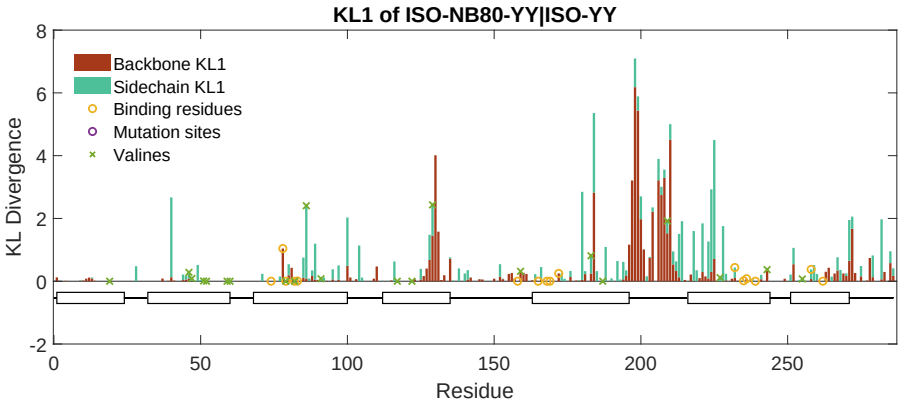


(a) KL per residue where target ensemble is isoprenalol bound β1AR-YY and reference ensemble is isoprenalol bound β1AR-TS. Large KL signals can be seen in the vicinity of the mutation sites (TM5 and NPxxY motif).

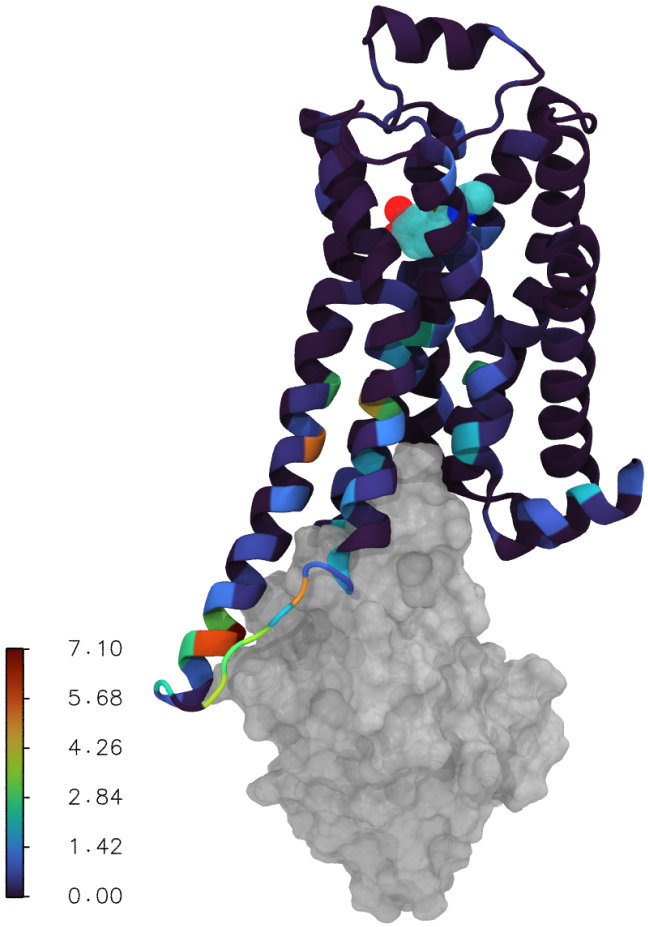


(b) Same KL shown above mapped on the ISO-bound β1AR-TS structure, where the backbone is shown in cartoon representation and colored according to the KL value. Ligand (ISO) is shown in transparent sphere representation. Mutation sites are shown in red licorice.

Figure 2.26: KL_1 **of isoprenalol bound β1AR to quantify differences between YY and TS sequence variants.**



(a) KL per residue where target ensemble is isoprenalol and nb80 bound β1AR-YY and reference ensemble is isoprenalol bound β1AR-YY.



(b) Same KL shown above mapped on the ISO- and Nb80-bound β1AR-TS structure, where the backbone is shown in cartoon representation and colored according to the KL value. Ligand (ISO) is shown in transparent sphere representation. Nb80 is shown in transparent surface representation.

Figure 2.27: KL_1 **of isoprenalol bound β1AR-YY to quantify differences due to nanobody binding.**

affinity (V314(6.59)), tail volume[1] (V103(2.65)), and insertion depth[2] (V125(3.36)). Additional information regarding how these properties were obtained can be found in extended Data table 1 in Isogai et al. (202). In this section, we report KL_1 as sum of contributions from both backbone (ϕ, ψ) and sidechain (χ_1) of valine residues, unless otherwise mentioned. KL_1 of residue V226(5.57) in β1AR bound to different ligands with β1AR-TS apo as reference has an excellent correlation with efficacy of ligands to Gs (Fig. 2.28b), with $R^2 = 0.943$. Note that the full agonists were simulated with nanobody bound (as shown in the figure legend) to mimic an active like state. Another property that correlated with KL_1 at the same site as that observed in NMR experiments is ligand insertion depth measured at V125(3.36) with $R^2 = 0.746$ (Fig. 2.29b). The correlations between KL_1 and Gs efficacy and ligand insertion depth are maintained upon change of reference state of KL-divergence from apo state β1AR-TS to antagonist bound β1AR-TS (CAR) (Fig. 2.30).

On the other hand, correlations are not observed for other ligand specific properties with the same valines reported in Isogai et al., as seen in Fig. 2.29c and d. A possible explanation for the lack of correlation of ligand affinity (which correlated with V314(6.59)) is that the ligand is already in the ligand binding pocket in the simulations, and thus the simulations contain no information on ligand affinity. The lack of correlation with tail volume is puzzling, however, since that is information that is present in the simulations. After scanning the residues close to the ligand binding site, we found an excellent correlation between KL_1 at residue N7.39 and tail volume for the simulated systems with $R^2 = 0.982$ (Fig. 2.31a). This correlation is also present if we consider β1AR-TS bound to the agonists (ISO and DOB) without the nanobody with $R^2 = 0.926$, hinting that this is a robust correlation that is not sensitive to the exact simulation conditions (Fig. 2.31b).

[1] The tail volumes were calculated by the Molinspiration Property Calculation Service for the tail group including the amino moiety.

[2] The insertion depth of the ligand was taken as the distance between the β-carbon atom of the ligand amino group and the amide nitrogen atom of V125 (V117 for β2AR) in the crystal structures of turkey β1AR in complexes with isoprenaline (PDB ID: 2Y03), dobutamine (PDB ID: 2Y00), carvedilol (PDB ID: 4AMJ), and cyanopindolol (PDB ID: 4BVN) as well as of human β2AR in complex with alprenolol (PDB ID: 3NYA).

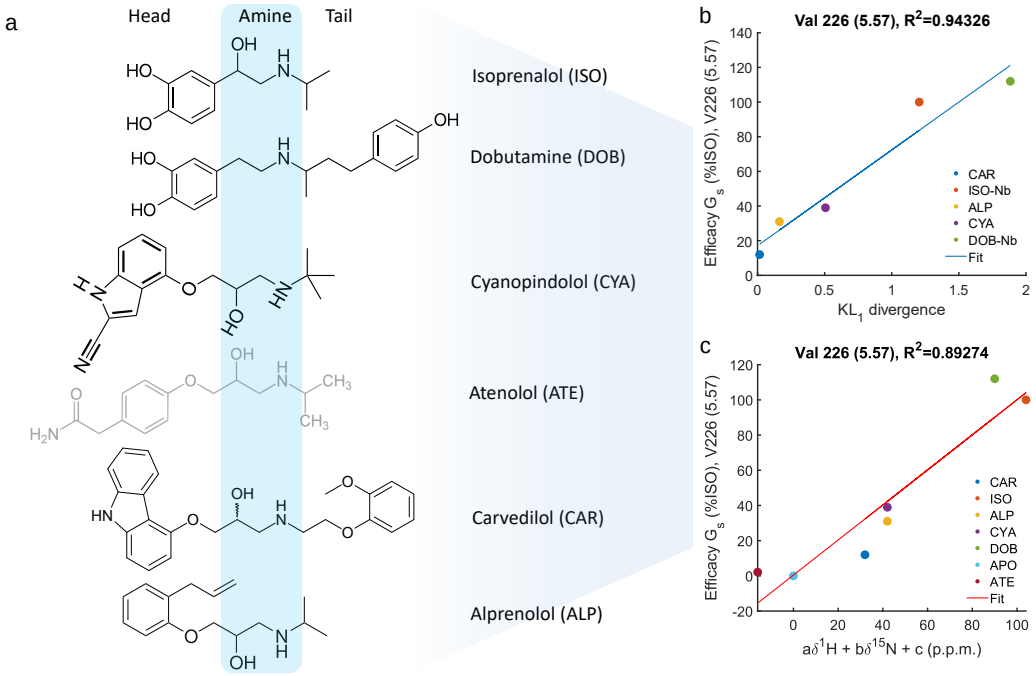


Figure 2.28: **Ligand chemical structure and fitting of data to experimental efficacies of said ligands: (a)** Chemical structures of the β1AR ligands used in this study. The ligand atenolol was not simulated, but is present in the reference study (202). **(b)** and **(c)** G-s efficacy measured as percentage of activation of the full agonist isoprenalol plotted against KL_1 for residue V226 (5.57) calculated with β1AT-TS apo as reference **(b)** and a best fit linear combination of the V226 (5.57) chemical shifts ($a = -515, b = -31.7, c = 8.4 \times 10^3$) **(c)**.

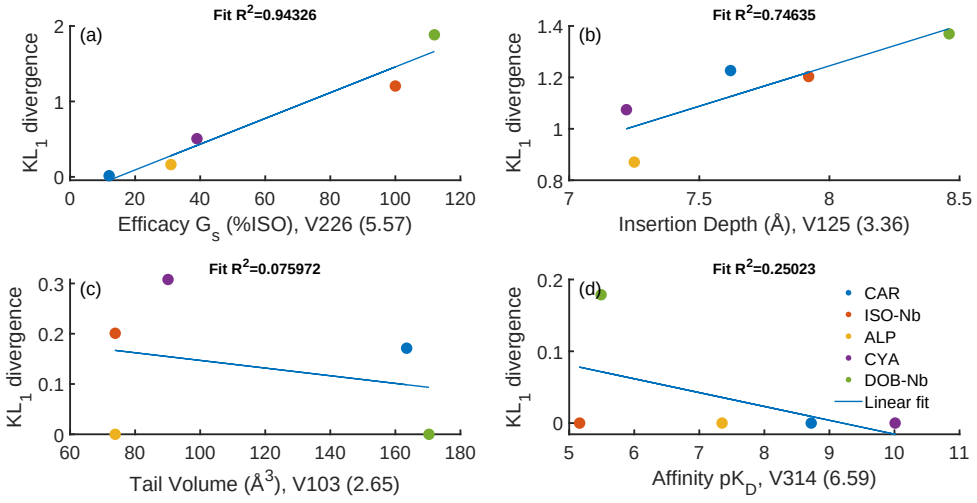


Figure 2.29: **Fitting of KL_1 with β1AR-TS apo as reference to different experimental observables correlated with NMR chemical shifts: (a)** efficacy of ligands for Gs signalling pathway, **(b)** ligand insertion depth, **(c)** ligand tail volume, and **(d)** ligand affinity.

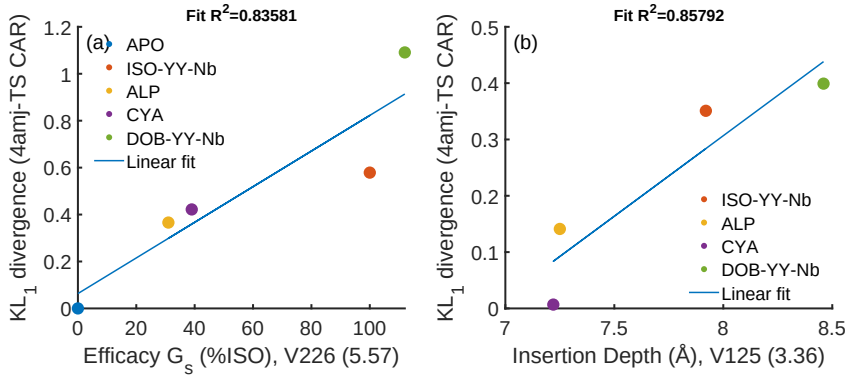


Figure 2.30: **Robustness of correlation to change of KL_1 reference state:** Fitting of KL_1 with antagonist bound β1AR-TS as reference to **(a)** efficacy of ligands for Gs signalling pathway and **(b)**) ligand insertion depth.

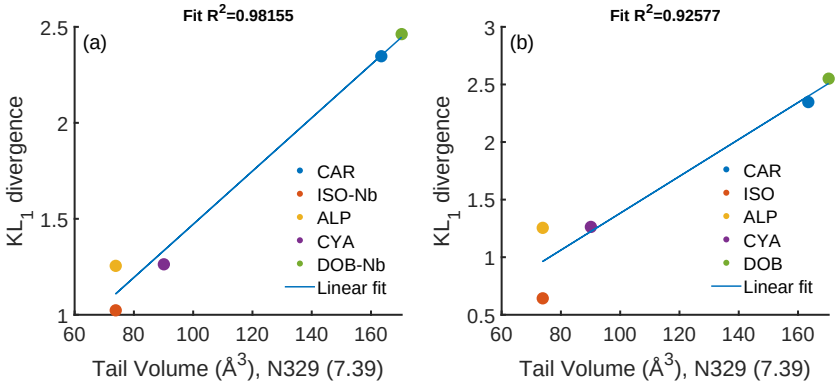


Figure 2.31: **Fitting of KL_1 of residue N329 (7.39) with β1AR-TS apo as reference to tail volume of simulated ligand: (a)** correlation using nanobody bound simulations for agonist ligands and **(b)** correlation using β1AT-TS without any bound intracellular binding partner.

Comparison of ligand efficacy with KL-divergence at probed valine sites The observed correlation between KL_1 and chemical shifts reported at V226 (5.57) with ligand efficacy hints that this residue contains dynamical information pertaining to the agonism of the ligands. V5.57 is one position away from the conserved Y5.58, which, along with Y7.53, are necessary for binding of G-protein-mimicking nanobody and thus receptor activation (202; 63). We probed KL_1 at the different valine sites previously reported, and then attempted to correlate KL_1 with ligand efficacy at sites other than V226 (5.57). Figures 2.32 and 2.33 show the correlation plots of reported efficacy in Isogai et al. (202) with backbone KL_1 and backbone+sidechain KL_1 respectively. Interestingly, V298 (6.43), which is one site away from the conserved F6.44 of the PIF motif in class A GPCRs, shows excellent correlation with ligand efficacy. Other noted valine sites that correlate well are V172(4.56) and V230(5.61). These findings hint that the dynamical information pertaining to ligand efficacy can be found at certain conserved sites and selected valine sites.

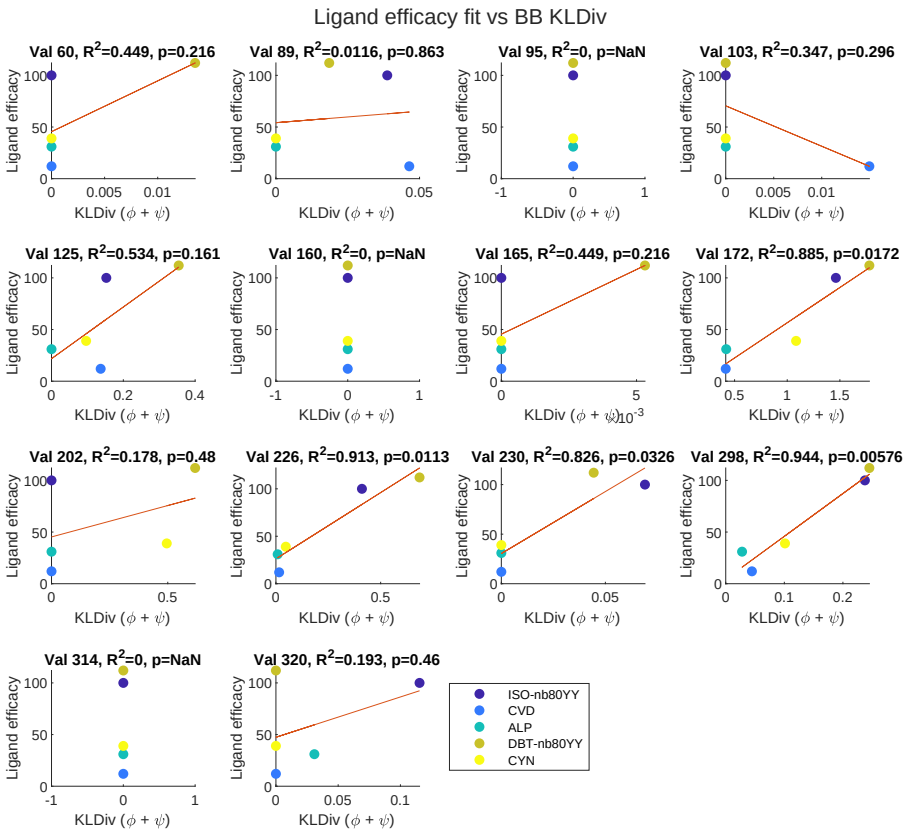


Figure 2.32: **β1AR simulations fitting of backbone KL_1 for valine residues using β1AR-TS apo as reference to ligand efficacies reported in Isogai et al. (202):** correlation R^2 and p-values are reported in subplot titles. NaNs mean that all KL_1 at a given residues are zero.

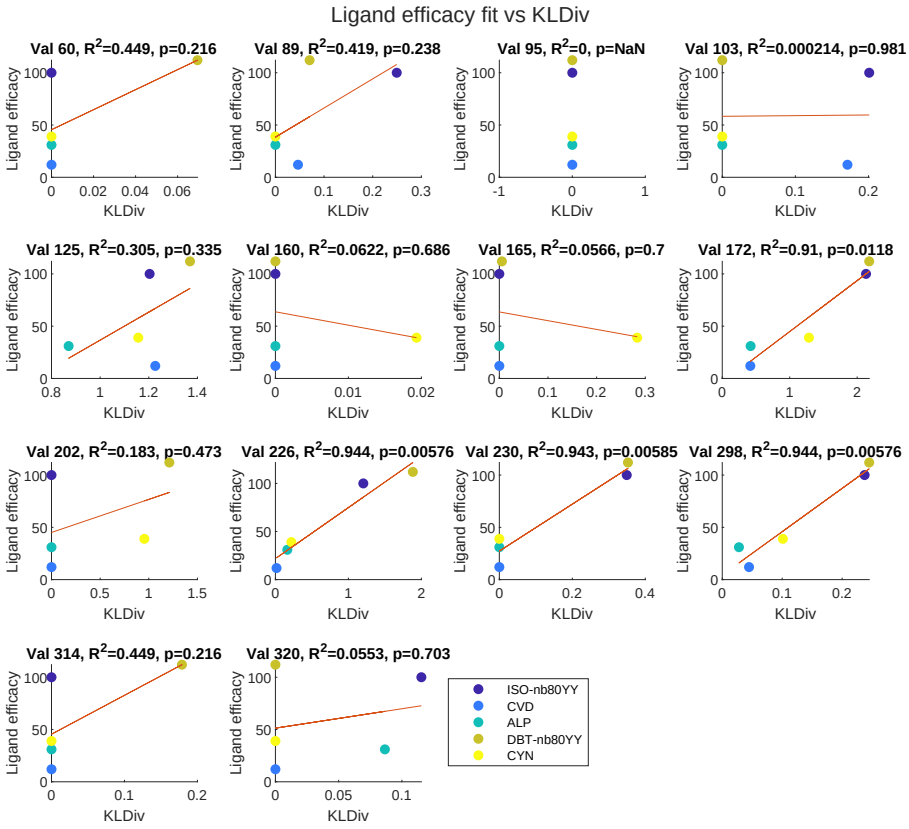


Figure 2.33: **β1AR simulations fitting of backbone KL_1 for valine residues using β1AR-TS apo as reference to ligand efficacies reported in Isogai et al. (202):** correlation R^2 and p-values are reported in subplot titles. NaNs mean that all KL_1 at a given residues are zero.

Comprehensive comparison with valine NMR chemical shifts To take the analysis further, we investigated the correlation between KL_1 and chemical shifts of valine residues reported in Grahl et al. (63). To that end, we fit a linear function $a\delta^1 H + b\delta^{15} N + c$ to KL_1 with β1AR-TS apo as reference state for the set of TS simulations (Fig. 2.34) and YY simulations (Fig. 2.35) separately. As mentioned before, we do not expect KL_1 to reproduce the chemical shifts, as they are extracted from simulations performed at an insufficient level of theory (using molecular mechanics forcefields). It would nonetheless be instructive to see how far we can find correspondence between the experimental and computational sides. For each of the variants tested (TS and YY), a few of the valines show significant or (bordering on significance, assuming $\alpha = 0.05$) correlations with the NMR fits. Note that correlations are meaningless of the KL_1 value is small (less than ≈ 0.2 in this case). For TS simulations, valine positions 103 and 202 show large enough KL_1 and significant correlations, while positions 172 and 202 are highlighted for the YY variants, with position 125 bordering on significance. One observation is that the correlation observed before for V226 (5.57), which is indiative of Gs efficacy is not observed here. The main reason is absence of nanobody-agonist bound simulations in the comparison performed here. In short, while there a few positions with significant correlations, there is no one to one correspondence between KL_1 and valine 1H and 15N chemical shifts for β1AR.

2.6.3 Discussion

Over the last sections, we showed that KL_1 of selected valine residues correlated well with ligand efficacy along the Gs pathway, ligand insertion depth, and had limited correlation with NMR chemical shifts. In addition, we report a ligand binding residue where KL_1 correlates well with tail volume (N7.39). we have also shown that some of the correlations are robust to change of reference states (Fig. 2.30) and correlations regarding ligand volume are present independent of the presence of an intracellular nanobody in the simulations (Fig. 2.31). At the same time, we have seen that correlation with Gs efficacy is sensitive to the presence of the bound nanobody, since efficacy of a given ligand is a fundamentally allosteric property in GPCRs (allosteric regulation). All this demonstrates that KL_1 is a suitable tool to quantify effects of perturbations in simulations and to probe allosteric effects of said perturbations.

In addition, the current KL-divergence formulation allows decomposition of KL_1 into backbone and sidechain contributions. If we consider only the backbone contribution (φ, ψ) to KL_1, we still observe correlations with ligand efficacy toward Gs and insertion depth, albeit with a reduced KL_1 signal (Fig. A.8). However, correlation of residue N329 (7.39) to ligand tail volume is worse and loses significance (Fig. A.9). This decomposition shows that the relevant signal of N7.39 for the sake of the aforementioned comparison is mainly its sidechain contribution.

When comparing to chemical shifts, it is crucial to remember that 15N chemical shifts have contributions from hydrogen bonding, backbone conformation, sidechain orientation, and

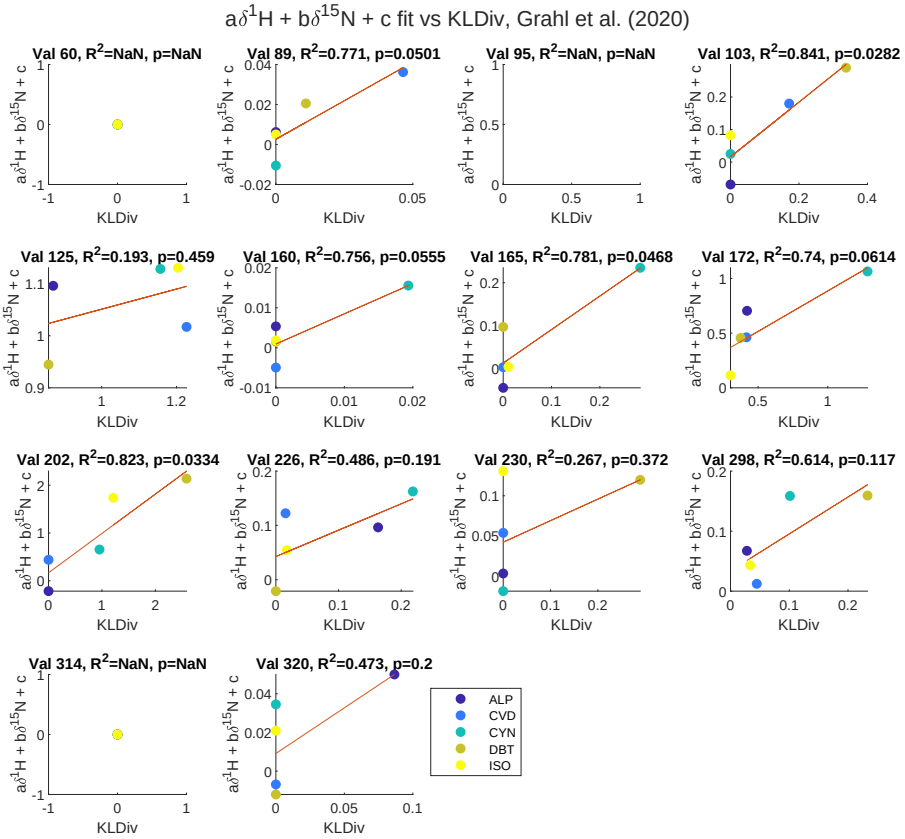


Figure 2.34: **Thermostabilized (TS) β1AR simulations fitting of KL_1 for valine residues using β1AR-TS apo as reference to NMR chemical shifts ($a\delta^1H + b\delta^{15}N + c$) from Grahl et al. (63):** correlation R^2 and p-values are reported in subplot titles. NaNs mean that all KL_1 at a given residues are zero.

neighboring residues (285). A future continuation of this work in the direction of completing comparison with 1H-15N NMR ought to study effects of not just backbone conformation and sidechain orientation (which are already captured by (ψ, φ, χ_x) in the current KL-divergence formulation) but also hydrogen bond lengths. The main challenge to adding hydrogen bond length is the level of theory currently used in large-scale molecular simulations, which lacks the ability to accurately model sub angstrom distance shifts that contribute to chemical shifts (286).

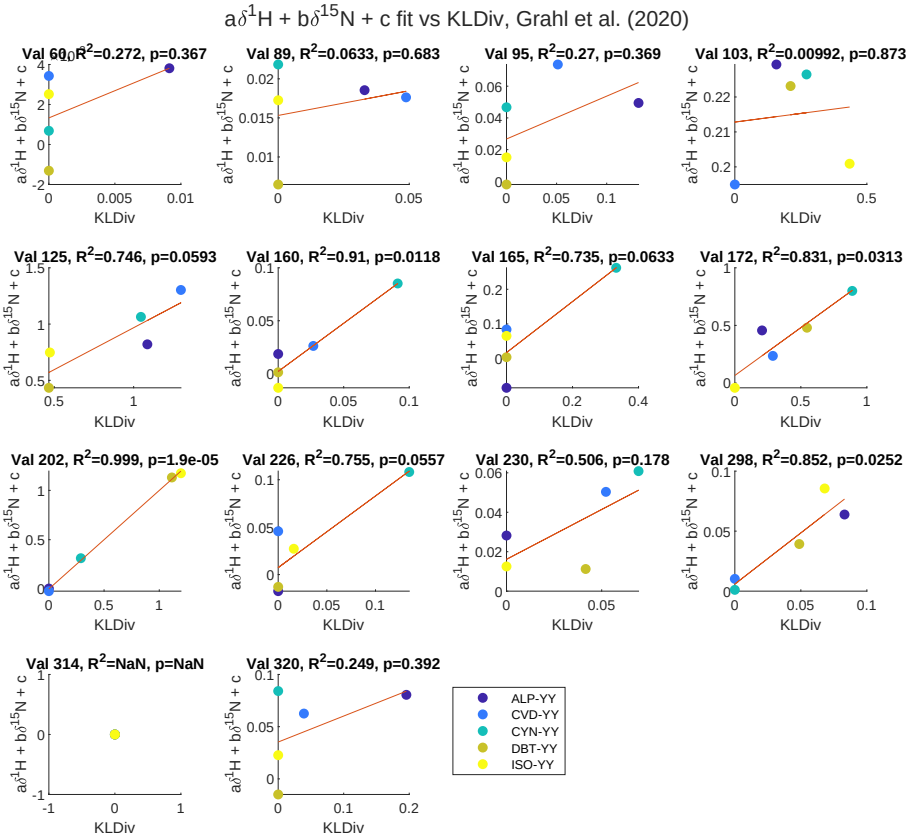


Figure 2.35: **A5.58Y-L7.53Y (YY) β1AR simulations fitting of KL_1 for valine residues using β1AR-TS apo as reference to NMR chemical shifts ($a\delta^1H + b\delta^{15}N + c$) from Grahl et al. (63):** correlation R^2 and p-values are reported in subplot titles. NaNs mean that all KL_1 at a given residues are zero.

2.6.4 Methods

Structure preparation

Systems simulated for wild turkey β1AR are shown in Tab. 2.1, where TS is the thermostabilized variant as described in Isogai et al .(202) and YY is the A5.58Y-L7.53Y reverted double mutant starting from TS. Structure preparation started by rebuilding missing loops from the reference structures using RosettaRemodel (170) after constrained relaxation using Rosetta (168). The loops were built to match the sequence reported in the reference (202). For every structure, 2000 decoys with 5 trajectories each were generated. The lowest scoring decoys were used for all subsequent analyses.

Ref. PDB	Variant	In complex	Ligand	Ligand type	Comment
4AMJ	TS	None	None	NA	Apo state
6H7J	YY	Nb80	Isoprenalol	Agonist	
2Y03	TS	None	Isoprenalol	Agonist	
2Y03	YY	None	Isoprenalol	Agonist	
6H7L	YY	Nb6B9	Dobutamine	Agonist	
2Y01	TS	None	Dobutamine	Agonist	
2Y01	YY	None	Dobutamine	Agonist	
6H7O	YY	Nb6B9	Cyanopindolol	Partial agonist	
4BVN	TS	None	Cyanopindolol	Partial agonist	
4BVN	TS	None	Cyanopindolol	Partial agonist	
4AMJ	TS	None	Cyanopindolol	Partial agonist	Ligand from 4BVN
4AMJ	YY	None	Carvedilol	Antagonist	
4AMJ	TS	None	Carvedilol	Antagonist	
4AMJ	YY	None	Alprenolol	Partial agonist	Ligand from 3NYA
4AMJ	TS	None	Alprenolol	Partial agonist	Ligand from 3NYA

Table 2.1: Systems simulated for wild turkey β1AR. TS is the thermostabilized variant as described in Isogai et al .(202). YY is the A5.58Y-L7.53Y reverted double mutant starting from TS. Ligands from different reference structures were overlayed after structural alignment and then relaxed with MD.

Wild turkey β1AR TS and YY models were generated using RosettaMembrane (171). The starting structures are listed in Tab. 2.1. Receptor and nanobody (when present) chains were kept from the starting structure during in-silico mutagenesis (ISM). Residues of interest were mutated and adjacent residues within 5 Å were subjected to alternating cycles of sidechain repacking and backbone relaxation through Rosetta's Monte Carlo-based energy minimization algorithm. 200 decoys were generated per design to ensure score convergence. The lowest scoring decoys were used for all subsequent analyses.

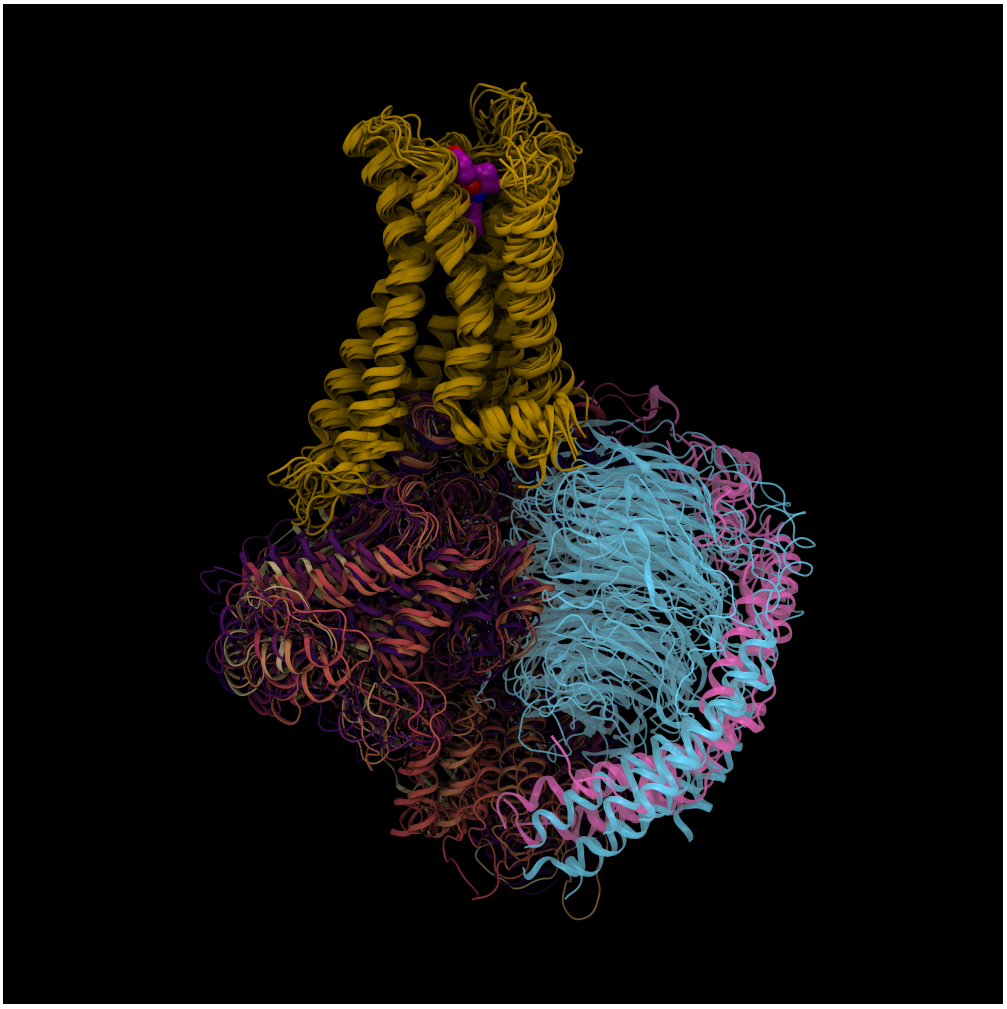
After ISM, ligands were replaced in the orthosteric binding site. Ligands from different reference structures were overlayed after structural alignment and then relaxed with MD. Ligand protonation state was predicted using Protoss (280).

Molecular dynamics simulations

The receptor-ligand-nanobody (when present) complexes were inserted in a regular hexagonal POPC lipid bilayer where the distance between parallel sides is 90Å solvated by layer of water above and below the bilayer (40 Å for systems without nanobody and 50 Å for systems with) with 0.15 M of Na^+ and Cl^- ions using CHARMM-GUI bilayer builder (287; 283; 281). Parameters for the ligands were generated using CGenFF (288). Simulations were performed with GROMCAS 2020.5 (289; 290) with CHARMM36 forcefield (291) in an NPT ensemble at 310K and 1 bar using a Nose-Hoover thermostat (with a period of temperature fluctuations of 1.0 ps) and Parrinello-Rahman barostat (with semi-isotropic coupling at a relaxation time of 5 ps) respectively. Equations of motion were integrated with a timestep of 2 fs using a leap-frog algorithm. Each system was energy minimized using a steepest descent algorithm for 5000 steps, and then equilibrated with the atoms of the ligand-receptor-nanobody (when present) complex and lipids restrained using a harmonic restraining force in 6 steps as shown in Tab. A.19. After constrained equilibration, 4 to 5 replicas of 400 ns were run for all systems except 4AMJ-TS-CYA, 4AMJ-TS-ALP, and 4AMJ-YY-ALP, for which 500 ns replicas were run (Tab. 2.1, those are the systems where the ligand was overlayed from a different structure). The first 50 to 125 ns of every simulation was discarded for equilibration of Cα and ligand RMSD, and the rest of the simulation was used for calculating statistics.

AlloDy calculation

The *KLDiv* module of AlloDy was used to calculate Kullback–Leibler divergences from the simulations. The calculations were performed as described in Sec. 2.5.2. Unless otherwise mentioned, the reference state used for the calculation was the *apo* state simulation, and both backbone and sidechain dihedrals were used for the calculation. When mentioned, the antagonist bound reference state refers to the 4AMJ-TS-CVD simulation (Tab. 2.1).



Artwork 3: *Dopamine D2 receptor (yellow) bound to bromocriptine interacting with Gαi (beige thru purple), Gβ (cyan) and Gγ (magenta). This simulation uses the structure from 6VMS (132). The structure was determined using a D2 receptor stabilized in the active state designed previously in the lab (16).*

Applications Part II

3 Computational rewiring of allosteric pathways reprograms GPCR selective responses to ligands in dopamine receptors

"What I cannot create, I do not understand"
— Richard Feynman

Author contributions: M.H. contributed to the study design and performed MD simulations, data analysis, interpretation, and figure design. M.H. wrote the chapter. This work was done in collaboration with Dr. Daniel Keri and Aurélien Oggier.

3.1 Introduction

The importance of G-protein-coupled receptors (GPCRs) as valuable drug targets underscores the necessity to comprehend and predict how ligand binding to a GPCR triggers specific signaling responses, particularly within GPCR families with divergent receptor roles. Allostery enables ligand-induced changes in protein structure and dynamics to be efficiently transmitted to distant sites and is a widespread regulation mechanism of protein function (see (220; 27; 292) and Ch. 1). Owing to the lack of high-resolution dynamical measurements on GPCRs, how allostery is encoded into receptor sequence, structure and dynamics is not well understood. Computational studies using sequence co-evolution inference or molecular dynamics simulations can detect networks of functionally and dynamically coupled receptor residues that may provide efficient communication pathways(277; 293; 25; 294) . However, how distinct ligands engage these networks to elicit precise and selective signaling responses remains elusive.

In this chapter, we employ computational protein design techniques previously developed in the lab (16) coupled with dynamic analysis from AlloDy to explore the allosteric functions

of GPCRs, unveiling mechanistic relationships between agonist ligand chemistry, receptor sequence, structure, dynamics, and allosteric signaling across the dopamine receptor family using dopamine D1 and D2 receptors as a model systems. The designed receptor variants exhibited correlated responses to structurally similar ligand agonists and displayed selective reactions across members of the dopamine receptor family. Computational analysis identified distinct topologies of allosteric signal transmission pathways for various agonist-receptor pairs, perturbed differently by the designs. By harnessing these insights, we rewired ligand-receptor specific pathways and engineer receptors with highly selective ligand responses. Our study suggests that diverse ligand agonists activate a given signaling effector through specific "allosteric activator" moieties, which engage partially independent signal transmission networks in GPCRs. These allosteric activators have evolved to optimize either binding affinity or signaling efficacy based on the receptor's function. Additionally, we investigated the impact of allosteric modulators and demonstrated the ability of receptor design to emulate the effects of positive modulators. These results furnish a mechanistic framework for comprehending and predicting the influence of sequence polymorphism on receptor pharmacology, providing valuable insights for selective drug design and rational receptor engineering for both fundamental research and therapeutic applications.

3.2 Results

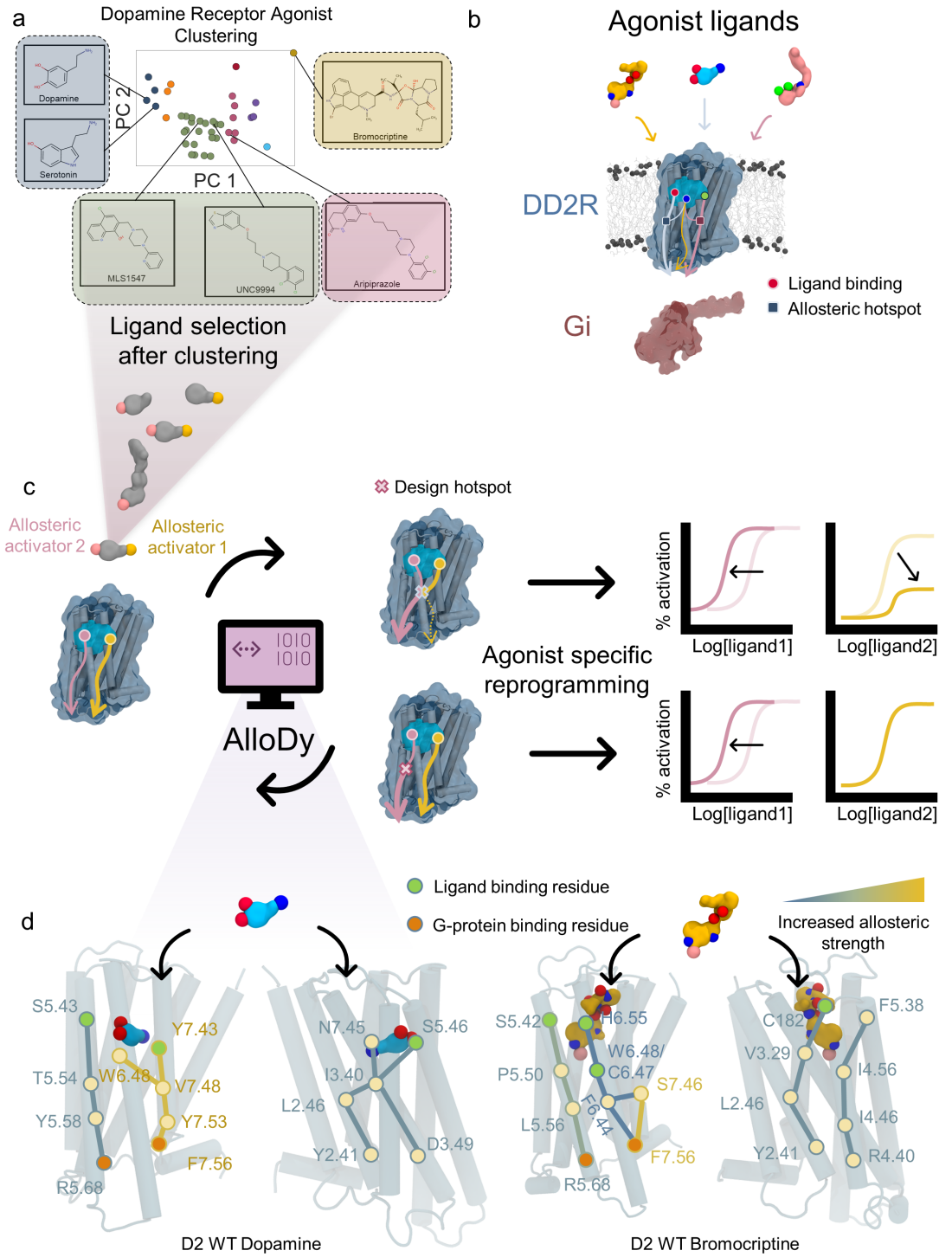
In a previous study, we had identified a class of allosteric sites ("allosteric propagators") that connect highly conserved microswitches into fully wired allosteric pathways. We were able to fine-tune GPCR signaling responses through novel allosteric "propagator" microswitches designed on several TMHs, which suggested the existence of multiple allosteric signal transduction pathways running through the TM region of the receptor (16). The designed microswitches at propagator sites enhanced the sensitivity to both dopamine and serotonin, implying that receptor responses to both ligands may involve the same path. Since signal transduction pathways connect the extracellular ligand to the intracellular effector binding sites, we reasoned that ligands with similar structure should engage similar paths through overlapping contacts with the receptor. Conversely, structurally distinct agonists that bind to the receptor through different "allosteric activator" chemical groups should involve alternative pathways and therefore be sensitive to different allosteric "propagators".

3.2.1 Ligand functional effects correlate with ligand clustering in dopamine D2 receptors

To test this hypothesis, we sought to explore the mechanistic relationships between agonist ligand chemistry and receptor signaling through the atomic-resolution mapping of allosteric pathways and quantification of signal transductions in ligand-receptor systems (Fig. 3.1). We took a multi-disciplinary approach combining ligand structure clustering for mapping chemical space, molecular dynamics simulations (MD) for allosteric pathway discovery and

perturbation response quantification, computational protein design for reprogramming allosteric responses and a battery of cell signaling assays for measurement of receptor responses to multiple ligands (Fig. 3.1c). If we understand how ligands activate a receptor through engagement of specific allosteric pathways, we should be able to rationally rewire these paths and design novel selective ligand-GPCR responses (Fig. 3.1c).

We selected the dopamine D2 receptor as a system for the first part of this study because it performs critical neurological functions (295; 296), has been structurally characterized in several signaling states and is regulated by numerous partial and full ligand agonists. We first analyzed the structural similarity of a comprehensive set of 39 characterized D2 agonists using standard structural clustering approaches and identified 4 major clusters (Fig. 3.1b, **methods**). Consistent with our expectations, serotonin (SE) and dopamine (DA) adopt very similar chemical structure and therefore belong to the same cluster. D2 agonists from other clusters are often much larger than DA and SE, populate distinct chemical subspaces and can presumably contact the receptor through multiple additional sites.



(Caption on next page.)

Figure 3.1: **Outline of computational and experimental workflow: a** Principal component analysis plot of clustered dopamine receptor agonists. Highlighted agonists were chosen for experimental testing and are colored based on cluster. **b** Different agonists engage different set of binding residues that each contribute to allosteric pathways connecting the protein structure. **c** Computational and experimental workflow performed in this work. After selection of agonists from clustering, molecular dynamics (MD) simulations of ligand-bound receptors are input to our software AlloDy that predicts ligand-specific allosteric pathways. Ligand specific allosteric hotspots are targeted for mutagenesis. Chosen mutants are tested in live cells to confirm ligand selective effect. **d** Highlights of allosteric pathways extracted for dopamine bound and bromocriptine bound dopamine D2 receptor.

To investigate the relationship between ligand chemical space and long-range allosteric signaling responses, we first assessed how several allosteric propagator microswitches previously-designed in the TM core of D2 in addition to newly designed ligand specific variants modulate the receptor response to structurally-distinct ligands (Fig. 3.2). The previously designed variants (T5.54M, F6.44I, C6.47L, and T5.54M-C6.47L) were designed with parameters for DA-Gi pathway, without any consideration for ligand selectivity (16), while the additional designs reported in this work considered two sets of parameters, one for DA and another for BRC. The parameters involved are the modeled structures that are used as a starting point for in-silico mutagenesis (ISM) and the allosteric hotspots used to approximate dynamics from DCCM. Because these sites are far away from the ligand binding site, they can probe long-range interactions connecting ligands and allosteric pathways running though the receptor structure. We selected representative members of the 4 largest ligand clusters and measured *in vitro* the D2-mediated activation of the G protein Gi2 upon ligand stimulus using HEK reporter cell lines stably expressing a TRP channel (Fig. 3.2a). Nine designed D2 receptors incorporating distinct allosteric propagator microswitches at TMHs 3, 4, 5 and 6 (Fig. 3.2d and e) were transiently expressed in stable HEK cells and incubated with the following 6 agonist ligands: the full agonists dopamine (DA), bromocriptine (BRC), and the partial agonists serotonin (SE), aripiprazole (AR), MLS1547 (MLS) and UNC9994 (UNC). Dose titrations revealed very distinct effects of the designed microswitches on the assayed ligands. The large (i.e. from 1 to 2 orders of magnitude) increases in potency for dopamine and potency and efficacy for serotonin were not observed for the other ligands (Fig. 3.2b and d). While the designed microswitches still behaved as gain of function for the partial agonists, smaller increases in efficacy and potency were observed except for mutations I4.46N and L3.41G (residues are numbered according to the Ballesteros–Weinstein numbering scheme (1)). I4.46N is dead for all surveyed ligands except DA, for which it is a loss of function, while L3.41G breaks the pattern by being a gain of efficacy for BRC and a loss of function for the partial agonists. L3.41H, on the other hand, displayed strong gain of function for DA/SE cluster, a slight gain of function for MLS/UNC cluster, and no effect on the others. F6.44I displayed gain of function for all ligands except BRC, and F6.44M signals with all ligands except for BRC, which loses all signaling. The double mutant T5.54M-C6.47L appears to have an additive effect for the DA/SE cluster, a mixed effect on MLS/UNC/AR, and a loss on BRC when observing from the lens of potency shift (Fig. 3.2d).

Interestingly, specific designed effects tend to be similar for ligands from the same cluster and could be classified into 4 distinct classes; i.e. DA/SE, UNC/MLS, AR, BRC (Fig. 3.2f). The results suggest that, while each allosteric propagator microswitch has the potential to modulate ligand signaling differently, the effects correlate with how ligands activate the receptor and trigger allosteric pathways. Importantly, the allosteric coupling profiles could not be explained by the differences in efficacy of the ligands for the WT receptor as DA and BRC are both full agonists and SE is a very weak agonist for D2.

3.2.2 Molecular exploration of WT dopamine D2 receptor behavior using AlloDy

These observations suggested the existence of several allosteric pathways running through distinct TMH interfaces that would be potentiated differently by the two full agonists to activate Gi, hence prompting us to investigate the behavior of the DA and BRC-bound D2 receptor using molecular dynamics over 2 μs simulation time. Consistent with large differences in ligand size, BRC displayed more contacts than DA, especially on TMH 3, ECL2, TMH 6, and TMH 7 (Fig. 3.3). In addition, BRC-D2 structures displayed significantly lower RMSF than DA-D2 structures especially in regions directly in contact with the ligand (Fig. A.10).

To assess how the differences in ligand contacts, receptor rigidity, and amino acid substitutions affect the allosteric signaling pathways running through the receptors, we used AlloDy (Ch. 2, (297)). Allosteric signals are predicted to be best propagated by residues that exchange the most information (298), so we assign residues allosteric scores σ_m proportional to the number of pathways going through them weighted by the pathway endpoint MI. We also used AlloDy to quantify Kullback-Leibler divergences (KL) between target and reference simulations. Examples of perturbations could be amino-acid substitutions or different ligands binding the receptor. Perturbation responses are divided into two classes, those close to the perturbation site (direct effect) and others far from the perturbation site (allosteric effect). During the analysis process, AlloDy also performs principal component analysis (PCA) of ligand binding poses and extracts consensus class A GPCR activation features such as interhelical (i.e., TM3-TM6 and TM3-TM7) distances on the intracellular side of the receptor and the RMSD for the NPxxY motif (Ch. 2).

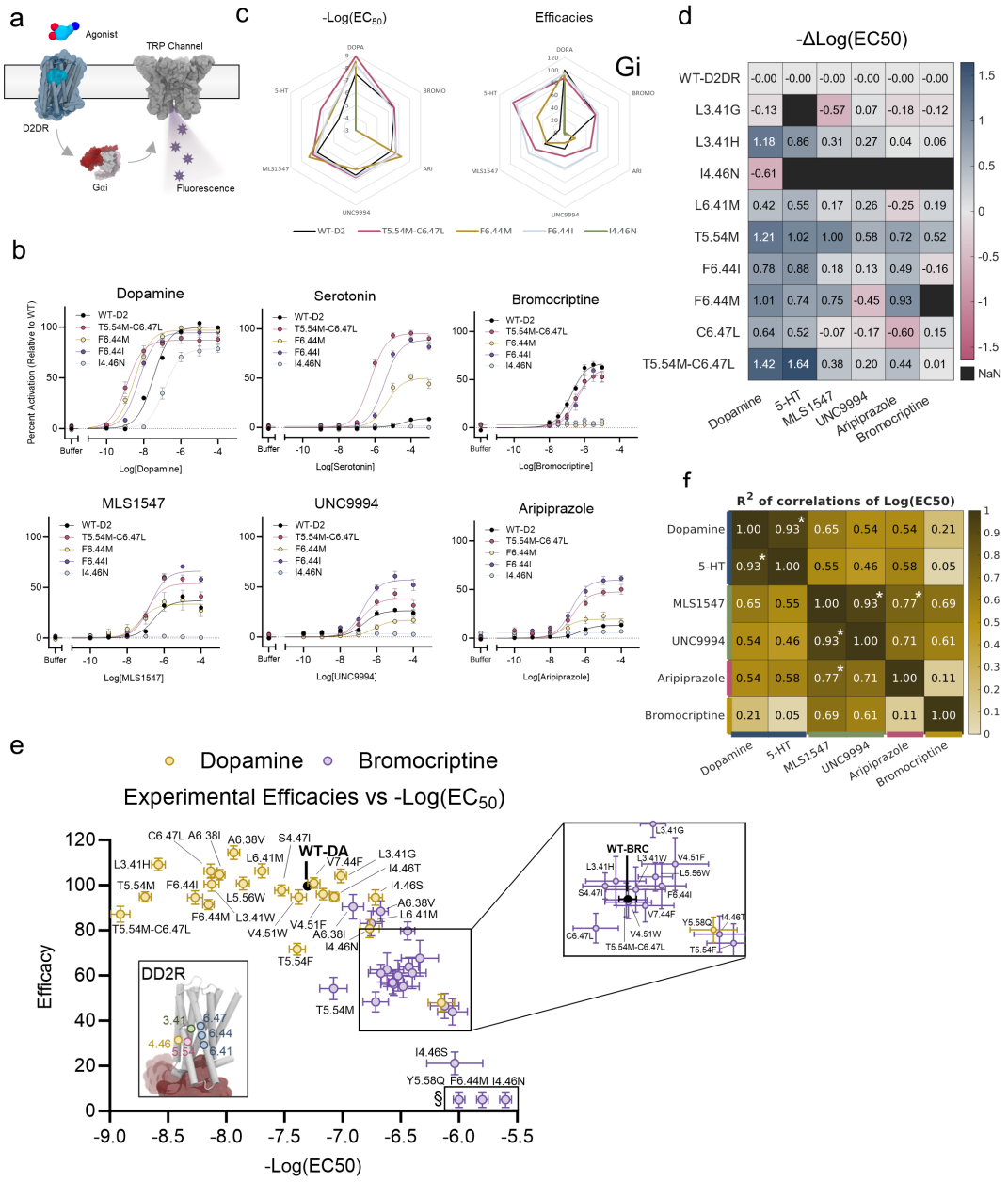
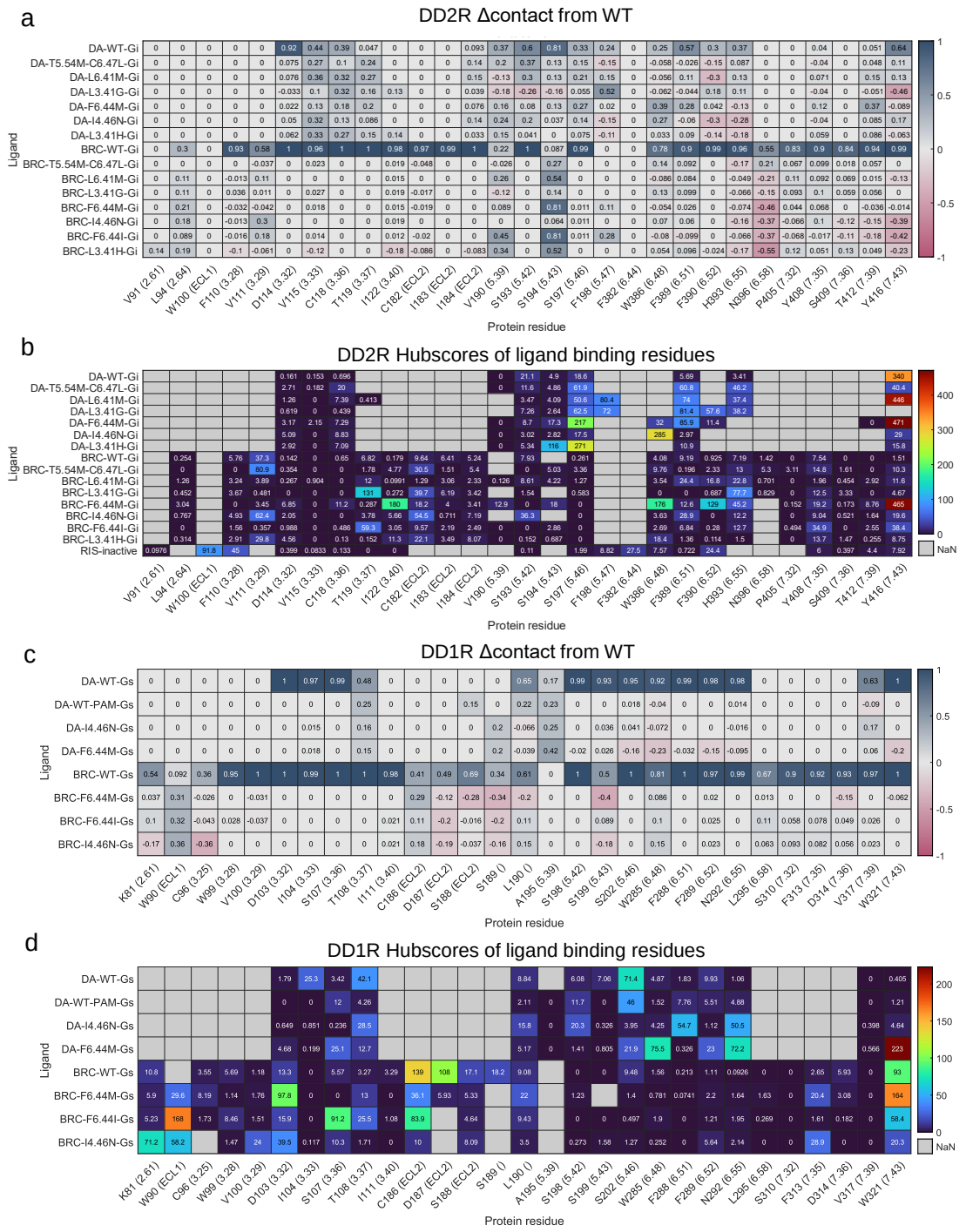


Figure 3.2: **Functional output, classes of mutations and of allosteric effects: a** Schematic of the TRP channel assay used to produce the data shown in this figure. **b** Dose response curves of selected mutations for the six tested ligands show a range of ligand selectivity. **c** Same data represented as radar plots. **d** $-\Delta\log(EC50)$ for the nine tested mutants for all tested ligands. NaNs represent a dead receptor that does not signal. **e** Normalized efficacies and $-\log(EC50)$ values for D2 designs in TRP channel assays measuring Gi response. Positions of the ligand-agnostic and ligand-selective designs on the dopamine D2 receptor structure are shown in the inset. **f** Pearson correlations of EC50 shifts between agonists for tested dopamine D2 receptor designs. *p < 0.05. Rows including NaNs were not included in the correlation calculation. Colored bars represent clustering from Fig. 3.1a.



(Caption on next page.)

Figure 3.3: **Ligand binding residues and their allosteric hubscores for DD1R and DD2R: a** Difference of ligand contact persistence between mutant and WT for simulated variants of DD2R. WT values represent fraction of simulation in which ligand is in contact with a given residue. Contact is defined via a dual cutoff scheme where two heavy atoms (non-hydrogen) come in contact when they are within 3.5 Å and leave contact when the atoms are further than 5 Å apart **b** Allosteric hubscores σ_m of ligand binding residues for DA-bound, BRC-bound, and RIS-bound states. NaNs represent residues not having significant contact with the ligand. Significant cutoff is taken to be 30% of all simulated time. **c** Same as **a** but for DD1R. **d** Same as **c** but for DD1R.

Using the aforementioned features extracted using AlloDy, we analyzed WT D2 receptor bound to different ligands and in different states: DA-D2WT-GiH5; BRC-D2WT-GiH5; and risperidone(RIS)-D2WT, where the DA and BRC bound simulations start from the active state (6VMS) and RIS bound D2 simulations start from the inactive state (6CM4). GiH5 represents the H5 C-terminal helix of the G-protein which interfaces with the IC cleft of a GPCR. Comparing active to inactive state simulations indicates markers of activation in DD2R, while comparing the two active ligand bound simulations informs us on the ligand sepcific activation mechanisms.

Starting from the inactive state simulations, RIS kept an almost perfect unbroken contact with I3.40, F5.47, F6.44, and W6.48, where the contacts with F5.47 and F6.44 are unique to RIS in our simulation set, compared to DA and BRC (Fig. 3.3). In other DD2R inactive state structures, ligands haloperidol (299) and spiperone (300) also contact F6.44 in inactive state D2 structures, which makes sense since one of the ways the antagonists/inverse agonist are blocking activation is through the PIF motif. RIS also formed contacts with W100(23.50) for more than half of the total simulation time (RIS forms a T-stack in the crystal structure, although this interaction is not maintained during the simulation). On the other hand, RIS lacks contact with activating residues on TM5, S5.42 and S5.43.

As for the activation state, the markers of activation are taken to be TM3-6 and TM3-7 distances defined between Cαs of residues 3.50-6.30 and 3.50-7.53 and RMSD of NPxxY motif to the inactive reference, 6CM4. The inactive state simulations sampled a single well defined well that includes the inactive state reference, while DA-D2WT-GiH5 displayed a major and a minor population that are both far from the inactive reference (Fig. 3.4a).

Looking more closely at differences between DA and RIS-bound simulations at the degree of freedom level, we observe significant divergences beyond the expected ones at the ligand binding site. On the ECL side, ECL1 exhibits backbone and sidechain divergences at positions W100(23.50) and F102(23.52), while ECL2 shows mostly backbone divergences (expected with the majorly different conformation in ECL2 between the two ensembles), with the highest divergence being on the C182(45.50). Other expected highlighted features are the conserved motifs that are hallmark markers of class A GPCR activation, including I3.40/F6.44 from the PIF motif, W6.48, R3.50 from the DRY motif, and Y7.53, which all also display higher MI in the

active state. The backbone KL_1 signal at the NPxxY motif represents the kink observed in active state GPCRs, and this is combined with TM7 having a significantly larger MI contribution in DA active state simulations. Other significant divergences are seen on Y2.41, Y5.62, and E6.30. C3.37 and L3.43 are mostly sidechain changes that corroborate with I3.40 changes. Y5.58 stabilizes the active state in an interaction with Y7.53 and has been shown to be essential for tertiary complex formation in β1AR (202).

Looking at the ligand binding, activation landscapes, MI, and divergences altogether, we get a rigorous picture of markers of activation and ensemble differences between active and inactive ensemble.

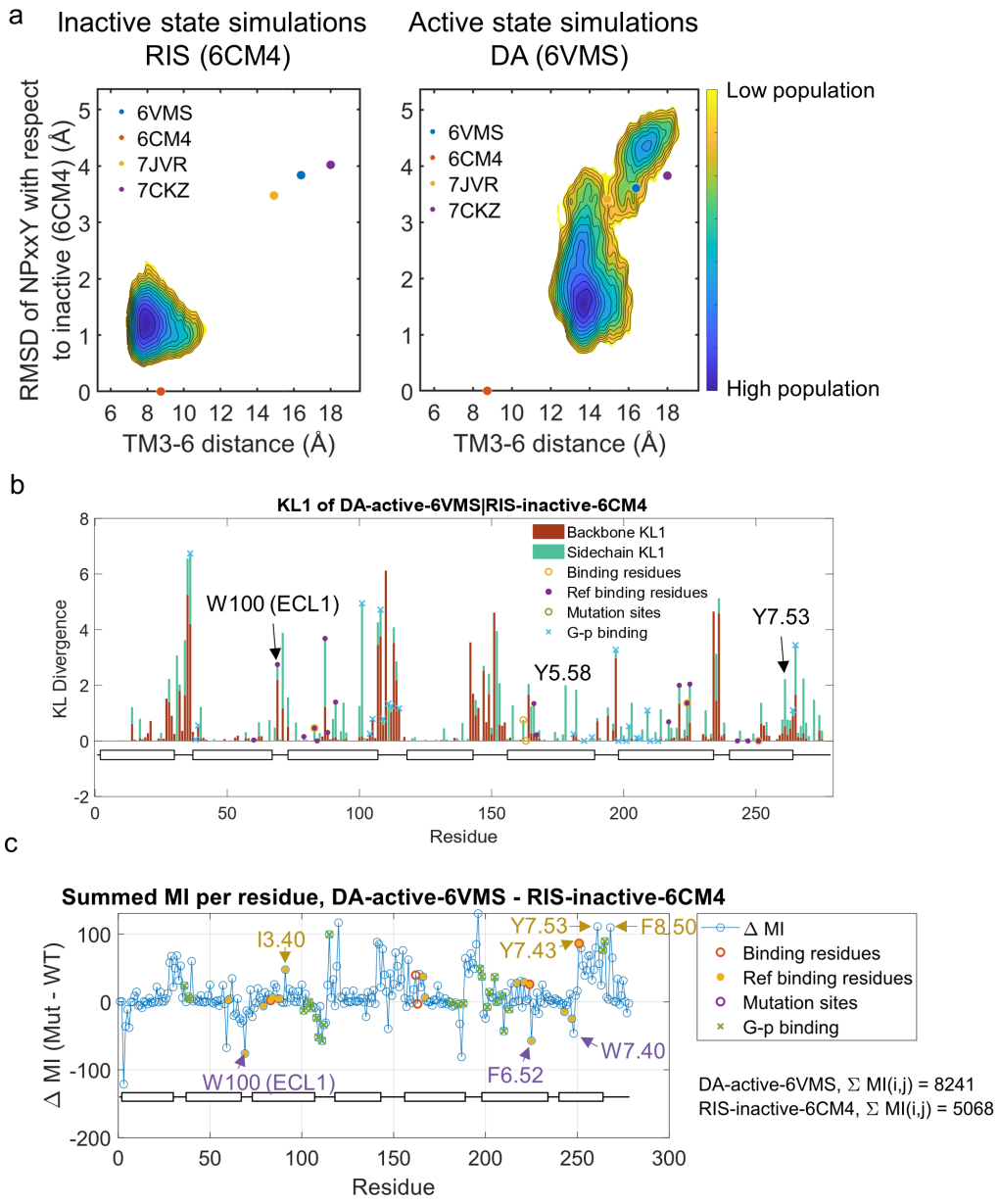


Figure 3.4: **Inactive state RIS-bound and active state DA-bound DD2R comparison: a** Activation landscapes for inactive state RIS-D2WT simulations (left) and active state DA-D2WT-GiH5 simulations (right). TM3-6 distance is measured between residues R3.50 and E6.30. **b** KL-divergence with DA-D2WT-GiH5 as target ensemble and RIS-D2WT as reference ensemble. Brown and green bars represent backbone and sidechain contributions to the KL respectively. Ligand binding residues in the target ensemble are in yellow disks and those in the reference ensemble are in purple circles. **c** Difference of residue summed mutual information (MI) between target and reference ensembles. Ligand binding residues and G-protein binding residues are marked on the figure. Rectangles represent the seven TM helices of GPCRs.

Moving on to the ligand specific differences in active state simulations of WT D2 systems (DA and BRC-bound). Major allosteric pathways connecting extracellular to the intracellular regions are initiated from a set of divergent ligand binding residues. With the exception of common allosteric contacts with conserved serines on TMH 5 (S5.42/6), DA initiated pathways from residues on TMH 6 (F6.51) and TMH 7 (Y7.43), while BRC initiated pathways from TMH 6 (H6.55 and W6.48), TMH 7 (Y7.35) and ECL2 (C182). Both systems have a pathway running through TMH 5 connecting to the G-protein binding interface. The major differences lie in DA-bound D2 having strong cross TMH connectivity between TMHs 6 and 7, which is weaker in the case of BRC (Fig. 3.1d), and in having stronger allosteric pathways, as is signified by higher overall MI and MI of conserved class A residues of DA-D2WT over BRC-D2WT (Fig. 3.5b). KL-divergence analysis between the two simulations shows obvious differences at the binding site (residues F6.52 and Y7.35 show high signal), and others such as I3.40, Y5.48, ECL3, and W7.40. Interestingly, only very slight changes are seen in the protein core or G-protein binding region, with KL<1 (Fig. 3.5a).

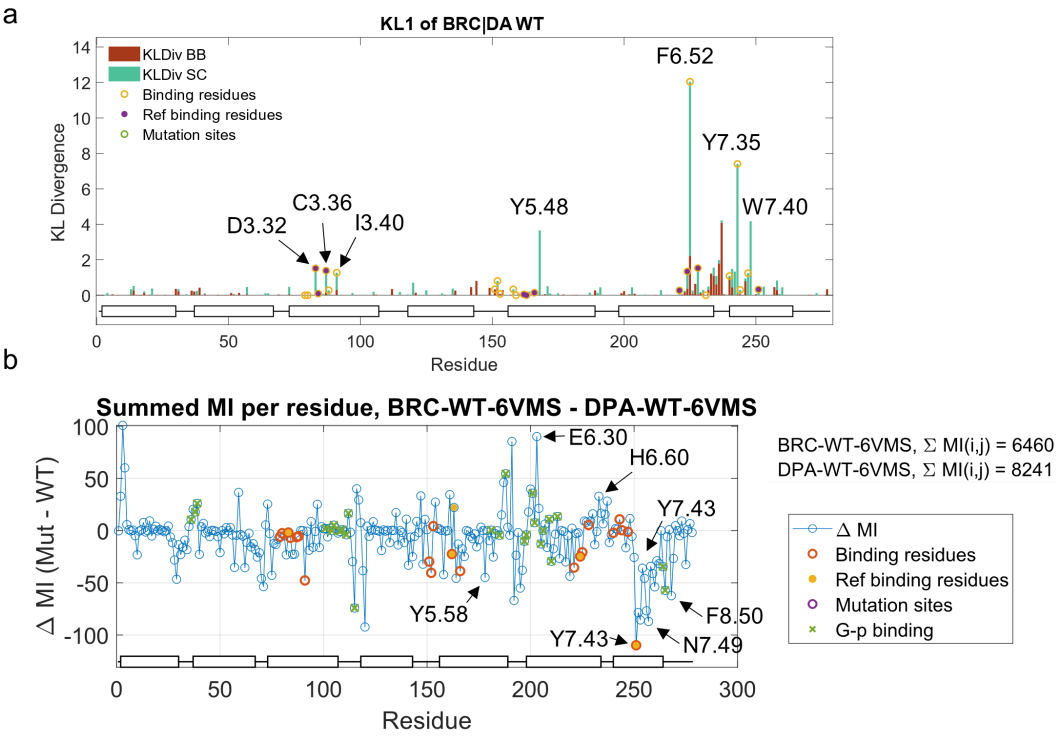


Figure 3.5: **KL-divergence and MI differences for DA and BRC-bound active state D2 simulations: a** KL-divergence with BRC-D2WT-GiH5 as target ensemble and DA-D2WT-GiH5 as reference ensemble. Brown and green bars represent backbone and sidechain contributions to the KL respectively. Ligand binding residues in the target ensemble are in yellow disks and those in the reference ensemble are in purple circles **b** Difference of residue summed mutual information (MI) between target and reference ensembles. Ligand binding residues and G-protein binding residues are marked on the figure. Rectangles represent the seven TM helices of GPCRs.

3.2.3 Allosteric design using distinct ligand specific pathways in dopamine D2 receptors

If DA and BRC exploit distinct pathways to activate the same signaling effector, we should be able to create D2 receptors with high ligand selectivity by reprogramming long-range communication along ligand-specific allosteric channels. We first identified the main allosteric sites through which DA or BRC-specific paths were predicted to run through and defined DA or BRC-selective allosteric hubs (Fig. 3.1d).

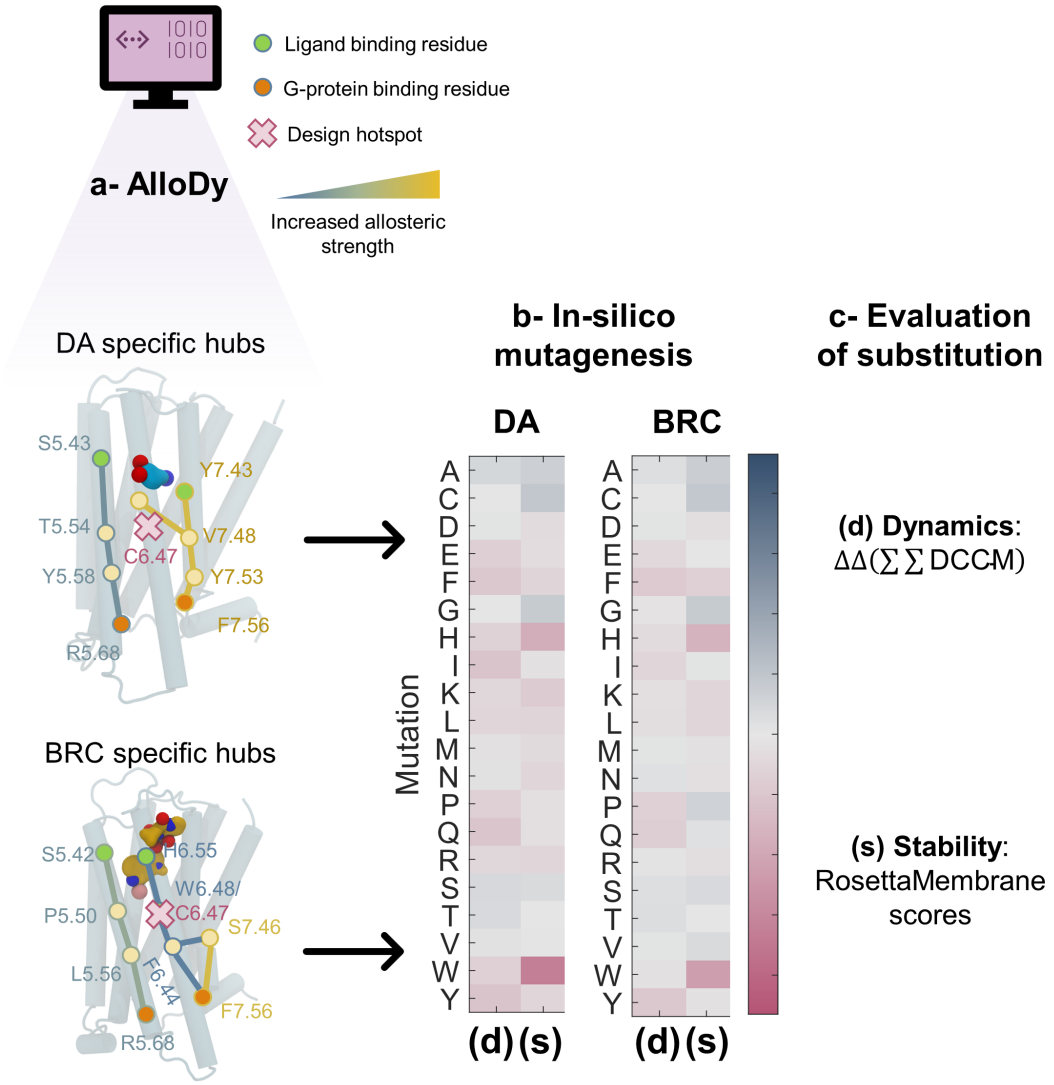


Figure 3.6: **Computational protein design strategy: a** Allosteric pathways are extracted using AlloDy. **b** Design hotspots are mutated to all 20 possible amino acids using RosettaMembrane (171). **c** Amino acid substitution is chosen based on stability (**s**) and coarse grained dynamics (**d**) calculated using Eq. 1.2.

We then applied our computational allosteric design approach to select in silico novel combi-

nations of amino-acids at these hubs predicted to enhance or weaken the allosteric coupling mediated by the ligand selective paths (Fig. 3.6). Design hotspots are chosen as residues with high allosteric score σ_m for either ligand and residues not directly contacting the ligand or the G-protein. DCCM values are calculated as difference of differences between (active and inactive state) and (mutant and WT) receptors. The allosteric hubs extracted from AlloDy are used in the sum in Eq. 1.2. The list of mutations, measured experimental efficacies and potencies, in addition to the calculated $\Delta\Delta DCCM$ values are summarized in Tab. 3.1.

Table 3.1: DD2R Gi2 pathway experimental responses for dopamine (DA) and bromociptine (BRC). SEM = standard error of the mean. Design column shows the origin of the design, phase 1 are taken from Chen et al. (16), while phase 2 are the variants designed in this work. ΔΔ(DCCM) is the sum described in Eq. 1.2.

Ligand	Mutation	LogEC50	SEM	Efficacy	SEM	Design	ΔΔ(DCCM)
DA	WT	-7.489	0.038	99.959	1.086	NA	NA
BRC	WT	-6.563	0.042	57.135	1.036	NA	NA
DA	T5.54M-C6.47L	-8.909	0.082	87.114	2.022	Phase 1 (16)	0.74
BRC	T5.54M-C6.47L	-6.570	0.100	56.907	2.934	Phase 1 (16)	NA
DA	L6.41M	-7.910	0.048	104.182	1.598	Phase 2	-0.00096
BRC	L6.41M	-6.756	0.096	83.078	3.205	Phase 2	-0.13
DA	L3.41G	-7.361	0.042	103.273	1.439	Phase 2	-0.11
BRC	L3.41G	-6.442	0.061	79.724	2.307	Phase 2	-0.048
DA	L3.41H	-8.674	0.061	97.338	1.674	Phase 2	-0.0094
BRC	L3.41H	-6.620	0.144	62.564	4.253	Phase 2	0.014
DA	F6.44M	-8.499	0.050	97.083	1.264	Phase 2	0.017
BRC	F6.44M	nan	nan	nan	nan	Phase 2	-0.0064
DA	F6.44I	-8.267	0.065	94.520	1.771	Phase 1 (16)	0.44
BRC	F6.44I	-6.401	0.118	61.130	4.011	Phase 1 (16)	NA
DA	I4.46N	-6.884	0.076	78.628	2.186	Phase 2	0.01
BRC	I4.46N	-7.114	0.684	3.288	1.000	Phase 2	-0.35
DA	C6.47L	-8.909	0.082	106.241	1.819	Phase 1 (16)	-0.51
BRC	C6.47L	-6.716	0.108	48.392	2.561	Phase 1 (16)	NA

Our aim was to engineer receptors with more ligand selective responses or gain of function variants for BRC. We targeted a wide variety of sites in the TM core of the receptor, including TMHs 3 and 4 that were not covered in our previous study (16). We validated our selected designs by measuring ligand-stimulated Gi-activation responses in TRP-HEK cells as described above. Consistent with our intentions, we were able to engineer D2 receptors with a high variation in response to BRC, ranging from gain of efficacy to total loss of signaling. In particular, designed microswitches at sites 3.41 and 6.41 enhanced BRC efficacy by up to 39.54% and 45.41%, respectively (Fig. 3.2e). Remarkably, these effects were highly selective for BRC in the L3.41G design which did not show any difference in DA responses as compared to WT. Conversely, sequence changes at the 4.46 hub had a slight loss of function effect on DA but considerably decreased the response to BRC. The extent of the loss of function depended on the exact sequence substitution, with I4.46T displaying the smallest effect, and I4.46N

displaying no Gi activation signals even at submillimolar concentrations of BRC. Lastly, F6.44M strongly enhanced the sensitivity for DA while suppressing all Gi-mediated response to BRC. This receptor achieved a very high level of ligand selectivity and, together with the other designs, demonstrates that the rational rewiring of selective allosteric pathways enables the fine-tuning and control of ligand mediated receptor functions (Fig. 3.2b and e).

Given the experimental findings, we expanded the molecular dynamics simulations set to include the following ligand-receptor variants of D2-GiH5 and extracted aforementioned features using AlloDy: DA-T5.54M-C6.47L; DA-L3.41G; DA-L3.41H; DA-I4.46N; DA-L6.41M; DA-F6.44M; BRC-T5.54M-C6.47L; BRC-L3.41G; BRC-L3.41H; BRC-I4.46N; BRC-L6.41M; BRC-F6.44M; and BRC-F6.44I.

From the total set of simulations, we attempted to extract global features that could correlate with experimental observations. We report that allosteric pathways passing through TMH 5 in dopamine systems show a correlation with the shift in potency in response to the mutation of the system (Fig. 3.8a). This agrees with the presence of consensus serines on TMH 5 ligand binding site in aminergic class A GPCRs (Fig. 3.7) and with previous studies with mutation of these sites to alanine is detrimental to G-protein signaling in dopamine (301) and adrenergic receptors (302; 303). Additionally, the largest difference in the ligand binding site between active and inactive β2AR is an inward bulge of TM5 centered at S5.46 (57). We observed this shift via backbone KL_1 between the active and inactive ensembles at positions 5.44, 5.45, and 5.46 (Fig. 3.4b).

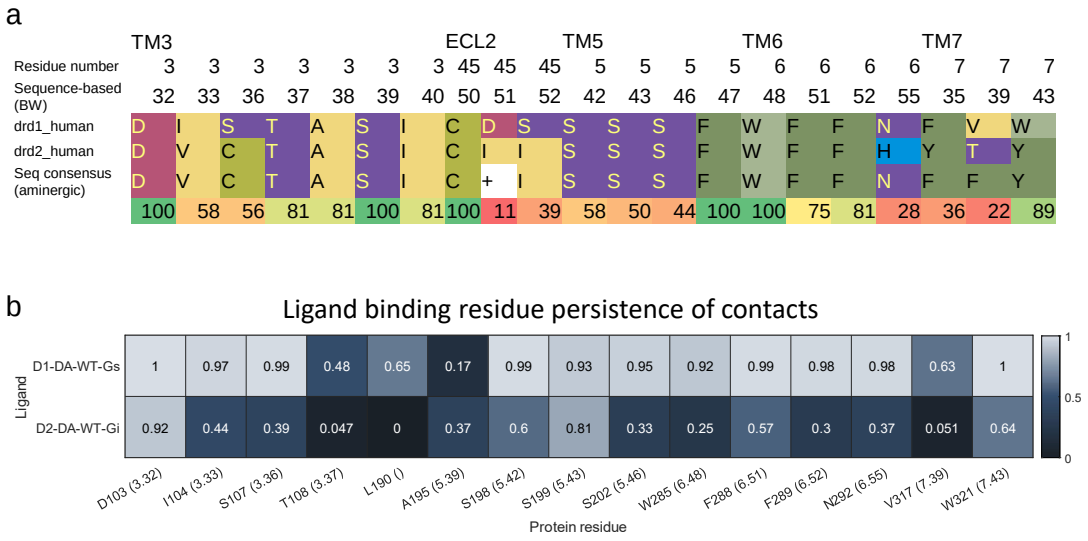
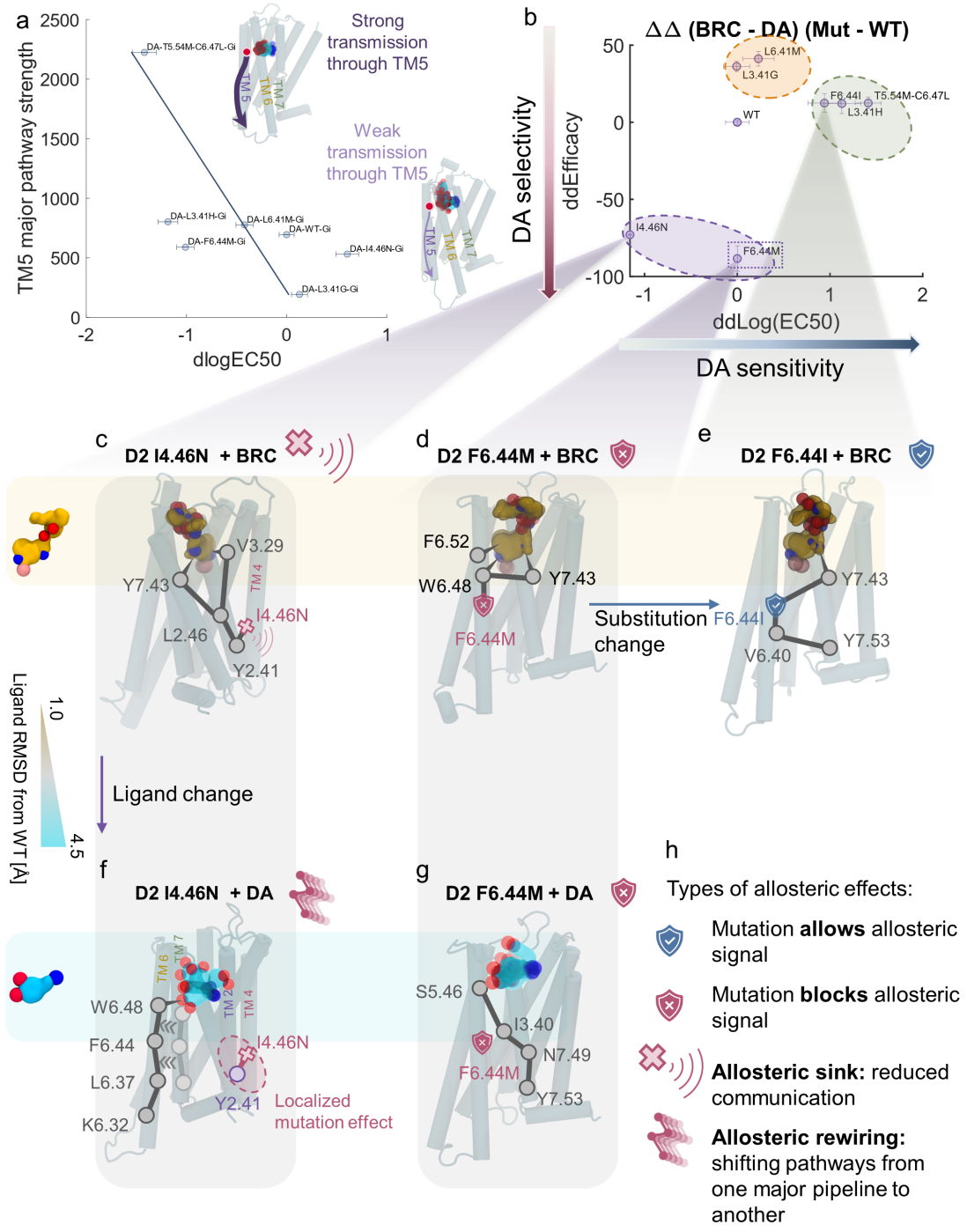


Figure 3.7: **Ligand binding residue conservation and persistence of contacts in DD1R and DD2R: a** Residues of ligand binding region for DD1R and DD2R. The sequence consensus among aminergic class A GPCRs. **b** Ligand binding residue persistence of contacts for DD1R and DD2R. Note that the amino acid identity and numbering shown on the x-axis is that of DD1R.



(Caption on next page.)

Figure 3.8: **DD2R allosteric effects explain ligand selectivity: a** Changes in potency for dopamine bound DD2R correlate with strength of major allosteric pathways passing through TM5. The correlation has an R^2 = 0.436 **b** Difference of differences of efficacy and potency between mutant and WT and bromocriptine and dopamine. **c-g** Models of allosteric effect in DD2R in response to mutations and ligands (bromocriptine: panels **c-e**, and dopamine: panels **f-g**. Ligands are colored according to their RMSD from their respective WT. **h** Legend for allosteric effects observed in the simulations

3.2.4 Molecular origins of ligand selectivity upon mutation in dopamine D2 receptors

Beyond WT systems, we have classified mutations into distinct clusters in ligand sensitivity-selectivity space (Fig. 3.8b), with the most DA selective variants being I4.46N and F6.44M. We will discuss these two variants in more detail to investigate the origins of ligand selectivity:

BRC-D2-I4.46N: Among the I4.46 substitutions, mutation I4.46N kills all signaling with BRC. We begin the investigation by comparing mutual information between the mutant and WT systems. While there is a reduction in the total MI of the system, there exists a set of selected residues, including the mutation site, neighboring sites (Y2.41, L2.46), and distant residues (W7.40, Y7.43) that have increased residue summed MI. We explore further by investigating KL_1, allosteric scores σ_m, and allosteric pathways. The asparagine mutation forms H-bonds with Y2.41 and S2.45, changing their side chain dihedral distributions as well as that of L2.46 (evidence from KL_1, Fig. 3.9c), which leads to L2.46 and Y2.41 acting as allosteric hubs (evidence from σ_m) at the junction of allosteric pathways connecting ligand binding regions (TM3, V3.29, TM6, F6.51, TM7, Y7.43, from allosteric pathway topology, Fig. 3.9a) to TM2 and the mutation site. This funnels information into TM2, as is seen from MI distributions, where only TM2 has higher MI than WT, while other TMs have lower summed MI (Tab. 3.2). A consequence of this is a reduction in the communication between G-p binding regions and the rest of the receptor (reduction of ~ 25%, from summed MI). The mutation also allosterically modifies the conformation of ECL3, G4.63 and sidechain of T7.39 (evidenced by KL_1). No major difference in ligand binding poses or contacts is observed. Synthesizing all of this information, we propose the effect of I4.46N on BRC as an allosteric sink mechanism (Fig. 3.8c), where information is diverted toward the neighborhood of the mutation and the ligand thus does not lead to signaling.

Table 3.2: **Summed MI over every TM helix in BRC-D2-WT-GiH5 and BRC-D2-I4.46N-GiH5.**

Summed MI	TM1	TM2	TM3	TM4	TM5	TM6	TM7
BRC-WT-Gi	717.2	319.7	374.6	613.7	782.1	797.7	387.2
BRC-I4.46N-Gi	686.7	386.5	326.8	352.4	619.1	536.0	275.9

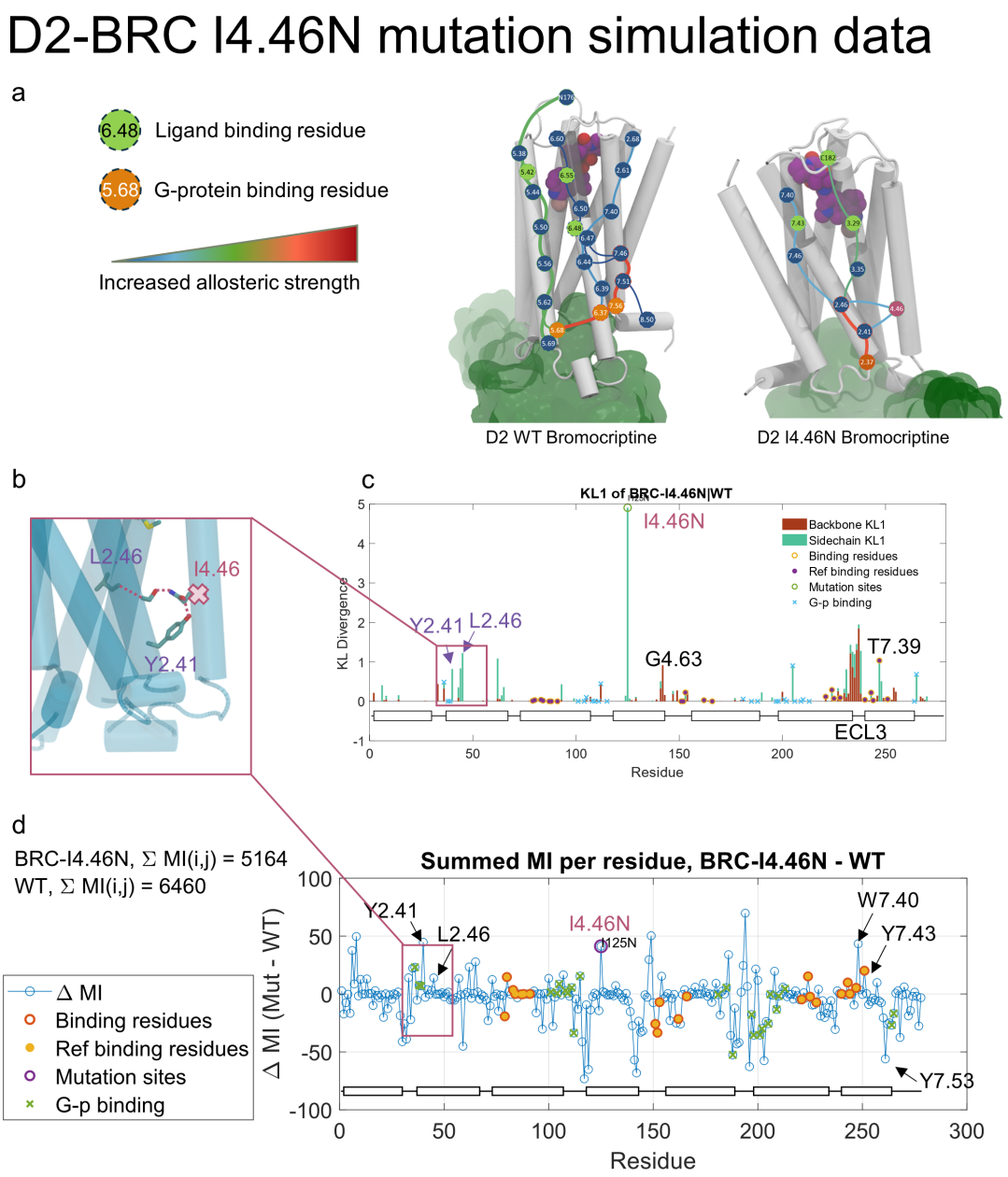


Figure 3.9: **BRC-D2-I4.46N-GiH5 simulation data: a** Selection of top ten allosteric pathways for BRC-D2-WT-GiH5 (left) and BRC-D2-I4.46N-GiH5 (right). **b** N4.46 forming polar interactions with Y2.41 and S2.45. **c** KL-divergence with BRC-D2-I4.46N-GiH5 as target ensemble and BRC-D2WT-GiH5 as reference ensemble. **d** Difference of residue summed mutual information (MI) between target and reference ensembles. Ligand binding residues and G-protein binding residues are marked on the figure. Rectangles represent the seven TM helices of GPCRs.

DA-D2-I4.46N: On the other hand, DA is still able to signal through Gi despite the mutation being a loss of function both in terms of efficacy and potency. Similarly to BRC, the asparagine mutation forms H-bonds with Y2.41 and S2.45, changing their side chain dihedral distributions (mostly Y2.41, evidenced by KL_1, Fig. 3.10b and c). Unlike the case of BRC, this does not translate to an allosteric signal through TM2 (Fig. 3.10a). In fact, Y2.41 has a loss of total MI in I4.46N (Fig. 3.10d). Thus, we conclude that the effect of mutation is structurally localized (evidenced by KL_1). However, the ligand (DA) reacts to the mutation by changing binding pose (separate cluster from WT simulations in Fig. A.12), which is reflected in the contact persistence map and allosteric scores of ligand binding residues (Fig. 3.3a and b). Residue Y7.43 has a weaker allosteric score despite not losing contact, and the role is taken instead by W6.48, which has increased contact persistence and allosteric score. This change leads to a shift in allosteric communication from TM7 to TM6, which is reflected in MI of TM6 and TM7 for I4.46N (Fig. 3.10d), which we term allosteric rewiring (Fig. 3.8f).

Table 3.3: **Summed MI over every TM helix in DA-D2-WT-GiH5 and DA-D2-I4.46N-GiH5.**

Summed MI	TM1	TM2	TM3	TM4	TM5	TM6	TM7
DA-WT-Gi	629.7	412.8	586.8	821.9	934.0	743.8	1155.5
DA-I4.46N-Gi	607.1	418.0	447.3	876.5	919.2	965.8	690.1

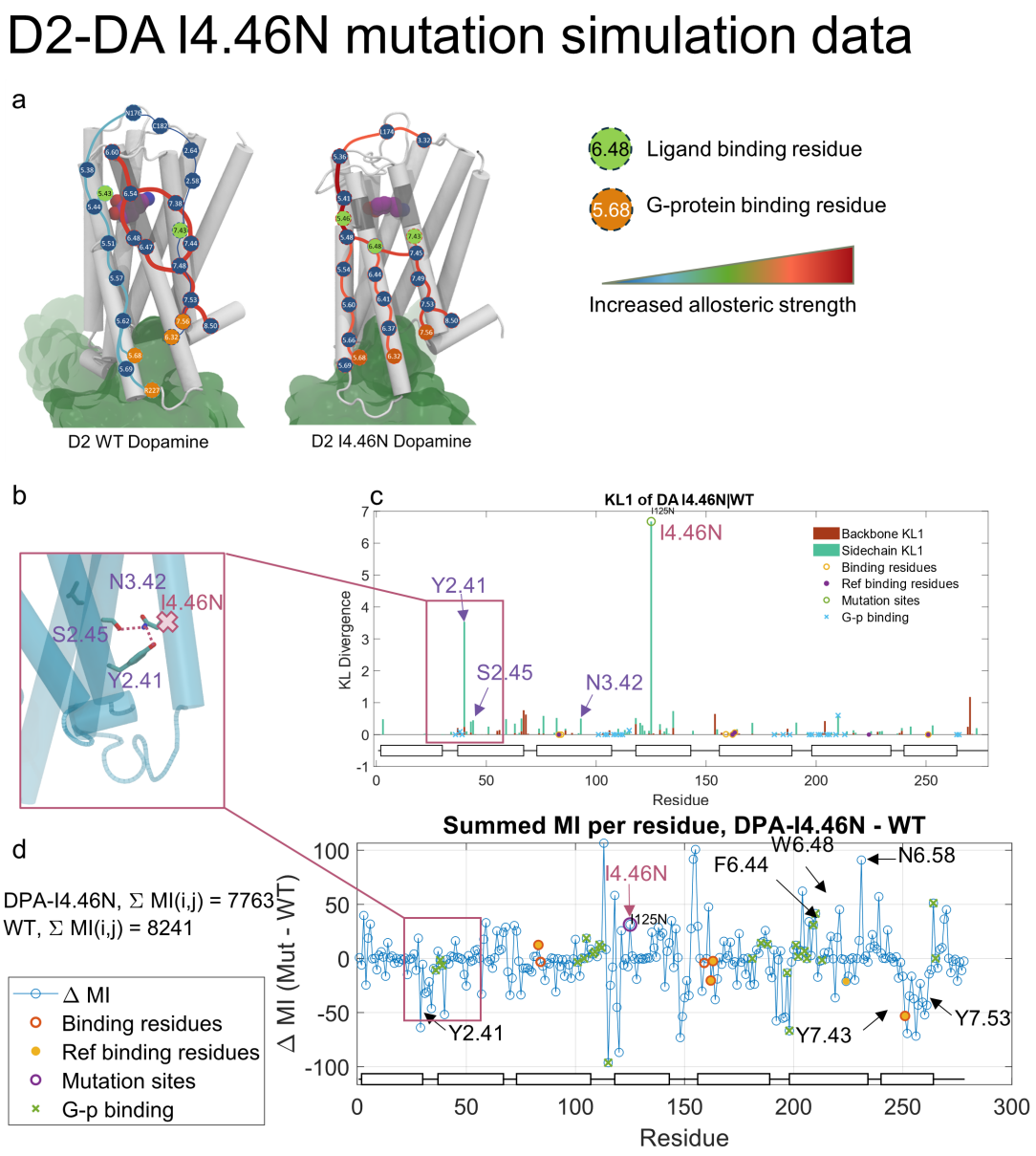


Figure 3.10: **DA-D2-I4.46N-GiH5 simulation data: a** Selection of top ten allosteric pathways for DA-D2-WT-GiH5 (left) and DA-D2-I4.46N-GiH5 (right). **b** N4.46 forming polar interactions with Y2.41 and S2.45. **c** KL-divergence with DA-D2-I4.46N-GiH5 as target ensemble and DA-D2WT-GiH5 as reference ensemble. **d** Difference of residue summed mutual information (MI) between target and reference ensembles. Ligand binding residues and G-protein binding residues are marked on the figure. Rectangles represent the seven TM helices of GPCRs.

Another highly selective mutation is F6.44M, which is dead for BRC but increases potency toward DA:

BRC-D2-F6.44M and F6.44I: In the case of BRC-D2-F6.44M, we observe changes in rotameric states of I3.40 and L6.41 (evidenced by KL_1, Fig. 3.11b). Contrary to the WT case, allosteric pathways jump from W6.48 to I3.40, skipping position 6.44 (Fig. 3.11a). This is supported by M6.44 having lower σ_m than WT with BRC, but slightly higher with DA. In addition, substitution F6.44I with BRC displays larger σ_m than WT, and this is reflected in I6.44 not blocking allosteric transmission (Fig. 3.8d-e). Another piece of evidence is that *apo* state simulations of D2 WT and F6.44M starting from the active state have shown that F6.44M accesses an inactive-like state more readily than WT in terms of TM3-6 distance (Fig. A.11).

Table 3.4: **Summed MI over every TM helix in BRC-D2-WT-GiH5, BRC-D2-F6.44M-GiH5, and BRC-D2-F6.44I-GiH5.**

Summed MI	TM1	TM2	TM3	TM4	TM5	TM6	TM7
BRC-WT-Gi	717.2	319.7	374.6	613.7	782.1	797.7	387.2
BRC-F6.44M-Gi	877.1	640.5	426.3	687.1	909.3	1206.6	731.0
BRC-F6.44I-Gi	559.2	532.3	489.7	740.2	821.8	674.3	761.5

DA-D2-F6.44M: In the case of DA, no divergence is observed at position I3.40 (evidenced by KL_1). DA changes binding pose (separate cluster from WT simulations in Fig. A.12), and shows stronger allosteric contacts with TM5 (S5.46), TM6, (F6.51), and TM7 (Y7.43) (Fig. 3.3), which would hint at the increase in potency. In terms of allosteric transmission, pathways show communication between TMH 3 -TMH 7 and TMH 6 (residues 6.47- 6.48) – TMH 7 that lead to the NPxxY motif and to G-protein binding residues in TMHs 6 and 7. M6.44 still acts as an allosteric block, but DA is able to adapt its binding pose, leading to signaling regardless of the mutation (Fig. 3.8g).

Table 3.5: **Summed MI over every TM helix in DA-D2-WT-GiH5 and DA-D2-F6.44I-GiH5.**

Summed MI	TM1	TM2	TM3	TM4	TM5	TM6	TM7
DA-WT-Gi	629.7	412.8	586.8	821.9	934.0	743.8	1155.5
DA-F6.44M-Gi	594.3	614.0	523.4	825.7	862.1	924.1	959.5

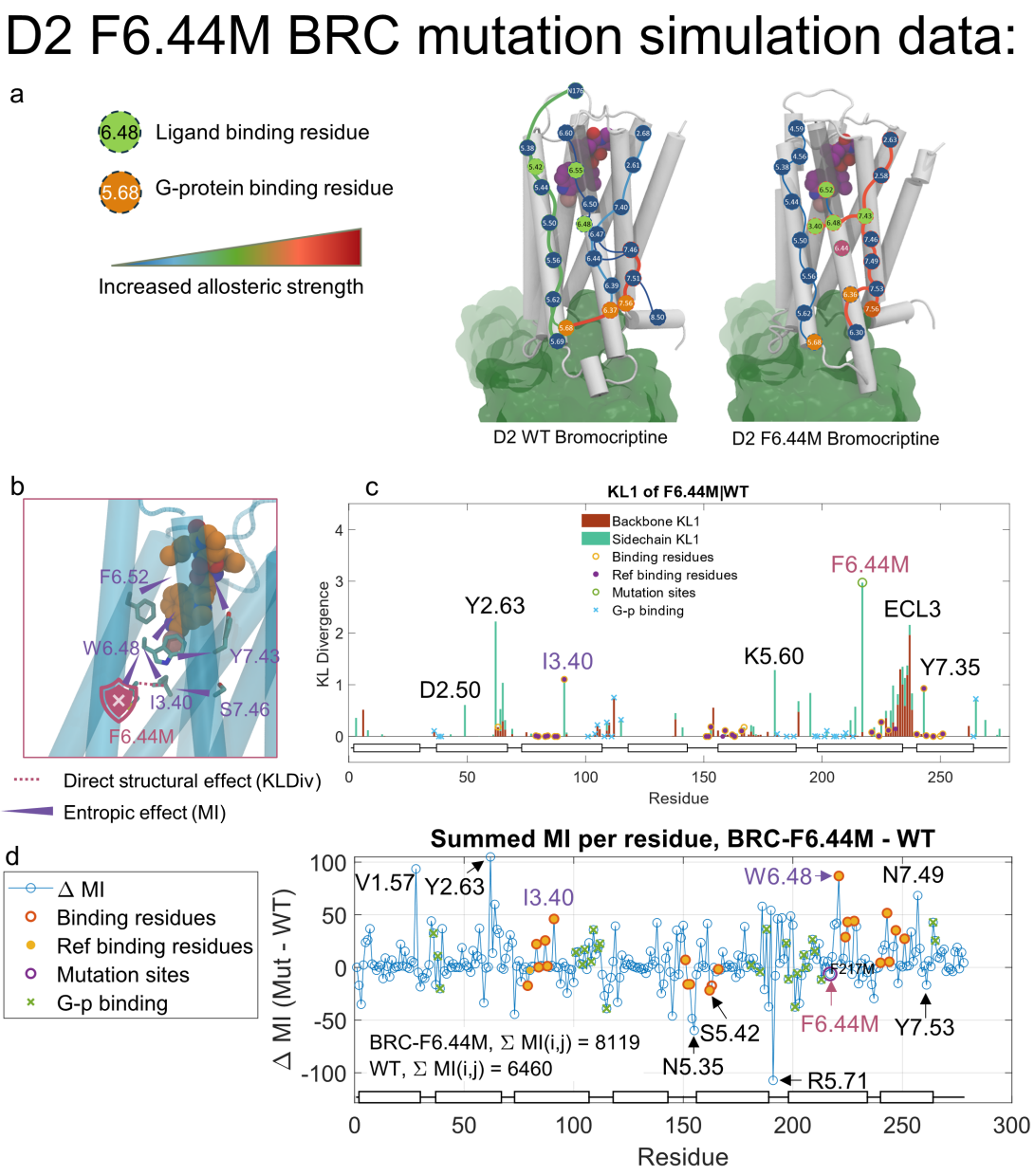


Figure 3.11: **BRC-D2-F6.44M-GiH5 simulation data: a** Selection of top ten allosteric pathways for BRC-D2-WT-GiH5 (left) and BRC-D2-F6.44M-GiH5 (right). **b** KL-divergence with BRC-D2-F6.44M-GiH5 as target ensemble and BRC-D2WT-GiH5 as reference ensemble. **c** Difference of residue summed mutual information (MI) between target and reference ensembles. Ligand binding residues and G-protein binding residues are marked on the figure. Rectangles represent the seven TM helices of GPCRs.

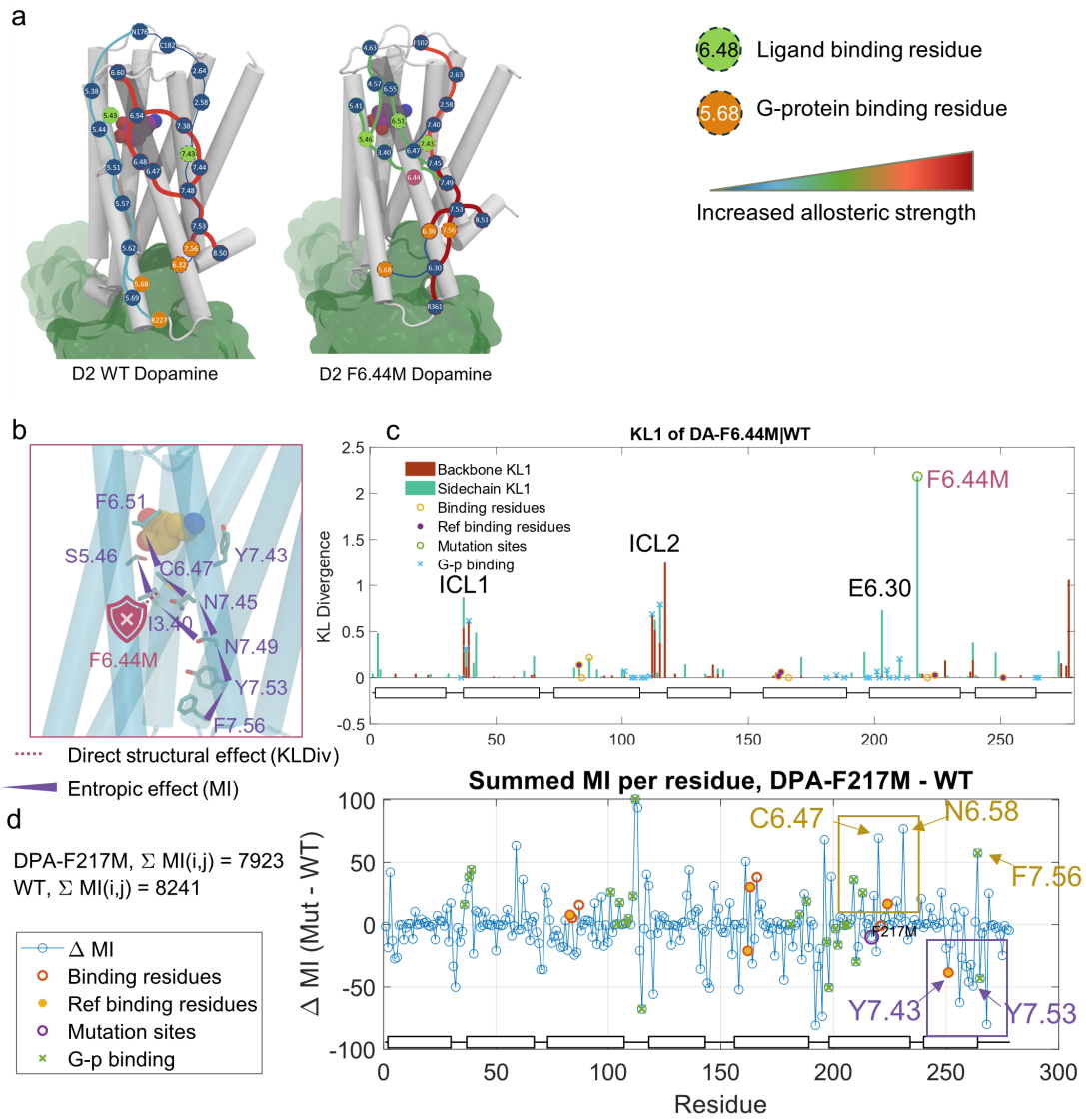


Figure 3.12: **DA-D2-F6.44M-GiH5 simulation data: a** Selection of top ten allosteric pathways for DA-D2-WT-GiH5 (left) and DA-D2-F6.44M-GiH5 (right). **b** KL-divergence with DA-D2-F6.44M-GiH5 as target ensemble and DA-D2WT-GiH5 as reference ensemble. **c** Difference of residue summed mutual information (MI) between target and reference ensembles. Ligand binding residues and G-protein binding residues are marked on the figure. Rectangles represent the seven TM helices of GPCRs.

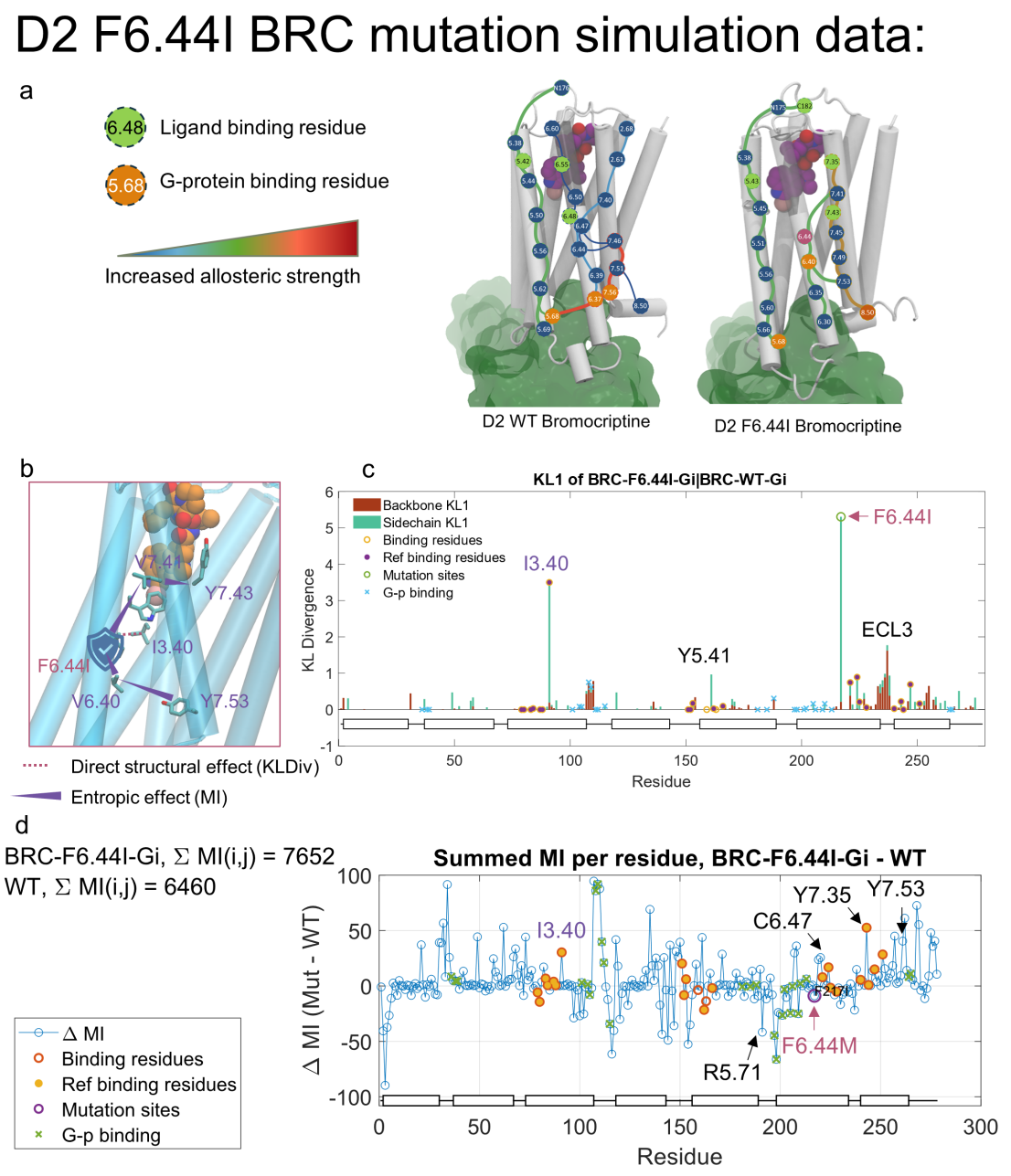


Figure 3.13: **BRC-D2-F6.44I-GiH5 simulation data: a** Selection of top ten allosteric pathways for BRC-D2-WT-GiH5 (left) and BRC-D2-F6.44I-GiH5 (right). **b** KL-divergence with BRC-D2-F6.44I-GiH5 as target ensemble and BRC-D2WT-GiH5 as reference ensemble. **c** Difference of residue summed mutual information (MI) between target and reference ensembles. Ligand binding residues and G-protein binding residues are marked on the figure. Rectangles represent the seven TM helices of GPCRs.

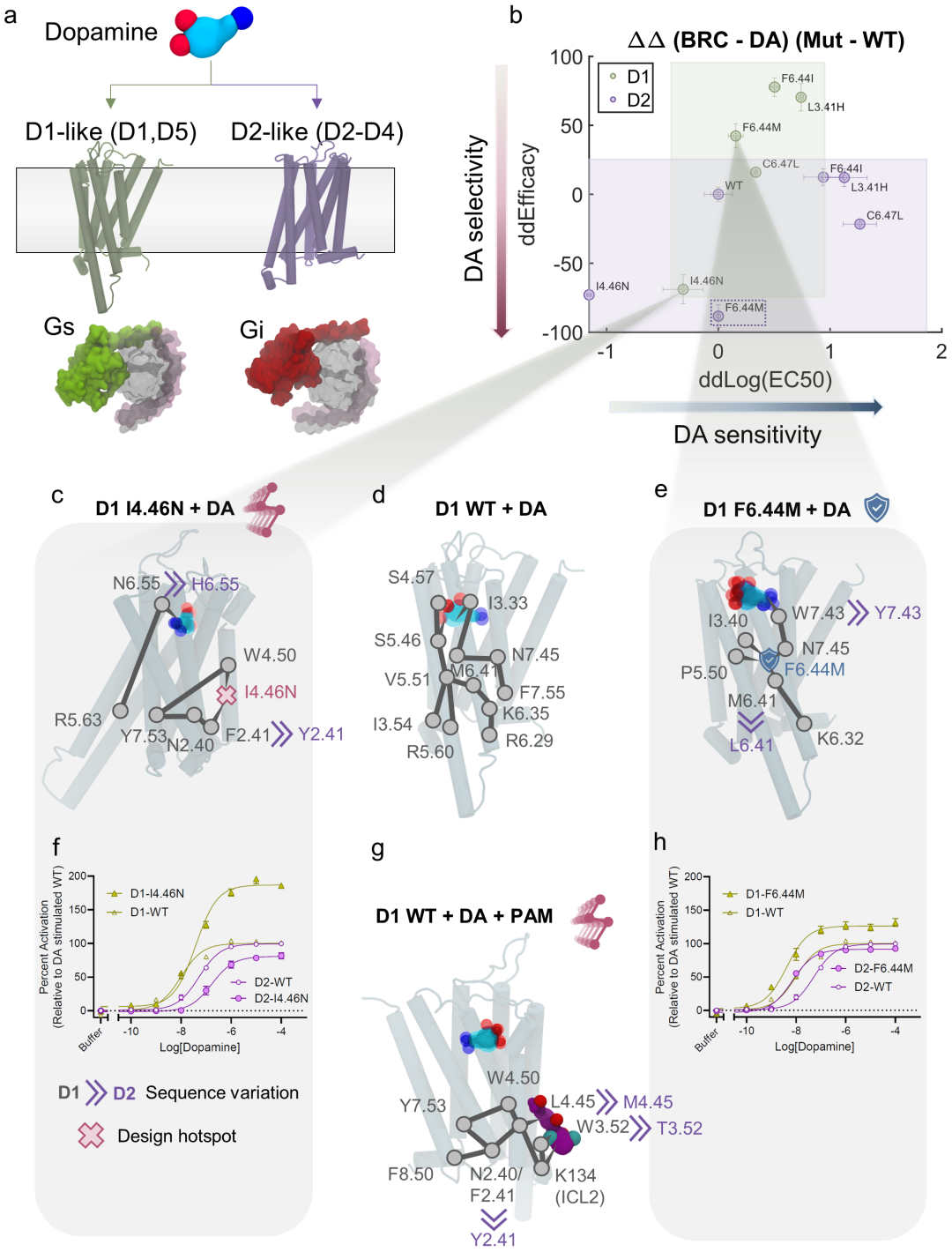
3.2.5 Allostery across the dopamine family, case of dopamine D1 receptor

To investigate the effects of the most selective mutations in a different sequence context, we computationally and experimentally studied a subset of the mutations and ligands chosen for the D2 receptors in D1. Dopamine receptor family includes five receptors divided into two sub-families, the D1-like (D1 and D5), which binds to the stimulatory G-protein Gs, and D2-like (D2, D3, and D4), which primary signal through the inhibitory G-protein G-i/o. D1 and D2 are the most abundant in the central nervous system, and are thus the target of this study (Fig. 3.14a) (304; 305)

We selected three ligands (DA, BRC, and 5-HT) and five mutations (L3.41H, I4.46N, F6.44I, F6.44M, C6.47L) and measured *in vitro* the D1-mediated activation of the G protein Gs upon ligand stimulus using HEK reporter cell lines using the EPAC cAMP assay. Dose titrations revealed very distinct effects of the designed microswitches on the assayed ligands compared to their responses to D2-Gi.

The first significant observation is that none of the tested variants exhibited any loss of function behavior. They all ranged from neutral to gain of function. The second difference is that DA responses exhibit a significant gain of efficacy that was not observed in D2 variants (all tested D2 variants exhibited no effect or loss of efficacy). Finally, and most interestingly, a general loss of function variant (I4.46N) became gain of function (DA) or neutral (BRC, 5-HT), and variant F6.44M, which is dead for D2-BRC, becomes a gain of function for D1-BRC (Fig. 3.14b).

Similar to D2, we combined molecular simulations with AlloDy analysis to investigate the molecular mechanisms of differences between D1 and D2 responses. Furthermore, sequence differences in ligand binding site and in receptor core could play a role in this divergent response. To be able to compare how dopamine engages allosteric signaling in DD1R vs DD2R, we would first need to identify key differences in the ligand binding site and compare with the consensus of aminergic receptor binding sites (Fig. 3.7). Consensus serines on TM5, positions 5.42, 5.43, and 5.46 (conservation percentages 58, 50, and 44 in class A aminergic GPCRs) interact with the two hydroxyl groups on DA (also known as a catechol motif). Interactions with these serines is crucial for allosteric signal transmission in both receptors, and even non-catechol based agonists, such as PW0464 still interact with these serines via fluorine H-bonding (note that ligand PW0464 activates Gs but does not recruit arrestins) (306; 307). Fully conserved aspartic acid on 3.32 forms a salt bridge with a charged amine common to aminergic receptors. Allosteric pathway calculations with AlloDy show that this is a purely structural interaction with very little to no role in allosteric communication (Fig. 3.3b and d). On TM6, F6.51 interacts in both simulation sets and participates in allosteric signal on TM6 with the other binding residues. F6.52 is common to both, but it only forms a T-stacking interaction with the dopamine ring in D1.



(Caption on next page.)

Figure 3.14: **Ligands and mutations reacting in different contexts, case of Dopamine D1 and D2 receptors: a** dopamine D1 and D2 receptors diverge evolutionarily (sequence identity of 19% in humans) and bind to different G-proteins. **b** Difference of differences of efficacy and potency between mutant and WT and bromocriptine and dopamine for D1 and D2 receptors. The x and y-axes can be interpreted as DA sensitivity and DA selectivity respectively. **c** Model of allosteric enhancing effect of I4.46N in DA-D1-GsH5 simulations. **d** Major allosteric pathways in DA-D1WT-GsH5 simulations. **e** Model of allosteric effect of F6.44M in DA-D1-GsH5 simulations. **f** and **h** Dose response curves for D1 mediated Gs activation and D2 mediated Gi activation in response to DA for I4.46N (**f**) and F6.44M (**h**) variants. **g** Model of allosteric enhancing effect of positive allosteric modulator (PAM) binding to D1.

On the other hand, the main differences in binding residues are Y/W7.43, H/N6.55, and C/S3.36, where the first residue is that of D2 and the second residue is that of D1. Y7.43 is a major allosteric hub in D2 active state simulations, while W7.43 is almost absent from allosteric paths and MI signal (W7.43 average residue summed MI is 0.0114 while that of Y7.43 is 0.3996 in DA-bound WT systems). Residue 6.55 is absent from D1 simulations and is weakly present in D2 simulations. 6.55 has been reported to be a "modulator" residue for arrestin signaling via sequence variation (139). Finally, 3.36 contributes very weakly to allosteric pathways in D2 (C3.36), while S3.36 is involved in D1-DA (Fig. 3.3).

Insights from DD1R WT simulations

Three sets of WT DD1R simulations were prepared starting from previously solved DD1R structures (PDB code 7CKZ (306)): DD1R-GsH5 bound to DA and the positive allosteric modulator (PAM) LY3154207, DD1R-GsH5 bound to DA, and *apo* DD1R simulations.

Starting with the general features of D1 simulations, both DA-GsH5 bound simulations exhibited a single activation state in TM3-6, TM3-7, and NPxxY RMSD to inactive space with a single well that includes the coordinates of the starting structure (Fig. 3.15a and b). Dopamine was stable in the binding pocket with ~ 1.4 Å RMSD and assumed one major and one minor binding pose in both systems. This was in contrast with DD2R-DA-6VMS simulations, where DA had more flexibility and covered a larger span of binding poses (Fig. 3.16).

The effect of the PAM is a subtle but important one. The PAM binds in the cleft between TM3/4 towards ICL2 and has little structural effect on the receptor (Fig. 3.15). In contrast, the PAM effect could be clearly seen in the mutual information and allosteric pathways. MI per residue were significantly larger in PAM bound simulations, and the effect permeated beyond the PAM binding region. Experimentally, the PAM has been reported to increase the potency of D1 to DA but not its efficacy and can even slightly activate the receptor at high enough concentrations (306).

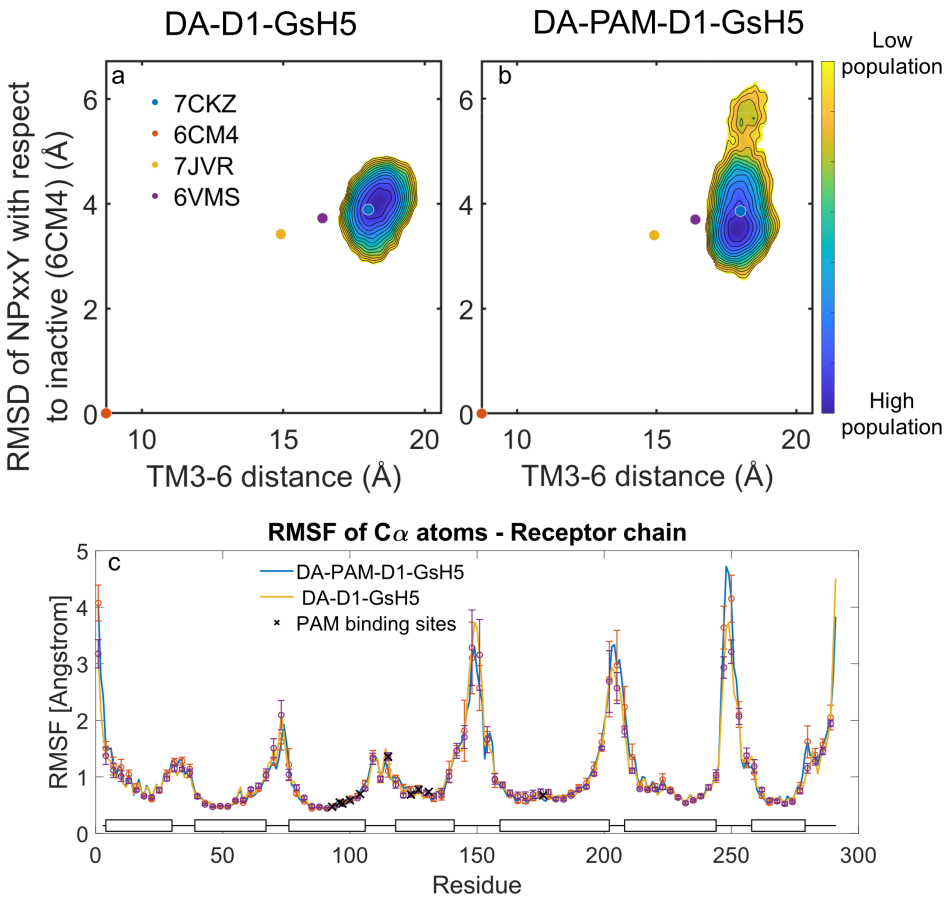


Figure 3.15: **DA-D1-GsH5 simulations and effect of the PAM: a** and **b** Activation landscapes for DA-D1-GsH5 simulations (**a**) and DA-PAM-D1-GsH5 simulations (**b**). TM3-6 distance is measured between residues R3.50 and E6.30. **c** RMSF plots of DA-DA-GsH5 with and without PAM present in the simulations. Rectangles represent the seven TM helices of GPCRs.

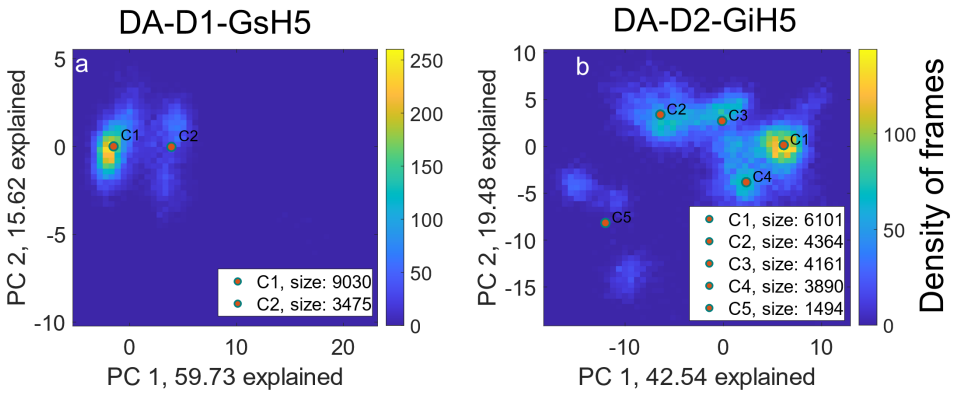


Figure 3.16: **PCA of ligand binding poses for DA-bound WT DD1R and DD2R: a** DA-D1-GsH5 simulations and **b** DA-D2-GiH5. PCA was performed on ligand heavy atoms. Data was clustered with 2 PCs using a K-means algorithm.

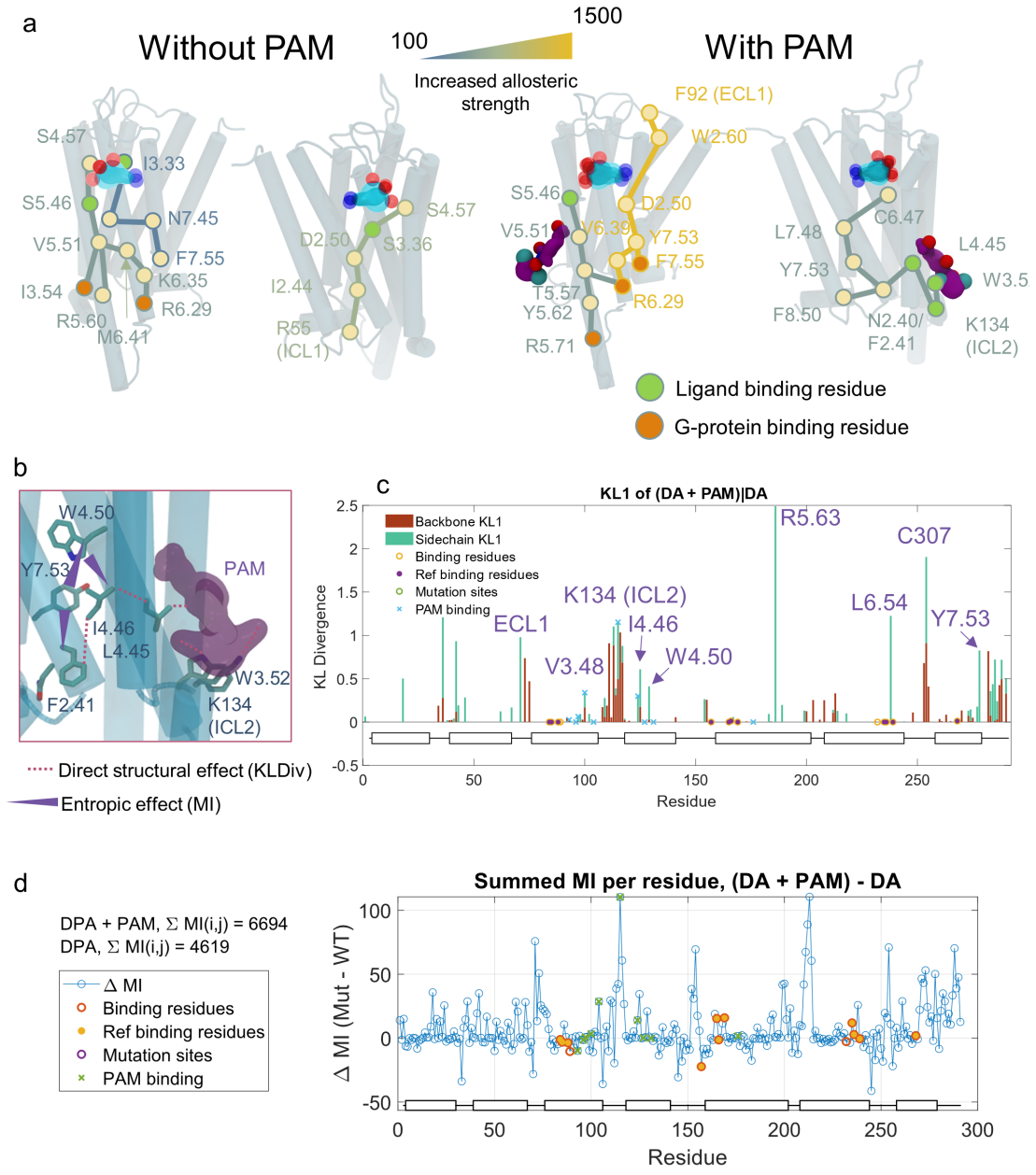


Figure 3.17: **DA-D1WT-GsH5 simulation data with and without PAM: a** Selection of top ten allosteric pathways for DA-D1WT-GsH5 (left) and DA-PAM-D1WT-GsH5 (right). **b** PAM binding site interactions. **c** KL-divergence with DA-PAM-D1WT-GsH5 as target ensemble and DA-D1WT-GsH as reference ensemble. **d** Difference of residue summed mutual information (MI) between target and reference ensembles. Ligand binding residues and G-protein binding residues are marked on the figure. Rectangles represent the seven TM helices of GPCRs.

Beyond WT D1 systems, models of allostery in D1 variants

DA-D1-I4.46N: Unlike the case of DD2R, I4.46N shows a gain of function for DA in DD1R (Fig. 3.14f) when it is a loss of function in D2. Qualitatively, the mutation shifts allosteric pathways toward a more active-like topology in an unusual way. We can observe top-ranking pathways connecting to the IC binding site going through the mutation neighborhood in addition to the classical pathways connecting ligand binding to IC binding site. Moreover, top ranking allosteric hubs shift toward the mutation neighborhood of TM4 (residues W4.50, N4.46, and L4.45) and include Y7.53, which was weaker in WT in terms of hubscore σ_m and MI (Fig. 3.18). Combined MI, KL_1, and σ_m analysis shows a clear set of connections from I4.46N to F2.41 and then R55 (ICL1) all the way to Y7.53 (Fig. 3.14c). Note that a very similar pattern is seen in the PAM bound simulations, which suggests that the mutation and the PAM affect the activation of D1 in a similar way.

DA-D1-I4.46N: Mutation F6.44M exhibits gain of function effects for all ligands tested with D1. Simulations of F6.44M-D1-DA-GsH5 show that position M6.44 allows allosteric signaling through it (Fig. 3.14e), starting from ligand binding site W7.43 and connecting to M6.44 through N7.45 and then continuing along TMH 6 to M6.41 toward the G-p binding interface. Residue summed MI is significantly increased for all the aforementioned residues (Fig. 3.19), and ligand binding hubscores σ_m are significantly stronger for W7.43 and ligand binding residues in TMH 6 for the F6.44M variant. This behavior is in stark contrast with D2 simulations, where M6.44 acted as an allosteric block. Sequence variation at the ligand binding site (Y7.43 to W7.43) and in the mutation neighborhood (L6.41 to M6.41) could play an important role in this divergent behavior.

Table 3.6: **DD1R Gs pathway experimental responses for dopamine (DA) and bromociptine (BRC). SEM = standard error of the mean.**

Ligand	Mutation	LogEC50	SEM	Efficacy	SEM
DA	WT	-8.003	0.059	100.120	1.668
BRC	WT	-5.603	0.094	55.983	2.339
DA	I4.46N	-7.419	0.054	186.842	2.891
BRC	I4.46N	-5.334	0.200	65.906	6.086
DA	F6.44M	-8.381	0.089	126.321	2.793
BRC	F6.44M	-5.825	0.121	94.318	5.060
DA	F6.44I	-8.919	0.052	151.426	1.809
BRC	F6.44I	-6.016	0.067	128.217	3.361
DA	L3.41H	-8.862	0.152	176.491	6.308
BRC	L3.41H	-5.722	0.137	138.044	7.673
DA	C6.47L	-8.680	0.077	121.148	2.118
BRC	C6.47L	-5.945	0.053	76.706	1.785

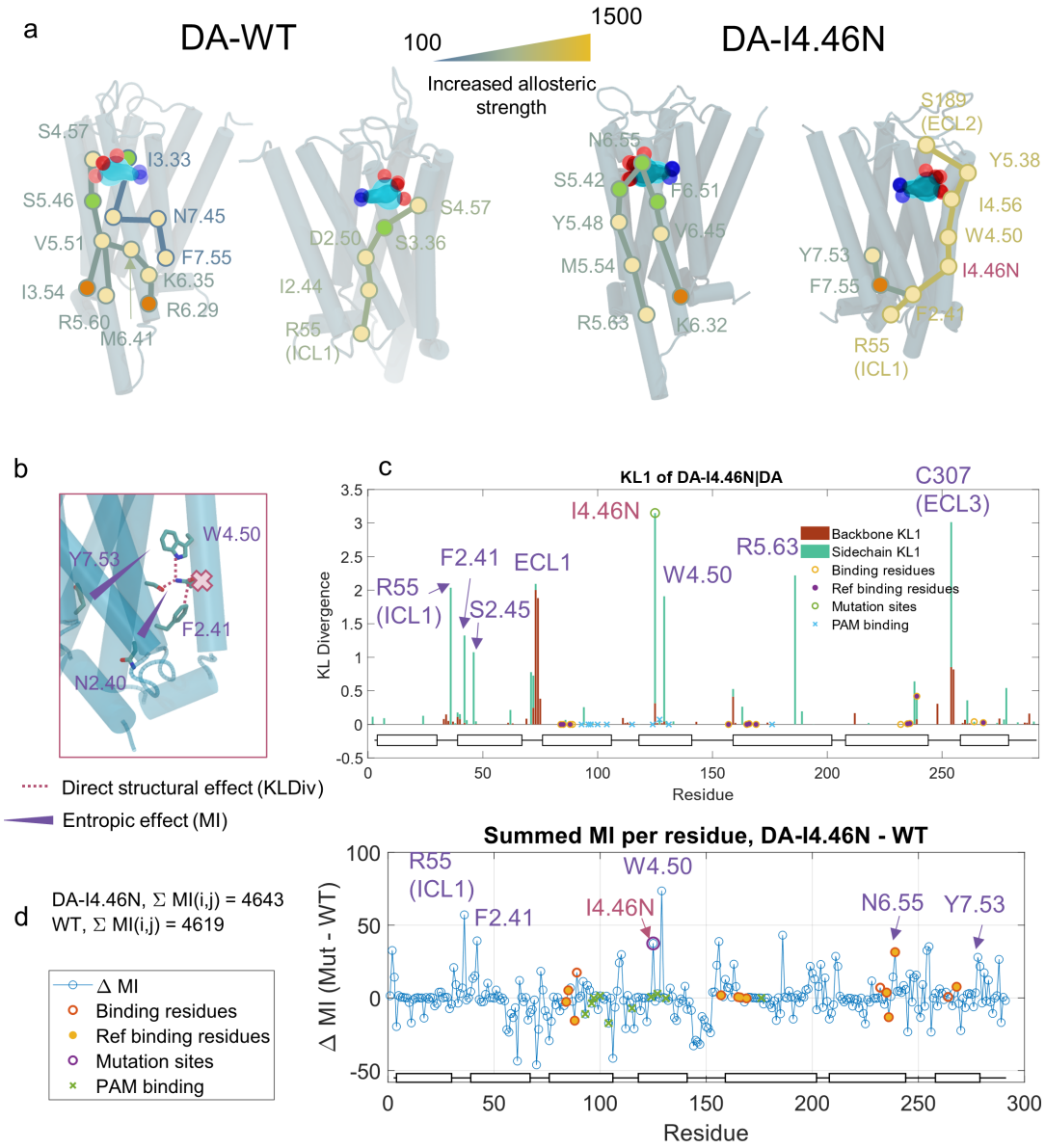


Figure 3.18: **DA-D1-I4.46N-GsH5 simulation data: a** Selection of top ten allosteric pathways for DA-D1WT-GsH5 (left) and DA-D1-I4.46N-GsH5 (right). **b** Mutation site interactions. **c** KL-divergence with DA-D1-I4.46N-GsH55 as target ensemble and DA-D1WT-GsH as reference ensemble. **d** Difference of residue summed mutual information (MI) between target and reference ensembles. Ligand binding residues and G-protein binding residues are marked on the figure. Rectangles represent the seven TM helices of GPCRs.

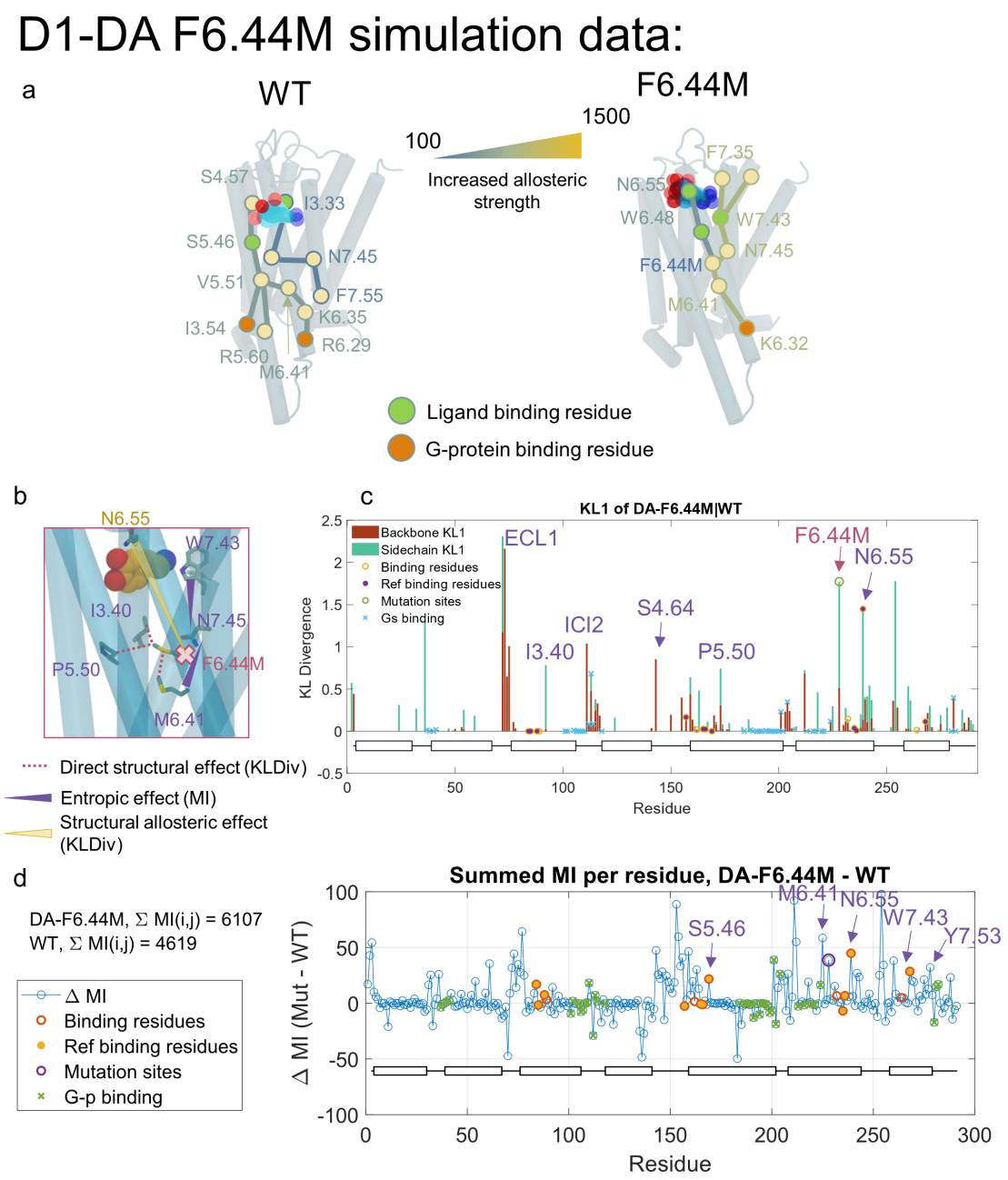


Figure 3.19: **DA-D1-F6.44M-GsH5 simulation data: a** Selection of top ten allosteric pathways for DA-D1WT-GsH5 (left) and DA-D1-F6.44M-GsH5 (right). **b** Mutation site interactions. **c** KL-divergence with DA-D1-F6.44M-GsH5 as target ensemble and DA-D1WT-GsH as reference ensemble. **d** Difference of residue summed mutual information (MI) between target and reference ensembles. Ligand binding residues and G-protein binding residues are marked on the figure. Rectangles represent the seven TM helices of GPCRs.

3.3 Discussion

In this study, we explored the relationships between agonist ligand chemical space and sequence variation in dopamine receptor signaling responses through the lens of receptor conformational dynamics, allostery and protein design. We found that allosteric microswitches designed in the receptor TM core to enhance ligand-mediated Gi activation had gain of function effects that were specific to ligand agonists. A strong correlation was observed between ligand structural similarity and the functional shifts measured for ligand-D2 pairs and prompted us to further investigate the two full agonist ligands (DA and BRC) that displayed the most divergent Gi-mediated responses to designed microswitches. Analysis by AlloDy revealed a distinct topology of allosteric pathways connecting each ligand to the Gi binding site and different path perturbation and rewiring upon designed microswitches. We were able to design variants that preferentially signal through one ligand but not the other.

Furthermore, we have observed that sequence variation among dopamine receptors leads to divergent responses to mutations. Our study suggests that distinct ligand agonists can activate a given signaling pathway through specific "allosteric activator" moieties that engage partially independent allosteric pathways running through the receptor (Fig. 3.20a). These allosteric activators diverge among receptors of the same family in terms of both sequence and allosteric strength, such as the case of Y/W7.43 between DD2R/DD1R. Combining our data, DA ligand affinity data, and mutagenesis data reported in the literature (306; 301), we hypothesize that DD1R has evolved to optimize for affinity and not signaling (efficacy), while DD2R has evolved for signaling but not affinity. Three pieces of evidence are presented for this statement: 1- Mutagenesis studies of DD1R ligand binding site: mutating DD1R residues to their DD2R counterpart reduces EC50 and increases efficacy ((301), Figure S6 mutations N6.55H and W7.43Y). 2- Our mutagenesis data reported in this work and in Chen et al. (16), we observed increased EC50 but not increased efficacy for DD2R. While our mutagenesis study is limited in scope, it suggests that DD2R is already optimized for signal transduction/efficacy. 3- The presence of PAMs for DD1R signifies that DD1R by itself is not ideal for signaling and could be improved upon by allosteric modulators. This is clear from our allosteric calculations, where there is a general increase of information transfer across the receptor upon PAM binding. This could be related to their biological functions, with DD1R activating a stimulatory G-protein and DD2R binding an inhibitory G-protein. Overall, our ability to rationally design receptor sequence variants with fine-tuned signaling responses to specific ligands paves the way for engineering very selective ligand biosensors and predicting the effects of receptor polymorphisms on drug pharmacology for enhanced personalized medicine.

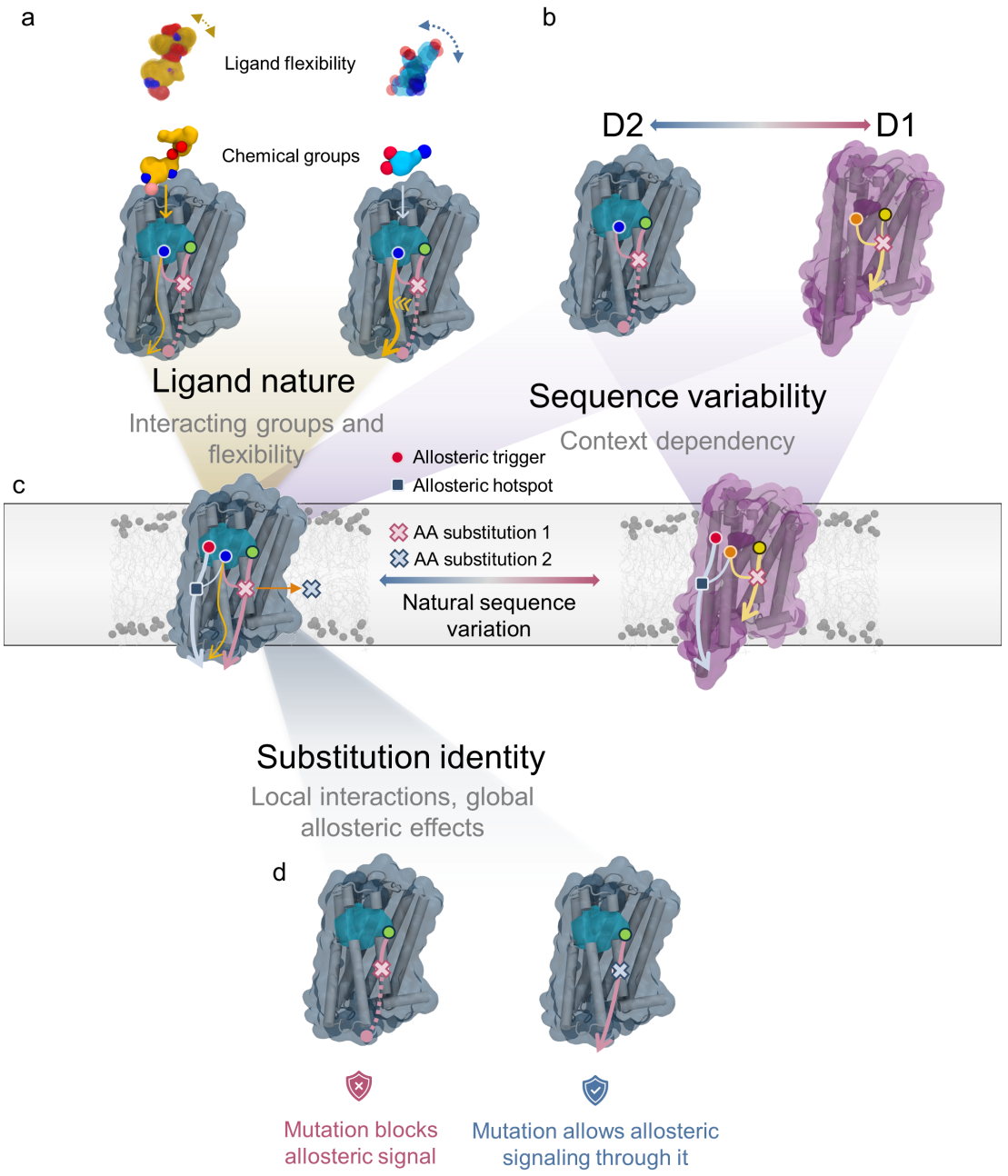


Figure 3.20: **Ligand properties, sequence variation, and substitution identity play major roles in receptor response to substitutions: a** The role of the ligand is played through its level of flexibility and chemical groups interacting with ligand binding sites. **b** Sequence context dependence of an amino acid substitution can change its effect from a loss to a gain of function by allowing allosteric signal through it. **c** Synthesis of all the mentioned effects. **d** Substitution identity changes local interactions or subtle rotameric states that lead to divergent allosteric responses.

3.4 Methods

3.4.1 D2 ligand clustering

A Dopamine D2 receptor ligand list was obtained from the CHEMBL database (https://www.ebi.ac.uk/chembl/) and only agonists were kept. The physicochemical properties of all agonists were summarized using the JOELib cheminformatics library. The resulting data were clustered using density-based clustering (DBSCAN) (308) and finally dimensionality was reduced using principal component analysis (PCA) for visualization.

3.4.2 G-alphaI TRP channel cell-based assay

HEK-293 cells stably expressing the TrpC4β channel (generously provided by Dr. Michael X. Zhu) were maintained in DMEM supplemented with 10% fetal bovine serum (FBS) and 50μg/mL geneticin as a selective antibiotic and grown at 37°C and 5% CO2. FLIPR Membrane potential assays (Molecular Devices) were performed as previously described1. Briefly, the assay relies on the detection of a membrane-permeable fluorescent dye coupled to a non-permeable quencher. The non-selective cation TRP channel changes the membrane potential upon activation by Gαi, which enables the selective influx of the dye. 24 hours prior to the assay, 150'000 cells/well were seeded into clear-bottom 96-well plates and were reverse transfected with an optimized quantity of receptor DNA (present in the pcDNA3.1+ vector) and 0.5μL Lipofectamine 2000 per well. Prior to reading the transfected plates, the media was removed and the FLIPR dye was applied. The relevant drug was transferred into the plates during plate reading on a Flexstation3 multi-mode plate reader and changes in fluorescence (emission at 535nm, excitation at 565nm) were measured for a period of a maximum of 4 minutes. Controls were removed and maximum fluorescence values were reported as function of the logarithm of the drug concentration in GraphPad PRISM10.

3.4.3 G-alphaS BRET-EPAC cAMP assay

HEK-293T cells (gift from Prof. Ted Wensel at Baylor College of Medicine) were maintained in DMEM supplemented with 10% fetal bovine serum (FBS) and grown at 37°C and 5% CO2. Upon agonist stimulation of the dopamine receptor D1, GDP-GTP exchange will promote GαS dissociation from the Gβ and Gγ subunits. GαS will subsequently activate adenylyl cyclase, which will increase the concentration of cAMP in the cell. The well-characterized BRET sensor CAMYEL (cAMP sensor using YFP-EPAC-RLuc) based on the exchange protein directly activated by cAMP (EPAC) will change conformation and BRET ratios will decrease upon cAMP increase. 24 hours prior to the assay, 75'000 cells/well were seeded into white-bottom 96-well plates and reverse transfected with an optimized quantity of DNA to match WT levels as well as an optimized quantity of the CAMYEL biosensor using 0,5μL Lipofectamine 2000 per well. Right before reading, the media was removed and cells were washed once with HBSS and 40μL HBSS was added in each well. Coelenterazine h was added on top of the cells

and incubated for 5min. After a first read, drugs were added in each well and change in light emission was recorded using a Mithras2 multimode plate reader. Controls were removed and changes in BRET ratios were plotted as function of the logarithm of the drug concentration in GraphPad PRISM10.

3.4.4 Enzyme-linked Immunosorbent Assay (ELISA)

To ensure all receptor variants express at a level similar to the WT receptor control, ELISAs were performed against the 3xHA N-terminal tag present on each receptor variant. For each of the aforementioned cell-based assay, an accompanying ELISA plate was also reverse transfected in parallel using the same conditions as the main assay plate. On the day of the assay, the media was removed from the wells and the cells were fixed with a 4% paraformaldehyde (PFA) solution for 10 minutes. Fixation was followed by a 2% bovine serum albumin (BSA) solution incubation, anti-HA mouse primary antibody incubation, and an anti-mouse secondary HRP-linked antibody incubation. Each for a period of 1 hour with three PBS washes between each step. SuperSignal chemiluminescent substrate (Thermo Fisher) was added to each well and plates were incubated for 5 minutes before a luminescent reading on a Flexstation3 plate reader. Mock-transfected wells were used to determine and subtract the baseline signal from the remaining wells.

3.4.5 Ligand docking

Dopamine docking onto the solved dopamine D2 receptor active state structure (6VMS (132)) was accomplished by the established IPhold Rosetta protocol (309). Only the receptor chain was used during all steps; heteroatoms and additional chains were removed. The overall protocol consists of sequential coarse-grained docking coupled with structural relax, decoy clustering, high resolution docking, and ligand clustering steps. During the first docking and relax step 10,000 decoys are generated, of which the lowest 10% scoring are used in the subsequent structure clustering step to diversify target receptor conformation. High resolution docking was performed on the cluster centers of the largest 6 clusters, again generating 10,000 decoys for each cluster center model and only using the lowest 10% scoring decoys for subsequent analysis. An additional filter was added to only keep the lowest 50% scoring decoys in terms of interface energy. The remaining decoys were used for ligand binding mode clustering using a DBSCAN algorithm with ligand heavy atom coordinates as the input (308). The largest binding mode cluster was designated as the putative native binding mode and used for further analyses.

3.4.6 In silico mutagenesis

Dopamine D2 variant models were generated using RosettaMembrane (171), a Rosetta-based protein structure prediction software utilizing a Monte Carlo gradient descent energy mini-

mization algorithm enhanced with an anisotropic implicit membrane scoring functionality. The recently released active-state dopamine D2 receptor structure (6VMS (132)) served as the initial starting structure. All heteroatoms and non-receptor or G-protein chains were removed from the starting structure. Residues of interest were mutated and adjacent residues within 5 Å were subjected to alternating cycles of sidechain repacking and backbone relaxation through Rosetta's Monte Carlo-based energy minimization algorithm. 200 decoys were generated per design to ensure score convergence. The lowest scoring decoys were used for all subsequent analyses.

3.4.7 Molecular dynamics simulations

The starting structures used for MD simulations are: 1- BRC bound 6VMS with the last 20 residues of the C-terminal helix of Gαi and the sequence re-mutated back to WT, 2- the DA bound docked model based on 6VMS, 3- RIS bound 6CM4 (101), and 4- DA bound 7CKZ (306) with the last 20 residues of the C-terminal helix of Gαs. Mutant variants were generated based on the WT structures using RosettaMembrane (171). The receptor-ligand-helix complex was inserted in a $90 \times 90 Å^2$ POPC lipid bilayer solvated by 22.5 Å layer of water above and below the bilayer with 0.15 M of Na^+ and Cl^- ions using CHARMM-GUI bilayer builder (287; 283; 281). Parameters for the two ligands (dopamine and bromocriptine) were generated using CGenFF (288). Simulations were performed with GROMACS 2019.4 for DD2R simulations and GROMCAS 2020.5 for DD1R simulations (289; 290) with CHARMM36 forcefield (291) in an NPT ensemble at 310K and 1 bar using a velocity rescaling thermostat (with a relaxation time of 0.1 ps) and Parrinello-Rahman barostat (with semi-isotropic coupling at a relaxation time of 5 ps) respectively. Equations of motion were integrated with a timestep of 2 fs using a leap-frog algorithm. Each system was energy minimized using a steepest descent algorithm for 5000 steps, and then equilibrated with the atoms of the ligand-receptor-G-p helix complex and lipids restrained using a harmonic restraining force in 6 steps as shown in Tab. A.19. After constrained equilibration, 5 to 10 replicas of 250 to 400 ns were run for each system, where the first 50 ns of every simulation was discarded for equilibration of Cα RMSD, and the rest of the simulation was used for calculating statistics.

4 Computational study of designed dynamic receptor—peptide signaling complexes applied to chemotaxis

"There is nothing new to be discovered in physics now.
All that remains is more and more precise measurement"
— Lord Kelvin, 1897

Author contributions: This chapter is partially based on the following publication: Jefferson, R. E., Oggier, A., Füglistaler, A., Camviel, N., Hijazi, M., Rico-Villarreal, A. R., Arber, C., and Barth, P. (2023). Computational design of dynamic receptor—peptide signaling complexes applied to chemotaxis. *Nature Communications*, 14(1), 2875. M.H. developed AlloDy and ran the MD simulations and all related calculations.

In this chapter, we study the dynamic origins of designed flexible peptide-receptor complexes by combining clustering of flexible peptide conformations and allosteric pathway analysis. The main goal of the study is to develop a computational strategy for designing signaling complexes between conformationally dynamic proteins and peptides. The design strategy aims to address the challenges of engineering protein biosensors that sensitively respond to specific biomolecules, and the designs are able to elicit chemotaxis in primary human T cells. To complement the design approach and add a layer of explainability to the designs, we run MD simulations and couple that with AlloDy to explore the conformational space sampled by the WT peptide ligand and the designs, as well as relate this flexibility to allosteric signaling. In this chapter, we will include the introduction and a summary of the results from the paper, as well as the work that we contributed.

4.1 Introduction

Designing biosensors with arbitrary input and output behaviors is a grand challenge of synthetic biology. Current approaches focus on engineering binding to structurally well-defined

protein (310) and small-molecule chemical cues (311), and couple molecular recognition to synthetic optical reporters that are built-in modular biosensor scaffolds. While this strategy provides elegant solutions to the design of *in vitro* diagnostics, applications for *in vivo* detection and synthetic cell biology rely on coupling the molecular sensor to the precise activation and orchestration of complex intracellular signaling functions that often cannot be recapitulated *de novo*. Harnessing synthetic sensing to fine-tuned native signaling functions in a biosensor scaffold is limited by our poor mechanistic understanding of allosteric signal transduction and lack of techniques to rationally engineer these properties.

Computational approaches for the design of protein-protein recognition have produced a wide array of therapeutic proteins including potent inhibitors and vaccines mostly through the optimization of binding interactions between static protein surfaces (311; 312; 313). However, several classes of proteins including signaling receptors and peptides display high levels of conformational plasticity and binding of these molecules often involves large structural rearrangements through conformational selection and mutual induced fit (309; 314; 315; 316). The rational design of dynamic binding complexes remains particularly challenging and has not been reported to date.

Peptides mediate close to 40% of cell signaling functions through ubiquitous interactions with membrane receptors and soluble proteins (317; 318). Unbound peptide ligands are often partially disordered in solution, which challenges structure determination and the computational sampling of the vast space of peptide conformations. In contrast to rigid protein binders and small-molecule ligands, structural information on peptide binding is scarce and limits supervised training and validation of deep-learning (319; 165; 320) and physics-based (321) protein–peptide complex structure prediction approaches. Consequently, the mechanistic underpinnings of peptide-mediated functions remain also poorly understood.

A recent comparative genomics study of peptidergic GPCRs revealed important features of the peptide-GPCR network (322). Peptide-binding GPCRs typically involve larger binding cavities and ligand contact areas than receptors binding to small molecules. The peptidergic signaling network is often characterized by GPCRs sensing an array of peptide ligands and peptides capable of activating several receptors, which complicates the prediction of binding and signaling determinants. The specific receptor–peptide modeling and engineering problem is further complicated by the high flexibility of both receptor and peptide ligand, which often mutually adopt a new conformation through induced fit to reach the active state and initiate signal transduction (323).

In this study, we first develop a computational strategy for modeling the binding between flexible peptides and structurally uncharacterized proteins and designing signaling membrane receptors with high binding sensitivity to peptide ligands. To validate the approach, we create chemokine receptor–peptide pairs that elicit potent intracellular signaling in human cells and chemotactic responses in primary T cells. Lastly we carry out molecular dynamics simulations on the complexes to assess the conformational sampling of each of the designed

peptides, relate peptide conformations to allosteric transmission, and uncover mechanistic determinants of GPCR–peptide recognition and signaling.

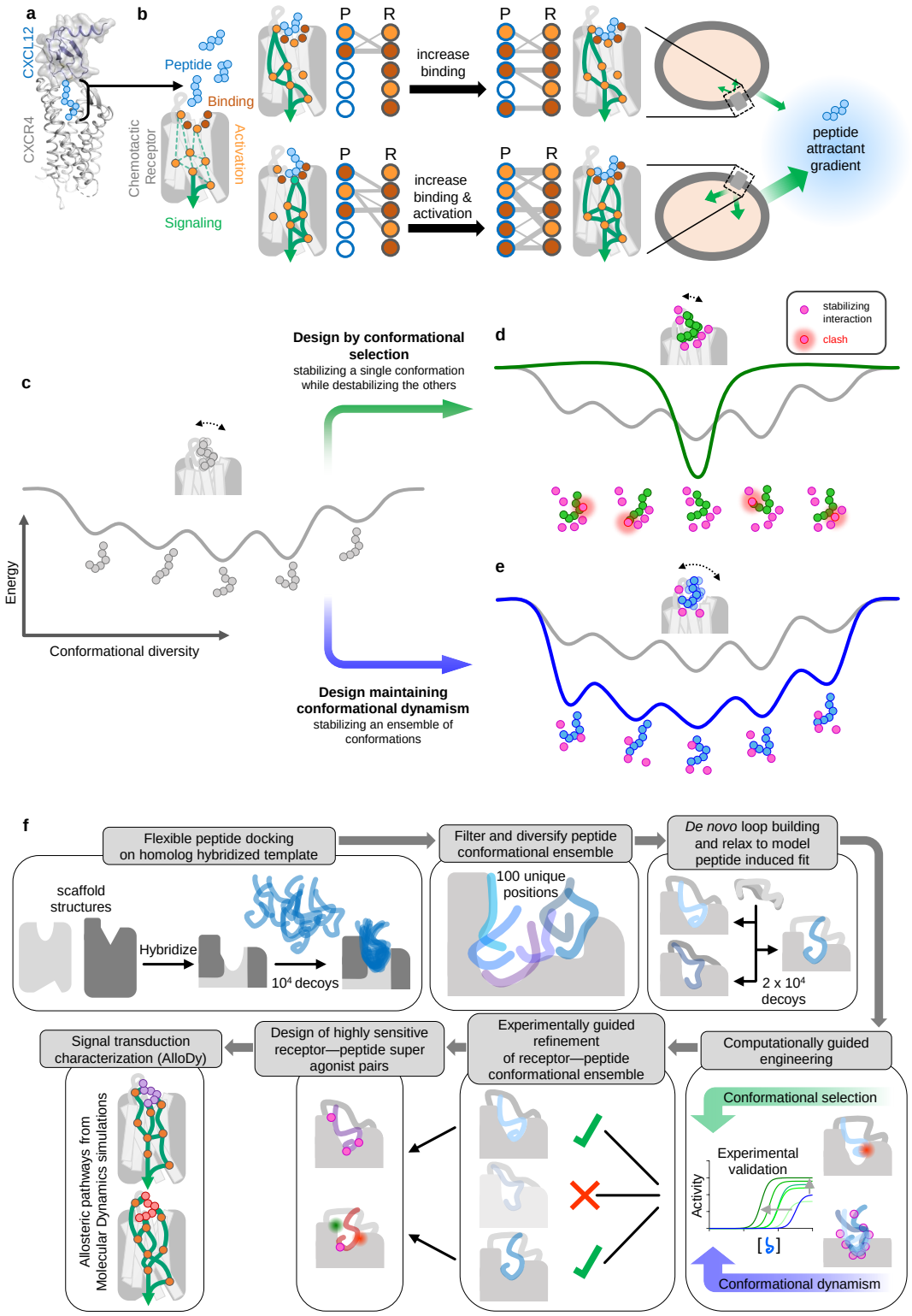
4.2 Results

4.2.1 Overall rationale and goal of the study

In the long run, we aim to design custom-built modular biosensors that can link binding of a flexible peptide input signal to fine-tuned and complex cellular responses through genetically encoded single-receptor domains. We define this designed class of biosensors as CAPSens, which stands for Conformationally Adaptive Peptide BioSensors. Such an approach would enable the reprogramming of cellular functions upon a wide range of environmental cues and would impact cellular therapies that rely on cell trafficking, including cancer immunotherapies.

Toward that goal, we developed a method that can build flexible receptor–peptide conformational ensembles and model peptide-mediated receptor signaling pathways. Unlike previous work that mainly optimizes binding and models receptors as rigid target structures (324), this approach enables the modeling of signaling active states and the design of dynamic complexes with altered binding contacts and allosteric networks enhancing both recognition sensitivity and signaling response (Fig. 4.1a, b).

To demonstrate this strategy, we targeted the chemokine receptor CXCR4–CXCL12 peptide signaling axis. We selected that signaling complex because CXCR4, upon sensing its native ligand CXCL12, regulates important physiological functions, including cell chemotaxis (i.e., cell migration along a gradient of CXCL12), but remains structurally uncharacterized in the active signaling state. Using the approach, we modeled and designed CXCR4 variants with high binding sensitivity to the native CXCL12 and also created receptor–peptide binding pairs that triggered potent signaling and cell migration (Fig. 4.1b).



(Caption on next page.)

Figure 4.1: **a** Cartoon representation of the CXCR4–CXCL12 complex (schematic model of receptor and chemokine structures aligned to CXCR2:CXCL8 complex structure. PDB IDs: 4RWS, 4UAI, 6LFO) targeted for the design of chemotactic receptors with enhanced binding and signaling responses towards peptide attractants. The CXCL12 chemokine ligand consists of a folded domain and a flexible 8 residue-long N-terminal tail (sequence: KPVSLSYR represented with light blue spheres) inserted into the receptor binding pocket. **b** The peptide ligand can adopt distinct bound conformations through specific contacts with receptor pocket residues that are classified as drivers of binding (red) or activation (orange). Dotted green lines correspond to putative allosteric signal transduction pathways running through the receptor. Plain green lines schematically represent the specific pathways engaged by each peptide conformation bound to the receptor. Through design, receptor–peptide connectivity (represented as an interaction graph between peptide (P) and receptor (R) residues) can be rewired to promote binding (top), activation, or both (bottom) to ultimately reprogram the cell migratory response. c–e General overview of the protein-peptide binding design strategies employed in the study. **c** Schematic view of a conformational energy landscape describing the binding of a flexible peptide to a receptor. The peptide is represented with 6 light gray spheres and adopts distinct conformations in each local energy minimum. **d** Design by conformational selection stabilizes one favored receptor–peptide conformation, while destabilizing others. Destabilizing interactions are represented as steric clashes. **e** Design to preserve dynamism selects amino-acid substitutions stabilizing multiple receptor–peptide conformations, hence maintaining conformational entropy at the binding interface. **f** Pipeline of the modeling & design strategy involving receptor–peptide modeling, rational design, experimental validation, refinement of receptor-peptide models, design of peptide super-agonists and analysis of allosteric signal transduction properties.

4.2.2 Computational modeling and design framework of GPCR–peptide signaling complexes

Despite tremendous progress in protein structure determination, experimental structures of signaling receptor–peptide complexes remain scarce. In absence of structures of the interacting partners, the design of binding complexes necessitates a method that both models the conformations of the bound molecules and engineers functional binding interactions. Molecular recognition between flexible peptide and signaling receptors likely involves significant structural rearrangements of both molecules through conformational selection (i.e., selection from an ensemble of unbound conformations) and induced fit (i.e., conformational changes occurring upon binding) effects. Therefore, we first reasoned that an effective method for modeling receptor–peptide structures should explore a vast conformational binding space, including the large ensemble of conformations explored by the flexible peptide but also the diverse receptor conformational changes triggered by peptide binding. We also hypothesized that maintaining a high level of conformational flexibility or dynamism at the binding interface may be critical for evolving complexes that optimize both peptide recognition and long-range allosteric response, necessitating interactions between multiple functional sites. Hence, to test this hypothesis, we sought to carry out and compare design calculations that either stabilize specific receptor–ligand bound conformations through conformational selection (Fig. 4.1c, d) or maintain high levels of conformational entropy by enabling the binding of a wide range of peptide conformations (Fig. 4.1c, e).

Our computational strategy was developed with these ideas in mind and proceeds in the main steps outlined in Fig. 4.1f). The first part is the protein modeling stage while the second part refers to the protein design stage.

The approach involves exploring a wide conformational binding space for both the flexible peptide and receptor, aiming to optimize peptide recognition and allosteric response. The computational strategy includes building receptor scaffolds, peptide docking, filtering and diversifying peptide-bound positions, loop remodeling, and finally relaxation of receptor–peptide complexes.

Two design approaches are presented: conformational selection for stabilizing specific conformations and a design that maintains conformational flexibility. The design process is experimentally validated, refined, and used to create super-agonist pairs. Finally, the allosteric properties of the designed complexes are characterized through molecular dynamics simulations with AlloDy.

4.2.3 Design of hyper-sensitive CAPSens for the native CXCL12 chemokine

We aimed to model and design peptide ligand agonists, with a focus on the N-terminal region of the chemokine CXCL12 that activates the CXCR4 receptor. We employed computational design strategies, including conformational selection (Csel) and maintaining conformational

flexibility (Cdyn), to optimize binding (we also termed a combined designed Csedy). In the first round of our conformational selection design (Csel1), we improved interfacial contact density between the receptor and peptide, leading to enhanced sensitivity to CXCL12 in cell-based assays (Fig. 4.2). Later on, we introduced a second binding motif, resulting in a substantially improved receptor (Csel2) with enhanced calcium mobilization and Gαi-coupling. While Csel1 is focused on increasing contact density in the binding pocket (and optimizing contacts with the three N-terminal residues of the peptide ligand), Csel2 adds contacts on the extracellular part, mainly contacting ECL2. We tested a total of 19 designs with a 37% success rate in achieving enhanced binding and signaling properties.

We then sought to design dynamic receptor-peptide complexes that could maintain high conformational entropy at the binding interface. We created a library of variant designs by rationally mutating predicted peptide binding and allosteric residues from the initial ensemble of receptor-peptide models. This library was tested for calcium mobilization and Gαi coupling, resulting in the identification of activating mutations on TM1, TM3, and ECL2. The Cdyn receptor variant was considerably more sensitive, with an almost 11-fold increase in Gαi potency and a 20% increase in efficacy compared to the starting CXCR4 WT scaffold. We then combined the binding hotspot motifs from Csel2 with the activating sites identified in Cdyn, resulting in the Csedy sensor with a more than 9-fold increase in Gαi response. The study demonstrated that our approach could design highly sensitive sensors for the WT CXCL12 chemokine-derived peptide, with Cdyn and Csedy achieving larger improvements in potency compared to the conformational selection approach, indicating that maintaining conformational flexibility had the potential to identify more effective binding interactions that trigger receptor activation.

4.2.4 Design of CAPSen chemotactic peptide super-agonist pairs

We aimed to create selective receptor-peptide pairs, focusing on designing peptide super-agonists for synthetic sensor-response systems. Our computational models identified key sites on the peptide scaffolds, and mutations at P7 and P3 significantly enhanced binding interactions with receptor designs (Csel1 and Csel2). These mutations led to increased Gαi efficacy and signaling potency in the designed sensors. Combining mutations at P3 and P7 produced the most substantial enhancements in both potency and efficacy. These results showcased the potential of our computational approach for engineering highly sensitive receptor-peptide pairs and suggested that the native binding interface may not be optimized for binding and signaling potency.

4.2.5 Designed receptor–peptide pairs enhanced human T cell chemotaxis

We tested our ultra-sensitive CAPSens to see if they could induce cell migration in response to chemokines. Chemotaxis relies on complex intracellular pathways governing receptor oligomerization, cell motility, and adhesion following G-protein activation by chemokines.

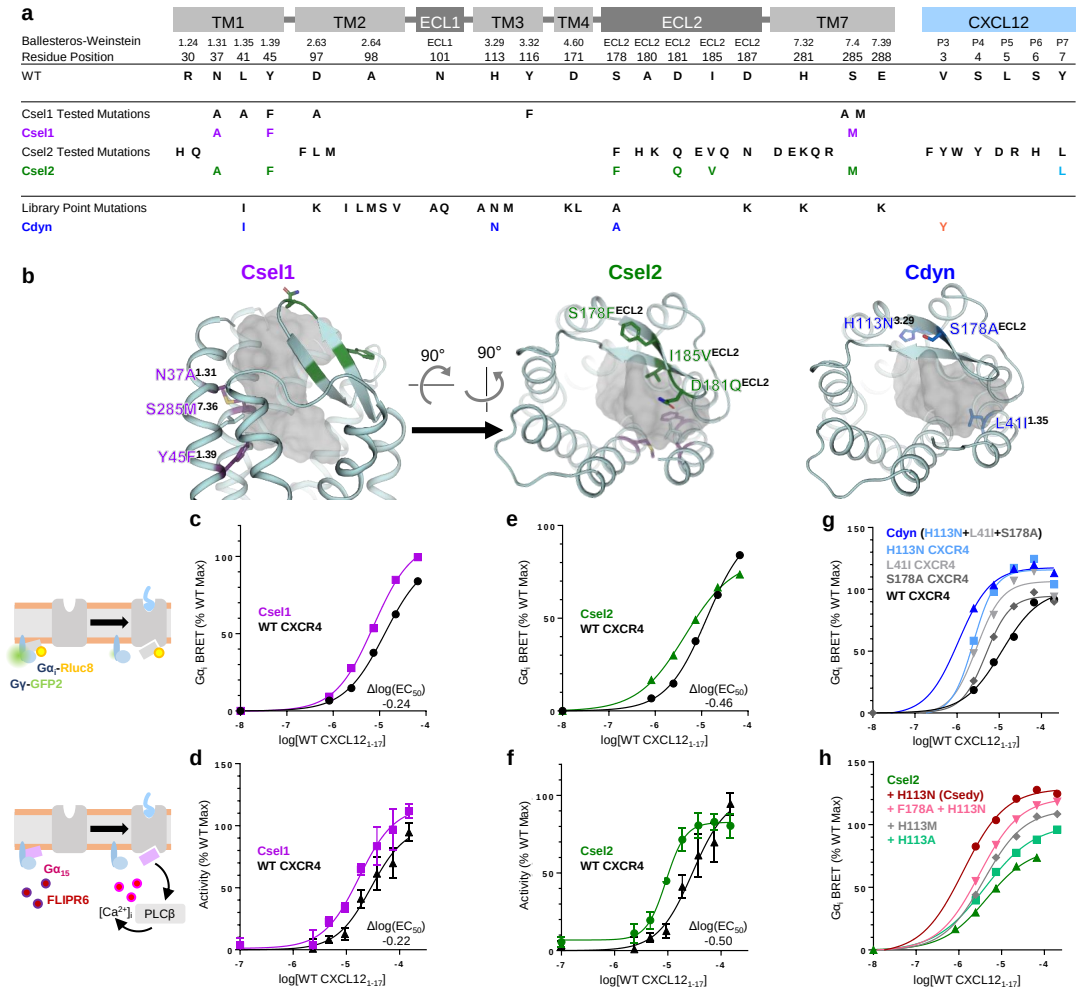


Figure 4.2: (**a**) Table describing the mapping on the receptor topology and numbering of receptor and peptide residues targeted for design. ECL and TM refer to extracellular loop and transmembrane helix, respectively. BW refers to Ballesteros Weinstein notation. Csel1, Csel2, Cdyn represent the sequences of the selected designs. (**b**) Location of the designed residues (shown in sticks) mapped onto the backbone structure of receptor peptide binding site (shown in cartoon). The WT CXCL12 peptide is represented as a gray-colored surface. (**c–h**) WT peptide-induced cell signaling responses of designed receptors measured through Gαi activation and calcium release: Gαi BRET of Csel1 design (mean, $n = 2$ technical replicates) (**c**), Csel2 design (mean, $n = 2$ technical replicates) (**e**), and library-screened mutations (mean, $n = 2$ technical replicates) (**g**). Calcium mobilization of Csel1 design (mean ± s.e.m., $n = 3$ technical replicates) (**d**) and Csel2 design (mean ± s.e.m., $n = 3$ technical replicates) (**f**). Gαi BRET of single-point library mutations in Csel2 design background (mean, $n = 2$ technical replicates) (**h**).

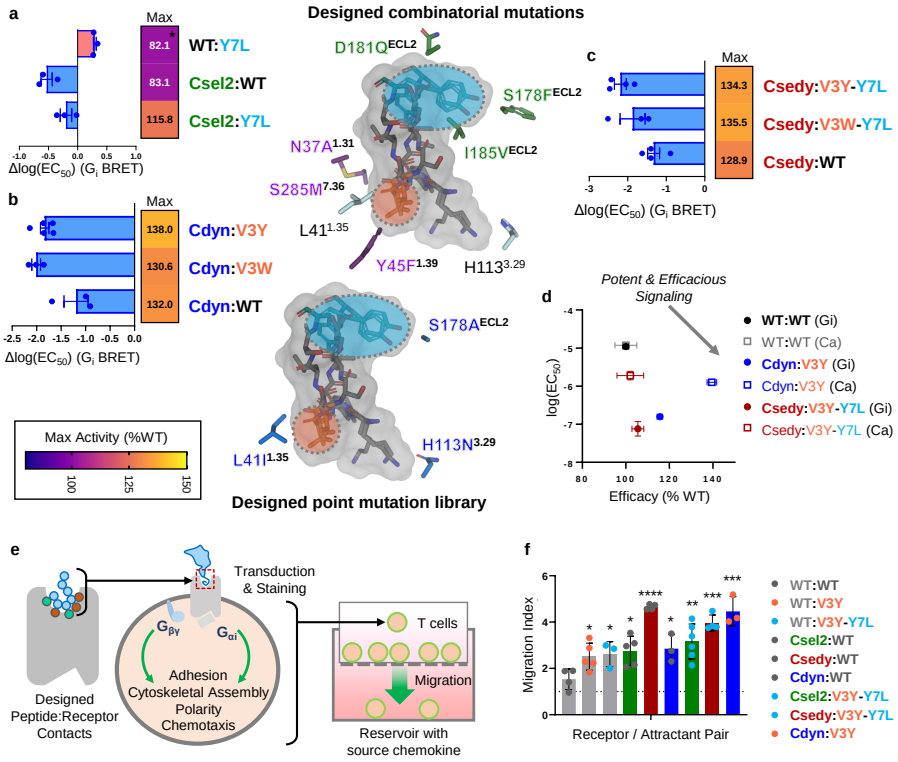


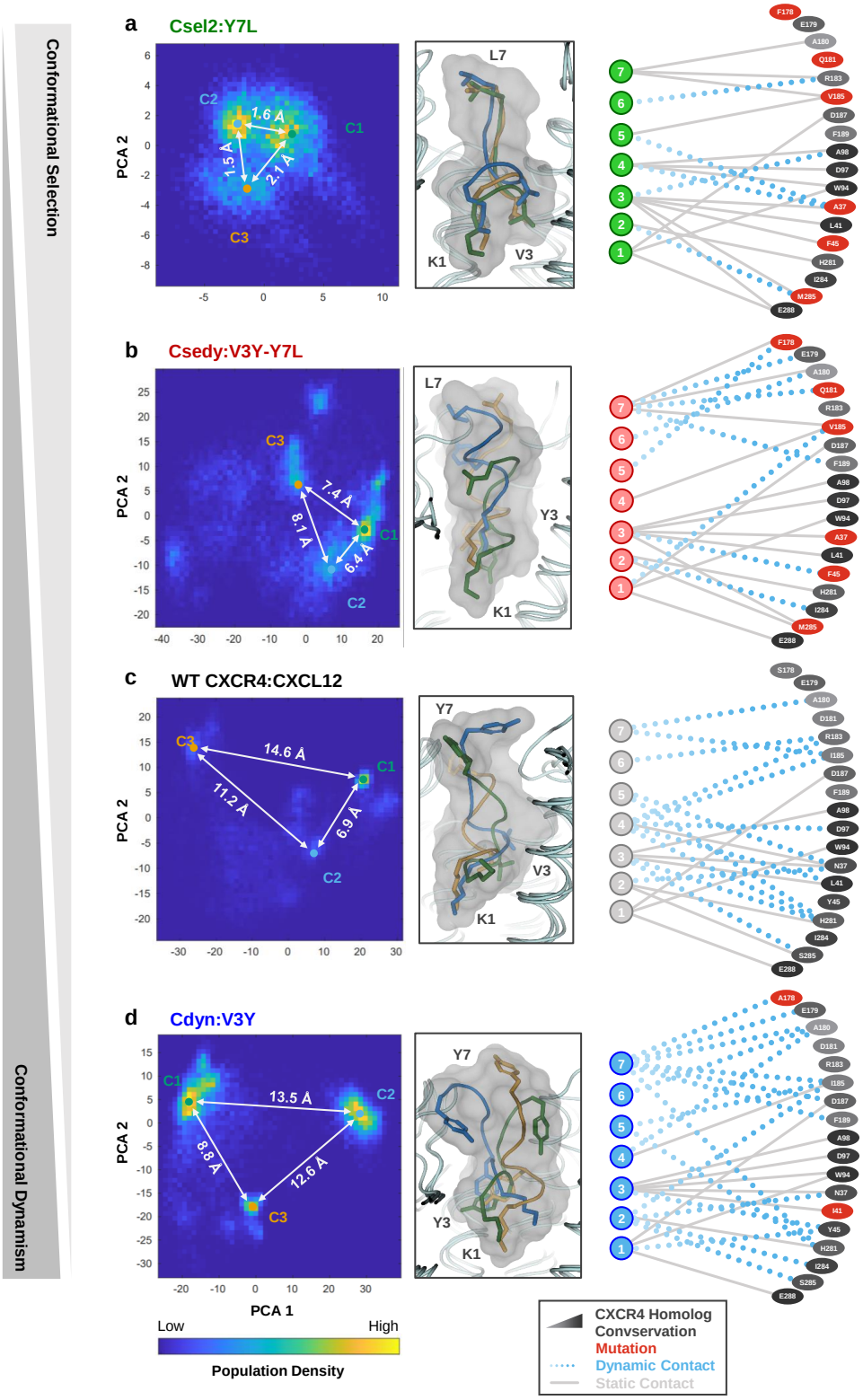
Figure 4.3: (**a-d**) Shifts in sensitivity (mean fitted value of dose-response curve fits ± s.e.m., $n = 5$ for Cdyn:V3Y, $n = 4$ for Csedy:V3Y-Y7L and Csedy:WT, $n = 3$ for all other pairs) and maximum activity (fitted value) for various receptor–peptide pairs involving the following designed peptides: a CXCL12 Y7L variant, (**b**) CXCL12 V3 substitutions, (**c**) CXCL12 V3Y/W-Y7L. (**d**) Changes in potency and efficacy across three separate experiments (mean ± s.e.m., $n = 3$ independent experiments). (**e**) Schematic of Boyden chamber migration assay of T cells transduced with engineered receptors and (**f**) migratory responses of transduced primary human T cells towards full-length chemokine. Bars are colored according to the transduced CXCR4 variant, and individual points are colored according to the CXCL12 variant (mean ± s.d., $n = 3$ for WT:V3Y-Y7L, Cdyn:WT, Cdyn:V3Y; mean ± s.d., $n = 4$ for WT:WT, Csedy:WT, Csedy:V3Y-Y7L; mean ± s.d., $n = 5$ for WT:V3Y, Csel2:WT; mean ± s.d., $n = 6$ for Csel2:V3Y-Y7L). Significance shown with two-sided unpaired t-test p values to WT:WT migration. $^{*}p \leq 0.05$, ** $p \leq 0.01$, $^{***}p \leq 0.001$, $^{****}p \leq 0.0001$ ($p = 0.0297$ for WT:V3Y, $p = 0.0334$ for WT:V3Y-Y7L, $p = 0.0141$ for Csel2:WT, $p < 0.0001$ for Csedy:WT, $p = 0.0205$ for Cdyn:WT, $p = 0.005$ Csel2:V3Y-Y7L, $p = 0.0001$ for Csedy:V3Y-Y7L, $p = 0.0008$ for Cdyn:V3Y)

This validation demonstrated our ability to engineer cell behavior in response to environmental cues. We transduced human T cells with designed sensors and observed significantly increased migration towards both WT and engineered chemokines. Our engineered CAPSens enhanced T cell migration by up to 4.7-fold (Fig. 4.3f), demonstrating that the designed signaling properties also affect cell functions and phenotypes. This approach focusing on the flexible peptide region of chemokines appears to be generalizable for designing biosensors responding to full-length chemoattractants, indicating its potential to enrich receptor-peptide binding interfaces for improved sensitivity and potency in receptor signaling.

4.2.6 Highly conformationally adaptive designed receptor–peptide binding interfaces through mutual induced fit

Our designed receptor–peptide agonist pairs offer a unique opportunity to uncover the structural and dynamics underpinnings of receptor–peptide binding and agonism. Experimental structures by X-ray crystallography or cryo-electron microscopy of our designs would only provide snapshots of the conformational ensemble and not reveal whether our designs achieved their functions through the intended modulation of the binding dynamics. Therefore, we decided to instead carry out molecular dynamics (MD) simulations of the designs to investigate the sequence-structure-dynamics relationships underlying their functions. Since the computational design was performed using knowledge-based potentials of the Rosetta software, MD simulations using molecular mechanics force fields provide an orthogonal validation of the design calculations.

Starting from the refined design models that best agreed with the experimental data, we run up to 1.9 microsecond long equilibrium MD simulations in explicit lipids (see Methods). Within this timescale, the peptide-bound receptor complex remains in the active state as assessed by the local conformation of consensus class A GPCR activation features such as interhelical (i.e., TM3-TM6 and TM3-TM7) distances on the intracellular side of the receptor and the RMSD for the NPxxY motif (Fig. 4.9). However, the simulations are long enough for the peptide and receptor to explore distinct bound conformations and enable a qualitative comparison of dynamic ensembles between variants (Fig. 4.4, Fig. 4.10).



(Caption on next page.)

Figure 4.4: Population density of ligand poses sampled in all-atom MD simulations are plotted in PC space (left panels). Inter-cluster RMSD of the most populated ligand conformations shown. Representative peptide poses are shown for the three most populated conformational substates of each variant (middle panels). Only sidechains for positions 1, 3, and 7 of the peptide are shown for clarity. The top 15 most frequent contacts from each substate are depicted schematically (right panels). Strong static contacts (solid gray lines) are prevalent in two or more substates, while weaker dynamic contacts (dotted blue lines) are unique to a single substate. Receptor–peptide variants are shown in increasing order of conformational dynamism of the complexes: **a** the Csel2:Y7L complex, **b** the Csedy:V3Y-Y7L complex, **c** the WT CXCR4—CXCL12 complex, **d** and the Cdyn:V3Y complex.

We analyzed the conformational space of the complexes using principle component analysis (PCA) of the MD trajectories. The analysis revealed the existence of distinct families (i.e., clusters) of conformations for the WT and all engineered receptor–peptide pairs (Fig. 4.4). The WT and Cdyn binding interfaces were characterized by high peptide ligand RMSD (up to 14.6Å distance between clusters of peptide conformations, Fig. 4.4a, b) (Tab. A.17) and suggest that the design approach was able to maintain high levels of peptide conformational diversity at the binding interface. On the other hand, the Csel2 design displayed substantially lower conformational diversity (only up to 2.1Å inter-cluster RMSD, Fig. 4.4c), consistent with that design strategy stabilizing a subset of the receptor–peptide structures through conformational selection. As expected for the hybrid design strategy, the V3Y-Y7L peptide explored an intermediate conformational space in the binding pocket of Csedy (up to 8.1Å, Fig. 4.4d). Overall, the different levels of peptide structural heterogeneity identified by the MD simulations are consistent with the intended modulation of the conformational space by our 2 design strategies. To rule out that the observed conformational heterogeneity results solely from potential inaccuracies in our predicted models, we carried out the same MD analysis starting from the experimental structure of the related complex between the N-terminal peptide of RANTES (325) and the CCR5 chemokine receptor. We observed a similar diversity in the conformations of the bound chemokine peptide (Fig. 4.5), suggesting that native chemokine receptors may actually bind agonist peptides with a significant degree of conformational dynamism.

We then analyzed in detail the network of binding contacts engaged by the distinct families of peptide conformations. Contacts were defined as dynamic if they were unique to one cluster or static if they were observed for at least two peptide binding modes. Throughout the MD trajectories, the peptide engaged with Csel2 through 16 strong static versus 5 weaker dynamic binding contacts (Fig. 4.4). The number of static contacts dropped to 12 and 11 while the dynamic ones raised to 11 and 13 in the Csedy and Cdyn complexes, respectively (Fig. 4.4). These observations further confirm that Cdyn and Csedy complexes involve a more dynamic binding interface than Csel2.

Conformational diversity was also noticeable on the receptor side and best quantified using a volumetric analysis of the peptide binding pocket. To simplify the analysis, representative

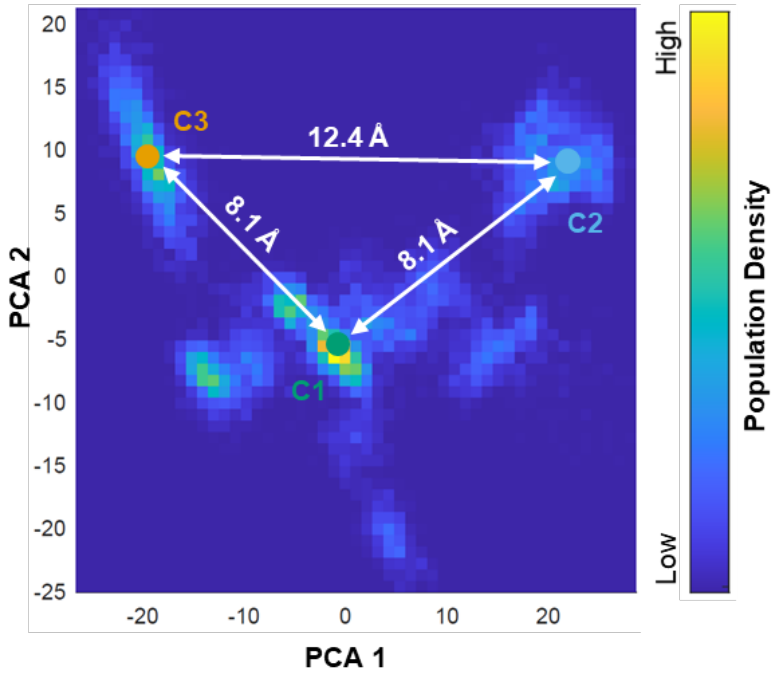


Figure 4.5: **Conformational flexibility of the ligand in the CCR5-RANTES complex.** Population density and cluster centers in PC space. Inter-cluster RMSD of the most populated ligand conformations shown. The highest density areas are colored yellow, while the lowest density areas are blue. The first 2 PCs explain 53.66% and 19.09% of the variability in the data respectively. Total simulation time is 1500 ns.

members of each cluster were selected and cross-sectional areas were calculated at different depths of the binding site. This analysis highlighted significant conformational adaptation of the binding surface (e.g., by up to 52% at a cavity depth of 10.25Å) in response to the different peptide conformations and sequences (Fig. 4.6a, b). When we mapped the distribution of the largest cluster of conformations (i.e., cluster C1) onto a 3D map of the structure-function relationship (Fig. 4.6c), we observed that the designed pairs occupy subspaces that are far apart in both receptor binding pocket and peptide conformation dimensions.

Overall, these findings suggest that the high conformational plasticity of the CXCR4–CXCL12 binding interface may facilitate the adaptation of contact networks in response to even limited changes in receptor and peptide sequence space through mutual induced fit. Although this analysis implies that higher conformational flexibility at the binding interface correlates with stronger signaling efficacy, it does not provide mechanistic insights into how such structurally distinct binding complexes could trigger potent signaling responses.

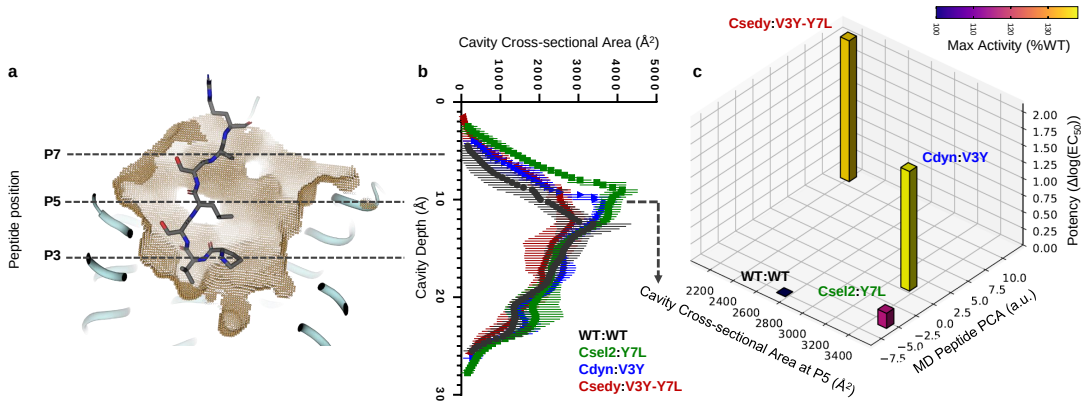


Figure 4.6: **a** WT CXCR4 ligand-binding cavity with depths marked for N-terminal residues P3, P5, and P7. **b** Cross-sectional area across cluster centers from molecular dynamics simulations. (mean ± s.d.) **c** 3D map of structure-function relationship. Activity shifts from WT of individual receptor–peptide pairs (z-axis for potency and bars colored according to maximal activity) are plotted as a function of conformational shifts of the peptide (y-axis: calculated by Principal Component Analysis on bound peptide ensembles (see Methods)) and conformational shifts of the receptor binding pocket (x-axis: calculated by cross-sectional area at the P5 depth, 10.25Å (see Methods)) for the center of the largest cluster of conformations.

4.2.7 Potent signaling achieved through substantially rewired but robust allosteric pathways

To address that question, we sought to investigate how peptide binding initiates signal transductions across the receptor. Since the inference of allosteric pathways using experimental approaches remains very challenging and would require extensive measurements by NMR spectroscopy, we relied on predictions from MD simulations coupled with AlloDy (AlloDy v1.0.0 was used in this paper) to carry out the analysis (Ch. 2, Fig. 4.8). Since MD simulations in this study are performed on peptide-bound receptor complexes in the active state, they do not carry out information on the transition of the receptor from inactive to active states, but inform on how the extracellular and intracellular sites communicate when the peptide is bound and the receptor occupies the active state. Within that framework, effective signal transductions should translate into strong allosteric pipelines running through the receptor structure and connecting the intra- and extracellular receptor sides.

As shown in Fig. 4.7, AlloDy identified several allosteric pipelines for the WT CXCR4 and designed CAPSens that flowed from the agonist peptide through a layer of receptor residues at the pocket interface, termed 'allosteric triggers', down to a conserved set of 'allosteric transmission hubs' located in the TM region away from the shell of ligand-binding residues. These transmission hubs include highly conserved class A GPCR motif residues W252$^{6.48}$ (W toggle) and N298$^{7.49}$ (NPxxY), as well as key highly conserved activation residues in CXCR4: F87$^{2.53}$, L120$^{3.36}$, H203$^{5.42}$, mutation of which has been shown to impair calcium mobilization

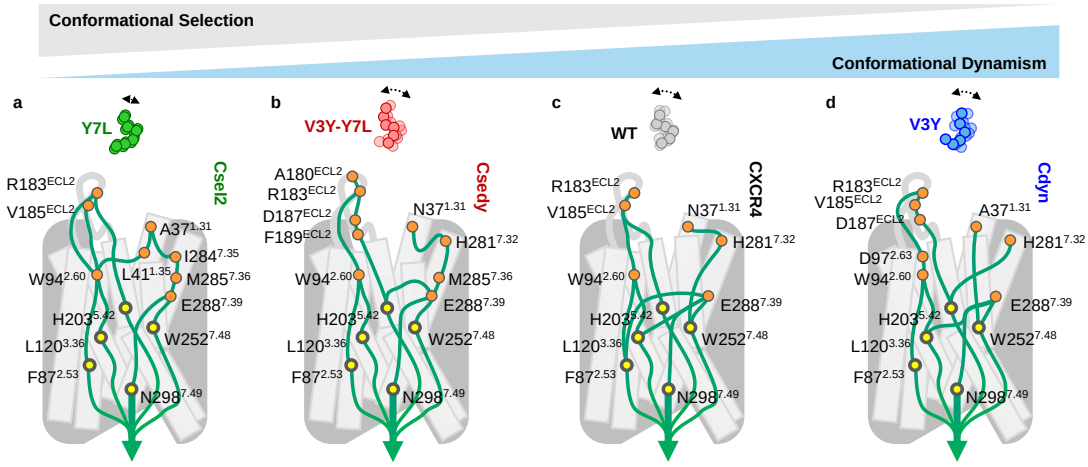


Figure 4.7: Signal transduction pathways calculated by AlloDy from peptide-contacting residues (orange circles). The allosteric pathways utilize **a** diversity of receptor pocket residues to communicate information, but ultimately propagate the activation signals through a common set of allosteric determinants in the core of the receptor (yellow circles: $F87^{2.53}$, $L120^{3.36}$, $H203^{5.42}$, $W252^{6.48}$, $N298^{7.49}$). (See Figs. S7–S10 for a complete mapping of the allosteric couplings.) Receptor–peptide variants are shown in increasing order of conformational dynamism of the complexes: **a** the Csel2:Y7L complex, **b** the Csedy:V3Y-Y7L complex, **c** the WT CXCR4–CXCL12 complex, and **d** the Cdyn:V3Y complex.

activity (326; 327; 328). Though the relative amounts of information passing through this conserved set of transmission hubs varies between the designs (Tab. A.18), the topology of the allosteric pipelines leading to G-protein signaling remain very similar between variants (Figs. A.13, A.14, A.15, A.16).

By contrast, how allosteric signals initiated by the peptide reach the common set of transmission hubs is highly dependent on the designed receptor–peptide interface. In fact, each variant utilizes a distinct set of allosteric triggers to connect to the activating allosteric pipelines (WT: $R183^{ECL2}$, $I185^{ECL2}$, $W94^{2.60}$, $N37^{1.31}$, $H281^{7.32}$, $E288^{7.39}$, see Fig. A.13; Csel2: $R183^{ECL2}$, $I185^{ECL2}$, $W94^{2.60}$, $A37^{1.31}$, $L41^{1.35}$, $I284^{7.35}$, $M285^{7.36}$, $E288^{7.39}$, see Fig. A.14; Cdyn: $R183^{ECL2}$, $I185^{ECL2}$, $D187^{ECL2}$, $D97^{2.63}$, $W94^{2.60}$, $N37^{1.31}$, $H281^{7.32}$, see Fig. A.15; Csedy: $A180^{ECL2}$, $R183^{ECL2}$, $V185^{ECL2}$, $D187^{ECL2}$, $F189^{ECL2}$, $W94^{2.60}$, $A37^{1.31}$, $H281^{7.32}$, $I284^{7.35}$, $M285^{7.36}$, $E288^{7.39}$, see Fig. A.16). Though there is some overlap in the allosteric triggers, the specific interactions with the peptide agonist that confer activation are unique across WT and the CAPSens and reflects the high diversity of binding contacts engineered at the binding interface.

Overall, while there is substantial dynamic rearrangement of activating contacts between variants (Fig. 4.7), they propagate the activation signals through a common set of allosteric determinants in the core of the receptor. These findings suggest that high conformational adaptability of the binding site together with robustness of the allosteric transmission layer are critical features for the evolution of signaling receptor–peptide pairs.

4.3 Discussion

Current approaches for the design of protein-protein complexes mostly follow the classical lock and key paradigm and optimize interactions between static binding surfaces. The lack of dynamic treatment of protein interactions hampers the effective design of complexes involving flexible proteins and ligands, which represent a large fraction of the molecules regulating cellular functions. In particular, peptide-binding receptors constitute one of the most abundant signaling systems in humans. However, these complexes have been particularly challenging to study and engineer partly owing to their high conformational flexibility. Here, we developed a computational framework for the design of dynamic protein complexes involving adaptable conformational ensembles of the molecules. We demonstrate that this approach is critical to achieve optimal sensing of flexible ligands and strong allosteric signaling responses that necessitate the interactions with multiple functional sites on the receptor surface.

Previous efforts to engineer high-affinity peptide ligands to the native CXCR4 receptor have mostly generated antagonists. Library-selected CXCL12-based antagonists(325) identified single N-terminal amino acid additions of Leu and Met that may antagonize CXCR4 in a similar mode as vMIP-II(329), IT1t(104) or Met-CXCL12(330), that intercalate between W94$^{2.60}$ and H113$^{3.29}$, a local hydrophobic inter-TM groove not occupied by our agonist peptide ensemble. These antagonists may lock W94$^{2.60}$ in an inactive conformation, preventing allosteric signal propagation through key downstream transmission hubs, suggesting that selecting solely for high-affinity receptor–peptide interactions is susceptible to conformational selection of inactive receptor states.

Our designs were able to explore regions of sequence space not enriched by such binding-selective approaches, suggesting that the computational method can explore and engineer alternative active states not commonly accessed by the WT receptor. Unlike most binding interfaces between globular proteins, our designs displayed considerable structural adaptation to sequence changes. Remarkably, the diversity of designed allosterically coupled residue networks at the ligand-binding pockets is large among variants, and even between conformational substates of the same variant. While the designed allosteric triggers still operate and funnel signals through the same set of conserved transmission hubs as WT, they considerably enhance signal transduction through optimally rewired dynamic couplings. Our CAPSens are capable of ultrasensitive responses, not by enriching a high density of strong contacts around a particular active conformational substate, but by preserving conformational dynamism, as observed in our MD simulations. Our findings support a receptor–peptide recognition model where conformational flexibility is essential for the bound molecules to engage a multitude of functional interactions triggering effective allosteric responses. In this model, high levels of conformational entropy enable the shifting of active state ensembles and the rewiring of allosteric coupling via contacts not commonly accessed by the WT complex (Fig. 4.7, Figs. A.13, A.14, A.15, A.16). As such, the high conformational adaptability of the native CXCR4–CXCL12 binding interface is critical in accommodating and rewiring allosteric entry points to the transmission layer. Overall, our study suggests that the combination of a flexible

sensing layer coupled with a robust signal transmission layer may be a common hallmark of GPCRs, providing potential mechanistic insights into the high evolvability of sensing and signaling properties in this receptor family. While our computational findings are consistent with the experimental results, the MD simulations were performed on design models and not experimental structures. Hence, we cannot rule out that the precise details of the simulated binding complexes may be affected by inaccuracies in the starting structural models.

This work was started before AlphaFold2 was released. A recent study (331) indicates that AlphaFold can predict peptide–protein interactions despite not being trained on this task, suggesting that the main features of peptide binding can be implicitly captured as an extension of folding. Our method is geared towards modeling flexible peptide interactions which do not involve strong patterns of unique static contacts such as those characteristic of folded polypeptide chains. Therefore, our study should provide a complementary and useful approach to neural network based methods trained on protein folded structures.

In the long run, we expect that designed chemotactic signaling systems should prove useful in a wide variety of therapeutic contexts. Chemotactic peptides are attractive targets since directional movement of cells in response to gradients of these molecules (i.e., chemotaxis) is essential throughout biology and control over cell migration represents a key challenge in synthetic cell biology. For example, efficient immune cell homing to and into cancers is one of the main bottlenecks in modern immunotherapy (332; 333; 334; 335; 336). Hence, these therapeutic approaches would benefit from engineered cytotoxic lymphocytes with enhanced chemotaxis toward tumor sites. Overall, our results suggest that engineered receptors could trigger migration towards cancer-prone sites at longer distances with shallower chemokine gradients when compared to native chemotactic systems. Our designed CAPSen:hyper-agonist peptide pairs open the door to bringing cell migration under exogenous and spatiotemporal control, providing a promising synthetic cell biology tool.

Most biosensor design approaches have focused on engineering protein domains for optimal recognition of structurally well-defined molecules. Previous studies have repurposed designer receptors exclusively activated by a designer drug (DREADDs) to elicit chemotaxis towards the small molecule clozapine-N-oxide (CNO) (337), but the direct in vivo application of this approach is limited by the delivery of CNO and the inherent lack of utility as a gradient-generating homing molecule. Our CAPSens exhibit some degree of orthogonality in the Csel2:Y7L pair, and future iterations could be developed as genetically encodable orthogonal receptor–peptide pairs allowing for biological expression of the homing signal by cells that would enable synthetic transmitter-receiver cell systems and precise spatiotemporal control of cell homing. By targeting flexible and structurally uncharacterized peptides, our design platform significantly expands the range of molecules that can be detected by biosensors. Unlike approaches that rely on multi-domain sensor reconstitution upon ligand sensing, our method optimizes the coupling between molecular recognition and allosteric response in a single protein domain within the restricted design space of the ligand pocket interface and can generate CAPSens with strongly enhanced dynamic and sensitive responses. Carving

biosensors into versatile GPCR scaffolds offers key additional advantages. GPCRs can now be engineered to trigger a wide range of intracellular functions through reprogrammed coupling to diverse effectors including G-proteins and arrestins (338; 339). Alternatively, inserting fluorescent protein domains into GPCR scaffolds enables fast and direct optical detection of ligand molecules (340). As such, our approach lays a foundation for a wide range of synthetic biology, diagnostics, and therapeutic applications that would benefit from sensor systems that trigger complex cellular outputs or enable direct highly sensitive detection of chemical cues.

4.4 Methods

The parts of the methods that are not directly related to my contribution, which include experimental details and flexible peptide design, are left to the reference (297).

4.4.1 Molecular dynamics (MD) simulations

The final selected models for CXCR4: WT:WT, Cdyn:V3Y, Csel2:Y7L, and Csedy:V3Y-Y7L complexes were used as starting input poses for MD simulations. CCR5-RANTES simulations were started from the 11 N-terminal residues of RANTES bound to the receptor extracted from the active state structure of the complex (PDB: 7F1R). The receptor-ligand complex was inserted into a regular hexagonal POPC lipid bilayer with 90 Å perpendicular distance between any parallel sides and solvated by 22.5 Å layer of water above and below the bilayer with 0.15 M of Na+ and Cl- ions using CHARMM-GUI bilayer builder (287; 283; 281). Simulations were performed with GROMACS 2020.5 (289; 290) with CHARMM36 forcefield (291) in an NPT ensemble at 310K and 1 bar using a Nosé–Hoover thermostat (independently coupled to three groups: protein, membrane, and solvent with a relaxation time of 1 ps for all three) and Parrinello-Rahman barostat (with semi-isotropic coupling at a relaxation time of 5 ps respectively). Equations of motion were integrated with a timestep of 1 fs for the first three steps of equilibration and then 2 fs using the leap-frog algorithm. Each system was energy minimized using the steepest descent algorithm for 5000 steps, and then equilibrated with the atoms of the ligand-receptor complex and lipids restrained using a harmonic restraining force in 6 steps (Tab. A.19). After constrained equilibration, 5 to 7 independent trajectories of 200 or 300 ns (Tab. A.20) were run for each system. The first 50 ns of the simulations were discarded as the time needed for the system to equilibrate, as shown by the Cα RMSD of the receptors and the ligands. The total simulated time was defined to ensure convergence of the 1st and 2nd order entropies calculations in the top three PCA clusters in every system (see sections below).

4.4.2 Principle component analysis (PCA) of bound peptide conformational ensemble

All MD trajectories sampled the receptor active state as assessed by the structural distribution of consensus class A GPCR activation features such as the TM3-TM6 and TM3-TM7 interhelical distances on the intracellular side of the receptor, except for the C2 and C3 substates of the variant Cdyn for those we also observed a minor alternative minimum not representative of a true active state (i.e. distinct from the well containing the experimental active state structures). The frames corresponding to these minor populations were filtered out to ensure that subsequent conformational analysis were truly reflecting receptor active states. This filtering process yielded 999, 1109, 1277, and 1057 ns of simulated time for subsequent analysis of WT:WT CXCR4, Csel2:Y7L, Cdyn:V3Y, and Csedy:V3Y:Y7L, respectively. PCA was performed on the Cartesian coordinates of Cα and Cβ atoms of peptide ligands from receptor—peptide conformations selected by combining molecular dynamics trajectories from each of the studied systems. Representative models from the molecular dynamics trajectories were chosen as the highest density points in the space of principal components (PCs) 1 and 2. The first 2 PCs explain 43.7% and 19.0% of the variability of the data, respectively. PCA was also performed individually for each of the systems studied with MD on the Cartesian coordinates of Cα of peptide ligands. The PCA space was then clustered using a k-means clustering algorithm, with the optimal number of clusters being evaluated by the Calinski-Harabasz criterion. The variability explained by the first 2 PCs is shown in Tab. A.21. For each cluster, contact frequency between receptor and peptide residues was calculated as the percent of frames for which a heteroatom of a given receptor residue is within 5 Å of a heteroatom of a given peptide residue.

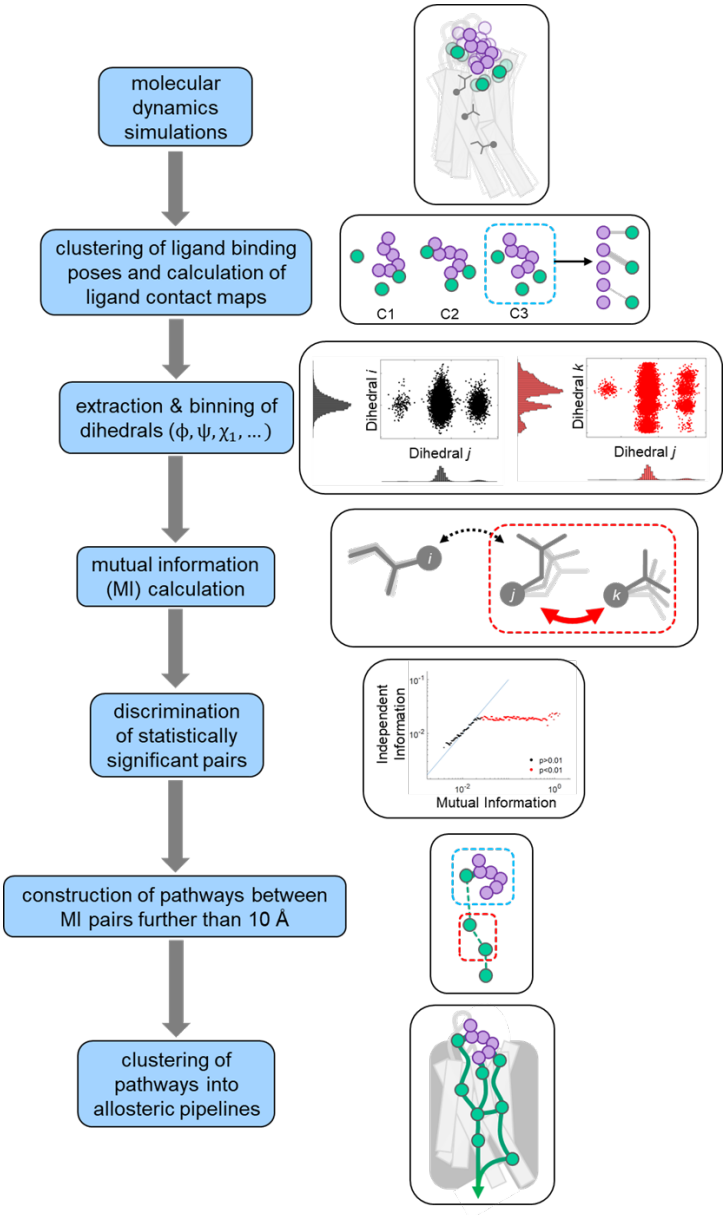


Figure 4.8: **Schematic of the steps performed by AlloDy v1.0.0 to extract correlated motions from simulations and cluster them into allosteric pipelines.** Ligand binding poses are clustered from molecular dynamics simulations using principal component analysis. A particular ligand pose is highlighted in the blue box with contact mapping to the right. Then dihedrals are extracted for every cluster separately. For every cluster, mutual information (MI) is calculated with finite size corrections and statistically filtered to remove uncoupled pairs (black) and include highly coupled pairs (red) before being summed over residue pairs. Pathways are then constructed using a shortest distance algorithm between residue pairs (red box) that have significant MI and are more than 10 Å apart stemming from ligand contacts (blue box). Constructed pathways are clustered into allosteric pipelines.

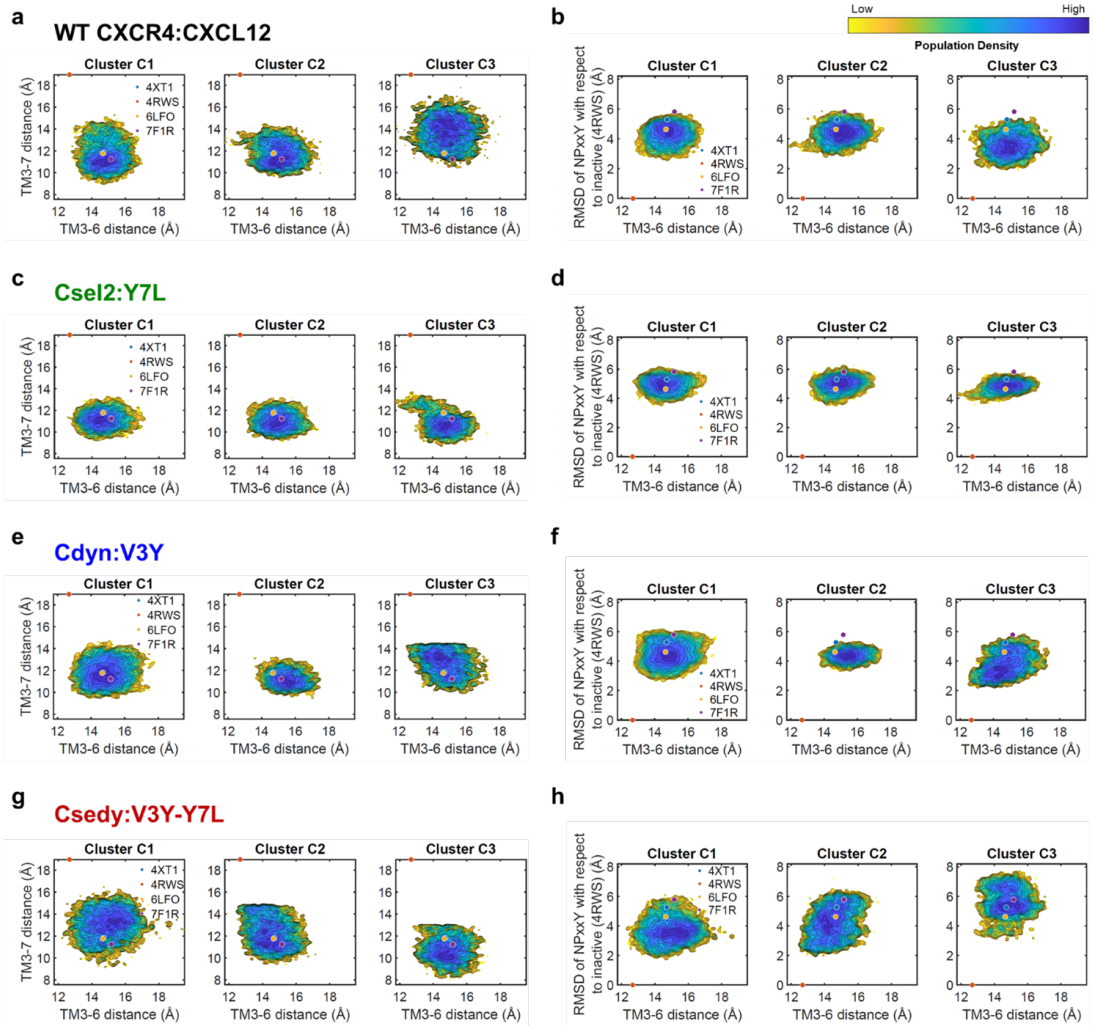


Figure 4.9: **Activation landscapes of molecular dynamics simulations.** Landscapes representing the density of frames in the activation conformational space described by interhelical distances of TM3-6 and either TM3-7 on the intracellular side of the receptor (left) or the RMSD of the NPxxY motif from the inactive reference structure (right) for (a,b) CXCR4 WT:WT, (c,d) Csel2:Y7L, (e,f) Cdyn:V3Y, and (g,h) Csedy:V3Y-Y7L. Distances were calculated between alpha carbons of residues R3.50, K6.30, and Y7.53 for TM3, TM6, and TM7 respectively. Landscapes were calculated in a fashion similar to 2D potential of mean force, with densities defined by kernel density estimates with gaussian functions. The color bars are in arbitrary units, where blue (lowest quantity) represents the highest density of frames, and yellow (highest quantity) represents the edge of the populated space. Experimental structures of inactive CXCR4 (PDB ID: 4RWS) and active US28, CXCR2, CCR5 (PDB IDs: 4XT1, 6LFO, and 7F1R, respectively) chemokine receptor structures are shown for reference.

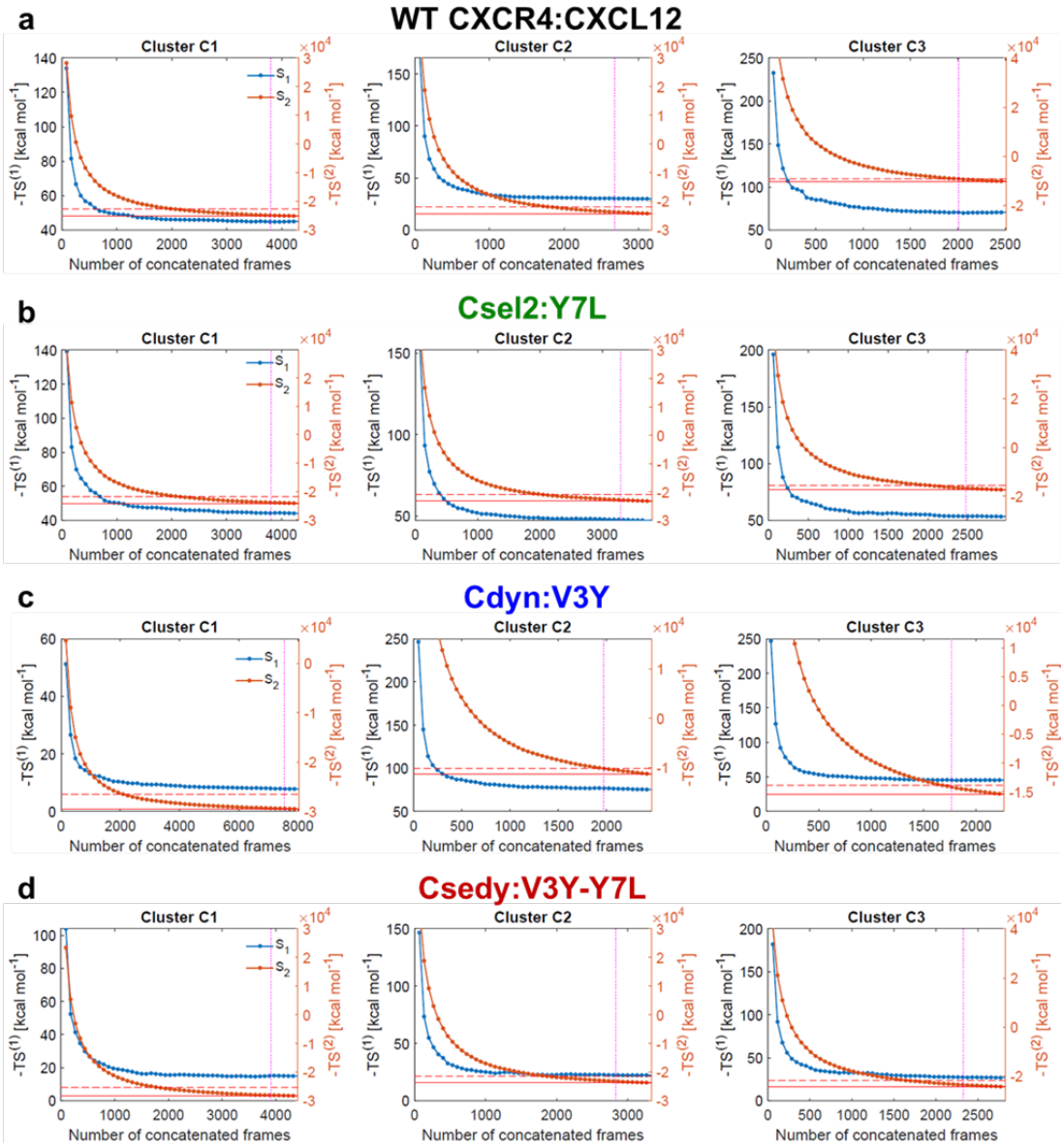


Figure 4.10: **Convergence of 1st and 2nd order entropies for WT and CAPSen receptor—peptide complexes.** Convergence plots of 1st order (blue, left Y-axis) and 2nd order (red-brown, right Y-axis) entropies as a function of number of frames in every cluster for (a) CXCR4 WT:WT, (b) Csel2:Y7L, (c) Cdyn:V3Y, and (d) Csedy:V3Y-Y7L. 1st order entropy is defined as the sum of marginal entropies of all individual dihedrals (ϕ, psi, and χ's) and 2nd order entropy is the sum of joint entropy of dihedral pairs formed by the top 300 dihedrals with the highest summed MI. 1st order entropies converge much faster than 2nd order ones, so we consider 2nd order entropies for the convergence criterion. Entropies are considered converged when there is less than 10% variability from the final entropy value over the last 500 frames (50 ns) considered. The 10% limit is represented by the dashed horizontal red line. The final entropy value is marked by the solid horizontal red line. The last 500 frames are marked with the vertical dotted magenta line.

5 Integration of genetic variation and allostery in class A GPCR signaling

"The capacity to blunder slightly is the real marvel of DNA. Without this special attribute, we would still be anaerobic bacteria and there would be no music."
— Lewis Thomas

Author contribution: M.H. managed the project, chose research direction, prepared MD simulations with all subsequent allosteric analysis, and wrote the chapter. This work has been done in collaboration with a summer research program student at the lab, Mariia Minaeva.

5.1 Introduction

The ability to predict the effects of amino acid substitutions on protein stability and function holds paramount significance in molecular biology, drug discovery, evolutionary biology, and disease research. Amino acid substitutions, arising from genetic mutations, can substantially influence the biological properties of proteins (as we have seen throughout this work), thus impacting cellular processes and organismal health. Consequently, the prediction of these effects has garnered considerable attention. This chapter explores the relationship between amino acid substitutions, protein function, and allosteric metrics.

While recent advancements, such as AlphaMissense (341), have made great strides toward predicting whether a mutation would be benign or pathogenic, the analysis we present here attempts to understand the effect of genetic variation in the light of protein function (and specifically, allostery). To this end, we predict evolutionary scores using GEMME (342) and allosteric scores using AlloDy for dopamine receptors D1 and D2, and β2-adrenergic receptor. We then looked at occurrence of single nucleotide single (SNVs) within allosteric hotspot positions. Finally, we perform a comparison between predicted GEMME scores and reported

deep scanning mutagenesis functional outcomes for β2AR, relating the reported clusters of "tolerance to mutations" to evolutionary scores (183).

5.2 Results

5.2.1 Relationship between evolutionary scores and allosteric scores in dopamine receptors

As a first level of the analysis, we plot allosteric scores from AlloDy (σ_m) and GEMME scores (averaged over all possible amino acid substitutions at any given residue) on the same axis for DD1R (Fig. 5.1a) and DD2R (Fig. 5.2a) to get a general overview of both distributions. We have one allosteric score per residue to the lack of feasibility of running MD simulations for every mutation, while GEMME scores span every possible amino acid substitution at every residue, thus we average GEMME scores when comparing them with allosteric scores at any given position. Visually, there is little overlap between the two sets, this is confirmed by the very weak (if any) correlation between σ_m and averaged GEMME scores (Fig. 5.1b and c, Fig. 5.2b and c). Note that disordered regions of the receptors (long ICL3 in DD2R and C-terminal tail in DD1R) have not been simulated and thus have no allosteric score. While the lack of correlation is no proof of the independence of the two measures, it hints that these two metrics describe different aspects of the system being studied.

5.2.2 Single nucleotide variants (SNVs) and allosteric residues

To further investigate the role of allosteric residues, we extracted missense variants from The Genome Aggregation Database (gnomAD) v3.1.2 (343) and mapped them along with allosteric scores σ_m in Fig. 5.3. We observe higher count of amino acid substitutions in disordered regions of the receptors (although it is good to keep in mind that substitution counts are small, ranging between one and four at a given position). Interestingly, there is little overlap between the SNVs and allosteric hotspots. A possible explanation is that allosteric hotspots are critical for function and that any variation in these positions will not be tolerated.

In addition, we looked into the variants reported in ClinVar (344), a public archive of interpretations of clinically relevant variants. In the case of DD2R, most of the missense variants in Clinvar are in loop regions (ICL1, ICL3, and ECL2), while the few positions in TM region have an allosteric strength of zero (for DA-bound DD2R simulations) with the exception of SNV V5.72I, which had significant allosteric strength (Tab. 5.1).

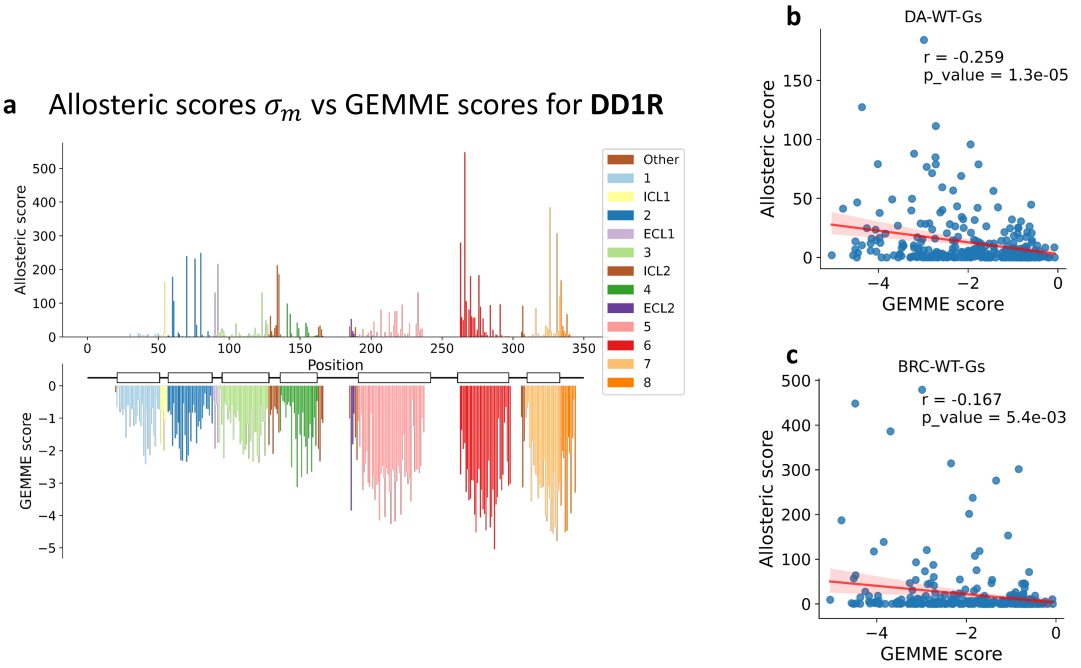


Figure 5.1: **Allosteric scores and averaged GEMME scores are not correlated for DD1R: (a)** Allosteric scores from AlloDy σ_m for DA-DD1R bound to a positive allosteric modulator (top) and GEMME scores (averaged over all amino acid substitutions) plotted against residue numbers. Rectangles and lines show TM helices and loops. **(b)** Correlation between allosteric scores σ_m for DA-bound DD1R simulations and and averaged GEMME scores. **(c)** Same as panel **b** but with BRC-bound DD1R simulations.

Protein change	Clinical significance	dbSNP ID	Protein region	BW	Allosteric score
A410T, A381T	Benign	rs758683320	TM7	7.37	0
A64V	Likely benign	rs201137518	ICL1	NA	9.34806
G261R	Uncertain	rs1415830775	ICL3	NA	NA
H287P, H316P	Uncertain	rs1182082677	ICL3	NA	NA
K327E, K298E	Benign	rs71653614	ICL3	NA	NA
N176D	Uncertain	rs776148708	ECL2	NA	11.72111
P271H, P300H	Uncertain	rs765357874	ICL3	NA	NA
P310S, P281S	Conflicting	rs1800496	ICL3	NA	NA
Q337K, Q366K	Uncertain	rs779477138	TM6	6.28	0
R245Q, R274Q	Uncertain	rs200184730	ICL3	NA	NA
R294W, R265W	Uncertain	rs758884516	ICL3	NA	NA
S311C, S282C	Benign	rs1801028	ICL3	NA	NA
V154I	Uncertain	rs104894220	TM4	4.44	0
V223I	Uncertain	rs764968856	TM5	5.72	143.1067

Table 5.1: **Clinical missense variants for DD2R:** protein change shows the sequence for long and short isoforms of DD2R. BW represents generic class A numbering. Allosteric scores were calculated from DA-bound DD2R WT simulations.

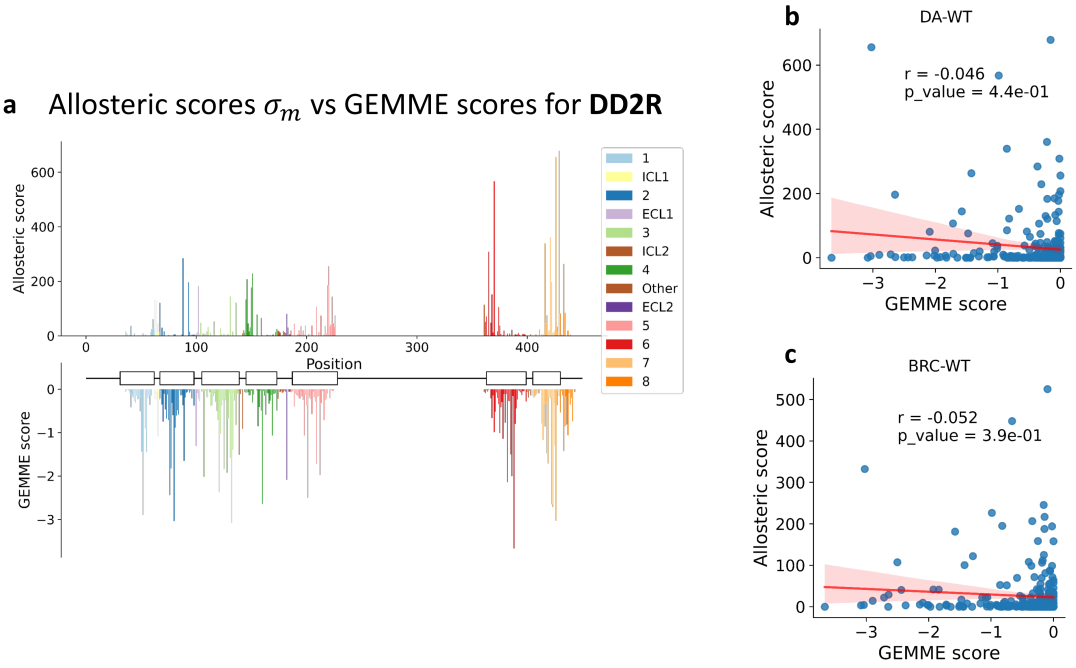


Figure 5.2: **Allosteric scores and averaged GEMME scores are not correlated for DD2R: (a)** Allosteric scores from AlloDy σ_m (top) and GEMME scores (averaged over all amino acid substitutions) plotted against residue numbers. Rectangles and lines show TM helices and loops. **(b)** Correlation between allosteric scores σ_m for DA-bound DD1R simulations and and averaged GEMME scores. **(c)** Same as panel **b** but with BRC-bound DD2R simulations.

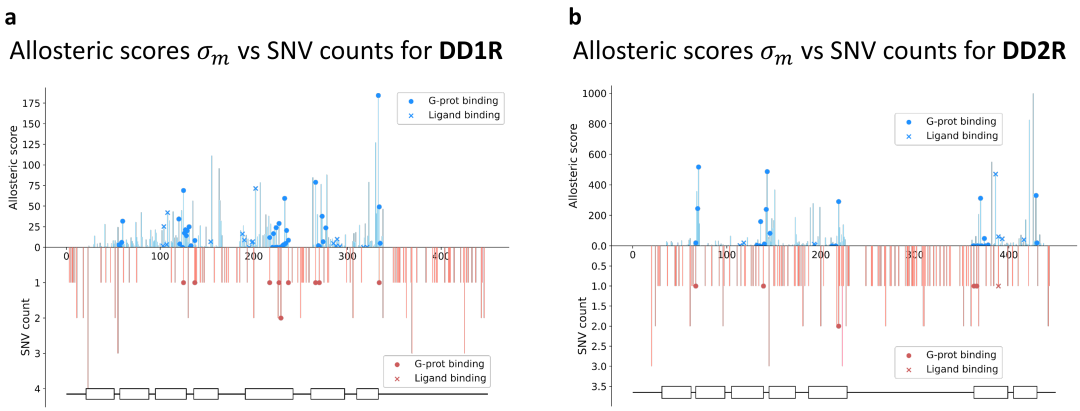


Figure 5.3: **Single nucleotide variant (SNV) count (red) and allosteric scores σ_m (blue):** for **(a)**DD1R and **(b)** DD2R. SNVs were extracted from The Genome Aggregation Database (gnomAD) v3.1.2 by filtering missense variants.

5.2.3 Relationship between evolutionary scores, allosteric scores, and function in beta2-adrenergic receptors

The next frontier in this analysis is comparing evolutionary couplings and allosteric scores with functional data. Jones et al. (183) have reported deep scanning mutagensis data for β2AR in response to isoproterenol (also known as isoprenaline) for cyclic AMP (cAMP) dependent pathway (one of the main signaling modalities of Gs-coupled receptors). They then applied dimensionality reduction followed by clustering to unveil functionally relevant groups of residues, which were divided into six clusters. These clusters were interpreted functionally according to mutation sensitivity, ranging from globally intolerant to charge sensitive to tolerant residues, as seen in Fig. 5.4a.

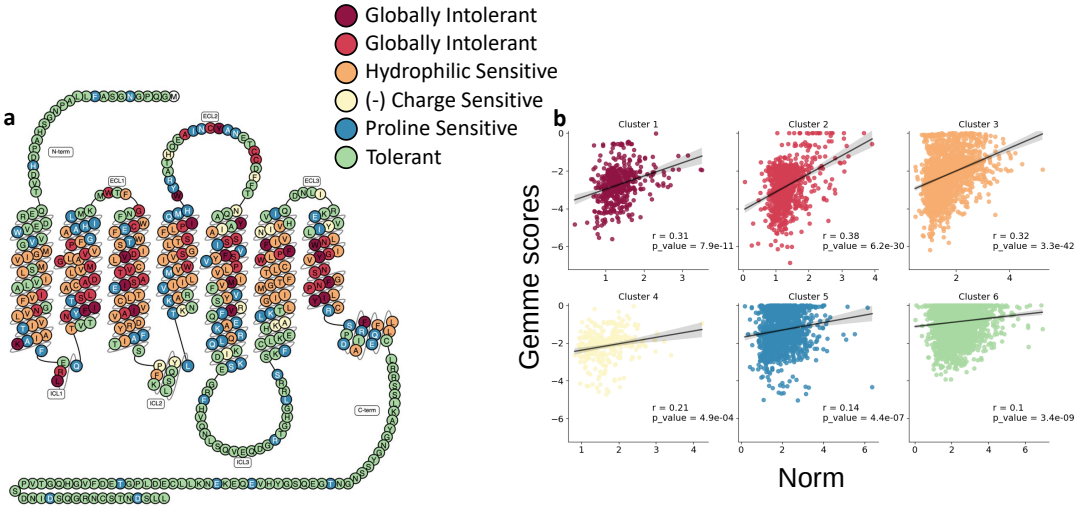


Figure 5.4: **Intolerant β2AR residues correlated better with evolutionary scores: (a)** Snake plot of β2AR where residues are colored according to their clustering from deep scanning mutagenesis study. Figure was recreated using data from Jones et al. (183). **(b)** Correlation plots of deep scanning mutagensis data (Supplementary file 2, average activity (183)) with GEMME scores show stronger correlations for less intolerant substitutions and the opposite for tolerant ones.

We hypothesized that the correlation between evolutionary scores and functional readout will depend on the cluster: the more intolerant a residue is to mutation, the higher the correlation of evolutionary couplings to functional outcomes will be. We test this by calculating linear fits between GEMME scores and functional readout for every amino acid substitution at each of the residue positions reported in Jones et al. (Supplementary file 2, average activity). We find that, as expected (Fig. 5.4b), tolerant and proline sensitive positions have low correlation coefficients (R^2 = 0.1 and R^2 = 0.14 respectively), charge sensitive cluster has a middling correlation coefficient (R^2 = 0.21), and the rest of less tolerant clusters have higher correlation

coefficients (R^2 = 0.32 for hydrophilic sensitive, and R^2 = 0.38, R^2 = 0.31 for the globally intolerant clusters).

In the space of GEMME and allosteric scores (Fig. 5.5a), positions with high allosteric and conservation scores are more prevalent in intolerant clusters (clusters 1 to 3), while residues with high allosteric score and low conservation are more prevalent in tolerant clusters (clusters 5 to 6). This allows us to divide important residues into functionally and structurally important highly conserved ones and more variable residues that potentially modulate subfamily or even receptor specific function, while leaving some leeway for evolvability.

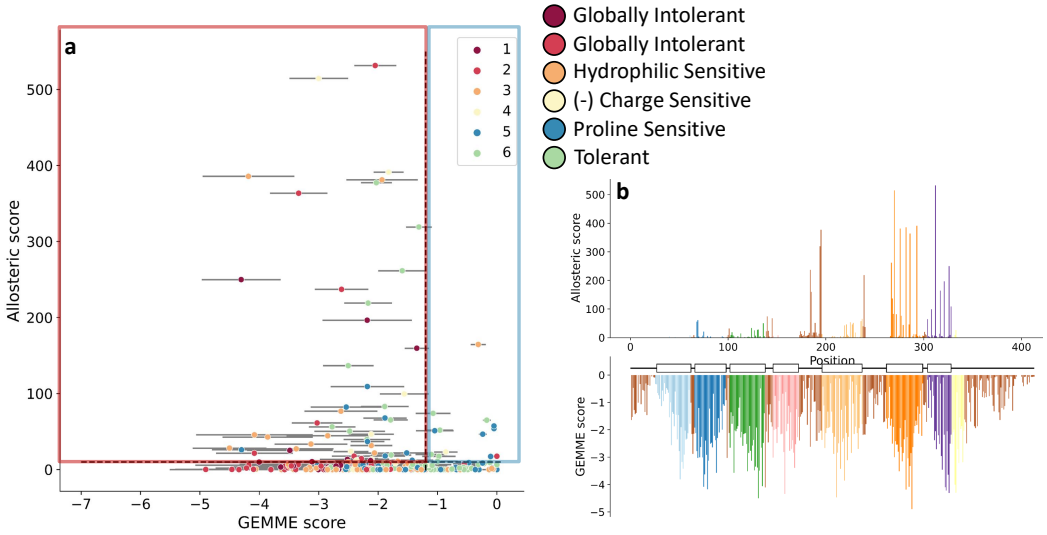


Figure 5.5: **Gemme scores, tolerance clusters, and allosteric scores σ_m: (a)** scatter of allosteric versus GEMME scores where residues are colored by the clustering extracted from Jones et al. (183). The red square represents residues with top 25% σ_m and high evolutionary score, while the blue square represents area with top 25% σ_m and low evolutionary score. The line separating red and blue squares is the median for evolutionary scores. **(b)** Allosteric scores σ_m and GEMME scores versus protein residues. The bars are colored by protein domain. Rectangles and lines represent TM helices and loop regions respectively.

5.3 Methods

5.3.1 GEMME Overview

We utilized GEMME (Genome-wide Estimation of Mutation Effect) (342) as a quick tool for estimating mutation effects. GEMME relies on Multiple Sequence Alignment (MSA) and offers the advantage of using a concise set of biologically meaningful parameters. Initially, it constructs an evolutionary tree including the query sequence and its homologous sequences from the MSA, then calculates the conservation level for each position, indicating its tolerance to mutations. When analyzing a mutation, GEMME assesses the position's conservation in evolution and estimates the required evolutionary fit for accommodating mutations. It combines the mutation's frequency and physicochemical similarities with the minimum evolutionary fit needed, determined by how far one must go in the evolutionary tree to find a natural sequence with the mutation. Consequently, mutations requiring more changes across the sequence are predicted as more deleterious.

It demonstrates performance similar to statistical inference-based (345) and deep learning-based techniques, including family-specific models and high-capacity protein language models (346; 347; 348). Importantly, GEMME explicitly models the protein's evolutionary history, generating an evolutionary score denoted as $\Delta\Delta E$. The calculated $\Delta\Delta E$ scores range from 0 (conservative substitutions) to -7 (incompatible substitutions based on the MSA). In summary, GEMME leverages evolutionary insights and conservation assessments to predict mutation deleteriousness and its impact on protein function and structure.

5.3.2 Procedure for Evolutionary Score Calculation

To construct MSAs, we downloaded DD1R, DD2R, and β2AR sequences from UniProt (P21728, P14416, and P07550 respectively) and applied HHblits (349) with specific parameters (`-e 1e-10 -p 20 -B 20000`) over the Uniclust (350) database built upon UniRef30_2023_02. Subsequently, we degapped the MSAs by removing positions that aligned with gaps in the query sequence, a prerequisite for GEMME's operation, and formatted them to comply with the input requirements of the GEMME tool. Finally, we ran GEMME locally with default parameters to calculate mutant evolutionary scores utilizing the provided docker image.

5.3.3 Molecular dynamics simulations and AlloDy

MD simulations of dopamine receptors (DA-DD1R-GsH5, DA-DD1R-PAM-GsH5, DA-DD2R-GiH5, BRC-DD2R-GiH5) were taken from the computational rewiring of dopamine receptors project (Ch. 3). All the details of the simulations are explained there.

As for β2AR, the starting structures used for MD simulations of are: Isoprenaline bound 7DHR (302) with the last 20 residues of the C-terminal helix of Gs and the sequence re-mutated back to WT, and carazolol bound 2RH1 (351). The setup of the simulations is identical to

that described for the dopamine systems (Ch. 3:**Methods**). After constrained equilibration, 4 replicas of 1000 ns were run for each system, the first 125 ns of every simulation was discarded for equilibration of C-alpha RMSD, and the rest of the simulation was used for calculating statistics.

Calculation of mutual information (MI), allosteric scores (σ_m), and allosteric pathways has been performed on dopamine D1, dopamine D2, and β2 adrenergic receptors as described in the methods chapter (Ch. 2).

6 Conclusions and contributions

6.1 Conclusion

This thesis aims to investigate and design allostery by first modeling allosteric behavior in GPCRs from a dynamical perspective and then incorporating the model descriptors with traditional protein design methods to rationally design minimal perturbations that would elicit a modified allosteric response.

Building upon the knowledge in the literature on GPCRs, dynamical simulations, and allosteric modeling, **Part I** introduces AlloDy, a MD simulations analysis package that assembles an ensemble of metrics for quantifying allosteric transmission. AlloDy allows for a holistic analysis of a set of simulations by including basic simulation diagnostics, ligand binding contacts, ligand pose clustering, GPCR specific activation state determination, mutual information calculation, allosteric pathway extraction, and perturbation response quantification. These metrics can be used individually, or combined as seen in **Part II** to design allostery in dopamine receptors or validate already designed flexible peptide agonists in chemokine receptors.

In an age of *de novo* protein design (hallucination, generative methods, etc), advancing rational design is of paramount importance. In this work, we have taken allosteric design of GPCRs a step further by introducing ligand specificity via mutation of hotspots distant to the ligand binding site. This ligand specificity could be introduced via single point substitutions, which tells us that with proper knowledge of the system, minimum perturbation is required to reach the design goal. In addition, the descriptors developed in this work that were used in the process of design have pushed forward understanding of the underpinnings of allostery in GPCRs and proteins in general.

This molecular understanding requires incorporating dynamic information into design, whether through different ligand peptide binding poses, correlation metrics (MI), allosteric pathways, or ensemble differences (Fig. 6.1). This reinforces the idea that structural information is not enough for (1) understanding and (2) designing allosteric behavior.

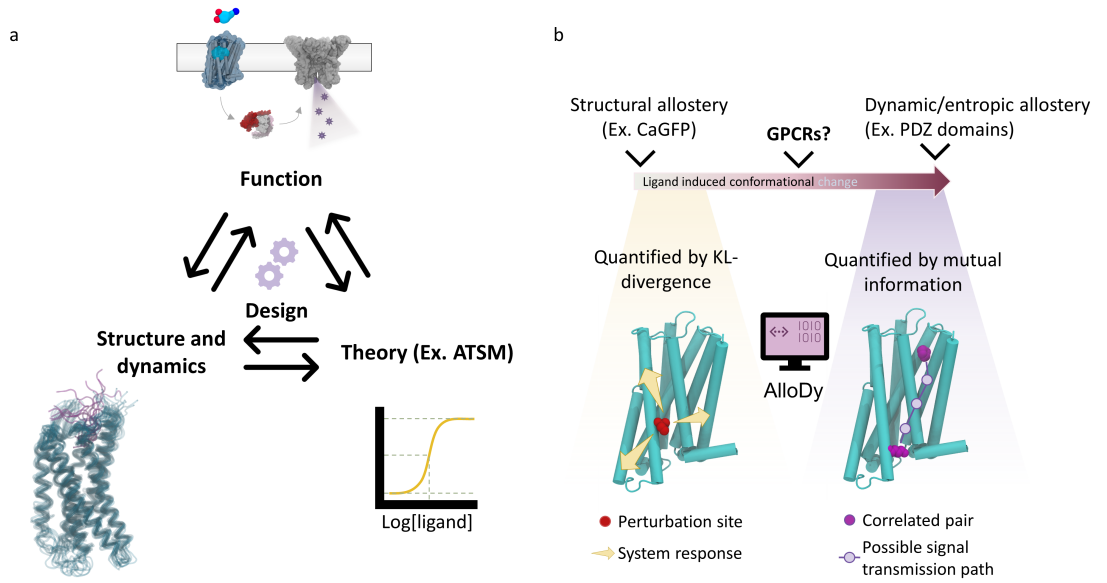


Figure 6.1: **a** The trio of theory, structure/dynamics, and function with design being at the center and requiring knowledge of the three aspects. **b** Description of allosteric mechanisms as a function of an order parameter (177). AlloDy uses metrics that can describe both ends of the spectrum.

Finally, we explore the relationship between genetic variation, allosteric metrics, evolutionary scoring, and functional readouts in GPCRs. We find that there is no clear relationship between the surveyed allosteric metrics and evolutionary scores, and that genetic variants found in the population and clinical databases generally do not overlap with allosteric hotspots.

To synthesize, this thesis puts forward a relationship between evolvability of proteins and conformational dynamics (352; 353). We show that a receptor such as dopamine D2 is one point mutation away from achieving total ligand selectivity, and that chemokine receptors are a few mutations away from having a rigid or a flexible binding mode with modified chemokine peptides. The conformational dynamics of these receptors allow a level of functional promiscuity that can be modulated via point substitutions. Knowledge of the dynamics and allosteric transmission within the studied receptors allow for a level of control of the functional modulation.

6.2 Contributions

Hereby we summarize the contributions that this thesis provides to the field:

AlloDy: Developing a publicly available tool, AlloDy, to study allostery in proteins using diverse metrics from information and graph theories. AlloDy provides scores at a protein level (allosteric pathways) and residue level (allosteric scores, MI, KL-divergences) that act as

valuable input for the protein design process.

Ligand selective design: First allosterically ligand selective designs in GPCRs, which form a basis for selective drug design and rational receptor engineering for both fundamental research and therapeutic applications

Evidence for allosteric pathways in GPCRs: the fact that we were able to achieve ligand specific responses by exploiting unique allosteric pathways is evidence for the pathway view of allostery in GPCRs.

Expanding the available set of GPCR MD simulations: Comprehensive set of MD simulations to be contributed to the already fantastic work being done at gpcrmd (354). Simulation sets contain simulations of GPCRs in different states, binding various ligands, interacting with various IC binding partners, and mutated GPCRs.

6.3 Future directions

6.3.1 Method development

In light of the advancements delineated within this thesis, it is evident that there remains a substantial body of research yet to be undertaken to unify allosteric description of G protein-coupled receptors (GPCRs) with the overarching objective of protein design.

On the side of AlloDy development, there is still some improvements to be done regarding accessibility and ease of use. Furthermore, correlation metrics emerge as a pivotal facet in this kind of analysis. Beyond the conventionally employed measure of mutual information (MI), there may arise a need for directional metrics such as transfer entropy, which would lead to construction of directional graphs that would elucidate directionality in the intricate mechanisms underlying GPCR allosteric regulation and modulation. After the construction of the protein graph, the appropriateness of considering shortest paths followed by path clustering as the descriptive paradigm begs further scrutiny, with an exploration of employing simpler, yet more interpretable metrics for measuring allosteric strength of residues.

Moreover, the precise nature of the simulations conducted and their subsequent interpretation still requires consideration. The KL-divergence formulation as a measure of perturbation response is a step in the direction of interpretation of simulations in different states. However, the current simulation setup lacks free energy insight into transitions from and to receptor active state under the effect of mutation (we currently estimate free energy differences between mutants and WT using RosettaMembrane, which does not include transition information).

Lastly, and relating to the previous point, implementation of enhanced sampling techniques may serve improving sampling in simulations under the current framework (where we sample equilibrium states). Replica exchange methods with solute scaling/tempering (247) or accelerated MD (244) can serve this purpose. For simulating transitions, metadynamics protocols

have been successfully employed to study GPCR systems (355), and can prove invaluable to study free energy differences along transition paths in mutant receptors.

6.3.2 Applications

One of the steps forward in this thesis is moving from allosteric design in GPCRs toward endogenous ligand and cognate G-protein (16) to ligand selective design. A natural extension to this is design of bias toward an intracellular binding partner (either G-protein or β-arrestin, or bias between G-protein subtypes) allosterically. This follows a similar underlying assumption that different IC effectors will engage a set of common and unique allosteric pathways, where we can target effector unique pathways for design. Initial simulations of arrestin bound systems analyzed by AlloDy find separate sets of pathways when compared with G-protein bound simulations. The main challenge would be deciding the exact amino acid substitutions at the design hotspots, since the design strategy depending on NMA calculations seems to have reached its limit with ligand selectivity calculations.

A longer term goal is using this understanding that we gain from allosteric design to move toward drug/peptide design. The allosteric description provided here allows specific targeting of hotspots in the receptor ligand binding region that initiate allosteric pathways connecting to the effector binding region, opening the door for designing specific interactions for function.

A Appendix: supplementary figures

A.1 Supplementary figures: Development: AlloDy

This section contains supplementary figures for chapter 2.

A.1.1 Md2path: calculating allosteric pathways from MD simulations

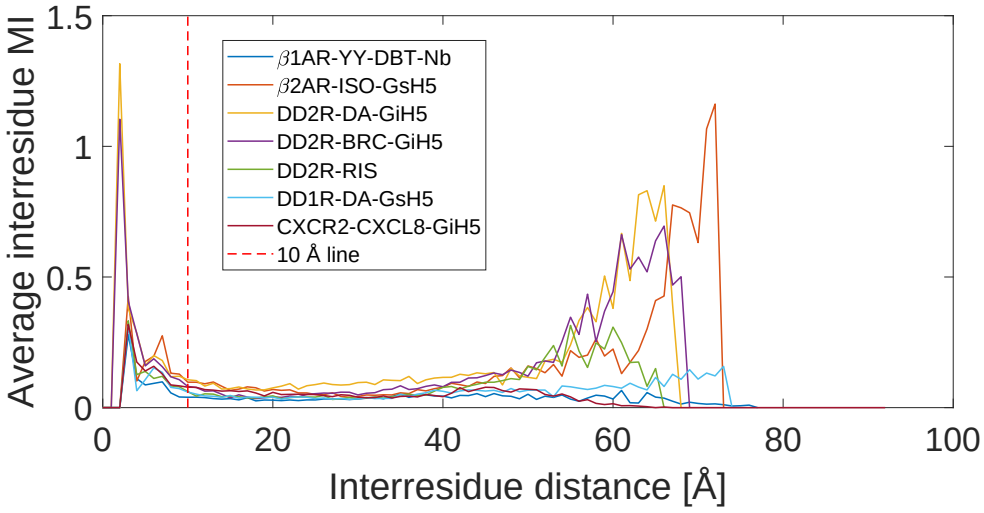


Figure A.1: **Average interresidue MI vs interresidue distance** for a set of simulated class A GPCRs. The 10 Å line is shown for clarity. The first peak at 2 or 3 Å represents "direct communication" between residues in close proximity, while the further peak between 640 and 72 Å represents "allosteric communication".

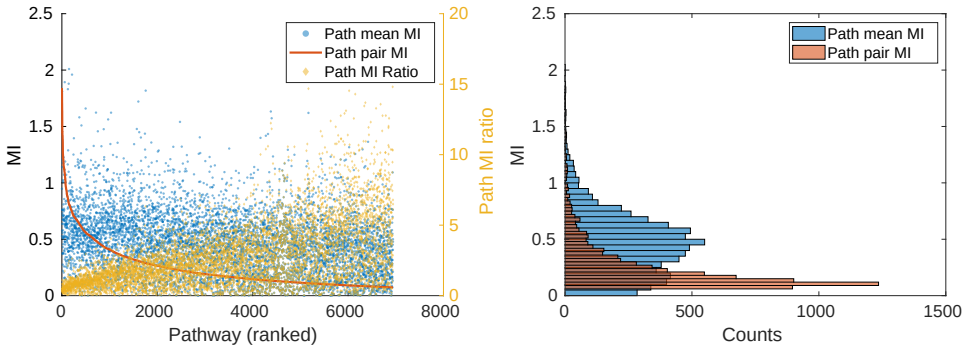


Figure A.2: **Distribution of path pair and mean MI:** (left) path pair MI (red solid line: end node MI of a pathway) and path mean MI (yellow rhombus, mean over all nodes in a pathway) as a function of pathway indices ranked by pair MI. (right) distribution of path pair MI (red) and path mean MI (blue).

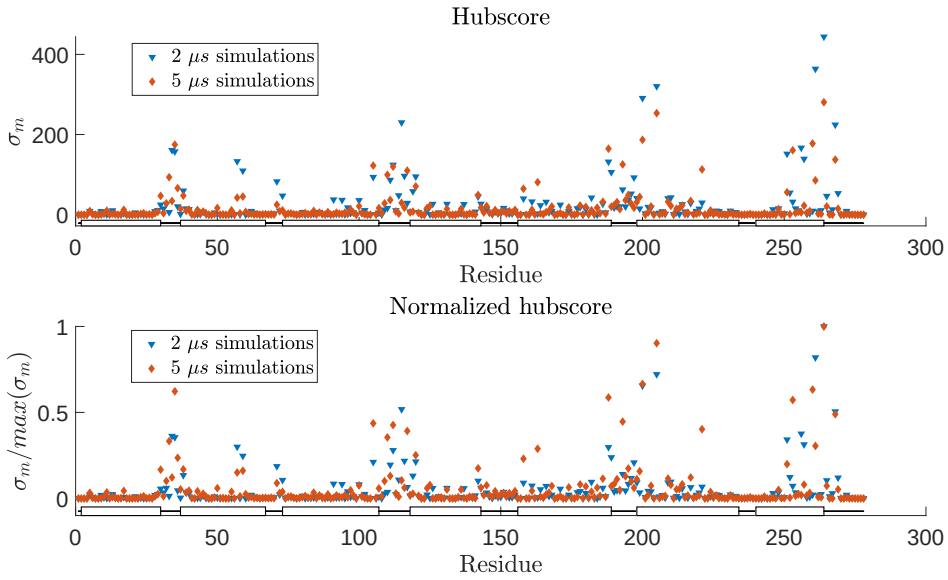


Figure A.3: **Allosteric strength scores in two sets of independent DA-DD2R simulations of different lengths:**(top) Raw allosteric strength scores σ_m are plotted. (bottom) Allosteric strength scores normalized by maximum score for every analyzed set of simulations.

A.1.2 Higher order KL terms: amino acid substitution in bromocriptine-bound DD2R:I4.46N and WT

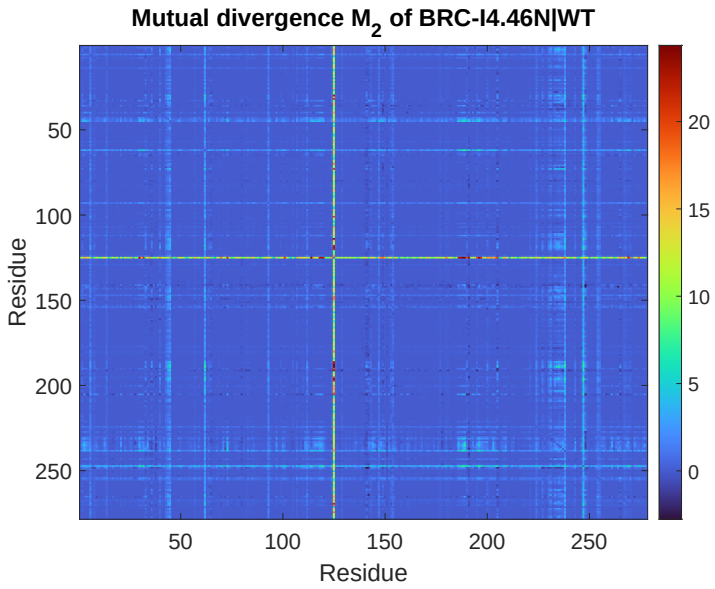
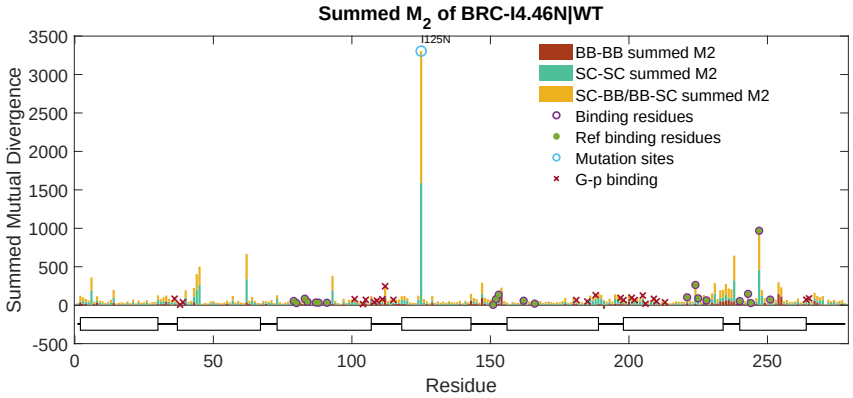
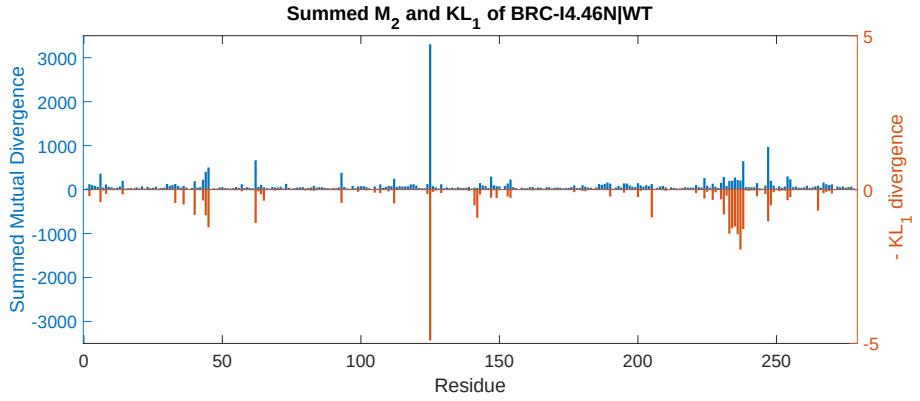


Figure A.4: Mutual divergence M_2 for DD2R bound to BRC I4.46N substitution (test ensemble) and BRC WT (reference ensemble).



(a) Mutual divergence M_2 summed over residues for DD2R system with the WT BRC-bound as reference and I4.46N BRC-bound as test ensembles. M_2 contributions are color-coded according to backbone-backbone (brown), sidechain-sidechain (green), and backbone-sidechain (yellow). Ligand binding residues and G-protein binding residues are marked. The largest divergences are in the ligand binding region. The 7TM helices of GPCRs are highlighted.



(b) Same residue summed M_2 (blue, left y-axis) plotted above with KL_1 plotted in the negative direction (red, right y-axis).

Figure A.5: M_2 and KL_1 comparison of BRC-bound DD2R I4.46N and WT.

A.1.3 Higher order KL terms: Gi-helix5 and dopamine-bound DD2R (active state) and risperidone-bound DD2R (inactive state)

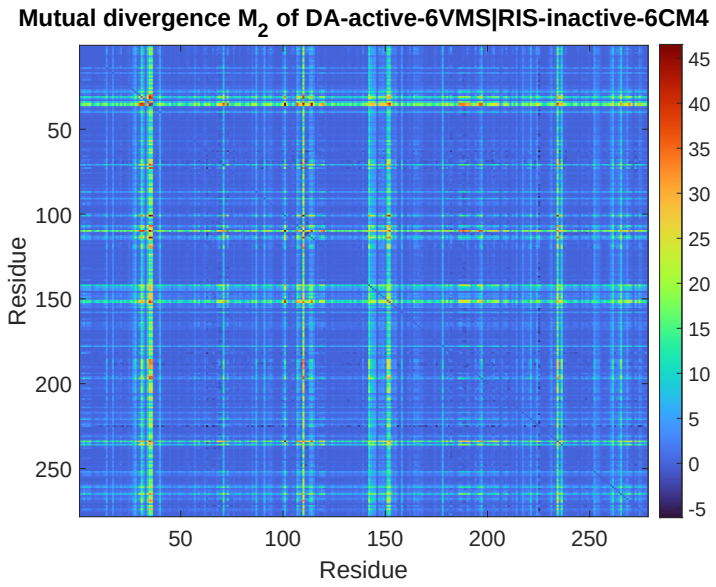
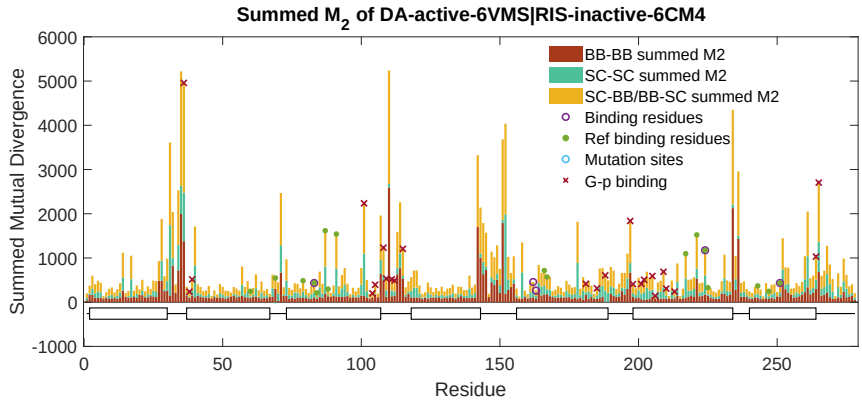
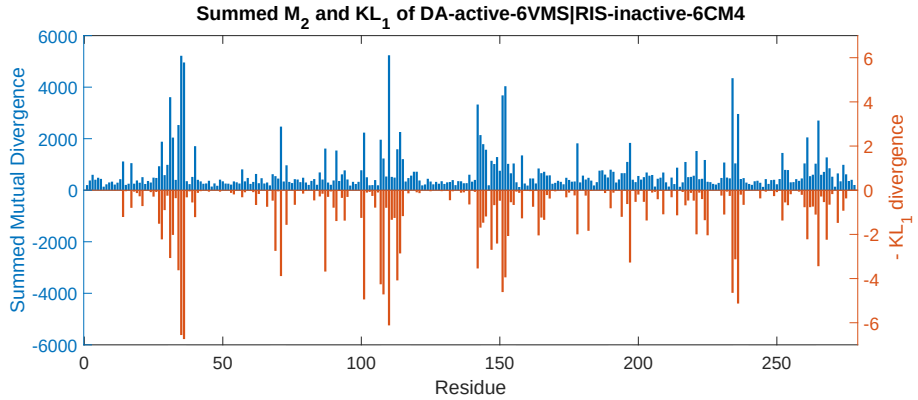


Figure A.6: Mutual divergence M_2 for DA-DD2R-Gi helix 5 complex (test ensemble) and RIS-bound DD2R (reference ensemble).



(a) Mutual divergence M_2 summed over residues for DD2R system with the RIS-bound as reference and DA-DD2R-Gi helix 5 complex as test ensembles. M_2 contributions are color-coded according to backbone-backbone (brown), sidechain-sidechain (green), and backbone-sidechain (yellow). Ligand binding residues and G-protein binding residues are marked. The largest divergences are in the ligand binding region. The 7TM helices of GPCRs are highlighted.



(b) Same residue summed M_2 (blue, left y-axis) plotted above with KL_1 plotted in the negative direction (red, right y-axis).

Figure A.7: M_2 and KL_1 comparison of DA-DD2R-Gi helix 5 complex (test ensemble) and RIS-bound DD2R (reference ensemble).

A.1.4 Relationship of KL-divergences to experimental observables fitting using backbone divergences

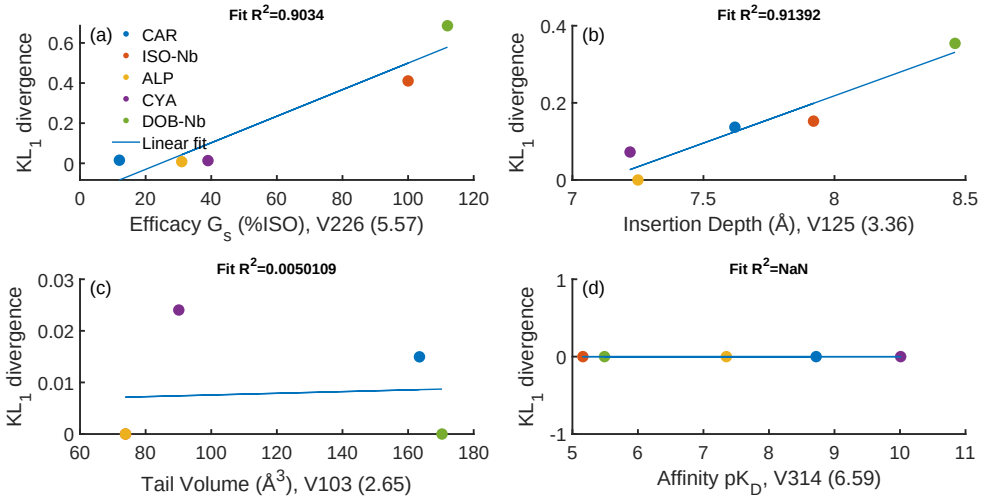


Figure A.8: **Fitting of backbone KL_1 with β1AR-TS apo as reference to different experimental observables correlated with NMR chemical shifts: (a)** efficacy of ligands for Gs signalling pathway, **(b)** ligand insertion depth, **(c)** ligand tail volume, and **(d)** ligand affinity.

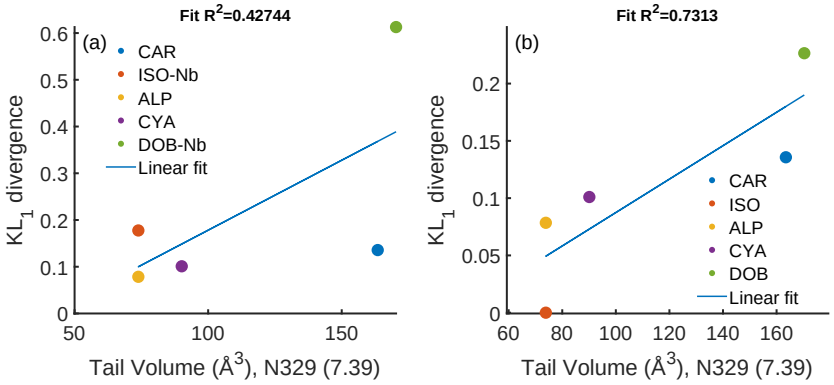


Figure A.9: **Fitting of backbone KL_1 of residue N329 (7.39) with β1AR-TS apo as reference to tail volume of simulated ligand: (a)** correlation using nanobody bound simulations for agonist ligands and **(b)** correlation using β1AT-TS without any bound intracellular binding partner.

A.2 Supplementary figures: Computational rewiring of allosteric pathways reprograms GPCR selective responses to ligands in dopamine receptor

This section contains supplementary figures for chapter 3.

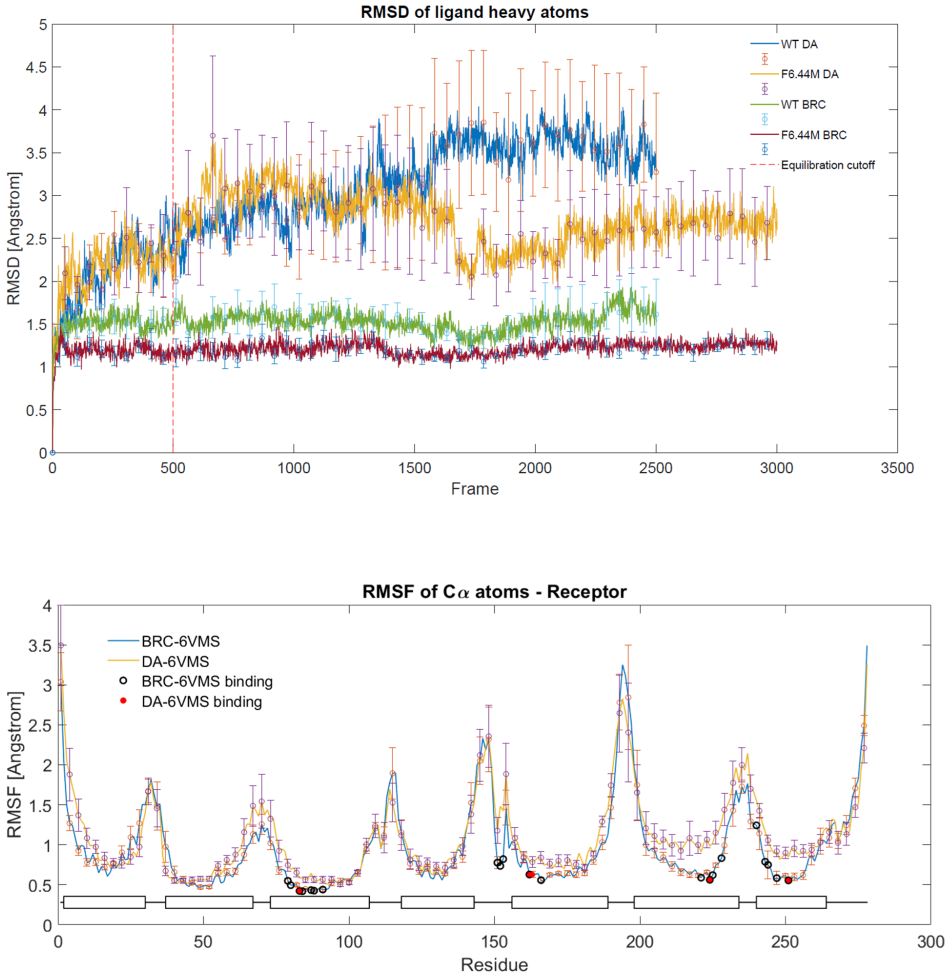


Figure A.10: **RMSD and RMSF of WT and F6.44M DD2R. a** Root mean square deviations (RMSD) of the ligands DA and BRC during MD simulations for D2-WT (blue: DA, green: BRC) and D2-F6.44M mutation (yellow: DA, crimson: BRC). The first 500 frames of the simulations are discarded as equilibration (red dashed line). **b** Root mean square fluctuations (RMSF) of the DA and BRC-bound D2 receptors during MD simulations for WT. The first 500 frames of the simulations are discarded as equilibration. Ligand binding residues for every ligand are marked on the plot (red dots for DA binding residues and black circles for BRC binding residues). TMHs are represented as rectangles at the bottom of the plot space.

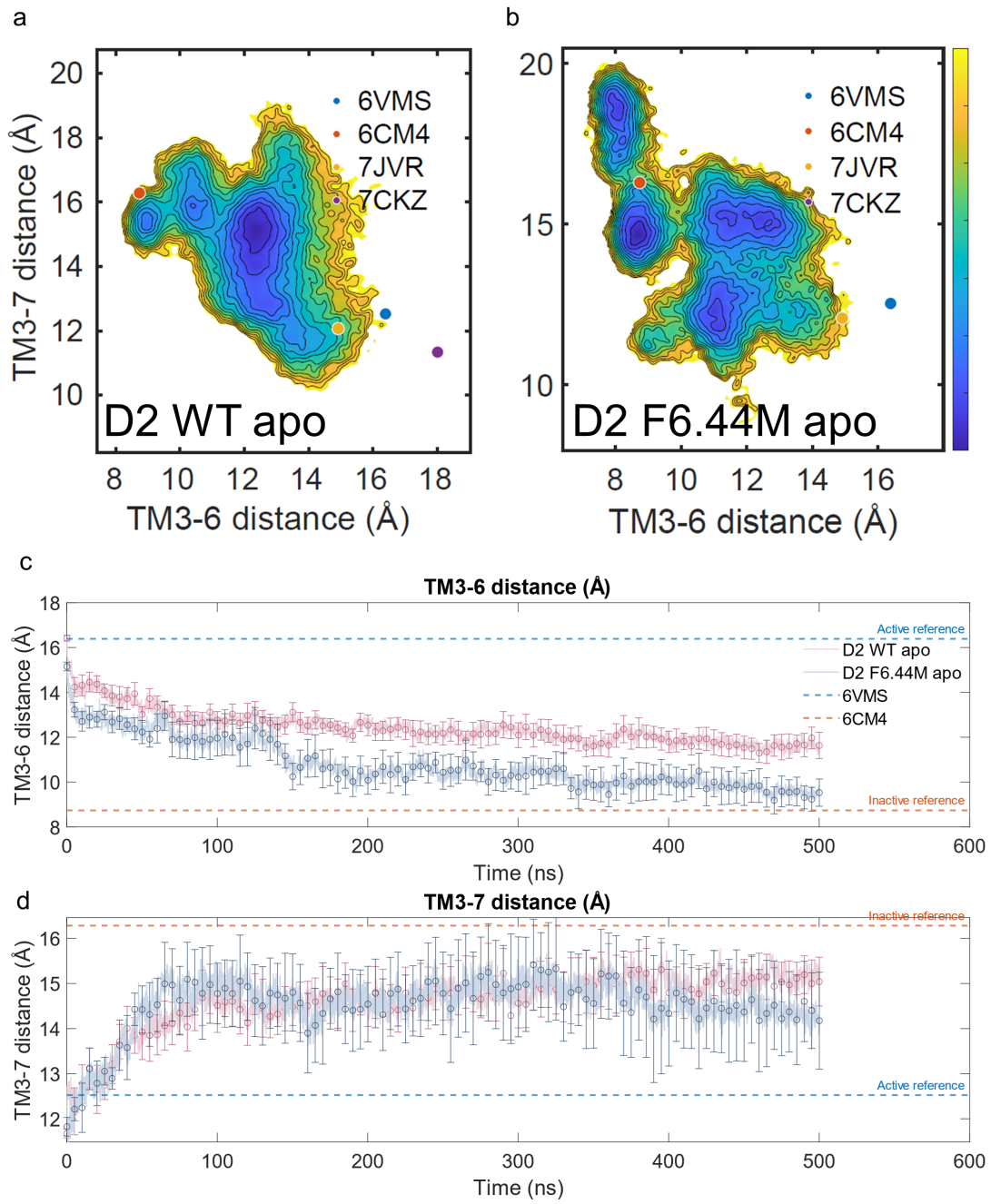


Figure A.11: **Activation states of WT and F6.44M *apo* state simulations starting from active states: a** and **b** activation landscapes of dopamine D2 WT apo state ($n = 10$) starting from an active like state (**a**) and D2 F6.44M apo state ($n = 6$) starting from an active like state (**b**). Reference inactive state (6CM4) and active states (6VMS, 7JVR, and 7CKZ) are highlighted on the plots. **c** and **d** Time series plots of TM3-6 (**c**) and TM3-7 (**d**) distances for the aforementioned systems. Dashed lines represent active (6VMS, blue) and inactive (6CM4, red) references. TM3-6 distances were calculated between C-alphas of residues R3.50 and E6.30, while TM3-7 distances were calculated between C-alphas of resides R3.50 and Y7.53.

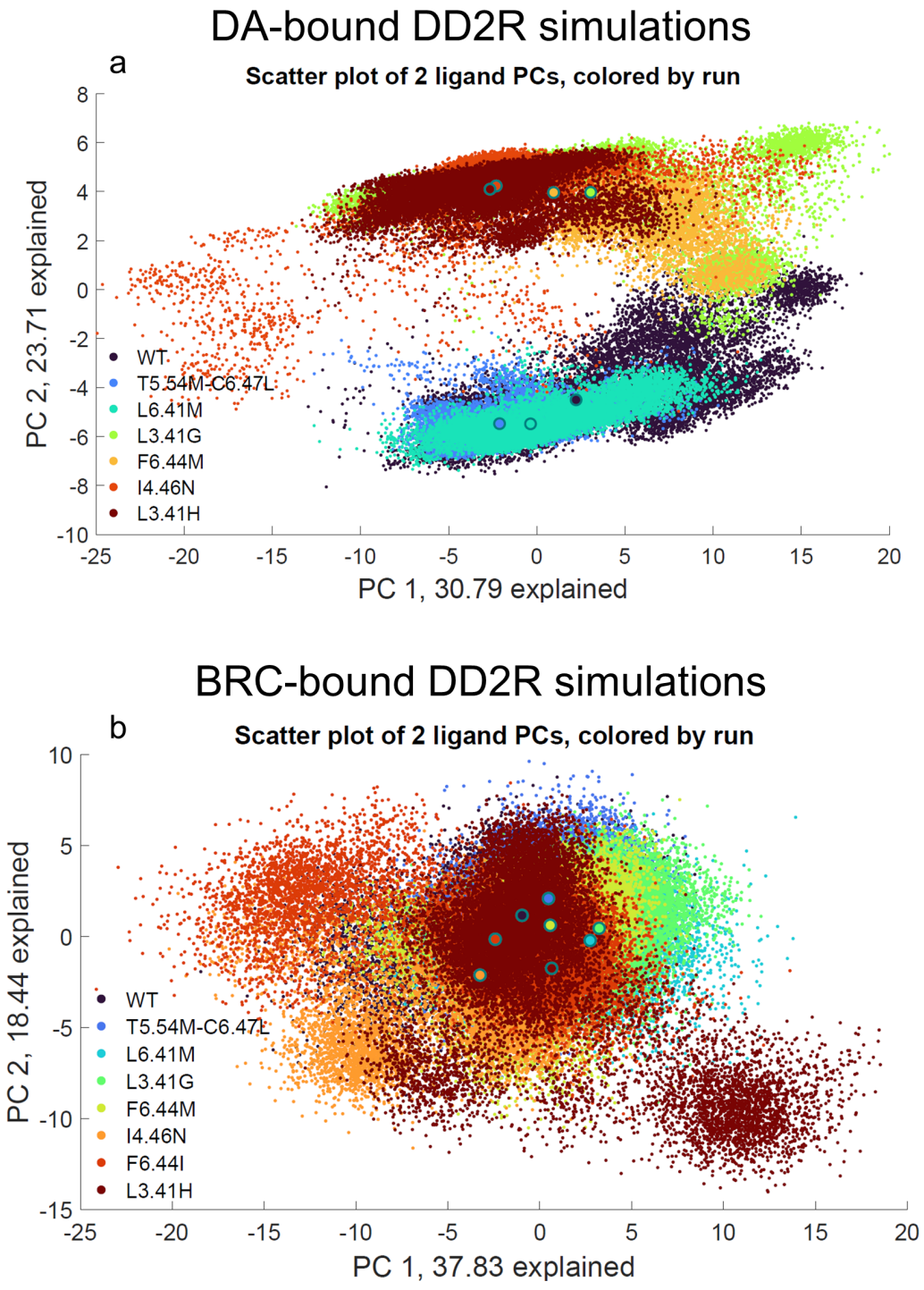


Figure A.12: **Ligand binding pose PCA in DD2R simulated systems: a** DA-bound DD2R simulations and **b** BRC-bound DD2R simulations. PCA was performed on heavy atom coordinates.

A.3 Supplementary figures: Computational study of designed dynamic receptor—peptide signaling complexes applied to chemotaxis

This section contains supplementary figures for chapter 4.

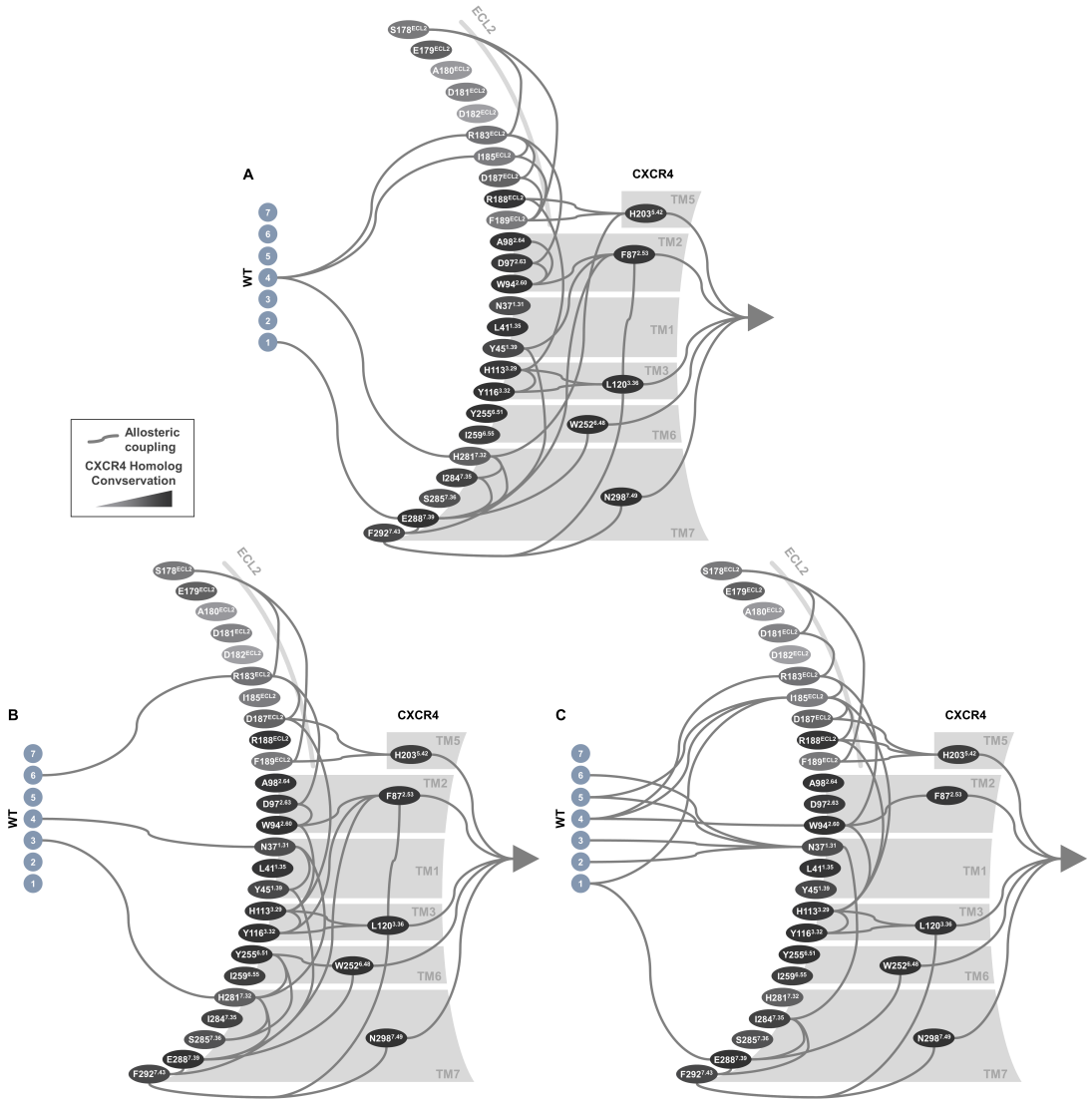


Figure A.13: **Predicted allosteric couplings in the WT CXCR4:WT CXCL12 complex. (a-c)** Predicted allosteric pipelines (solid lines) starting from the peptide and running through the receptor towards the intracellular side are calculated for substates C1 **(a)**, C2 **(b)**, and C3 **(c)** and represented schematically as follows using 3 layers of residues from left to right. Left layer: Peptide residues shown as grey spheres. Middle layer: Receptor residues in the extracellular peptide binding pocket shown as ovals. Those allosterically coupled to peptide residues (connected by a solid line) are defined as allosteric triggers. Right layer: Allosteric transmitter residues coupled to allosteric triggers shown as ovals in the receptor core and located in distinct transmembrane helices (TM 2,3,5,6,7). Receptor residues are colored according to their level of sequence conservation in CXCR4.

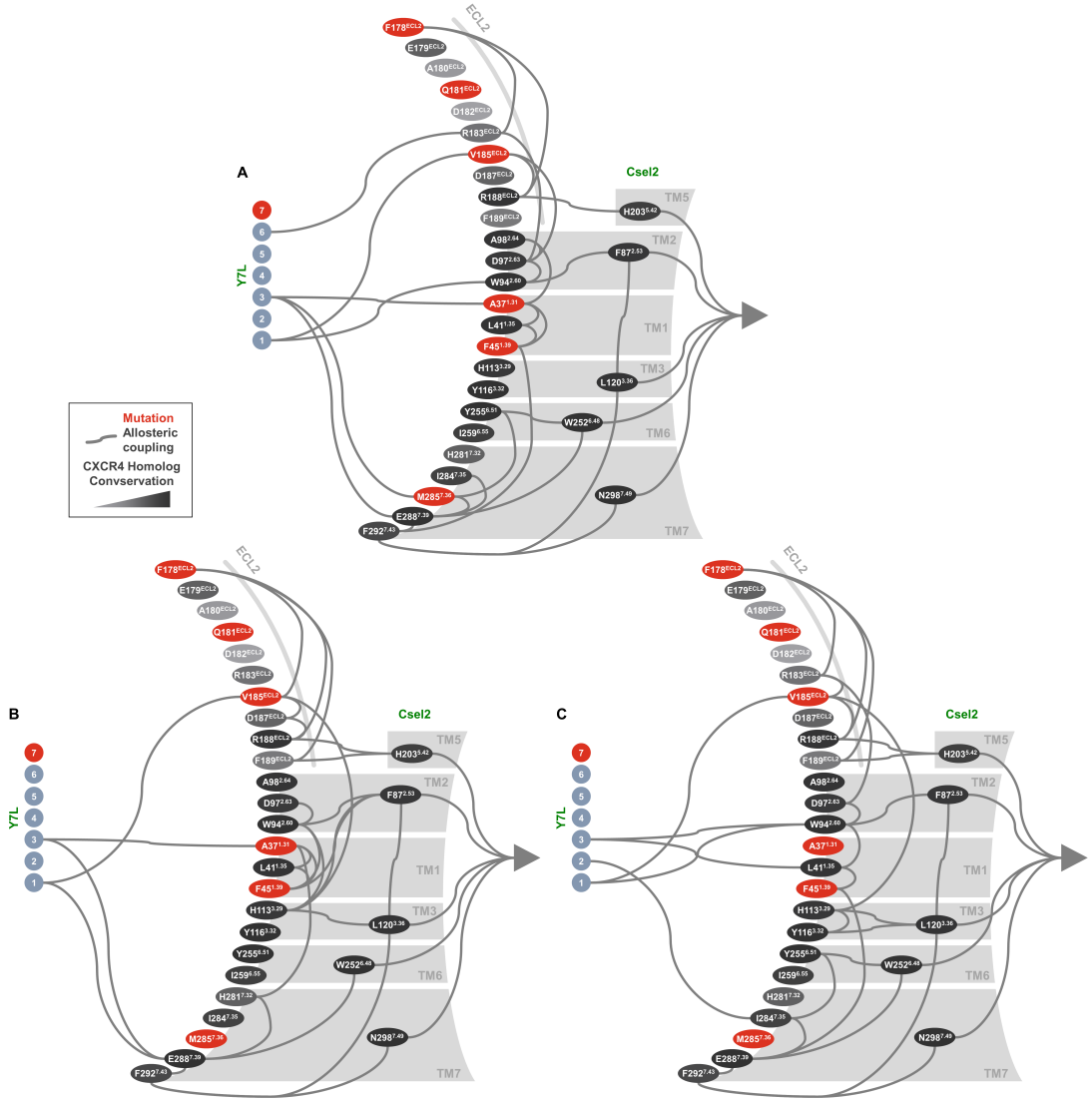


Figure A.14: **Predicted allosteric couplings in the Csel2:Y7L CXCL12 complex. (a-c)** Predicted allosteric pipelines (solid lines) starting from the peptide and running through the receptor towards the intracellular side are calculated for substates C1 **(a)**, C2 **(b)**, and C3 **(c)** and represented schematically as follows using 3 layers of residues from left to right. Left layer: Peptide residues shown as grey spheres. Middle layer: Receptor residues in the extracellular peptide binding pocket shown as ovals. Those allosterically coupled to peptide residues (connected by a solid line) are defined as allosteric triggers. Right layer: Allosteric transmitter residues coupled to allosteric triggers shown as ovals in the receptor core and located in distinct transmembrane helices (TM 2,3,5,6,7). Receptor residues are colored according to their level of sequence conservation in CXCR4. Mutated peptide and receptor residues are colored in red.

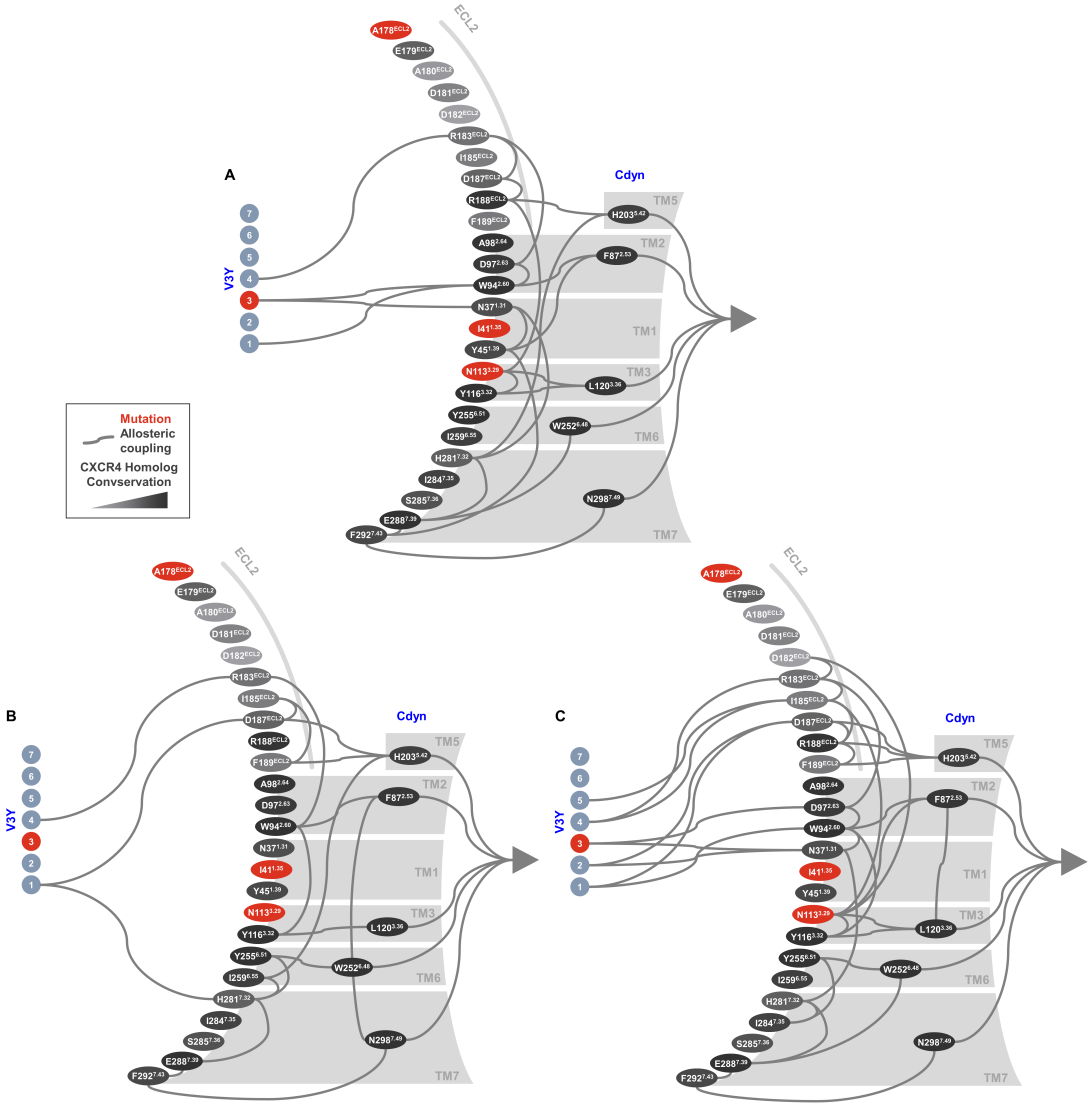


Figure A.15: **Predicted allosteric couplings in the Cdyn:V3Y CXCL12 complex. (a-c)** Predicted allosteric pipelines (solid lines) calculated for substates C1 **(a)**, C2 **(b)**, and C3 **(c)** are represented schematically as follows using 3 layers of residues from left to right. Left layer: Peptide residues shown as grey spheres. Middle layer: Receptor residues in the extracellular peptide binding pocket shown as ovals. Those allosterically coupled to peptide residues (connected by a solid line) are defined as allosteric triggers. Right layer: Allosteric transmitter residues coupled to allosteric triggers shown as ovals in the receptor core and located in distinct transmembrane helices (TM 2,3,5,6,7). Receptor residues are colored according to their level of sequence conservation in CXCR4. Mutated peptide and receptor residues are colored in red.

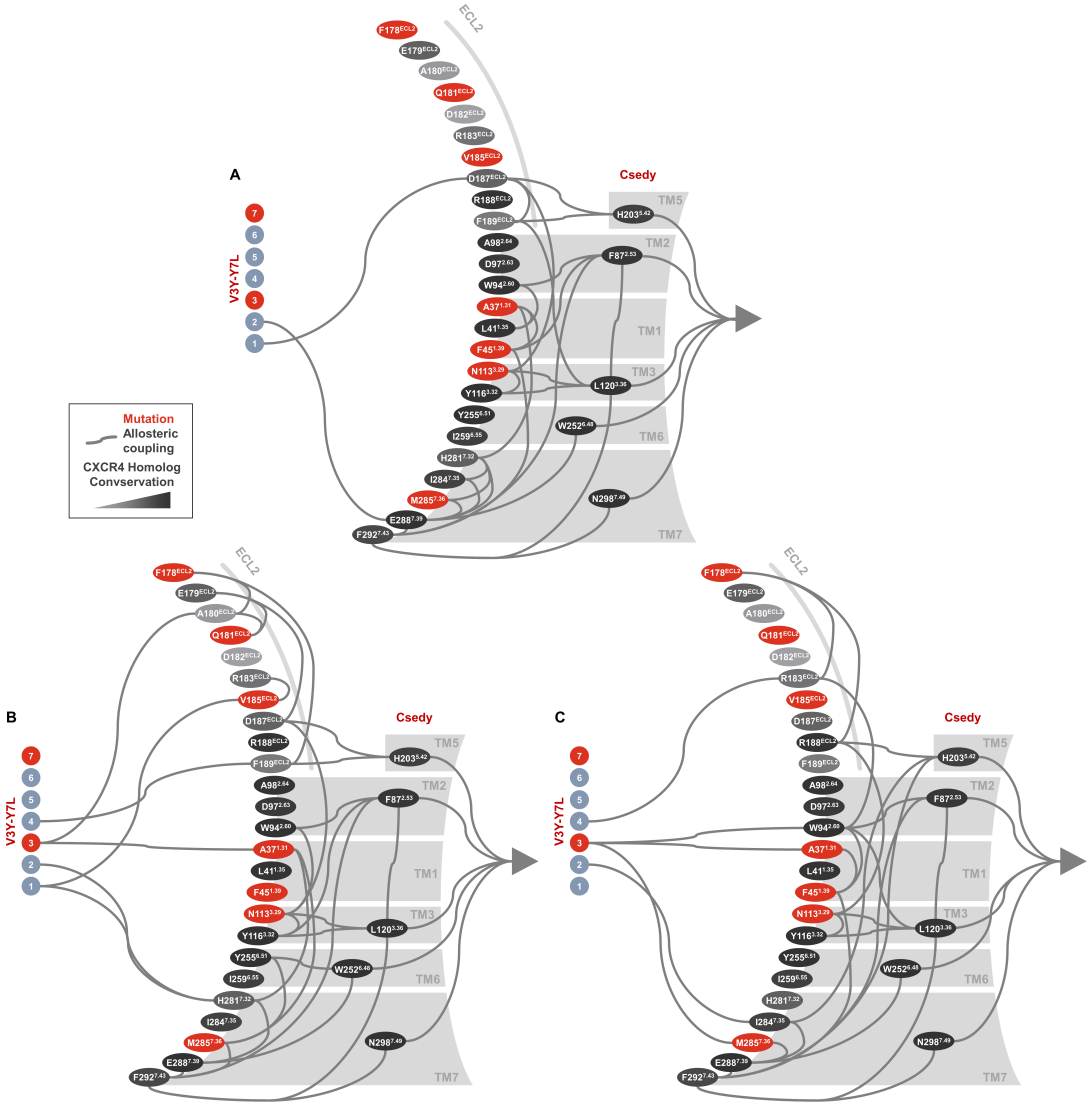


Figure A.16: **Predicted allosteric couplings in the Csedy:V3Y-Y7L CXCL12 complex. (a-c)** Predicted allosteric pipelines (solid lines) starting from the peptide and running through the receptor towards the intracellular side are calculated for substates C1 **(a)**, C2 **(b)**, and C3 **(c)** and represented schematically as follows using 3 layers of residues from left to right. Left layer: Peptide residues shown as grey spheres. Middle layer: Receptor residues in the extracellular peptide binding pocket shown as ovals. Those allosterically coupled to peptide residues (connected by a solid line) are defined as allosteric triggers. Right layer: Allosteric transmitter residues coupled to allosteric triggers shown as ovals in the receptor core and located in distinct transmembrane helices (TM 2,3,5,6,7). Receptor residues are colored according to their level of sequence conservation in CXCR4. Mutated peptide and receptor residues are colored in red.

Variant / Substate	Path Density (% total)	Occupancy (% total)	Intra-cluster RMSD (Å)	
			Complex	Pocket
WT				
C1	51.3	29.6	2.43	2.54
C2	31.9	21.9	2.55	2.63
C3	16.8	17.3	2.32	2.30
Csel2				
C1	30.6	34.5	2.14	1.99
C2	31.7	30.4	2.07	2.09
C3	37.7	23.8	2.18	2.08
Cdyn				
C1	51.3	51.9	3.01	3.61
C2	13.4	15.9	3.01	3.47
C3	35.2	14.6	2.69	3.31
Csedy				
C1	42.8	30.4	2.51	2.76
C2	18.0	23.0	2.47	3.04
C3	39.1	19.5	2.48	2.68

Figure A.17: **Conformational dynamics and structural characterization of clustered substates of WT and CAPSen designs from molecular dynamics simulations.** Pathway density distribution across substates selected as input for AlloDy. Occupancy calculated from total simulated frames. Intra-cluster RMSD calculated from receptor and peptide (Complex) and contacting residues across all variants including peptide (Pocket). Contact frequency thresholds used to define binding and allosteric contacts normalized to intra-cluster complex RMSD of each substate.

Hub	F87$^{2.53}$	L120$^{3.36}$	H203$^{5.42}$	W252$^{6.48}$	N298$^{7.49}$
WT	670 (0.27)	792 (0.32)	310 (0.12)	517 (0.21)	220 (0.09)
Csel2:Y7L	621 (0.22)	378 (0.13)	482 (0.17)	338 (0.12)	1017 (0.36)
Cdyn:V3Y	302 (0.14)	648 (0.29)	395 (0.18)	156 (0.07)	733 (0.33)
Csedy:V3Y-Y7L	881 (0.21)	1422 (0.34)	772 (0.18)	171 (0.04)	990 (0.23)

Figure A.18: **Allosteric strength of transmission hubs across variants.** Allosteric hubscores calculated from AlloDy for transmission hub residues conserved among WT and designed variants. Hubscores measure the total number of allosteric pathways running through a residue. Since pathways are constructed from pairs of residues exchanging significant amount of mutual information, comparison of hubscores between transmission hubs gives an indication of the relative amount of information passing through these sites (see Methods for detailed calculations). The fractional hubscore among the transmission hubs for each variant is shown in parentheses to enable comparison between variants.

Step	1	2	3	4	5	6
Time [ps]	125	125	125	250	250	250
Timestep [fs]	1	1	1	2	2	2
Ensemble	NVT	NVT	NPT	NPT	NPT	NPT
Position restraints [kJ/(mol·nm)]						
Backbone	4000	2000	1000	500	200	50
Side-chain	2000	1000	500	200	50	0
Lipid & Ligand	1000	400	400	200	40	0
Dihedral	1000	400	200	200	100	0

Figure A.19: **Equilibration restraints used for molecular dynamics simulations of receptor—peptide complexes.** Timing and position restraints used for all simulated systems during equilibration in 6 steps.

System (Receptor:Peptide)	Number of Runs	Simulated Time of Every Run	Total Simulated Time
WT:WT	7	3 x 200 ns 4 x 300 ns	1.8 µs
Cdyn:V3Y	7	2 x 200 ns 5 x 300 ns	1.9 µs
Csedy:V3Y-Y7L	7	3 x 200 ns 4 x 300 ns	1.8 µs
Csel2:Y7L	5	5 x 300 ns	1.5 µs
CCR5:RANTES	5	5 x 300 ns	1.5 µs

Figure A.20: **Simulated time of each system.** Number of runs and simulated time for each run and total simulation time for each receptor—peptide system.

System (Receptor:Peptide)	Variability explained by PC1 (%)	Variability explained by PC2 (%)
WT:WT	59.95	16.75
Cdyn:V3Y	67.47	14.72
Csedy:V3Y-Y7L	52.37	21.34
Csel2:Y7L	28.33	20.52
CCR5:RANTES	53.66	19.09

Figure A.21: **Variability of each system explained by principle component analysis.** Variability explained by the first and second principle components (PC1, PC2) for each receptor—peptide system.

Bibliography

[1] 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 in neurosciences*, volume 25, pages 366–428. Elsevier, 1995.

[2] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: a plausible model. *J Mol Biol*, 12(1):88–118, 1965.

[3] Jacques Monod. *Chance and necessity an essay on the natural philosophy of modern biology*. New York, Knopf, 1971.

[4] Linus Pauling. The oxygen equilibrium of hemoglobin and its structural interpretation. *Proceedings of the National Academy of Sciences of the United States of America*, 21(4):186, 1935.

[5] Jacques Monod, Jean-Pierre Changeux, and Francois Jacob. Allosteric proteins and cellular control systems. *Journal of molecular biology*, 6(4):306–329, 1963.

[6] Hans Frauenfelder, Benjamin H McMahon, Robert H Austin, Kelvin Chu, and John T Groves. The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Proceedings of the National Academy of Sciences*, 98(5):2370–2374, 2001.

[7] Joseph Adams, Kenneth Johnson, Robert Matthews, and Stephen J Benkovic. Effects of distal point-site mutations on the binding and catalysis of dihydrofolate reductase from escherichia coli. *Biochemistry*, 28(16):6611–6618, 1989.

[8] Susan M Green and David Shortle. Patterns of nonadditivity between pairs of stability mutations in staphylococcal nuclease. *Biochemistry*, 32(38):10131–10139, 1993.

[9] James E Mace, Barry J Wilk, and David A Agard. Functional linkage between the active site of α-lytic protease and distant regions of structure: Scanning alanine mutagenesis of a surface loop affects activity and substrate specificity. *Journal of molecular biology*, 251(1):116–134, 1995.

[10] David B Olsen, Mark W Stahlhut, Carrie A Rutkowski, Hilary B Schock, Lawrence C Kuo, et al. Non-active site changes elicit broad-based cross-resistance of the hiv-1 protease to inhibitors. *Journal of Biological Chemistry*, 274(34):23699–23701, 1999.

[11] Samy O Meroueh, Pierre Roblin, Dasantila Golemi, Laurent Maveyraud, Sergei B Vakulenko, Yun Zhang, Jean-Pierre Samama, and Shahriar Mobashery. Molecular dynamics at the root of expansion of function in the m69l inhibitor-resistant tem β-lactamase from escherichia c oli. *Journal of the American Chemical Society*, 124(32):9422–9430, 2002.

[12] TAJ Duke, N Le Novere, and D Bray. Conformational spread in a ring of proteins: a stochastic approach to allostery. *Journal of molecular biology*, 308(3):541–553, 2001.

[13] Sebastiano Pasqualato, Julie Ménétrey, Michel Franco, and Jacqueline Cherfils. The structural gdp/gtp cycle of human arf6. *EMBO reports*, 2(3):234–238, 2001.

[14] Olga V Makhlynets, Elizabeth A Raymond, and Ivan V Korendovych. Design of allosterically regulated protein catalysts. *Biochemistry*, 54(7):1444–1456, 2015.

[15] Anita K Nivedha, Christofer S Tautermann, Supriyo Bhattacharya, Sangbae Lee, Paola Casarosa, Ines Kollak, Tobias Kiechle, and Nagarajan Vaidehi. Identifying functional hotspot residues for biased ligand design in g-protein-coupled receptors. *Molecular pharmacology*, 93(4):288–296, 2018.

[16] Kuang-Yui Michael Chen, Daniel Keri, and Patrick Barth. Computational design of g protein-coupled receptor allosteric signal transductions. *Nature Chemical Biology*, pages 1–10, 2019.

[17] Linus Pauling, Harvey A Itano, Seymour J Singer, and Ibert C Wells. Sickle cell anemia, a molecular disease. *Science*, 110(2865):543–548, 1949.

[18] Ruth Nussinov, Mingzhen Zhang, Ryan Maloney, Yonglan Liu, Chung-Jung Tsai, and Hyunbum Jang. Allostery: allosteric cancer drivers and innovative allosteric drugs. *Journal of molecular biology*, page 167569, 2022.

[19] Jean-Pierre Changeux. Allostery and the monod-wyman-changeux model after 50 years. *Annual review of biophysics*, 41:103–133, 2012.

[20] William A Eaton, Eric R Henry, James Hofrichter, and Andrea Mozzarelli. Is cooperative oxygen binding by hemoglobin really understood? *Nature structural biology*, 6(4):351–358, 1999.

[21] TAJ Duke and Dennis Bray. Heightened sensitivity of a lattice of membrane receptors. *Proceedings of the National Academy of Sciences*, 96(18):10104–10108, 1999.

[22] Jean-Pierre Changeux and Stuart J Edelstein. Allosteric mechanisms of signal transduction. *Science*, 308(5727):1424–1428, 2005.

[23] Daniel E Koshland Jr, George Némethy, and David Filmer. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5(1):365–385, 1966.

[24] A Cooper and DTF Dryden. Allostery without conformational change. *European Biophysics Journal*, 11(2):103–109, 1984.

[25] Nikolay V Dokholyan. Controlling allosteric networks in proteins. *Chemical reviews*, 116(11):6463–6487, 2016.

[26] Vincent J Hilser, James O Wrabl, and Hesam N Motlagh. Structural and energetic basis of allostery. *Annual review of biophysics*, 41:585–609, 2012.

[27] Hesam N Motlagh, James O Wrabl, Jing Li, and Vincent J Hilser. The ensemble nature of allostery. *Nature*, 508(7496):331–339, 2014.

[28] Nataliya Popovych, Shangjin Sun, Richard H Ebright, and Charalampos G Kalodimos. Dynamically driven protein allostery. *Nature structural & molecular biology*, 13(9):831–838, 2006.

[29] David D Boehr, Ruth Nussinov, and Peter E Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology*, 5(11):789–796, 2009.

[30] Qiang Cui and Martin Karplus. Allostery and cooperativity revisited. *Protein science*, 17(8):1295–1307, 2008.

[31] Stefano Gianni and Per Jemth. Allostery frustrates the experimentalist. *Journal of Molecular Biology*, page 167934, 2022.

[32] Diego U Ferreiro, Elizabeth A Komives, and Peter G Wolynes. Frustration, function and folding. *Current opinion in structural biology*, 48:68–73, 2018.

[33] Candice Gautier, Louise Laursen, Per Jemth, and Stefano Gianni. Seeking allosteric networks in pdz domains. *Protein Engineering, Design and Selection*, 31(10):367–373, 2018.

[34] Alexander S Hauser, Sreenivas Chavali, Ikuo Masuho, Leonie J Jahn, Kirill A Martemyanov, David E Gloriam, and M Madan Babu. Pharmacogenomics of gpcr drug targets. *Cell*, 172(1-2):41–54, 2018.

[35] Daniel Wacker, Raymond C Stevens, and Bryan L Roth. How ligands illuminate gpcr molecular pharmacology. *Cell*, 170(3):414–427, 2017.

[36] Alfred G Gilman. G proteins: transducers of receptor-generated signals. *Annual review of biochemistry*, 56(1):615–649, 1987.

[37] Nigel Chaffey. Alberts, b., johnson, a., lewis, j., raff, m., roberts, k. and walter, p. molecular biology of the cell. 4th edn., 2003.

[38] N Dhanasekaran and Jonathan M Dermott. Signaling by the g12 class of g proteins. *Cellular signalling*, 8(4):235–245, 1996.

[39] Najeah Okashah, Qingwen Wan, Soumadwip Ghosh, Manbir Sandhu, Asuka Inoue, Nagarajan Vaidehi, and Nevin A Lambert. Variable g protein determinants of gpcr coupling selectivity. *Proceedings of the National Academy of Sciences*, 116(24):12054–12059, 2019.

[40] Manbir Sandhu, Anja M Touma, Matthew Dysthe, Fredrik Sadler, Sivaraj Sivaramakrishnan, and Nagarajan Vaidehi. Conformational plasticity of the intracellular cavity of gpcr- g-protein complexes leads to g-protein promiscuity and selectivity. *Proceedings of the National Academy of Sciences*, 116(24):11956–11965, 2019.

[41] Luis Jaimes Santiago and Ravinder Abrol. Understanding g protein selectivity of muscarinic acetylcholine receptors using computational methods. *International Journal of Molecular Sciences*, 20(21):5290, 2019.

[42] Evan H Hurowitz, James M Melnyk, Yu-Jiun Chen, Hosein Kouros-Mehr, Melvin I Simon, and Hiroaki Shizuya. Genomic characterization of the human heterotrimeric g protein α, β, and γ subunit genes. *DNA research*, 7(2):111–120, 2000.

[43] DR Brandt and EM Ross. Gtpase activity of the stimulatory gtp-binding regulatory protein of adenylate cyclase, gs. accumulation and turnover of enzyme-nucleotide intermediates. *Journal of Biological Chemistry*, 260(1):266–272, 1985.

[44] Yuan Lin and Alan V Smrcka. Understanding molecular recognition by g protein $\beta\gamma$ subunits on the path to pharmacological targeting. *Molecular pharmacology*, 80(4):551–557, 2011.

[45] Diomedes E Logothetis, Yoshihisa Kurachi, Jonas Galper, Eva J Neer, and David E Clapham. The $\beta\gamma$ subunits of gtp-binding proteins activate the muscarinic k+ channel in heart. *Nature*, 325(6102):321–326, 1987.

[46] Wei-Jen Tang and Alfred G Gilman. Type-specific regulation of adenylyl cyclase by g protein $\beta\gamma$ subunits. *Science*, 254(5037):1500–1503, 1991.

[47] Jonathan L Blank, Kathleen A Brattain, and John H Exton. Activation of cytosolic phosphoinositide phospholipase c by g-protein beta gamma subunits. *Journal of Biological Chemistry*, 267(32):23069–23075, 1992.

[48] Alan V Smrcka and Paul C Sternweis. Regulation of purified subtypes of phosphatidylinositol-specific phospholipase c beta by g protein alpha and beta gamma subunits. *Journal of Biological Chemistry*, 268(13):9667–9674, 1993.

[49] Gerald W Zamponi, Emmanuel Bourinet, Donald Nelson, Joel Nargeot, and Terry P Snutch. Crosstalk between g proteins and protein kinase c mediated by the calcium channel α1 subunit. *Nature*, 385(6615):442–446, 1997.

[50] Shahriar M Khan, Adam Min, Sarah Gora, Geeda M Houranieh, Rhiannon Campden, Mélanie Robitaille, Phan Trieu, Darlaine Pétrin, Ashley M Jacobi, Mark A Behlke, et al. G$\beta 4\gamma 1$ as a modulator of m3 muscarinic receptor signalling and novel roles of g$\beta 1$ subunits in the modulation of cellular signalling. *Cellular signalling*, 27(8):1597–1608, 2015.

[51] Robert J Lefkowitz and Sudha K Shenoy. Transduction of receptor signals by ß-arrestins. *Science*, 308(5721):512–517, 2005.

[52] Vsevolod V Gurevich and Eugenia V Gurevich. The structural basis of arrestin-mediated regulation of g-protein-coupled receptors. *Pharmacology & therapeutics*, 110(3):465–502, 2006.

[53] Yanyong Kang, X Edward Zhou, Xiang Gao, Yuanzheng He, Wei Liu, Andrii Ishchenko, Anton Barty, Thomas A White, Oleksandr Yefanov, Gye Won Han, et al. Crystal structure of rhodopsin bound to arrestin by femtosecond x-ray laser. *Nature*, 523(7562):561–567, 2015.

[54] Catherine AC Moore, Shawn K Milano, and Jeffrey L Benovic. Regulation of receptor trafficking by grks and arrestins. *Annu. Rev. Physiol.*, 69:451–482, 2007.

[55] Yuri K Peterson and Louis M Luttrell. The diverse roles of arrestin scaffolds in g protein–coupled receptor signaling. *Pharmacological reviews*, 69(3):256–297, 2017.

[56] Patrick Scheerer, Jung Hee Park, Peter W Hildebrand, Yong Ju Kim, Norbert Krauß, Hui-Woog Choe, Klaus Peter Hofmann, and Oliver P Ernst. Crystal structure of opsin in its g-protein-interacting conformation. *Nature*, 455(7212):497–502, 2008.

[57] Søren GF Rasmussen, Brian T DeVree, Yaozhong Zou, Andrew C Kruse, Ka Young Chung, Tong Sun Kobilka, Foon Sun Thian, Pil Seok Chae, Els Pardon, Diane Calinski, et al. Crystal structure of the β 2 adrenergic receptor–gs protein complex. *Nature*, 477(7366):549, 2011.

[58] Yanyong Kang, Oleg Kuybeda, Parker W de Waal, Somnath Mukherjee, Ned Van Eps, Przemyslaw Dutka, X Edward Zhou, Alberto Bartesaghi, Satchal Erramilli, Takefumi Morizumi, et al. Cryo-em structure of human rhodopsin bound to an inhibitory g protein. *Nature*, 558(7711):553, 2018.

[59] Antoine Koehl, Hongli Hu, Shoji Maeda, Yan Zhang, Qianhui Qu, Joseph M Paggi, Naomi R Latorraca, Daniel Hilger, Roger Dawson, Hugues Matile, et al. Structure of the μ-opioid receptor–g i protein complex. *Nature*, 558(7711):547, 2018.

[60] Javier García-Nafría, Rony Nehmé, Patricia C Edwards, and Christopher G Tate. Cryo-em structure of the serotonin 5-ht 1b receptor coupled to heterotrimeric g o. *Nature*, 558(7711):620, 2018.

[61] Javier García-Nafría, Yang Lee, Xiaochen Bai, Byron Carpenter, and Christopher G Tate. Cryo-em structure of the adenosine a2a receptor coupled to an engineered heterotrimeric g protein. *Elife*, 7:e35946, 2018.

[62] Shoji Maeda, Qianhui Qu, Michael J Robertson, Georgios Skiniotis, and Brian K Kobilka. Structures of the m1 and m2 muscarinic acetylcholine receptor/g-protein complexes. *Science*, 364(6440):552–557, 2019.

[63] Anne Grahl, Layara Akemi Abiko, Shin Isogai, Timothy Sharpe, and Stephan Grzesiek. A high-resolution description of β1-adrenergic receptor functional dynamics and allosteric coupling from backbone nmr. *Nature communications*, 11(1):2216, 2020.

[64] Shuya Kate Huang, Omar Almurad, Reizel J Pejana, Zachary A Morrison, Aditya Pandey, Louis-Philippe Picard, Mark Nitz, Adnan Sljoka, and R Scott Prosser. Allosteric modulation of the adenosine a2a receptor by cholesterol. *Elife*, 11:e73901, 2022.

[65] Shuya Kate Huang, Louis-Philippe Picard, Rima SM Rahmatullah, Aditya Pandey, Ned Van Eps, Roger K Sunahara, Oliver P Ernst, Adnan Sljoka, and R Scott Prosser. Mapping the conformational landscape of the stimulatory heterotrimeric g protein. *Nature Structural & Molecular Biology*, pages 1–10, 2023.

[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 β2-adrenergic receptor. *Proceedings of the National Academy of Sciences*, 108(46):18684–18689, 2011.

[67] Claudio M Costa-Neto, Lucas T Parreiras-e Silva, and Michel Bouvier. A pluridimensional view of biased agonism. *Molecular pharmacology*, 90(5):587–595, 2016.

[68] Carmen Klein Herenbrink, David A Sykes, Prashant Donthamsetti, Meritxell Canals, Thomas Coudrat, Jeremy Shonberg, Peter J Scammells, Ben Capuano, Patrick M Sexton, Steven J Charlton, et al. The role of kinetic context in apparent biased agonism at gpcrs. *Nature communications*, 7:10842, 2016.

[69] Daniel Wacker, Chong Wang, Vsevolod Katritch, Gye Won Han, Xi-Ping Huang, Eyal Vardy, John D McCorvy, Yi Jiang, Meihua Chu, Fai Yiu Siu, et al. Structural features for functional selectivity at serotonin receptors. *Science*, 340(6132):615–619, 2013.

[70] Laura M Wingler, Matthias Elgeti, Daniel Hilger, Naomi R Latorraca, Michael T Lerch, Dean P Staus, Ron O Dror, Brian K Kobilka, Wayne L Hubbell, and Robert J Lefkowitz. Angiotensin analogs with divergent bias stabilize distinct receptor conformations. *Cell*, 176(3):468–478, 2019.

[71] Zoran Rankovic, Tarsis F Brust, and Laura M Bohn. Biased agonism: An emerging paradigm in gpcr drug discovery. *Bioorganic & medicinal chemistry letters*, 26(2):241–250, 2016.

[72] Peter Kolb, Terry Kenakin, Stephen PH Alexander, Marcel Bermudez, Laura M Bohn, Christian S Breinholt, Michel Bouvier, Stephen J Hill, Evi Kostenis, Kirill A Martemyanov, et al. Community guidelines for gpcr ligand bias: Iuphar review 32. *British journal of pharmacology*, 179(14):3651–3674, 2022.

[73] Laura M Bohn, Robert J Lefkowitz, Raul R Gainetdinov, Karsten Peppel, Marc G Caron, and Fang-Tsyr Lin. Enhanced morphine analgesia in mice lacking β-arrestin 2. *Science*, 286(5449):2495–2498, 1999.

[74] Amal El Daibani, Joseph M Paggi, Kuglae Kim, Yianni D Laloudakis, Petr Popov, Sarah M Bernhard, Brian E Krumm, Reid HJ Olsen, Jeffrey Diberto, F Ivy Carroll, et al. Molecular mechanism of biased signaling at the kappa opioid receptor. *Nature Communications*, 14(1):1338, 2023.

[75] Edward R Siuda, Richard Carr III, David H Rominger, and Jonathan D Violin. Biased mu-opioid receptor ligands: a promising new generation of pain therapeutics. *Current opinion in pharmacology*, 32:77–84, 2017.

[76] Gáspár Pándy-Szekeres, Jimmy Caroli, Alibek Mamyrbekov, Ali A Kermani, György M Keserű, Albert J Kooistra, and David E Gloriam. Gpcrdb in 2023: state-specific structure models using alphafold2 and new ligand resources. *Nucleic Acids Research*, 51(D1):D395–D402, 2023.

[77] AJ Venkatakrishnan, Xavier Deupi, Guillaume Lebon, Christopher G Tate, Gebhard F Schertler, and M Madan Babu. Molecular signatures of g-protein-coupled receptors. *Nature*, 494(7436):185–194, 2013.

[78] James S Davidson, Colleen A Flanagan, Peter D Davies, Janet Hapgood, David Myburgh, Ricardo Elario, Robert P Millar, Wynn Forrest-Owen, and Craig A McArdle. Incorporation of an additional glycosylation site enhances expression of functional human gonadotropin-releasing hormone receptor. *Endocrine*, 4:207–212, 1996.

[79] Christoffer K Goth, Ulla E Petaja-Repo, and Mette M Rosenkilde. G protein-coupled receptors in the sweet spot: glycosylation and other post-translational modifications. *ACS pharmacology & translational science*, 3(2):237–245, 2020.

[80] Andrija Sente, Raphael Peer, Ashish Srivastava, Mithu Baidya, Arthur M Lesk, Santhanam Balaji, Arun K Shukla, M Madan Babu, and Tilman Flock. Molecular mechanism of modulating arrestin conformation by gpcr phosphorylation. *Nature structural & molecular biology*, 25(6):538–545, 2018.

[81] X Edward Zhou, Yuanzheng He, Parker W de Waal, Xiang Gao, Yanyong Kang, Ned Van Eps, Yanting Yin, Kuntal Pal, Devrishi Goswami, Thomas A White, et al. Identification of phosphorylation codes for arrestin recruitment by g protein-coupled receptors. *Cell*, 170(3):457–469, 2017.

[82] Yasushi Fukushima, Toshihito Saitoh, Motonobu Anai, Takehide Ogihara, Kouichi Inukai, Makoto Funaki, Hideyuki Sakoda, Yukiko Onishi, Hiraku Ono, Midori Fujishiro, et al. Palmitoylation of the canine histamine h2 receptor occurs at cys305 and is important for cell surface targeting. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1539(3):181–191, 2001.

[83] Jennifer Greaves, Gerald R Prescott, Oforiwa A Gorleku, and Luke H Chamberlain. The fat controller: roles of palmitoylation in intracellular protein trafficking and targeting to membrane microdomains. *Molecular membrane biology*, 26(1-2):67–79, 2009.

[84] Sergio Oddi, Enrico Dainese, Simone Sandiford, Filomena Fezza, Mirko Lanuti, Valerio Chiurchiù, Antonio Totaro, Giuseppina Catanzaro, Daniela Barcaroli, Vincenzo De Laurenzi, et al. Effects of palmitoylation of cys415 in helix 8 of the cb1 cannabinoid receptor on membrane localization and signalling. *British journal of pharmacology*, 165(8):2635–2651, 2012.

[85] Margherita Persechino, Janik Björn Hedderich, Peter Kolb, and Daniel Hilger. Allosteric modulation of gpcrs: From structural insights to in silico drug discovery. *Pharmacology & Therapeutics*, page 108242, 2022.

[86] Jeffery M Klco, Gregory V Nikiforovich, and Thomas J Baranski. Genetic analysis of the first and third extracellular loops of the c5a receptor reveals an essential wxfg motif in the first loop. *Journal of Biological Chemistry*, 281(17):12010–12019, 2006.

[87] Stewart D Clark, Ha T Tran, Joanne Zeng, and Rainer K Reinscheid. Importance of extracellular loop one of the neuropeptide s receptor for biogenesis and function. *Peptides*, 31(1):130–138, 2010.

[88] Michael J Rizzo, John P Evans, Morgan Burt, Cecil J Saunders, and Erik C Johnson. Unexpected role of a conserved domain in the first extracellular loop in g protein-coupled receptor trafficking. *Biochemical and biophysical research communications*, 503(3):1919–1926, 2018.

[89] Vindhya Nawaratne, Katie Leach, Christian C Felder, Patrick M Sexton, and Arthur Christopoulos. Structural determinants of allosteric agonism and modulation at the m4 muscarinic acetylcholine receptor: identification of ligand-specific and global activation mechanisms. *Journal of Biological Chemistry*, 285(25):19012–19021, 2010.

[90] Kwang H Ahn, Alexander C Bertalovitz, Dale F Mierke, and Debra A Kendall. Dual role of the second extracellular loop of the cannabinoid receptor 1: ligand binding and receptor localization. *Molecular pharmacology*, 76(4):833–842, 2009.

[91] Arun K Shukla, Garima Singh, and Eshan Ghosh. Emerging structural insights into biased gpcr signaling. *Trends in biochemical sciences*, 39(12):594–602, 2014.

[92] Matthew Conner, Stuart R Hawtin, John Simms, Denise Wootten, Zoe Lawson, Alex C Conner, Rosemary A Parslow, and Mark Wheatley. Systematic analysis of the entire

second extracellular loop of the v1a vasopressin receptor: key residues, conserved throughout a g-protein-coupled receptor family, identified. *Journal of Biological Chemistry*, 282(24):17405–17412, 2007.

[93] Line Barington, Pia C Rummel, Michael Lückmann, Heidi Pihl, Olav Larsen, Viktorija Daugvilaite, Anders H Johnsen, Thomas M Frimurer, Stefanie Karlshøj, and Mette M Rosenkilde. Role of conserved disulfide bridges and aromatic residues in extracellular loop 2 of chemokine receptor ccr8 for chemokine and small molecule binding. *Journal of Biological Chemistry*, 291(31):16208–16220, 2016.

[94] Vimesh A Avlani, Karen J Gregory, Craig J Morton, Michael W Parker, Patrick M Sexton, and Arthur Christopoulos. Critical role for the second extracellular loop in the binding of both orthosteric and allosteric g protein-coupled receptor ligands. *Journal of Biological Chemistry*, 282(35):25677–25686, 2007.

[95] Alaa Abdul-Ridha, Laura Lopez, Peter Keov, David M Thal, Shailesh N Mistry, Patrick M Sexton, J Robert Lane, Meritxell Canals, and Arthur Christopoulos. Molecular determinants of allosteric modulation at the m1 muscarinic acetylcholine receptor. *Journal of Biological Chemistry*, 289(9):6067–6079, 2014.

[96] Antonio G Soto, Thomas H Smith, Buxin Chen, Supriyo Bhattacharya, Isabel Canto Cordova, Terry Kenakin, Nagarajan Vaidehi, and JoAnn Trejo. N-linked glycosylation of protease-activated receptor-1 at extracellular loop 2 regulates g-protein signaling bias. *Proceedings of the National Academy of Sciences*, 112(27):E3600–E3608, 2015.

[97] Karen J Gregory, Nathan E Hall, Andrew B Tobin, Patrick M Sexton, and Arthur Christopoulos. Identification of orthosteric and allosteric site mutations in m2 muscarinic acetylcholine receptors that contribute to ligand-selective signaling bias. *Journal of Biological Chemistry*, 285(10):7459–7474, 2010.

[98] Mark Wheatley, Denise Wootten, Matthew T Conner, John Simms, R Kendrick, Ryan T Logan, David R Poyner, and James Barwell. Lifting the lid on gpcrs: the role of extracellular loops. *British journal of pharmacology*, 165(6):1688–1703, 2012.

[99] Jeffery M Klco, Christina B Wiegand, Kirk Narzinski, and Thomas J Baranski. Essential role for the second extracellular loop in c5a receptor activation. *Nature structural & molecular biology*, 12(4):320–326, 2005.

[100] Andrea N Naranjo, Amy Chevalier, Gregory D Cousins, Esther Ayettey, Emily C McCusker, Carola Wenk, and Anne S Robinson. Conserved disulfide bond is not essential for the adenosine a2a receptor: Extracellular cysteines influence receptor distribution within the cell and ligand-binding recognition. *Biochimica Et Biophysica Acta (BBA)-Biomembranes*, 1848(2):603–614, 2015.

[101] Sheng Wang, Tao Che, Anat Levit, Brian K Shoichet, Daniel Wacker, and Bryan L Roth. Structure of the d2 dopamine receptor bound to the atypical antipsychotic drug risperidone. *Nature*, 555(7695):269, 2018.

[102] Chong Wang, Yi Jiang, Jinming Ma, Huixian Wu, Daniel Wacker, Vsevolod Katritch, Gye Won Han, Wei Liu, Xi-Ping Huang, Eyal Vardy, et al. Structural basis for molecular recognition at serotonin receptors. *Science*, 340(6132):610–614, 2013.

[103] Nicolas A Heyder, Gunnar Kleinau, David Speck, Andrea Schmidt, Sarah Paisdzior, Michal Szczepek, Brian Bauer, Anja Koch, Monique Gallandi, Dennis Kwiatkowski, et al. Structures of active melanocortin-4 receptor–gs-protein complexes with ndp-α-msh and setmelanotide. *Cell Research*, 31(11):1176–1189, 2021.

[104] Beili Wu, Ellen YT Chien, Clifford D Mol, Gustavo Fenalti, Wei Liu, Vsevolod Katritch, Ruben Abagyan, Alexei Brooun, Peter Wells, F Christopher Bi, et al. Structures of the cxcr4 chemokine gpcr with small-molecule and cyclic peptide antagonists. *Science*, 330(6007):1066–1071, 2010.

[105] Haitao Zhang, Hamiyet Unal, Cornelius Gati, Gye Won Han, Wei Liu, Nadia A Zatsepin, Daniel James, Dingjie Wang, Garrett Nelson, Uwe Weierstall, et al. Structure of the angiotensin receptor revealed by serial femtosecond crystallography. *Cell*, 161(4):833–844, 2015.

[106] Dorothea Jäger, Caroline Schmalenbach, Stefanie Prilla, Jasmin Schrobang, Anna Kebig, Matthias Sennwitz, Eberhard Heller, Christian Tränkle, Ulrike Holzgrabe, Hans-Dieter Höltje, et al. Allosteric small molecules unveil a role of an extracellular e2/transmembrane helix 7 junction for g protein-coupled receptor activation. *Journal of Biological Chemistry*, 282(48):34968–34976, 2007.

[107] Maren Claus, Holger Jaeschke, Gunnar Kleinau, Susanne Neumann, Gerd Krause, and Ralf Paschke. A hydrophobic cluster in the center of the third extracellular loop is important for thyrotropin receptor signaling. *Endocrinology*, 146(12):5197–5203, 2005.

[108] Alexandre Connolly, Brian J Holleran, Élie Simard, Jean-Patrice Baillargeon, Pierre Lavigne, and Richard Leduc. Interplay between intracellular loop 1 and helix viii of the angiotensin ii type 2 receptor controls its activation. *Biochemical Pharmacology*, 168:330–338, 2019.

[109] Wanchao Yin, Zhihai Li, Mingliang Jin, Yu-Ling Yin, Parker W De Waal, Kuntal Pal, Yanting Yin, Xiang Gao, Yuanzheng He, Jing Gao, et al. A complex structure of arrestin-2 bound to a g protein-coupled receptor. *Cell research*, 29(12):971–983, 2019.

[110] Roland R Franke, Bernd König, Thomas P Sakmar, H Gobind Khorana, and Klaus P Hofmann. Rhodopsin mutants that bind but fail to activate transducin. *Science*, 250(4977):123–125, 1990.

[111] Jose Manuel Perez-Aguilar, Jufang Shan, Michael V LeVine, George Khelashvili, and Harel Weinstein. A functional selectivity mechanism at the serotonin-2a gpcr involves ligand-dependent conformations of intracellular loop 2. *Journal of the American Chemical Society*, 136(45):16044–16054, 2014.

[112] Fredrik Sadler, Ning Ma, Michael Ritt, Yatharth Sharma, Nagarajan Vaidehi, and Sivaraj Sivaramakrishnan. Autoregulation of gpcr signalling through the third intracellular loop. *Nature*, pages 1–8, 2023.

[113] Chunmin Dong, Catalin M Filipeanu, Matthew T Duvernay, and Guangyu Wu. Regulation of g protein-coupled receptor export trafficking. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1768(4):853–870, 2007.

[114] Jean-Michel Neumann, Alain Couvineau, Samuel Murail, Jean-Jacques Lacapere, Nadege Jamin, and Marc Laburthe. Class-b gpcr activation: is ligand helix-capping the key? *Trends in biochemical sciences*, 33(7):314–319, 2008.

[115] Julie Kniazeff, Laurent Prézeau, Philippe Rondard, Jean-Philippe Pin, and Cyril Goudet. Dimers and beyond: The functional puzzles of class c gpcrs. *Pharmacology & therapeutics*, 130(1):9–25, 2011.

[116] David Jonathan Wasilko, Zachary Lee Johnson, Mark Ammirati, Ye Che, Matthew C Griffor, Seungil Han, and Huixian Wu. Structural basis for chemokine receptor ccr6 activation by the endogenous protein ligand ccl20. *Nature communications*, 11(1):3031, 2020.

[117] Polina Isaikina, Ching-Ju Tsai, Nikolaus Dietz, Filip Pamula, Anne Grahl, Kenneth N Goldie, Ramon Guixà-González, Camila Branco, Marianne Paolini-Bertrand, Nicolas Calo, et al. Structural basis of the activation of the cc chemokine receptor 5 by a chemokine agonist. *Science advances*, 7(25):eabg8685, 2021.

[118] James LJ Coleman, Tony Ngo, and Nicola J Smith. The g protein-coupled receptor n-terminus and receptor signalling: N-tering a new era. *Cellular signalling*, 33:1–9, 2017.

[119] Demet Araç, Antony A Boucard, Marc F Bolliger, Jenna Nguyen, S Michael Soltis, Thomas C Südhof, and Axel T Brunger. A novel evolutionarily conserved domain of cell-adhesion gpcrs mediates autoproteolysis. *The EMBO journal*, 31(6):1364–1378, 2012.

[120] Gabriel S Salzman, Sarah D Ackerman, Chen Ding, Akiko Koide, Katherine Leon, Rong Luo, Hannah M Stoveken, Celia G Fernandez, Gregory G Tall, Xianhua Piao, et al. Structural basis for regulation of gpr56/adgrg1 by its alternatively spliced extracellular domains. *Neuron*, 91(6):1292–1304, 2016.

[121] Fabian Pohl, Florian Seufert, Yin Kwan Chung, Daniela Volke, Ralf Hoffmann, Torsten Schoneberg, Tobias Langenhan, Peter W Hildebrand, and Norbert Strater. Structural basis of gain domain autoproteolysis and cleavage-resistance in the adhesion g-protein coupled receptors. *bioRxiv*, pages 2023–03, 2023.

[122] Louis Dumas, Matthieu Marfoglia, Byeongseon Yang, Mahdi Hijazi, Amede Larabi, Kevin Lau, Florence Pojer, Michael Nash, and Patrick BARTH. Uncovering and engineering

the mechanical properties of the adhesion gpcr adgrg1 gain domain. *bioRxiv*, pages 2023–04, 2023.

[123] Brian L Zhong, Christina E Lee, Vipul T Vachharajani, Thomas C Sudhof, and Alexander R Dunn. Piconewton forces mediate gain domain dissociation of the latrophilin-3 adhesion gpcr. *bioRxiv*, pages 2023–01, 2023.

[124] Chaoyu Fu, Wenmao Huang, Qingnan Tang, Minghui Niu, Shiwen Guo, Tobias Langenhan, Gaojie Song, and Jie Yan. Step-wise mechanical unfolding and dissociation of the gain domains of adgrg1/gpr56, adgrl1/latrophilin-1 and adgrb3/bai3: insights into the mechanical activation hypothesis of adhesion g protein-coupled receptors. *bioRxiv*, pages 2023–03, 2023.

[125] Gerti Beliu, Steffen Altrichter, Ramon Guixà-González, Mareike Hemberger, Ina Brauer, Anne-Kristin Dahse, Nicole Scholz, Robert Wieduwild, Alexander Kuhlemann, Hossein Batebi, et al. Tethered agonist exposure in intact adhesion/class b2 gpcrs through intrinsic structural flexibility of the gain domain. *Molecular cell*, 81(5):905–921, 2021.

[126] Cornelius Krasel, Ulrike Zabel, Kristina Lorenz, Susanne Reiner, Suleiman Al-Sabah, and Martin J Lohse. Dual role of the β2-adrenergic receptor c terminus for the binding of β-arrestin and receptor internalization. *Journal of Biological Chemistry*, 283(46):31840–31848, 2008.

[127] Dean P Staus, Hongli Hu, Michael J Robertson, Alissa LW Kleinhenz, Laura M Wingler, William D Capel, Naomi R Latorraca, Robert J Lefkowitz, and Georgios Skiniotis. Structure of the m2 muscarinic receptor–β-arrestin complex in a lipid nanodisc. *Nature*, 579(7798):297–302, 2020.

[128] Yang Lee, Tony Warne, Rony Nehmé, Shubhi Pandey, Hemlata Dwivedi-Agnihotri, Madhu Chaturvedi, Patricia C Edwards, Javier García-Nafría, Andrew GW Leslie, Arun K Shukla, et al. Molecular basis of β-arrestin coupling to formoterol-bound β1-adrenoceptor. *Nature*, 583(7818):862–866, 2020.

[129] Julien Bous, Aurélien Fouillen, Hélène Orcel, Stefano Trapani, Xiaojing Cong, Simon Fontanel, Julie Saint-Paul, Joséphine Lai-Kee-Him, Serge Urbach, Nathalie Sibille, et al. Structure of the vasopressin hormone–v2 receptor–β-arrestin1 ternary complex. *Science Advances*, 8(35):eabo7761, 2022.

[130] Can Cao, Ximena Barros-Álvarez, Shicheng Zhang, Kuglae Kim, Marc A Dämgen, Ouliana Panova, Carl-Mikael Suomivuori, Jonathan F Fay, Xiaofang Zhong, Brian E Krumm, et al. Signaling snapshots of a serotonin receptor activated by the prototypical psychedelic lsd. *Neuron*, 110(19):3154–3167, 2022.

[131] Jie Heng, Yunfei Hu, Guillermo Pérez-Hernández, Asuka Inoue, Jiawei Zhao, Xiuyan Ma, Xiaoou Sun, Kouki Kawakami, Tatsuya Ikuta, Jienv Ding, et al. Function and dynamics of the intrinsically disordered carboxyl terminus of β2 adrenergic receptor. *Nature Communications*, 14(1):2005, 2023.

[132] Jie Yin, Kuang-Yui M Chen, Mary J Clark, Mahdi Hijazi, Punita Kumari, Xiao-chen Bai, Roger K Sunahara, Patrick Barth, and Daniel M Rosenbaum. Structure of a d2 dopamine receptor–g-protein complex in a lipid membrane. *Nature*, 584(7819):125–129, 2020.

[133] William I Weis and Brian K Kobilka. The molecular basis of g protein–coupled receptor activation. *Annual review of biochemistry*, 87:897–919, 2018.

[134] Krzysztof Palczewski, Takashi Kumasaka, Tetsuya Hori, Craig A Behnke, Hiroyuki Motoshima, Brian A Fox, Isolde Le Trong, David C Teller, Tetsuji Okada, Ronald E Stenkamp, et al. Crystal structure of rhodopsin: Ag protein-coupled receptor. *science*, 289(5480):739–745, 2000.

[135] Tod D Romo, Alan Grossfield, and Michael C Pitman. Concerted interconversion between ionic lock substates of the β2 adrenergic receptor revealed by microsecond timescale molecular dynamics. *Biophysical journal*, 98(1):76–84, 2010.

[136] Louise Valentin-Hansen, Marleen Groenen, Rie Nygaard, Thomas M Frimurer, Nicholas D Holliday, and Thue W Schwartz. The arginine of the dry motif in transmembrane segment iii functions as a balancing micro-switch in the activation of the β2-adrenergic receptor. *Journal of Biological Chemistry*, 287(38):31973–31982, 2012.

[137] Begonia Y Ho, Andreas Karschin, Theresa Branchek, Norman Davidson, and Henry A Lester. The role of conserved aspartate and serine residues in ligand binding and in function of the 5-ht1a receptor: a site-directed mutation study. *FEBS letters*, 312(2-3):259–262, 1992.

[138] Martijn Bruysters, Heinz H Pertz, Aloys Teunissen, Remko A Bakker, Michel Gillard, Pierre Chatelain, Walter Schunack, Henk Timmerman, and Rob Leurs. Mutational analysis of the histamine h1-receptor binding pocket of histaprodifens. *European journal of pharmacology*, 487(1-3):55–63, 2004.

[139] Tomasz Maciej Stepniewski, Arturo Mancini, Richard Ågren, Mariona Torrens-Fontanals, Meriem Semache, Michel Bouvier, Kristoffer Sahlholm, Billy Breton, and Jana Selent. Mechanistic insights into dopaminergic and serotonergic neurotransmission–concerted interactions with helices 5 and 6 drive the functional outcome. *Chemical Science*, 12(33):10990–11003, 2021.

[140] Tony Warne, Patricia C Edwards, Andrew S Doré, Andrew GW Leslie, and Christopher G Tate. Molecular basis for high-affinity agonist binding in gpcrs. *Science*, 364(6442):775–778, 2019.

[141] Manbir Sandhu, Aaron Cho, Ning Ma, Elizaveta Mukhaleva, Yoon Namkung, Sangbae Lee, Soumadwip Ghosh, John H Lee, David E Gloriam, Stéphane A Laporte, et al. Dynamic spatiotemporal determinants modulate gpcr: G protein coupling selectivity and promiscuity. *Nature Communications*, 13(1):7428, 2022.

[142] Alexander S Hauser, Charlotte Avet, Claire Normand, Arturo Mancini, Asuka Inoue, Michel Bouvier, and David E Gloriam. Common coupling map advances gpcr-g protein selectivity. *Elife*, 11:e74107, 2022.

[143] Mickey Kosloff, Emil Alexov, Vadim Y Arshavsky, and Barry Honig. Electrostatic and lipid anchor contributions to the interaction of transducin with membranes: mechanistic implications for activation and translocation. *Journal of Biological Chemistry*, 283(45):31197–31207, 2008.

[144] Punita Kumari, Ashish Srivastava, Ramanuj Banerjee, Eshan Ghosh, Pragya Gupta, Ravi Ranjan, Xin Chen, Bhagyashri Gupta, Charu Gupta, Deepika Jaiman, et al. Functional competence of a partially engaged gpcr–β-arrestin complex. *Nature communications*, 7(1):13416, 2016.

[145] Kun Chen, Chenhui Zhang, Shuling Lin, Xinyu Yan, Heng Cai, Cuiying Yi, Limin Ma, Xiaojing Chu, Yuchen Liu, Ya Zhu, et al. Tail engagement of arrestin at the glucagon receptor. *Nature*, pages 1–7, 2023.

[146] Thomas J Cahill III, Alex RB Thomsen, Jeffrey T Tarrasch, Bianca Plouffe, Anthony H Nguyen, Fan Yang, Li-Yin Huang, Alem W Kahsai, Daniel L Bassoni, Bryant J Gavino, et al. Distinct conformations of gpcr–β-arrestin complexes mediate desensitization, signaling, and endocytosis. *Proceedings of the National Academy of Sciences*, 114(10):2562–2567, 2017.

[147] Jagannath Maharana, Parishmita Sarma, Manish K. Yadav, Sayantan Saha, Vinay Singh, Shirsha Saha, Mohamed Chami, Ramanuj Banerjee, and Arun K. Shukla. Structural snapshots uncover a key phosphorylation motif in gpcrs driving β-arrestin activation. *Molecular Cell*, 2023.

[148] Jak Grimes, Zsombor Koszegi, Yann Lanoiselée, Tamara Miljus, Shannon L O'Brien, Tomasz M Stepniewski, Brian Medel-Lacruz, Mithu Baidya, Maria Makarova, Ravi Mistry, et al. Plasma membrane preassociation drives β-arrestin coupling to receptors and activation. *Cell*, 186(10):2238–2255, 2023.

[149] Olgun Guvench and Alexander D MacKerell. Comparison of protein force fields for molecular dynamics simulations. *Molecular modeling of proteins*, pages 63–88, 2008.

[150] Peng Xu, Emilie B Guidez, Colleen Bertoni, and Mark S Gordon. Perspective: Ab initio force field methods derived from quantum mechanics. *The Journal of Chemical Physics*, 148(9):090901, 2018.

[151] Paraskevi Gkeka, Gabriel Stoltz, Amir Barati Farimani, Zineb Belkacemi, Michele Ceriotti, John D Chodera, Aaron R Dinner, Andrew L Ferguson, Jean-Bernard Maillet, Hervé Minoux, et al. Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems. *Journal of chemical theory and computation*, 16(8):4757–4775, 2020.

[152] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, 81(1):511–519, 1984.

[153] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

[154] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1):014101, 2007.

[155] Giovanni Bussi and Michele Parrinello. Accurate sampling using langevin dynamics. *Physical Review E*, 75(5):056707, 2007.

[156] Mahdi Hijazi, David M Wilkins, and Michele Ceriotti. Fast-forward langevin dynamics with momentum flips. *The Journal of Chemical Physics*, 148(18):184109, 2018.

[157] Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

[158] Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. Protein structure prediction using rosetta. In *Methods in enzymology*, volume 383, pages 66–93. Elsevier, 2004.

[159] Rhiju Das and David Baker. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.*, 77:363–382, 2008.

[160] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[161] Brian Kuhlman, Gautam Dantas, Gregory C Ireton, Gabriele Varani, Barry L Stoddard, and David Baker. Design of a novel globular protein fold with atomic-level accuracy. *science*, 302(5649):1364–1368, 2003.

[162] Peilong Lu, Duyoung Min, Frank DiMaio, Kathy Y Wei, Michael D Vahey, Scott E Boyken, Zibo Chen, Jorge A Fallas, George Ueda, William Sheffler, et al. Accurate computational design of multipass transmembrane proteins. *Science*, 359(6379):1042–1046, 2018.

[163] Lin Jiang, Eric A Althoff, Fernando R Clemente, Lindsey Doyle, Daniela Rothlisberger, Alexandre Zanghellini, Jasmine L Gallaher, Jamie L Betker, Fujie Tanaka, Carlos F Barbas III, et al. De novo computational design of retro-aldol enzymes. *science*, 319(5868):1387–1391, 2008.

[164] Daniel-Adriano Silva, Shawn Yu, Umut Y Ulge, Jamie B Spangler, Kevin M Jude, Carlos Labão-Almeida, Lestat R Ali, Alfredo Quijano-Rubio, Mikel Ruterbusch, Isabel Leung, et al. De novo design of potent and selective mimics of il-2 and il-15. *Nature*, 565(7738):186–191, 2019.

[165] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[166] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, pages 2022–12, 2022.

[167] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[168] Lucas Gregorio Nivón, Rocco Moretti, and David Baker. A pareto-optimal refinement method for protein design scaffolds. *PloS one*, 8(4):e59004, 2013.

[169] Hahnbeom Park, Philip Bradley, Per Greisen Jr, Yuan Liu, Vikram Khipple Mulligan, David E Kim, David Baker, and Frank DiMaio. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *Journal of chemical theory and computation*, 12(12):6201–6212, 2016.

[170] Po-Ssu Huang, Yih-En Andrew Ban, Florian Richter, Ingemar Andre, Robert Vernon, William R Schief, and David Baker. Rosettaremodel: a generalized framework for flexible backbone protein design. *PloS one*, 6(8):e24109, 2011.

[171] P Barth, Jack Schonbrun, and David Baker. Toward high-resolution prediction and design of transmembrane helical protein structures. *Proceedings of the National Academy of Sciences*, 104(40):15682–15687, 2007.

[172] Monique M Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical review letters*, 77(9):1905, 1996.

[173] Ivet Bahar, Timothy R Lezon, Ahmet Bakan, and Indira H Shrivastava. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chemical reviews*, 110(3):1463–1497, 2010.

[174] Basak Isin, AJ Rader, Harpreet Kaur Dhiman, Judith Klein-Seetharaman, and Ivet Bahar. Predisposition of the dark state of rhodopsin to functional changes in structure. *PROTEINS: Structure, Function, and Bioinformatics*, 65(4):970–983, 2006.

[175] Chung-Jung Tsai and Ruth Nussinov. A unified view of "how allostery works". *PLoS computational biology*, 10(2), 2014.

[176] Paul Leff. The two-state model of receptor activation. *Trends in pharmacological sciences*, 16(3):89–97, 1995.

[177] Patrick Weinkam, Jaume Pons, and Andrej Sali. Structure-based model of allostery predicts coupling between distant sites. *Proceedings of the National Academy of Sciences*, 109(13):4875–4880, 2012.

[178] Gerald M Carlson and Aron W Fenton. What mutagenesis can and cannot reveal about allostery. *Biophysical Journal*, 110(9):1912–1923, 2016.

[179] Li-Juan Ma, Iris Ohmert, and Vitya Vardanyan. Allosteric features of kcnq1 gating revealed by alanine scanning mutagenesis. *Biophysical Journal*, 100(4):885–894, 2011.

[180] Qingling Tang and Aron W Fenton. Whole-protein alanine-scanning mutagenesis of allostery: A large percentage of a protein can contribute to mechanism. *Human Mutation*, 38(9):1132–1143, 2017.

[181] Pradeep Bandaru, Neel H Shah, Moitrayee Bhattacharyya, John P Barton, Yasushi Kondo, Joshua C Cofsky, Christine L Gee, Arup K Chakraborty, Tanja Kortemme, Rama Ranganathan, et al. Deconstruction of the ras switching cycle through saturation mutagenesis. *Elife*, 6:e27810, 2017.

[182] Christopher JP Mathy, Parul Mishra, Julia M Flynn, Tina Perica, David Mavor, Daniel NA Bolon, and Tanja Kortemme. A complete allosteric map of a gtpase switch in its native cellular network. *Cell Systems*, 2023.

[183] Eric M Jones, Nathan B Lubock, AJ Venkatakrishnan, Jeffrey Wang, Alex M Tseng, Joseph M Paggi, Naomi R Latorraca, Daniel Cancilla, Megan Satyadi, Jessica E Davis, et al. Structural and functional characterization of g protein–coupled receptors with deep mutational scanning. *Elife*, 9:e54895, 2020.

[184] James R Thompson, Jessica K Bell, Judy Bratt, Gregory A Grant, and Leonard J Banaszak. V max regulation through domain and subunit changes. the active form of phosphoglycerate dehydrogenase. *Biochemistry*, 44(15):5763–5773, 2005.

[185] Igor A Shumilin, Chang Zhao, Ronald Bauerle, and Robert H Kretsinger. Allosteric inhibition of 3-deoxy-d-arabino-heptulosonate-7-phosphate synthase alters the coordination of both substrates. *Journal of molecular biology*, 320(5):1147–1156, 2002.

[186] Stephen M Soisson, Beth MacDougall-Shackleton, Robert Schleif, and Cynthia Wolberger. Structural basis for ligand-regulated oligomerization of arac. *Science*, 276(5311):421–425, 1997.

[187] Chung-Jung Tsai and Ruth Nussinov. Emerging allosteric mechanism of egfr activation in physiological and pathological contexts. *Biophysical journal*, 117(1):5–13, 2019.

[188] Roman A Laskowski, Fabian Gerick, and Janet M Thornton. The structural basis of allosteric regulation in proteins. *FEBS letters*, 583(11):1692–1698, 2009.

[189] Krzysztof Palczewski, Takashi Kumasaka, Tetsuya Hori, Craig A Behnke, Hiroyuki Motoshima, Brian A Fox, Isolde Le Trong, David C Teller, Tetsuji Okada, Ronald E Stenkamp, et al. Crystal structure of rhodopsin: Ag protein-coupled receptor. *science*, 289(5480):739–745, 2000.

[190] Bingfa Sun, Dan Feng, Matthew Ling-Hon Chu, Inbar Fish, Silvia Lovera, Zara A Sands, Sebastian Kelm, Anne Valade, Martyn Wood, Tom Ceska, et al. Crystal structure of dopamine d1 receptor in complex with g protein and a non-catechol agonist. *Nature communications*, 12(1):3305, 2021.

[191] Javier García-Nafría and Christopher G Tate. Structure determination of gpcrs: cryo-em compared with x-ray crystallography. *Biochemical Society Transactions*, 49(5):2345–2355, 2021.

[192] Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.

[193] Takanori Nakane, Abhay Kotecha, Andrija Sente, Greg McMullan, Simonas Masiulis, Patricia MGE Brown, Ioana T Grigoras, Lina Malinauskaite, Tomas Malinauskas, Jonas Miehling, et al. Single-particle cryo-em at atomic resolution. *Nature*, 587(7832):152–156, 2020.

[194] Javier García-Nafría and Christopher G Tate. Cryo-em structures of gpcrs coupled to gs, gi and go. *Molecular and cellular endocrinology*, 488:1–13, 2019.

[195] Qiuyan Chen, Manolo Plasencia, Zhuang Li, Somnath Mukherjee, Dhabaleswar Patra, Chun-Liang Chen, Thomas Klose, Xin-Qiu Yao, Anthony A Kossiakoff, Leifu Chang, et al. Structures of rhodopsin in complex with g-protein-coupled receptor kinase 1. *Nature*, 595(7868):600–605, 2021.

[196] Youwen Zhuang, Brian Krumm, Huibing Zhang, X Edward Zhou, Yue Wang, Xi-Ping Huang, Yongfeng Liu, Xi Cheng, Yi Jiang, Hualiang Jiang, et al. Mechanism of dopamine binding and allosteric modulation of the human d1 dopamine receptor. *Cell research*, 31(5):593–596, 2021.

[197] Makaia M Papasergi-Scott, Guillermo Perez-Hernnadez, Hossein Batebi, Yang Gao, Gozde Eskici, Alpay B Seven, Ouliana Panova, Daniel Hilger, Marina Casiraghi, Feng He, et al. Time-resolved cryo-em of g protein activation by a gpcr. *bioRxiv*, pages 2023–03, 2023.

[198] Ernesto J Fuentes, Channing J Der, and Andrew L Lee. Ligand-dependent dynamics and intramolecular signaling in a pdz domain. *Journal of molecular biology*, 335(4):1105–1115, 2004.

[199] Dzmitry Ashkinadze, Harindranath Kadavath, Aditya Pokharna, Celestine N Chi, Michael Friedmann, Dean Strotz, Pratibha Kumari, Martina Minges, Riccardo Cadalbert, Stefan Königl, et al. Atomic resolution protein allostery from the multi-state structure of a pdz domain. *Nature Communications*, 13(1):6232, 2022.

[200] Dean Strotz, Julien Orts, Harindranath Kadavath, Michael Friedmann, Dhiman Ghosh, Simon Olsson, Celestine N Chi, Aditya Pokharna, Peter Güntert, Beat Vögeli, et al. Protein allostery at atomic resolution. *Angewandte Chemie International Edition*, 59(49):22132–22139, 2020.

[201] Shuya Kate Huang and R Scott Prosser. Dynamics and mechanistic underpinnings to pharmacology of class a gpcrs: an nmr perspective. *American Journal of Physiology-Cell Physiology*, 322(4):C739–C753, 2022.

[202] Shin Isogai, Xavier Deupi, Christian Opitz, Franziska M Heydenreich, Ching-Ju Tsai, Florian Brueckner, Gebhard FX Schertler, Dmitry B Veprintsev, and Stephan Grzesiek. Backbone nmr reveals allosteric signal transduction networks in the β1-adrenergic receptor. *Nature*, 530(7589):237–241, 2016.

[203] Libin Ye, Chris Neale, Adnan Sljoka, Brent Lyda, Dmitry Pichugin, Nobuyuki Tsuchimura, Sacha T Larda, Régis Pomès, Angel E García, Oliver P Ernst, et al. Mechanistic insights into allosteric regulation of the a2a adenosine g protein-coupled receptor by physiological cations. *Nature communications*, 9(1):1372, 2018.

[204] Michael T Lerch, Rachel A Matt, Matthieu Masureel, Matthias Elgeti, Kaavya Krishna Kumar, Daniel Hilger, Bryon Foys, Brian K Kobilka, and Wayne L Hubbell. Viewing rare conformations of the β2 adrenergic receptor with pressure-resolved deer spectroscopy. *Proceedings of the National Academy of Sciences*, 117(50):31824–31831, 2020.

[205] Jiawei Zhao, Matthias Elgeti, Evan O'Brien, Cecilia Sar, Amal El Daibani, Jie Heng, Xiaoou Sun, Tao Che, Wayne L Hubbell, Brian Kobilka, et al. Conformational dynamics of the μ-opioid receptor determine ligand intrinsic efficacy. *bioRxiv*, pages 2023–04, 2023.

[206] M Tasumi, H Takeuchi, S Ataka, AM Dwivedi, and S Krimm. Normal vibrations of proteins: Glucagon. *Biopolymers: Original Research on Biomolecules*, 21(3):711–714, 1982.

[207] Michael Levitt, Christian Sander, and Peter S Stern. The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *International Journal of Quantum Chemistry*, 24(S10):181–199, 1983.

[208] Nobuhiro Go, Tosiyuki Noguti, and Testuo Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proceedings of the National Academy of Sciences*, 80(12):3696–3700, 1983.

[209] Konrad Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function, and Bioinformatics*, 33(3):417–429, 1998.

[210] Wenjun Zheng, Bernard R Brooks, and D Thirumalai. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proceedings of the National Academy of Sciences*, 103(20):7664–7669, 2006.

[211] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

[212] Turkan Haliloglu, Ivet Bahar, and Burak Erman. Gaussian dynamics of folded proteins. *Physical review letters*, 79(16):3090, 1997.

[213] Ali Rana Atilgan, SR Durell, Robert L Jernigan, Melik C Demirel, O Keskin, and Ivet Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

[214] M Gur, E Zomot, and I Bahar. Global motions exhibited by proteins in micro-to milliseconds simulations concur with anisotropic network model predictions. *The Journal of chemical physics*, 139(12):09B612_1, 2013.

[215] Ji Guo Su, Li Sheng Qi, Chun Hua Li, Yan Ying Zhu, Hui Jing Du, Yan Xue Hou, Rui Hao, and Ji Hua Wang. Prediction of allosteric sites on protein surfaces with an elastic-network-model-based thermodynamic method. *Physical Review E*, 90(2):022719, 2014.

[216] Adam T VanWart, John Eargle, Zaida Luthey-Schulten, and Rommie E Amaro. Exploring residue component contributions to dynamical network models of allostery. *Journal of chemical theory and computation*, 8(8):2949–2961, 2012.

[217] Chakra Chennubhotla and Ivet Bahar. Markov propagation of allosteric effects in biomolecular systems: application to groel–groes. *Molecular systems biology*, 2(1), 2006.

[218] Joe G Greener and Michael JE Sternberg. Allopred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC bioinformatics*, 16(1):1–7, 2015.

[219] Lei Yang, Guang Song, and Robert L Jernigan. How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophysical journal*, 93(3):920–929, 2007.

[220] Shoshana J Wodak, Emanuele Paci, Nikolay V Dokholyan, Igor N Berezovsky, Amnon Horovitz, Jing Li, Vincent J Hilser, Ivet Bahar, John Karanicolas, Gerhard Stock, et al. Allostery in its many disguises: from theory to applications. *Structure*, 27(4):566–578, 2019.

[221] Samuel Hertig, Naomi R Latorraca, and Ron O Dror. Revealing atomic-level mechanisms of protein allostery with molecular dynamics simulations. *PLoS computational biology*, 12(6), 2016.

[222] Toshiko Ichiye and Martin Karplus. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Bioinformatics*, 11(3):205–217, 1991.

[223] PH Hünenberger, AE Mark, and WF Van Gunsteren. Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *Journal of molecular biology*, 252(4):492–503, 1995.

[224] Oliver F Lange and Helmut Grubmüller. Generalized correlation for biomolecular dynamics. *Proteins: Structure, Function, and Bioinformatics*, 62(4):1053–1061, 2006.

[225] Christopher L McClendon, Gregory Friedland, David L Mobley, Homeira Amirkhani, and Matthew P Jacobson. Quantifying correlations between allosteric sites in thermodynamic ensembles. *Journal of chemical theory and computation*, 5(9):2486–2502, 2009.

[226] Aysima Hacisuleyman and Burak Erman. Entropy transfer between residue pairs and allostery in proteins: quantifying allosteric communication in ubiquitin. *PLoS computational biology*, 13(1):e1005319, 2017.

[227] Yuchen Yuan, Jiahua Deng, and Qiang Cui. Molecular dynamics simulations establish the molecular basis for the broad allostery hotspot distributions in the tetracycline repressor. *Journal of the American Chemical Society*, 144(24):10870–10887, 2022.

[228] Marcelo CR Melo, Rafael C Bernardi, Cesar De La Fuente-Nunez, and Zaida Luthey-Schulten. Generalized correlation-based dynamical network analysis: a new high-performance approach for identifying allosteric communications in molecular dynamics trajectories. *The Journal of chemical physics*, 153(13):134104, 2020.

[229] S. Bowerman and J. Wereszczynski. Chapter seventeen - detecting allosteric networks using molecular dynamics simulation. In Gregory A. Voth, editor, *Computational Approaches for Studying Enzyme Mechanism Part B*, volume 578 of *Methods in Enzymology*, pages 429–447. Academic Press, 2016.

[230] Elizabeth A Proctor, Pradeep Kota, Andrei A Aleksandrov, Lihua He, John R Riordan, and Nikolay V Dokholyan. Rational coupled dynamics network manipulation rescues disease-relevant mutant cystic fibrosis transmembrane conductance regulator. *Chemical science*, 6(2):1237–1246, 2015.

[231] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[232] Anurag Sethi, John Eargle, Alexis A Black, and Zaida Luthey-Schulten. Dynamical networks in trna: protein complexes. *Proceedings of the National Academy of Sciences*, 106(16):6620–6625, 2009.

[233] S. Bhattacharya and N. Vaidehi. Differences in allosteric communication pipelines in the inactive and active states of a gpcr. *Biophysical Journal*, 107(2):422 – 434, 2014.

[234] João Marcelo Lamim Ribeiro and Marta Filizola. Allostery in g protein-coupled receptors investigated by molecular dynamics simulations. *Current opinion in structural biology*, 55:121–128, 2019.

[235] Ron O Dror, Hillary F Green, Celine Valant, David W Borhani, James R Valcourt, Albert C Pan, Daniel H Arlow, Meritxell Canals, J Robert Lane, Raphaël Rahmani, et al. Structural basis for modulation of a g-protein-coupled receptor by allosteric drugs. *Nature*, 503(7475):295–299, 2013.

[236] Christopher J Draper-Joyce, Ravi Kumar Verma, Mayako Michino, Jeremy Shonberg, Anitha Kopinathan, Carmen Klein Herenbrink, Peter J Scammells, Ben Capuano, Ara M Abramyan, David M Thal, et al. The action of a negative allosteric modulator at the dopamine d2 receptor is dependent upon sodium ions. *Scientific Reports*, 8(1):1–12, 2018.

[237] Balaji Selvam, Zahra Shamsi, and Diwakar Shukla. Universality of the sodium ion binding mechanism in class ag-protein-coupled receptors. *Angewandte Chemie*, 130(12):3102–3107, 2018.

[238] Juan Manuel Ramírez-Anguita, Ismael Rodríguez-Espigares, Ramon Guixà-González, Agostino Bruno, Mariona Torrens-Fontanals, Alejandro Varela-Rial, and Jana Selent. Membrane cholesterol effect on the 5-ht2a receptor: Insights into the lipid-induced modulation of an antipsychotic drug target. *Biotechnology and applied biochemistry*, 65(1):29–37, 2018.

[239] Ramon Guixà-González, José L Albasanz, Ismael Rodriguez-Espigares, Manuel Pastor, Ferran Sanz, Maria Martí-Solano, Moutusi Manna, Hector Martinez-Seara, Peter W Hildebrand, Mairena Martín, et al. Membrane cholesterol access into a g-protein-coupled receptor. *Nature Communications*, 8(1):14505, 2017.

[240] Felix Rico, Laura Gonzalez, Ignacio Casuso, Manel Puig-Vidal, and Simon Scheuring. High-speed force spectroscopy unfolds titin at the velocity of molecular dynamics simulations. *science*, 342(6159):741–743, 2013.

[241] Steven Sheridan, Frauke Grater, and Csaba Daday. How fast is too fast in force-probe molecular dynamics simulations? *The Journal of Physical Chemistry B*, 123(17):3658–3664, 2019.

[242] Rommie E Amaro, Anurag Sethi, Rebecca S Myers, V Jo Davisson, and Zaida A Luthey-Schulten. A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase. *Biochemistry*, 46(8):2156–2173, 2007.

[243] Constantin Schoeler, Rafael C Bernardi, Klara H Malinowska, Ellis Durner, Wolfgang Ott, Edward A Bayer, Klaus Schulten, Michael A Nash, and Hermann E Gaub. Mapping mechanical force propagation through biomolecular complexes. *Nano letters*, 15(11):7370–7376, 2015.

[244] Donald Hamelberg, John Mongan, and J Andrew McCammon. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of chemical physics*, 120(24):11919–11929, 2004.

[245] Fuhui Zhang, Yuan Yuan, Haiyan Li, Liting Shen, Yanzhi Guo, Zhining Wen, and Xuemei Pu. Using accelerated molecular dynamics simulation to shed light on the mechanism of activation/deactivation upon mutations for ccr5. *RSC advances*, 8(66):37855–37865, 2018.

[246] Hung Nguyen Do, Jinan Wang, and Yinglong Miao. Deep learning dynamic allostery of g-protein-coupled receptors. *bioRxiv*, pages 2023–01, 2023.

[247] Lingle Wang, Richard A Friesner, and BJ Berne. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (rest2). *The Journal of Physical Chemistry B*, 115(30):9431–9438, 2011.

[248] Xiaojing Cong and Jérôme Golebiowski. Allosteric na+-binding site modulates cxcr4 activation. *Physical Chemistry Chemical Physics*, 20(38):24915–24920, 2018.

[249] Xiaojing Cong, Damien Maurel, Hélène Déméné, Ieva Vasiliauskaité-Brooks, Joanna Hagelberger, Fanny Peysson, Julie Saint-Paul, Jérôme Golebiowski, Sébastien Granier, and Rémy Sounier. Molecular insights into the biased signaling mechanism of the μ-opioid receptor. *Molecular Cell*, 81(20):4165–4175, 2021.

[250] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters*, 100(2):020603, 2008.

[251] Noureldin Saleh, Giorgio Saladino, Francesco Luigi Gervasio, and Timothy Clark. Investigating allosteric effects on the functional dynamics of β2-adrenergic ternary complexes with enhanced-sampling simulations. *Chemical Science*, 8(5):4019–4026, 2017.

[252] Derya Meral, Davide Provasi, and Marta Filizola. An efficient strategy to estimate thermodynamics and kinetics of g protein-coupled receptor activation using metadynamics and maximum caliber. *The Journal of chemical physics*, 149(22):224101, 2018.

[253] Donald Hamelberg, César Augusto F de Oliveira, and J Andrew McCammon. Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. *The Journal of chemical physics*, 127(15):10B614, 2007.

[254] Yinglong Miao, Sara E Nichols, and J Andrew McCammon. Free energy landscape of g-protein coupled receptors, explored by accelerated molecular dynamics. *Physical Chemistry Chemical Physics*, 16(14):6398–6406, 2014.

[255] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1-2):141–151, 1999.

[256] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[257] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[258] Sian Xiao, Gennady M Verkhivker, and Peng Tao. Machine learning and protein allostery. *Trends in Biochemical Sciences*, 2022.

[259] Christoph Wehmeyer and Frank Noé. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of chemical physics*, 148(24):241703, 2018.

[260] Mohammad M. Sultan and Vijay S Pande. tica-metadynamics: accelerating metadynamics by using kinetically selected collective variables. *Journal of chemical theory and computation*, 13(6):2440–2447, 2017.

[261] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

[262] Jingxuan Zhu, Juexin Wang, Weiwei Han, and Dong Xu. Neural relational inference to learn long-range allosteric interactions in proteins from molecular dynamics simulations. *Nature communications*, 13(1):1661, 2022.

[263] Josiah P Zayner, Chloe Antoniou, Alexander R French, Ronald J Hause Jr, and Tobin R Sosnick. Investigating models of protein function and allostery with a widespread mutational analysis of a light-activated protein. *Biophysical journal*, 105(4):1027–1036, 2013.

[264] Xinyi Liu, Shaoyong Lu, Kun Song, Qiancheng Shen, Duan Ni, Qian Li, Xinheng He, Hao Zhang, Qi Wang, Yingyi Chen, et al. Unraveling allosteric landscapes of allosterome with asd. *Nucleic acids research*, 48(D1):D394–D401, 2020.

[265] Zhen Wah Tan, Wei-Ven Tee, Enrico Guarnera, and Igor N Berezovsky. Allomaps 2: allosteric fingerprints of the alphafold and pfam-trrosetta predicted structures for engineering and design. *Nucleic Acids Research*, 51(D1):D345–D351, 2023.

[266] Lucas SP Rudden, Mahdi Hijazi, and Patrick Barth. Deep learning approaches for conformational flexibility and switching properties in protein design. *Frontiers in Molecular Biosciences*, page 840, 2022.

[267] Divesh Bhatt and Daniel M Zuckerman. Beyond microscopic reversibility: Are observable nonequilibrium processes precisely reversible? *Journal of chemical theory and computation*, 7(8):2520–2527, 2011.

[268] The MathWorks Inc. Matlab version: 9.11.0 (r2021b), 2022.

[269] The MathWorks Inc. Bioinformatics toolbox version: 4.15.2 (r2021b), 2022.

[270] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[271] Y Matsunaga and Y Sugita. Refining markov state models for conformational dynamics using ensemble-averaged data and time-series trajectories. *The Journal of chemical physics*, 148(24):241731, 2018.

[272] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[273] Benjamin J Killian, Joslyn Yudenfreund Kravitz, Sandeep Somani, Paramita Dasgupta, Yuan-Ping Pang, and Michael K Gilson. Configurational entropy in protein–peptide binding:: Computational study of tsg101 ubiquitin e2 variant domain with an hiv-derived ptap nonapeptide. *Journal of molecular biology*, 389(2):315–335, 2009.

[274] Hanspeter Herzel, AO Schmitt, and W Ebeling. Finite sample effects in sequence analysis. *Chaos, Solitons & Fractals*, 4(1):97–113, 1994.

[275] Ralf Steuer, Jürgen Kurths, Carsten O Daub, Janko Weise, and Joachim Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl_2):S231–S240, 2002.

[276] Aysima Hacisuleyman and Burak Erman. Information flow and allosteric communication in proteins. *The Journal of Chemical Physics*, 156(18), 2022.

[277] Supriyo Bhattacharya and Nagarajan Vaidehi. Differences in allosteric communication pipelines in the inactive and active states of a gpcr. *Biophysical journal*, 107(2):422–434, 2014.

[278] Midhun K Madhu, Kunal Shewani, and Rajesh K Murarka. Biased signaling in mutated variants of β2-adrenergic receptor: Insights from molecular dynamics simulations. *bioRxiv*, pages 2023–09, 2023.

[279] Christopher L McClendon, Lan Hua, Gabriela Barreiro, and Matthew P Jacobson. Comparing conformational ensembles using the kullback–leibler divergence expansion. *Journal of chemical theory and computation*, 8(6):2115–2126, 2012.

[280] Stefan Bietz, Sascha Urbaczek, Benjamin Schulz, and Matthias Rarey. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *Journal of cheminformatics*, 6:1–12, 2014.

[281] Sunhwan Jo, Taehoon Kim, and Wonpil Im. Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PloS one*, 2(9):e880, 2007.

[282] Sunhwan Jo, Joseph B. Lim, Jeffery B. Klauda, and Wonpil Im. Charmm-gui membrane builder for mixed bilayers and its application to yeast membranes. *Biophysical Journal*, 97(1):50 – 58, 2009.

[283] 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-gui membrane builder toward realistic biological membrane simulations. *Journal of Computational Chemistry*, 35(27):1997–2004, 2014.

[284] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, 2005.

[285] Xiao-Ping Xu and David A Case. Probing multiple effects on 15n, 13cα, 13cβ, and 13c chemical shifts in peptides using density functional theory. *Biopolymers: Original Research on Biomolecules*, 65(6):408–423, 2002.

[286] Stephan Grzesiek, Florence Cordier, Victor Jaravine, and Michael Barfield. Insights into biomolecular hydrogen bonds from hydrogen bond scalar couplings. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 45(3):275–300, 2004.

[287] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. Charmm-gui: A web-based graphical user interface for charmm. *Journal of Computational Chemistry*, 29(11):1859–1865, 2008.

[288] Kenno Vanommeslaeghe, Elizabeth Hatcher, Chayan Acharya, Sibsankar Kundu, Shijun Zhong, Jihyun Shim, Eva Darian, Olgun Guvench, P Lopes, Igor Vorobyov, et al. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of computational chemistry*, 31(4):671–690, 2010.

[289] Szilárd Páll, Mark James Abraham, Carsten Kutzner, Berk Hess, and Erik Lindahl. Tackling exascale software challenges in molecular dynamics simulations with gromacs. In *Solving Software Challenges for Exascale: International Conference on Exascale Applications and Software, EASC 2014, Stockholm, Sweden, April 2-3, 2014, Revised Selected Papers 2*, pages 3–27. Springer, 2015.

[290] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[291] Robert B Best, Xiao Zhu, Jihyun Shim, Pedro EM Lopes, Jeetain Mittal, Michael Feig, and Alexander D MacKerell Jr. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ1 and χ2 dihedral angles. *Journal of chemical theory and computation*, 8(9):3257–3273, 2012.

[292] Ruth Nussinov and Chung-Jung Tsai. Allostery in disease and in drug discovery. *Cell*, 153(2):293–305, 2013.

[293] Yun-Min Sung, Angela D Wilkins, Gustavo J Rodriguez, Theodore G Wensel, and Olivier Lichtarge. Intramolecular allosteric communication in dopamine d2 receptor revealed

by evolutionary amino acid covariation. *Proceedings of the National Academy of Sciences*, 113(13):3539–3544, 2016.

[294] Gürol M Süel, Steve W Lockless, Mark A Wall, and Rama Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature structural biology*, 10(1):59–69, 2003.

[295] Marianne O Klein, Daniella S Battagello, Ariel R Cardoso, David N Hauser, Jackson C Bittencourt, and Ricardo G Correa. Dopamine: functions, signaling, and association with neurological diseases. *Cellular and molecular neurobiology*, 39(1):31–59, 2019.

[296] Ernest P Noble. D2 dopamine receptor gene in psychiatric and neurologic disorders and its phenotypes. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 116(1):103–125, 2003.

[297] Robert E Jefferson, Aurélien Oggier, Andreas Füglistaler, Nicolas Camviel, Mahdi Hijazi, Ana Rico Villarreal, Caroline Arber, and Patrick Barth. Computational design of dynamic receptor—peptide signaling complexes applied to chemotaxis. *Nature Communications*, 14(1):2875, 2023.

[298] Michael V LeVine and Harel Weinstein. Nbit-a new information theory-based analysis of allosteric mechanisms reveals residues that underlie function in the leucine transporter leut. *PLoS computational biology*, 10(5):e1003603, 2014.

[299] Luyu Fan, Liang Tan, Zhangcheng Chen, Jianzhong Qi, Fen Nie, Zhipu Luo, Jianjun Cheng, and Sheng Wang. Haloperidol bound d2 dopamine receptor structure inspired the discovery of subtype selective ligands. *Nature communications*, 11(1):1074, 2020.

[300] Dohyun Im, Asuka Inoue, Takaaki Fujiwara, Takanori Nakane, Yasuaki Yamanaka, Tomoko Uemura, Chihiro Mori, Yuki Shiimura, Kanako Terakado Kimura, Hidetsugu Asada, et al. Structure of the dopamine d2 receptor in complex with the antipsychotic drug spiperone. *Nature communications*, 11(1):6442, 2020.

[301] Youwen Zhuang, Peiyu Xu, Chunyou Mao, Lei Wang, Brian Krumm, X Edward Zhou, Sijie Huang, Heng Liu, Xi Cheng, Xi-Ping Huang, et al. Structural insights into the human d1 and d2 dopamine receptor signaling complexes. *Cell*, 184(4):931–942, 2021.

[302] Fan Yang, Shenglong Ling, Yingxin Zhou, Yanan Zhang, Pei Lv, Sanling Liu, Wei Fang, Wenjing Sun, Liaoyuan A Hu, Longhua Zhang, et al. Different conformational responses of the β2-adrenergic receptor-gs complex upon binding of the partial agonist salbutamol or the full agonist isoprenaline. *National Science Review*, 8(9):nwaa284, 2021.

[303] Yanan Zhang, Fan Yang, Shenglong Ling, Pei Lv, Yingxin Zhou, Wei Fang, Wenjing Sun, Longhua Zhang, Pan Shi, and Changlin Tian. Single-particle cryo-em structural studies of the β2ar–gs complex bound with a full agonist formoterol. *Cell discovery*, 6(1):45, 2020.

[304] Cristina Missale, S Russel Nash, Susan W Robinson, Mohamed Jaber, and Marc G Caron. Dopamine receptors: from structure to function. *Physiological reviews*, 78(1):189–225, 1998.

[305] Jean-Martin Beaulieu and Raul R Gainetdinov. The physiology, signaling, and pharmacology of dopamine receptors. *Pharmacological reviews*, 63(1):182–217, 2011.

[306] Peng Xiao, Wei Yan, Lu Gou, Ya-Ni Zhong, Liangliang Kong, Chao Wu, Xin Wen, Yuan Yuan, Sheng Cao, Changxiu Qu, et al. Ligand recognition and allosteric regulation of drd1-gs signaling complexes. *Cell*, 184(4):943–956, 2021.

[307] Peng Zhou, Jianwei Zou, Feifei Tian, and Zhicai Shang. Fluorine bonding how does it work in protein- ligand interactions? *Journal of chemical information and modeling*, 49(10):2344–2355, 2009.

[308] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[309] Xiang Feng, Joaquin Ambia, Kuang-Yui M Chen, Melvin Young, and Patrick Barth. Computational design of ligand-binding membrane receptors with high selectivity. *Nature chemical biology*, 13(7):715, 2017.

[310] Alfredo Quijano-Rubio, Hsien-Wei Yeh, Jooyoung Park, Hansol Lee, Robert A Langan, Scott E Boyken, Marc J Lajoie, Longxing Cao, Cameron M Chow, Marcos C Miranda, et al. De novo design of modular and tunable protein biosensors. *Nature*, 591(7850):482–487, 2021.

[311] Anum A Glasgow, Yao-Ming Huang, Daniel J Mandell, Michael Thompson, Ryan Ritterson, Amanda L Loshbaugh, Jenna Pellegrino, Cody Krivacic, Roland A Pache, Kyle A Barlow, et al. Computational design of a modular protein sense-response system. *Science*, 366(6468):1024–1028, 2019.

[312] Anthony Marchand, Alexandra K Van Hall-Beauvais, and Bruno E Correia. Computational design of novel protein–protein interactions–an overview on methodological approaches and applications. *Current Opinion in Structural Biology*, 74:102370, 2022.

[313] Sarel J Fleishman, Timothy A Whitehead, Damian C Ekiert, Cyrille Dreyfus, Jacob E Corn, Eva-Maria Strauch, Ian A Wilson, and David Baker. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, 332(6031):816–821, 2011.

[314] Rie Nygaard, Yaozhong Zou, Ron O Dror, Thomas J Mildorf, Daniel H Arlow, Aashish Manglik, Albert C Pan, Corey W Liu, Juan José Fung, Michael P Bokoch, et al. The dynamic process of β2-adrenergic receptor activation. *Cell*, 152(3):532–542, 2013.

[315] Libin Ye, Ned Van Eps, Marco Zimmer, Oliver P Ernst, and R Scott Prosser. Activation of the a2a adenosine g-protein-coupled receptor by conformational selection. *Nature*, 533(7602):265–268, 2016.

[316] Peter Csermely, Robin Palotai, and Ruth Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in biochemical sciences*, 35(10):539–546, 2010.

[317] Evangelia Petsalaki and Robert B Russell. Peptide-mediated interactions in biological systems: new discoveries and applications. *Current opinion in biotechnology*, 19(4):344–350, 2008.

[318] Nir London, Barak Raveh, and Ora Schueler-Furman. Peptide docking and structure-based characterization of peptide binding: from knowledge to know-how. *Current opinion in structural biology*, 23(6):894–902, 2013.

[319] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[320] Yipin Lei, Shuya Li, Ziyi Liu, Fangping Wan, Tingzhong Tian, Shao Li, Dan Zhao, and Jianyang Zeng. A deep-learning framework for multi-level peptide–protein interaction prediction. *Nature communications*, 12(1):5465, 2021.

[321] Maciej Ciemny, Mateusz Kurcinski, Karol Kamel, Andrzej Kolinski, Nawsad Alam, Ora Schueler-Furman, and Sebastian Kmiecik. Protein–peptide docking: opportunities and challenges. *Drug discovery today*, 23(8):1530–1537, 2018.

[322] Simon R Foster, Alexander S Hauser, Line Vedel, Ryan T Strachan, Xi-Ping Huang, Ariana C Gavin, Sushrut D Shah, Ajay P Nayak, Linda M Haugaard-Kedström, Raymond B Penn, et al. Discovery of human signaling systems: pairing peptides to g protein-coupled receptors. *Cell*, 179(4):895–908, 2019.

[323] Oanh Vu, Brian Joseph Bender, Lisa Pankewitz, Daniel Huster, Annette G Beck-Sickinger, and Jens Meiler. The structural basis of peptide binding at class ag protein-coupled receptors. *Molecules*, 27(1):210, 2021.

[324] Nawsad Alam and Ora Schueler-Furman. Modeling peptide-protein structure and binding using monte carlo sampling approaches: Rosetta flexpepdock and flexpepbind. *Modeling Peptide-Protein Interactions: Methods and Protocols*, pages 139–169, 2017.

[325] Hui Zhang, Kun Chen, Qiuxiang Tan, Qiang Shao, Shuo Han, Chenhui Zhang, Cuiying Yi, Xiaojing Chu, Ya Zhu, Yechun Xu, et al. Structural basis for chemokine recognition and receptor activation of chemokine receptor ccr5. *Nature communications*, 12(1):4151, 2021.

[326] 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. *Proceedings of the National Academy of Sciences*, 113(35):9928–9933, 2016.

[327] Jeremiah D Heredia, Jihye Park, Riley J Brubaker, Steven K Szymanski, Kevin S Gill, and Erik Procko. Mapping interaction sites on human chemokine receptors by deep mutational scanning. *The Journal of Immunology*, 200(11):3825–3839, 2018.

[328] Won-Tak Choi, Shaomin Tian, Chang-Zhi Dong, Santosh Kumar, Dongxiang Liu, Navid Madani, Jing An, Joseph G Sodroski, and Ziwei Huang. Unique ligand binding sites on cxcr4 probed by a chemical biology approach: implications for the design of selective human immunodeficiency virus type 1 inhibitors. *Journal of virology*, 79(24):15398–15404, 2005.

[329] Ling Qin, Irina Kufareva, Lauren G Holden, Chong Wang, Yi Zheng, Chunxia Zhao, Gustavo Fenalti, Huixian Wu, Gye Won Han, Vadim Cherezov, et al. Crystal structure of the chemokine receptor cxcr4 in complex with a viral chemokine. *Science*, 347(6226):1117–1122, 2015.

[330] Otto O Yang, Stephen L Swanberg, Zhijian Lu, Michelle Dziejman, John McCoy, Andrew D Luster, Bruce D Walker, and Steven H Herrmann. Enhanced inhibition of human immunodeficiency virus type 1 by met-stromal-derived factor 1β correlates with down-modulation of cxcr4. *Journal of virology*, 73(6):4582–4589, 1999.

[331] Tomer Tsaban, Julia K Varga, Orly Avraham, Ziv Ben-Aharon, Alisa Khramushin, and Ora Schueler-Furman. Harnessing protein folding neural networks for peptide–protein docking. *Nature communications*, 13(1):176, 2022.

[332] Debora Vignali and Marinos Kallikourdis. Improving homing in t cell therapy. *Cytokine & Growth Factor Reviews*, 36:107–116, 2017.

[333] Stefano Garetto, Claudia Sardi, Diego Morone, and Marinos Kallikourdis. Chemokines and t cell trafficking into tumors: strategies to enhance recruitment of t cells into tumors. *Defects in T Cell Trafficking and Resistance to Cancer Immunotherapy*, pages 163–177, 2016.

[334] Robert Sackstein, Tobias Schatton, and Steven R Barthel. T-lymphocyte homing: an underappreciated yet critical hurdle for successful cancer immunotherapy. *Laboratory Investigation*, 97(6):669–697, 2017.

[335] Clare Y Slaney, Michael H Kershaw, and Phillip K Darcy. Trafficking of t cells into tumors. *Cancer research*, 74(24):7168–7174, 2014.

[336] Ignacio Melero, Ana Rouzaut, Greg T Motz, and George Coukos. T-cell and nk-cell infiltration into solid tumors: a key limiting factor for efficacious cancer immunotherapy. *Cancer discovery*, 4(5):522–526, 2014.

[337] Jason S Park, Benjamin Rhau, Aynur Hermann, Krista A McNally, Carmen Zhou, Delquin Gong, Orion D Weiner, Bruce R Conklin, James Onuffer, and Wendell A Lim. Synthetic control of mammalian-cell motility by engineering chemotaxis to an orthogonal bioinert chemical signal. *Proceedings of the National Academy of Sciences*, 111(16):5896–5901, 2014.

[338] M Young, T Dahoun, B Sokrat, C Arber, KM Chen, M Bouvier, and P Barth. Computational design of orthogonal membrane receptor-effector switches for rewiring signaling pathways. *Proceedings of the National Academy of Sciences*, 115(27):7051–7056, 2018.

[339] Justine S Paradis, Xiang Feng, Brigitte Murat, Robert E Jefferson, Badr Sokrat, Martyna Szpakowska, Mireille Hogue, Nick D Bergkamp, Franziska M Heydenreich, Martine J Smit, et al. Computationally designed gpcr quaternary structures bias signaling pathway activation. *Nature Communications*, 13(1):6826, 2022.

[340] Tommaso Patriarchi, Jounhong Ryan Cho, Katharina Merten, Mark W Howe, Aaron Marley, Wei-Hong Xiong, Robert W Folk, Gerard Joey Broussard, Ruqiang Liang, Min Jee Jang, et al. Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science*, 360(6396):eaat4422, 2018.

[341] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, page eadg7492, 2023.

[342] Elodie Laine, Yasaman Karami, and Alessandra Carbone. Gemme: a simple and fast global epistatic model predicting mutational effects. *Molecular biology and evolution*, 36(11):2604–2619, 2019.

[343] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[344] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, et al. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1):D1062–D1067, 2018.

[345] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135, 2017.

[346] Jeanne Trinquier, Guido Uguzzoni, Andrea Pagnani, Francesco Zamponi, and Martin Weigt. Efficient generative modeling of protein sequences using simple autoregressive models. *Nature communications*, 12(1):5800, 2021.

[347] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

[348] Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nature communications*, 12(1):2403, 2021.

[349] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173–175, 2012.

[350] Milot Mirdita, Lars Von Den Driesch, Clovis Galiez, Maria J Martin, Johannes Söding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research*, 45(D1):D170–D176, 2017.

[351] Vadim Cherezov, Daniel M Rosenbaum, Michael A Hanson, Søren GF Rasmussen, Foon Sun Thian, Tong Sun Kobilka, Hee-Jung Choi, Peter Kuhn, William I Weis, Brian K Kobilka, et al. High-resolution crystal structure of an engineered human β2-adrenergic g protein–coupled receptor. *science*, 318(5854):1258–1265, 2007.

[352] Nobuhiko Tokuriki and Dan S Tawfik. Protein dynamism and evolvability. *Science*, 324(5924):203–207, 2009.

[353] Antonio Iorio, Céline Brochier-Armanet, Caroline Mas, Fabio Sterpone, and Dominique Madern. Protein conformational space at the edge of allostery: turning a nonallosteric malate dehydrogenase into an "allosterized" enzyme using evolution-guided punctual mutations. *Molecular Biology and Evolution*, 39(9):msac186, 2022.

[354] Ismael Rodríguez-Espigares, Mariona Torrens-Fontanals, Johanna KS Tiemann, David Aranda-García, Juan Manuel Ramírez-Anguita, Tomasz Maciej Stepniewski, Nathalie Worp, Alejandro Varela-Rial, Adrián Morales-Pastor, Brian Medel-Lacruz, et al. Gpcrmd uncovers the dynamics of the 3d-gpcrome. *Nature Methods*, 17(8):777–787, 2020.

[355] Jacqueline C Calderón, Passainte Ibrahim, Dorothea Gobbo, Francesco Luigi Gervasio, and Timothy Clark. General metadynamics protocol to simulate activation/deactivation of class a gpcrs: Proof of principle for the serotonin receptor. *Journal of Chemical Information and Modeling*, 2023.

Mahdi Hijazi

Curriculum Vitae

Lausanne, Switzerland
(+41) 78 638 53 58
mahdi.hijazi.mh@gmail.com
github.com/mahdiofhijaz
19.08.1993 - B permit

Profile

- Computational scientist with diverse expertise spanning different disciplines: from algorithm development to protein design
- Developer of in-house code packages that provide solutions for the project and the team
- Certified effective communicator bridging computational and experimental domains
- Enthusiastic teacher and an eager learner with experience in mentoring and project management

Education

2019–2023 **EPFL**, *PhD in Materials Science and Engineering*, Laboratory of Protein and Cell Engineering.

2016–2018 **EPFL**, *MSc in Materials Science and Engineering*, Cumulative GPA of 5.42/6.

2011–2015 **American University of Beirut (AUB)**, *BEn in Mechanical Engineering*, Cumulative GPA of 88.83/100, equivalent to 3.91/4.0.

Research Experience

Jan 2019–Present **EPFL, Laboratory of Protein and Cell Engineering**, *Advisor: Prof. Patrick Barth*.

Computational protein design

- Computationally designed allosteric signaling in G-protein coupled receptors (GPCRs) using molecular dynamics simulations, Rosetta software suite, and graph theoretic metrics
- Collaborated with members of the wet lab to validate designed proteins using cell-based assays

Code development

- Developed a comprehensive computational package for analysis of biological systems with focus on ease of use, good coding practices, and experimental validation

March–August 2018 **IBM Research Zurich**, *Academic advisor: Prof. Nicola Marzari, IBM Supervisor: Dr. Teodoro Laino*.

- Simulated quantum biological systems using a sparse matrix-matrix multiplication algorithm with unprecedented massive scalability
- Re-parameterized the in-house developed semi-empirical molecular dynamics (SEMD) for biological applications

Sept 2016–March 2018 **EPFL, Laboratory of Computational Science and Modeling**, *Advisor: Prof. Michele Ceriotti, supervisor: Dr. David M. Wilkins*.

- Modified the Langevin thermostat in constant temperature molecular dynamics simulations so that it maintains optimal efficiency in the difficult high friction limit

Academic Positions of Responsibility

Teaching

Spring 2020 2021 Synthetic biology to senior undergraduate students at EPFL, tasks included: lecturing, exam writing and marking, and project design

Autumn 2020 Scientific literature analysis to masters students in computational biology at EPFL, tasks included: report grading, presentation evaluation, and mentoring on reading scientific literature

Management

Feb–July 2022 Simon Lietar (Masters project at EPFL), *Quantification of mutual information significance in allosteric pathway calculations in proteins*

Selected Publications

Robert E Jefferson, Aurélien Oggier, Andreas Füglistaler, Nicolas Camviel, **Mahdi Hijazi**, Ana Rico Villarreal, Caroline Arber, and Patrick Barth, "Computational design of dynamic receptor—peptide signaling complexes applied to chemotaxis," *Nature Communications* **14**(1), 2875 (2023)

Daniel Keri*, **Mahdi Hijazi***, Aurélien Oggier, and Patrick Barth, "Computational rewiring of allosteric pathways reprograms GPCR selective responses to ligands," *bioRxiv* (2022). *: co-first authorship

Jie Yin, Kuang-Yui M Chen, Mary J Clark, **Mahdi Hijazi**, Punita Kumari, Xiao-chen Bai, Roger K Sunahara, Patrick Barth, and Daniel M Rosenbaum, "Structure of a D2 dopamine receptor-G-protein complex in a lipid membrane," *Nature* **584**(7819), 125-129 (2020)

Mahdi Hijazi, David M. Wilkins, and Michele Ceriotti, "Fast-Forward Langevin Dynamics with Momentum Flips," *Journal of Chemical Physics* **148**(18), 184109 (2018). Awarded editor's pick

Awards

Sept 2022 Best poster award at the "Understanding function of G-Protein Coupled Receptors by atomistic and multiscale simulations" CECAM worksop, Lugano, Switzerland.

2016–2018 EPFL Excellence Fellowship for the academic years of 2016/2017 and 2017/2018.

June 2016 Advanced Communicator Bronze award, *Toastmasters International*, with specializations in "professionally speaking" and "storytelling".

Nov 2015 Advanced Leader Bronze award, *Toastmasters International*.

May 2015 AUB High Distinction honor awarded on graduation. (GPA above 90/100 for the last two years of study)

Skills

Bash, Matlab, Molecular dynamics simulations (GROMACS and NAMD), Protein design (Rosetta software suit), Molecular visualization (VMD and Pymol), High performance computing (Slurm and GPU servers), Git, C++, Python

Extracurricular Activities

Oct 2015–Feb 2016 **Trainer and Advisor at Olayan School of Business (AUB):** Trained a group of five Masters of Finance students for the Chartered Financial Analyst (CFA) Institute Research Challenge in order to improve their professional business communication and presentation skills.

July 2014–June 2015 **Trainer and Mentor at Toastmasters Speakers and Leaders (AUB):** Delivered 12 workshops on the topics of: body language, team building, and speech writing to a diverse audience of university students across Lebanon.

Languages

Arabic **Mother language**

English **Bilingual Fluency**

French **Level B1/B2**