DISCOVERY AND CHARACTERIZATION OF Y4R ALLOSTERIC MODULATORS USING COMPUTER-AIDED DRUG DESIGN

By

Oanh Vu Thi Ngoc

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

January 31, 2021

Nashville, Tennessee

Approved:

Jens Meiler, Ph.D.

Lauren Elizabeth Buchanan, Ph.D.

David Weaver, Ph.D.

Zhongyue (John) Yang, Ph.D.

This dissertation is dedicated to my family and my friends

ACKNOWLEDGMENTS

The work presented here was supported by the R01 GM080403, R01 DA046138, and R01 HL122010 grants from the National Institute of Health. My research on GPCRs was also made possible by a Deutsche Forschungsgemeinschaft grant through CRC1423, project number 421152132. I would like to thank the The Molecular Sciences Software Institute (MOLssI) for awarding me a fellowship for my method development research in cheminformatics. The computational resources were provided by the Advanced Computing Center for Research and Education and the Center for Structural Biology at Vanderbilt University.

I would like to thank my advisor, Dr. Jens Meiler, who inspired me to embark on research in computational drug design as a summer undergraduate intern, and continued to support and mentor me through my graduate studies. His dedication to science has inspired my own determination in continuing my career in scientific research.

I would like to thank other faculty members who fostered in my scientific growth while at Vanderbilt University, particularly my current and former committee members: Dr. Lauren Elizabeth Buchanan, Dr. David Weaver, Dr. Zhongyue (John) Yang, and Dr. Terry Lybrand. They have continually provided guidance and encouragement for my research progress.

The Vanderbilt Chemistry Department has been a steady and consistent source of support and the strongest advocates for my success. I'm thankful to Dr. David Cliffel for initially recruiting me into the program and for continual inspiration.

Current and former members of the Meiler lab have been invaluable as friends and colleagues. Jeff Mendenhall and Ben Brown have been wonderful friends and mentors. My thanks also go out Dr. Rocco Moretti for patiently helping me with Rosetta. I am grateful for the opportunities to mentor wonderful students, from whom I also learned a lot. I am incredibly thankful for everyone else in the lab with whom I have had fruitful scientific discussions, and I've been to the mountains and beach together.

I want to say a special thank you to Dr. Doaa Altarawy for her excellent mentorship. She taught me about machine learning and was instrumental in helping me on my work in cheminformatics. I am forever grateful for her guidance and all I learned from her about being a great scientist.

My graduate research works could not be possible without the contribution from my collaborators. People in the Beck-Sickinger lab, especially Dr. Annette Beck-Sickinger, Corinna Schüß, and Dr. Mario Schubert, have taught me everything I know about mutagenesis experiments, and validated my computational predictions. I would like to also thank the Weaver lab (Dr. David Weaver and Dr. Yu Du), the Emmittee lab (Dr. Kyle A. Emmitte and Dr. Nigam M. Mishra), and the Cox lab (Dr. Helen M. Cox and Dr. Iain R. Tough) for synthesizing and testing Y4 allosteric modulators.

I am certain I have the best Chemistry class, whom I consider my family here in Nashville, particularly Ellie, Jessica, and Souhrid. Additionally, my friends from around the world have been there for me in the best and worst of times and I am eternally grateful for their support: Amrita, Queenster, Linh Nguyen, Linh Trinh, Trang, Thu, Huong, Linh Phan, and many more.

I can not imagine how I could become who I am now without my family. My mom (Huong), dad (Lam), sister (Tram), my brother-in-law (Thanh), and my nephews (Nguyen and Minh) have always been my strongest supporters. I am blessed with a wonderful host family, Gary and Gretchen. They have always been there for me whenever I need them. Last but not least, my boyfriend, Fraidun, has been by my side as a source of love and support on this journey. I could not have accomplished any of this without him.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Summary

Computer-aided drug discovery/design (CADD) has been universally used to facilitate and expedite the development of small molecule-based therapeutics. Ligand-based virtual screening can save time and resources by quickly narrowing libraries of millions drug-like molecules to hundreds of compounds candidates for experi-mental tests and validation. On the other hand, structure-based methods such as ligand docking help characterize the interaction be-tween ligands and target proteins when the crystal structures are not available. The overall focus of this dissertation was to design and validate methods for the computational modeling of ligands in complex with G protein-coupled receptors (GPCRs) and apply the methods to the modeling of allosteric modulators for the neuropeptide Y 4 receptor (Y4R).

Chapter 2 introduces the importance of Peptide-binding class A GPCRs as a target for drug development, with Y4R being an especially important target for obesity treatment. While many of these peptides have been structurally characterized in their solution state, the few studies of peptides in their receptor-bound state suggest these peptides interact with a shared set of residues and undergo significant conformational changes. This review compares different peptide binding modes, lists important examples of peptide structural dynamics through focusing on the conformational changes observed during the binding event, and describes the implications for future studies based on the current studies. Our quantitative analysis discovered a common set of 14 residues that were shown to interact with peptide ligands among all available co-crystal structures. This shared binding site suggests a potential general pattern in peptide engagement among class A GPCRs. Portions of Chapter 2 came from a review entitled "The Structural Basis of Peptide Binding at Class A G Protein-Coupled Receptors" written by Oanh Vu*, Brian Joseph Bender*, Lisa Pankewitz, Daniel Huster, Annette G. Beck-Sickinger, and Jens Meiler.

The human neuropeptide Y (NPY) receptor family (Y1R, Y2R, Y4R, and Y5R) are comprised of G-protein coupled receptors (GPCR) that bind the ligands neuropeptide Y (NPY), polypeptide YY (PYY), and pancreatic polypeptide (PP). Those have been shown to play critical roles in regulating satisfaction after food in-take and energy homeostasis. Since Y4 is the only NPY receptor with much higher affinity towards PP than to other two peptide ligands, selective agonists of Y4 could be potentially used in obesity treatment. Due to various advantages of allosteric modulators (AMs). Since allosteric sites tend to be less preserved than the orthosteric counterparts, AMs are more likely to be selec-tive toward target proteins. Moreover, side effects of AMs might be less severe as (1) they only act when the native ligands are present, and (2) they tune the receptors' response instead of turning it on or off. structure-based methods, one of which is ligand

docking, characterize the interaction between AM hits and Y4 homology models. Chapter 3 presents the application of incorporating mutagenesis data into the comparative modeling and docking protocols of PP and VU0637120, a selective NAM, to Y4R. Parts of this chapter was taken from the paper: Corinna Schüß*, Oanh Vu*, Mario Schubert*, Yu Du, Nigam M. Mishra, Iain R. Tough, Jan Stichel, C. David Weaver, Kyle A. Emmitte, Helen M. Cox, Jens Meiler, and Annette G. Beck-Sickinger. Journal of Medicinal Chemistry 2021 64 (5), 2801-2814.

Ligand-based virtual screening such as quantitative structure-activity relationship (QSAR) can quickly narrow libraries of millions of drug-like molecules to hundreds of compounds candidates for experimental validation. 2D fragment-based similarity searching is one of the most robust CADD techniques for database searching and quantitative structure–activity relationship (QSAR) analysis. A large-scale benchmark study of eight different 2D fingerprint methods has shown that Molprint2D yields the best enrichments of active compounds on a diverse set of targets. BCL::Mol2D, like the MolPrint2D fingerprints, are developed based on atom environment, which encodes 2D information of surrounding atoms. However, BCL::Mol2D are reversible from the object to the string hashed representation, allowing direct extraction of atom environments that are crucial for de-scribing the pharmacophores of active molecules. Chapter 3 shows that BCL::Mol2D descriptors outperformed Molprint2D in both predictive performance and interpretability on QSAR tasks across 9 high throughput screening (HTS) PubChem datasets. Additionally, out of 600 compounds identified by our QSAR models, 20 show significant PAM activity. Parts of this chapter was taken from the paper: Vu O, Mendenhall J, Altarawy D, Meiler J. BCL::Mol2D-a robust atom environment descriptor for QSAR modeling and lead optimization. J Comput Aided Mol Des. 2019 May;33(5):477-486.

Chapter 5 includes an application of the GPCR comparative modeling, docking, and molecular similation refinement protocol to the Class A GPCR, P2Y6. This section is based on a manuscript entitled “Mapping the binding sites of UDP and prostaglandin E2 glyceryl ester in the nucleotide receptor P2Y6” by Zimmermann*, Antje Brüser*, Oanh Vu*, Gregory Sliwoski, Lawrence J. Marnett, Jens Meiler, and Torsten Schöneberg. The author of this dissertation provided the computational modeling presented in light of experimental studies performed by Dr. Anne Zimmermann and Dr. Antje Brüser.

CHAPTER 2

The Structural Basis of Peptide Binding at Class A G Protein-Coupled Receptors

This chapter is a collaborative work of Oanh Vu*, Brian Joseph Bender*, Lisa Pankewitz, Daniel Huster, Annette G. Beck-Sickinger, and Jens Meiler (*these authors contributed equally).

2.1 Introduction

G protein-coupled receptors (GPCRs) represent the largest membrane protein family and a significant target class for therapeutics. Receptors from GPCRs' largest class, class A, influence virtually every aspect of human physiology. About 45% of the members of this family endogenously bind flexible peptides or peptides segments within larger protein ligands. While many of these peptides have been structurally characterized in their solution state, the few studies of peptides in their receptor-bound state suggest these peptides interact with a shared set of residues and undergo significant conformational changes. For the purpose of understanding binding dynamics and the development of peptidomimetic drug compounds, further studies should investigate peptide ligands complexed to their cognate receptor.

2.2 Overview of Peptide Binding at Class A G Protein-Coupled Receptors

2.2.1 Peptide-Activated Receptors Are a Large Percentage of the GPCR Class A

Out of four classes of GPCR—A, B, C, or F—Class A is the largest and most diverse group in humans. This subfamily has been investigated most extensively in drug discovery due to their available structural and experimental data. They conform with the common GPCR structural fold, such as a seven-transmembrane (7TM) helices domain, three extracellular loops, and three intracellular loops with ligand-binding pockets and a G-protein-binding region located in the extracellular and intracellular ends of the helix bundle, respectively (Isberg et al., 2017). The variety of drugs targeting GPCRs reflects the diversity of chemical signals that can be transduced by GPCRs, including small molecules, lipids, ions, and proteins (Bockaert and Pin, 1999; Wacker et al., 2017). In particular, according to the data from the GPCRdb server (Isberg et al., 2017), the peptide- and protein-activated receptors are found to account for about 46% of all class A GPCRs in human. For this review, we consider GPCRs that recognize classical peptides and peptide-like segments within larger protein domains to belong to the same category of receptors. Peptide-activated receptors are found across all rhodopsin-like subfamilies (α, β, γ, and δ) and the entire secretin family (Fredriksson et al., 2003). Given this coverage, it is unsurprising that many of the blockbuster drugs mentioned above (e.g., olmesartan, buserelin, and valsartan) target members of this receptor group. While Olmesartan and Valsartan serve as

angiotensin II receptor blocker (ARB) in treating hypertension (Markham and Goa, 1997; Scott and McCormack, 2008), Buserelin, a luteinizing hormone— releasing hormone (LHRH) agonist, can be used to treat hormone responsive cancers such as prostate and breast cancer (Brogden et al., 1990). With such importance for therapeutic development, a full understanding of the structural and dynamical determinants of signaling for these molecules is necessary. This review covers what is known about these receptors structurally using various biophysical techniques and provides suggestions for future discovery routes.

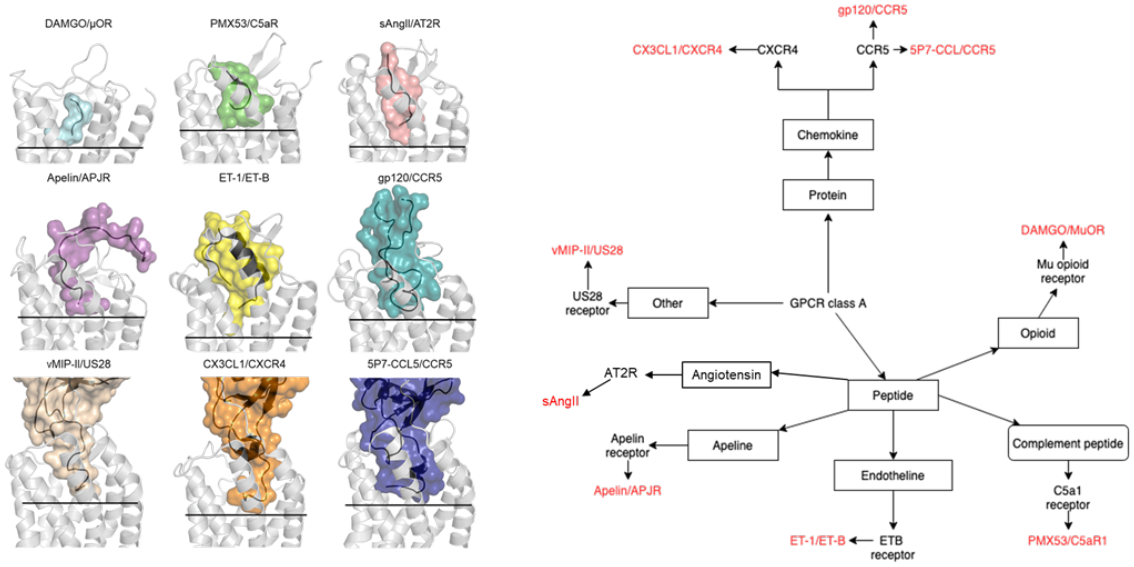


Figure 2.1: Overview of nine co-crystal structures of class A peptide-GPCR. (Left) Comparison of Peptide Binding Modes and Crystallized peptides DAMGO (cyan), PMX53 (green), sAngII (pink), apelin derivative (magenta), ET-1 (yellow), gp120 (teal), vMIP-II (wheat), CX3CL1 (orange), and 5P7-CCL5 (blue) at the recptors μ opioid receptor (μOR) (PDB ID: 6DDE), complement component 5a receptor (C5aR) (PDB ID: 6C1R), Angiotensin II type 2 receptor (AT2R) (PDB ID: 5XJM), Apelin receptor (APJR) (PDB ID: 5VBL), Endothelin B receptor (ET-B) (PDB ID: 5GLH), C-C chemokine receptor type 5 (CCR5) (PDB ID: 6MEO), US28 (PDB ID: 4XT1), CXC-chemokine receptor 4 (CXCR4) (PDB ID: 4RWS), and CCR5 (PDB ID: 5UIW), respectively. All receptors were aligned in the transmembrane region. Black bars illustrate the depth of penetration for each peptide ligand. (Right) Classification tree of eight class A GPCRs with their nine peptide ligands in those nine listed structures.

2.2.2 Diversity of Peptide Ligands

Peptide ligands come in a variety of lengths and structures though they share the common theme that they are ribosomally translated. Often, these peptide ligands are produced as pre-hormones that are subsequently processed to their active form. As a result, peptide ligands range in size from three amino acids (e.g., Thyrotropin-releasing hormone (TRH)) up to 100 amino acids (e.g., Chemokine ligand 23 (CCL23)). In addition to size differences, many peptide hormones undergo post-translational modifications. Some of these modifications are necessary to increase peptide half-life by inhibiting exopeptidases such as N-terminal pyroglutamation (e.g., TRH and Luteinizing hormone (LH) (Beck et al., 2001)) and C-terminal amidation (e.g.,

neuropeptide Y (NPY), pancreatic polypeptide (PP), and peptide YY (PYY) (Chance, 2003)). However, in some cases, these modifications serve dual purposes by acting as molecular recognition sites in their cognate receptors (Kaiser et al., 2015). Other types of post-translational modifications include lipidation, bromination, and disulfide bridge formation. A summary of modifications is found in Table 1. These modifications further increase the diversity of chemical space available to peptide hormones beyond the canonical 20 amino acids. The size, sequence, shape, charge, structural dynamics and chemical diversity allows for a vast degree of specificity between peptide hormones and their receptors. Further, it is common for a given peptide hormone to exist in multiple isoforms, such as the neuropeptide Y (NPY) family, which consists of NPY, peptide YY (PYY), and pancreatic polypeptide (PP) and the endothelin peptides ET-1, ET-2, and ET-3.

2.2.3 Reducing the Flexibility of Peptide Ligands is Crucial for Success in co-Crystallization

A significant challenge for interpretation of structures determined via crystallization of peptide-activated receptors in complex with their cognate peptide ligand is these peptides' inherent flexibility. Typically, small molecule antagonists and agonists will adopt a single conformation when interacting with a receptor and are fully encased in the receptor-binding pocket. Peptide ligands may adopt a single conformation in the binding pocket but, due to their length, the remainder of the ligand can remain outside the binding pocket and be flexible. This conformation change is likely why neurotensin 1 receptor (NTS1R) was crystallized with only residues 8-13 of the peptide as residues 1-7 are expected to extend above the receptor pocket and remain unconstrained (Da Costa et al., 2013; White et al., 2012). The peptide ligand of the apelin receptor, while full-length, was modified to incorporate a lactam ring, which significantly constrained the peptide's flexibility (Ma et al., 2017). Full-length chemokine crystallization is possible, as the portion of the chemokine that extends out of the binding pocket folds into a well-defined structural domain. However, the N-terminus of the receptor, known to recruit and bind the chemokines, has yet to be determined experimentally in its entirety (Burg et al., 2015; Qin et al., 2015; Zheng et al., 2017).

2.2.4 Complexity of Peptide Ligand and Receptor Interactions

In addition, as was recently classified, many peptide ligands target multiple receptors adding to their signaling complexity (Hauser et al., 2017). This complex selectivity of peptide ligand/receptor interactions results in peptide ligand biology's common theme: multi-ligand/multi-receptor systems. Evidence now shows that related ligands binding to the same receptor or the same ligand binding two different receptors can adopt different bound state conformations and sustain deviating interaction networks (Joedicke et al., 2018; Pedragosa-Badia et al., 2013a). This theme of multi-ligand/multi-receptor systems complicates the formulation of overarching binding and activation mechanisms that holistically explain this category of receptors,

unlike what is known about receptors activated by bioamines (Kooistra et al., 2013; Michino et al., 2015; Ngo et al., 2017). It also complicates the development of selective probes and therapeutic agents. As such, it is critical for a full understanding of receptor/hormone biology to study each peptide ligand/receptor combination in detail before attempting to formulate generalizations that can be used for future drug development. This task is monumental through many efforts that are ongoing to attempt this feat.

2.3 Comparison of Peptide Binding Modes across Class A GPCRs

2.3.1 Diversity in the Binding Modes of the Peptide Ligands to Class A GPCRs

The first crystal structure of a peptide-activated receptor was the CXCR4 receptor in 2010 (Wu et al., 2010). The receptor structure was determined in the inactive state bound to both a small molecule antagonist and a peptidomimetic. This receptor structure was similar to what had previously been seen for aminergic (Cherezov et al., 2007; Chien et al., 2010) and nucleotide (Jaakola et al., 2008) receptors. However, an interesting difference was the presence of a β-hairpin in extracellular loop 2 (ECL2), a motif that has been present in all peptide-activated receptor structures reported since that time (Wu et al., 2017). Two more years passed before another peptide-activated receptor structure was determined. The year 2012 was a watershed year for this family with the structure determination of all four opioid receptor (OR) members (δOR (Granier et al., 2012), κOR (Wu et al., 2012), μOR (Manglik et al., 2012), and NOP (Thompson et al., 2012)), the protease-activated receptor type 1 (PAR1) (Zhang et al., 2012), and the neurotensin type 1 receptor (NTS1R) (White et al., 2012). Notably, the NTS1R structure was the first structure determined of a peptide-activated receptor in complex with its endogenous peptide ligand. Interestingly, NT's binding depth (8-13) was not as pronounced as seen for the aminergic and nucleotide ligands, suggesting that peptide ligands bind more superficially and predominantly interact with the extracellular loops. As the extracellular loops are the most divergent region of GPCRs, this prevented the extrapolation of this binding mode to other peptide ligands.

Since 2012, additional peptide-activated receptor structures were determined. These included further chemokine receptors (CCR2 (Zheng et al., 2016), CCR5 (Tan et al., 2013), CCR9 (Oswald et al., 2016), and the viral US28 chemokine receptor (Burg et al., 2015)), both subtypes of the orexin (Yin et al., 2016, 2015) and angiotensin (Asada et al., 2018; Zhang et al., 2015b) receptors, the PAR2 receptor (Cheng et al., 2017), the endothelin-B receptor (Shihoya et al., 2016), the neuropeptide Y type 1 receptor (Yang et al., 2018), the neurokinin 1 receptor (Yin et al., 2018), and the C5a receptor (Robertson et al., 2018). The binding pockets of peptide-activated GPCRs are uniformly wide due to the structured ECL2 but display a variety of hydrophobic and electrostatic conditions (Wu et al., 2017). Of note, only a small subset of these structures has been determined with a peptide ligand bound. These include the chemokine receptors US28, CCR5, and CXCR4 (Burg et al., 2015; Qin et al., 2015; Shaik et al., 2019; Zheng et al., 2017), the endothelin-B

receptor (Shihoya et al., 2016), the apelin receptor (Ma et al., 2017), the μ opioid receptor (Koehl et al., 2018), the angiotensin type II receptor (Asada et al., 2018), and the C5a receptor (Liu et al., 2018). In contrast to the observed orientation of NT(8-13), these ligands' binding modes are very diverse, as seen in Figure 2.1. Peptide ligands can unwind their helix and adopt unstructured conformations to penetrate deep in the helical bundle via either their N- or C-terminus, such as apelin. They can bind with both termini folded into the binding pocket like ET-1, or in a horseshoe manner, presenting a curved surface to the receptor such as gp120. The ligands can bind deeply (sAngII) or closer to the surface (PMX53). However, conservation in peptide engagement mechanism among class A GPCRs has been investigated by combining earlier SAR studies and the alignment of interacting residues from recent GPCR-peptide structures. The authors suggested that common patterns in peptide-GPCR interactions were divided into four groups, depending on whether the peptide is cyclic or not and whether the GPCR interacts with the N- or the C-terminus of the peptide (Tikhonova et al., 2019). Superimposing the structures of the complexes, a common observation between the binding modes of different peptide ligands is that they often bind over an extended surface of the receptor (Figure 2.2A). More interestingly, we notice the peptides align surprisingly well at the core of the binding pocket (Figure 2.2B). Together with the conserved β-hairpin in ECL2, these observations suggest potential general themes conserved within GPCRs binding peptide-ligands.

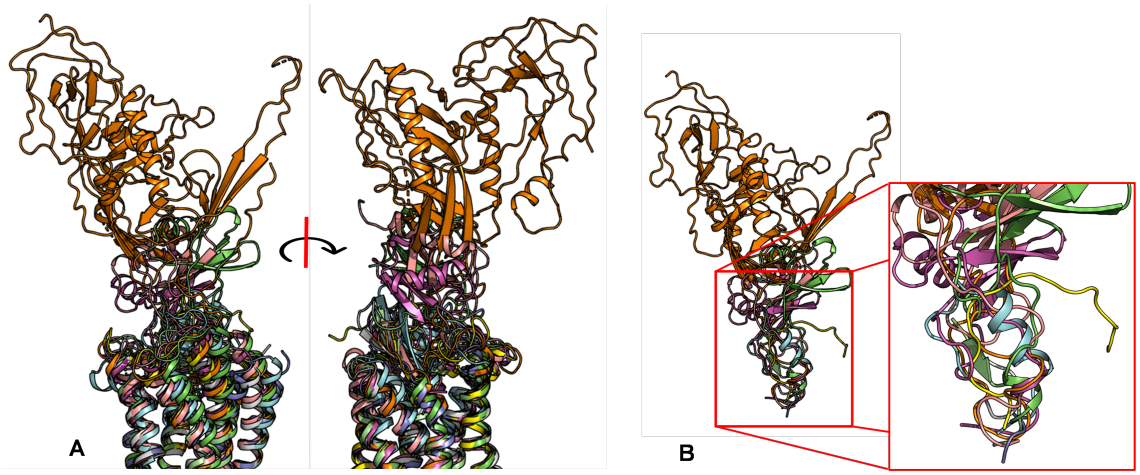


Figure 2.2: Despite the diversity in the peptide engagement, their overlapping region at the core of their binding pocket suggest common ligand-GPCR interactions. (A) Superimposition of the nine peptides/class A GPCR complexes. (B) Overlay of all peptide ligands and zoom-in of the peptide region at the cores of GPCRs.

2.3.2 Peptide Ligands Affect the Conformation of the Extracellular Surface

An essential consequence of the extended binding surface area of peptide ligands is that their presence affects not only the deep binding pocket but also the extracellular loops. This link between ligand engagement and

GPCR loop conformation was recently demonstrated by the endothelin receptor structures (Shihoya et al., 2016). This receptor was crystallized in the apo state and in complex with a peptide ligand. Interestingly, there was an extensive rearrangement of the extracellular domain in the peptide ligand presence (Figure 2.3A). This conformation rearrangement is expected to be the case for many peptide-activated receptor structures. In particular, the structural model of the Y1 receptor in complex with a small ligand found the N-terminus of the receptor lying over the binding pocket (Yang et al., 2018). Mutagenesis studies confirmed that this portion of the receptor did not affect the binding properties of either small molecule or endogenous peptide. It was implied that the N-terminus needed to be displaced from this crystallized orientation to allow binding of the much larger NPY ligand (Figure 2.3B). This implication was modeled and presented with the crystal structure with extensive use of orthogonal biophysical techniques, including NMR, cross-linking mass spectrometry, and mutagenesis. Additionally, the structure of the AT1R with a small molecule antagonist found the N-terminus lying over the ligand-binding pocket. In contrast, the AT2R structure, which was determined in the presence of a peptide analog sAngII, required the N-terminus to shift to allow access of the ligand to the orthosteric pocket (Figure 2.3C).

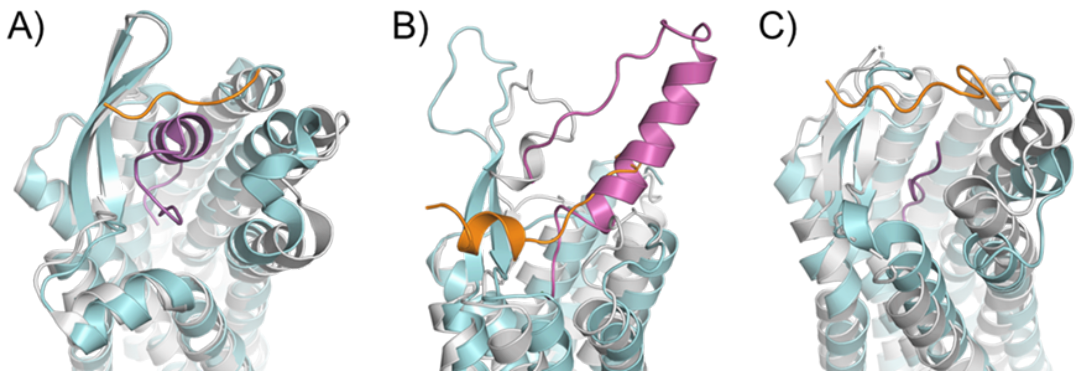


Figure 2.3: Rearrangements in the extracellular domain of peptide-activated GPCRs for peptide binding. (A) In the apo ET-B receptor (grey, PDB ID 5GLI) the N-terminus (orange) is lying over the ligand binding pocket. In the ET-1-bound state (cyan, PDB ID 5GLH), the bound ET-1 ligand (magenta) occupies the space of the N-terminus leading to its displacement (Shihoya et al., 2016). (B) The crystal structure of antagonist-bound Y1 receptor (grey, PDB ID 5ZBQ) also is found with the N-terminus (orange) lying over the ligand binding pocket. The modeled peptide-bound Y1R (cyan) places the NPY ligand (magenta) in this space displacing the N-terminus (Yang et al., 2018). (C) In the antagonist bound AT1 receptor (grey), the N-terminus (orange) extends over the pocket towards ECL2 (Zhang et al., 2015c). In the AT2 receptor (cyan) bound to sAngII (magenta), the peptide binds deep within the pocket and the N-terminus lays over ECL3 (Asada et al., 2018).

2.3.3 ECL1 and ECL2 Bound Conformation have Conversed across Class A Peptide-GPCRs

The superimposition of the three extracellular loops of peptide GPCR class A shows that the bound conformations of ECL1 and ECL2 are much more conserved than that of ECL3 (Figure 2.4). This observation

suggests that the first two extracellular loops could support a general interface for peptide binding. Together with the conserved β-hairpin in ECL2, details of the ECL mini-tertiary structure and orientation are critical for recognizing specific peptide ligands. Among nine class A GPCR structures that we investigated, four

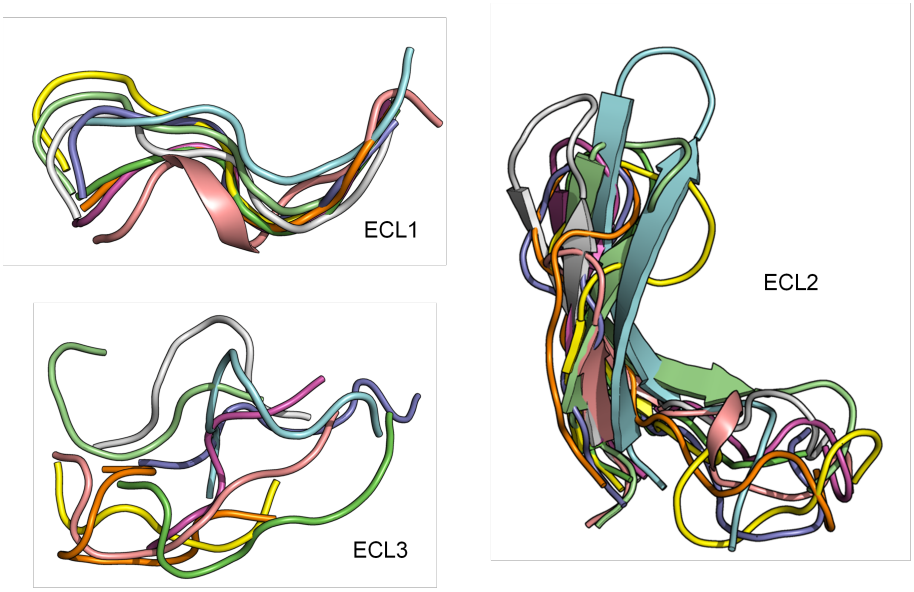
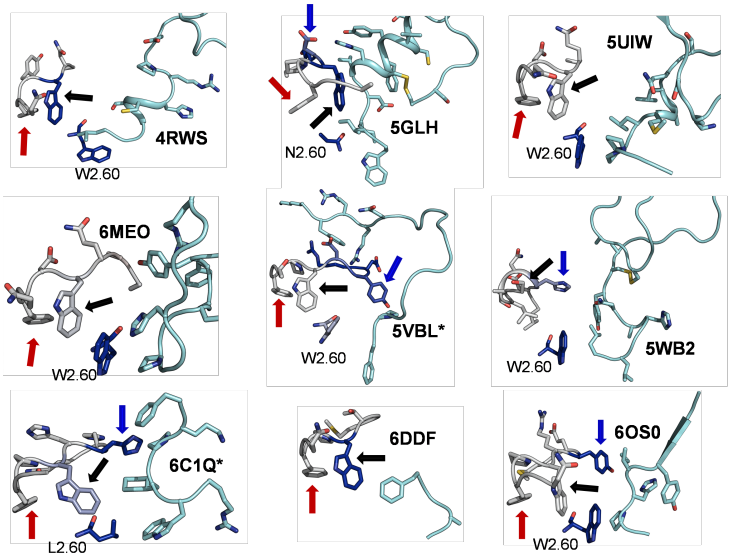


Figure 2.4: ECL1 and ECL2 have conserved bound conformation compared to ECL3. Overlay of three extracellular loops.

ECL1s have a common motif Y/HxWxF, and eight of them possess an xWxF motif. This motif, together with residue 2.60, interacts favorably with the bound conformation of the peptides. More specifically, the aromatic Y/H sidechain tends to form hydrogen bonds or hydrophobic interactions with the adjacent peptide sidechain or backbones. Moreover, the residue F23.52 forms a π-π to and stabilize W23.50's conformation, while W23.50 interacts with the peptide either directly through hydrophobic interactions or indirectly through π-π interaction with nearby W/L2.60 residue (Figure 2.5-Left). We also quantified the strength of the interactions we mentioned above by computing per residue $\Delta\Delta$G with Rosetta on the contacting GPCR residues. The $\Delta\Delta$G values of the three key residues (Y/H, W, and F) of the Y/HxWxF motif, together with the residue 2.60, are indicated by the colors on the images and reported in the table in figure 2.5. In general, the $\Delta\Delta$G values for those residues are negative, suggesting favorable interaction energy. This quantitative analysis further confirms our observations regarding common ECL1's mode of peptide engagement among the nine class A GPCR structures. Similarly, we conducted the $\Delta\Delta$G analysis on ECL2 residues and observed that all peptides interact favorably with the β-hairpin of this loop. Out of three extracellular loops, ECL2 tends to be the most structured with a distinctive secondary structure of a twisted beta-hairpin conformation. In all nine complex structures, ECL2 loops maintain the "open" conformation (Woolley and Conner, 2017), opening a "gate" and allowing the peptide ligand to enter the core of the TM bundle from the extracellular



Motif Y/HxWxF

PDB IDs	Receptor	23.47	23.48	23.49	23.50	23.51	23.52
4RWS	CXCR4	_	_	N	W	Y	F
5GLH	ETBR	_	E	D	W	P	F
5UIW, 6MEO	CCR5	_	_	Q	W	D	F
5VBL	apelinR	D	Y	D	W	P	F
5WB2	US28	D	H	N	S	L	A
6C1Q	C5a1R	H	H	H	W	P	F
6DDF	MuOR	G	_	T	W	P	F
6OS0	AT1R	R	Y	D	W	L	F

ECL1 sequence alignment

PDB IDs	Receptor	23.47	23.48	23.49	23.50	23.51	23.52
4RWS	CXCR4				-1.15		
5GLH	ETBR		-2.67		-1.33		
5UIW	CCR5						
6MEO	CCR5				-0.06		
5VBL*	apelinR		-0.82				
5WB2	US28		-0.34				
6C1Q*	C5a1R		-1.19		-0.41		
6DDF	MuOR				-1.24		
6OS0	AT1R		-2.05				

Rosetta ddG values of Y/H and W

Per-residue ΔΔG

-1 +1

Figure 2.5: ECL1 and the role of motif Y/HxWxF in peptide binding among class A GPCRs with peptide ligands. (Left) Interactions among ECL1, residue 2.60, and the peptide. The interacting peptide residues are colored in cyan. Residues on ECL1 and 2.60 are colored based on their computed per residue ΔΔG values (blue: negative ΔΔG, darkest blue: -1 or below; grey: ΔΔG value of 0, or no interactions; red: positive ΔΔG, darkest red: 1 and above). (Right) Tables show the sequence alignment of ECL1 and the three key residues in ECL1 motif Y/HxWxF are Y/H, W23.50, and F23.52, which are marked with blue, black, and red arrows, respectively.

region. Naturally, the peptides would interact with the β-hairpin of ECL2 β-hairpins at the "gate", which connects the extracellular space to the inside transmembrane domain. This observation is also reflected in the computed ΔΔG of interacting target residues. The ΔΔG analysis results suggest that the peptides generally engage with the β-hairpin of ECL2, especially at the tip where the three conserved residues (45.50, 45.51, and 45.52) are located (Figure 2.6 and Figure 2.7).

2.3.4 A List of 14 Common Interacting Residues Suggests a General Peptide Recognition and Binding Mechanism among 11 Class A GPCRs

Despite considerable diversity in size, sequence, secondary and tertiary structure of the nine peptide ligands, we observed significant overlap in the receptor region they bind to, particularly in a binding pocket between the outer leaflet portions of transmembrane helices (Figure 2.2). Using Rosetta (Conway et al., 2014; Smith and Kortemme, 2008), we calculated the per residue ΔΔG of the interacting residues on the transmembrane helices, two conserved ECL1 residues (23.49 and 23.50), and three conserved ECL2 residues (45.50, 45.51, and 45.52). The details of structure optimization and ΔΔG analysis protocols are listed in the supplementary material, and the ΔΔG values of all residues are listed in the supplement table S1. To make the optimization and ΔΔG analysis possible for the apelin/ApelinR (PDB ID: 5VBL) and the PMX53/C5aR (PDB ID: 6C1Q)

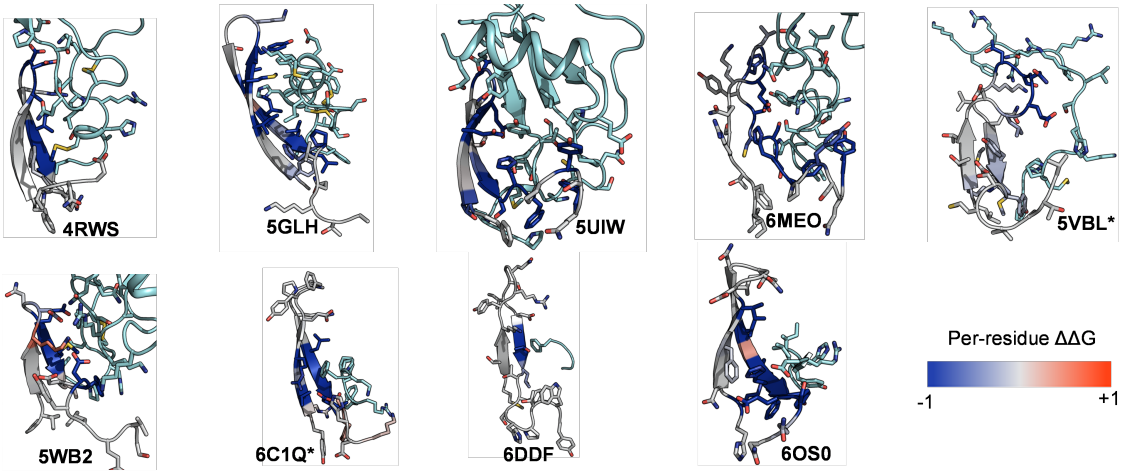


Figure 2.6: ECL2 β-hairpin and conserved residues interact with peptides of nine peptide/class A GPCR crystal structures.

structures, which contain non-nature peptide backbone, we generated 5VBL* and 6C1Q* as natural-backbone peptide analogs of those structures. More specifically, 5VBL* peptide ligand has the native apelin sequence, and the covalent bond between ornithine (ORN) at position 2 and the N-terminal acetyl group is omitted in the 6C1Q* peptide (Figure S3). The GPCR residues were then ranked based on their calculated ΔΔG. We selected 14 common residues with ΔΔG of less than -1 and contact peptide ligands in at least seven out of nine GPCR-peptide complexes. The details of the list and their locations are mapped in the structure of the ET-1/ETB receptor complex are shown in figure 2.6. This list of the top 14 residues implies a potential common peptide-binding mechanism among class A GPCRs. This common binding pocket encompasses two residues of TM2 (2.60 and 2.63), one from TM3 (3.32), three from TM6 (6.51, 6.55, and 6.58), five from TM7 (7.28, 7.32, 7.35, 7.36, and 7.39), and all three conserved ECL2 residues. More specifically, the common peptide engagement mechanism starts from the end of the β-hairpin of ECL2, extends to the tip of TM2, touches the extracellular half of TM7 and TM6, then ends at the core of TM3. A table that summarizes the non-Van Der Waal interactions are alsoAlthough additional structures of peptide-GPCR complexes are still needed to validate our hypothesis of common peptide binding pocket, this discovery could help guide future structural studies of this family of GPCRs.

We examine whether the common binding mechanism agrees with models of three class A GPCRs—Y1(Yang et al., 2018), Y2(Kaiser et al., 2015), and Ghrelin receptor(Bender et al., 2019)—and their endogenous peptide ligands—NPY and Ghrelin. In those studies, the peptide docking experiments were conducted using FlexPepDock (Raveh et al., 2011) with constraints from mutagenesis, cross-linking, and NMR data. For each complex, the ΔΔG analysis was performed on an ensemble of docking models. The per-residue ΔΔG values

Receptor	CXCR4	ETBR	CCR5	apelinR	US28	C5a1R	MuOR	CCR5	AT1R	Sum	Average
Residue #	4RWS	5GLH	5UIW	5VBL*	5WB2	6C1Q*	6DDF	6MEO	6OS0	ΔΔG	ΔΔG
7.39	-4.0	-3.7	-5.1	-1.3	-5.0	-3.3	-1.4	-0.6	-2.7	-27.1	-3.0
2.60	-5.4	-4.5	-3.8	-0.2	-2.4	-2.3	-1.7	-2.8	-3.1	-26.1	-2.9
45.51	-2.2	-4.9	-3.0	-0.2	-2.4	-3.0	-2.7	-1.9	-3.4	-23.8	-2.6
7.35	-2.2	0.0	-1.1	-5.6	-3.1	-5.4	-1.0	-1.9	-1.8	-22.2	-2.5
7.32	-2.8	-6.2	-3.3	-0.8	-2.5	0.1		-0.6	-6.0	-22.0	-2.4
45.52	0.0	-2.2	-4.1	-0.3	0.1	-5.4	-0.4	-4.1	-5.3	-21.6	-2.4
6.58	-4.7	-2.8	-0.9	-5.0	0.1	-1.6	-0.6	-1.7	-3.4	-20.6	-2.3
6.51	0.1	-2.8	0.0	-5.9	-3.8	-0.8	-3.1	-0.2	-0.1	-16.6	-1.8
2.63	-6.5	-3.6	-2.7	-0.2	-0.4	-0.2	0.1	-1.9	-0.6	-15.9	-1.8
3.32	0.0	-1.2	-2.0	-1.4	-1.2	-0.5	-5.7	-1.8	-1.9	-15.6	-1.7
7.36	-0.8	-4.8	-2.8	0.0	-3.4	-0.7	0.0	-2.6	-0.2	-15.3	-1.7
45.50	-3.5	-0.3	-0.6	-0.1	-1.2	-4.2	-1.6	-0.7	0.4	-11.8	-1.3
7.28	-1.7	-4.3	-0.6	-3.1	-0.6	-0.1			-0.6	-10.9	-1.2
6.55	0.0	-2.1	-0.4	-4.3	-1.2	0.1	-1.9	-0.9	0.0	-10.5	-1.2

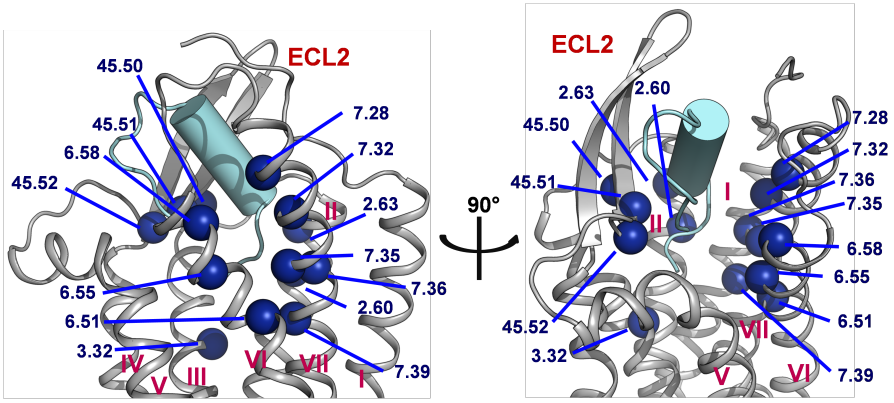


Figure 2.7: Residues with strongest interactions according to average computed ΔΔG suggest common binding pocket of peptide ligands. (Upper) a table shows a list of residues with top average computed ΔΔG values. The residues are numbered according on the Ballesteros-Weinsein numbering scheme (Isberg et al., 2015). For each residue position, the ΔΔG values is colored in the scale from -1 and less (blue) to 0 (white) to 1 and above (red). The absence of the ΔΔG values means the corresponding residues do not intereact with the peptide ligands. Two final columns of the table contain the sum and the average ΔΔG values across nine peptide-class A GPCR structures, respectively. The residue list is sorted in their ascending average ΔΔG order. (Lower) Front and side view of common peptide binding pocket toward the core of nine class A GPCR structures. The top residues in the upper table are mapped on the ET-1/ETB structure (PDB ID: 5GLH)(Shihoya et al., 2016). The important residues for peptide engagement across eight class A GPCRs are marked by blue spheres. The peptide ligand ET-1 are shown as a cyan cylinder with two unstructured extended regions.

were assigned to interacting residues of the GPCR targets. The peptides' binding pockets contain all the 14 common residues, except ghrelin does not contact the residue 7.36. Furthermore, most of the interactions between the common residues and NPY or ghrelin are favorable or at least neutral, except for the high ΔΔG value of residue 7.32 from Y2 (Figure 2.8). These results imply that the observation of the common peptide engagement pocket can also be applied to the docking study of peptide class A GPCRs, especially with limited experimental data.

A GPCR pharmacogenomics study has extracted polymorphism data for the coding-region of the 108 GPCR drug targets (Hauser et al., 2018). From the data provided by the authors, we found around 30 relevant GPCR mutants that were predicted to be deleterious by either SIFT (Sorting Intolerant From Tolerant) (Monahan et al., 1973) or Polyphen (Adzhubei et al., 2013). Those 30 genetic invariants have popu-

Residues	Y1	Y2	GhrelinR	Sum ΔΔG	Average ΔΔG
2.60	-0.8	-0.6	0.0	-1.4	-0.5
2.63	-1.6	-3.6	-0.5	-5.7	-1.9
3.32	-0.6	-7.0	-0.8	-8.4	-2.8
6.51	0.0	-2.6	-0.3	-2.9	-1.0
6.55	0.0	-3.8	-1.2	-5.0	-1.7
6.58	-0.4	-2.9	-3.1	-6.3	-2.1
7.28	-1.0	0.0	-0.9	-1.8	-0.6
7.32	0.2	1.4	-0.4	1.1	0.4
7.35	-0.5	-2.6	-0.2	-3.4	-1.1
7.36	0.0	0.0		0.0	0.0
7.39	0.0	-1.9	-0.4	-2.3	-0.8
45.50	0.2	-1.2	-0.1	-1.1	-0.4
45.51	-1.5	-4.1	-1.0	-6.6	-2.2
45.52	0.0	-1.4	-1.0	-2.4	-0.8

Figure 2.8: Models of peptide/class A GPCR complexes show the peptides interact with the top 14 common residues. (from left to right): A table lists ΔΔGs of the 14 common residues of Y1 (Yang et al., 2018), Y2 (Kaiser et al., 2015) and ghrelin receptors (Bender et al., 2019), as well as their sum and average values. The absence of the ΔΔG values means the corresponding residues do not interact with the peptide ligands. The residue ΔΔG cells as colored based on the ΔΔG values (negative: blue, neutral: white, and positive: red). Blank cells mean the residues do not interact with the peptide ligands. Models of NPY (cyan) binds with the Y1 receptor (Grey) [40] and the Y2 receptor (orange) [12], and Ghrelin (magenta) binds with ghrelin receptor (green).

lation allele frequencies of around 1 to 28 over 120,000 individuals and are related to shared peptide interacting residues or close to those residues. The table containing the information regarding the relevant mutants of peptide and protein binding class A GPCRs is summarized in the supplementary table Peptide_binding_pocket_genetic_variants.xlsx. The data suggests the great potential of the proposed common peptide-binding pocket as drug targets for class A GPCRs.

2.4 Structural Changes in Peptides Induced by Receptors are Critical for Binding

This theme of conformational change in peptides in their bound state is not unique to peptide-GPCR recognition. Studies of ubiquitin by X-ray crystallography bound to various substrates identified several unique conformations. However, NMR analysis revealed that all of these conformations existed simultaneously in solution, demonstrating that conformational selection drove the binding recognition event (Lange et al., 2008). Peptide binding sites have been characterized to require unique conformations of peptide ligands in GPCRs (Tyndall et al., 2005b), proteases (Tyndall et al., 2005a), and other systems, including antibodies and

major histocompatibility complex (Siligardi and Drake, 1995). To the best of our knowledge, there has not been a review of the conformational changes that peptide ligand must undergo from their unbound to bound states at GPCRs. These changes have relevance in the future determination of structure and dynamics and thus in peptidomimetic drug discovery. The following section will highlight examples of peptide structural dynamics focusing on the conformational changes observed during the binding event.

2.4.1 Neurotensin

Neurotensin (NT) is a tridecapeptide (Carraway and Leeman, 1975) with the C-terminal six residues known as NT(8-13) responsible for receptor activation (Henry et al., 1993). Original NMR studies of full-length NT in aqueous solution, methanol, and SDS (a membrane-mimic) found that under all conditions, the peptide was unstructured (Xu and Deber, 1991). In contrast, significant chemical shift perturbations were observed for the C-terminal NT (8-13) upon binding to the receptor, indicating a conformational change when bound (Williamson et al., 2002). This structural rearrangement was subsequently confirmed by determining the structures of free, membrane-bound, and receptor-bound NT(8-13) with solid-state NMR (Luca et al., 2003) and molecular dynamics (Heise et al., 2005). These studies found that both the solution and membrane-bound states contained no defined structure, while the receptor-bound peptide possessed an extended β-strand conformation. Knowledge of this extended binding pose allowed for the design of constrained peptides that reinforced the need for this conformation in the bound state. This further explained the reduced activity of end-to-end cyclization of NT(8-13) as it prevented the extended conformation (Van Kemmel et al., 1996).

2.4.2 Apelin

The apelin peptides are a family of peptides all formed from the same prohormone but with subsequent N-terminal proteolytic processing. Structure-activity relationships (SAR) analysis on the peptide identified a primary binding motif of the last five C-terminal residues with a secondary binding motif located four residues away (Fan et al., 2003; Medhurst et al., 2003; Murza et al., 2012). CD studies of the peptide revealed that in solution, the apelin peptides possessed no structured regions (Fan et al., 2003; Langelaan et al., 2009). Secondary structure could be induced either by lowering the temperature of the solution or the addition of membrane mimetics (Langelaan et al., 2009). The regions that became ordered under these conditions were the same regions that were previously identified in SAR studies as the binding motifs. When the structure of APJR bound to an apelin mimetic was determined, it was found that the apelin mimetic adopted a conformation that allowed for an ordered presentation of these two binding motifs at distinct regions of the receptor. Interestingly, mutagenesis and MD simulations of apelin-13 in the crystal structure revealed that native apelin peptide binds in a similar orientation as the crystalized ligand (Ma et al., 2017).

2.4.3 Endothelin

The endothelin peptides are a family of three 21-amino acid long peptides containing two internal disulfide bonds. Multiple NMR and X-ray studies have characterized the structure of these peptides to fold into a defined horseshoe orientation from residues 1 through 15 with residues 8-15 forming an αG-helix (Andersen et al., 1992; Janes et al., 1994; Aumelas et al., 1995; Atkins et al., 1995; Takashima et al., 2004; Bender et al., 2019). This horseshoe orientation is stabilized due to the disulfide bonds and is lost when the disulfides are interrupted (Hewage et al., 1999). The C-terminus beyond residue 15 is highly dynamic, adopting helical structures (Janes et al., 1994) or extended structures (Atkins et al., 1995; **?**) depending on the conditions of the experiment. In some cases, it is so poorly resolved that structure could not be assigned to these residues (Andersen et al., 1992; Aumelas et al., 1995). The C-terminus, however, is critical for activity at ETA/B and should bind in an ordered pocket within the receptor (Lattig et al., 2009). As seen in the ET-1/ETB co-crystal structure, the overall conformation of ET-1 remained mostly unchanged from that in solution because the two disulfide bonds reduce its structural flexibility (Shihoya et al., 2016). However, the C-terminus of the ligand unwinds to bind within the receptor core, while remaining close to the ligand N-terminus. This orientation of the C-terminus with respect to the N-terminus is found in two of the ten ensemble structures of a snake venom toxin with high sequence similarity and identical disulfide linkage as ET-1, suggesting the peptide can sample this conformation, albeit at a low population, in solution (Atkins et al., 1995). Interestingly, the receptor in the bound state folds its ECL2 and N-terminus over the ligand, explaining the extremely slow off-rates exhibited by these peptides in vivo (Hilal-Dandan et al., 1997; Takasuka et al., 1994). This structure clearly demonstrates that conformational changes in both binding partners are needed for full binding activity.

2.4.4 The Complement System Peptide Ligand C5a

The complement system is a peptide-receptor system comprising two ligands (C3a and C5a) and three receptors (C3aR, C5aR1, and C5aR2, previously known as GPR77). Both peptide ligands contain three conserved disulfide bonds that play a role in defining the overall helical bundle fold, which has been observed repeatedly by crystallography and NMR (Huber et al., 1980; Nettesheim et al., 1988; Zuiderweg et al., 1989; Williamson and Madison, 1990; Fredslund et al., 2008; Laursen et al., 2011; Bajic et al., 2013; Laursen et al., 2010; Cook et al., 2010; Schatz-Jakobsen et al., 2014; Zhang et al., 1997). While the full peptide is necessary for activation of the receptors, the C-terminal segment is the activation segment that binds at the receptor core (Siciliano et al., 1994). This C-terminal segment adopts a variety of conformations depending on the studied condition and lacks any secondary structure. One NMR study measured chemical shifts in of the C-terminal residues to find an αG-helix folding back onto the helix-bundle (Zhang et al., 1997), an unlikely conformation in the active state as this peptide must be ”presented” to the receptor for activation. Modeling

the C-terminus of C5a in a C5aR homology model also suggested that the endogenous peptide possessed a dramatically different conformation in the solution than in the bound state (Nikiforovich et al., 2008). In fact, this proposed binding mode was very similar to the bound conformation of the cyclic hexapeptide PMX53 (Liu et al., 2018). The ligand formed a beta-hairpin to interact directly with ECL2 via backbone hydrogen bonding. It is now understood that the cyclization enforces the conformation of the backbone orientation to predefine the backbone geometry needed for interaction with ECL2. Additional modeling studies have supported this extended conformation of C5a and derivative peptides (Rana and Sahoo, 2015; Sahoo et al., 2018).

2.4.5 Ghrelin

The ghrelin peptide is a 28 amino acid polypeptide with an octanoyl lipid modification at position Ser3(Kojima et al., 1999). This peptide is the only known lipid-modified peptide hormone in the human body, and it has been found that this lipid modification is critical for receptor activation (Bednarek et al., 2000; Kojima et al., 1999; Matsumoto et al., 2001a; **?**). Structure-function studies on ghrelin initially identified that the N-terminus of the peptide was critical for binding and activating the receptor via two main interactions: the positively charged amino head group and the hydrophobic octanoyl chain at Ser3 (Matsumoto et al., 2001a,b; Van Craenenbroeck et al., 2004). However, beyond these rules, little was known about the binding mode or conformation of ghrelin at its receptor. NMR and CD spectroscopy studies of the peptide in solution agreed that the peptide was highly disordered in the aqueous state (Silva Elipe et al., 2001). Increasing the hydrophobicity of the solution either with organic solvents or detergents seemed to increase the helicity of the central portion of the peptide while the termini remained highly flexible (De Ricco et al., 2013; Martin-Pastor et al., 2010; Staes et al., 2010; Vortmeier et al., 2015). However, recent NMR data of the peptide bound to its receptor revealed that a helix is found in the central peptide while the N-terminal binding portion converged to a well-defined extended structure (Bender et al., 2019; Ferré et al., 2019).

2.4.6 Gonadotropin-Releasing Hormone

Gonadotropin-releasing hormone (GnRH) is a decapeptide consisting of pyroGlu-His-Trp-Ser-Tyr-Gly-Leu-Arg-Pro-Gly-NH2. Evolutionary analysis reveals that the first four residues, the central Gly6 residue, and the last two residues are highly conserved (Millar, 2005). This pattern of conserved residues suggests a dual binding mode that requires both termini to come into close contact with the receptor. Extensive mutagenesis on both the peptide and receptor implies an inverted horseshoe binding motif for receptor activation (Sealfon et al., 1997). NMR studies of this peptide in solution failed to identify a single conformation (Chang et al., 1972; Chary et al., 1986; Grant and Vale, 1972; Monahan et al., 1973); however, peak sharpening increases in

the presence of membranes suggested a reduction in conformational dynamics (Chary et al., 1986). Computer simulations also revealed a broad population of conformations that could exist with many low energy states containing a β-turn conformation in residues 5-7 (Tanaka et al., 2003). The conformations of Gly6 adopt states that are inaccessible to any other L-amino acid but represent low energy conformations of D-amino acids (Hauser et al., 2018). Substitution of this residue with a D-amino acid enhances the likelihood of the β-turn, thereby prestabilizing the conformation for receptor binding. Interestingly, a Gly6 substitution with D-Trp can overcome the loss of binding in an Arg8 to Gln mutation (Millar et al., 1989). GnRH analogs, including goserelin, nafarelin, triptorelin, leuprorelin, buserelin, histrelin, and deslorelin, are used to treat hormone-sensitive diseases such as breast and prostate cancer (Kumar and Sharma, 2014) and often contain a D-amino acid substitution at position 6. NMR studies of nafarelin find that unlike GnRH, this peptide readily adopts a β-turn conformation in an aqueous solution (Andersen and Hammen, 1991). Similar results were obtained in the NMR analysis of leuprorelin (Laimou et al., 2010). All described findings imply that GnRH needs to select a particular conformation from the many accessible in the unbound state to bind at the GnRH receptor.

2.4.7 Neuropeptide Y

The neuropeptide Y (NPY) system consists of three 36 amino acid peptide amides (NPY, PYY, and PP) and four receptors (Y1, Y2, Y4, Y5) with differing affinities for the various peptide/receptor combinations (Pedragosa-Badia et al., 2013a). Initial studies to parse out the specific interactions of these peptides revealed that the C-terminal six residues were the primary binding and activation epitope within the NPY peptides (Beck-Sickinger and Jung, 1995). An X-ray crystal structure of avian PP revealed a disordered N-terminus with an αG-helix from residue 14-31 and a disordered C-terminus (Blundell et al., 1981). Solution NMR studies showed a different structure with the helix present through the C-terminal end of the peptide (Saudek and Pelton, 1990). Recent mutagenesis and docking study also suggested that the C terminus of PP needed to unwind to bind to Y4 receptor (Schüß et al., 2021a). Characterization of NPY in the membrane-bound state by NMR, CD, and EPR found the helix extending from residue 14 through the C-terminus (Bader et al., 2001; **?**). It was not until the peptide was structurally characterized in its Y2 receptor-bound state that it became clear that the C-terminus, though helical in its membrane-bound state, must unwind into an extended conformation for binding at the receptor (Kaiser et al., 2015). The conformational change of the C-terminus was also observed in a study of NPY binding at the Y1 receptor (Yang et al., 2018). However, in this study, photo-crosslinking revealed that the N-terminus of NPY was interacting with ECL2 instead of the central helix. This alteration of the second binding site interaction resulted in a distinct binding orientation of NPY at two if its four receptors (Figure 2.8). Further, studies will need to be pursued to contrast the binding mode

of NPY at the remaining receptors to understand the complete basis of subtype selectivity.

2.4.8 Opioid Peptides

The opioid receptor family, comprising of δOR, μOR, κOR, and NOP, responds to various endogenous peptides, including endorphins, dynorphins, and enkephalins. These peptides contain a common N-terminal motif of YGGF followed by diverging residues. It is suggested that the N-terminal motif is the activation sequence while the remaining residues confer receptor selectivity, the so-called "message-address" paradigm (Portoghese, 1989). Again, there is a conformational heterogeneity within the population of these peptides in both the aqueous- and membrane-bound states. A study of the peptide dynorphin B in the presence of the κOR found that the central portion of the peptide formed a well-defined αG-helical turn while the N- and C-terminal residues are structurally disordered (O'Connor et al., 2015). It was interesting that multiple conformations were found for the N-terminal motif in the bound state. This conformation diversity contrasts with molecular dynamics simulations run on the DAMGO peptide bound in the μOR-Gi cryo-EM structure with bound synthetic peptide (Koehl et al., 2018). Here, the researchers found that the peptide was relatively stable in its conformation over time within the binding pocket. At present, it is unclear if this conformational stability is due to the alterations of the peptide backbone in this synthetic peptide derivative, stabilization due to activation state, or a difference between the binding pockets of μOR and κOR.

2.5 Implications for Future Studies

The flexibility of the peptide ligands and the extracellular loops of the receptor mandate studying structure and dynamics of peptide-activated GPCRs in tandem. X-ray crystallography and Cryo-EM will provide critical snapshots that display key structural determinants of peptide/receptor interactions. However, these studies need to be complemented by spectroscopic investigations that study structure in the context of dynamics to gain a complete picture of the activation mechanism. While exciting progress in this area has been described over the past five years, we are only at the beginnings of these integrated approaches to study the structural dynamics of peptide-activated GPCRs. It is undeniable that interdisciplinary scientist teams are vital to the success of these studies, including experts in crystallography, spectroscopy, biochemistry, pharmacology, and modeling. Some of the computational technologies to integrate structural and dynamical data from various methods need to be optimized. However, since different methods introduce individual biases onto highly engineered systems, those systems need to be adequately considered when drawing conclusions for the wild-type ligand/receptor pair.

2.5.1 Peptides Need to be Characterized in their Bound State

Several peptide hormones have been examined to understand their structure via NMR or CD in solution. These include motilin (Andersson and Maler, 2002), prolactin-releasing peptide (Deluca et al., 2013; Rathmann et al., 2012), vasopressin (Lubecka et al., 2015), relaxin (Haugaard-Kedstrom et al., 2015; Rosengren et al., 2006), and somatostatin analogues (Grace et al., 2005, 2006, 2008, 2003). In contrast, relatively few examples exist of peptides studied in both their solution and bound states. These include the peptides neurotensin, NPY, ghrelin, and bradykinin (Bender et al., 2019; Denys et al., 1982; Joedicke et al., 2018; Kaiser et al., 2015; Lopez et al., 2008; Luca et al., 2003; Thomas et al., 2005; Vortmeier et al., 2015; Yang et al., 2018). A common theme in all these studies and the ones mentioned above is that the conformations of the peptides in their unbound states are distinct from their bound state (Figure 2.9). This conformation differentiation is perhaps unsurprising as the individual degrees of freedom in each amino acid are high in a peptide. In contrast, the receptor binding pocket imposes a stringent constraint on the conformation of these peptides. This theme of conformational sampling is analogous to the change in extracellular loop conformations in C5aR when bound to either a small molecule or peptide ligand (Liu et al., 2018; Robertson et al., 2018). Given these differences, it is necessary to study these peptides in the presence of their cognate receptors to develop a full understanding of the molecular basis of peptide recognition.

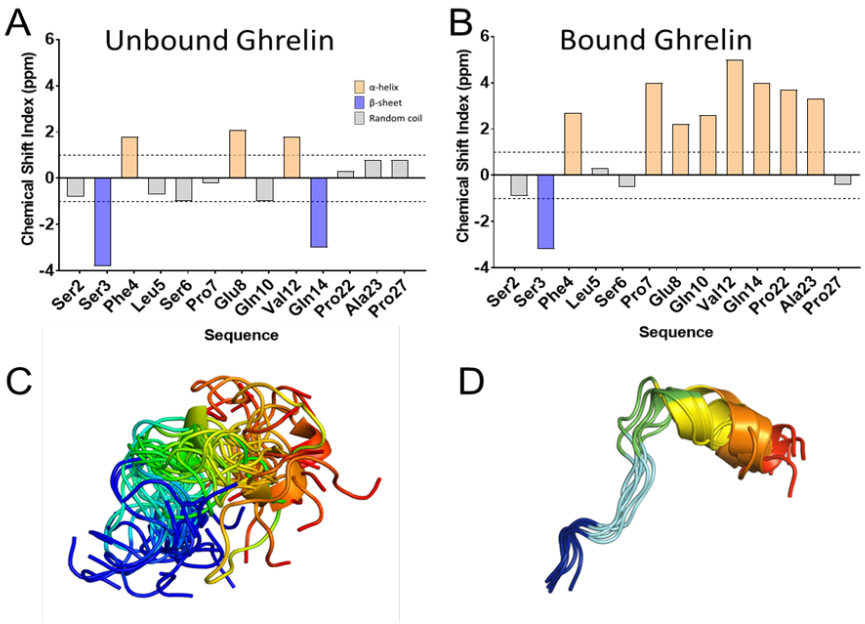
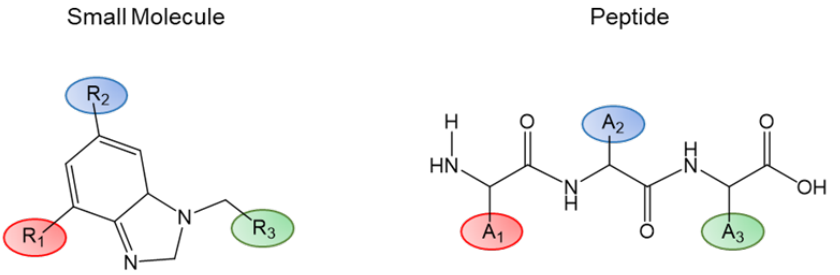


Figure 2.9: NMR measured conformational change in ghrelin upon binding receptor. (A,B) Chemical shift index measurements of select residues in the ghrelin peptide in the presence of empty membrane or membrane containing ghrelin receptor Bender et al. (2019); **?**. These measurements identify a degree of secondary structure formation in the presence of receptor. (C,D) The chemical shifts were used to build models of ghrelin peptide in its two states, colored blue to red from N- to C-terminus.

Of note, the studies described in the above section rely on a variety of biophysical techniques for structural characterization. While X-ray crystallography and, in some cases, cryo-EM can reveal the conformations of peptides binding to GPCRs, this is currently rare. This lack of structure availability is likely due to the inherent flexibility of peptide ligands, as described, which can hinder the crystallization process or identification of class averages. Complementary to these techniques, several studies have utilized NMR and CD to characterize the peptide structure. CD provides readily accessible information to the overall secondary structure changes in varying environments. However, often the structural details can only be assessed qualitatively and providing residue-based structural data on the basis of CD measurements is impossible. In contrast, NMR can provide detailed information on a residue and atomic level about the structural properties of these peptides. To this end, specific 13C and/or 15N labeling of the peptide ligand is usually required, which is easily done using solid state peptide synthesis. This way, detailed structural information for an individual residue can be obtained as described above. In addition to structural data, NMR, especially using saturation transfer difference (STD-) NMR, reveals information on population dynamics that may provide insights in the binding recognition process. Additional techniques used in the studies as mentioned earlier include EPR, H/DX-MS, cross-linking, and molecular modeling. Lastly, a powerful method used for decades in peptide ligand studies is the use of mutational analysis. Alanine scanning and backbone modification of peptides is analogous to traditional SAR studies of small molecule ligands (Figure 2.10). Future studies will likely need to combine multiple of these techniques to arrive at reliable understandings of these peptide-receptor complexes.



R_1	R_2	R_3
$-CH_2COOH$	$-CH(CH_3)_2$	-Phe(OH)
-H	$-CH_2$-imidazole	$-S-CH_3$
-guanidine	-H	-SH

A_1	A_2	A_3
-Asp	-Val	-Tyr
-Ala	-His	-Met
-Arg	-Gly	-Cys

Figure 2.10: Schematic of peptide structure-activity relationships (SAR). Much like swapping chemical moieties for small molecule (SAR), peptide mutagenesis and alanine scanning are important tools for determining peptide functionality at a given receptor.

2.5.2 Mimetics of the Bound-State Conformations can Aid in Structure Determination and Drug Discovery

As evidenced by the apelin, μOR, and neurotensin crystal structures, conformational stabilization or truncation of flexible components within the peptide ligands can assist in the crystallization of these complexes. While NT (8-13) and AMG3054 are less perturbed mimetics of the neurotensin and apelin, respectively, DAMGO represents a more dramatic change from the endogenous peptide ligand structure. Interpretation of these structures will need to be verified for the endogenous peptide ligands. SAR studies on peptides with no known crystal structures will be invaluable for understanding the conformational constraints required for these peptides in the bound states. Future crystallization trials with these conformationally constrained peptide derivatives will increase the likelihood of a stable crystal with interpretable density at the ligand binding site. Simultaneously, as the conformational constraints of these ligand binding sites become better understood, the development of more potent drug therapies may become more feasible. This vision has been evidenced clearly with the development of super-agonists for the gonadotropin-releasing hormone receptor. The addition of a D-amino acid enhanced the β-turn in the peptide that is needed for the bound state. It is suggested that the pre-orientation of the ligand conformation will reduce the entropic cost of binding, thereby increasing the affinity at the receptor. However, this suggestion has not yet been validated with stabilization attempts of neurotensin derivatives in which the best derivatives are still only on par with the endogenous peptide. Therefore, this theme will need to be investigated in future drug development to see if this consistently holds.

2.6 Conclusion

Recent studies imply that the receptors influence the conformation of their peptide ligands, and that the peptide ligand can alter the conformation of the receptors' extracellular loops. However, structural and dynamical studies on either the peptide ligands or receptors are often pursued independently. Our current understanding suggests that conformational selection is a prime driver of receptor recognition. As a result, it is essential to study the receptor in the presence of cognate ligand and design experiments to define that interface, since we see that the conformation of a peptide in the absence of the receptor does not predict its conformation in the receptor-bound state. They will likely also define essential differences in these systems' structural dynamics that evolved to allow for their diverse functions.

CHAPTER 3

The First Selective Y4 Receptor Antagonist Binds in a Deep, Allosteric Binding Pocket

Parts of this chapter was taken from the paper: Corinna Schüß*, Oanh Vu*, Mario Schubert*, Yu Du, Nigam M. Mishra, Iain R. Tough, Jan Stichel, C. David Weaver, Kyle A. Emmitte, Helen M. Cox, Jens Meiler, and Annette G. Beck-Sickinger Journal of Medicinal Chemistry 2021 64 (5), 2801-2814.

*Those authors contributed equally to this work

3.1 Summery

Human neuropeptide Y receptors (Y1R, Y2R, Y4R, and Y5R) belong to the superfamily of G protein-coupled receptors and play an important role in the regulation of food intake and energy metabolism. We identified and characterized the first selective Y4R allosteric antagonist (S) VU0637120, an important step towards validating Y receptors as therapeutic targets for metabolic diseases. To obtain insight into the antagonistic mechanism of (S)-VU0637120, we conducted a variety of in vitro, ex vivo, and in silico studies: competition binding, signal transduction at various Y4R/Y1R receptor chimeras, Y4R point mutations, in native intestinal tissue, and computational docking. These studies revealed that (S)-VU0637120 selectively inhibits Y4R in engineered cells lines, is active at inhibiting native Y4R function, and binds in an allosteric site located below the binding pocket of the endogenous ligand pancreatic polypeptide (PP) in the core of the Y4R transmembrane domains. Taken together our studies provide a first-of-its-kind tool for probing Y4R function and improve the general understanding of allosteric modulation, ultimately contributing to the rational development of allosteric modulators for peptide-activated G protein-coupled receptors (GPCR).

3.2 Introduction

The neuropeptide Y system is a multiligand/multireceptor family consisting of four Y receptors (Y1R, Y2R, Y4R and Y5R) and three peptide ligands: neuropeptide Y (NPY), peptide YY (PYY) and pancreatic polypeptide (PP). The NPY family is well known for its regulation of important physiological processes, including energy metabolism and food intake. Therefore, dysregulation of these processes can result in disorders, such as anorexia and obesity. Whereas the Y1R and Y5R induce orexigenic signals, the Y2R and Y4R display anorexigenic functions. As Y receptors have overlapping preferences for their peptidic agonists, the development of subtype-selective compounds is essential to enable a better understanding of the function of these receptors and to explore their therapeutic potential (Pedragosa-Badia et al., 2013b; Yulyaningsih et al., 2011). The Y1R and Y2R bind NPY and PYY with high affinity. The Y4R has a preference for PP compared

to NPY and PYY, and the Y5R is activated by all three ligands with comparable potency. Thus, it is very challenging to develop orthosteric ligands with high selectivity for one receptor subtype versus the other Y receptors. To date, a number of agonists and antagonists for Y receptors have been developed (Pedragosa-Badia et al., 2013b; Yulyaningsih et al., 2011). With respect to Y4R, the native ligand PP represents a potent and somewhat selective agonist that induces anorexigenic effects in humans (Batterham et al., 2003). The major challenge in the development of novel Y4R ligands remains obtaining selectivity with respect to the evolutionarily most closely related Y1R subtype (Hofmann et al., 2013; Parker et al., 1998). Selective Y4R agonists have been identified by crosslinking two peptide fragments derived from the C-terminal pentapeptide of PP/NPY (Balasubramaniam et al., 2006; Kuhn et al., 2016). However, no highly selective antagonists for Y4R have been described to date. Aiming at the development of small molecule probes to selectively target the human Y4R, approximately 220,000 compounds were tested via high-throughput screening (HTS). We recently reported on the discovery of Y4R positive allosteric modulators (PAM) from this experiment (Schubert et al., 2017a; Bryksin and Matsumura, 2010). In comparison to PAMs and agonists, the hit rate of negative allosteric modulators (NAM) and antagonists was very low (0.003%), reflecting the challenges in the development of Y4R antagonists (Keller et al., 2013; Kuhn et al., 2017). In this study, we present the in vitro, in silico, and ex vivo characterization of the first selective, small-molecule Y4R allosteric antagonist (S)-VU0637120 (Figure 3.1a) including the identification of its binding mode that allows further preclinical studies.

3.3 Method

3.3.1 Peptide ligands and compounds

The compound VU0637120 was purchased from Life Chemicals Inc. (Ukraine). The peptide ligands human pancreatic polypeptide (PP) and porcine neuropeptide Y (NPY) were synthesized by using automated solid-phase peptide synthesis and Fmoc (Fluorenylmethoxycarbonyl) strategy as described previously (Mäde et al., 2014; Schubert et al., 2017a; Sliwoski et al., 2016b). The enantiomers (R)- and (S)-VU0637120 were synthesized and purified individually. Detailed synthesis and characterization see Online Method section Synthesis and characterization of (S)- and (R)-VU0637120.

3.3.2 Cell culture

Wildtype COS7 cells were cultured in Dulbecco´s Modified Eagle Medium (DMEM, Lonza) supplemented with 10% heat-inactivated fetal bovine serum (FBS, Biochrom). The COS7 cell lines stably expressing one specific human Y receptor-eYFP fusion protein (Y1,2,4,5R-eYFP) and the chimeric G protein Δ6Gαqi4myr were created using pVitro2-hygro-mcs vectors as previously described (Mäde et al., 2014) and cultured in

DMEM supplemented with 10% FBS, 133 μg/ml hygromycin (Invivogen) and 1.5 mg/ml G418-sulfate (Amresco). Wildtype HEK293 cells were cultured in DMEM/Ham´s F12 (1:1 (vol/vol), Lonza) supplemented with 15% FBS. Stably transfected HEK293_Y4R-eYFP cells were generated using a hY4R-eYFP-pVitro2-neo-mcs vector as described (Böhme et al., 2008) and cultured in DMEM/Ham´s F12 (1:1, Lonza) with 15% FBS and 100 μ/ml hygromycin. All cell lines were cultured in a humidified atmosphere at 37°C and 5% CO2.

3.3.3 Plasmids

For the Y4R mutants screening, the cDNA of the Y4R, Y1R, Y4R/Y1R chimeras and Y4R mutants were cloned in a pEYFP_N1 vector (Clontech), C-terminally tagged with an enhanced yellow fluorescent protein (eYFP), as performed in previous studies (Merten et al., 2007; Pedragosa-Badia et al., 2014). The chimeric G protein Δ6Gαqi4myr cDNA was provided by E. Kostenis (Rheinische Friedrich-Wilhelm-Universität, Bonn, Germany) (Kostenis et al., 1997) and cloned into a pVitro2-mcs-neo vector. In the arrestin3 (arr3) recruitment studies, the Y4R was C-terminally tagged to Renilla luciferase 8 (Rluc8) and cloned in a pcDNA3 vector. Arr3 was N-terminally fused to venus fluorescent protein and cloned in a pcDNA3 plasmid (Schubert et al., 2017a; Wanka et al., 2017).

3.3.4 Generation of Y4R/Y1R chimeras and Y4R mutants

The chimeric Y1R/Y4R receptors were created using overlap extension PCR (Bryksin and Matsumura, 2010; Ho et al., 1989). In the first PCR, the N-terminal cDNA of the donor receptor subtype was amplified using a sense primer, which annealed before the N-terminal receptor sequence and contained a specific enzyme restriction site. The antisense primer annealed at the end of the desired segment and contained an overhang complementary to the cDNA of the second receptor subtype. The cDNA of the C terminal fragment was obtained in a second PCR using the cDNA from the second receptor subtype. Here, the sense primer annealed at the beginning of the desired segment and contained a 5´overhang sequence complementary to the cDNA of the N-terminal segment of the donor receptor, which was created in the first PCR. The antisense primer of the second PCR contained a restriction site sequence and annealed after the receptor sequence. The fragments obtained in both PCRs were purified using agarose gel electrophoresis and Promega Wizard® SV Gel and PCR Clean-Up System (Promega). In a third PCR, the receptor fragments were combined using the N-terminal and C-terminal fragments with sense and antisense primers annealing at the beginning or end of the respective fragments to obtain the chimeric cDNA. Afterwards, the PCR product was digested using BgIII and XbaI or SalI and NotI enzymes (ThermoFischer). The desired products were purified by gel electrophoresis and ligated into a pEYFP_N1 vector using T4 DNA ligase (ThermoFischer). The mutation of single amino

acids or 2-4 adjacent residues in the Y4R was performed by Quick change® site directed mutagenesis using PCR and appropriate primers. In the PCR, the Phusion (ThermoFischer) or Pfu Turbo DNA polymerase (Agilent) was used with 5X HF (Phusion) or 10X reaction (Pfu) buffer, 100 ng Y4R-eYFP_N1 template, 0.25-0.5 μM sense/antisense primers, 200 μM deoxyribonucleotide (dNTP) mix and 1-10% DMSO. The PCR conditions were adjusted according to the manufacturer's protocol of the DNA polymerase. Afterwards, the parental DNA was digested using Dpn1 (ThermoFischer). The PCR products were transformed into chemically competent E.coli DH5α by using heat shock (60 s, 42°C) and single clones were separated on Luria Broth (LB) medium-agarose plates containing kanamycin (30 μg/ml, Sigma-Aldrich). The plasmid DNA was prepared using Promega Wizard® SV Gel and PCR Clean-Up System or Promega PureyieldTM Midiprep System (Promega) according to manufacturer's protocol. Finally, the sequence of all Y4R/Y1R chimeras and Y4R mutants was confirmed by DNA sequencing.

3.3.5 Ca2+ flux assays

High-throughput screening of Y4R allosteric modulators was performed at the Vanderbilt HTS facility as previously described (Schubert et al., 2017a; Sliwoski et al., 2016b). The compound screening and the characterizations were performed with stably transfected COS7 cells expressing one specific Y receptor subtype (Y1,2,4,5R-eYFP) and the chimeric G protein Δ6Gαqi4myr as previously described (Schubert et al., 2017a). To investigate Y4R/Y1R chimera and Y4R mutants, COS7 cells were transiently transfected with one specific receptor construct (C-terminally tagged with eYFP) and the chimeric G protein Δ6Gαqi4myr using Metafectene® Pro transfection reagent. Therefore, COS7 cells were seeded in 25 cm2 culture flasks and grown to a confluency of 80%. Each flask was transfected with a total of 4 μg plasmid DNA (3 μg receptor-eYFP and 1 μg Δ6Gαqi4myr) by using 15 μl Metafectene® Pro according to manufacturer's protocol over night at 37°C. On the next day, 20,000 cells per well were reseeded in black, clear bottom 96 well plates (Greiner) and incubated over night at 37°C. For the assay, the cell culture medium was aspirated, and cells were incubated in 100 μl/well assay buffer (HBSS (Lonza), 20 mM HEPES (Sigma), 2.5 mM probenecid (Sigma), pH 7.4) containing 2.4 μM Fluo2 AM (Abcam) calcium dye. Afterwards, the dye solution was replaced by assay buffer and the Ca2+ signals were detected (excitation 485 nm; emission 525 nm) in a FlexStation III device (Molecular Devices). The compound VU0637120 and agonists (PP/NPY) were added automatically in a two-addition protocol. After 20 s baseline detection, DMSO control or VU0637120 were added to the cells, followed by the addition of PP or NPY after 110 s in a total run time of 140 s. The Ca2+ signal responses were analyzed as x-fold over basal values and normalized to the maximum activation obtained in the presence of 1 μM PP (Y4R, Y4R/Y1R chimeras and Y4R mutants) or NPY (Y1R, Y2R, Y5R).

3.3.6 Y4R membrane preparations and radioligand binding studies

Y4R membranes were generated from stably transfected HEK293_Y4R-eYFP cells, as previously described (Schubert et al., 2017a). Cells were suspended in DPBS and centrifuged at 1,800 rpm, 4°C for 5 min. The cell pellet was resuspended in Tris buffer (50 mM Tris (Sigma), 50 μM Pefabloc (Sigma), pH 7.5) and homogenized in a potter grinder (Potter SB Braun). The cell suspension was centrifuged at 2,400 rpm at 4°C for 20 min, the supernatant was removed and centrifuged for 60 min at 12,000 rpm, 4°C. The supernatant was discarded and the membrane pellet was resuspended in HEPES buffer (25 mM HEPES, 25 mM CaCl2 (Fluka), 1 mM MgCl2 (Fluka), 50 μM Pefabloc (Sigma), pH 7.4). After the homogenization in the potter grinder, the suspension was centrifuged at 12,000 rpm at 4°C for 60 min. The resulting pellet was resuspended in HEPES buffer and the protein concentration was determined using Bradford protein assay (Bradford, 1976). Membrane preparations were frozen in liquid nitrogen and stored at -20°C. For binding experiments, 0.3 μg of Y4R membrane preparation were suspended in HEPES buffer containing either DMSO as control or VU0637120. PP and 125I-PP (human) solutions were prepared in aqua dest. supplemented with 0.1% BSA (PAA Laboratories). Nonspecific binding was determined in the presence of 1 μM PP. Radioligand solutions (60 or 180 pM 125I-PP) were added to all samples and incubated for 5 h (at room temperature, 200 rpm). Afterwards, membrane bound 125I-PP was separated by filtration using GFC filter (PerkinElmer), presoaked with 0.1 % polyethylenimine (Sigma) in a MicroBeta 96-well filtermate harvester (PerkinElmer). Membranes were washed with cold PBS, dried for 15 min at 55°C and treated with MeltiLex scintillation sheets. For the quantification of radioactivity, membranes were measured in a MicroBeta scintillation counter (PerkinElmer).

3.3.7 IP-One assay

The IP-One assay was performed in COS7 cells stably expressing the Y4R-eYFP and the chimeric G protein Δ6Gαqi4myr. Therefore, 5,000 cells per well were seeded in white 384-well plates (Greiner) and incubated over night at 37°C. For the detection of IP production, the IP-one Gq assay kit (Cisbio) was used. The cell culture medium was removed, (S)-VU0637120 or DMSO control were added followed by the subsequent stimulation with increasing concentrations of PP for 1 h at 37°C. Afterwards, cells were lysed with lysis buffer containing antibody 1 (da-labeled IP1) and antibody 2 (anti-IP1-cryptate) according to the manufacturer's protocol. Signals were detected in a Tecan Spark plate reader (Tecan Group AG) measuring emission at 665 nm and 620 nm.

3.3.8 Arrestin3 recruitment assay

Arrestin3 recruitment to the Y4R was measured in a BRET assay in HEK293 cells transiently transfected with Y4R-Rluc8 and venus-arr3 fusion proteins as previously described (Schubert et al., 2017a). For kinetic

measurements, signals were measured continuously over 30 min after PP (100 nM) addition. BRET signals were observed as the ratio of fluorescence (venus-arr3) and luminescence (Y4R-Rluc8). NetBRET signals were calculated by subtraction of unstimulated controls.

3.3.9 Fluorescence microscopy

The internalization of the Y4R was investigated using fluorescence microscopy in HEK293 cells stably expressing the Y4R-eYFP fusion protein. 180,000 cells per well were seeded in μ-slide 8 well (IBIDI) chambers and incubated over night at 37°C. For microscopy, the cell culture medium was aspirated and nuclei of the cells were stained with 300 μl/well OptiMEM (serum-reduced media, Gibco) containing Hoechst33342 (Sigma) for 30 min at 37°C. Afterwards, the solution was aspirated and 150 μl/well OptiMEM containing 30 μM VU0637120 or DMSO control were added. 50 μl/well PP solution (10 nM PP or 100 nM PP final concentration) or OptiMEM (unstimulated samples) were added and the cells were stimulated for 30 min at 37°C. The stimulation solution was replaced by 200 μl OptiMEM. Microscopy was performed by using the AxioVert Observer Z1 (ZEISS) fluorescence microscope (YFP: filter set 46, DAPI: filter set 49, ApoTome, 63x/1.40 oil objective). The images were processed using ZEN2012 (ZEISS) and Axiovision SE64 (ZEISS) software. The detection of the eYFP fluorescence was performed with a constant exposure time (400 ms) to allow the quantification of cell surface receptors as performed previously (Schubert et al., 2017a). The cell surface fluorescence was quantified as pixel-intensity in black/white pictures by using ImageJ software and normalized to the unstimulated DMSO control.

3.3.10 Electrophysiological measurements

Vectorial ion transport was measured across mucosal preparations of mouse descending colon as described previously (Schubert et al., 2017a; Tough et al., 2006). Up to six adjacent preparations were dissected from each colon. Mucosae were placed individually between Ussing chamber halves and voltage-clamped at 0 mV (using a DVC1000; WPI, Sarasota, FL), as described previously. Mucosae with exposed areas of 0.14 cm^2 were bathed in oxygenated Krebs-Henseleit solution on both sides, at 37°C, and were voltage-clamped at 0 mV. A stable basal Isc was reached within 20 min after which DMSO (0.3%) or antagonist additions (300 nM – 30 μM of the active enantiomer (S)-VU637120 or 30 μM of inactive (R)-VU637120) were made to the basolateral reservoir. After an intermediary addition of the secretagogue VIP (10 nM, for 10-15 min) an optimal single rPP concentration (30 nM) (Tough et al., 2006) was added to activate epithelial Y4R and reduce Isc levels, followed 15 min later by PYY (10 nM). Changes in Isc to each drug addition (all to the basolateral reservoir) were pooled and presented as mean ± SEM. The pIC50 and statistical analysis was performed using GraphPad Prism software (version 7).

3.3.11 Homology modeling of Y4R

Sequences of Y4R and the templates (Table 3.1) were aligned using the Clustal Omega web server (Sievers et al., 2011). Multiple sequence alignment was then adjusted to ensure that the helix regions and reserved residues remain aligned, and to remove gaps within transmembrane α-helices. Y4R homology models were built using RosettaCM protocol (Song et al., 2013a) of the Rosetta3.9 software suites (Bender et al., 2016). Addition constraints were set up to account for the disulfide bond between Cys1.25 and Cys7.38, as well as between Cys3.25 and Cys5.25. The output Y4R homology models were evaluated by Rosetta total energy score, which includes knowledge-based energy terms such as hydrogen bonds, electrostatic interactions, and van der Waals packing (Song et al., 2013a). Top 10% of output models, that scored most favorably by Rosetta, were clustered based on RMSD cut-off of 2 Å using BCL::Cluster (Alexander et al., 2011). The clustering procedure was extensively discussed in precedent protocol paper (Combs et al., 2013). Finally, a Y4R model with the best Rosetta total energy score was chosen from each of ten largest clusters.

Template PDB ID and name	Seq. identity	Seq. similarity	Resolution (Å)
Y1R-xtal (neuropeptide Y1 receptor)	49	69	2.7
4s0v (human OX2 orexin receptor)	30	52	2.8
4zjc (human OX1 orexin receptor)	27	50	2.8
5dhg (human NOP receptor)	27	49	3
4n6h (human delta opioid receptor)	27	49	1.8
4djh (human kappa opioid receptor)	24	49	2.9
4dkl (mu opioid receptor)	26	48	2.8
5glh (human endothelin receptortype-B)	23	47	2.8
3odu (human chemokine receptor CXCR4)	23	47	2.5
5ndd (human protease activated receptor 2)	27*	48*	2.8

Table 3.1: **Ten class A GPCRs templates for building Y4R homology models**. Nine x-ray crystal structures of nine class A GPCRs were selected to be templates for Y4R homology modeling based on the sequence identity and similarity to Y4R as well as the quality of the crystal structures. Furthermore, TM1, TM2, and TM7 regions of PAR2 crystal structure was used as a template for the corresponding region of Y4R as the protein has a similar allosteric binding site. PAR2 was the only template in an intermediate state, the other templates were in inactive state. The sequence identity was computed using CLUSTER W v1.83 1 and the SequenceAlignment application of BCL v3.2.2 2 was used to calculate sequence similarity using BLOSUM62 matrix 3. The sequence identity and similarity between human protease activated receptor 2 (PAR2) and Y4R (marked with asterisk) were calculated based on sequence of TM1, TM2 and TM7 only.

3.3.12 Docking NAM to Y4R

As the experimental data suggested that the ligand could interact with residues located on TM1, TM2, and TM7, it is crucial to create enough space among those three transmembrane regions in the homology models before ligand docking. The output Y4R homology models from the first round of RosettaCM did not yield enough space for VU0637120 to approach and to form hydrogen bonds to the residue Y1.39. Hence, the Y4R models were re-hybridized with the ligand placed inside the space among important residues. To create a set of templates for the second rounds of RosettaCM, 65 different poses of the ligand were placed inside Y4R such that it interacted to the TM1, TM2, TM7, and in no more than 4 Å from Y1.39. This way, the interac-

tions between the ligand and the important residues according to the mutagenesis data would be weighed in the placement of helices in MC trajectory and optimization. The resulting models were then filtered for interactions to important residues and hydrogen bond to Y1.39 and clustered. The final ten cluster representative models were selected to be templates for the docking protocol based on the size of the corresponding clusters as well as the openness of the binding pocket suggested by the mutagenesis data.

Res #	cluster1	cluster 2	cluster 3
Y1.39	95.52	100.00	93.48
E1.42	0.00	6.15	0.00
T1.43	0.00	0.00	0.00
Q2.58	56.72	87.69	94.93
T2.61	100.00	100.00	100.00
Y2.64	100.00	100.00	100.00
W2.70	76.12	76.92	100.00
S3.28	82.09	98.46	22.46
Q3.32	98.51	100.00	100.00
F4.60	100.00	16.92	3.62
F7.35	100.00	95.38	100.00
L7.36	100.00	100.00	99.28
H7.39	100.00	100.00	100.00

	Cluster 1	Cluster 2	Cluster 3
Score:	33.87	35.56	32.20

Res#	cluster 1	cluster 2	cluster 3
C1.26	0.00	9.23	0.00
D1.27	0.00	100.00	0.00
S1.28	29.85	0.00	0.00
V1.29	0.00	0.00	0.00
V1.31	0.00	23.08	0.00
M1.32	94.03	66.15	0.00
V1.33	0.00	0.00	0.00
F1.34	0.00	0.00	0.00
I1.35	62.69	100.00	15.94
V1.36	58.21	0.00	0.00
T1.37	0.00	0.00	0.00
S1.38	0.00	0.00	0.00
S1.40	0.00	0.00	0.00
V1.44	0.00	0.00	0.00
C2.54	0.00	3.08	5.80
L2.55	0.00	0.00	0.00
L2.56	0.00	0.00	0.00
L2.60	5.97	23.08	0.72
T2.65	100.00	100.00	65.94
I2.66	2.99	0.00	0.00
D2.68	35.82	89.23	0.00
M3.27	0.00	0.00	0.00
A3.29	100.00	86.15	19.57
C3.33	86.57	53.85	65.94
M3.34	0.00	0.00	0.00
S3.35	0.00	9.23	97.10
V3.36	11.94	7.69	93.48
A4.62	0.00	0.00	0.00
N4.63	0.00	0.00	0.00
S4.64	97.01	18.46	2.17
I5.3	14.93	0.00	0.00

Res #	cluster 1	cluster 2	cluster 3
I5.5	0.00	0.00	0.00
K5.22	0.00	4.62	0.00
V5.23	0.00	0.00	25.36
V5.24	25.37	3.08	98.55
T5.26	89.55	10.77	42.75
E5.27	92.54	4.62	0.00
S5.28	0.00	0.00	0.00
W5.29	10.45	0.00	0.00
L5.31	0.00	0.00	0.00
H5.34	0.00	0.00	0.00
R5.35	71.64	0.00	0.00
T5.36	0.00	0.00	0.00
I5.37	0.00	0.00	0.00
Y5.38	2.99	0.00	0.00
T5.39	73.13	1.54	2.17
F5.41	0.00	0.00	0.00
L5.42	10.45	0.00	1.45
W6.48	4.48	0.00	86.96
H6.52	4.48	0.00	47.10
F6.54	17.91	0.00	0.00
N6.55	97.01	60.00	89.86
E6.58	94.03	30.77	92.75
D6.59	4.48	0.00	0.00
W6.60	0.00	0.00	0.00
I7.26	4.48	0.00	16.67
I7.28	0.00	0.00	0.00
N7.32	40.30	100.00	84.06
L7.40	2.99	55.38	20.29
L7.41	0.00	0.00	0.00
M7.43	8.96	96.92	100.00

Table 3.2: **Interaction contact between Y4R and (S)-VU0637120 in residue level.** The first column is the Ballesteros-Weinstein number 4 of 13 important residues (green) and 61 unimportant residues (black). The last three columns are the percentages of the cluster models that have the ligand within 5 A from those 74 residues; the cells are colored according to the magnitude of the percentages such that blue is closer to 0%, and red is closer to 100%.

A set of 106 conformations of VU0637120 was generated using the ConformerGenerator application of BCL v3.2.2 (Kothiwale et al., 2015). RosettaLigand was used to dock VU0637120 to Y4R homology models. The docking protocol is described in a previous publication (Combs et al., 2013). In addition to Rosetta total energy score, the interface_delta score, which is relative to the predicted binding energy between VU0637120 and Y4R, was also extracted for all 20,000 output docking models. We selected models with total Rosetta score of at least 95% of the best total Rosetta score attained and with top 10% interface delta score. We applied an experimental filter on the selected docking models. The filter makes sure that

the selected ligand poses that contact nine out of the 11 residues that were shown to be important to the antagonistic activity by mutagenesis data. A residue is determined to contact the ligand if at least one atom of the residue is at most 5 Å away from any atom of the ligand. Those filtered models were then clustered based the ligand position with 3 Å RMSD cut-off. For each of five largest clusters, the interaction contact and strength were computed for each cluster such that the scores are higher when an important residue has high contact frequency to the ligand or favorable binding energy toward the ligand. The interaction contact score is computed as: $Interactioncontact(t) = \sum ct_i \times 4.7 - ct_n$,where ct_i and ct_n are the percent of docking poses that have the ligand contact with residues that have been identified as important and residues that were not confirmed as critical for binding residue, respectively. The weight of 4.7 was chosen to balance the impact of 13 critical with 61 non-critical residues. Similarly, the interaction strength of a cluster t is calculated as $Interactionstrength(t) = \sum ddg_i \times -4.7 + ddg_n$, where dd_gi and ddg_n are the predicted binding energies between the ligand and an important residue and non-critical residue, respectively. In the cluster with the most favorable interaction contact and strength scores, a model that best explains the mutagenesis data will be chosen as a putative binding pose of VU0637120 and Y4R. Per residue contract and interaction strength are listed in the Table 3.2 and Table 3.3.

3.3.13 Predicting gain-of-function Y4R mutants

The residues that interacts with VU0637120 were redesigned with the Rosetta binding pocket design protocol (Moretti et al., 2016). Predicted mutants that are located on TM3, TM4, and TM5 are selected only if they specifically interact with the methylsulfonyl tail and the six-member ring head of the pose of VU0637120 cluster 1. Single mutant models of the predicted gain-of-function mutants are then rebuilt with RosettaCM to verify favorable selective interactions of new residues to VU0637120 cluster 1.

3.3.14 Docking PP to Y4R

PP was docked to Y4R homology models (built in step 10) through four rounds of RosettaCM and Flex-PepDock with gradually increasing length of PP (last 5, last 8, and last 23 residues of PP). We removed the ECL2 from the Y4R homology models (HM) before docking 5mer PP, and added the ECL2 back in after docking 23mer PP. Similar protocol can be found in a previous publication (Yang et al., 2018). The docking was guided by experimental restraints from mutagenesis studies (Pedragosa-Badia et al., 2014). Details of the restraints are described in Table 3.4.

Res#	cluster1	cluster 2	cluster 3
Y1.39	-0.27	-0.38	0.02
E1.42	0.00	0.15	0.33
T1.43	0.01	-0.01	-0.03
Q2.58	-0.33	0.40	0.46
T2.61	-0.36	-0.05	0.00
Y2.64	-0.42	-0.95	-0.62
W2.70	0.01	-0.09	-0.43
S3.28	0.12	0.94	0.18
Q3.32	-0.94	1.10	-0.09
F4.60	-0.17	-0.02	0.00
F7.35	-0.16	-0.09	-0.07
L7.36	0.37	-1.08	0.24
H7.39	-0.03	-0.30	-0.49

	Cluster 1	Cluster 2	Cluster 3
Score	8.79	2.94	5.36

Res#	cluster 1	cluster 2	cluster 3
C1.26	0.00	0.04	0.00
D1.27	0.00	-0.07	0.00
S1.28	0.01	0.00	0.00
V1.29	-0.01	0.00	0.00
V1.31	0.00	0.03	0.01
M1.32	-0.20	0.01	-0.01
V1.33	0.00	0.00	0.00
F1.34	0.08	0.00	0.00
I1.35	-0.10	-0.02	0.35
V1.36	-0.13	0.03	0.04
T1.37	-0.02	0.00	0.00
S1.38	0.05	0.06	0.09
S1.40	0.00	0.00	0.00
V1.44	0.00	0.00	0.00
C2.54	0.02	-0.10	0.00
L2.55	0.00	0.00	0.00
L2.56	-0.02	0.00	0.02
L2.60	-0.01	0.02	-0.01
T2.65	-0.22	-0.62	0.02
I2.66	-0.15	0.00	0.00
D2.68	-0.02	0.00	0.00
M3.27	0.12	0.01	0.01
A3.29	-0.36	-0.03	0.01
C3.33	-0.15	-0.08	0.01
M3.34	0.04	0.05	0.04
S3.35	0.01	0.12	0.46
V3.36	-0.03	-0.10	-0.12
A4.62	0.00	0.00	0.00
N4.63	-0.03	0.00	0.00
S4.64	-0.13	0.00	0.00
I5.3	-0.01	0.00	0.00

Res #	cluster 1	cluster 2	cluster 3
I5.5	0.01	0.14	0.19
K5.22	-0.01	0.03	0.08
V5.23	0.01	-0.01	-0.04
V5.24	-0.05	0.13	-0.02
T5.26	-0.08	0.23	0.32
E5.27	-0.05	0.03	0.07
S5.28	0.06	0.03	0.05
W5.29	0.01	0.00	0.00
L5.31	0.00	0.00	0.00
H5.34	0.00	0.00	0.00
R5.35	0.01	-0.01	-0.02
T5.36	0.00	0.00	0.00
I5.37	0.00	0.00	0.00
Y5.38	-0.01	0.00	0.00
T5.39	-0.04	0.00	0.00
F5.41	0.00	0.00	0.00
L5.42	-0.04	0.00	0.00
W6.48	0.13	0.00	-0.02
H6.52	0.01	-0.01	-0.01
F6.54	-0.09	0.02	0.03
N6.55	0.02	-0.01	0.03
E6.58	0.20	0.06	0.02
D6.59	0.01	0.00	0.00
W6.60	0.00	0.00	0.00
I7.26	0.01	-0.04	-0.05
I7.28	0.00	0.00	0.00
N7.32	-0.16	-0.14	-0.03
L7.40	-0.13	0.87	0.60
L7.41	0.00	0.28	0.23
M7.43	0.02	0.21	0.59

Table 3.3: **Interaction strength between Y4R and (S)-VU0637120 in residue level.** The first column is the Ballesteros-Weinstein number 4 of 13 important residues (green) and 61 unimportant residues (black). The last three columns are the average binding energy score that is broken down to each of those 74 residues; the cells are colored according to the magnitude of binding energy, blue is more negative (more favorable), red is more positive (less favorable).

Y4 Residues	PP residues	Centroid restraints	Full atom restraints	Proposed interaction
Y2.64 or/and Y2.69	Y27	CB within 7A	CB within 7A	Unknown
D6.59	R35	CB within 7A	OD1-NH1 within 4A OD2-NH2 within 4A	Salt bridge/Hbond
D7.32	R33	CB within 7A	• Distance: combination of OD1 and 1(2)HH1(2)/HE within 4A • Angle: combination of possible acceptor-H-donor > 90	Hbond
F7.35	R33	CB within 7A	CZ-NH1 within 7A CG-NH1 within 7A	Pi-cation
F7.35	Y36	CB within 7A	CZ-CZ within 7A	Unknown
D2.68	??	None	None	Unknown
W2.70	??	None	None	Unknown

Table 3.4: **Experimental restraints used to guide docking of PP to Y4R.**

3.3.15 Data analysis

Data analysis was performed using GraphPad Prism 5.0 (or version 7.0, for electrophysiology) and python matplotlib packages. EC50 and Emax values of concentration-response curves were obtained by nonlinear regression analysis (curve fit). The quantification of the effect of VU0637120 in Ca2+ flux assays was performed using an operational model of allosterism (Leach et al., 2007) (1). The parameters of the model describe: Em – maximum system response; n – related to the slope of response curves; [A] and [B] – concentration of agonist (A) and modulator (B); τA and τB - intrinsic agonism (efficacy) of the agonist (A) and the compound (B); α and β – effect of the modulator on the affinity (α) or the efficacy (β) of the agonist A. In the calculations, the KA of the agonists PP (logKA -10.2) was constrained to values determined in previous binding assays (Schubert et al., 2017a).

$$E = \frac{E_m(\tau_A[A](K_B + \alpha\beta[B]) + \tau_B[B]K_B)^n}{([A]K_B + K_AK_B + K_A[B] + \alpha[A][B])^n + (\tau_A[A](K_B + \alpha\beta[B]) + \tau_B[B]K_A)^n)} \tag{3.1}$$

The arr3 recruitment rates k were calculated from kinetic traces using the one-phase association equation, as performed previously (Schubert et al., 2017a).

$$Y = Y_0 + (E_m - Y_0)(1 - e^{-kx}) \tag{3.2}$$

Schild analysis for allosteric antagonist was performed using equation (3), whereas [B] describes the concentration of the allosteric antagonist, EC50 refers to the EC50 of the DMSO control and EC50 to the EC50 in the presence of (S)-VU0637120 (Kenakin, 2006).

$$\log\left(\frac{EC'_{50}}{EC_{50} - 1}\right) = \log\left[\frac{[B](1-\alpha)}{\alpha[B] + K_B}\right] \tag{3.3}$$

3.3.16 Synthesis and characterization of (S)- and (R)-VU0637120

Synthesis and purification. Air sensitive reactions were carried out under a nitrogen atmosphere (Airgas Catalog No. NI UHP300). The following solvents were employed for chemical reactions: dichloromethane (99.9%, Extra Dry, AcroSealTM, Acros Organics Catalog No. 610300010) and N,N-dimethylformamide (Anhydrous, 99.8%, packaged under Argon in resealable ChemSealTM bottles, Alfa Aesar Catalog No. 43997). The following solvents were employed for working up reactions and/or extractions: ethyl acetate (Certified ACS grade, Fisher Chemical Catalog No. E145-20), ethyl ether (Anhydrous, BHT stabilized, Certified ACS, Fisher Chemical Catalog No. E138-4), and dichloromethane (Not Stabilized, HPLC grade, Fisher Chemical Catalog No. D150-4). Brine was prepared from deionized water and sodium chloride (Reagent

grade, Fisher Chemical Catalog No. S25541B). Anhydrous sodium sulfate (Lab grade, Fisher Chemical Catalog No. S25568A) was employed for drying organic extracts. Thin layer chromatography (TLC) was conducted on glass plates coated with Silica Gel 60 F254 from Millipore Sigma (Catalog No. 1057150001). Normal phase flash chromatography was carried out on either a CombiFlash® EZ Prep or CombiFlash® Rf+ automated flash chromatography system, both from Teledyne ISCO. Normal phase flash chromatography was carried out using RediSep® Rf normal phase disposable flash columns (40-60 micron) from Teledyne ISCO (Catalog Nos. 69-2203-304, 69-2203-312, 69-2203-324, 69-2203-340, 69-2203-380, and 69-2203-320). The following solvents were employed for TLC and normal phase chromatography: hexanes (Certified ACS grade, Fisher Chemical Catalog No. H292-20), ethyl acetate (Certified ACS grade, Fisher Chemical Catalog No. E145-20), dichloromethane (Not Stabilized, HPLC grade, Fisher Chemical Catalog No. D150-4), and methanol (HPLC grade, Fisher Chemical, Catalog No. A452-4). Reverse phase preparative HPLC was carried out on a CombiFlash® EZ Prep automated flash chromatography system equipped with a RediSep® Prep C18 10 x 250 mm, 100Å, 5 μm HPLC preparative column from Teledyne ISCO (Catalog No. 692203809). The following solvents were employed for reverse phase chromatography: acetonitrile (HPLC grade, Fisher Chemical Catalog No. A998SK-4) and water purified using a Milli-Q® Advantage A10 Water Purification System from Millipore Sigma.

Characterization. All NMR spectra were recorded on a 300 MHz Bruker Fourier 300HD NMR spectrometer equipped with a dual 1H and 13C probe with Z-Gradient and automatic tuning and matching, full computer control of all shims with TopShimTM, 24-sample SampleCaseTM automation system, and TopSpinTM software. All NMR samples were prepared with either methyl sulfoxide-d6 with 0.03% TMS, 99.8 atom % D, Acros Organics Catalog No. 360000100) or chloroform-d with 0.03% TMS, 99.8+ atom % D, Acros Organics Catalog No. 209561000). 1H and 13C chemical shifts are reported in δ values in ppm downfield with tetramethylsilane (TMS) as the internal standard. Data are reported as follows: chemical shift, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, b = broad, m = multiplet), integration, coupling constant (Hz). High resolution mass spectrometry was conducted on an Agilent 6230 Accurate-Mass Time-of-Flight (TOF) LC/MS with ESI source equipped with MassHunter Walkup software. MS parameters were as follows: fragmentor: 175 V, capillary voltage: 3500 V, nebulizer pressure: 35 psig, drying gas flow: 11 L/min, drying gas temperature: 325 ºC. Samples were introduced via an Agilent 1260 Infinity UHPLC comprised of a G4225A HiP Degasser, G1312B binary pump, G1367E ALS, G1316A TCC, and G1315C DAD VL+ with a 5 μL semi-micro flow cell with a 6 mm path length. UV absorption was observed at 220 nm and 254 nm with a 4 nm bandwidth. Column: Agilent Zorbax SB-C18, Rapid Resolution HT, 1.8 μm, 2.1 x 50 mm. Gradient conditions: Hold at 5% CH3CN in H2O (0.1% formic acid) for 1.0 min, 5% to 95% CH3CN in H2O (0.1% formic acid) over 5 min, hold at 95% CH3CN in H2O (0.1% formic acid) for 1.0 min,

0.5 mL/min. All final analogs were at least 95% pure according to these analytical methods.

Synthesis and Characterization of (S)-VU0637120[a]

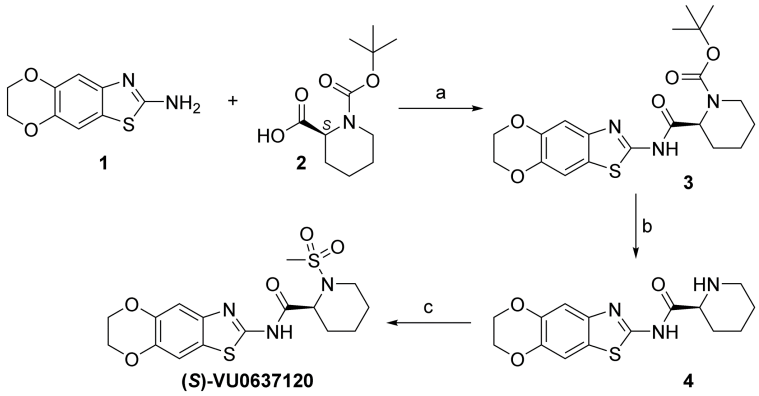
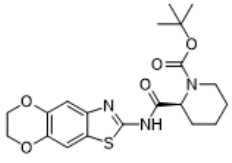


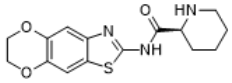
Figure 3.1

[a]*Reagents and conditions:* (a) DIEA, HATU, DMF, 86%; (b) HCl, dioxane, CH2Cl2, 89%; (c) NEt3, MeSO2Cl, CH2Cl2, 60%.



tert-Butyl(S)-2-((6,7-dihydro-[1,4]dioxino[2',3':4,5]benzo[1,2-d]thiazol-2-yl) carbamoyl) piperidine-1-carboxylate (3).

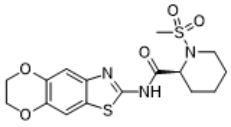
N,N-Diisopropylethylamine (DIEA) (331 μL, 1.92 mmol) was added to a mixture of N-Boc-L-pipecolinic acid (242 mg, 1.06 mmol), 1-[bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxid hexafluorophosphate (HATU) (734 mg, 1.92 mmol) in DMF (4 mL) and stirred for 20 min. A solution of 1 (200 mg, 960 mmol) in DMF (2 mL) was added dropwise, and the solution was allowed to stir for 12 h. After completion of the reaction as judged by TLC, 15 mL of water was added. The mixture was extracted with ethyl acetate (2x), washed with brine, and dried over anhydrous sodium sulfate. The crude product was purified by flash chromatography using ethyl acetate and hexane as eluent, and the title compound was obtained as an off-white solid (346 mg, 86%). 1H NMR (300 MHz, CDCl3) δ 9.50 (br s, 1H), 7.26 (s, 2H), 5.02 (br s, 1H), 4.30 (s, 4H), 4.06 (m, 1H), 2.91 – 2.75 (m, 1H), 2.37 (m, 1H), 1.78 – 1.36 (m, 5H), 1.51 (s, 9H). 13C NMR (75 MHz, CDCl3) δ 171.18, 169.79, 156.48, 143.42, 143.14, 141.92, 125.10, 108.59, 108.49, 81.52, 64.33, 60.41, 28.36, 24.97, 21.08, 20.56, 14.22. HRMS, calc'd for C20H26N3O5S+ [M+H], 420.1588; found 420.1600.



(S)-N-(6,7-dihydro-[1,4]dioxino[2',3':4,5]benzo[1,2-d]thiazol-2-yl)piperidine-2-carboxamide (4).

Compound 3 (340 mg, 0.811 mmol) was dissolved in dichloromethane (6 mL) and cooled to 0 ºC in

an ice bath. Hydrogen chloride solution (4 mL, 4.0 M in dioxane, 16 mmol) was added dropwise, and the reaction mixture was stirred at room temperature for 2 h. After completion of reaction as judged by TLC, the majority of the solvent was removed in vacuo, and the product was precipitated by the addition of diethyl ether. The crude product was filtered, washed with diethyl ether (2x), dried, and purified by preparative HPLC using acetonitrile and water as eluent (20% acetonitrile in water). The title compound was obtained as off-white solid (230 mg, 89%). 1H NMR (300 MHz, DMSO-d6) δ 9.90 (d, J = 10.0 Hz, 1H), 9.13 (m, 1H), 7.51 (s, 1H), 7.26 (s, 1H), 4.29 (s, 4H), 4.15 (m, 1H), 3.30 (d, J = 11.8 Hz, 1H), 2.96 (m, 1H), 2.29 (d, J = 11.4 Hz, 1H), 1.93 – 1.41 (m, 5H). 13C NMR (75 MHz, DMSO) δ 168.73, 156.70, 143.67, 143.08, 142.04, 124.55, 109.28, 108.37, 64.48, 57.44, 43.80, 27.18, 22.03, 21.54. HRMS, calc'd for C15H18N3O3S+ [M+H], 320.1063; found 320.1067.

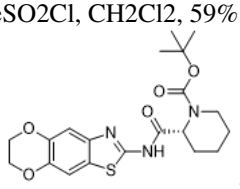


(S)-N-(6,7-dihydro-[1,4]dioxino[2',3':4,5]benzo[1,2-d]thiazol-2-yl)-1-(methyl sulfonyl)-piperidine-2- carboxamide ((S)-VU0637120)).

Triethylamine (128 μL, 0.936 mmol) was added to a solution of compound 4 in dichloromethane (2 mL) at 0 ºC in an ice bath and stirred for 5 min. Methanesulfonyl chloride (40.0 μL, 0.376 mmol) was added dropwise, and the reaction was warmed to room temperature and stirred for an additional 1 h. After completion of the reaction, water (15 mL) was added to the reaction mixture, and it was extracted with dichloromethane (2x). The crude product was purified by flash chromatography using ethyl acetate and hexane as eluent, and the title compound was obtained as an off-white solid (75 mg, 60%). 1H NMR (300 MHz, DMSO-d6): δ 12.39 (br s, 1H), 7.45 (s, 1H), 7.21 (s, 1H), 4.72 (d, J = 4.6 Hz, 1H), 4.27 (s, 4H), 3.67 – 3.47 (m, 2H), 2.93 (s, 3H), 1.89 – 1.34 (m, 6H). 13C NMR (75 MHz, CDCl3) δ 169.13, 156.52, 143.55, 142.80, 142.06, 124.86, 108.62, 108.37, 64.31, 64.26, 55.95, 43.34, 39.66, 26.35, 24.29, 19.73. HRMS, calc'd for C16H20N3O5S2+ [M+H], 398.0839; found 398.0848. [α]25D –66º (c 0.098, CHCl3).

Synthesis and Characterization of (R)-VU0637120[a]

[q]*Reagents and conditions:* (a) DIEA, HATU, DMF, 85%; (b) HCl, dioxane, CH2Cl2, 93%; (c) NEt3, MeSO2Cl, CH2Cl2, 59%.



tert-Butyl(S)-2-((6,7-dihydro-[1,4]dioxino[2',3':4,5]benzo[1,2-d]thiazol-2-yl) carbamoyl) piperidine-1-carboxylate (6).

Intermediate 6 was synthesized analogous to intermediate 3 using N-Boc-D-pipecolinic acid. 1H NMR (300 MHz, CDCl3) δ 9.51 (br s, 1H), 7.27 (s, 2H), 5.02 (br s, 1H), 4.30 (s, 4H), 4.06 (m, 1H), 2.89 – 2.75

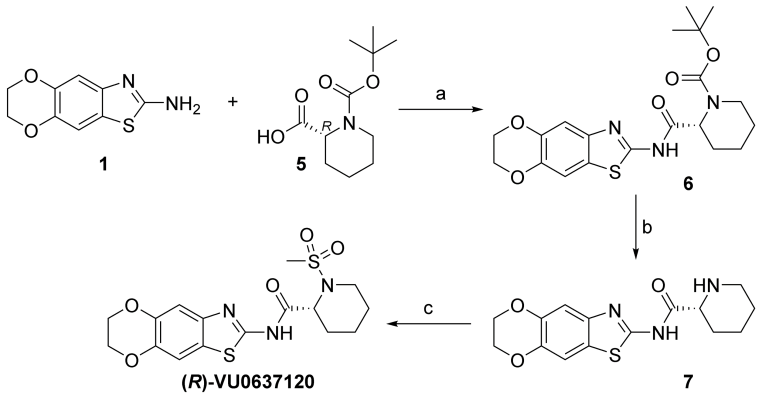
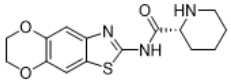


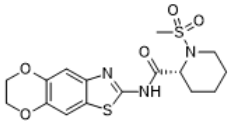
Figure 3.2

(m, 1H), 2.39 (m, 1H), 1.77 – 1.36 (m, 5H), 1.51 (s, 9H). 13C NMR (75 MHz, CDCl3) δ 169.80, 156.53, 143.43, 143.07, 141.94, 140.90, 125.07, 108.57, 108.50, 81.52, 64.31, 60.42, 28.36, 24.69, 21.08, 20.57, 14.22. HRMS, C20H26N3O5S+ [M+H], 420.1588; found 420.1607.



(S)-N-(6,7-dihydro-[1,4]dioxino[2',3':4,5]benzo[1,2-d]thiazol-2-yl) piperidine-2-carboxamide (7).

Intermediate 7 was synthesized analogous to intermediate 4. 1H NMR (300 MHz, DMSO-d6) δ 9.83 (d, J = 9.8 Hz, 1H), 9.13 (m, 1H), 7.51 (s, 1H), 7.26 (s, 1H), 5.26 (br s, 1H), 4.29 (s, 4H), 4.13 (m, 1H), 3.30 (d, J = 12.0 Hz, 1H), 3.07 (m, 1H), 2.96 (m, 1H), 2.28 (d, J = 11.4 Hz, 1H), 1.88 – 1.47 (m, 5H). 13C NMR (75 MHz, DMSO-d6) δ 168.73, 156.71, 143.72, 143.07, 142.05, 124.55, 109.30, 108.38, 64.49, 57.47, 43.81, 27.18, 22.04, 21.56. HRMS, calc'd for C15H18N3O3S+ [M+H], 320.1063; found 320.1075.



(R)-N-(6,7-dihydro-[1,4]dioxino[2',3':4,5]benzo[1,2-d]thiazol-2-yl)-1-(methyl sulfonyl)-piperidine-2-carboxamide ((R)-VU0637120)).

(R)-VU0637120 was synthesized analogous to (S)-VU0637120. 1H NMR (300 MHz, DMSO-d6): δ 12.39 (br s, 1H), 7.45 (s, 1H), 7.21 (s, 1H), 4.72 (d, J = 4.2 Hz, 1H), 4.28 (s, 4H), 3.62 – 3.47 (m, 2H), 2.93 (s, 3H), 1.89 – 1.19 (m, 6H). 13C NMR (75 MHz, CDCl3) δ 168.88, 156.36, 143.57, 142.91, 142.09, 124.95, 108.70, 108.40, 64.34, 64.29, 56.03, 43.43, 39.86, 26.06, 24.26, 19.82. HRMS, calc'd for C16H20N3O5S2+ [M+H], 398.0839; found 398.0849. [α]25D +73º (c 0.098, CHCl3).

3.4 Results

3.4.1 (S)-VU0637120 selectively inhibits Y4R activation by PP through an allosteric mechanism

The compound VU0637120 was identified as the most potent Y4R small molecule antagonist in the HTS and is characterized in detail within this study. VU0637120 was first investigated for its effect to modulate the G protein signaling pathway at the Y4R performing Ca2+ flux assays. G protein activation studies confirmed that VU0637120 effectively decreases Y4R activation by PP in a concentration-dependent manner. This is indicated by the rightward-shift of PP concentration-response curves with increasing VU0637120 concentrations (Figure 3.1b). The activity of VU0637120 at the Y4R was quantified using an operational model of allosterism. This model allows the calculation of individual parameters like the equilibrium binding constant (KB) of VU0637120 as well as the effect of VU0637120 on the modulation of orthosteric agonist potency (α) and efficacy (β) (Figure 3.1b) (Leach et al., 2007). VU0637120 has a negative allosteric effect on the affinity of PP to its Y4R orthosteric binding pocket (α = 0.02) and slightly reduces signaling efficacy (β = 0.77). The affinity (KB) of VU0637120 to its allosteric binding pocket is in the range of 300 – 400 nM (pKB 6.4 ± 0.08). VU0637120 has one chiral center (Figure 3.1a). To determine whether both stereoisomers are active, we synthesized enantiopure (S)- and (R) VU0637120 and tested them for their ability to reduce a submaximal PP response (PP EC80) at the Y4R (Figure 3.1c). Here, (S)-VU0637120 was confirmed as the active stereoisomer of the compound, as it is able to effectively decrease PP response at the Y4R with an IC50 value of 2.8 μM (pIC50 5.5 ± 0.03). In contrast, (R)-VU0637120 is inactive at the Y4R up to the top concentration tested.

To validate the effect of the active enantiomer (S)-VU0637120, we also tested (S)-VU0637120 on the G protein signaling pathway at the Y4R using different functional assays: Ca2+ flux assays, a hemi-equilibrium kinetic assay and inositol phosphate (IP)-One assay as an equilibrium assay (Bdioui et al., 2018). We demonstrate that (S)-VU0637120 reduces PP response at the Y4R in a concentration-dependent manner in Ca2+ flux assay, as was also shown for the racemic form of VU0637120 (Figure 3.1b), and IP-One assay. Schild analysis of the data reveal that the antagonistic effect of (S)-VU0637120 is saturable at concentrations larger than 10 μM, indicating that the compound does not act as a pure orthosteric antagonist and support an allosteric mode of action (Figure 3.4.1a, b).

As previous studies have already shown the importance of the arrestin3 (arr3) pathway for Y4R signaling (Wanka et al., 2017), we investigated the activity of VU0637120 on the arr3 recruitment as an alternative downstream event to G protein signaling. PP induced arr3 recruitment to the Y4R was analyzed in kinetic experiments by detecting arr3 recruitment constantly to the Y4R over 30 min (Figure 3.1d). The study demonstrates, that VU0637120 slows arr3 recruitment to the Y4R in response to PP compared to the DMSO

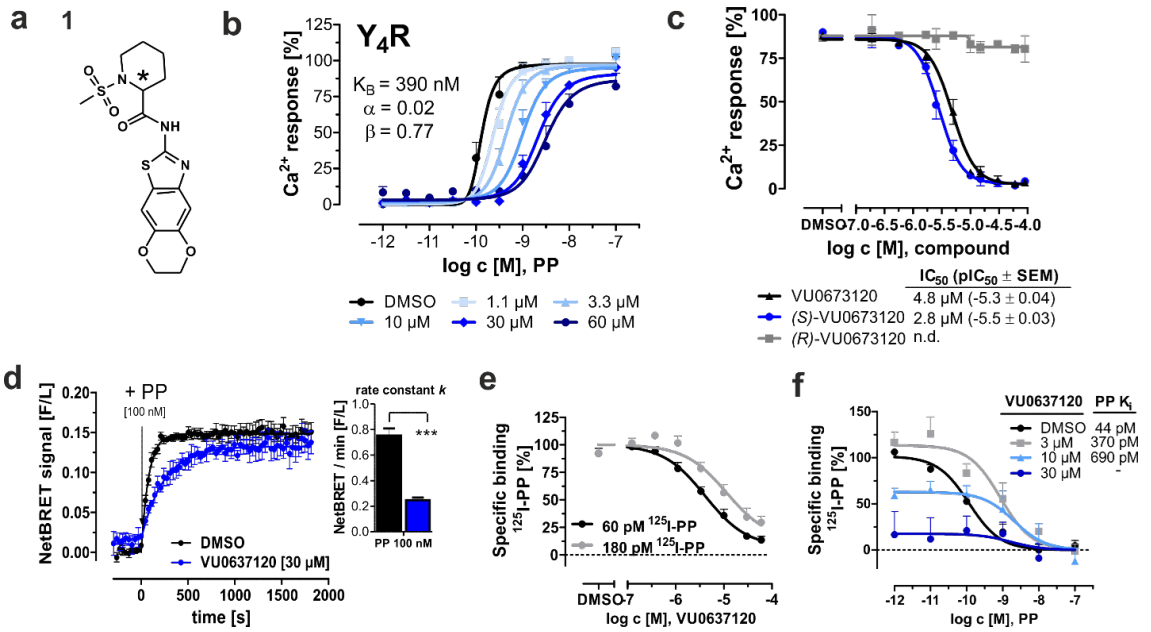


Figure 3.3: Characterization of VU0637120. a) Structure of VU0637120 (1), chiral C-atom is labeled by an asterisk. b) VU0637120 effectively decreases Y4R activation by PP in Ca2+ flux assays using stably transfected COS7_hY4R-eYFP_Δ6$G\alpha_{qi4myr}$ cells. Data represent mean ± SEM from N $\geq$ 2 independent experiments. VU0637120 effect on PP potency (α) and efficacy (β) was quantified using an operational model of allosterism (Leach et al., 2007). c) Effect of increasing concentrations of VU0637120 and its enantiopure isomers on Y4R activation in response to a submaximal PP concentration in COS7_hY4R-eYFP_Δ6$G\alpha_{qi4myr}$ cells. Data are shown as mean ± SEM from at least N $\geq$ 2 independent experiments. d) Investigation of arr3 recruitment to Y4R in response to 100 nM PP in presence of 30 μM of VU0637120. Kinetic bioluminescence resonance energy transfer (BRET) experiments were performed in HEK293 cells transiently expressing Y4R-Rluc8 and venus-arr3. Rate constant k of arr3 recruitment is summarized as bar graph. Data represent the mean ± SEM from N $\geq$ 4 independent experiments. Statistical analysis was performed using unpaired t-test, *** P $<$ 0.001. e) VU0637120 reduces the specific binding of 125I-PP to Y4R membrane preparations. Binding was investigated in competition binding assays with 60 or 180 pM 125I-PP. Unspecific binding was determined in the presence of 1 μM PP. Data represent the mean ± SEM from N $\geq$ 4 independent experiments. f) PP competition binding in presence of VU0637120 using 125I-PP (180 pM) at Y4R membrane preparations. Data are shown as mean ± SEM from N $\geq$ 4 independent experiments.

control, which can be confirmed by the significantly reduced rate constant k of arr3 recruitment to the Y4R in the presence of 30 μM of VU0637120. Furthermore, the effect of VU0637120 on the internalization of the Y4R was investigated (Figure 3.4.2). Data show that VU0637120 itself has no effect on the membrane localization of the Y4R as the amount of quantified cell surface receptors is comparable to the DMSO control. Stimulation of Y4R with either 10 nM or 100 nM PP in the presence of DMSO control induces internalization of the Y4R indicated by a reduced amount of Y4R cell surface receptors. In contrast, VU0637120 blocks Y4R internalization induced by 10 nM PP and only slightly decreases Y4R internalization by 100 nM PP compared to DMSO.

To further characterize the effect of VU0637120 on the endogenous ligand PP at the Y4R, radioactive

ligand binding studies were performed. Competition binding studies revealed that increasing concentrations of VU0637120 reduce PP binding at the Y4R (Figure 3.1e). Whereas high concentrations of VU0637120 reduce specific binding of 60 pM 125I-PP (1 x KD) down to about 15%, the competition curve observed with 180 pM failed to reach a bottom plateau, and even high concentrations of the compound were not able to fully displace the binding of the orthosteric radioligand. Further binding studies showed that the Ki of PP increases in the presence of VU0637120, and confirm that the compound modulates the activity of the endogenous ligand (Figure 3.1f). These experiments verify an allosteric mechanism of VU0637120, as the binding of a NAM to its receptor often induces a low-affinity state for the binding of the orthosteric ligand, which can result in reduced radioligand binding in competition binding studies (Burford et al., 2011). This finding is also in agreement with the effect of VU0637120 on the reduction of PP affinity to its Y4R orthosteric binding pocket ($\alpha = 0.02$), calculated by the operational model of allosterism (Leach et al., 2007). Interestingly, VU0637120 is structurally distinct from all previously described Y receptor antagonists and much smaller than the dimeric Y4R antagonists, also supporting the proposition of an allosteric mechanism of VU0637120 (Keller et al., 2013).

3.4.2 Ex vivo assays on mouse tissue confirm the selective allosteric effect of VU0637120

With respect to the different functions of Y receptors in the regulation of satiety and energy metabolism, a high selectivity of VU0637120 to the Y4R is important (Zhang et al., 2011, 2010). We first investigated the effect of VU0637120 on the Y1R, Y2R and Y5R in vitro. A high concentration of 30 μM of VU0637120 has no effect on the efficacy (EMax) and potency (EC50) of NPY-induced response at the Y1R, Y2R and Y5R, which demonstrates the selectivity of VU0637120 to the Y4R subtype (Figure 3.2a).

3.4.3 Rosetta docking guided by mutagenesis identified a partial allosteric binding pocket of (S)-VU0637120

Next, the activity of the compound was investigated in mucosal preparations of mouse descending colon, tissue shown previously to express native Y4R (Tough et al., 2006). Electrophysiological measurements of vectorial ion transport showed a decrease in the size of rat PP (rPP) responses after exposure of mucosae to increasing concentrations of (S)-VU0637120, exhibiting an IC50 value of 3.8 μM (pIC50 5.42 ± 0.27). Consistent with results from the Ca2+ flux assays, the (R)-VU0637120 enantiomer (30 μM) had no effect on rPP signaling (Figure 3.2b). To further study the selectivity of VU0637120 in mouse colon mucosa, PYY-induced activation of Y1R and Y2R, which are also expressed in these mucosal preparations, was examined. PYY responses were unaffected by either enantiomer (Figure 3.2b) showing selectivity for the Y4R, rather than Y1R or Y2R PYY-induced signaling in these native preparations.

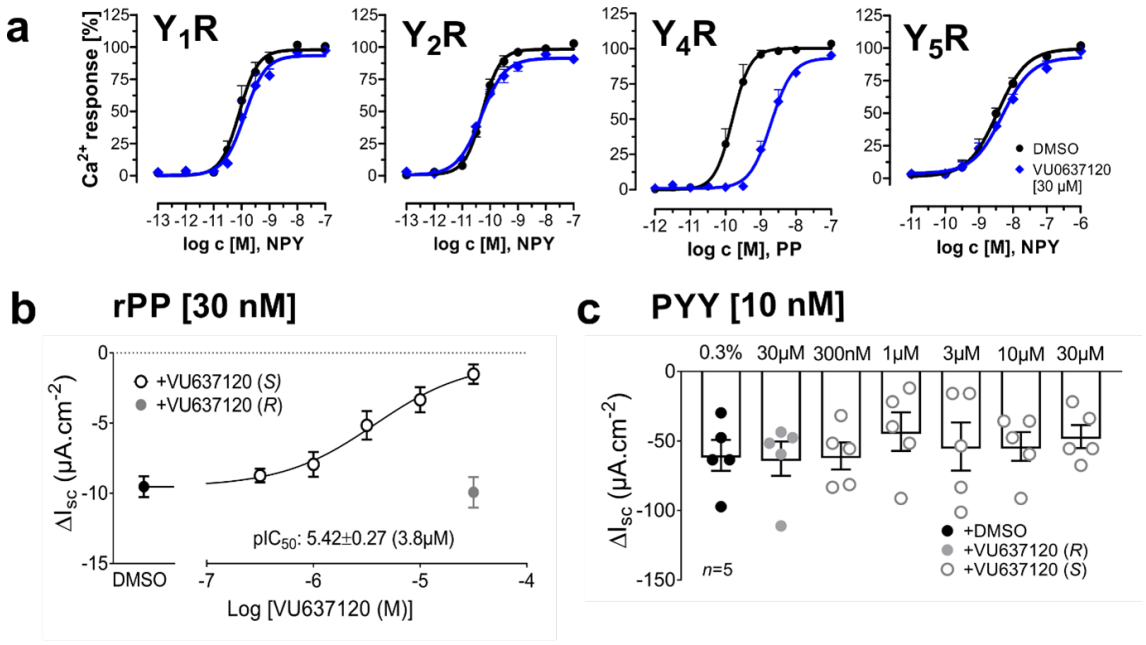


Figure 3.4: **VU0637120 is selective for the Y4R subtype in vitro and ex vivo.** a) Activation of Y1R, Y2R and Y5R by NPY and Y4R by PP in presence of 30 μM of VU0637120. Receptor signaling was measured in Ca2+ flux assays using COS7 cells stably expressing one specific Y receptor (Y1,2,4,5R-eYFP) and the chimeric G protein $\Delta 6G\alpha_{qi4myr}$. Data represent the mean ± SEM from N $\geq$ 2 independent experiments. b) (S)-VU637120 inhibits rPP responses in a concentration-dependent manner, while 30 μM of (R)-VU637120 have no effect on the Y4R agonist in mucosal preparations of mouse descending colon. c) Subsequent PYY signals are resistant to either enantiomer of VU0637120 at the different concentrations shown, compared to the DMSO control.

To characterize the interaction between VU0637120 and the Y4R and to gain insight into the mechanism of allosteric compounds, we first tested the activity of VU0637120 at eight Y4R/Y1R chimeras (Figure 3.3a) to elucidate Y4R domains, which are important for the interaction with the compound. The activity of VU0637120 at all Y4R/Y1R chimeras is displayed as the reduction of PP response in the presence of 15 μM antagonist, compared to the DMSO control (bar graphs, Figure 3.3b). The results suggest an important role of TM1, TM2 and ECL2 for the activity of VU0637120, as the exchange of these Y4R regions by the corresponding sequences of the Y1R drastically reduced the antagonistic activity (Figure 3.3b).

Based on these data, we performed an initial screen of 77 variants of Y4R with segmental and single point mutations (Figure 3.4). For the initial mutant screen, the racemic mixture of VU0637120 was used, as it is commercially available and thus, less expensive compared to the enantiopure compounds. Based on these results, the systematic investigation of the Y4R extracellular domain identified eleven positions, including Y1.39, Q2.58, T2.61, Y2.64, W2.70, S3.28, A3.29, Q3.32, F4.60, F7.35 and L7.36 to be critical for the antagonistic activity of VU0637120 (Figure 3.4). To consider the effects of a mutation on the functionality of the receptor, all constructs were characterized for membrane localization and PP activation (Supplementary

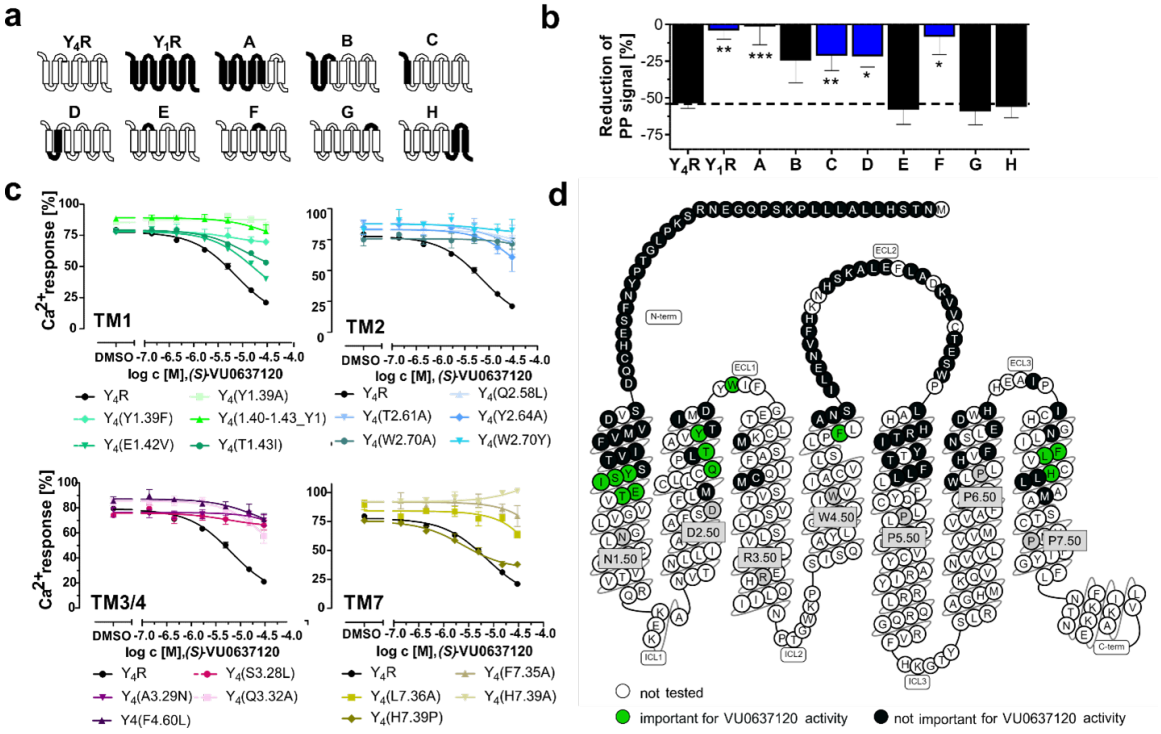


Figure 3.5: **Investigation of the VU0637120 binding site at the Y4R.** a) Schemes of Y4R/Y1R chimeras, whereby Y4R segments are shown in white and Y1R segments are colored in black. b) Loss of submaximal PP response (PP EC80) in presence of 15 µM VU0637120 at Y4R, Y1R and chimeras. Receptor activation was investigated in COS7 cells transiently expressing one specific receptor-eYFP construct and the chimeric G protein $\Delta 64\alpha_{qi4myr}$. Data are shown as mean ± SEM of N $\geq$ 3 independent experiments. Statistical analysis was performed using one-way ANOVA and Dunnett´s posttest: * P $<$ 0.05, ** P $<$ 0.01, *** P $<$ 0.001. c) Effect of increasing concentrations of (S)-VU0637120 at important Y4R residues in TM1, TM2, TM3/4 and TM7 in response to a submaximal PP concentration (PP EC80) studied in Ca2+ flux assays in transiently transfected COS7 cells expressing a definite Y4R mutant and the chimeric G protein $\Delta 64\alpha_{qi4myr}$. Each data point is shown as mean ± SEM of N $\geq$ 3 independent experiments. Residues are numbered according to the nomenclature of Ballesteros Weinstein (Ballesteros and Weinstein, 1995). d) Y4R snake plot, residues important for (S)-VU0637120 activity are highlighted in green, residues not important for VU0637120 are marked in black. Adapted from GPCRdb.org (Isberg et al., 2017).

Table 3.S1).

In order to more thoroughly characterize the binding pocket of the active (S)-VU0637120 enantiomer at the Y4R, positions that have been identified in the initial screen to be important for VU0637120, were tested with (S)-VU0637120. The antagonistic activity of (S)-VU0637120 was determined by the measurement of submaximal receptor activation (PP EC80) in the presence of increasing (S)-VU0637120 concentrations. We confirmed the importance of position Y1.39, Q2.58, T2.61, Y2.64, W2.70, S3.28, A3.29, Q3.32, F4.60, F7.35 and L7.36 for (S)-VU0637120 as mutations at these positions lead to a dramatic loss of antagonistic activity of (S)-VU0637120 (Figure 3.3c). A summary of the results of the mutagenesis data highlighting important positions for VU0637120 is given in Figure 3.3d.

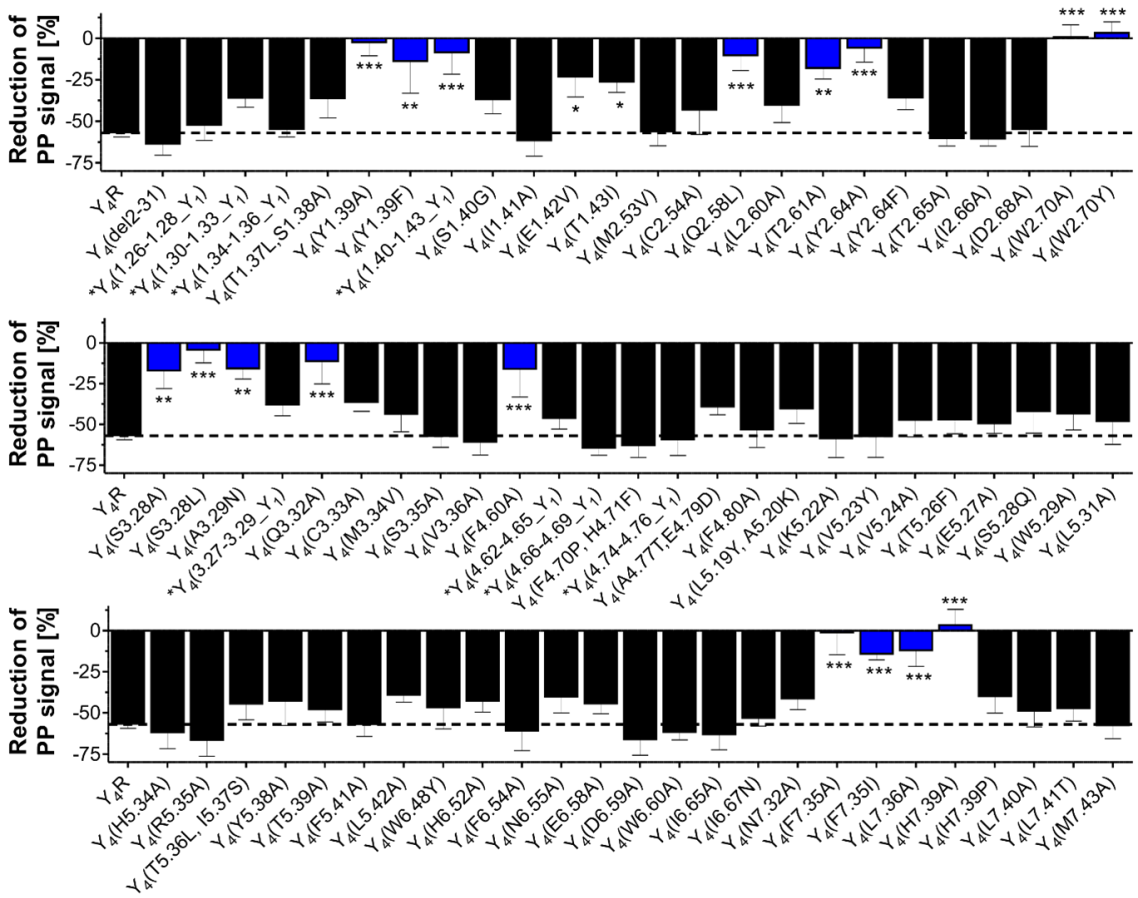


Figure 3.6: **Y4R mutant screen for the investigation of the VU0637120 binding pocket.** Inhibition of a submaximal PP activation (PP EC80) of Y4R mutants by VU0637120. Receptor activation was measured in Ca2+ flux assay in presence of DMSO or 15 µM VU0637120 using COS7 cells transiently expressing one specific receptor mutant and the chimeric G protein $\Delta G\alpha_{qi4myr}$. Data are shown as mean ± SEM of N $\geq$ 3 independent experiments. Statistical analysis was performed using one-way ANOVA and Dunnett´s posttest: * P $<$ 0.05, ** P $<$ 0.01, *** P $<$ 0.001. Residues are numbered according to the nomenclature of Ballesteros-Weinstein (Ballesteros and Weinstein, 1995).

To further delineate the structural determinants of activity, (S)-VU0637120 was docked using RosettaLigand (Combs et al., 2013; Yang et al., 2018; Burford et al., 2011) into a Y4R homology model constructed with RosettaCM (Song et al., 2013a). A similar approach was performed previously to investigate the binding mode of PP (Pedragosa-Badia et al., 2014). The resulting docking models were prioritized based on Rosetta predicted binding energy (Table 3.2) and the proximity of (S)-VU0637120 to the residues that were shown to be important in mutagenesis experiments (Figure 3.4). Initial placement of (S)-VU0637120 inside Y4R was guided by significant loss in antagonistic activity of the compound to the mutants Y4(Y1.39A) and Y4(Y1.39F) (Figure 3.3c, Figure 3.4). The best-scoring binding pose (Figure 3.6; Figure 3.8, cluster 1) places the ligand parallel to the membrane plane deep inside the transmembrane region. This binding mode

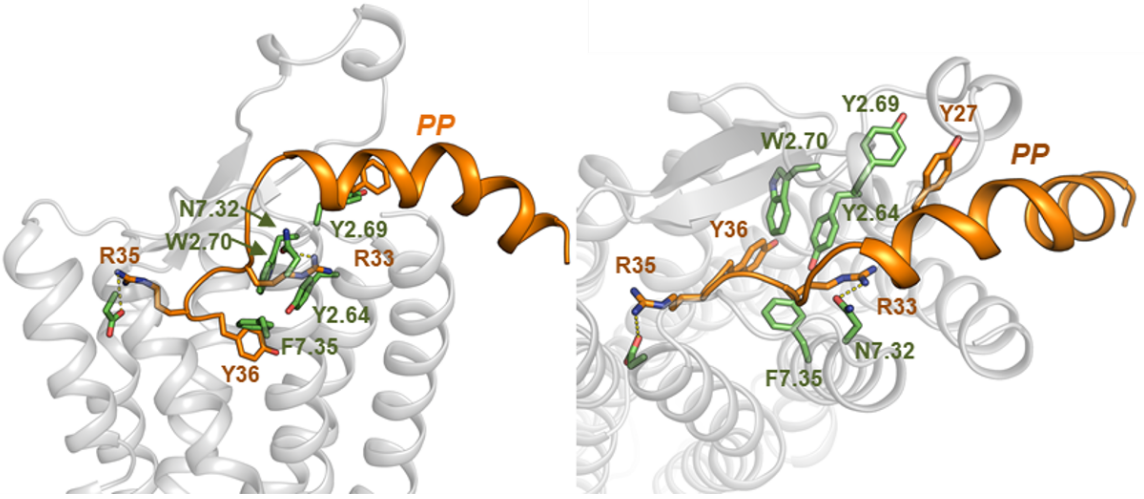


Figure 3.7: **Docking models of PP to Y4R.** Y4R residues that were determined to be important to the PP-Y4R activation are colored in green.

is supported by the sensitivity of VU0637120 to mutation of F4.60 and A3.29. The loss of VU0637120 function by the introduction of the bigger, hydrophilic asparagine at position A3.29 confirms that the space on top of TM3, below the ECL2 is essential for VU0637120 activity (Figure 3.3, Figure 3.4). A second round of mutagenesis was conducted to confidently exclude two alternative binding poses (Figure 3.6, cluster2/3; Figure 3.7). The data reveal that (S)-VU0637120 docking poses of cluster 2 and cluster 3 are unlikely, as (S) VU0637120 activity is insensitive to mutations at positions predicted to be important for the compound in cluster 2 (D1.27, T2.65, D2.68, N7.32 and M7.43) or cluster 3 (M2.53, S3.35, V3.36, V5.24, W6.48, N7.32 and M7.43).

3.4.4 Rosetta docking guided by mutagenesis identified a partial allosteric binding pocket of (S)-VU0637120

The docking model of cluster 1 is consistent with the mutagenesis data. It shows that (S) VU0637120 forms polar contacts with Y1.39, Q3.32 and H7.39, CH-π interactions with F4.60 and F7.35, and non-polar contacts to Y2.64 and L7.36 (Figure 3.8). Thereby, residues Y2.64, W2.70, F7.35 and L7.36 form the top surface area of the binding site. The distances among the hydroxyl group of Y2.64, the amine group in the indole ring of W2.70 and the oxygen in the amide group of (S)-VU0637120 are close to the range of significant polar contacts (Figure 3.8b). It is possible that those polar interactions exist, changing the orientation of the sidechain of W2.70 and Y2.64, and destabilize the PP binding mode.

The model further predicts the role of TM1, TM2 and the ECL2 for Y4R selectivity of VU0637120, which was observed by the investigation of different Y4R/Y1R chimeras (Figure 3.3a). Mutation of specific

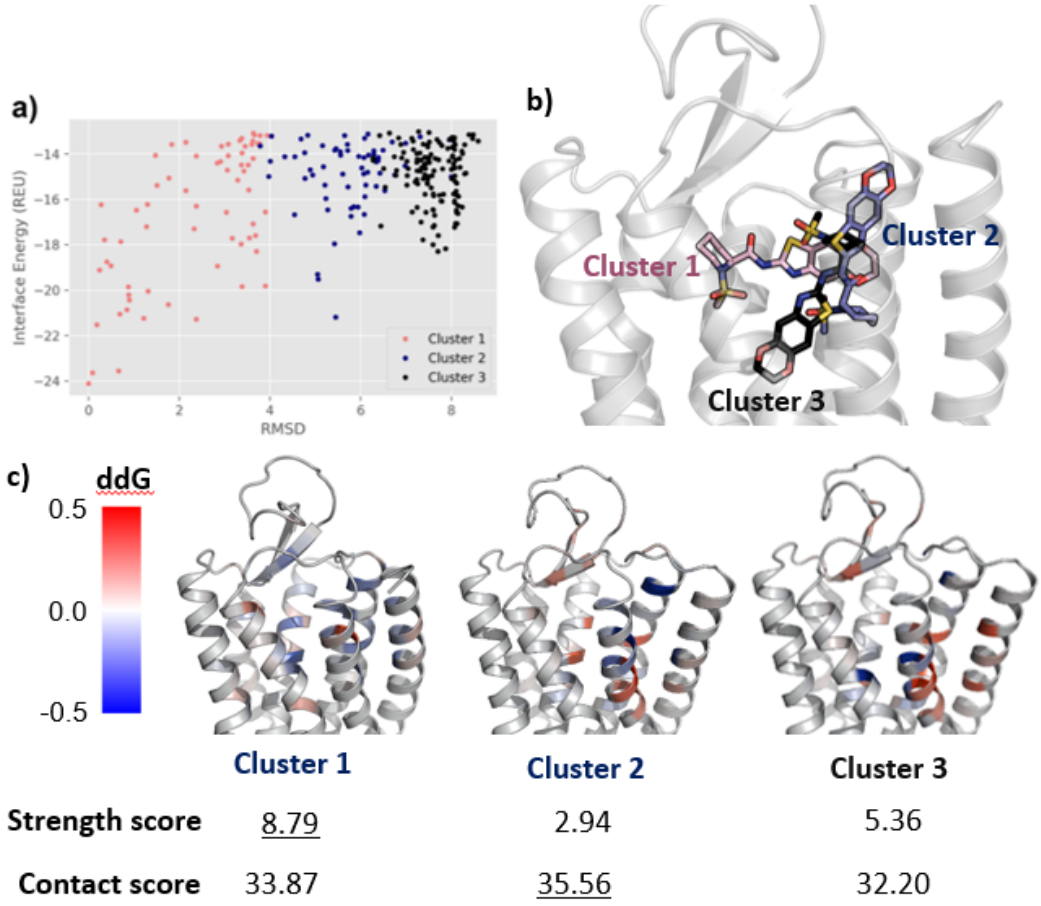


Figure 3.8: **RosettaLigand docking outputs clusters into three different poses.** a) Interface energy and RSMD plot of (S)-VU0637120-Y4R docking models. Scatter plot between the interface delta (interface energy) score on Y-axis and RMSD to the model with the best score on the X-axis. b) All three binding poses locate (S)-VU0637120 deep inside the transmembrane region. c) Calculated per residue ddq are mapped on the binding pocket of each docking pose. The interaction strength or contact scores are calculated based on per residue ddGs and ligand contacts, respectively.

fragments and positions identified region 1.40-1.43 and Q2.58 to be highly important (Figure 3.3c, Figure 3.4). In the model, these positions are in close proximity and form polar contacts between TM1 and TM2. In addition, position H7.39 has polar interactions with Y1.39 (Figure 3.8c, d). Thus, these residues might be important for the TM1-TM2 and TM1-TM7 orientation to shape the binding pocket of (S)-VU0637120. The role of the ECL2 was investigated by the mutation of almost the entire loop sequence (Figure 3.4). However, none of these mutations had a strong effect on VU0637120 activity. Thus, the loss of antagonistic activity at the Y4R chimera containing the ECL2 of the Y1R (chimera H, Figure 3.3a, b) is likely due to conformational effects, which is in agreement with the high sequence variation of the loops between Y4R and

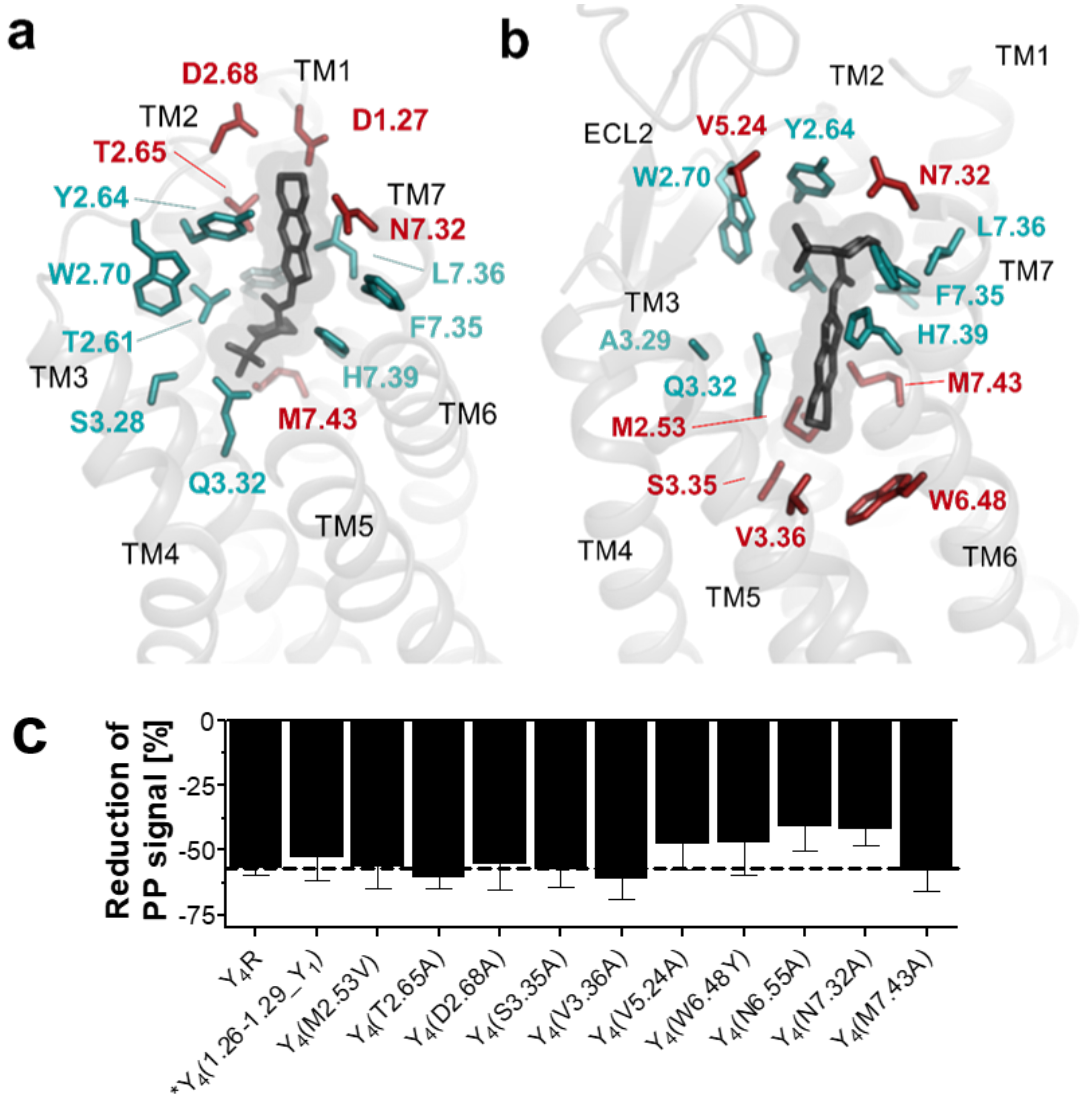


Figure 3.9: **Exclusion of (S)-VU0637120-Y4R docking poses of cluster 2 and cluster 3.** (S)-VU0637120 docking pose at the Y4R of a) cluster 2 and b) cluster 3. Residues that are important for the activity of VU0637120 and are involved in the binding of VU0637120 in all docking clusters are shown in cyan. Positions shown in red are predicted to be only important for a specific binding pose and thus enable the discrimination between the different binding modes. c) Inhibition of a submaximal PP activation (PP EC80) by VU0637120 at Y4R mutants that are predicted to be specifically important for the binding poses 2 and 3 (red residues in a and b). Receptor activation was measured in Ca2+ flux assays in presence of DMSO or 15 µM VU0637120 in COS7 cells transiently expressing one specific receptor mutant and the chimeric G protein $\Delta G\alpha_{qi4myr}$. Data are shown as mean $\mp$ SEM of N $\geq$ 3 independent experiments. Residues are numbered according to the nomenclature of Ballesteros-Weinstein (Ballesteros and Weinstein, 1995). The mutation of residues that are predicted to be important for the binding poses 2 and 3 had no significant effect on the activity of VU0637120, and thus cluster 2 and cluster 3 can be excluded as possible binding poses.

Y1R (Supplementary Figure 3.S2). The docking model locates (S)-VU0637120 directly below the conserved disulfide-bridge and suggests a structural role of the ECL2 in the formation of the binding pocket (Figure 3.8a).

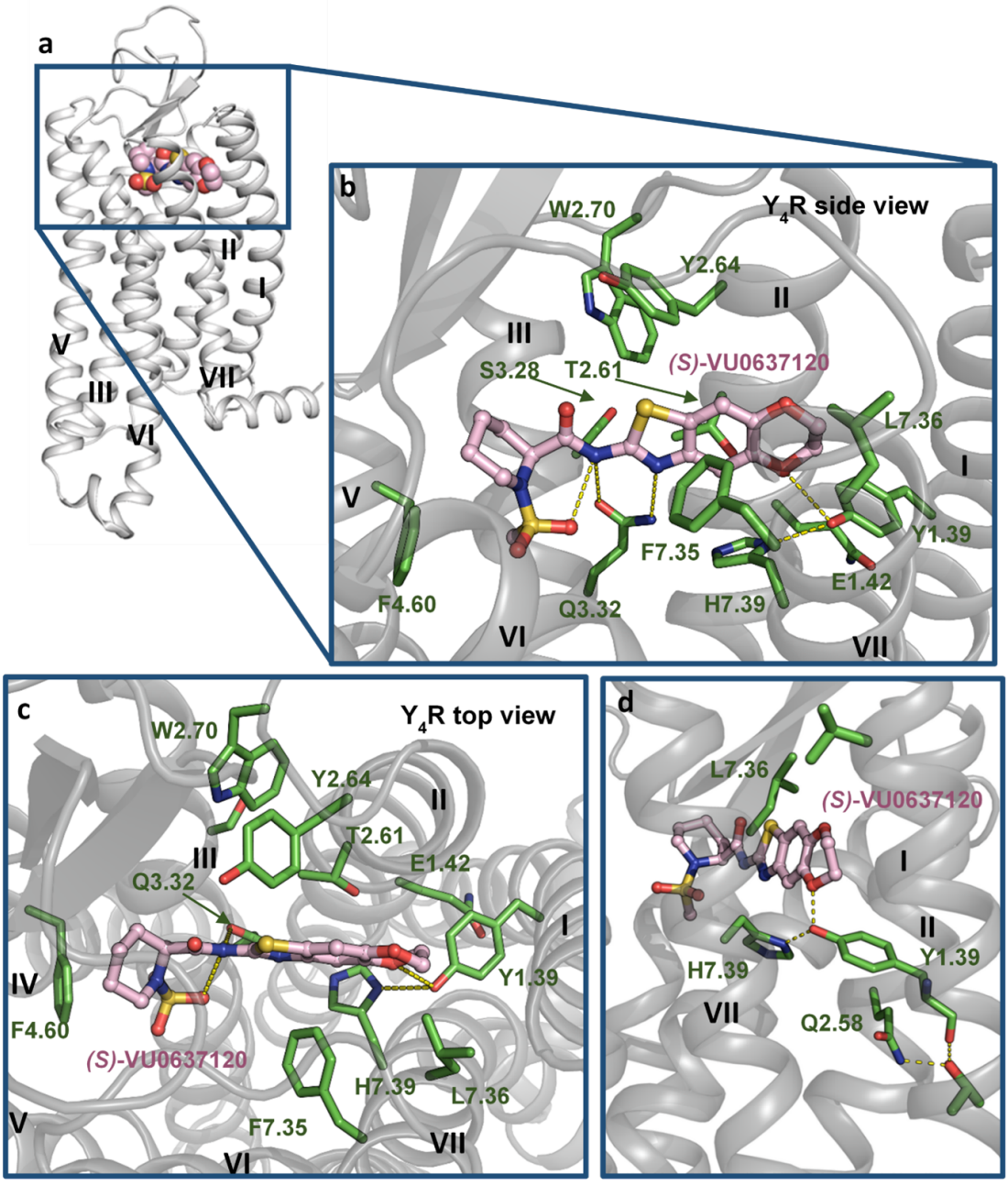


Figure 3.10: **Predicted binding mode of (S)-VU0637120.** a) Computational docking of (S)-VU0637120 identified an allosteric binding site located bellow the extracellular interface. b-c) Contacts of (S)-VU0637120 with most important residues (side view and top view). d) Interaction network among TM1, TM2, TM7 and (S)-VU0637120.

3.4.5 Identification of Y4R gain-of-function mutants for (S)-VU0637120

Based on the docking of (S)-VU0637120 to the Y4R, we used RosettaDesign to predict Y4R mutants that are favorable for the activity of (S)-VU0637120. The top mutants predicted by RosettaDesign were inspected visually. The E6.58R mutant was selected because the mutated residue is predicted to form an additional hydrogen-bond interaction to the methanesulfonamide group of (S)-VU0637120. Additionally, A3.29V and A3.29P mutants were chosen, as both valine and proline could form better hydrophobic interactions to the six-member ring of (S) VU0637120. Those predicted favorable interactions were confirmed in the docking model of (S) VU0637120 to single Y4R mutants, which were reconstructed with RosettaCM (Song et al., 2013a) and RosettaLigand (Combs et al., 2013; Yang et al., 2018; Burford et al., 2011) (Figure 3.9c).

To check whether we can verify the RosettaDesign prediction in in vitro assays, we created the Y4R gain-of-function mutants by mutagenesis and tested the effect of (S)-VU0637120 at these mutants. Here, we demonstrate that the antagonistic effect of (S)-VU0637120 on PP response is significantly increased at the mutants Y4(E6.58R), Y4(A3.29P) and Y4(A3.29V) compared to the DMSO control (Figure 3.9a, b). These experiments confirm the accuracy of the (S)-VU0637120 binding mode to the Y4R as we were able to predict and confirm favorable interactions between distinct Y4R residues and functional groups of (S)-VU0637120.

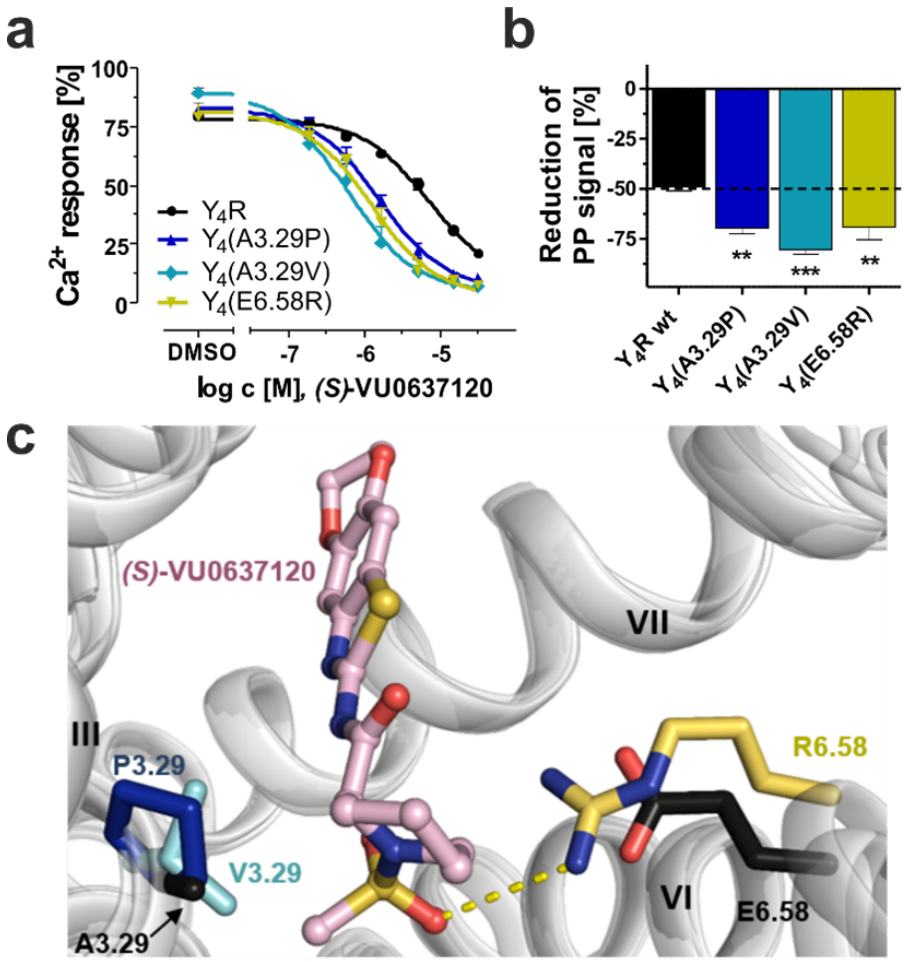


Figure 3.11: **Gain-of-function of (S)-VU0637120 at specific Y4R mutants.** a) Concentration-response curves of increasing concentrations of (S)-VU0637120 in response to a submaximal PP concentration (PP EC80) at specifically identified Y4R residues. Ca2+ flux assays were performed in COS7 cells transiently expressing one specific Y4R-eYFP mutant and the chimeric G protein $\Delta G\alpha_{qi4myr}$. Data are shown as mean ± SEM of at least N $\geq$ 3 independent experiments. b) Effect of 15 µM of (S)-VU0637120 on a submaximal PP concentration (PP EC80) at the Y4R gain-of-function mutants, Data from Fig 5a. Comparison of (S) VU0637120 activity at Y4R and the mutants was performed using one-way ANOVA and Dunnett´s posttest: ** $P < 0.01$, *** $P < 0.001$. c) Point mutant models made by Rosetta illustrate the improvement in hydrophobic interactions between (S)-VU0637120 and P3.29 or V3.29, and the additional hydrogen bond between (S)-VU0637120 and R6.58. Residues are numbered according to the nomenclature of Ballesteros-Weinstein (Ballesteros and Weinstein, 1995).

3.5 Discussion

In this study, we highlight the characterization of the first-in-class allosteric antagonist (S)-VU0637120 that exhibits clear Y4R selectivity in engineered cell lines and in proven native Y4R-expressing mouse colon mucosa. For the enantio-selective (S)-VU0637120 compound, we identified an allosteric binding pocket in the core of the Y4R transmembrane domains below the endogenous ligand binding site, that ultimately strengthen our understanding of allosteric modulation of the Y4R.

To map the binding pocket of the allosteric antagonist VU0637120 in comparison to the orthosteric agonist PP, mutagenesis experiments were performed for both ligands (Supplementary Table 3.S1, Figure 3.4). The models of PP docked to Y4R were reconstructed as we remade Y4R homology models with a more realistic ECL2 loop and an updated list of templates, which include the Y1R crystal structure, discovered in 2018 (Yang et al., 2018) (Table 3.1).

Previous studies and our most recent mutagenesis data revealed that positions D2.68, W2.70, N6.55, D6.59, N7.32, F7.35 and M7.43 are important for Y4R activation by PP and demonstrated that PP binds at the Y4R in a flat binding mode (Merten et al., 2007; Pedragosa-Badia et al., 2014) (Figure 3.10, Figure 3.5). Additionally, positions T2.61, W2.70 and F7.35 were identified to be critical for both, the activity of the orthosteric agonist PP and the allosteric antagonist VU0637120. In contrast, most of the positions that influence VU0637120 activity, including Y1.39, Q2.58, Y2.64, S3.28, A3.29, F4.60 and L7.36, are located in a deeper pocket and fail to affect PP activation (Figure 3.10). These results indicate that VU0637120 binds in a secondary, allosteric binding pocket at the Y4R with a slight overlap with the binding site of the endogenous ligand PP. Whereas the Y4R-PP docking model indicates a more open conformation of the Y4R in complex with PP, the binding of VU0637120 might stabilize a more closed receptor conformation. Thus, it follows that VU0637120 binding might create a low-affinity binding state for the endogenous ligand PP with reduced agonist binding.

Interestingly, the binding pocket of (S)-VU0637120 at the Y4R overlaps with that of orthosteric antagonists and NAMs of other peptide GPCRs (Figure 3.11). The crystal structure of the related Y1R in complex with orthosteric antagonists revealed that the binding pocket of the antagonist BMS-193885 extends from the peptide binding site to the center of the transmembrane domain among TM3, TM4, TM5 and TM6, and partially overlaps with the binding site of VU0637120 at its six-member ring and methanesulfonamide group (Yang et al., 2018). In contrast, the binding site of (S)-VU0637120 stretches horizontally towards TM1, TM2, and TM7. This region is known as an allosteric pocket in other rhodopsin-like GPCRs that bind to endogenous peptide ligands such as C-C chemokine receptor type 5 (CCR5) (Tan et al., 2013) and protease activated receptor 2 (PAR2) (Cheng et al., 2017). The interaction with these receptor regions might explain the allosteric character of (S)-VU0637120, which was confirmed for the modulation of PP affinity in radioligand binding studies (Figure 3.1), as well as in the Schild plot analysis of Y4R activation by PP (Figure 3.4.1). Furthermore, positions in TM1 and TM2 were shown to be critical for Y4R selectivity of VU0637120. These findings highlight the potential to develop highly selective Y4R ligands by targeting this binding pocket.

The discovery of (S)-VU0637120 defines a novel class of Y4R antagonists that is smaller and more selective than the most promising compounds described to date. A small molecule compound with weak antagonistic activity at the Y4R is UR-AK49, but this compound is not Y4R selective and can also bind to

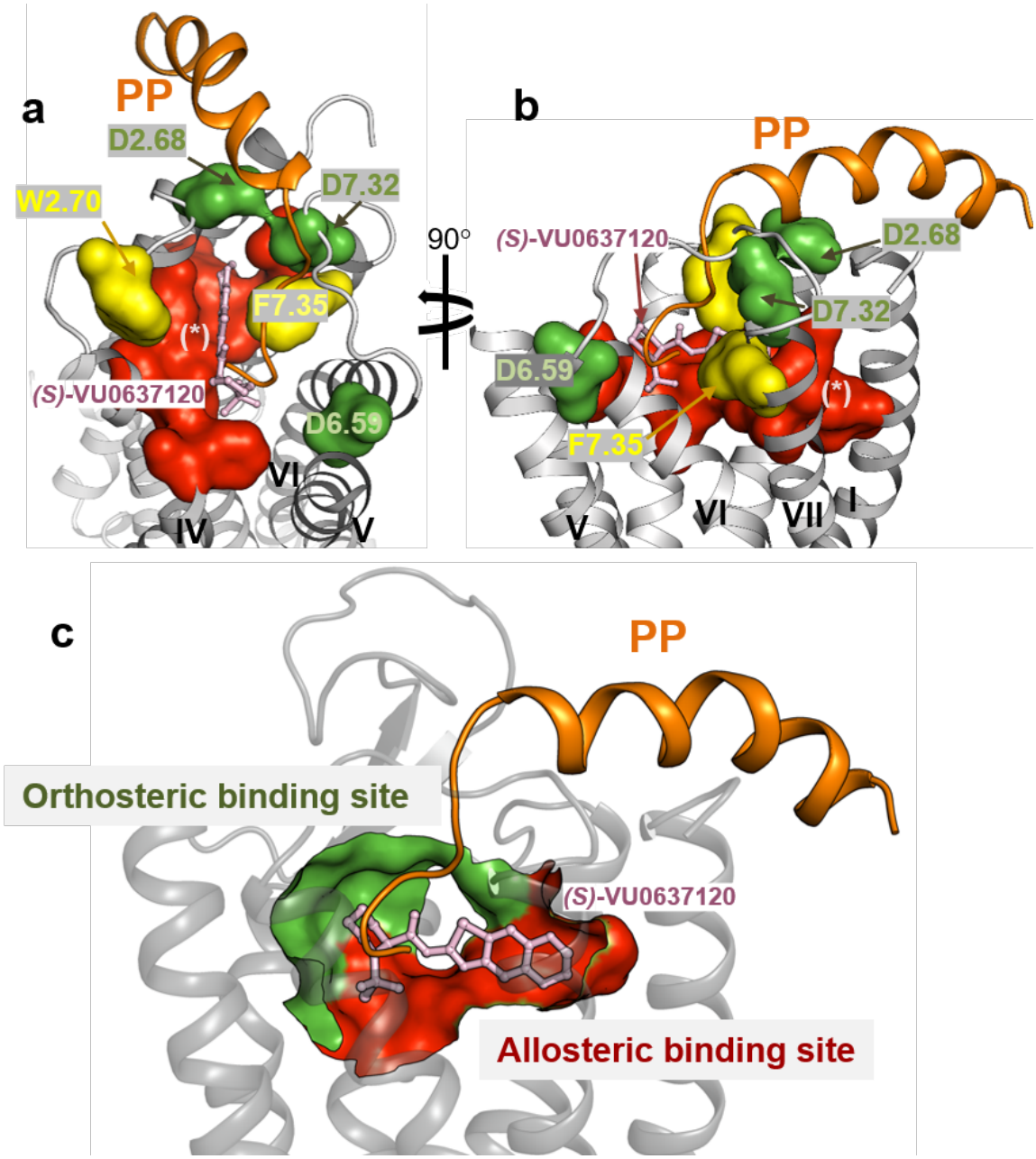


Figure 3.12: **Comparing the binding pocket of PP and (S)-VU0637120.** a-b) Superimposition of docking models of PP and(S)-VU0637120 to Y4R. Residues that are important to the activity of PP, (S)-VU0637120, and both are colored in green, red and yellow, respectively. Group of residues that are represented in red surface: (*): Y1.39, E1.42, Q2.58, T2.61, S3.28, A3.29, Q3.32, F4.60, L7.36, H7.39. c)(S)-VU0637120 binds in a secondary, allosteric binding pocket (red) at the Y4R, but this site might slightly overlap with theorthostericbinding site of PP (green).

Y1R and Y5R (Ziemek et al., 2007). Previous Y4R antagonists were generated by modifying Y1R antagonists (BIBP3226 (Rudolf et al., 1994), BIBO3304 (Wieland et al., 1998)). But the most active compounds developed by this approach still display substantial activity at Y1R (UR-MEK381) or Y2R (UR MEK388), respectively (Keller et al., 2013).

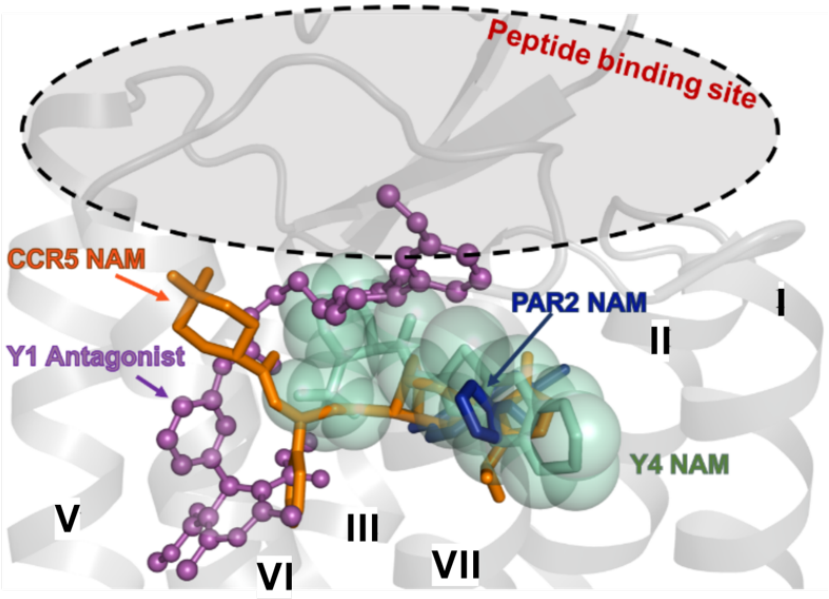


Figure 3.13: **Docking models of PP to Y4R.** Y4R residues that were determined to be important to the PP-Y4R activation are colored in green.

In summary, allosteric modulators provide a promising new class of ligands to target peptide GPCRs. As such, molecules bind to allosteric sites, which are often less conserved compared to orthosteric ligand binding sites of receptor subtypes, and a higher subtype selectivity can be achieved (Gao and Jacobson, 2013). Furthermore, allosteric molecules of peptide GPCRs have the advantage to modulate or fine-tune the receptor response by the endogenous peptide and consequently maintain the normal physiological regulation (Wootten et al., 2013). This study will support the analysis of this relevant class of modulators and facilitate drug development for peptide GPCRs in general.

3.6 Author Contributions

Corinna Schüß planned and performed in vitro experiments (Ca2+ flux assays, cloning of Y4R mutants, fluorescence microscopy studies) analyzed the experimental data, and wrote the manuscript.

Oanh Vu planned and performed docking studies of VU0637120 in a Y4R homology model, predicted specific mutagenesis experiments and wrote the manuscript.

Dr. Mario Schubert planned and performed in vitro experiments (Ca2+ flux, BRET assays, fluorescence microscopy, radioligand binding experiments, cloning of Y4R mutants), analyzed the experimental data, and

wrote the manuscript.

Dr. Yu Du performed HTS of 220.000 compounds, identified VU0637120 as a selective Y4R antagonist and validated the specificity of VU0637120 in screenings of different cell lines and receptors.

Dr. Nigam M. Mishra synthesized, purified, and characterized (R)-VU0637120 and (S)-VU0637120 and wrote and reviewed the synthetic chemistry section of the SI. Dr. Iain Tough performed electrophysiological measurements at mucosal preparations of mouse descending colon.

Dr. Jan Stichel cloned Y4R/Y1R chimera that were used in this study and reviewed the manuscript.

Prof. Dr. C. David Weaver planned and performed the HTS to identify Y4R modulators, analyzed and discussed experimental data, and reviewed and edited the manuscript.

Prof. Dr. Kyle Emmitte designed and supervised the stereoselective synthesis of (R)-VU0637120 and (S)-VU0637120, reviewed and edited the manuscript, and wrote and reviewed the synthetic chemistry section.

Prof. Dr. Helen Cox planned and analyzed ex vivo studies in mucosal preparations of mouse descending colon, discussed experimental data, and reviewed and edited the manuscript.

Prof. Dr. Jens Meiler planned and performed docking studies of VU0637120 in a Y4R homology model, contributed in data discussion, planning of experiments and data analysis, and wrote, reviewed and edited the manuscript.

Prof. Dr. Annette G. Beck-Sickinger planned and analyzed experiments to investigate the activity and binding site of VU0637120 (Ca2+ flux studies, BRET assays, fluorescence microscopy, cloning of Y4R mutants), and wrote, reviewed and edited the manuscript.

3.7 Acknowledgements

The authors thank K. Löbner, J. Schwesinger and C. Dammann for technical assistance. We kindly acknowledge the Vanderbilt HTS facility for the help in the identification of the Y4R modulators. Dr. A. Rodriguez, B. J. Bender and N. Kühn are acknowledged for discussions about data analysis and work in docking studies. This work was funded by the National Institutes of Health (Grant R01 DK0973376 to J.M. and C.D.W. and grant R01 DA046138 to J.M.), the DFG (Grant BE1264-16, SFB 1052/A3 to A.G.B.-S.), the European Union and the Free State of Saxony (Grants 100148835 and 143213128452 to A.G.B.-S.) and, the BBSRC (BB/N006763/1 to H.M.C).

3.8 Supplementary Material

Receptor	Chimera Sequence[a]	pEC_{50} ± SEM	EC_{50} [nM]	x-fold over Y_4R[b]	E_{max} ± SEM[c]
Y_4R	Y_4R wildtype	9.29 ± 0.05	0.5	1	100 ± 1
Y_1R	Y_1R wildtype	9.58 ± 0.10	0.3	1	96 ± 2
A	Y_4(1-241_Y_1)	8.65 ± 0.21	2.2	4	267 ± 18
B	Y_4(1-114_Y_1)	8.96 ± 0.10	1.1	2	208 ± 7
C	Y_4(1-65_Y_1)	8.51 ± 0.11	3.1	6	87 ± 4
D	Y_4(66-99_Y_1)	8.78 ± 0.13	1.6	3	118 ± 6
E	Y_4(99-114_Y_1)	9.16 ± 0.14	0.7	1	129 ± 6
F	Y_4(176-213_Y_1)	9.49 ± 0.21	0.3	1	115 ± 9
G	Y_4(291-300_Y_1)	9.10 ± 0.14	0.8	2	114 ± 5
H	Y_4(242-C term_Y_1)	6.85 ± 0.27	139.9	270	64 ± 11

[a] Amino acid are numbered according to the Y_4R sequence, exchanges were made based on sequence alignment Supplementary Figure 8
[b] EC_{50} ratio of mutant/Y_4R wildtype
[c] All data were normalized to PP maximum response at the wildtype Y_4R.

Table 4.S2: **Activation of Y4R, Y1R and receptor chimera.** Activation of the Y4R, Y1R and chimera was measured in a Ca2+ flux assay in transiently transfected COS7 cells expressing one receptor mutant and the chimeric G protein $\Delta 6G_{\alpha qi4myr}$. The Y4R and all chimera were stimulated with PP, the Y1R was activated with NPY. Data represent the mean $\mp$ SEM from N $\geq$ 3 independent experiments.

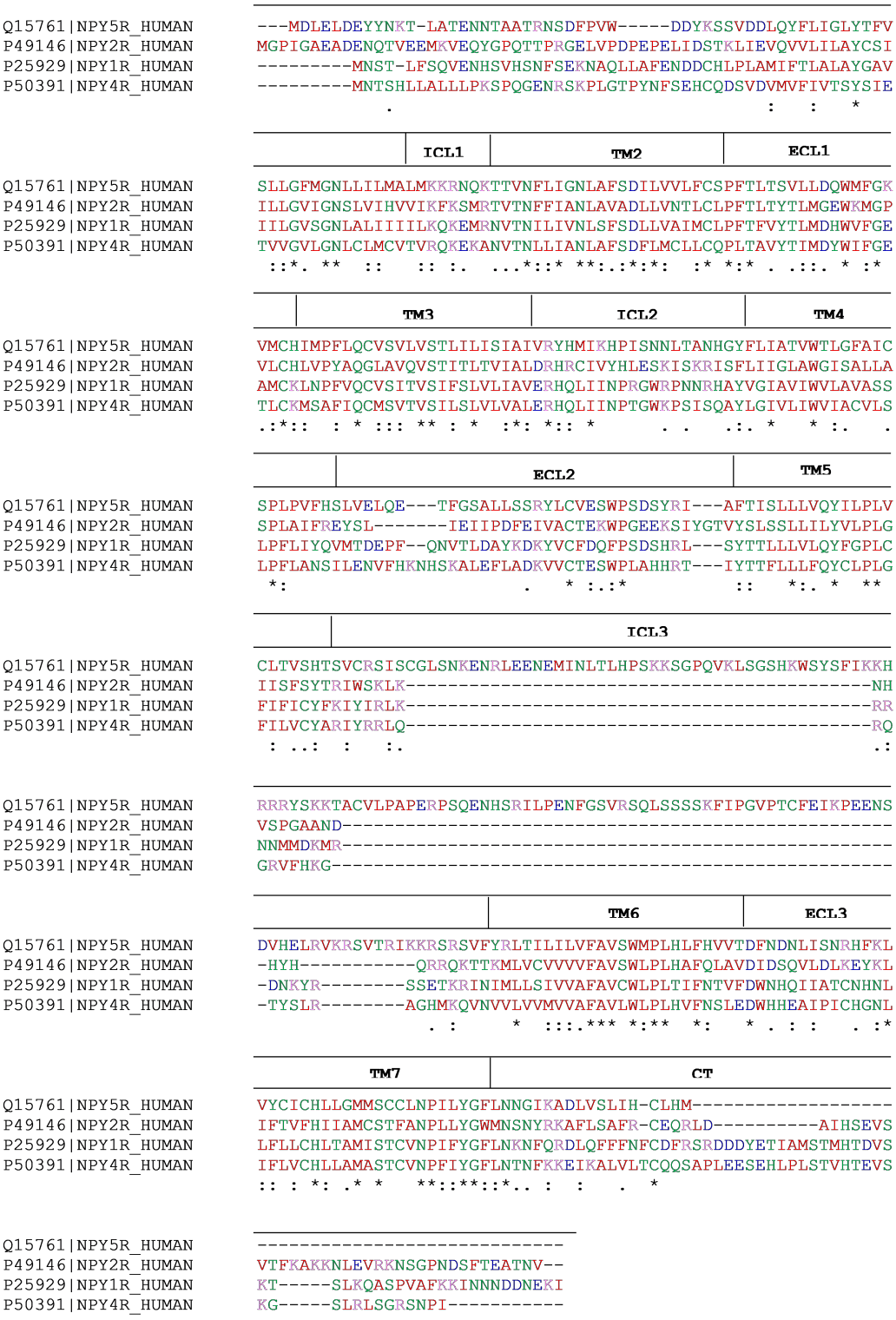


Figure 3.14: **Sequence alignment of human Y receptors.** Alignment was performed using CLUSTAL O(1.2.4) multiple sequence alignment tool 5. Termini, loops and TMs are marked based on the location in the Y4R homology model. Colors indicate the property of the amino acid (blue – hydrophobic, red – positive charge, magenta – negative charge, green – polar).

CHAPTER 4

BCL::Mol2D – a robust atom environment descriptor for QSAR modeling and pharmacophore mapping

Parts of this chapter was taken from the paper: Vu O, Mendenhall J, Altarawy D, Meiler J. BCL::Mol2D-a robust atom environment descriptor for QSAR modeling and lead optimization. J Comput Aided Mol Des. 2019 May;33(5):477-486.

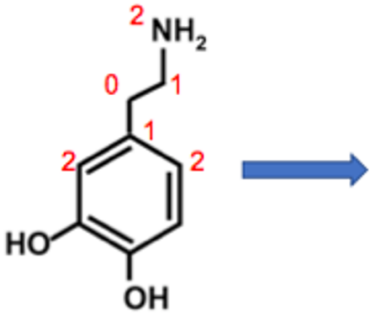
4.1 Abstract

Comparing fragment based molecular fingerprints of drug-like molecules is one of the most robust and frequently used approaches in computer-assisted drug discovery (CADD). Molprint2D, a popular atom environment (AE) descriptor, yielded the best enrichment of active compounds across a diverse set of targets in a recent large-scale study. We present here BCL::Mol2D descriptors that outperformed Molprint2D on nine PubChem datasets spanning a wide range of protein classes. Because BCL::Mol2D records the number of AEs from a universal AE library, a novel aspect of BCL::Mol2D over the Molprint2D is its reversibility. This property enables decomposition of predicted activities from machine learning models to particular molecular substructures. Artificial neural networks (ANNs) with dropout, when trained on BCL::Mol2D descriptors outperform those trained on Molprint2D descriptors by up to 26% in logAUC metric. When combined with the Reduced Short Range (RSR) descriptor set, our previously published set of descriptors optimized for QSARs, BCL::Mol2D yields a modest improvement. Finally, we demonstrate how the reversibility of BCL::Mol2D enables visualization of a 'pharmacophore map' that could guide lead optimization for serine/threonine kinase 33 (STK33) inhibitors.

4.2 Introduction

Ligand-based computer aided drug design (LB-CADD) relies on the observation that small molecule ligands often share a defined set of molecular features that promote molecular recognition of a ligand by a target protein – the so-called pharmacophore (Carlsson et al., 2009; Gates et al., 2008). While structurally unrelated chemotypes can represent the same pharmacophore, it is also correct that often molecules of similar structure share the pharmacophore required for targeting a protein. One advantage of LB-CADD methods is that the comparison of small molecule structures is independent of the knowledge of the three-dimensional structures of the target protein and its dynamics (Cramer, 2012). Two fundamental approaches of LB-CADD include similarity search and quantitative structure activity relationship (QSAR) models. While the for-

mer selects molecules that have similar structures to known actives, the latter infers a relationship between physicochemical properties of molecules and the bioactivity of interest and uses this relationship to select for molecules with high predicted bioactivity (Sliwoski et al., 2014). Due to its ability to rapidly screen libraries of compounds and significantly improve the discovery rate of actives, LB-CADD has become an increasingly popular in silico approach. A typical LB-CADD model is comprised of two major components: 1) a quantitative representation of chemical structures (descriptors) and 2) a similarity metric or, in the case of QSAR models, a mathematical function to compute bioactivity from these descriptors, often a machine-learning algorithm. While the former quantifies the similarity between input descriptors, the latter predicts bioactivity of compounds from the molecular descriptors (Sliwoski et al., 2014).



Layers	0	1	2
	C	=C	-N
		-C	=C
		-C	-C

Figure 4.1: **Illustration of an atom environment.** (Left) Molecule configuration with the heavy atoms being indexed based on its "layer". 0-center atom; 1, 2 -neighbor atoms that are 1 or 2 bonds away. (Right) Connectivity table of the atom environment

Molprint2D is a 2D similarity search method based on atom environments (AE), which encode atomic properties, such as element types and bond types, of surrounding atoms within two bonds distance from the atom of interest (height=2, Figure 4.1) (Bender et al., 2004). A largescale benchmark study of eight different 2D fingerprint methods has shown that Molprint2D fingerprint generated by the CANVAS software package yielded the best enrichments of active compounds on a diverse set of targets (Sastry et al., 2010). Each binary bit in a Molprint2D fingerprint only documents the presence or absence of a unique AE (Bender et al., 2004). In the current work, we test the hypothesis that in addition to presence also the number of AEs could be important to distinguish substructures with similar AE composition (e.g., six-member rings vs. five-member rings).

The Molprint2D defines different AEs based on the element type of atoms bound up to two bonds away from the central atom of interest (height=2). This description is highly overlapping as every atom will be represented in many AEs. We hypothesized that a more fine-grained list of AEs that includes hybridization state, i.e. electron configuration (Gasteiger and Marsili, 1980), in addition to element type but ventures only

one bond around the atom of interest (height=1) would provide a more information dense description of the AE. We set out to test this idea in the present work. Furthermore, Molprint2D generates AE set from the training dataset. We also hypothesized, that focusing on the most likely AEs in drug-like molecules can remove any bias from the training data. This AE library enables the model to be readily applicable to scaffold-hop into new chemical space and reduce the length of the descriptor vector. Thus, we generated a list of common AEs (i.e. the AE library) from a large database of over 900,000 drug-like compounds.

Artificial Neural Networks (ANN) are one of the most commonly used non-linear classifiers in QSAR models for LB-CADD due to their strong predictive power (Mendenhall and Meiler, 2016; Gregory et al., 2010). We have previously shown that ANN-QSAR models outperformed fingerprint-similarity searches on Molprint2D descriptors (Mendenhall and Meiler, 2016). However, their advantage in predictive capacity comes with the pitfall of their "black-box" nature (Cherkasov et al., 2014); it is difficult to map which structural features contribute to the activity. Previous efforts at interpreting QSAR models to aid molecular design used sensitivity analysis to rank importance of each descriptor on ANN training (Guha and Jurs, 2005; **?**; **?**). Yet the success of characterizing an ANNs' internal function also depends on the nature of the descriptors used in the QSAR studies (Cherkasov et al., 2014). We hypothesize that fingerprint descriptors are particularly well-suited to interpretation when used for training an ANN as they are reversible; each input number refers to one specific structural motive. It is one goal of the present study to test this hypothesis.

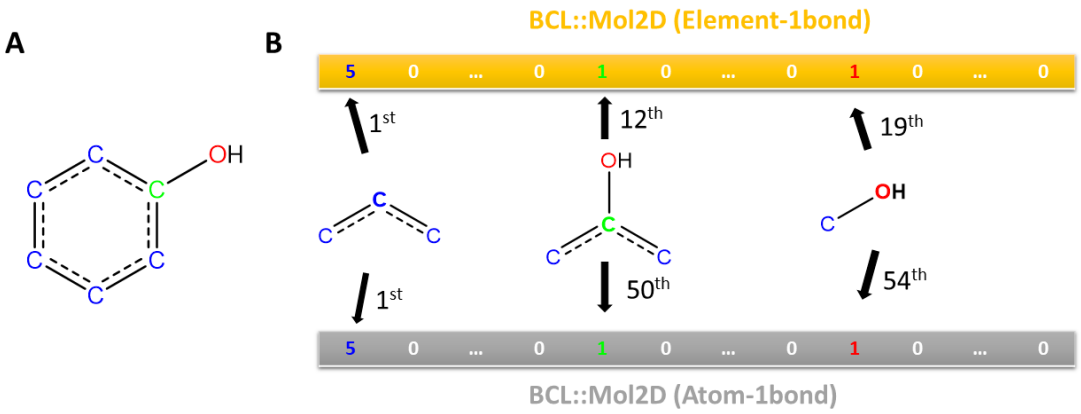


Figure 4.2: **Illustration of BCL::Mol2D fingerprint.** (Element type) of Taxol. The list of numbers on the right is the first 100 entries of Taxol's BCL::Mol2D fingerprint. Each descriptor stores the count of a unique atom environment. Each Substructures are marked with an unique colored circles

In this paper, we introduce BCL::Mol2D, which significantly outperforms Molprint2D in predictive capacity and ANN interpretability. BCL::Mol2D documents the counts of common AEs, in which atoms are

classified based on their element types and hybridization states (Figure 4.2). There are two atomic encoding schemes for BCL::Mol2D descriptors: the 'Element type' enciphers atoms based on their atomic numbers and bond orders, while the 'Atom type' further distinguishes between elements with different orbital configurations (Gasteiger and Marsili, 1980). ANNs, with drop-outs (Srivastava et al., 2014), trained on BCL::Mol2D descriptors perform significantly better than ANNs trained on Molprint2D descriptors. Moreover, we demonstrate the potential of BCL::Mol2D in the interpretation of ANN-QSAR models. The BCL::Mol2D descriptors are reversible from their numerical representation to the original chemical structures. Therefore, they allow extraction of the AEs that are crucial for optimization of predicted bioactivity of compound candidates. BCL::Mol2D descriptor method has been added to the BCL::Cheminfo package, which is free for non-commercial users. Finally, BCL::Mol2D are also combined with our previously-described best performing reduced short-range 3D descriptor set (BCL::3D-RSR) (Mendenhall and Meiler, 2016). The resulting hybrid descriptor set modestly, albeit consistently, improve the performance of QSAR models.

4.3 Methods

4.3.1 Data curation

A previously established QSAR benchmark of nine HTS datasets (table 4.1) was used to evaluate performance of ANNs. These datasets are comprised of compounds from HTS scans on eight protein targets: two class A G-protein coupled receptor (GPCRs), three ion channels, one transporter, one kinase, and one enzyme. As previously detailed (Butkiewicz et al., 2013, 2017; Sliwoski et al., 2016a), molecules were labeled as active compounds only if their activity was verified in follow up confirmatory assays, selectivity assays, and dose response experiments. Each data set contained more than 170 active and 61,000 inactive compounds. Three-dimensional conformations were generated with Corina version 3.60 (Gasteiger et al., 2003), with the driver options of adding hydrogens (wh) and removing molecules from which the software could not generate 3D structures from (r2d).

Target protein	PubChem AID	# active	# inactive	A/I ratio[a] (%)
Orexin 1 receptor antagonists	743306	233	217925	0.11
M1 muscarinic receptor agonists	652178	187	61646	0.30
M1 muscarinic receptor antagonists	1053187	362	61394	0.59
$Kir_{2.1}$ K+ channel inhibitors	743120	172	301321	0.06
KCNQ2 K+ channel potentiates	1159610	213	302192	0.07
Cav3 T-type Ca^{2+} inhibitors	1053190	703	100172	0.70
Serine/threonine kinase 33 inhibitor	743321	172	319620	0.05
Tyrosyl-DNA phosphodiesterase inhibitors	489007	281	341084	0.08

[a]Ratio between number of active and the inactive compounds

Table 4.1: **Nine PubChem HTS datasets used in the benchmark study**

4.3.2 Generation of descriptors

BCL::Mol2D: To generate each AE, one of two atomic encoding schemes (Element or Atom) is assigned. All neighbor heavy atoms that are up to either two bonds (e.g. AE height=2) or one bond away (AE height=1) from the central atom are included in each AE. The shorter AE height of one was tested to see whether an increase in information density was beneficial for ANN training. Atom type and Element type AE libraries contain AEs were generated from in-house database of 900,000 drug-like small molecules, and with more than 100 counts. There are 574 AEs in the Atom type AE library and 240 in the Element type counterpart when AE height is set to 1. The BCL::Mol2D fingerprints are then generated to document counts of AEs in the AE library (Figure 4.2).

Molprint2D: The descriptors were generated with ElemRC atom types using the Schrodinger Canvas software suit, consistent with the optimal settings in the 2D fingerprint benchmark (Sastry et al., 2010). The AEs that appear from 1% to 90% molecules of each PubChem dataset were selected. The length of the Molprint2D ranges from 300 to 334 across the nine datasets.

Reduced SR (BCL::3D-RSR) descriptor set: This is a shortened version of the short range (SR) descriptor set introduced in a precedent study (Mendenhall and Meiler, 2016). The SR, containing 1315 descriptors in total, calculates six atomic properties for both signed and unsigned 2D/3D autocorrelation descriptors (Broto et al., 1984; Sliwoski et al., 2016a). The BCL::3D-RSR set reduces the number of descriptors down to 391: 23 scalar, 132 short range 2DA_Sign and 240 3DA_Sign descriptors (Supplementary Table 4.S1). Most of reduction is a result of using only signed versions of 2D and 3D autocorrelation (2DA_Signed and 3DA_Signed) (Sliwoski et al., 2016a) and only four atomic properties are calculated.

Hybrid fingerprint of BCL::3D-RSR set and BCL::Mol2D descriptor: descriptors from the BCL::3D-RSR set and BCL::Mol2D descriptors (height = 1) were combined to create hybrid fingerprints. BCL::Mol2D(Atom) +BCL::3D-RSR hybrid fingerprints comprise of 965 descriptor values, while BCL::Mol2D(Element)+BCL::3D-RSR hybrid fingerprints contain 631 descriptor values.

4.3.3 ANN-QSAR model training and evaluation

The performance of artificial neural network (ANN) – QSAR models using BCL::Mol2D descriptors was compared with those with the Molprint2D on each of the nine datasets. All ANN-QSAR models were trained with simple back propagation using a sigmoid transfer function with $\eta = 0.05$ and $\eta = 0.5$. The architecture of the ANNs consisted of a single hidden layer of 32 neurons and drop-out rates (Mendenhall and Meiler, 2016; Srivastava et al., 2014) of 0.05 for visible (input-layer) neurons, and 0.25 for hidden neurons, as previously optimized, with full connectivity to the input and output layer of the ANN.

QSAR models were evaluated with logAUC score (Mysinger and Shoichet, 2010), which is area under the

curve of the logarithmic receiver operating characteristic curve (logROC) between false positive rates of 0.001 to 0.1 (Mendenhall and Meiler, 2016). Each QSAR experiment was bootstrapped with replacement 2000 times to obtain logAUC mean and confident intervals using BCL v3.5, model:ComputeStatistics application. Average logAUC were computed across the nine HTS datasets for each descriptor condition. Two tailed two-sample t-test was then conducted to compare the average logAUC of different descriptor configurations. In each dataset, the active compounds were upsampled such that for each representation of an active compound is followed by presentation of 10 inactive compounds (A:I ratio = 0.1).

4.3.4 Cross-validation

Five-fold cross-validation was used throughout the evaluation of the QSAR models. After each of the nine PubChem datasets was randomized, it is split into fifths. The ANNs was trained on four of the parts (e.g. the training set), and made predictions on the last fifth (e.g. the test set). The folds then rotate so that all folds are used for training and testing the model. Each of the five parts was a test set for one set of four ANN models with different random seeds. The final performance metrics are averaged across the four estimates of each test fold.

4.3.5 Sensitivity analysis

We used the QSAR model trained on the dataset 2689 (Butkiewicz et al., 2013), which contains compounds from bioassays scanning for inhibitors of serine/threonine kinase 33 (STK33) (Liao et al., 2006) because this model yielded highest logAUC score. Each BCL::Mol2D descriptor, which corresponds to a specific substructure of the molecule, was evaluated for output sensitivity, i.e. the influence of small changes (+1 or -1) in the descriptor on the output of the ANN. Sensitivity score, S_{o,d_i}, of a descriptor d_i for a molecule m, was defined to be the discrete derivative of the output f with respect to the value of that descriptor, and was calculated as

$$S^m_{f(d_i),d_i} = \frac{(f(d_i+\delta) - f(d_i))}{\delta}$$

where δ is the change applied to d_i. $S^m_{f(d_i),d_i}$ is referred to as the decrement derivative when δ is -1 and increment derivative when δ is 1. To investigate the effects of removing and adding different AEs on altering the predicted activity of a structure, we selected eight actives for which there was a corresponding inactive molecule in the dataset with at least 90% of substructure in common. Predicted activity (ANN output) of the active compounds are greater than 0.95 and that of the inactive compounds is lower than 0.60. PubChem IDs and predicted activity of 16 compounds used in the sensitivity analysis are included in Table S5.

4.4 Results

This study is comprised of a benchmark and sensitivity analysis to evaluate performance and functionality of BCL::Mol2D descriptors. BCL::Mol2D (height=2) was first benchmarked against Molprin2D to examine the effects of changing the descriptor value from presence/absence to count of unique AEs. Then, we verified that decreasing the height of BCL::Mol2D from 2 to 1 would not significantly alter performance. The sensitivity analysis on BCL::Mol2D (height=1) aims at estimating how alteration of a certain substituent affects its corresponding molecular prediction output. Discrete derivatives of AEs were computed and mapped on the pharmacophore to signify potential impacts of adding or removing their corresponding substituents. In the final stage of the benchmark, we determined if adding BCL::Mol2D descriptor into the BCL::3D-RSR set improves its performance. Different descriptor configurations were evaluated through logAUC scores of trained QSAR-ANN models.

4.4.1 ANN-QSAR benchmarks on BCL::Mol2D in comparison to Molprint2D

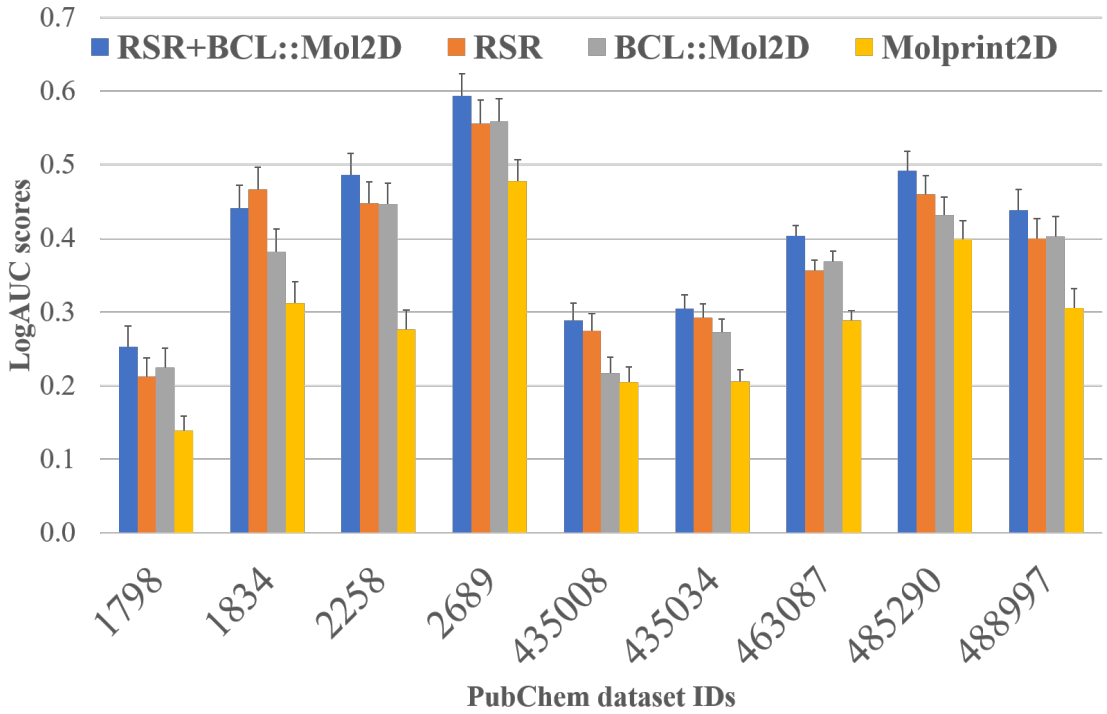


Figure 4.3: **BCL::Mol2D (green line) outperforms Molprint2D (red round dot line) by 26.7%, and improves RSR (yellow short dash line)'s performance by 6.8% when combined with RSR (purple long dash line).** BCL::Mol2D descriptors are atom typed with height=1. RSR+BCL::Mol2D are hybrid fingerprints from combining BCL::Mol2D and the BCL::3D-RSR descriptor set. Black nonagons represent various levels of logAUC score. Datasets, located at vertices of the nonagons, are referred by their PubChem assay IDs

ANN-QSAR models were trained to compare the performance of BCL::Mol2D vs. Molprint2D (Bender

et al., 2004; Sastry et al., 2010) across nine HTS PubChem datasets (Butkiewicz et al., 2013). table 4.2 summarizes the details of descriptor configurations and their average logAUC scores, and figure 4.3 illustrates the performance of those descriptors broken down into individual datasets. Following the design of Molprint2D, our initial implementation of BCL::Mol2D with the AE height of 2. While Molprint2D (height=2) contains binary bits that record the presence/absence of Element type AEs, BCL::Mol2D documents counts of either Element or Atom type AEs. The logAUC scores of ANNs trained with either of those two descriptors were measured across nine HTS PubChem datasets. Compared to the performance of ANNs with Molprint2D, BCL::Mol2D significantly improved ANN predictive power up 22.52% for Atom type, and 25.93% for Element type (p-values < 0.01).

Descriptor name	Atomic encoding scheme of AEs	AE height	AE Value type	logAUC mean	logAUC SD	AUC mean	AUC SD	Number of descriptors	Software
Molprint2D	Element[a]	2	Presence	0.290	7.8E-03	0.785	5.7E-03	300 to 334	CANVAS
BCL::Mol2D	Element	2	Count	0.365	8.3E-03	0.787	6.5E-03	5117	BCL
BCL::Mol2D	Atom	2	Count	0.355	8.5E-03	0.763	6.8E-03	8080	BCL
BCL::Mol2D	Element	1	Count	0.337	8.4E-03	0.816	5.4E-03	240	BCL
BCL::Mol2D	Atom	1	Count	0.367	8.5E-03	0.822	5.5E-03	574	BCL
BCL::Mol2D	Element+Atom	1	Count	0.368	8.5E-03	0.822	5.5E-03	814	BCL
BCL::3D-RSR	NA	NA	NA	0.385	8.5E-03	0.835	5.2E-03	391	BCL
BCL::3D-RSR +BCL::Mol2D	Element	1	Count	**0.406**	8.6E-03	**0.842**	5.2E-03	631	BCL
BCL::3D-RSR +BCL::Mol2D	Atom	1	Count	**0.411**	8.7E-03	**0.841**	5.3E-03	965	BCL

[a]Element type of Molprint2D also has the information of whether the atoms are in aromatic/non-aromatic rings.

Table 4.2: **Average logAUCs, AUCs and their SDs across nine PubChem datasets and number of descriptors for different descriptor configurations**

4.4.2 Reducing the AE height from two bonds to one shrinks the size of the BCL::Mol2D fingerprint without reducing the QSAR performance.

Since each BCL::Mol2D descriptors documents count of an unique AE, length of BCL::Mol2D fingerprint equals the size of the AE library. The AE library was built by collecting AEs that appeared more than 100 times among 900,000 drug-like small molecules. However, this common AE list contains several thousand AEs if the height is set to 2. The resulting fingerprints were likewise very sparse – less than 0.7% of all descriptor values were non-zero. We hypothesized that this fingerprint is unnecessarily large to encode even the most complex, drug-like, molecules with less than 100 unique AEs. Reducing the AE height to 1 reduces the length of the BCL::Mol2D fingerprint. It is 14-fold for Atom type and 20-fold for Element type (table 4.2). The sparsity of the descriptors is also reduced – now up to 25% of all descriptor values are non-zero.

ANN-QSAR models were trained on BCL::Mol2D descriptors with either Atom or Element atom encod-

ing scheme, which built AEs of either one or two-bond limit. The results (table 4.2) suggested that reducing the bond limit in BCL::Mol2D had no significant effects on predictive power of the QSAR models with Atom type (+3.4%, p-value ¿ 0.05), although doing so would moderately lower performance of BCL::Mol2D Element type fingerprints across nine HTS datasets (-7.5%, p-value ¡ 0.05). Compared to the logAUC scores of ANNs with Molprint2D, BCL::Mol2D (height=1) still significantly improved ANN predictive power up 26.7% for Atom type (Figure 4.3), and 16.5% for Element type (p-values¡0.01).

4.4.3 BCL::Mol2D moderately improves logAUC when combined the BCL::3D-RSR set

The BCL::Mol2D (height=1) descriptor was also tested in conjunction with a previously optimized descriptor set, BCL::3D-RSR, that utilizes a mix of 2D and 3D auto-correlation functions, along with scalar molecular descriptors (Mendenhall and Meiler, 2016). More specifically, the combined sets modestly, though consistently, performed better than just the BCL::3D-RSR set alone. The performance improvement from adding BCL::Mol2D to the BCL::3D-RSR set is 6.8% (p-value ¡0.01) when adding Atom type descriptors, and 5.5% (p-value ¡ 0.01) for the Element type ones (table 4.2). Additionally, combining the BCL::3D-RSR set with BCL::Mol2D consistently performed better than BCL::Mol2D alone. In particular, adding BCL::3D-RSR set improved the average logAUC score by 20.3% for Element type, and 12.0% for Atom type (p-values¡0.01) (table 4.2 and Figure 4.3). Means, standard deviations, and 95% confident intervals of logAUC scores from ANN-QSAR models trained on all descriptor configurations mentioned above are summarized in the supplementary Table 4.S2.

4.4.4 Pharmacophore mapping through sensitivity analysis of the ANNs

Unlike Molprint2D descriptors, BCL::Mol2D descriptor values correspond to the counts of molecular substructures that they represent. Hence, we can estimate the effects of adding or removing a certain substituent based on sensitivity analysis of the AEs that the substituent encompasses. Discrete derivatives of molecular predicted activity were computed for each unique AE when adding (increment derivative) or removing (decrement derivative) that AE. Eight pairs of an active and an inactive with less than 10% difference in structure (according to Tanimoto index (Rogers and Tanimoto, 1960a)) were selected for this analysis as described in the methods section. We investigated whether the ANN model can predict which structural differences between those active and inactive compounds cause significant differences in their bioactivity.

The decrement derivatives of AEs are mapped onto their corresponding atom for each of 16 compounds in the analysis (Figure 4.4, first column). The results show that removal of substructures with positive decrement derivatives (blue region) often boost the predicted activity, while addition of ones with negative decrement derivatives (red regions) leads to worsen value of the output value of the compound. The corresponding fig-

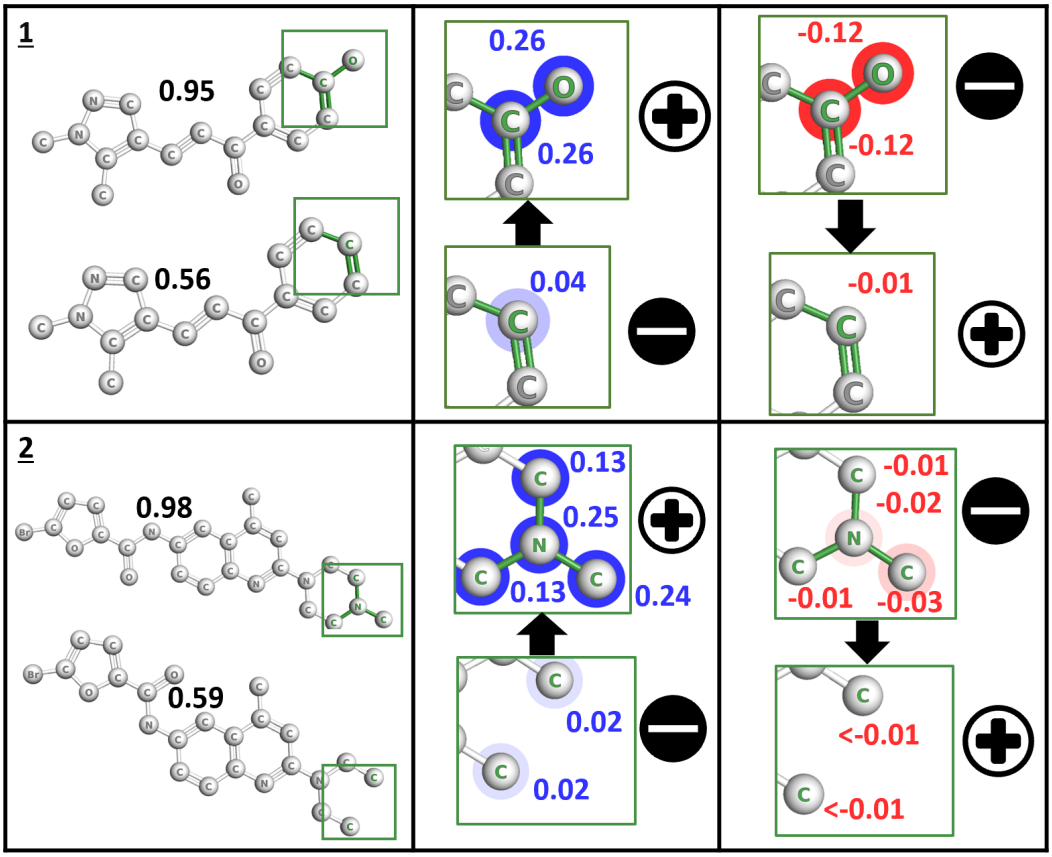


Figure 4.4: **Mapping partial contributions of AEs to the ANN prediction output of STK33 inhibitors using BCL::Mol2D (Atom type, height=1).** The first column contains general structures of two pairs of compounds (one active and one inactive) with their corresponding ANN predicted activities. The atoms that are different between active and inactive compounds are colored in green (green rectangles). The second and third columns illustrate the transformation from active to inactive and from inactive to active, respectively. The directions of the transformation are shown in black arrows. The atoms that are highlighted in green are colored based on the finite differences of their corresponding AEs. Red circles mean negative values, and blue circles have positive values. The decrement derivatives (marked with minus signs) are represented on the deleted substructures, and the increment derivatives (marked with plus signs) are represented on added substructures in each transformation. Additional examples are reported in figure S5.

ures for seven of the eight actives illustrate that altering the blue regions while keeping red regions intact lead to higher predicted activity, with only the 3rd scaffold appearing to contradict this conclusion. The transformation from inactive 3 to active 3 is the only one that involved removal of a AE with negative decrement derivative. However, taking away the other AEs with larger positive decrement derivative from this molecule could offset that negative effect. More specifically, removing the hydroxyl oxygen AE (decrement derivative = 0.354) increases the predicted activity about 4.7x more than the reduction caused by removal of the aromatic carbon (decrement derivative = -0.076).

To identify the type of changes in molecular structures that significantly affect their predicted activity,

sums of increment and derivatives were computed for added and removed AEs, respectively for transformation from inactive to active, and from active to inactive. Generally, transformation from inactive to active compounds replaced AEs, whose sum of decrement derivatives is positive, by AEs with positive sum of increment derivatives (Figure 4.4, second column). Again, in the case of the scaffold 5, the benefit of adding the chloride groups on the inactive 5 (with the sum of increment derivatives of 0.727) might outweigh the effect of the removed the carbons (decrement derivative of -0.043). A similar trend is shown in the transformation from active to inactive compounds. The values of removed and added AEs for each scaffold are listed in the supplement Tables S2 and S3.

4.4.5 A case study of applying Lead optimization through derivatization

We illustrate here an example of applying knowledge from the sensitivity analysis of the BCL::Mol2D descriptors on derivatization to improve the ANN prediction output (Fig. 5). From an inactive STK inhibitor (the inactive compound of the compound pair 5 from the sensitivity analysis, figure S5), we have created a new compound with improvement in ANN prediction output from 0.42 to 1.0 after two steps of modification. In each step, we manually select the added and removed functional group with positive decrement and increment derivatives, respectively. One exception is the added in aromatic carbon atom the second step, which has a negligible increment derivative. However, the transformation in step 2 still improves the ANN prediction output of the compound because the effect of removing the chloride group (decrement derivative sum = 0.12) outweighs the impact of adding the aromatic carbon (increment derivative = -0.01). The values of removed and added AEs for each step are listed in the Table S3.

4.5 Discussion

In this study, BCL::Mol2D descriptors were tested using an established QSAR benchmark of nine large high-throughput screens to ensure general applicability of the method. We observe consistent improvements in logAUC scores over all datasets when the QSAR models when trained with BCL::Mol2D instead of Molprint2D, even though the height of AEs of BCL::Mol2D was reduced from two to one. This observation suggests that 2D information of an additional layer of neighboring atoms fails to improve the useful informational content to the ANNs. Interestingly, performance of Element type BCL::Mol2D fingerprints, although with fewer AEs, is significantly higher than performance of Molprint2D. This improvement suggests that counts of unique AEs, even with a height of one, provide more information than just their presence/absence with height of two, like Molprint2D, perhaps by helping distinguish substructures with similar AE composition. Adding electron configuration of atoms, which further differentiates AEs with the same element type, was also shown to improve performance of the fingerprints.

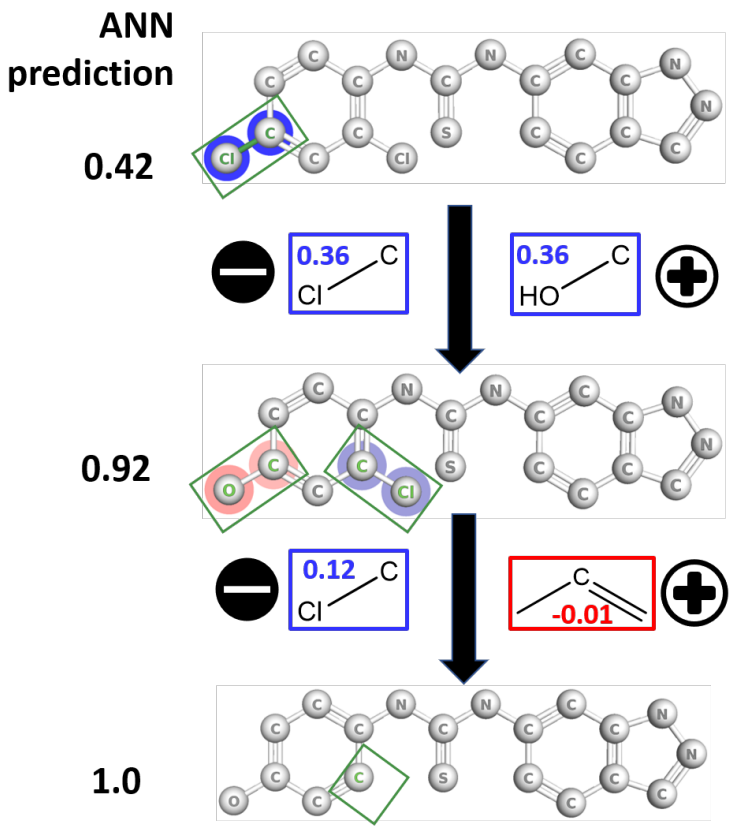


Figure 4.5: **Applying sensitivity analysis of BCL::Mol2D descriptor to lead optimization through derivatization.** Starting from the inactive compound from the STK inhibitor HTS, we remove functional groups with favorable decrement (marked with black minus signs) and add functional groups with favorable increment (mark with black plus signs) derivatives. The process results in a known active compound much higher ANN prediction output (denoted by black numbers on the most left side). The substructures that are modified in each step are labeled and framed in green, and colored based on the decrement derivatives of their corresponding AEs (positive: blue; negative: red). Between molecular structures: added or removed substructures in each step are framed according to the sum of increment and decrement derivatives, respectively (blue: positive value, red: negative value).

Combining BCL::Mol2D and the BCL::3D-RSR descriptors sets yields a modest, but consistent performance improvement over the optimized BCL::3D-RSR set alone. This suggests possible partial information overlap between two descriptor sets. Future studies could consider performing descriptor selection analysis on the hybrid fingerprints to prune out the descriptors that do not provide meaningful information to the model.

Since the values of BCL::Mol2D descriptors directly relates to atomic fragments of the molecular structures, derivatives of individual AEs can be used to estimate the effects of removal/addition of functional groups on the scaffolds. Previous studies (Baskin et al., 2002; Carlsson et al., 2009; Marcou et al., 2012) have attempted to estimate/rank the global importance of descriptors based on their partial derivatives. How-

ever, since we are interested in extracting information from the ANNs to optimize specific scaffolds, we only focused on effects of changes in local substructures to a specific prediction. Furthermore, as values BCL::Mol2D descriptors are discrete integers, their decrement and increment "discrete derivatives" are likely to be different than those computed using a traditional continuous derivative calculation. Hence, using these two types of discrete derivatives would distinguish the effects of removing and adding a particular AE in scaffold optimization. Carlson et al. (2009) suggested that changes of important regions (with high sensitivity scores) would alter the predicted activity of that molecules (Carlsson et al., 2009). However, equivocating descriptor importance with their partial first derivatives for a molecule is not always meaningful because descriptor values could have a partial first derivative of zero while retaining a large second derivative when they are at local optima. Likewise, we did not measure the centered first derivatives of the ANN with respect to descriptors in order to rank their importance, instead looking at discrete increments or decrements of the descriptor. We propose that, AEs with positive decrement derivatives should be replaced by AEs with positive increment derivatives to improve the predicted activity.

We propose that AEs with positive decrement derivatives should be replaced by AEs with positive increment derivatives to improve the ANN prediction output. We tested this proposal by applying the sensitivity analysis of BCL::Mol2D on transform an inactive STK inhibitor to a novel compound with more than substantial improvement in ANN prediction output. We hence demonstrated that we can use BCL::2D descriptors to leverage knowledge from the QSAR-ANN models to optimize lead compounds. Although in this study, we only focus on the derivatization aspect of the lead optimization, performing more central and dramatic modifications on the scaffolds should be possible, though the success of such an approach will depend heavily on whether the training data contained molecules with similar structures.

4.6 Acknowledgment

This study is funded by Molecular Science Software Institute (MolSSI) Fellowship and NIH. MolSSI is funded by the NSF grant (ACI-1547580). Work in the Meiler laboratory is supported through NIH (R01 GM099842, R01 DK097376) and NSF (CHE 1305874). The author would like to thank Dr. Francois Berenger for discussion regarding descriptor design.

4.7 Author Contributions

OV, J. Mendenhall and J. Meiler designed the study. OV implemented the descriptor, performed the benchmark and analysis, and wrote the manuscript. J. Meiler and DA supervised the project. J. Mendenhall, DA and J. Meiler edited the manuscript. All authors read and approved the final manuscript.

4.8 Supplementary Material

Descriptor types	Molecular/atomic properties
Scalar (one value for each molecular property)	Molecular weight # hydrogen bond donors and acceptors LogP-octanol/water coefficient Molecular total charge # rotatable bonds # aromatic rings # rings Total polar surface area of molecule (TPSA) Maximum number of bond between two atoms (Bond girth) Widest diameter of the molecule (Girth) # atoms in the largest ring # bridge atoms in fused rings # bridge atoms in fused aromatic rings Min, max, absolute sum of σ charges Min, max, absolute sum of V charges
2DA_Sign (11 bonds, 36 values for each atom property)	σ charges V charges IsHTernary[a] IsInAromaticRingTernary[b]
3DA_Sign (0.25 Å, 6 Å max, 72 values/atom properties)	σ charges V charges IsHTernary[+] IsInAromaticRingTernary[b]

Table 4.S1: **List of descriptors in the BCL::3D-RSR set**

Descriptors	**Element-2bond**				**Element-1bond**			
Datasets	Average	SD	95% CI		Average	SD	95% CI	
1798	0.222	0.025	0.18	0.264	0.197	0.025	0.156	0.239
1834	0.355	0.029	0.308	0.402	0.363	0.03	0.314	0.413
2258	0.448	0.029	0.401	0.496	0.396	0.029	0.348	0.443
2689	0.556	0.032	0.504	0.609	0.505	0.033	0.451	0.56
435008	0.243	0.022	0.207	0.281	0.197	0.021	0.163	0.232
435034	0.262	0.018	0.232	0.291	0.274	0.018	0.245	0.304
463087	0.344	0.014	0.321	0.369	0.347	0.014	0.325	0.371
485290	0.438	0.025	0.395	0.478	0.42	0.025	0.379	0.46
488997	0.415	0.026	0.372	0.458	0.337	0.025	0.296	0.379
Descriptors	**Atom-2bond**				**Atom-1bond**			
Datasets	Average	SD	95%CI		Average	SD	95% CI	
1798	0.208	0.026	0.168	0.251	0.224	0.027	0.18	0.269
1834	0.349	0.029	0.303	0.4	0.382	0.031	0.331	0.434
2258	0.434	0.029	0.387	0.481	0.446	0.029	0.398	0.494
2689	0.529	0.032	0.476	0.581	0.559	0.031	0.508	0.611
435008	0.238	0.023	0.201	0.277	0.217	0.022	0.182	0.254
435034	0.252	0.018	0.223	0.281	0.273	0.018	0.244	0.302
463087	0.327	0.015	0.303	0.351	0.368	0.015	0.345	0.393
485290	0.443	0.025	0.4	0.483	0.432	0.024	0.389	0.47
488997	0.415	0.027	0.371	0.459	0.402	0.027	0.359	0.448
Descriptors	**Molprint2D**				**RSR**			
Datasets	Average	SD	95% CI		Average	SD	95% CI	
1798	0.139	0.02	0.106	0.173	0.212	0.025	0.171	0.254
1834	0.312	0.029	0.263	0.359	0.466	0.03	0.417	0.515
2258	0.276	0.026	0.233	0.318	0.448	0.029	0.4	0.496
2689	0.478	0.029	0.429	0.527	0.556	0.032	0.505	0.61
435008	0.205	0.021	0.17	0.24	0.274	0.024	0.236	0.313
435034	0.205	0.016	0.179	0.233	0.292	0.018	0.262	0.324
463087	0.289	0.013	0.268	0.31	0.356	0.014	0.334	0.38
485290	0.398	0.026	0.356	0.44	0.46	0.025	0.417	0.5
488997	0.306	0.026	0.265	0.35	0.399	0.027	0.356	0.445
Descriptors	**RSR+Atom-1bond**				**RSR+Element-1bond**			
Datasets	Average	SD	95% CI		Average	SD	95% CI	
1798	0.252	0.028	0.205	0.299	0.234	0.027	0.189	0.278
1834	0.441	0.031	0.389	0.492	0.458	0.030	0.407	0.508
2258	0.486	0.03	0.437	0.536	0.481	0.030	0.433	0.532
2689	0.593	0.03	0.544	0.644	0.577	0.031	0.526	0.628
435008	0.288	0.024	0.249	0.328	0.285	0.024	0.247	0.325
435034	0.305	0.019	0.275	0.336	0.308	0.018	0.277	0.339
463087	0.403	0.015	0.379	0.428	0.400	0.015	0.376	0.425
485290	0.492	0.026	0.447	0.532	0.484	0.026	0.441	0.525
488997	0.438	0.028	0.393	0.485	0.427	0.028	0.383	0.474
Descriptors	**(Element+Atom)-1bond**							
Datasets	Average	SD	95% CI					
1798	0.215	0.026	0.171	0.259				
1834	0.389	0.03	0.341	0.439				
2258	0.458	0.029	0.412	0.507				
2689	0.548	0.032	0.495	0.602				
435008	0.22	0.022	0.183	0.259				
435034	0.273	0.018	0.243	0.304				
463087	0.366	0.014	0.344	0.39				
485290	0.436	0.025	0.394	0.474				
488997	0.403	0.027	0.36	0.449				

Table 4.S2: **Average, standard deviation (SD), and 95% confidence interval (CI) of logAUC scores of different BCL::Mol2D descriptor configurations (atom hash and height) across nine PubChem datasets**

Descriptors	**Element-2bond**				**Element-1bond**			
Datasets	Average AUC	SD	95% CI		Average AUC	SD	95% CI	
1798	0.656	0.025	0.615	0.696	0.671	0.022	0.635	0.706
1834	0.773	0.024	0.734	0.812	0.861	0.015	0.835	0.886
2258	0.812	0.020	0.778	0.844	0.842	0.016	0.815	0.869
2689	0.826	0.024	0.786	0.864	0.878	0.017	0.849	0.905
435008	0.737	0.020	0.703	0.770	0.733	0.018	0.704	0.761
435034	0.804	0.013	0.781	0.826	0.831	0.012	0.811	0.851
463087	0.872	0.008	0.860	0.885	0.901	0.006	0.892	0.911
485290	0.808	0.017	0.778	0.837	0.828	0.016	0.801	0.854
488997	0.798	0.020	0.764	0.830	0.795	0.017	0.767	0.823
Descriptors	**Atom-2bond**				**Atom-1bond**			
Datasets	Average AUC	SD	95%CI		Average AUC	SD	95% CI	
1798	0.636	0.025	0.596	0.677	0.691	0.022	0.655	0.727
1834	0.751	0.024	0.710	0.792	0.855	0.016	0.829	0.881
2258	0.795	0.021	0.761	0.830	0.836	0.017	0.808	0.865
2689	0.820	0.023	0.782	0.858	0.883	0.017	0.855	0.911
435008	0.712	0.021	0.677	0.746	0.754	0.018	0.724	0.784
435034	0.760	0.016	0.734	0.786	0.833	0.012	0.813	0.853
463087	0.821	0.010	0.805	0.836	0.903	0.006	0.893	0.912
485290	0.775	0.020	0.741	0.806	0.832	0.016	0.806	0.857
488997	0.801	0.019	0.769	0.830	0.806	0.018	0.776	0.836
Descriptors	**Molprint2D**				**RSR**			
Datasets	Average AUC	SD	95% CI		Average AUC	SD	95% CI	
1798	0.654	0.023	0.617	0.692	0.688	0.022	0.651	0.722
1834	0.794	0.019	0.761	0.825	0.884	0.016	0.858	0.909
2258	0.765	0.020	0.732	0.797	0.849	0.017	0.821	0.876
2689	0.866	0.018	0.833	0.895	0.915	0.012	0.894	0.935
435008	0.768	0.017	0.740	0.795	0.799	0.018	0.769	0.828
435034	0.779	0.013	0.757	0.800	0.841	0.012	0.821	0.860
463087	0.860	0.008	0.847	0.872	0.898	0.006	0.888	0.908
485290	0.810	0.015	0.783	0.835	0.852	0.014	0.827	0.874
488997	0.772	0.018	0.743	0.801	0.785	0.019	0.754	0.816
Descriptors	**RSR+Atom-1bond**				**RSR+Element-1bond**			
Datasets	Average AUC	SD	95% CI		Average AUC	SD	95% CI	
1798	0.708	0.022	0.671	0.745	0.703	0.022	0.665	0.738
1834	0.874	0.018	0.844	0.901	0.901	0.014	0.878	0.923
2258	0.855	0.017	0.828	0.882	0.857	0.016	0.830	0.883
2689	0.901	0.016	0.874	0.926	0.895	0.017	0.867	0.922
435008	0.805	0.017	0.777	0.833	0.809	0.016	0.783	0.835
435034	0.837	0.013	0.816	0.858	0.842	0.012	0.822	0.861
463087	0.913	0.006	0.903	0.922	0.917	0.005	0.908	0.926
485290	0.866	0.014	0.843	0.887	0.850	0.015	0.825	0.875
488997	0.813	0.018	0.784	0.842	0.808	0.018	0.778	0.838
Descriptors	**(Element+Atom)-1bond**							
Datasets	Average AUC	SD	95% CI					
1798	0.694	0.022	0.658	0.731				
1834	0.851	0.017	0.822	0.878				
2258	0.831	0.019	0.800	0.862				
2689	0.886	0.017	0.858	0.912				
435008	0.743	0.018	0.714	0.772				
435034	0.834	0.012	0.814	0.854				
463087	0.897	0.006	0.887	0.907				
485290	0.844	0.015	0.818	0.868				
488997	0.821	0.017	0.793	0.848				

Table 4.S3: **Average, standard deviation (SD), and 95% confidence interval (CI) of logAUC scores of different descriptor configurations (RSR and hybrid) across nine PubChem datasets**

Scaffold	Active/inactive	PubChem ID	Predicted activity by ANNs
1	active	5736857	0.95
1	inactive	6015853	0.56
2	active	3240925	0.98
2	inactive	9550320	0.59
3	active	5334477	0.93
3	inactive	5344140	0.57
4	active	1725836	0.93
4	inactive	24761375	0.19
5	active	2396745	0.98
5	inactive	2396755	0.42
6	active	5767865	0.96
6	inactive	5749690	0.49
7	active	1918543	0.95
7	inactive	1916919	0.45
8	active	5409364	0.96
8	inactive	5439215	0.21

Table 4.S4: **Derivatives of AEs involved in the transformation from inactive to active compounds for eight scaffolds in the sensitivity analysis**

CHAPTER 5

Mapping the binding sites of UDP and prostaglandin E2 glyceryl ester in the nucleotide receptor P2Y6

Parts of this chapter was taken from the manuscript by Anne Zimmermann*, Antje Brüser*, Oanh Vu*, Gregory Sliwoski, Lawrence J. Marnett, Jens Meiler, and Torsten Schöneberg.

*Those authors contributed equally to this work

5.1 Abstract

Cyclooxygenase-2 catalyzes the biosynthesis of prostaglandins from arachidonic acid and the biosynthesis of prostaglandin glycerol esters (PG-Gs) from 2-arachidonoylglycerol. PG-Gs are mediators of several biological actions such as macrophage activation, hyperalgesia, synaptic plasticity, and intraocular pressure. Recently, the human UDP receptor P2Y6 was identified as a target for prostaglandin E2 glycerol esters (PGE_2-G). Here, we show that UDP and PGE_2-G are evolutionary conserved endogenous agonists at vertebrate P2Y6 orthologs. Using sequence comparison of P2Y6 orthologs, homology modeling, and ligand docking studies, we proposed several receptor positions participating in agonist binding. Site-directed mutagenesis and functional analysis of these P2Y6 mutants revealed that both UDP and PGE_2-G share in parts one ligand-binding site. Thus, the convergent signaling of these two chemically very different agonists has already been manifested in the evolutionary design of the ligand-binding pocket.

5.2 Introduction

Cyclooxygenases (COX) catalyze the rate-limiting step of prostaglandin biosynthesis. Prostaglandins are potent bioactive lipid messengers that realize their functions via activation of G protein-coupled receptors (GPCRs)(Woodward et al., 2011). Besides this well-studied enzymatic function of COX isoenzymes, the inducible COX-2 selectively oxygenates 2-arachidonoylglycerol to form prostaglandin glycerol esters (PG-Gs)(Alhouayek and Muccioli, 2014; Kozak et al., 2001). Due to the rapid degradation of PG-Gs, there is limited knowledge about their biological function(Kingsley et al., 2019). Previous studies suggested that the PG-Gs PGE_2-G and PGF2-G may activate GPCRs in the murine macrophage-like cell line RAW264.7 and the human lung's adenocarcinoma cell line H1819(Nirodi et al., 2004; Richie-Jannetta et al., 2010). The fast Ca2+ response observed with both cell lines indicated specific signal transduction via unknown Gq- and/or Gi protein-coupled receptors. Using a subtractive screening approach, where mRNA from PGE_2-G response-positive and -negative cell lines was subjected to transcriptome-wide RNA sequencing analysis, we identified the UDP receptor P2Y6 as the target of PGE_2-G(Bruser et al., 2017).

Because P2Y6 is expressed in the spleen, thymus, intestine, leukocytes, and aorta, and PGE_2-G is involved in inflammation and macrophage activation, there is accumulating evidence that the P2Y6/PGE_2-G pair functions in an auto-/paracrine mode. Studies with P2Y6-deficient mice have shown that P2Y6 is involved in the UDP-dependent contraction and endothelium-dependent relaxation of the aorta(Bar et al., 2008). P2Y6 is also reported to have high relevance in the immune system(Le Duc et al., 2017). For example, it was demonstrated using P2Y6-deficient mice that the receptor fine-tunes the activation of T cells in allergen-induced pulmonary inflammation(Garcia et al., 2014; Giannattasio et al., 2011) and reduces macrophage-mediated cholesterol uptake in atherosclerotic lesions(Stachon et al., 2014). PGE_2-G is known to induce hyperalgesia(Hu et al., 2008). Experiments with a mouse model of sickle cell disease revealed elevated COX-2 and PGE_2-G levels responsible for persistent inflammation and hyperalgesia. Pharmacological COX-2 or P2Y6 inhibition suggested the P2Y6/PGE_2-G pair as a mediator of pain in this animal model(Khasabova et al., 2019).

Currently, it is hypothesized that P2Y6 integrates the two different chemical signals, UDP and PGE_2-G, to a shared intracellular response. Here, nucleotides are released into the extracellular space upon injury and inflammation to serve as a "danger" signal exerting pro-inflammatory effects(Kepp et al., 2017). Cell lysis results in an immediate release of nucleotides to reach concentrations ¿100 nM and recruitment of macrophages via stimulation of P2Y receptors(Communi et al., 2000; Lazarowski et al., 2003). Similarly, PGE_2-G acts via P2Y6 to regulate the fast and efficient recruitment of macrophages. Previous studies revealed an extremely low EC50 value in the range of 1 pM for PGE_2-G at its receptor(Bruser et al., 2017; Nirodi et al., 2004; Richie-Jannetta et al., 2010). Physiologically, this seems reasonable because PGE_2-G only occurs in low amounts and is rapidly hydrolyzed to PGE2(Kozak et al., 2001). UDP has been shown to lower intraocular pressure via activation of P2Y6 expressed in the ciliary body making the receptor a promising target for glaucoma treatment(Jacob et al., 2018; Shinozaki et al., 2017). Interestingly, PGE_2-G also reduces intraocular pressure in dogs and monkeys(Woodward et al., 2016), and one can speculate that this effect is mediated via P2Y6.

Identification of P2Y6 as receptor targeted by PGE_2-G was a first critical step to characterize the physiological function of PG-Gs and to manipulate this signaling system pharmacologically. However, the structural basis of the promiscuity to at least two structurally not related endogenous agonists is still enigmatic. Our initial studies addressed whether UDP and PGE_2-G share the binding pocket or bind at different sites. Current data supported the hypothesis that UDP and PGE_2-G most probably share receptor interaction sites, but additional determinants private to each agonist may contribute to the individual binding pockets(Bruser et al., 2017). In this study, we extended our initial structure-function relation studies by predicting potential interaction sites between the agonists and human P2Y6 with the help of molecular docking and by perform-

ing site-directed mutagenesis studies. We found that the agonist specificity of P2Y6 is evolutionarily old and was already established for both UDP and PGE_2-G in fish orthologs. Ortholog comparison, homology modelling, ligand docking, and molecular dynamics simulation proposed several receptor positions participating in agonist binding. Functional analysis of mutant P2Y6 revealed an overlapped binding pocket of both endogenous agonists.

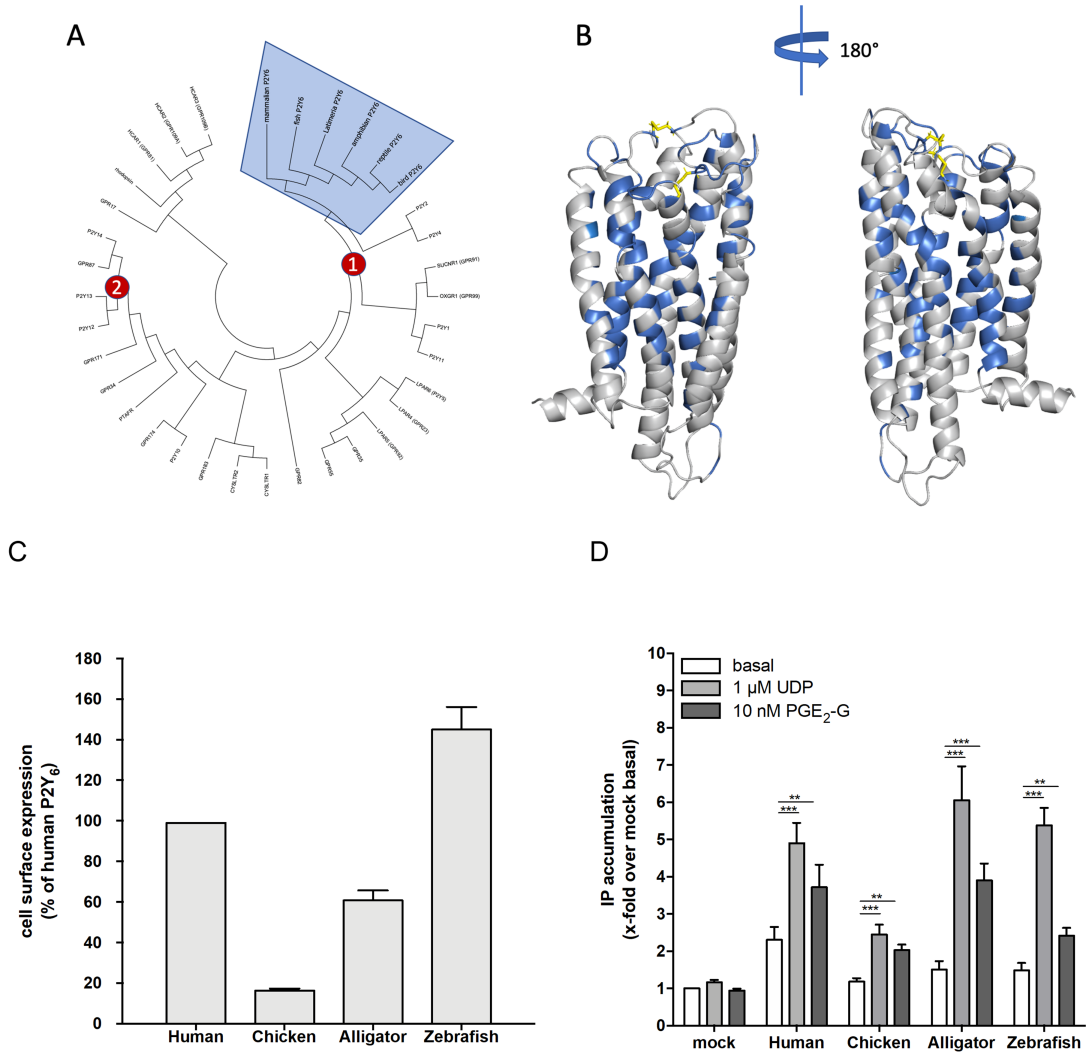


Figure 5.1: **Phylogenetic relation and structural functional conservation of vertebrate P2Y6.** (A) The amino acid sequence of 233 P2Y6 orthologs were aligned using the MUSCLE algorithm (Edgar, 2004). When compared to all other human P2Y-like sequences all orthologs cluster at the expected position in the phylogenetic tree. The evolutionary history was inferred using the Neighbor-Joining method Saitou and Nei (1987). Cluster 1 (red circle 1) represents the P2Y1-like receptor subgroup and Cluster 2 (red circle 2) the P2Y12-like receptor subgroup. (B) Using a homology modeling approach, the 3D structure of the human P2Y6 was generated Bruser et al. (2017) and the 100% conserved positions from the vertebrate P2Y6 alignment (A) are depicted in blue and yellow (disulfide bridges). (C) HEK293T cells were transiently transfected with either HA-tagged version of the indicated vertebrate P2Y6 orthologs and the expression levels of receptors were measured by a cell surface ELISA (see Methods). (D) HEK293 cells transfected with the indicated vertebrate P2Y6 orthologs were used for intracellular IP measurements (see Methods). The basal IP1 levels of mock-transfected cells was 21.7 $\pm$ 1.7 nM. All data are given as means $\pm$ SEM of four (A) and three (B) independent experiments each performed in triplicate. *p $<$0.05, **p $<$0.01, ***p $<$0.001 (paired Student's t test).

5.3 Results

5.3.1 Evolutionary conservation of agonist promiscuity.

In our recent study, we found PGE_2-G as an endogenous agonist for human and mouse P2Y6, in addition to UDP(Bruser et al., 2017). Present study aims to identify positions within the human P2Y6 relevant for mediating this promiscuous agonist profile. As a first step towards this goal, we followed an evolutionary approach to predict the functional relevance of each position within a given protein on the basis of sequence data from orthologs(Coster et al., 2012). Sequence divergence of a given amino acid position in a protein is the result of an evolutionary process characterized by the continuous accumulation of mutations, which are subsequently accepted or rejected by natural selection. This process leaves a signature of divergence (high evolutionary rate) or conservation (low evolutionary rate) for each position in the protein sequences. As tested in a phylogenetic analysis (Figure 5.1A), P2Y6 is an ideal candidate for such an analysis because it is present in almost all vertebrates with one-to-one orthology from fish to mammals. Aligning 233 full-lengths vertebrate orthologs from mammals, birds, reptiles, amphibians, and fishes (accession numbers and alignment are given in the supplementary material file P2Y6 orthologs.fas), we found 99 amino acid positions (30.2% of all positions in human P2Y6) that are 100% conserved between all orthologs. As shown in Figure 5.1B, these fully conserved positions localize preferentially to the transmembrane helixes 1-7 (TM1-7) with side chains pointing inside the fold stabilizing interactions between TMs and contributing to putative binding pockets for agonists. We tested if these residues provide conserved agonism of both UDP and PGE_2-G in distantly related orthologs using functional assays.

Although the functionality of P2Y6 upon UDP stimulation has been proven in fish(Li et al., 2018), salamander(Reifel Saltzberg et al., 2003), and chicken(Li et al., 1998; Webb et al., 1996), it is unknown whether PGE_2-G agonism at P2Y6 is preserved at non-mammalian P2Y6 orthologs. Therefore, we cloned P2Y6 orthologs from zebrafish, alligator, and chicken and measured the cell surface expression of N-terminally HA-tagged receptors in a cellular ELISA. As shown in Figure 5.1C, except for the chicken P2Y6 ortholog (only 20% cell surface expression), all other variants are well-expressed at the cell surface compared to the human receptor allowing for functional assays. P2Y6 couples to Gq/11 proteins, and activation increases intracellular inositol phosphate (IP) levels(Bruser et al., 2017). Functional analysis in an IP1 accumulation assay with saturating concentrations of UDP (1 μM,(Bruser et al., 2017)) revealed the expected responses in HEK293T cells transiently transfected with the different P2Y6 orthologs (Figure 5.1D). The lower UDP-induced IP1 levels in cells transfected with the chicken ortholog correlated with its lower cell surface expression (Figure 5.1C). Next, P2Y6 orthologs were tested with saturating concentration of PGE_2-G (10 nM,(Bruser et al., 2017)) to determine whether its agonistic property is conserved during evolution. PGE_2-G-induced IP1-formation was

seen for the human, alligator, chicken, and zebrafish orthologs (Figure 5.1D). Our assay results are consistent with the presence of COX-2, the main prostaglandin-endoperoxide synthase, which is capable of generating PGE_2-G(Kozak et al., 2001) in all species investigated (see NCBI sequence database). Since agonism was seen in fish, reptiles, birds, and mammals, the common molecular architecture of P2Y6 orthologs must have preserved the conserved agonist- and signal transduction specificity.

5.3.2 PDE2-G and UDP have a partially overlapping binding pocket.

Currently, there is no experimental structure available for P2Y6. To estimate whether the two different agonists, UDP and PGE_2-G, may share structural determinants when interacting with the receptor, we simulated binding by docking the agonists into a comparative model of P2Y6(Bruser et al., 2017). Based on this initial modeling and docking study we have formed a hypothesis that the two ligands UDP and PGE_2-G may have an overlapping binding pocket flanked by transmembrane helices (TM) 3, 5, 6, and 7 with PGE_2-G extending further to TM2, ECL2, the extracellular tip of TM6, and the core of TM3 (suppl. Figure 5.S1). The model suggested that UDP and PGE_2-G share a number of interaction sites with others being specific for one of the two agonists. For example, both UDP and PGE_2-G form hydrogen bonds with positions R103 and R287 and orient their phosphate moieties and glycerol ester moieties, respectively, towards these positively charged amino acid residues of P2Y6. A precedent docking study already predicted that R103 and R287 contribute to UDP binding(Costanzi et al., 2005). Previously, we identified Y262 mainly participating in UDP binding(Bruser et al., 2017).

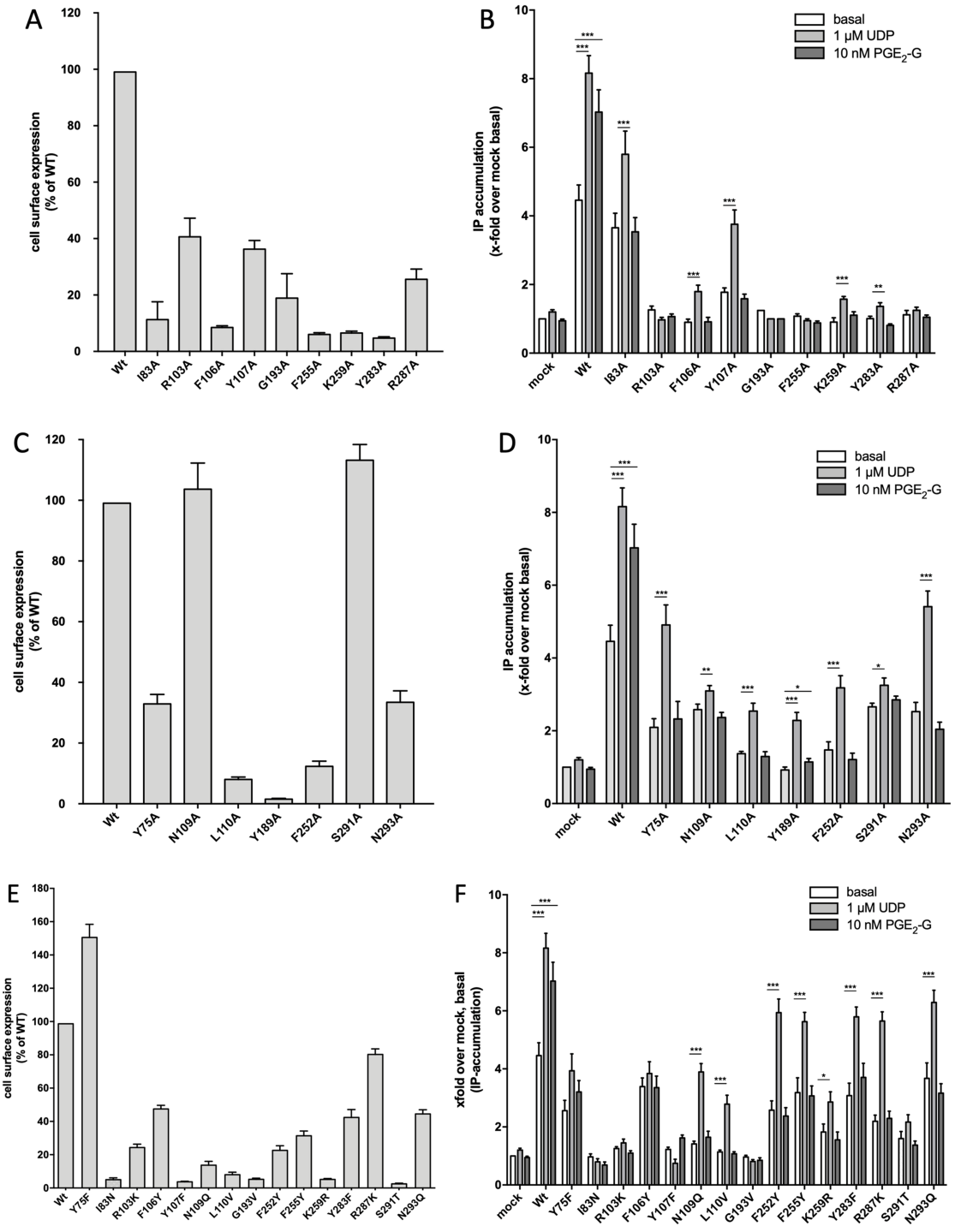


Figure 5.2: **UDP and PGE_2-G have overlapping agonist binding sites at P2Y6.** (A, B) Positions predicted to interact with both, UDP and PGE_2-G were individually mutated to alanine. (C, D) Positions predicted to preferentially interact with PGE_2-G but not with UDP were individually mutated to alanine. (E, F) Most positions mutated to alanine were also mutated to physicochemically related amino acids. HEK293T cells were then transfected with wildtype (Wt) and mutant P2Y6. (A, C, E) Cell surface expression of mutant P2Y6 receptors was determined as described. Optical density (OD) is given as percentage of P2Y6 Wt minus OD of mock-transfected cells. Data are given as means ± SEM of three independent experiments performed in triplicate. (B, D, F) Transfected HEK293 cells were stimulated with UDP (1 µM) and PGE_2-G (10 nM) and tested in IP1 accumulation assays as described. All data are means ± SEM of three to five independent experiments, each performed in triplicate. *p ¡0.05, **p ¡0.01, ***p ¡0.001 (paired Student's t test).

To study the functional relevance of the individual positions predicted to be involved in agonist binding, we performed mutagenesis studies changing the positions individually to Ala and testing the mutants in IP1 accumulation assays. First, we studied the positions proposed to be important to the agonism of in both ligands, UDP and PGE_2-G. Seven of the nine predicted positions are 100% conserved among vertebrates. The only exceptions are I83 and Y283, which are substituted by Val and Phe in some vertebrates. As shown in Figure 5.2A, the substitution of all investigated positions with Ala led to a reduction in receptor cell surface expression. Only R103A, Y107A, and R287A showed reasonable cell surface expression levels between 30-40% of the wildtype P2Y6. Testing the mutants in IP assays revealed that none of the mutants showed any response to PGE_2-G (Figure 5.2B). The mutants I83A, F106A, Y107A, K259A, and Y283A significantly responded to UDP but with extents that mainly correlated to their cell surface expression levels (Figure 5.2A/B). One exception was I83A displaying low cell surface expression but an almost unchanged basal activity and response to UDP. Considering only those mutants that appear at the cell surface to a significant amount, R103 and R287 participate in the agonistic activities of both ligands, whereas Y107 contributes only to PGE_2-G activity.

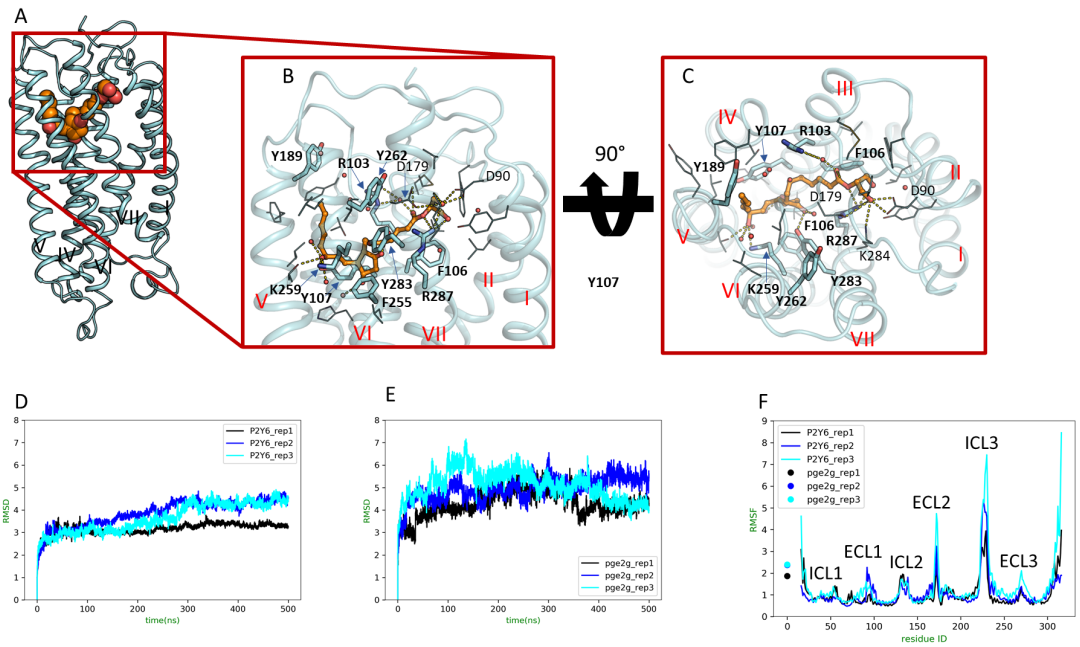


Figure 5.3: **P2Y6-PGE_2-G molecular dynamics-refined docked model.** PGE_2-G as docked to P2Y6 homology models, then the selected docked model was further refined with total of more than 2 µs of molecular dynamics. Lateral (A-B) and extracellular (C) views of the MD-refined model of PGE_2-G docked in the comparative model of the human P2Y6. Hydrogen bonds are indicated as dashed yellow lines, and sidechains of residues that are important to PGE_2-G activity are shown in sticks. Transmembrane helices (TM) are numbered from N- to C-terminal. 2D diagram of interactions of between the ligand and P2Y6 residues (D) was created using MOE (version 2020.09) (Vilar et al., 2008). Plots of RMSD to the starting docked model throughout the MD simulation (E) and per-residue RMSF after discarding the first 400 ns of the MD simulation (F).

Furthermore, all positions that were previously predicted to participate mainly in PGE_2-G binding (Figure 5.2A) were mutated to Ala and tested in cell surface and IP1 assays. The mutant Y75A, N109A, S291A, and N293A were expressed at the cell surface at detectable levels and above (Figure 5.3C). Y75A and N293A were still active upon UDP incubation but not in the presence of PGE_2-G (Figure 5.3D). N109A and S291A displayed half of the basal activity of the wildtype P2Y6 but were marginally activated by UDP. This data set revealed Y75 and N293 as residues that might involve in PGE_2-G but not in UDP agonist activation. Thus, N109 and S291 are necessary for directly or indirectly forming the binding site of both agonists.

5.3.3 Iterative refinement of P2Y6 models binding PGE_2-G and UDP.

For further model refinement, we constructed a new P2Y6 homology model in an iterative process(Bender et al., 2020) using an updated list of GPCR template structures (suppl. Table S1). We also performed docking of UDP and PGE_2-G with a new induced-fit docking protocol and incorporate the experimental restraints as the constraints to guide the positioning of the ligands during docking(Bender et al., 2019). By breaking down the Rosetta binding energy at the residue level, we examined the contribution of each residue to the interaction between the ligand and P2Y6 in each docking pose. We used the binding strength to prioritize the docking poses that encompass the favorable interactions between the ligand and the residues critical to the ligand activity according to mutagenesis results(Schüß et al., 2021b). Then, the selected docked models were subjected to conventional molecular dynamics (MD) for a total of 1.5 μs in three replicates to confirm kinetic stability of the observed binding poses. The structures are stable during the MD simulation as the RMSD, and the RMSF of both P2Y6 transmembrane helices and the ligand are in the reasonable ranges (Figures 3 and 4) Finally, we examined the involvement of different P2Y6 residues in the engagement of two agonists through the frequency of pairwise interactions between the ligand of the GPCR receptor. We calculated the relative contact strength (Figure 5) as the sum of atom pair interaction.

We selected final docking poses based on agreement with the experimental data and Rosetta interface energy score for both, UDP and PGE_2-G (suppl. Figure 5.S2, suppl. Table S2). Based on the co-crystal structures of P2Y1 with the antagonists MRS2500(Zhang et al., 2015a), and P2Y12 with its agonist ADP and ATP(Zhang et al., 2014a), we hypothesized that the corresponding positively charged residues in P2Y6, R103 (3.29) and R287 (7.39), form hydrogen bonds and electrostatic interactions to the group with the highest electron density of UDP and PGE_2-G, the diphosphate residue group and the glycerol ester group, respectively (Figures 3B-C and 4B-C). Our hypothesis is consistent with our mutagenesis data. Mutation of R103 and R107 to Ala abolished UDP activation of the mutant P2Y6 (Figure 5.2B), although the receptor mutants were still expressed at the cell surface (Figure 5.2A), indicating no gross structural alterations of the receptor. Similarly, the R103A and R287A mutants could not be activated by PGE_2-G (Figure 5.2B). Docking

PGE_2-G into the P2Y6 homology model revealed only one cluster where both R103 and R287 participate in ligand binding. In refined docked models, both Arg residues coordinate the glyceryl moiety and carbonyl oxygen of the ester form hydrogen bonds together with K284 (TM7), D179 (ECL2) , and D90 (TM1) (Figures 3B-C). To further characterize the relevance of these residues, we separately mutated both positions to lysine, which kept the positive charge but reduced the number of possible hydrogen bond donors (R103K, R287K). As shown in Figure 5.2F, both mutants are incompatible with PGE_2-G activation, but UDP still activated R287K. This discrepancy in agonism of those Arg to Lys mutants indicated that the glyceryl moiety of PGE_2-G might be benefited from alternately interacting with multiple hydrogen bond donors of R287. At the same time, UDP only needs to form a salt-bridge with a positively charged side chain at this position.

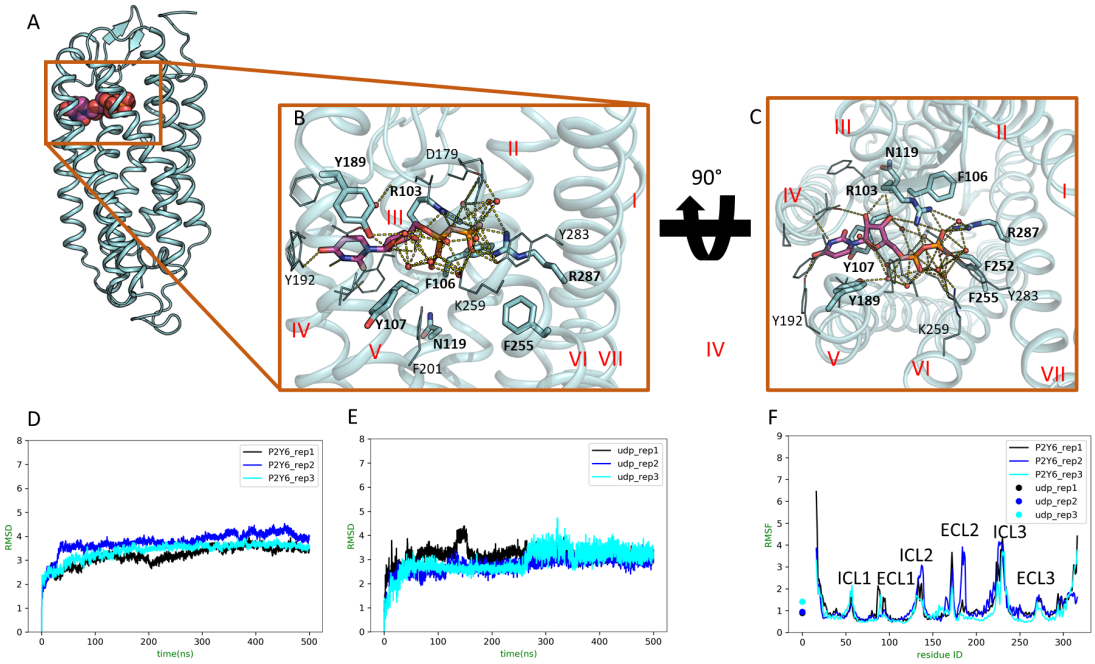


Figure 5.4: **P2Y6-UDP molecular dynamics-refined docked model.** UDP was docked to P2Y6 homology models, then the selected docked model was further refined with total of more than 2 us of molecular dynamics. Lateral (A-B) and extracellular (C) views of the MD-refined model of UDP docked in the comparative model of the human P2Y6. Hydrogen bonds are indicated as dashed yellow lines, and sidechains of residues that are important to UDP activity are shown in sticks. Plots of RMSD to the starting docked model throughout the MD simulation (D-E), and per-residue RMSF after discarding the first 100 ns of the MD simulation (F).

According to the computational models, while Y107 shields the binding pocket of PGE_2-G toward the cytosolic half of the receptor (Figure 5.3B), this residue, together with Y189, formed $\pi-\pi$ interactions and confined the movement of UDP's pyrimidine ring (Figure 5.4B). Additionally, the residue Y192 Here, hydrophobic interactions with the aliphatic backbone of the PGE2 and pyrimidine moiety of UDP, respectively, are possible (Figures 3 and 4). However, Y107 seems to be necessary only for activation with PGE_2-G.

Therefore, we also asked whether Y107 can be replaced by Phe only to keep the aromatic ring. As shown in Figure 5.2E, Y107F abolished cell surface expression of P2Y6 so that both agonists cannot activate the receptor (Figure 5.2F).

We also mutated other positions, which we had already mutated to alanine (Figure 5.2A-D), by changing them into more conservative mutations (Y75F, I83N, F106Y, N109Q, L110V, G193V, F252Y, F255Y, K259R, Y283F, N293Q). These mutants were functionally tested to check whether more distinct physicochemical changes are compatible with receptor functionality. I83N, Y107F, G193V, and S291T were purely expressed at the cell surface and showed no response to UDP or PGE_2-G (Figure 5.2E/F). Y75F showed a similar functionality as the wildtype receptor, however, with significantly lower IP1 responses to both agonists. In the model, this residue is located far below the binding site of both agonists (as viewed from extracellular) and, most likely, contributes indirectly to the formation of the binding pocket. In contrast, F106 is located in the model in the vicinity of both agonists, and mutation to Tyr abolished activation by both agonists. Still, the mutation failed to interfere with basal receptor activity and reduced cell surface expression only to 50% of the wildtype P2Y6 (Figure 5.2E/F). It is, therefore, likely that F106 contributes to the coordination of both ligands within the binding pocket. Our models suggest that this F106 is in contact with both ligands (Figures 3D, 4D, and 5A). N109, S291, and N293 cluster below the proposed bindings site are essential for either stabilizing the binding pocket or the downstream propagation of the activation pathway. Interestingly, the residue N293 is one of the four residues of the Na+ binding pocket switch, which has been shown to be essential for the activation mechanism of many other class A GPCR agonists(Liu et al., 2012; Zhou et al., 2019).

In both models, Y283 formed hydrogen bonds with the P2Y6 agonists (Figures 3B, 4B, and S2). Exchange of this residue with Phe only interferes with activation by PGE_2-G (Figure 5.2E). Thus, it is possible that the bound conformation of UDP can still be sufficiently stabilized with the hydrogen/salt bridge network between its diphosphate groups and close-by positively charged residues K25, R103, R287, and K259 (Figures 3B-D). Furthermore, F252 and F255 are directly located below Y283, probably forming a $\pi-$electron stack stabilizing Y283 in its position (Figure 5.3B). Mutation of both residues to Tyr retained UDP activation but abolished PGE_2-G induced receptor activity. These results indicate that PGE_2-G-mediated binding and/or activation depends on the correct orientation of this aromatic stack formed by Y283 (TM7), F252 (TM6), and F255 (TM6).

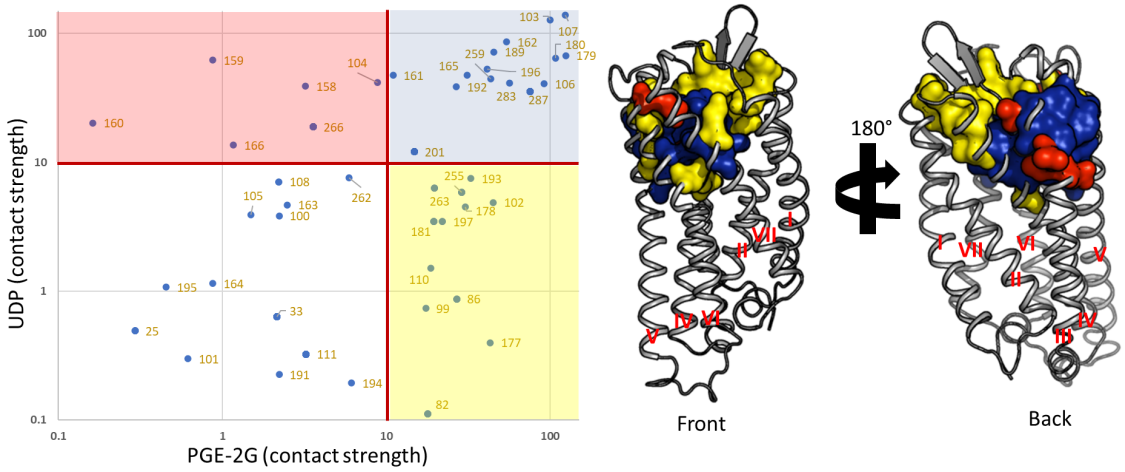


Figure 5.5: **Computed per-residue relative contact strengths of UDP and PGE_2-G to P2Y6 suggest overlapping binding pocket of those two agonists.** Relative contract strength is the sum of atom pair contact frequency between each agonist and P2Y6 residues. A relative contact strength of a particular residue is equal to 1 means that there is an atom from the ligand interact with a sidechain atom of the corresponding in all frames of the simulation, on average. A contact strength of 10 is considered to be significant. (A) Relative contact strength between P2Y6 residues and two agonists computed on the last threshold of 10 is marked in red lines on the plot. The x and y axis are shown in a log scale. Residues with significant interaction strength to both PGE_2-G and UDP and only to PGE_2-G are shown in blue and red dots, respectively. (B) Front and back view of the overlapped binding pockets of two agonists. Residues shared between both agonists are shown in blue, while the residue that only show strong relative contact strength to either UDP or PGE_2-G are colored in red and yellow, respectively.

To examine the overlapping area between the binding pockets of two agonists, we identified a list of P2Y6 residues that frequently interact with both ligands (relative contact strength of more than 10) during the last 400 ns of the MD simulations. Those residues are shown as dots on blue area on Figure 5.5A and blue surfaces in Figure 5.5B. The overlapping area of the binding pockets spans across the extracellular half of TM3, TM5, TM6, and TM7, and the tip of TM2, TM4, and TM5. Out of 15 identified common residues, three residues (R103, F107, and R287) were confirmed by mutagenesis studies. Eleven of the remaining twelve residues are in close proximity with those residues that were confirmed to be important for activation by both agonists. The only exception is Y189, as discussed above. A list of six residues that interacted with UDP more frequently than with PGE_2-G during the MD simulations were marked in dots on red surface and red surface on. We also identify a list of 12 residues that interacted with PGE_2-G more frequently than with UDP (yellow area and surface) (Figure 5.5A-B). These two residue lists implied in the binding pocket of PGE_2-G might expand to the tip of TM1, TM2, TM3, and TM4, and the core of TM6, while the binding pocket of UDP expand to the core of TM3 and TM5.

5.4 Discussion

We recently discovered PGE_2-G as an additional endogenous agonist for the human P2Y6(Bruser et al., 2017). It is not unusual that a given GPCR has more than one physiological ligand as it was shown, e.g., for the TSH receptor having TSH and thyrostimulin as agonists(Nakabayashi et al., 2002). In most cases of multiple agonism, the ligands are structurally or chemically related. However, in the case of P2Y6, the two agonists identified so far are chemically distinct, and the potencies differ by factor 50,000(Bruser et al., 2017). This difference suggests that P2Y6 integrates distinct physiological signals related to immune functions(Koizumi et al., 2007; Le Duc et al., 2017; Li et al., 2014; Warny et al., 2001) and pain(Hu et al., 2008; Khasabova et al., 2019). Furthermore, P2Y6 orthologs exist in all vertebrates investigated (Figure 5.1A), but only 30% of all amino acid positions are fully conserved (Figure 5.1B). Therefore, the question is whether agonist preferences were also conserved during evolution. Here, we demonstrated that PGE_2-G- and UDP induce activation of mammalian, bird, reptile, and fish P2Y6 orthologs (Figure 5.1C/D). This evolutionary conservation is well in line with the fact that both P2Y6 and COX-2 are preserved in vertebrates. Our data, therefore, suggest that P2Y6 kept its dual agonist specificity for both PGE_2-G and UDP during the entire vertebrate evolution.

Functional data suggest that UDP and PGE_2-G have an overlapping agonist binding site(Bruser et al., 2017). However, in functional assays (Figure 5.1D), the Emax values of PGE_2-G are lower than UDP, suggesting a potentially different binding mode at P2Y6 as seen for partial agonists(Yao et al., 2006). To further address whether UDP and PGE_2-G share a binding pocket or bind at two separate sites, we used a homology model of the human P2Y6 and performed computer-aided ligand docking to predict the binding mode of both agonists.

The predicted binding pockets of UDP and PGE_2-G revealed shared and specific determinants for ligand orientation (suppl. Figure 5.S1). Mutation of these residues to alanine and experimental testing of these mutants (Figure 5.2) identified only a few positions that could be adequately evaluated because of sufficient cell surface expression (>25% of the wildtype). It should be noted that we failed to perform saturation binding assays and concentration-response curve experiments, methods that are usually engaged for detailed characterization of the mutants. UDP and PGE_2-G are unsuitable for radioligand-binding studies because of high background noise due to nucleotide binding to many cellular targets and PGE_2-G's very lipophilic nature, respectively. Furthermore, performing concentration-response curves with UDP in the used heterologous cell system is limited because UDP concentrations > 10 µM produced an endogenous signal in IP1 assays (data not shown). With these limitations, we found two categories: i) loss of activation by both agonists and ii) loss/strong reduction of activation by PGE_2-G. Except for the previously characterized mutant Y262A

(Bruser et al., 2017), we did not identify any other mutation that caused a loss of UDP activation but not PGE_2-G agonism. It, therefore, seems that PGE_2-G mainly occupies most of the UDP binding side but recruits additional interaction partners.

Regarding the evolutionary aspect, Y75, R103, N109, Y262, R287, S291, and N293 are fully conserved among vertebrate P2Y6 orthologs indicating their structural and functional importance as suggested in our docking models. R103 (R3.29), Y283 (Y7.35), and R287 (R7.39) are conserved in the P2Y1-like receptor subgroup but not in the P2Y12-like receptor subgroup (Figure 5.1A). Those observations are in line with the fact that, within the crystal structures of the ADP-bound P2Y1 and P2Y12 receptors, the agonist binding sites significantly differ between both receptors(Zhang et al., 2015a, 2014a,b). In P2Y12-like cluster 2, the respective positions are S/A3.29, K7.35, and L7.39. N109, S291, and N293 are also found in other receptors shown in Figure 5.1A at the corresponding positions indicating more general structural functions. Our new mutagenesis data residues reported that residues, such as Y75 and N293, mediated only the agonism of PGE_2-G with P2Y6. Interestingly, both of those residues did not form significant contacts to either ligand based on the models. Furthermore, N293 is one of the four residues that constitute the Na+ binding pocket, whose repack switching is essential to the activation of many other class A GPCRs(Zhou et al., 2019). For the P2Y6 mutants with sufficient cell surface expression, most conserved residues either directly interact directly with both ligands (R103 and R287) or indirectly stabilize the PGE_2-G's binding pocket (Y75, N109, S291, N293).

Docking studies and MD simulations provided a potential atomic-detail explanation to our mutagenesis results. The modeling data suggest that the diphosphate group of UDP and the glycerol ester moiety of PGE_2-G form a hydrogen interaction network with two positively charged residues R103 and R287 (Figures 3 and 4), and nearby residues such as K259 and Y283. Furthermore, the pyrimidine ring of UDP and the $\omega-$lipophilic chain of PGE_2-G form hydrophobic interaction with Y107. Unfortunately, mutating K259 and Y283 to Ala caused the P2Y6 cell expression insufficient to reliably detect their effects on the activation of UDP and PGE_2-G. However, the importance of R103, R287, and Y107 in the agonistic activity of UDP and PGE_2-G was confirmed with mutagenesis (Figure 5.2). Our refined models suggest that these two P2Y6 agonists have partially overlapped binding pockets, stretching from the extracellular half of TM3, TM5, TM6, and TM7, to the tip of TM2, TM4, and TM5 (Figure 5).

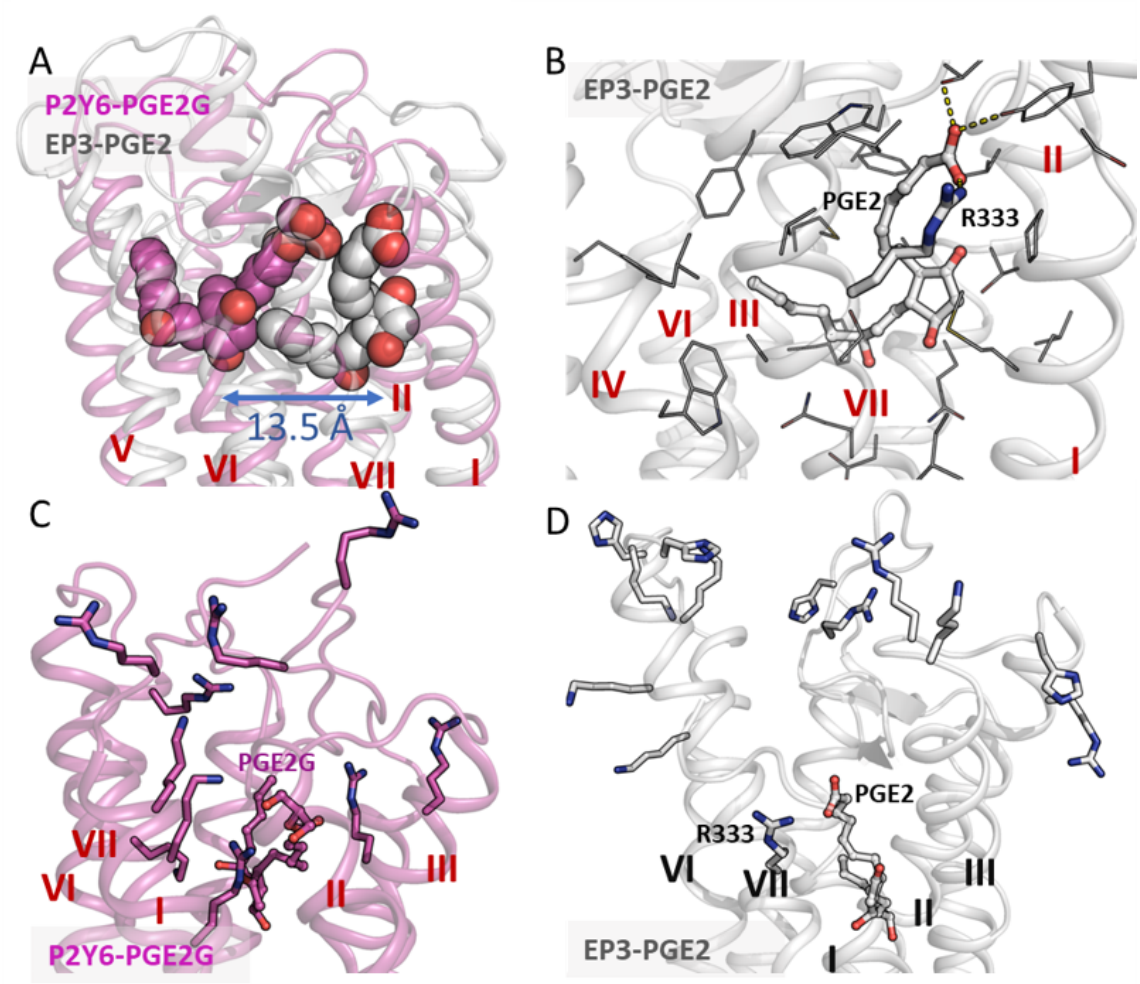


Figure 5.6: **Comparison between the P2Y6-PGE$_2$−G and the EP3-PGE2 complexes (PDB ID: 6AK3).** (A) Overall view of position of the binding pockets of PGE$_2$−G (grey spheres) and PGE2 (magenta spheres). The ring of PGE$_2$−G shifts around 13.5Å toward the TM5 compared to that of PGE2. (B) A close-up look of EP3 residues (grey lines) that interacted with PGE2 (grey balls and sticks). The only positively charged sidechain, R333, that located close to the ligand was also shown in grey balls and sticks. (C-D) The transmembrane region of the P2Y6 (C-magenta) has significantly more positively charged sidechains (shown in sticks) than does that of EP3 (D-grey), enabling the shift and elongation of the binding pose of PGE$_2$−G (magenta balls and sticks). In contrast, the only positively charged residue in the extracellular half of the transmembrane region of EP3 is R333 (grey sticks) on TM7.

Interestingly, our proposed agonist engagement modes of P2Y6 resemble the general activation mechanism previously proposed from the crystal structures of P2Y12(Zhang et al., 2014a) than from a structural study on P2Y1(Yuan et al., 2016). More specifically, ADP's negatively charged diphosphate group forms hydrogen bonds/salt bridge residues on the extracellular half of both P2Y1 and P2Y12 receptors, stabilizing the proximity between TMs 3-4 and TMs 6-7 (the "closed state"), which is consistent with our models. However, unlike the agonist engagement mechanism of P2Y1 proposed by Yuan et al.(Yuan et al., 2016), the binding pocket of P2Y6 agonists in our models locate deeper toward the core of the intermembrane helices. Hence,

UDP and PGE_2-G maintain the "closed state" of P2Y6 while being buried inside the helical bundle, similar to the binding mode of ADP and P2Y12. Further structural studies are needed to confirm our observation regarding the similarity and differences in agonistic activation mechanism among P2Y6, P2Y1, and P2Y12. Furthermore, compared to the structures of prostaglandin EP receptors (EP2, EP3, and EP4) with PGE2, an analog of PGE_2-G, the binding pocket of PGE_2-G shifts around 13.5Å toward TM4/5 (Figure 5.6A). This might be due to there are substantially more positive charged residues in the extracellular half of the transmembrane region of P2Y6 than in that of EP receptors (Figure 5.6 B-D). In fact, since the sequences similarity and identities were so low (Table 6) that we did not include any of the EP receptors as template to reconstruct homology model of P2Y6.

In conclusion, P2Y6 is the target of two endogenous agonists, UDP and PGE_2-G, since over 450 million years of vertebrate evolution. Two polar residues, R103 (R3.29) and R287 (R7.39), interact to phosphate and glyceryl moieties of those two ligands, respectively. In contrast, the pyrimidine and PGE2 moieties interact with a more hydrophobic environment of the ligand-binding site. Via this shared binding pocket, P2Y6 could integrate different chemical signals into a Gq/11 protein-mediated intracellular signal transduction. To extend the understanding of P2Y6 activation beyond our currently known uniform IP1 and Ca2+ responses, future studies should investigate these two agonists gradually differ in their induced signal transduction, e.g., in the kinetics of signaling or by recruiting other G proteins and arrestins.

5.5 Methods

5.5.1 Materials

If not stated otherwise, all chemicals were purchased from Sigma-Aldrich (Germany), and cell culture materials were provided by Life Technologies GmbH (Germany).

Generation of receptor constructs. cDNA from H1819 cells was used to amplify and clone the human P2Y6 coding sequence(Bruser et al., 2017). In addition, genomic DNA from chicken, alligator, and zebrafish was used to amplify the respective coding sequences of P2Y6. All sequences were double-tagged with an N-terminal HA epitope and a C-terminal FLAG epitope and, for transient transfection, introduced into the mammalian expression vector pcDps(Sangkuhl et al., 2002). All mutant constructs were generated by a PCR-based site-directed mutagenesis and fragment replacement strategy. All constructs were verified by sequencing.

Cell culture, transfection, measurement of intracellular inositol phosphates. For functional assays, receptor constructs were heterologously expressed in human embryonic kidney (HEK293T) cells upon transient transfection. Cells were grown in DMEM/F12 supplemented with 10 % FBS, 100 units/ml penicillin, and 100 µg/ml streptomycin at 37°C and 5 % CO2

An indirect cellular ELISA was used to estimate cell surface expression of heterologously expressed receptors carrying an N-terminal HA tag(Schoneberg et al., 1998). According to the manufacturer's protocol, to measure IP1, HEK293T cells were split into 96-well plates (20,000 cells/well) and transfected with 100 ng vector constructs using Lipofectamine® (Invitrogen). Empty vector (mock) served as a negative control. Then, 48 h after transfection, cells were stimulated 30 min at 37 °C with 35 µl 1x IP1 stimulation buffer (Cisbio) containing the respective reagents. Next, cells were lysed by adding 30 µl lysis buffer (Cisbio) per well and kept frozen at -20 °C until measurement. IP1 measurements using the Cisbio IP-one Tb kit (Cisbio, Codolet, France) were performed in ProxiPlate-384 Plus microplates (Perkin Elmer) with the EnVision Multilabel Reader (Perkin Elmer). The assays were performed with a final concentration of 1 %DMSO.

5.5.2 Generation of P2Y6 comparative models

A comparative model of P2Y6 was constructed using the protein structure prediction software package, ROSETTA version 3.12(Bender et al., 2016; Leaver-Fay et al., 2011), using multiple GPCR templates(Bender et al., 2020). The X-ray crystal structures of P2Y1 and P2Y12 (Protein Data Bank ID: 4xnw, 4ntj)(Zhang et al., 2015a, 2014b) were chosen as main templates based on high sequence similarity to P2Y6. To increase conformational sampling, these templates were supplemented with protease-activated receptors (PARs) PAR1 and PAR2 (3vw7 and 5ndd)(Cheng et al., 2017; Zhang et al., 2012), angiotensin II type I and type II ATI and ATII (6do1 and 5ung)(Wingler et al., 2019; Zhang et al., 2017), kappa opioid receptor (6b73)(Che et al., 2018), free fatty acid receptor (FFAR) 1 (5tzr)(Lu et al., 2017), platelet-activating factor receptor (PAFR) (5zkp)(Cao et al., 2018), and endothelin B receptor (ETBR) (6igk)(Shihoya et al., 2018). The information of the name and PDB IDs, as well as the sequence identity and similarity to P2Y6, are summarized in the suppl. Table S1. An initial sequence alignment of 11 GPCR receptors was created using the GPCRdb structure-based sequence alignment application(Kooistra et al., 2021). Adjustments were then made to ensure that all secondary structure elements were properly aligned while moving significant gaps to loop regions. In addition, the first 15 and last 12 residues of the P2Y6 sequence were truncated as they are not crucial for the binding of the ligands(Bruser et al., 2017).

After assigning coordinates to P2Y6 residues from each template alignment using Rosetta's partial-thread application, RosettaCM(Song et al., 2013b) 'hybridizer' was used to combine segments across all templates in a metropolis Monte Carlo with a simulated annealing approach to arrive at energetically favorable compositions. In brief, RosettaCM exchanges template fragments into a starting model to achieve energetically favorable hybrid template models. Any residues still lacking coordinates were modeled de novo using 3mer and 9mer fragments extracted from the PDB fragment database. Transmembrane segments, as predicted using the OCTOPUS server(Viklund and Elofsson, 2008) and adjusted to match with the transmembrane spans

of theP2Y1 and P2Y12 helices according to the calculation made by the PPM server(Lomize et al., 2012), were modeled within Rosetta's implicit membrane potential(Yarov-Yarovoy et al., 2006). The resulting full sequence models were subjected to eight iterative cycles of sidechain repacking and gradient minimization within the membrane potential. P2Y6, P2Y1, and P2Y12 share a conserved disulfide bond between the N-terminal C18 and C273 in extracellular loop 3(Deflorian and Jacobson, 2011). Therefore, disulfide bond constraints were introduced between these residues as well as C99 and C177. Secondary constraints were also applied to the extracellular loop 2 (ECL2) of P2Y6 models so that its beta-hairpin structure is maintained during loop modeling. In total, 20,000 P2Y6 homology models were generated. The top 10% of all generated models by pose score were clustered by Cα RMSD using K-means clustering into eight clusters. The top ten scored models from each of those eight clusters were selected for docking.

5.5.3 Rosetta ligand docking

Ligand docking into the comparative model of P2Y6 with UDP and PGE_2-G was performed with Rosetta Ligand(Lemmon and Meiler, 2012; Meiler and Baker, 2006). One hundred conformations of PGE_2-G and 100 conformations of UDP were generated with BCL::Conf(Mendenhall et al., 2021). This application builds small-molecule conformations from substructures seen in experimentally elucidated structures. A starting position was selected for both ligands based on the average of ligands present in all GPCR templates. The induced-fit docking protocol started with an initial docking round with high constraint weight to penalize the ligand placements that were far away from the residues deemed important to the molecule's activity according to the mutagenesis data. Then, another round of relaxing the backbone of the residues surrounding the ligand to mimic the induced fit effect, and a final refinement docking with low constraint weight to optimize the ligand-receptor atomic interactions. The docking protocol included a low resolution (centroid mode) phase consisting of 500 cycles sampling ligand conformers in 4 Å translation search and complete reorientation search, which are constraints by preset distance-based constraints from the mutagenesis results, and a high-resolution phase consisting of six cycles of sidechain refinement with small perturbations of ligand poses and conformation. During the refinement phase, the translation search was reduced to 1 Å, and the constraint weigh score was reduced to 1. This phase finds an energetically favorable pose by combining minor ligand conformational flexibility with sidechain refinement simultaneously. For each ligand, the top 10% models by interface delta score were collected in each of three rounds of induced fit docking. Those top models were then clustered, and the top 10 models were selected for the next round of docking. The Rosetta interface scores versus ligand RMSDs graphs after the final round of induced fit docking are shown in the suppl. Figure 5.S2.

For each selected pose cluster, a $\Delta\Delta G$ value, the change in free energy with and without ligands bound to

P2Y6, was calculated for each residue in the receptor. A binding strength score, which measures the linear sum of $\Delta\Delta G$ of residues that are favorable and unfavorable to the activity of the ligands, was calculated as binding strength $= \sum(\Delta\Delta G_{nni}) - \sum(\Delta\Delta G_i)$, where $\Delta\Delta G_{nni}$ is the computed Rosetta $\Delta\Delta G$ value for the residues that are not important for the ligand activity based on the mutagenesis data and also have a negative $\Delta\Delta G$ value. $\Delta\Delta G_i$ is the computed Rosetta $\Delta\Delta G$ value of the residues that were shown experimentally to affect the activity of the ligands. Essentially, the binding strength score measures the relative agreement between a particular docking pose and the mutagenesis data(Schüß et al., 2021b). The supplementary table S4 shows the computed average per residue $\Delta\Delta G$ values for each cluster. For UDP docking models, the pose cluster with highest binding strength score was be selected.

5.5.4 Molecular simulation and anlysis

MD simulation of the selected docking models. Selected docking models were then refined with molecular dynamics simulation. For each ligand, two independent replicates with around two µs in total simulation time were conducted. All membrane systems were built with the membrane building tool PackMol-Memgen(Schott-Verdugo and Gohlke, 2019). Downser++(Morozenko and Stuchebrukhov, 2016) were then used to dock waters inside the transmembrane region of P2Y6 in the presence of the ligands. The bimembrane system contained POPC and Cholesterol with a molecule number ratio of 10:1. Proteins, lipids, TIP3P water, and ions were modeled with the FF19SB(Maier et al., 2015) and Amber Lipid17(Gould et al., 2018) force fields, and the ligands were modeled with the GAFF2 small molecule force field(He et al., 2020; Wang et al., 2004). A TIP3P water layer of 25 Å was included, and $Cl-$ or $K+$ ions were added to neutralize the charge of the system. Each bilayer system was first minimized for 5,000 steps using steepest descent followed by 15,000 steps of conjugate gradient minimization. During heating, the protein backbone and sidechain atoms, lipid and water were restrained to their starting coordinates with harmonic force constants of 10 and 5, heated to 10 K over 1,000 steps with a step size of 1 fs using constant boundary conditions and Langevin dynamics with a rapid collision frequency of 10,000 ps^{-1}. The system was then heated to 100 K over 50,000 steps with constant volume dynamics and the collision frequency set to 1000 ps^{-1} and, finally, to 303 K over 100,000 steps with constant pressure dynamics and anisotropic pressure scaling turned on, while the positional restraints on the system were gradually removed. The system was then run with the protein-complex held fixed for another one ns at 303 K. Production MD was conducted for 500 ns at 303K using a step size of 4 fs with hydrogen mass repartitioning(Hopkins et al., 2015), constant pressure periodic boundary conditions (NPT system), semi-anisotropic pressure scaling, and Langevin dynamics. MD trajectories were analyzed using CPPTRAJ (version 18.0) and PTRAJ (version 2.0.2.dev0)(Roe and Cheatham, 2013), as well as VMD (visual molecular dynamics; version 1.9)(Humphrey et al., 1996). The first 100 ns of the simulation

was removed before we performed the calculation of RMSF, atomic contact, and Molecular Mechanics with a Poisson-Boltzmann/Surface Area solvent (MM-PBSA). Relative contract strength is calculated as the sum of atom pair contact frequency between the agonist and each P2Y6 residue. The values of per-residue relative contact strengths are listed in the supplementary table 5.S3.

5.6 Acknowledgements

We thank Katja Ettig for excellent technical assistance. This work was supported by the German Research Foundation CRC1423 project number 421152132 (TS, JM), and XXX (Vanderbilt University). Anne Zimmermann was supported by the MD program of the Medical Faculty, University Leipzig.

5.7 Author contributions

A.Z., A.B. performed the experiments. O.V. performed the generation of a receptor homology modeling, ligand docking, and molecular dynamic simulation. A.B., A.Z., O.V., J.M., T.S. analyzed the data. A.B., L.J.M., O.V., G.S., J.M., T.S. designed the study and wrote the paper with contributions from all authors.

5.8 Supplementary information

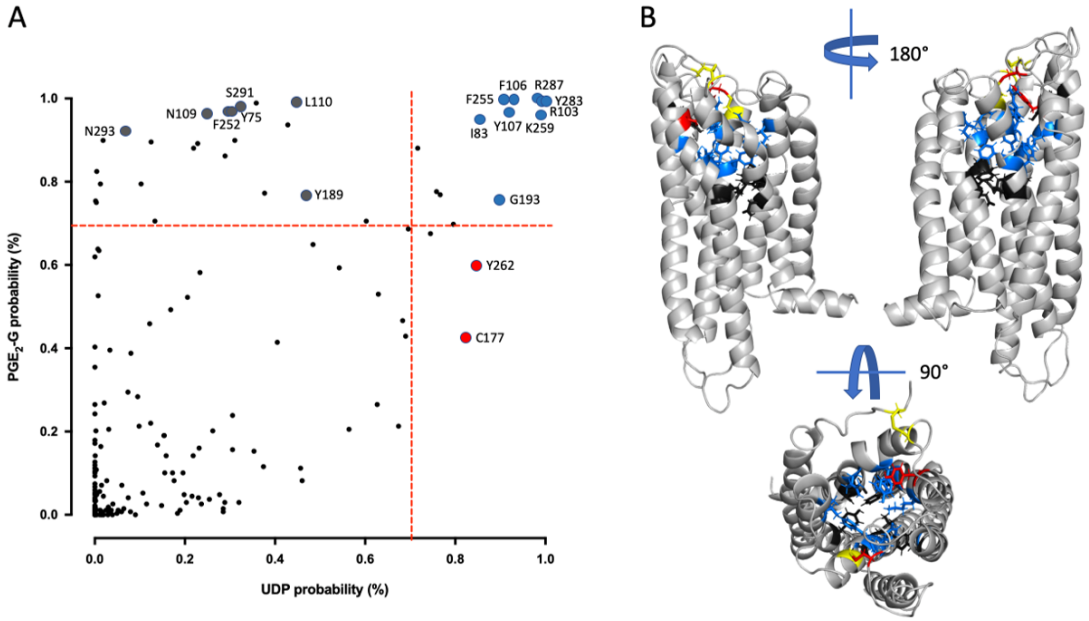


Figure 5.S1: **Prediction of P2Y6 residues involved in UDP- and PGE2-G binding.** (A) The A homology model of the human P2Y6 was generated and ligand docking of both agonists, UDP and PGE2-G, were performed. Based on these docking studies (Bruser et al., 2017) the probability to interact with every residue of P2Y6 was calculated and plotted. (B) The model of the human P2Y6 is shown where residues private for PGE2-G and UDP are shown in black and red, respectively. Residues shared between both agonists are shown in blue. The conserved cysteine bridges connecting the N terminus with ECL3 and between ECL1 and ECL2 are depicted in yellow.

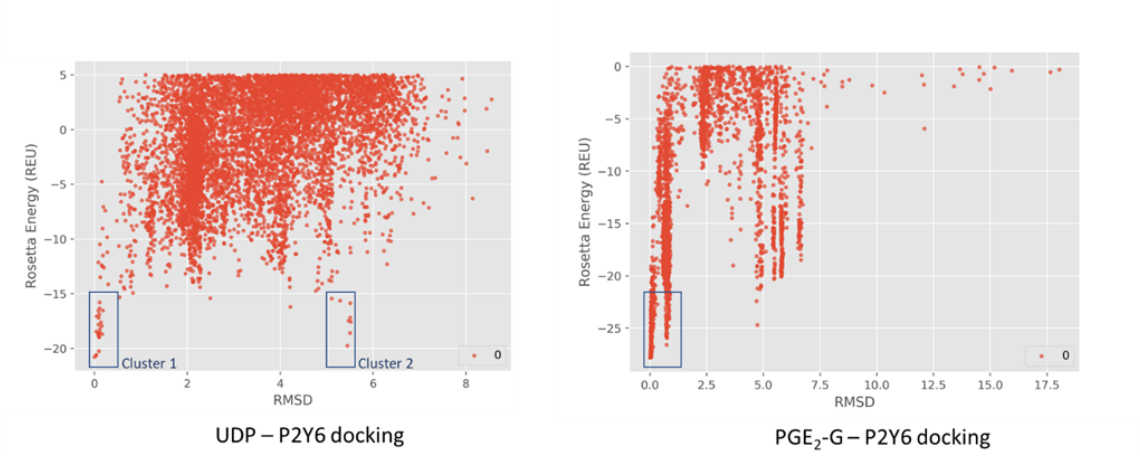


Figure 5.S2: **Docking interface score vs RMSD plots of output models from the final round of the induced-fit docking of PGE2-G and UDP to P2Y6 homology models.** P2Y6 homology models were built with the RosettaCM protocol. The induced-fit docking experiment includes initial placement of the ligand into the potential binding pocket, relaxing the backbone and sidechain of interaction P2Y6 residues with the presence of the ligand inside the binding pocket, and a final refinement docking round. For each ligand, clusters of docked models were selected from the final refinement docking round (positioned within the blue rectangles in the graphs) for further $\Delta\Delta G$ analysis, and MD simulation.

Name	% sequence similarity	% sequence identity	pdb xtal structure ID	State	resolution	GPCR class
P2Y1 receptor	53	38	4xnv	Intermediate	2.2	A
PAR2	42	27	5ndd	Intermediate	2.8	A
AT1 receptor	45	27	6do1	Active	2.9	A
PAR 1	41	26	3vw7	Intermediate	2.2	A
K receptor	40	25	6b73	Active	3.1	A
FFA1 receptor	35	24	5tzr	Intermediate	2.2	A
AT2 receptor	38	24	5ung	Active	2.8	A
PAF receptor	40	23	5zkp	Active	2.8	A
P2Y12 receptor	37	22	4pxz	Intermediate	2.5	A
ETB receptor	36	21	6igk	Intermediate	2	A

Table 5.S1: **Ten structures of GPCR class A that were used as the templates to build the homology models of P2Y6.** Comparative models of P2Y6 was constructed using the protein structure prediction software package, ROSETTA version 3.12 (Bender et al., 2016; Leaver-Fay et al., 2011), using multiple GPCR templates Robertson et al. (2018). The sequence similarity and identity percentages of 11 GPCR receptors to P2Y6 were calculated using the GPCRdb structure-based sequence alignment application (Kooistra et al., 2021).

Residue	model ID	UDP ΔΔG values	
		Cluster 1	Cluster 2
75	60	0	-0.59
83	68	0	0
103	88	-0.42	0.73
106	91	0.23	-0.68
107	92	-0.91	-0.24
109	94	0	0
193	178	-0.07	0
252	237	0	0.7
262	247	0	0.05
287	272	-3.28	-2.48
291	276	0	-0.07
293	278	0	0
	Strength scores	2.63	2.28

Residue	model ID	PGE-2G ΔΔG values
75	60	0
83	68	0
103	88	-5.5225
106	91	0.0024
107	92	-1.7071
109	94	0
193	178	-0.0526
252	237	0
262	247	-0.1374
287	272	-1.8945
291	276	0
293	278	0
	Strength scores	9.3117

Table 5.S2: **$\Delta\Delta G$ analysis of docking model clusters.** (Right – PGE2-G; Left – UDP). For each selected docking pose cluster (see Figure S2), a $\Delta\Delta G$ value, the change in free energy with and without ligands bound to P2Y6, was calculated for each residue in the receptor. A binding strength score, which measures the linear sum of $\Delta\Delta G$ of residues that are favorable and unfavorable to the activity of the ligands, was calculated as binding strength$= \sum \Delta\Delta G_{nni} - \sum \Delta\Delta G_i$, where $\Delta\Delta G_{nni}$ are the computed Rosetta $\Delta\Delta G$ values for the residues that are not important for the ligand activity based on the mutagenesis data and also have a negative $\Delta\Delta G$ value. $\Delta\Delta G_i$ is the computed Rosetta $\Delta\Delta G$ values of the residues that were shown experimentally to affect the activity of the ligands. Essentially, the binding strength score measures the relative agreement between a particular docking pose and the mutagenesis data.

Energy terms	P2Y6 – UDP complex			P2Y6 – PGE2G complex		
VDWAALS	-29.1272	-29.7626	-31.9360	-61.5590	-50.9681	-56.6685
EEL	-15.1285	-12.0547	-13.4128	0.9327	0.3908	1.0425
ENPOLAR	-3.7993	-4.0941	-3.6725	-5.8512	-5.7313	-5.8455
DELTA TOTAL	**-48.0550**	**-45.9113**	**-49.0213**	**-66.4775**	**-56.3086**	**-61.4715**

Table 5.S2: **Calculation of the free energy of binding (DELTA TOTAL) of the complexes between P2Y6 and its two agonists** (UDP and PGE2-G) using MM-PBSA (without entropy calculation and with membrane related options turned on) over the last 400ns of three MD replicates. The energy terms VDVWAALS, EEL, and ENPOLAR are the energy contribution of non-bonded van der Waals, electronstatic, and Non-polar solvation energy to DELTA TOTAL, respectively.

Similarities are on the lower-left side of the table, and identities on the upper-right.

	1	2	3	4
1. [Human] EP_2 receptor	-	27	30	15
2. [Human] EP_3 receptor	40	-	28	15
3. [Human] EP_4 receptor	48	44	-	17
4. [Human] $P2Y_6$ receptor	27	27	27	-

Table 5.S2: **Sequence identities and similarities among P2Y6, EP2, EP3, and EP4 on helices and three extracellular loops.** The similarities are on the lower-left side of the table, and the identities are on the upper-right side. The table was generated using the similarity matrix application of the GPCRdb server.

CHAPTER 6

Conclusion

6.1 Summary and Implication

G protein-coupled receptors (GPCRs) represent the largest membrane protein family and a significant target class for therapeutics. Receptors from GPCRs' largest class, class A, influence virtually every aspect of human physiology. About 45% of the members of this family endogenously bind flexible peptides or peptides segments within larger protein ligands. While many of these peptides have been structurally characterized in their solution state, the few studies of peptides in their receptor-bound state suggest these peptides interact with a shared set of residues and undergo significant conformational changes. For the purpose of understanding binding dynamics and the development of peptidomimetic drug compounds, further studies should investigate peptide ligands complexed to their cognate receptor.

Out of four classes of GPCR—A, B, C, or F—Class A is the largest and most diverse group in humans. This subfamily has been investigated most extensively in drug discovery due to their available structural and experimental data. They conform with the common GPCR structural fold, such as a seven-transmembrane (7TM) helices domain, three extracellular loops, and three intracellular loops with ligand-binding pockets and a G-protein-binding region located in the extracellular and intracellular ends of the helix bundle, respectively (Isberg et al., 2017). The variety of drugs targeting GPCRs reflects the diversity of chemical signals that can be transduced by GPCRs, including small molecules, lipids, ions, and proteins (Bockaert and Pin, 1999; Wacker et al., 2017). In particular, according to the data from the GPCRdb server (Isberg et al., 2017), the peptide- and protein-activated receptors are found to account for about 46% of all class A GPCRs in human. For this review, we consider GPCRs that recognize classical peptides and peptide-like segments within larger protein domains to belong to the same category of receptors. Peptide-activated receptors are found across all rhodopsin-like subfamilies (α, β, γ, and δ) and the entire secretin family (Fredriksson et al., 2003). Given this coverage, it is unsurprising that many of the blockbuster drugs mentioned above (e.g., olmesartan, buserelin, and valsartan) target members of this receptor group. With such importance for therapeutic development, a full understanding of the structural and dynamical determinants of signaling for these molecules is necessary. This review covers what is known about these receptors structurally using various biophysical techniques and provides suggestions for future discovery routes. Peptide-binding GPCRs represent nearly a quarter of the druggable human GPCR superfamily. Our analysis discovered a common set of 14 residues that were shown to interact with peptide ligands among all available co-crystal structures. This shared binding site suggests a

potential general pattern in peptide engagement among class A GPCRs.

Obesity has been the fastest growing health problem in America. According to the report from the National Health and Nutrition Examination Survey (2013-2014), more than 2 in 3 adults were overweighed or obese. Obese patients are at greater risk for heart disease, type II diabetes, high blood pressure, stroke, cancer, and osteoarthritis (Ogden CL, 2015). The economic burden of the obesity epidemic is also staggering. The estimated health care costs could range from $147 billion to $210 billion per year. Furthermore, lower productivity can cost em-ployers about $500 per obese employees per year(Gates et al., 2008).

Neuropeptide (NPY), peptide YY (PYY), and pancreatic peptide (PP) belong to a hormone family that regulate feeding behaviors and energy expenditure. These peptides bind and activate the Y receptors, class A G-protein coupled receptors (GPCRs), to elicit their bioactivities. In human, there are four Y receptor subtypes: Y1, Y2, Y4 and Y5 with a wide range of affinity to NPY and NPY-like peptides(Lindner et al., 2008a). Understanding the molecular mechanisms of this multi-receptor/multi-ligand system is essential to elucidate their physiological effects. activation of Y4 by PP has been verified to promote postprandial satiety in precedent literatures. Reduction if food intake and body weights were observed after administration of PP into normal mice, but not in Y4-knockout mice (Balasubramaniam et al., 2006), proving the critical roles of Y4 in these effects. Furthermore, intra-peritoneal injection of BVD-74D, an Y4 agonist, reduce feeding activities and weight gain with a dose dependent fashion in normal mice, but not in Y4-knockout mice(Wren and Bloom, 2007). Those evidences proved potential of Y4 as target for the treatment of obesity.

The neuropeptide Y4 receptor (Y4) is a 375 amino acid G-protein coupled receptor (GPCR) that is expressed mainly in peripheral tissues and the brain stem(Lindner et al., 2008b; Lundell et al., 1995; Schwartz, 1983). Y4 is the only receptor subtype with low affinity for NPY and PYY and high, picomolar affinity for PP(Cabrele and Beck-Sickinger, 2000). Thus, selective agonists of Y4 could be promising candidates for obesity therapeutics(Wren and Bloom, 2007). In fact Obinepitide (TM-30338), a variant of PP and PYY, is currently in phase II clinical trials as a treatment for obesity. However, as Obinepitide is a peptide, issues of stability and bioavailability remain. The objective of the present drug discovery project is to develop these confirmed hit compounds into small molecule probes of the Y4 receptor and leverage these molecules to understand mechanism of action, ligand bias, and determinants of selectivity within the Y receptor family. Allosteric modulators of GPCRs have a higher chance to be selective as allosteric binding sites tend to be evolutionary less conserved between receptor subtypes(**?**). The therapeutic potential of allosteric modulators is further increased by their ability to tune the receptor re-sponse instead of simply turning it on or off. Side effects may also be reduced for allosteric potentiators be-cause the therapeutic only acts when the receptor is engaged by its native ligand. Therefore, we focused our efforts in this dissertation on these allosteric molecules as identification of small molecule allosteric modulators of Y4 will allow future development of

pharmacological probes and could eventually seed a drug discovery program in obesity.

The investigation of the allosteric binding site via mutagenesis of the Y4 represents another milestone in the understanding of small molecule effects on this receptor with respect to the molecular mechanism and binding mode. To achieve this, the results of experimental studies on Y4 mutants was be combined with computa-tional modeling of the Y4 in and docking of the allosteric modulators in an iterative approach. This resulted in structural models that enable the development and structure-based optimization of novel compounds. Our proposed binding pose of (S)-VU0637120, a Y4 partial NAM, highlighted patterns of locations among allosteric pockets across different GPCR class A subfamilies despite their low sequence identity. Looking at a broader picture, this study demonstrates that targeting this allosteric binding pocket holds promise for the development of novel, selective compounds as tools for the investigation of peptide GPCRs, which represent about 15% of all GPCRs (Wu et al., 2017). However, peptide ligands that activate peptide GPCRs often face many challenges with dosing and pharmacokinetic properties such as rapid clearance, poor metabolic stability, and consequently poor bioavailability, as well as weak membrane permeability (Craik et al., 2013). Additionally, it is very challenging to develop selective peptide ligands within multiligand/multireceptor systems. Such difficulties have been noted with PP, the native peptide ligand of the Y4R, which possesses a half-life of about 7 min (Adrian et al., 1978) and shows a high affinity to the Y5R (Gerald et al., 1996). Within this study, we were able to identify and characterize the first small molecule, allosteric antagonist (S)-VU0637120 that selectively activates the Y4R with no effect on the evolutionarily-related Y1R, Y2R and Y5R subtypes. We also obtained a model for the (S)-VU0637120/Y4R complex that not only agrees with data from systematical screening of the Y4R extracellular domain by site-directed mutagenesis but also predicted three 'gain-of-function' mutations that increased the activity of (S) VU0637120.

Regarding our LB-CADD efforts, We presented the BCL::Mol2D molecular descriptor that significantly improves both predictive power and interpretability of the ANN-QSARs compared to Molprint2D and our previous-best descriptor set. 2D fragment-based similarity searching is one of the most robust CADD techniques for database searching and quantitative structure–activity relationship (QSAR) analysis. We further illustrate how BCL::Mol2D can be used to identify potential modification of a given inactive molecule to improve its predicted activity. Therefore, ANN-QSAR models trained on BCL::Mol2D could be employed in conjunction with a Monte Carlo or genetic algorithm as a structure generator (Meiler and Will, 2002; Sliwoski et al., 2014) to automate the process of rational combinatorial drug-like molecule design. The sensitivity analysis on BCL::Mol2D can guide medicinal chemists in the design of focused libraries (Zheng et al., 1998) to optimize new derivatives by filtering out unfruitful scaffold modification. This will potentially reduce the number of compounds for synthesis and testing in drug discovery campaigns.

Although they were not mentioned in the previous chapters, additional methods that study of structure

and dynamics may also reveal how specific peptide-receptor recognition may formulate a general mechanism of activation for peptide-bound class A GPCRs. Molecular dynamics simulation studies can be conducted to sample the energy landscape of the peptide activation mechanism . Additionally, the wealth of ligand-GPCR interactions data available enable deep learning models to be trained on and predict potential peptide-GPCR interactions or design novel potent biologic targeting GPCRs. These common peptide-GPCR interactions could help guide future exploration of the ensembles of protein-ligand conformations through computational modeling and various experimental techniques. The strength of computation lies in its ability to accurately use sparse experimental data to predict these types of interactions. Therefore, an iterative approach between computational sampling and energy minimization can be combined with restraints derived from a diversity of experimental methods. Incorporating a wide variety of complementary experimental techniques allows the integration of each method's advantages in providing less ambiguous restraints: NMR provides dynamic restraints, X-ray provides rigid high-resolution restraints, and mutational studies and cross-linking allow single residue-specific restraints. These experimental restraints limit the search space of possible conformations, allowing for more accurate sampling in modeling. These predictions can then be used to guide the design of future experiments.

6.2 Future Directions

6.2.1 Ligand- and structure-based virtual screening in drug discovery for Y4 receptor

Previous high throughput screening (HTS) experiments isolated a rich set of 19 confirmed Y4 positive allosteric modulators (PAM)(Schubert et al., 2017b; Sliwoski et al., 2016b). Due to various advantages of allosteric modulators (AMs), furute efforts should develope these confirmed hits into small molecule in vitro probes for the Y4 receptor. Since allosteric sites tend to be less preserved than the orthosteric counterparts, AMs are more likely to be selective toward target proteins. Moreover, side effects of AMs might be less severe as (1) they only act when the na-tive ligands are present, and (2) they tune the receptors' response instead of turning it on or off(Conn et al., 2014; Gregory et al., 2010). Computer-aided drug discovery/design (CADD) has been universally used to facilitate and expedite the development of small molecule-based therapeutics. Ligand-based virtual screening can save time and resources by quickly narrowing libraries of millions drug-like molecules to hundreds of compounds candidates for experimental tests and validation. For this project, ligand based (LB) and structure based (ST) computer-aided drug discovery (CADD) could be combined to discover small molecule modulators to Y4R. Future efforts should incorporate BCL::Mol2D fingerprint into artificial neural network (ANN) employed to predict Y4 AM activity from 3D images of molecular electron density and electrostatic fields. On the other hand, structure-based methods such as RosettaLigand docking help characterize the interaction between ligands and target proteins when the crystal structures are

not available(Sliwoski et al., 2014).

6.2.2 Applicability Domain to optimize QSAR effort

A current limitation of many QSAR approaches is lack of tools to quantify the reliability of the models' prediction on new set of compounds. Hence, I am planning to develop a distance-based applicability domain(Minovski et al., 2013) method into BCL::ChemInfo package to assess whether the chemical space of the external compounds fit into the scope of the QSAR models. Since drug-like chemical space is much larger than the region covered by the training dataset, an assessment like AD tool is needed to evaluate reliability of QSAR prediction on novel compounds. Clustering-based AD has succeeded in projects on smaller datasets(Weaver and Gleeson, 2008). Likewise, this confidence metric will be useful to assess whether a model's prediction can be considered as reliable as it is for typical compounds in the training dataset. : K-means cluttering(Rogers and Tanimoto, 1960b) divides the training compound set into clusters of molecules based on structural information and bio-physical properties encoded in a set of 23 scalar molecular descriptors(Bender et al., 2004). Then, an AD score is computed for each compound in the testing compound set by estimating the distance of that com-pound to the closest cluster of the training set. The testing compounds are ranked based on their AD scores. We will test whether compounds stratified by decreasing applicability to the model show corresponding decreases in prediction accuracy. Additionally, we will also investigate if removal of the least applicable parts of the training dataset can improve model performance on the remainder of the dataset. We will assess these metrics on an independent set of nine large high-throughput screens (Butkiewicz et al., 2013).

6.2.3 A Monte-Carlo based Algorithm that Utilizes ANN-QSAR Models and Pharmacophore Mapping Features of BCL::Mol2D Descriptors to Design Docused Libraries for Synthesis

Combinatorial organic chemistry has emerged as means to synthesize libraries of structurally diverse compounds via combining available molecular fragments. In drug discovery campaigns, those libraries are often subject to high throughput screening (HTS) for targeted biological assays. However, synthesis and testing of compound libraries require substantial time and resources. In order to minimize the experimental efforts, I propose an innovative computational focused library design method to combine fragment building blocks into a diverse set of compounds that are synthesizable and likely to be active1. Therefore, the objective of this project is implementing FocusMolLib, a novel ligand-based focused library design method that employs metropolis Monte Carlo (MC) to generate combinatorial compounds. The molecule fragments and compound products will be evaluated by quantitative structure activity relationship (QSAR) models.

A molecular fragment library is generated from 900,000 drug-like molecules of our in-hour database.

A QSAR model will be trained on the hybrid fingerprints (SR+Atom), and its applicability domain (AD) (aim II) will be calculated. To translate the QSAR models into a focused library for a scaffold of interest, a MC random walk structure generator will be implemented. The algorithm replaces one fragment within the scaffold with a chemically different fragment from the fragment library. Fragment selection and replacement are guided by the sensitivity analysis of BCL::Mol2D. In an iterative process, several thousand derivatives are created. If the output 3D structures are within the QSAR model's AD, biological activity of those designed derivatives is then predicted using the QSAR model. To assess the reliability of the method, FocusMolLib will be used to generate derivatives for positive allosteric modulators (PAMs) of the human Y4 receptor(Sliwoski et al., 2016b). Generated compounds will be synthesized, and their bioactivity will be experimentally validated.

CHAPTER 7

APPENDIX - Protocol Capture

7.1 Protocol Capture for Chapter 2: The Structural Basis of Peptide Binding at Class A G Protein-Coupled Receptors

7.1.1 Structure preparation

Nine structures are downloaded and the coordinates of the GPCR targets and the peptide ligands were extracted. The complex structures were then minimized by Rosetta backrub applied to the interface residues followed by two cycles of fast relax. Backrub movement mimics the backbone fluctuations observed in the crystal lattice. The BackrubDD mover combined backrub movements with metropolis Monte Carlo to sample low energy backbone conformation that were close to the starting crystal structures in the context of the Rosetta all-atom force field. The FastRelax protocol found low-energy backbone and side-chain conformations near a starting conformation by applying ten repeats of five rounds of packing and minimizing, with the repulsive weight in the scoring function gradually increased from a very low value to the normal value from one round to the next. For each complex, ten optimized model were generated. The sequence numbering table for each GPCR class A were extracted from GPCRdb database.

The Rosetta script file to relax the starting cleaned PDB files :

```
<ROSETTASCRIPTS>
   <SCOREFXNS>
   </SCOREFXNS>
   <MOVERS>
      <BackrubDD name="backrub" interface_distance_cutoff="8.0">
      </BackrubDD>
      <FastRelax name="relax" repeats="2">
      </FastRelax>
      <InterfaceAnalyzerMover name="analyze" packstat="0" pack_input="0"
      pack_separated="1" fixedchains="B" />
   </MOVERS>
   <PROTOCOLS>
      <Add mover_name="backrub"/>
      <Add mover_name="relax"/>
      <Add mover_name="analyze"/>
   </PROTOCOLS>
   <OUTPUT scorefxn="REF2015"/>
</ROSETTASCRIPTS>
```

This bash script file **relax_no_ligand.sh** will take two parameters, which are the input starting PDB file and the path to the working directory. It will run the optimization protocol on the starting refined PDB of a peptide-Class A GPCR complex structure.

```
#!/bin/bash
template=`readlink -e $1`
work_dir=`readlink -e $2`
prefix=$3
xml=relax_interface_analyze_no_ligand.xml
nstruct=10

cd $work_dir
/path/to/rosetta/main/source/bin/rosetta_scripts.linuxgccrelease \
   -parser:protocol $xml \
   -out:pdb_gz \
   -out:prefix ${prefix}_ \
   -nstruct $nstruct \
   -out:file:scorefile ${prefix}_scores.out \
   -s $template \
   -native $template \
   -relax:constrain_relax_to_start_coords \
   -relax:constrain_relax_to_native_coords
```

7.1.2 Incorporate NCAAs into peptide modeling for 6C1Q

A rotamer library of 1000 conformers of 3-cyclohexyl-L-alanine were generated using BCL:Conf. We also used the built-in ornithine rotamer library from the Rosetta database. Those non-natural amino acids were then incorporated into the structural optimization step of the complex.

Here are the files needed to optimize the complex structure: The params file of 3-cyclohexyl-L-alanine:

```
NAME A01
IO_STRING A01 X
TYPE POLYMER
AA UNK
ATOM N Nbb NH1 -0.61
ATOM CA CAbb CT1 -0.09
ATOM C CObb C 0.62
ATOM O OCbb O -0.55
ATOM CB CH2 CT2 -0.18
ATOM CG CH1 CT1 -0.09
ATOM CD1 CH2 CT2 -0.18
```

```
ATOM CD2 CH2 CT2 -0.18
ATOM CE1 CH2 CT2 -0.18
ATOM CE2 CH2 CT2 -0.18
ATOM CZ CH2 CT2 -0.18
ATOM H HNbb H 0.43
ATOM HA Hapo HB 0.10
ATOM 1HB Hapo HA 0.10
ATOM 2HB Hapo HA 0.10
ATOM 1HG Hapo HA 0.10
ATOM 1HD1 Hapo HA 0.10
ATOM 2HD1 Hapo HA 0.10
ATOM 1HD2 Hapo HA 0.10
ATOM 2HD2 Hapo HA 0.10
ATOM 1HE1 Hapo HA 0.10
ATOM 2HE1 Hapo HA 0.10
ATOM 1HE2 Hapo HA 0.10
ATOM 2HE2 Hapo HA 0.10
ATOM 1HZ Hapo HA 0.10
ATOM 2HZ Hapo HA 0.10
BOND C CA
BOND C O
BOND N CA
BOND N H
BOND CA CB
BOND CA HA
BOND CB CG
BOND CB 1HB
BOND CB 2HB
BOND CG CD1
BOND CG CD2
BOND CG 1HG
BOND CD1 CE1
BOND CD1 1HD1
BOND CD1 2HD1
BOND CE1 CZ
BOND CE1 1HE1
BOND CE1 2HE1
BOND CZ CE2
BOND CZ 1HZ
BOND CZ 2HZ
BOND CE2 CD2
```

```
BOND CE2 1HE2
BOND CE2 2HE2
BOND CD2 1HD2
BOND CD2 2HD2
LOWER_CONNECT N
UPPER_CONNECT C
CHI 1 N CA C O
CHI 2 N CA CB CG
CHI 3 CA CB CG CD1
NBR_ATOM CB
NBR_RADIUS 7.205283
FIRST_SIDECHAIN_ATOM CB
PROPERTIES PROTEIN L_AA
ICOOR_INTERNAL N 0.000000 0.000000 0.000000 N CA C
ICOOR_INTERNAL CA 0.000000 180.000000 1.455349 N CA C
ICOOR_INTERNAL C 0.000000 71.221549 1.523303 CA N C
ICOOR_INTERNAL O 145.761188 58.740641 1.228210 C CA N
ICOOR_INTERNAL UPPER -34.237210 64.254358 1.336084 C CA N
ICOOR_INTERNAL CB -125.479170 68.825479 1.516646 CA N C
ICOOR_INTERNAL CG -179.617102 66.644119 1.519925 CB CA N
ICOOR_INTERNAL CD1 62.004514 71.498542 1.578395 CG CB CA
ICOOR_INTERNAL CE1 68.977451 70.719321 1.545436 CD1 CG CB
ICOOR_INTERNAL CZ 53.512025 66.968504 1.520904 CE1 CD1 CG
ICOOR_INTERNAL CE2 -56.341648 68.262637 1.526339 CZ CE1 CD1
ICOOR_INTERNAL CD2 55.815984 67.254791 1.524263 CE2 CZ CE1
ICOOR_INTERNAL 1HD2 66.938982 70.864542 1.070017 CD2 CE2 CZ
ICOOR_INTERNAL 2HD2 119.629237 70.864046 1.069952 CD2 CE2 1HD2
ICOOR_INTERNAL 1HE2 120.475944 71.359493 1.069955 CE2 CZ CD2
ICOOR_INTERNAL 2HE2 119.053916 71.362573 1.070015 CE2 CZ 1HE2
ICOOR_INTERNAL 1HZ 120.329887 71.106791 1.070047 CZ CE1 CE2
ICOOR_INTERNAL 2HZ 119.341556 71.104542 1.070031 CZ CE1 1HZ
ICOOR_INTERNAL 1HE1 -120.512548 71.432812 1.069977 CE1 CD1 CZ
ICOOR_INTERNAL 2HE1 -118.969202 71.434096 1.069971 CE1 CD1 1HE1
ICOOR_INTERNAL 1HD1 119.958689 70.490558 1.070089 CD1 CG CE1
ICOOR_INTERNAL 2HD1 120.082795 70.483462 1.070002 CD1 CG 1HD1
ICOOR_INTERNAL 1HG -118.258663 70.499256 1.070042 CG CB CD1
ICOOR_INTERNAL 1HB -120.565157 71.519929 1.070009 CB CA CG
ICOOR_INTERNAL 2HB -118.871071 71.519344 1.070043 CB CA 1HB
ICOOR_INTERNAL HA -117.870256 71.615741 1.070019 CA N CB
ICOOR_INTERNAL LOWER -45.370585 55.080156 1.333945 N CA C
ICOOR_INTERNAL H -179.998550 62.464039 0.980014 N CA LOWER
```

```
PDB_ROTAMERS A01_conformers.pdb
```

The resfile that specifies the mutants' locations

```
NATAA # default is repacking only


start
2 B PIKAA X[ORN]
4 B PIKAA X[A01]
```

The Rosetta XML file contains the protocol of mutating the residues and optimize the complex structure

```
<ROSETTASCRIPTS>
   <SCOREFXNS>
   </SCOREFXNS>
   <PACKER_PALETTES>
      DefaultPackerPalette name="NCAA_expanded" />
      <CustomBaseTypePackerPalette name="base_ncaa" additional_residue_types="%%res_type
          ↪ %%" />
   </PACKER_PALETTES>
   <TASKOPERATIONS>
      Include rotamer options from the command line
      <InitializeFromCommandline name="ifcl" />
         Design and repack residues based on resfile
      <ReadResfile name="rrf" filename="%%resf_file%%"/>
   </TASKOPERATIONS>
   <MOVERS>
      FavorNativeResidue name="favor_native" bonus="0.75"/>
      Design the antibody interface
      <PackRotamersMover name="design" scorefxn="REF2015" task_operations="ifcl,rrf"
          ↪ packer_palette="base_ncaa"/>
         PackRotamersMover name="design" scorefxn="REF2015" task_operations="ifcl,rrf" />
         Analyze the resulting interface
      <InterfaceAnalyzerMover name="analyze" scorefxn="REF2015" packstat="0" pack_input
          ↪ ="0" pack_separated="1" fixedchains="A" />
      <PeptideCyclizeMover name="crosslink" >
         <Torsion res1="302" res2="305" res3="305" res4="305" atom1="CG" atom2="CD" atom3
             ↪ ="NZ" atom4="CE" cst_func="CIRCULARHARMONIC 3.141592654 0.005" />
         <Angle res1="302" atom1="CG" res_center="302" atom_center="CD" res2="305" atom2="
             ↪ NZ" cst_func="HARMONIC 2.01000000 0.01" />
         <Angle res1="302" atom1="CD" res_center="305" atom_center="NZ" res2="305" atom2="
             ↪ CE" cst_func="HARMONIC 2.14675498 0.01" />
```

```
        <Distance res1="302" res2="305" atom1="CD" atom2="NZ" cst_func="HARMONIC 1.32865
            ↪ 0.01" />
        <Bond res1="302" res2="305" atom1="CD" atom2="NZ" add_termini="true" />
    </PeptideCyclizeMover>
    <BackrubDD name="backrub" interface_distance_cutoff="6.0">
    </BackrubDD>
    <FastRelax name="relax" scorefxn="REF2015" repeats="2"/>
  </MOVERS>
  <FILTERS>
    <Ddg name="ddg_f" scorefxn="REF2015" threshold="-75" jump="1" repack="false" repeats
        ↪ ="1" />
    <ShapeComplementarity name="sc_f" min_sc="0.5" jump="1" />
  </FILTERS>
  <APPLY_TO_POSE>
  </APPLY_TO_POSE>
  <PROTOCOLS>
    Run the design protocol
    Add mover="favor_native" />
    <Add mover="design" />
    <Add mover_name="backrub"/>
    <Add mover_name="relax"/>
      Calculate interface metrics for the final sequence
    <Add mover="analyze" />
    Add filter="ddg_f" />
    Add filter="sc_f" />
  </PROTOCOLS>
  <OUTPUT scorefxn="REF2015" />
</ROSETTASCRIPTS>
```

Then we can run the shell script containing the Rosetta command line

```
#!/bin/bash

#Rosetta version
ROSETTA=/path/to/rosetta/main/source/bin/rosetta_scripts.linuxgccrelease

cd 6C1Q
# asign prefix, names of residue, and parame files
resf_file=design.resf
xml_file=relax.xml
res1=A01
param2=3-cyclohexyl-L-alanine.params
```

```
# Run design with top 10 relaxed models
$ROSETTA -parser:protocol $xml_file \
   -out:pdb_gz \
   -out:prefix ${seed}_\
   -nstruct 10 \
   -out:file:scorefile ${seed}_scores.out \
   -s 6C1Q_refined.pdb \
   -native 6C1Q_refined.pdb \
   -ex1 \
   -ex2 \
   -parser:script_vars res_type=${res1},${res2} resf_file=${resf_file} \
   -extra_res_fa ${param} \
   -chemical:exclude_patches LowerDNA UpperDNA SpecialRotamer VirtualBB ShoveBB
       ↪ VirtualDNAPhosphate VirtualNTerm CTermConnect sc_orbitals pro_hydroxylated_case1
       ↪  pro_hydroxylated_case2 ser_phosphorylated thr_phosphorylated tyr_phosphorylated
       ↪  tyr_sulfated lys_dimethylated lys_monomethylated lys_trimethylated
       ↪ lys_acetylated glu_carboxylated cys_acetylated tyr_diiodinated C_methylamidated
       ↪ MethylatedProteinCterm Cterm_amidation
```

7.1.3 Modeling of native peptide of 5VBL

Since the structure of the ligand in the structure is derived from the C-terminus of apelin, we generated the model of the native apelin peptide and the apelin receptor using fix backbone design application in Rosetta before the structural optimization step.

Shell sript file to run relax on 5VBL complex

```
#!/bin/bash
#SBATCH --ntasks=1
#SBATCH --nodes=1
#SBATCH --mem-per-cpu=4G
#SBATCH --time=10:00:00
#SBATCH --array=1-100

seed=$SLURM_ARRAY_TASK_ID
#Rosetta version
ROSETTA=/Path/To/Rosetta/main/source/bin/rosetta_scripts.default.linuxgccrelease

cd 5VBL
# asign prefix, names of residue, and parame files
resf_file=design.resf
```

```
xml_file=relax.xml
res1=R01
param1=./L-homo-arginine.params
res2=P01
param2=./L-Octahydroindole-2-carboxylic-acid.params
res3=L01
param3=/home/vuot2/bin/ncaa_lib/external_database/leucine_dev/L-norleucine.params
res4=023
param4=./4-Chloro-L-phenylalanin.params
res5=A01
param5=~/bin/ncaa_lib/external_database/alanine_dev/3-cyclohexyl-L-alanine.params

# Run design with top 10 relaxed models
$ROSETTA -parser:protocol $xml_file \
   -out:pdb_gz \
   -out:prefix ${seed}_ \
   -nstruct 1 \
   -out:file:scorefile ${seed}_scores.out \
   -s 5VBL_refined.pdb \
   -native 5VBL_refined.pdb \
   -ex1 \
   -ex2 \
   -parser:script_vars res_type=${res1},${res2},${res3},${res4},${res5} resf_file=${
       ↪ resf_file} \
   -extra_res_fa ${param1} ${param2} ${param3} ${param5} ${param4}
```

The accompanying Rosetta XML script file is

```
<ROSETTASCRIPTS>
   <SCOREFXNS>
      <ScoreFunction name="score_cst" weights="ref2015" >
         <Reweight scoretype="coordinate_constraint" weight="1" />
         <Reweight scoretype="atom_pair_constraint" weight="1" />
         <Reweight scoretype="dihedral_constraint" weight="1" />
         <Reweight scoretype="angle_constraint" weight="1" />
      </ScoreFunction>
   </SCOREFXNS>
   <PACKER_PALETTES>
      DefaultPackerPalette name="NCAA_expanded" />
      <CustomBaseTypePackerPalette name="base_ncaa" additional_residue_types="%%res_type
          ↪ %%" />
   </PACKER_PALETTES>
```

```
<TASKOPERATIONS>
   Include rotamer options from the command line
   <InitializeFromCommandline name="ifcl" />
      Design and repack residues based on resfile
   <ReadResfile name="rrf" filename="%%resf_file%%"/>
</TASKOPERATIONS>
<MOVERS>
   FavorNativeResidue name="favor_native" bonus="0.75"/>
   Design the antibody interface
   <PackRotamersMover name="design" scorefxn="REF2015" task_operations="ifcl,rrf"
       ↪ packer_palette="base_ncaa"/>
      PackRotamersMover name="design" scorefxn="REF2015" task_operations="ifcl,rrf" />
      Analyze the resulting interface
   <InterfaceAnalyzerMover name="analyze" scorefxn="REF2015" packstat="0" pack_input
       ↪ ="0" pack_separated="1" fixedchains="A" />
   <BackrubDD name="backrub" interface_distance_cutoff="6.0">
   </BackrubDD>
   <FastRelax name="relax" scorefxn="score_cst" ramp_down_constraints="false" repeats
       ↪ ="1" />
      Define the covalent bond between two sidechain
   <PeptideCyclizeMover name="crosslink" >
      <Torsion res1="302" res2="302" res3="305" res4="305" atom1="CG" atom2="CD" atom3
          ↪ ="NZ" atom4="CE" cst_func="CIRCULARHARMONIC 3.141592654 0.005" />
      <Angle res1="302" atom1="CG" res_center="302" atom_center="CD" res2="305" atom2="
          ↪ NZ" cst_func="HARMONIC 2.01000000 0.01" />
      <Angle res1="302" atom1="CD" res_center="305" atom_center="NZ" res2="305" atom2="
          ↪ CE" cst_func="HARMONIC 2.14675498 0.01" />
      <Distance res1="302" res2="305" atom1="CD" atom2="NZ" cst_func="HARMONIC 1.32865
          ↪ 0.01" />
      <Bond res1="302" res2="305" atom1="CD" atom2="NZ" add_termini="true" />
   </PeptideCyclizeMover>
</MOVERS>
<FILTERS>
   <Ddg name="ddg_f" scorefxn="REF2015" threshold="-75" jump="1" repack="false" repeats
       ↪ ="1" />
   <ShapeComplementarity name="sc_f" min_sc="0.5" jump="1" />
</FILTERS>
<APPLY_TO_POSE>
</APPLY_TO_POSE>
<PROTOCOLS>
   Run the design protocol
```

```
      Add mover="favor_native" />
      <Add mover="design" />
      <Add mover="crosslink" />
      <Add mover_name="backrub"/>
      <Add mover_name="relax"/>
      Calculate interface metrics for the final sequence
      <Add mover="analyze" />
   </PROTOCOLS>
   <OUTPUT scorefxn="REF2015" />
</ROSETTASCRIPTS>
```

7.1.4 ΔΔG Analysis

Residue-pair Rosetta interaction energy between GPCR targets and peptide ligands were computed. For each model of the optimized structure ensemble, the ΔΔG value of each interaction target residue is the sum of computed pairwise interaction energy. The ΔΔG values of each residue were then averaged over 10 structures.

The shell script file that performs the ΔΔG Analysis is

```
#!/bin/bash

###############
# OANH VU
# 06 05 2020
###############

THREAD_NUM=10

# Check the number of the parameters
if [ "$#" -ne 4 ]
then
   echo "Wrong number of arguements. Usage: $0 <model_list of pdb.gz files> <prefix of
       ↪ output name> <2 chain names in alphabetical order>"
   echo "This script calculates average per-residue-ddg for residues in the protein-
       ↪ protein interface. The model will use ref2015 as the scoring fxns"
   echo "the first model in the model list will be used for the output pdb file"
   exit 1
else
   # Read in variables
   model_list=$1
   prefix=$2
   target_chain=$3
```

```
    ligand_chain=$4
    ROSETTA=/path/to/rosetta/main/source/bin/residue_energy_breakdown.linuxgccrelease

    echo Computing per_residue ddg
    $ROSETTA -l $model_list -out:file:silent ${prefix}_ddg_score.out
    echo extracting the average ddg values for each cluster
    model_num=`cat $model_list | wc -l`
    model_num_cutoff=$((model_num / 5))
    #Choose the residues that have interaction to at least 20% of the selected model
    grep -v onebody ${prefix}_ddg_score.out | grep "[0-9]${target_chain} "| grep "[0-9]${
        ↪ ligand_chain} "| awk '{print $4,$7}'|sort -g | uniq -c| awk -v n=
        ↪ $model_num_cutoff '$1 > n {print $2, $3}' > ${prefix}_ddg_interface.res.lst

    # extract all ddg values
    #grep "[0-9]${target_chain} " ddg_out.sc| grep "[0-9]$ligand_chain " | grep -v oneboy|
        ↪ awk '{print $4, $5, $7, $8, $(NF-1), $2}' > interaction.ddg.sc
    grep "[0-9]${target_chain} " ${prefix}_ddg_score.out| grep "[0-9]$ligand_chain " | grep
        ↪  -v onebody| awk '{print $4, $7, $(NF-1), $NF}' |sed "s/${target_chain} / /g"|
        ↪ sed "s/${ligand_chain} / /g" > ${prefix}_${ligand_chain}_${target_chain}
        ↪ _interaction.ddg.sc
    #grep -v onebody ddg_score.out| grep ' '| awk '{print $4, $(NF-1)}' > cluster_${cluster
        ↪ }.res.ddg.sc
    # take average of ddg on each residues of the ligand and the target
    for res in `awk '{print $1}' ${prefix}_ddg_interface.res.lst|sort|uniq`; do echo $res `
        ↪ grep "^${res} " ${prefix}_${ligand_chain}_${target_chain}_interaction.ddg.sc |
        ↪ awk '{print $3}' | scripts/ave_columns`; done | sed "s/${target_chain} / /g" > $
        ↪ {prefix}_chain${target_chain}.ave.res.ddg.sc
    for res in `awk '{print $2}' ${prefix}_ddg_interface.res.lst|sort|uniq`; do echo $res `
        ↪ grep " ${res} " ${prefix}_${ligand_chain}_${target_chain}_interaction.ddg.sc |
        ↪ awk '{print $3}' | scripts/ave_columns`; done | sed "s/${ligand_chain} / /g" > $
        ↪ {prefix}_chain${ligand_chain}.ave.res.ddg.sc
fi
```

7.2 Protocol capture for chapter 3: The First Selective Y4 Receptor Antagonist Binds in a Deep, Allosteric Binding Pocket

7.2.1 Building Homology Model of Y4 with RosettaCM

Sequences of Y4R and the templates were aligned using the Clustal Omega web server (Sievers et al., 2011). Multiple sequence alignment was then adjusted to ensure that the helix regions and reserved residues remain aligned, and to remove gaps within transmembrane α-helices. Addition constraints were set up to

account for the disulfide bond between Cys1.25 and Cys7.38, as well as between Cys3.25 and Cys5.25 in the Y4R.disulfide file. The span file that mark the transmembrane region of Y4 is:

```
TM region prediction for Y4R
7 307
antiparallel
n2c
   5 31 5 31
  45 69 45 69
  79 104 79 104
 122 144 122 144
 179 202 179 202
 233 254 233 254
 267 289 267 289
```

Y4R homology models were built using RosettaCM protocol (Song et al., 2013a) of the Rosetta3.9 software suites (Bender et al., 2016). The XML file that details the RosettaCM protocol is:

```
<dock_design>
      <TASKOPERATIONS>
      </TASKOPERATIONS>
      <SCOREFXNS>
            <stage1 weights="input_files/stage1_membrane.wts" symmetric=0>
                  <Reweight scoretype=atom_pair_constraint weight=1/>
            </stage1>
            <stage2 weights="input_files/stage2_membrane.wts" symmetric=0>
                  <Reweight scoretype=atom_pair_constraint weight=0.5/>
            </stage2>
            <fullatom weights="input_files/stage3_rlx_membrane.wts" symmetric=0>
                  <Reweight scoretype=atom_pair_constraint weight=0.5/>
            </fullatom>
            <membrane weights="membrane_highres_Menv_smooth" symmetric=0>
                  <Reweight scoretype=cart_bonded weight=0.5/>
                  <Reweight scoretype=pro_close weight=0/>
            </membrane>
      </SCOREFXNS>
      <FILTERS>
      </FILTERS>
      <MOVERS>
            <Hybridize name=hybridize stage1_scorefxn=stage1 stage2_scorefxn=stage2
                ↪ fa_scorefxn=fullatom batch=1 stage1_increase_cycles=1.0
                ↪ stage2_increase_cycles=1.0 linmin_only=1 realign_domains=0 disulf_file
```

```
                ↪ ="input_files/Y4R-fasta.disulfide">
                  <Fragments 3mers="input_files/Y4R-frags.200.3mers" 9mers="input_files/
                      ↪ Y4R-frags.200.9mers"/>
                  <Template pdb="threaded_pdbs/4dkl_out.pdb" cst_file="AUTO" weight= 1.000
                      ↪ />
                  <Template pdb="threaded_pdbs/4djh_out.pdb" cst_file="AUTO" weight= 1.000
                      ↪ />
                  <Template pdb="threaded_pdbs/4n6h_out.pdb" cst_file="AUTO" weight= 1.000
                      ↪ />
                  <Template pdb="threaded_pdbs/4s0v_out.pdb" cst_file="AUTO" weight= 1.000
                      ↪ />
                  <Template pdb="threaded_pdbs/4zjc_out.pdb" cst_file="AUTO" weight= 1.000
                      ↪ />
                  <Template pdb="threaded_pdbs/5dhg_out.pdb" cst_file="AUTO" weight= 1.000
                      ↪ />
                  <Template pdb="threaded_pdbs/5glh_out.pdb" cst_file="AUTO" weight= 1.000
                      ↪ />
                  <Template pdb="threaded_pdbs/3odu_out.pdb" cst_file="AUTO" weight= 1.000
                      ↪ />
            </Hybridize>
            <ClearConstraintsMover name=clearconstraints/>
            <FastRelax name=relax scorefxn=membrane repeats=1 dualspace=1 bondangle=1/>
      </MOVERS>
      <APPLY_TO_POSE>
      </APPLY_TO_POSE>
      <PROTOCOLS>
            <Add mover=hybridize/>
            <Add mover=clearconstraints/>
            <Add mover=relax/>
      </PROTOCOLS>
      <OUTPUT scorefxn=membrane/>
</dock_design>
```

And the options for the Rosetta command line are listed in this rosetta_cm.options file as showed below.

```
-database /path/to/Rosetta/main/database/ #### path to Rosetta database


# i/o
-in:file:fasta input_files/Y4R.fasta #### fasta of final sequence to be modeled
-parser:protocol input_files/rosetta_cm.xml #### path to XML script
-out:path:all output_files/ #### designates where to put pdbs/silent files/scorefiles/etc
# output styles
```

```
-out:pdb #### specifiies output format as pdbs
#-out:pdb_gz #### specifies output formats as gzipped pdbs
#-out:file:silent cm_models.silent #### specifies output format as silent files (Rosetta-
    ↪ specific compressed file)
#-out:file:silent_struct_type binary #### specifies silent file (if used) to be in type '
    ↪ binary'
#-out:file:scorefile y4r_scores.out #### gives specific name for scorefile (default is
    ↪ scores.sc)
-nstruct 5  #### specifies number of models to be created
# membrane options
-in:file:spanfile input_files/Y4R.span   #### lists transmembrane spanning regions for
    ↪ membrane scoring
-membrane:no_interpolate_Mpair
-membrane:Menv_penalties
-rg_reweight .1

# relax options
-relax:minimize_bond_angles
-relax:minimize_bond_lengths
-relax:jump_move true
-default_max_cycles 200
-relax:min_type lbfgs_armijo_nonmonotone
-score:weights input_files/stage3_rlx_membrane.wts   #### path to membrane weights file
-use_bicubic_interpolation
-hybridize:stage1_probability 1.0
-sog_upper_bound 15

# reduce memory footprint
-chemical:exclude_patches LowerDNA UpperDNA Cterm_amidation SpecialRotamer VirtualBB
    ↪ ShoveBB VirtualDNAPhosphate VirtualNTerm CTermConnect sc_orbitals
    ↪ pro_hydroxylated_case1 pro_hydroxylated_case2 ser_phosphorylated thr_phosphorylated
    ↪  tyr_phosphorylated tyr_sulfated lys_dimethylated lys_monomethylated
    ↪ lys_trimethylated lys_acetylated glu_carboxylated cys_acetylated tyr_diiodinated
    ↪ N_acetylated C_methylamidated MethylatedProteinCterm

-linmem_ig 10
```

We now can run the command line to generate Y4 homology models

```
id= $1
/dors/meilerlab/apps/rosetta/rosetta-3.9/main/source/bin/rosetta_scripts.linuxgccrelease -
    ↪ database /dors/meilerlab/apps/rosetta/rosetta-3.9/main/database/ @ input_files/
```

```
    ↪ rosetta_cm.options -out:prefix y4r_${id}_ -out:file:scorefile y4r_${id}_scores.out
```

The output Y4R homology models were evaluated by total Rosetta total energy score, which includes knowledge-based energy terms such as hydrogen bonds, electrostatic interactions, and van der Waals packing (Song et al., 2013a).

Top 10% of output models, that scored most favorably by Rosetta, were clustered based on RMSD cut-off of 2 Å using BCL::Cluster (Alexander et al., 2011).

```
grep SCORE: *.out | awk '{print $2, $NF".pdb.gz"}' | sort -n |head -1000 >
    ↪ top10_percent_models.sc
```

Finally, a Y4R model with the best Rosetta total energy score was chosen from each of ten largest clusters.

7.2.2 Induced-fit docking of (S)-VU0637120 to Y4

The Y4R models were re-hybridized with the ligand placed inside the space among important residues. To create a set of templates for the second rounds of RosettaCM, 65 different poses of the ligand were placed inside Y4R such that it interacted to the TM1, TM2, TM7, and in no more than 4 Å from Y1.39. This way, the interactions between the ligand and the important residues according to the mutagenesis data would be weighed in the placement of helices in MC trajectory and optimization. Those 65 ligand-Y4 complexes were used as the templates for re-hybridization. A alignment entry was placed at the begining of each of the template PDB

```
sed -i '1s/^/REMARK query_anchored_aln
    ↪ CQDSVDVMVFIVTSYSIETVVGVLGNLCLMCVTVRQKEKANVTNLLIANLAFSDFLMCLLCQPLTAVYTIMDYWIFGETLCKM
SAFIQCMSVTVSILSLVLVALERHQLIINPTGWKPSISQAYLGIVLIWVIACVLSLPFLANSILENVFHKNHSKALEFLADKV
VCTESWPLAHHRTIYTTFLLLFQYCLPLGFILVCYARIYRRLQRQGRVFHKGYSLRAGHMKQVNVVLVVMVVAFAVLWLPLHV
FNSLEDWHHEAIPICHGNLIFLVCHLLAMASTCVNPFIYGFLNTNFKKEIKALVLTC\n/' NAM-Y4-*.pdb
```

The Rosetta XML script with the protocol of performing re-hybridization with the ligand placed inside the binding pocket

```
<ROSETTASCRIPTS>
   <TASKOPERATIONS>
   </TASKOPERATIONS>
   <SCOREFXNS>
      <ScoreFunction name="stage1" weights="input_files/stage1_membrane.wts" symmetric
          ↪ ="0">
         <Reweight scoretype="atom_pair_constraint" weight="1"/>
      </ScoreFunction>
```

```
      <ScoreFunction name="stage2" weights="input_files/stage2_membrane.wts" symmetric
          ↪ ="0">
         <Reweight scoretype="atom_pair_constraint" weight="0.5"/>
      </ScoreFunction>
      <ScoreFunction name="fullatom" weights="input_files/stage3_rlx_membrane.wts"
          ↪ symmetric="0">
         <Reweight scoretype="atom_pair_constraint" weight="0.5"/>
      </ScoreFunction>
      <ScoreFunction name="membrane" weights="membrane_highres_Menv_smooth" symmetric="0">
         <Reweight scoretype="cart_bonded" weight="0.5"/>
         <Reweight scoretype="pro_close" weight="0"/>
      </ScoreFunction>
   </SCOREFXNS>
   <FILTERS>
   </FILTERS>
   <MOVERS>
      <Hybridize name="hybridize" add_hetatm="1" stage1_scorefxn="stage1" stage2_scorefxn
          ↪ ="stage2" fa_scorefxn="fullatom" batch="1" stage1_increase_cycles="1.0"
          ↪ stage2_increase_cycles="1.0" linmin_only="1" realign_domains="0" disulf_file
          ↪ ="input_files/Y4R.disulfide">
         <Fragments three_mers="input_files/Y4R-frags.200.3mers" nine_mers="input_files/
             ↪ Y4R-frags.200.9mers"/>
         <Template pdb="%%template%%" cst_file="AUTO" weight="1.000" />
      </Hybridize>
      <ClearConstraintsMover name="clearconstraints"/>
      <FastRelax name="relax" scorefxn="membrane" repeats="1" dualspace="1" bondangle
          ↪ ="1"/>
   </MOVERS>
   <APPLY_TO_POSE>
   </APPLY_TO_POSE>
   <PROTOCOLS>
      <Add mover="hybridize"/>
      <Add mover="clearconstraints"/>
      <Add mover="relax"/>
   </PROTOCOLS>
   <OUTPUT scorefxn="membrane"/>
</ROSETTASCRIPTS>
```

Similar to the previous step of generating Y4 homology model, there was an option file for running the Rosetta command line

```
-database /dors/meilerlab/apps/rosetta/rosetta-3.9/main/database/ #### path to Rosetta
    ↪ database

# i/o
-in:file:fasta input_files/Y4R.fasta #### fasta of final sequence to be modeled
-parser:protocol input_files/rosetta_cm.xml #### path to XML script
-out:path:all output_files/ #### designates where to put pdbs/silent files/scorefiles/etc
# output styles
-out:pdb #### specifiies output format as pdbs
#-out:pdb_gz #### specifies output formats as gzipped pdbs
#-out:file:silent cm_models.silent #### specifies output format as silent files (Rosetta-
    ↪ specific compressed file)
#-out:file:silent_struct_type binary #### specifies silent file (if used) to be in type '
    ↪ binary'
-out:file:scorefile Y4R_scores.out #### gives specific name for scorefile (default is
    ↪ scores.sc)
-nstruct 10 #### specifies number of models to be created



# membrane options
-in:file:spanfile input_files/Y4R.span #### lists transmembrane spanning regions for
    ↪ membrane scoring
-membrane:no_interpolate_Mpair
-membrane:Menv_penalties
-rg_reweight .1

# relax options
-relax:minimize_bond_angles
-relax:minimize_bond_lengths
-relax:jump_move true
-default_max_cycles 200
-relax:min_type lbfgs_armijo_nonmonotone
-score:weights input_files/stage3_rlx_membrane.wts #### path to membrane weights file
-use_bicubic_interpolation
-hybridize:stage1_probability 1.0
-sog_upper_bound 15

# reduce memory footprint
-chemical:exclude_patches LowerDNA UpperDNA Cterm_amidation SpecialRotamer VirtualBB
    ↪ ShoveBB VirtualDNAPhosphate VirtualNTerm CTermConnect sc_orbitals
    ↪ pro_hydroxylated_case1 pro_hydroxylated_case2 ser_phosphorylated thr_phosphorylated
```

```
    ↪  tyr_phosphorylated tyr_sulfated lys_dimethylated lys_monomethylated
    ↪ lys_trimethylated lys_acetylated glu_carboxylated cys_acetylated tyr_diiodinated
    ↪ N_acetylated C_methylamidated MethylatedProteinCterm

-linmem_ig 10

# run multiple processors to produce output for one file
#-multiple_processes_writing_to_one_directory
```

The command line to generate the re-hybridized models (were ran in parallalel in 10 processors) is:

```
template= $1
rep= $2
/dors/meilerlab/apps/rosetta/rosetta-3.9/main/source/bin/rosetta_scripts.linuxgccrelease @
    ↪  input_files/rosetta_cm.options -out:prefix test_1 -nstruct 100 -in:file:
    ↪ extra_res_fa NAM.fa.params -in:file:extra_res_cen NAM.cen.params -score:
    ↪ extra_improper_file NAM.fa.tors -out:file:scorefile ${rep}_scores.out -
    ↪ ignore_zero_occupancy false -parser:script_vars template= ${template}
```

The resulting models were then filtered for at least 10 interactions to important residues and hydrogen bond to Y1.39.

```
python filter_based_on_interaction_num.py -l Y4-NAM_models.lst -r important_res -j 10 -m
    ↪ 10 -o good_models.lst
```

The filter_based_on_interaction_num.py compute the number of residue-ligand contacts in each docked model. This script can run on multiple CPUs in parallel.

```
import numpy as np
import argparse

parser = argparse.ArgumentParser(description='filter the docked model that contact to at
    ↪ least a number of residues')
parser.add_argument('-l', '--list', required=True,help='list of input docking models. Make
    ↪  sure the target chain is A\
              and the ligand chain is X', dest='list')
parser.add_argument('-r', '--res', required=True, help='list of important residues', dest
    ↪ ='res')
parser.add_argument('-j', '--n_jobs', type=int, default=1, help='number of \
              threads used, default: 1', dest='n_jobs')
parser.add_argument('-d', '--distance', default=6, type=float, dest='distance',
              help='max distance to be considered a contact')
parser.add_argument('-m', '--min', dest='min', required=True, type=float,
```

```
                help='min number of contact')
parser.add_argument('-o', '--output', dest='output', default="output.txt",
                help='output file name')
args = parser.parse_args()

from Bio.PDB.PDBParser import PDBParser
parser = PDBParser()
model_lst = np.loadtxt(args.list, dtype="S100")
res_list = np.loadtxt(args.res, dtype="i4")

# Return the distance between two atoms
def atm_distance(atom1, atom2):
   sum = 1000
   if (atom1.element != 'H') & (atom2.element != 'H'):
      sum = 0
      for i in range(0,3):
         sum = sum + (atom1.get_coord()[i] -
                  atom2.get_coord()[i])**2
      sum = sum**0.5
   return sum

# return the min atom pairwise distance between two residues
def res_distance(res1, res2):
   min_dis = 1000
   for atom1 in res1:
      for atom2 in res2:
         min_dis = min(min_dis, atm_distance(atom1, atom2))
   return min_dis

# returns the number of contacted important residues of a model
def contacted_res_num(model):
   pdb = parser.get_structure("pdb", model)[0]
   ligand = pdb['X'].get_list()[0]
   target = pdb['A'].get_list()
   contact_num = 0
   for resi in res_list:
      residue = target[resi]
      if res_distance(ligand, residue) <= args.distance:
         contact_num = contact_num + 1
   return contact_num
def test_serial_version():
```

```
   for model in model_lst:
      print contacted_res_num(model)
def main():
   #test_serial_version()
   import multiprocessing as mp
   import random

   random.seed(123)

   pool = mp.Pool(processes=args.n_jobs)
   results = [pool.map(contacted_res_num, model_lst)]
   f = open(args.output, 'w')
   f.write('{}\t{}\n'.format('model', '#contacts'))
   for i in range(0, len(model_lst)):
      if results[0][i] >= args.min:
         f.write('{}\t{}\n'.format(model_lst[i], results[0][i]))
   f.close()
if __name__ == "__main__":
   main()
```

The output models were then clustered into ten clusters based on the distance matrix of the set of the important residues, which were hypothesized to be lining the binding pocket. The representative models were selected to be templates for the docking protocol based on the size of the corresponding clusters as well as the openness of the binding pocket suggested by the mutagenesis data.

```
bcl.exe protein:Compare -quality RMSD -pdb_list pdb.lst -specify_residues res_lst -
    ↪ output_dir distance_matrix/ -prefix Y4R -scheduler PThread 9
bcl.exe bcl:Cluster -distance_input_file Y4RRMSD.txt -input_format TableLowerTriangle -
    ↪ output_file models.out -linkage Complete -output_format Rows Centers -scheduler
    ↪ PThread 10 -remove_nodes_below_size 40 -remove_internally_similar_nodes 1.5
```

A set of 106 conformations of VU0637120 was generated using the ConformerGenerator application of BCL v3.2.2 (Kothiwale et al., 2015).

```
bcl.exe molecule:ConformerGenerator -conformation_comparer SymmetryRMSD 0.0 -
    ↪ max_iterations 8000 -top_models 106 -cluster -ensemble_filenames NAM.sdf -
    ↪ conformers_single_file NAM_rotamers.sdf -explicit_aromaticity
```

RosettaLigand was used to dock VU0637120 to Y4R homology models. Here is the Rosetta XML script file:

```
<ROSETTASCRIPTS>
```

```
<SCOREFXNS>
   <ScoreFunction name="ligand_soft_rep" weights="ligand_soft_rep">
   </ScoreFunction>
   <ScoreFunction name="hard_rep" weights="ligand">
      <Reweight scoretype="fa_intra_rep" weight="0.005"/>
   </ScoreFunction>
</SCOREFXNS>

<LIGAND_AREAS>
   <LigandArea name="inhibitor_dock_sc" chain="X" cutoff="6.0" add_nbr_radius="true"
       ↪ all_atom_mode="false"/>
   <LigandArea name="inhibitor_final_sc" chain="X" cutoff="6.0" add_nbr_radius="true"
       ↪ all_atom_mode="false"/>
   <LigandArea name="inhibitor_final_bb" chain="X" cutoff="7.0" add_nbr_radius="false"
       ↪ all_atom_mode="true" Calpha_restraints="0.3"/>
</LIGAND_AREAS>

<INTERFACE_BUILDERS>
   <InterfaceBuilder name="side_chain_for_docking" ligand_areas="inhibitor_dock_sc"/>
   <InterfaceBuilder name="side_chain_for_final" ligand_areas="inhibitor_final_sc"/>
   <InterfaceBuilder name="backbone" ligand_areas="inhibitor_final_bb" extension_window
       ↪ ="3"/>
</INTERFACE_BUILDERS>

<MOVEMAP_BUILDERS>
   <MoveMapBuilder name="docking" sc_interface="side_chain_for_docking" minimize_water
       ↪ ="false"/>
   <MoveMapBuilder name="final" sc_interface="side_chain_for_final" bb_interface="
       ↪ backbone" minimize_water="false"/>
</MOVEMAP_BUILDERS>

<SCORINGGRIDS ligand_chain="X" width="35">
   <ClassicGrid grid_name="classic" weight="1.0"/>
</SCORINGGRIDS>

<MOVERS>
   <StartFrom name="start" chain="X">
      <File filename="%%start_file%%" />
   </StartFrom>
   <Transform name="transform" chain="X" box_size="5.0" move_distance="0.04" angle="4"
```

```
        ↪ cycles="500" repeats="1" temperature="10" initial_perturb="3" />
      <HighResDocker name="high_res_docker" cycles="6" repack_every_Nth="3" scorefxn="
        ↪ ligand_soft_rep" movemap_builder="docking"/>
      <FinalMinimizer name="final" scorefxn="hard_rep" movemap_builder="final"/>
      <InterfaceScoreCalculator name="add_scores" chains="X" scorefxn="hard_rep" native
        ↪ ="%%templateligand%%"/>
   </MOVERS>
   <FILTERS>
      <Ddg name="ddg" scorefxn="hard_rep" chain_num="2" threshold="%%maxddg%%"/>
   </FILTERS>

   <PROTOCOLS>
      <Add mover_name="start"/>
      <Add mover_name="transform"/>
      <Add mover_name="high_res_docker"/>
      <Add mover_name="final"/>
      <Add mover_name="add_scores"/>
      <Add filter_name="ddg"/>
   </PROTOCOLS>
</ROSETTASCRIPTS>
```

And the Rosetta command line is

```
#!/bin/bash
template=$1
params_f=$2
work_dir=$3
maxddg=0
prefix=$5
start_file=$6
xml=dock.xml
ntruct=100

cd $work_dir
mkdir -p output_files_${prefix}
/dors/meilerlab/apps/rosetta/rosetta-3.12/main/source/bin/rosetta_scripts.linuxgccrelease
    ↪ \
   -parser:protocol $xml \
   -out:file:silent ${prefix}_silent.out \
   -parser:script_vars maxddg=${maxddg} templateligand=$template start_file=$start_file \
   -out:prefix ${seed}_ \
   -nstruct $ntruct \
```

```
   -s $template \
   -score:analytic_etable_evaluation true \
   -out:file:scorefile ${seed}_scores.out \
   -mistakes \
   -restore_pre_talaris_2013_behavior true \
   -out:path:all ./output_files_${prefix}/ \
   -packing:ex1 \
   -packing:ex2 \
   -packing:no_optH false \
   -packing:flip_HNQ true \
   -packing:ignore_ligand_chi true \
   -in:file:extra_res_fa $params_f
```

In addition to Rosetta total energy score, the interface_delta score, which is relative to the predicted binding energy between VU0637120 and Y4R, was also extracted for all 20,000 output docking models. We selected models with total Rosetta score of at least 95% of the best total Rosetta score attained and with top 10% interface delta score.

```
awk '$2 < -1100 {print $0}' *.out | grep -v "delta" |grep -v 'SEQUENCE:' |awk '$59 < -13 {
    ↪ print $0}' > good_models.sc
```

```
grep temp\_ filter\_9.txt | xargs -n1 -I@ grep @ good_models.sc >> filter\_9.sc
```

We applied an experimental filter on the selected docking models. The filter makes sure that the selected ligand poses that contact nine out of the 11 residues that were shown to be important to the antagonistic activity by mutagenesis data. A residue is determined to contact the ligand if at least one atom of the residue is at most 5 Å away from any atom of the ligand.

```
python filter_ligand_poses.py good_models.sc -g all.groups -e 15 -s True -p ../NAM.params
    ↪ -i ../important_res -n 9 -o filter_9.txt
grep temp_ filter_9.txt | xargs -n1 -I@ grep @ good_models.sc > filter_9.sc
```

The python script filter_ligand_poses.py is

```
#!/usr/bin/env python
#written by: Oanh Vu, Benjamin K. Mueller


# Reads in a standard Rosetta scorefile and list of important residues
import argparse
import pandas as pd
import os
import gzip
```

```
import sys
import numpy as np
def print_interaction_list(inter_list, f):
   if len(inter_list) < 1:
      print "None."
   else:
      labels = ["groups","residues"]
      df = pd.DataFrame(inter_list, columns=labels)
      groups = df.groups.unique()
      resis = df.residues.unique()

      groups.sort()
      resis.sort()

      resi_row = "\t"
      for r in resis:
         resi_row += '{:>6}'.format(r)
      f.write("%s\n" % resi_row)
      for g in groups:
         group_row = str(g)+"\t"
         for r in resis:
            group_row += '{:>6}'.format(len(df.loc[(df['groups'] == g) & (df["residues"]
                ↪ == r)]))
         f.write("%s\n" % group_row)

class System:
   def __init__(self):
      self.residue_dictionary = {}
      self.last_residue_number = 0

   def add_atom(self, atomLine):
      residue_number = atomLine[22:26].strip()
      residue_name = atomLine[17:20].strip()

      if (int(residue_number) > self.last_residue_number):
         self.last_residue_number = int(residue_number)

      if residue_number in self.residue_dictionary:
         self.residue_dictionary[residue_number].add_atom(atomLine)
      else:
         tmpRes = Residue(residue_number, residue_name)
```

```
            self.residue_dictionary[residue_number] = tmpRes
            self.residue_dictionary[residue_number].add_atom(atomLine)

    def get_Residue(self, resi_name):
        return self.residue_dictionary[resi_name]

    def get_last_residue_number(self):
        return self.last_residue_number

    def check_if_residue_exists_in_system(self, resi_num):
        if str(resi_num) in self.residue_dictionary:
            return 1
        else:
            return 0

    def get_all_residues(self):
        return self.residue_dictionary

    def get_all_atoms(self):
        all_atoms = {}
        for resi in self.residue_dictionary:
            all_atoms.update(self.residue_dictionary[resi].get_all_atoms())

        return all_atoms



class Residue:
    def __init__(self, resiNum, resiName):
        self.num = resiNum
        self.resi_type = resiName
        self.atom_dictionary = {}

    def add_atom(self, atomLine):
        tmpAtom = Atom(atomLine)
        self.atom_dictionary[tmpAtom.getAtomType()] = tmpAtom

    def get_Atom(self, atom_name):
        return self.atom_dictionary[atom_name]

    def get_Resi_Type(self):
        return self.resi_type
```

```
   def number_of_atoms(self):
      return len(self.atom_dictionary)

   def find_nearest_distance_to_point(self, coor1):
      shortest_dist = sys.float_info.max
      shortest_dist_atom_type = ""
      for atoms in self.atom_dictionary:
         coor2 = self.atom_dictionary[atoms].getCoor()
         dist = (((coor1[0]-coor2[0])**2) + ((coor1[1]-coor2[1])**2) + ((coor1[2]-coor2
             ↪ [2])**2))**0.5;

         if (dist < shortest_dist):
            shortest_dist = dist
            shortest_dist_atom_type = self.atom_dictionary[atoms].getAtomType()

      return dist

   def is_residue_atom_within_distance(self, coor1, cutoff):
      #print len(self.atom_dictionary)
      #print self.resi_type
      for atoms in self.atom_dictionary:
         coor2 = self.atom_dictionary[atoms].getCoor()
         dist = (((coor1[0]-coor2[0])**2) + ((coor1[1]-coor2[1])**2) + ((coor1[2]-coor2
             ↪ [2])**2))**0.5;
         #print coor2
         if (dist <= cutoff):
            return 1

      return 0

   def get_all_atoms(self):
      return self.atom_dictionary


class Atom:
   def __init__(self, atomLine):
      self.x = float(atomLine[30:38].strip())
      self.y = float(atomLine[38:46].strip())
      self.z = float(atomLine[46:54].strip())
```

```
      self.resiNum = atomLine[22:26].strip()
      self.resiName = atomLine[17:20].strip()
      self.atomType = atomLine[12:16].strip()
      self.chain = atomLine[21:22].strip()


   def getX(self):
      return self.x

   def getY(self):
      return self.y

   def getZ(self):
      return self.z

   def getCoor(self):
      coordinates = [self.x,self.y,self.z]
      return coordinates

   def getResiNum(self):
      return self.resiNum

   def getResiName(self):
      return self.resiName

   def getAtomType(self):
      return self.atomType

   def getElement(self):
      return self.atomType.lstrip('0123456789')[0:1]

   def getChain(self):
      return self.chain

def distance(coor1, coor2):
   dist = (((coor1[0]-coor2[0])**2) + ((coor1[1]-coor2[1])**2) + ((coor1[2]-coor2[2])**2))
       ↪ **0.5;
   return dist

def calcRMSD(lig1, lig2):
   distSum = 0
```

```
   ligand_1_atoms = lig1.get_all_atoms()
   ligand_2_atoms = lig2.get_all_atoms()
   for atoms in ligand_1_atoms:
      if (ligand_1_atoms[atoms].getElement() != "H"):
         coor1 = ligand_1_atoms[atoms].getCoor()
         coor2 = ligand_2_atoms[atoms].getCoor()
         dist = ((coor1[0]-coor2[0])**2) + ((coor1[1]-coor2[1])**2) + ((coor1[2]-coor2[2])
             ↪ **2);
         distSum += dist

   return ((1.0/len(ligand_1_atoms))*distSum)**0.5

def import_scorefile(args):
   selection = pd.read_csv(args.scorefile, delim_whitespace=True)[["description", "
       ↪ interface_delta_X"]]
   ligand_energy_threshold = selection["interface_delta_X"].min()+ args.energy_from_best
   return selection[ selection["interface_delta_X"] < ligand_energy_threshold]

def import_groupings(file_name):
   group_lines = open_file(file_name)

   group_list = []
   for lines in group_lines:
      atoms = lines.split()
      group_list.append(atoms)

   return group_list

def open_file(file_name):
   if (os.path.exists(file_name)):
      if file_name.endswith('.gz'):
         fileObj = gzip.GzipFile(file_name, 'rb')
         lines = fileObj.readlines()
         fileObj.close()
         return lines
      else:
         f = open(file_name, "r")
         lines = f.readlines()
         f.close()
         return lines
   else:
```

```
      print "Cannot open file "+file_name+", exiting...\n"
      sys.exit(1)

# check if the protein and ligand have accepted number of interaction
def check_important_interaction(protein, ligand, num, important_res):
   count = 0
   # Keep track of which residues already contact ligand
   res_check=np.zeros(len(important_res))
   for group in range(0, len(group_list)):
      #print group
      #print group_list[group]
      for atom in group_list[group]:
         #print atom
         lig_atom_coor = ligand.get_all_atoms()[atom].getCoor()
         #print lig_atom_coor
         i=0
         while i < len(important_res):
            while res_check[i] == 1:
               i=i+1
               if i == len(important_res):
                  break
            if i == len(important_res):
               break
            resnum = important_res[i]
            res = protein.residue_dictionary[resnum]
            #print res.get_Resi_Type()
            #print group_list[0]
            if res.is_residue_atom_within_distance(lig_atom_coor, args.distance):
               res_check[i] = 1
               count = count + 1
               #print count
               if (count >= num) :
                  return True
            i=i+1
   return False

##########Code in action#####
# Read in arguments
parser = argparse.ArgumentParser(description='From a Rosetta scorefile determine
    ↪ statistics of ligand pose clusters')
parser.add_argument("scorefile", default="score.sc", help="file name of the rosetta
```

```
    ↪ scorefile")
parser.add_argument("-g","--groupings", help="divide ligands into groups and show stats by
    ↪  group")
parser.add_argument("-d","--distance", type=float, default=5, help="distance cutoff")
parser.add_argument("-e","--energy_from_best", type=int, default=5, help="interface energy
    ↪  included from best score")
#parser.add_argument("-r","--rmsd", type=float, default=150, help="rmsd to cluster ligand
    ↪ poses by")
parser.add_argument("-p","--params", help="ligand params file")
parser.add_argument("-i","--residues", help="list of important residues")
parser.add_argument("-n","--inteaction_num", type=int, help="minimum numberof interactions
    ↪  to important residues")
parser.add_argument("-s","--structure", type=bool, default=False, help="True/False: print
    ↪ out the list of structures")
parser.add_argument("-o","--output", default="output.txt", help="name of output file")

args = parser.parse_args()

# Import files
scorefile_table = import_scorefile(args)
group_list = import_groupings(args.groupings)
param_file_lines = open_file(args.params)
important_residues = import_groupings(args.residues)[0]
#print len(important_residues)
atom_name_to_rosetta_atom_type = {}

for i in param_file_lines:
   if i[0:6] == "ATOM ":
      atom_line_info = i.split()
      atom_name_to_rosetta_atom_type[atom_line_info[1]] = atom_line_info[2]

# initialize lists of ligands, protein, energy, pose name
ligand_dict = {}
protein_dict = {}
energy_dict = {}
pose_list = []

for index, row in scorefile_table.iterrows():
   pdbfilename = row["description"]+".pdb"
   interface_energy = row["interface_delta_X"]
   pose_name = row["description"]
```

```
   #print pose_name

   pdb_file_lines = open_file(pdbfilename)
   #read file and import protein and ligand
   protein = System()
   ligand = System()
   for i in pdb_file_lines:
      if i[0:6] == "ATOM ":
         protein.add_atom(i)
      if i[0:6] == "HETATM" and i[21:22] == "X":
         ligand.add_atom(i)
   #print ligand.get_all_residues()
   if check_important_interaction(protein, ligand, args.inteaction_num, important_residues
       ↪ ):
      #updates lists of ligand and protein and energy
      ligand_dict[pose_name] = ligand
      protein_dict[pose_name] = protein
      energy_dict[pose_name] = interface_energy
      pose_list.append(pose_name)
#print len(pose_list)
# print out the tables
hbond_interaction_list = []
resi_interaction_list = []
for pose in pose_list:
   #print pose
   curr_prot_pose = protein_dict[pose]
   curr_lig_pose = ligand_dict[pose]

   # Calc Contact Frequency
   for residue_num in curr_prot_pose.residue_dictionary:
      residue = curr_prot_pose.residue_dictionary[residue_num]
      #print residue_num
      for group in range(0, len(group_list)):
         for atom in group_list[group]:
            lig_atom_coor = curr_lig_pose.get_all_atoms()[atom].getCoor()
            if (residue.is_residue_atom_within_distance(lig_atom_coor, args.distance)):
               resi_interaction_list.append([int(group+1), int(residue_num)])
               break

   # Calc Hbond contacts
   ligand_atoms = curr_lig_pose.get_all_atoms()
```

```
   for atom in ligand_atoms:
      #print atom
      lig_atm_type = atom_name_to_rosetta_atom_type[atom]
      if (lig_atm_type == "ONH2" or lig_atm_type == "OH" or lig_atm_type == "Oaro" or
          ↪ lig_atm_type == "OOC"):
         protein_residues = curr_prot_pose.get_all_residues()
         for residue in protein_residues:
            resi_type = protein_residues[residue].get_Resi_Type()
            polar_resi_list = ["ARG", "ASN", "GLN", "HIS_D", "HIS", "LYS", "SER", "THR", "
                ↪ TYR"]
            if ( resi_type in polar_resi_list):
               resi_atoms = protein_residues[residue].get_all_atoms()
               polar_hydrogens = ["HG","HH","HG1", "1HH1", "2HH1", "1HH2", "2HH2", "HE",
                   ↪ "1HD2", "2HD2", "1HE2", "2HE2", "HD1", "HE2", "1HZ", "2HZ", "3HZ"]
               for ph in polar_hydrogens:
                  if ph in resi_atoms.keys():
                     hbond_dist = distance(ligand_atoms[atom].getCoor(), resi_atoms[ph].
                         ↪ getCoor())
                     if ( hbond_dist < 3.0):
                        hbond_interaction_list.append([atom, int(resi_atoms[ph].getResiNum
                            ↪ ())])


      if (lig_atm_type == "Hpol"):
         protein_residues = protein.get_all_residues()
         for residue in protein_residues:
            resi_type = protein_residues[residue].get_Resi_Type()
            polar_resi_list = ["ASP", "GLU", "ASN", "GLN", "SER", "TYR", "THR"]
            if ( resi_type in polar_resi_list):
               resi_atoms = protein_residues[residue].get_all_atoms()
               polar_oxygen = ["OD1","OD2","OE1","OE2","OG","OG1","OXT","OH"]
               for po in polar_oxygen:
                  if po in resi_atoms.keys():
                     hbond_dist = distance(ligand_atoms[atom].getCoor(), resi_atoms[po].
                         ↪ getCoor())
                     if ( hbond_dist < 3.0):
                        hbond_interaction_list.append([atom, int(resi_atoms[po].getResiNum
                            ↪ ()) ] )
f = open(args.output, 'w')
if (args.structure == True):
   for item in pose_list:
      f.write("%s\n" % item)
```

```
# Print Contact Frequency stats
f.write("\nRESIDUE INTERACTION TABLE\n")
print_interaction_list(resi_interaction_list, f)
f.write("\n")

#Print Hbond stats
f.write("\nHBOND CONTACT TABLE\n")
print_interaction_list(hbond_interaction_list, f)
f.close()
```

Those filtered models were then clustered based the ligand position with 3 Å RMSD cut-off. For each of five largest clusters, the interaction contact and strength were computed for each cluster such that the scores are higher when an important residue has high contact frequency to the ligand or favorable binding energy toward the ligand.

The interaction contact score is computed as: $Interactioncontact(t) = \sum ct_i \times 4.7 - ct_n$,where ct_i and ct_n are the percent of docking poses that have the ligand contact with residues that have been identified as important and residues that were not confirmed as critical for binding residue, respectively. The weight of 4.7 was chosen to balance the impact of 13 critical with 61 non-critical residues. Similarly, the interaction strength of a cluster t is calculated as $Interactionstrength(t) = \sum ddg_i \times -4.7 + ddg_n$, where dd_gi and ddg_n are the predicted binding energies between the ligand and an important residue and non-critical residue, respectively. In the cluster with the most favorable interaction contact and strength scores, a model that best explains the mutagenesis data will be chosen as a putative binding pose of VU0637120 and Y4R.

7.2.3 Docking result analysis : Docking and ddG analysis

7.3 Protocol capture for chapter 4 : BCL::Mol2D – a robust atom environment descriptor for QSAR modeling and pharmacophore mapping

7.3.1 Download and install BCL

The Bio Chemical Library (BCL) is a software package that provides a suite of cheminformatics tools that allow construction of quantitative structure-activity-relation (QSAR) models for virtual screening, pharmacophore mapping, and drug design. BCL is free of charge for non-comercial use.

Follow the instruction on meilerlab.org/index.php/servers/bcl-academic-license to obtain access to the source code of BCL.

7.3.2 Benchmark different descriptor configurations in QSAR tasks

In the benchmark folder:

Make dataset bin files Download the smi string file of nine PubChem datasets at http://www.meilerlab.org/jobs/downloadfile/name/qsar_benchmark_smiles.zip. Convert the smi string files to SDF files with 3D conformation using OpenBabel and Corina as described in the method section of the paper. Store the sdf files in the data folder. We will use the BCL descriptor:GenerateDataset application to convert the SDF files to binary files contains the molecular fingerprints for the training compounds

```
#!/bin/bash
bcl=/path/to/BCL/build/linux64_release/bin/bcl-apps-static.exe

# Number of cores used
cpu_n=10

# Make bin files for active and inactive sdf files
for f in '1798' '1834' '2258' '2689' '435008' '435034' '463087' '485290' '488997'
do
 for descriptor in 'Element1RSR' 'Atom1RSR' 'Element2' 'Atom2' 'Element1' 'Atom1' 'RSR'
 do
   $bcl descriptor:GenerateDataset -source 'SdfFile(filename='$f'.active.sdf.gz)' -
       ↪ feature_labels ../feature_labels/${descriptor}.object -result_labels "1" -output
       ↪  ${f}_active_${descriptor}.bin -scheduler PThread $cpu_n
   $bcl descriptor:GenerateDataset -source 'SdfFile(filename='$f'.inactive.sdf.gz)' -
       ↪ feature_labels ../feature_labels/${descriptor}.object -result_labels "0" -output
       ↪  ${f}_inactive_${descriptor}.bin -scheduler PThread $cpu_n

   # Combine and then randomize the active and inactive molecules from each set (see for
       ↪ loop)
   $bcl descriptor:GenerateDataset -source 'Randomize(Combined(Subset(filename='$f'
       ↪ _active_'${descriptor}'.bin), Subset(filename='$f'_inactive_'${descriptor}'.bin)
       ↪ ))' -output ${f}_${descriptor}.bin -scheduler PThread $cpu_n
   rm ${f}_active_${descriptor}.bin ${f}_inactive_${descriptor}.bin
 done
done
```

In the data folder, make the bin data files by running:

```
./make_bin_files.sh
```

7.3.3 Train the QSAR models

Train and benchmark the QSAR models with different descriptor configurations by running this script:

There are 4 object files corresponding to 4 different descriptor sets

Atom1.object

```
Combine(
  UMol2D(
  atom hashing type=Atom,
  feature size=574,
  Atom environment height=1
  )
)
```

Element1.object

```
Combine(
  UMol2D(
  atom hashing type=Element,
  feature size=240,
  Atom environment height=1
  )
)
```

Atom2.object

```
Combine(
      MolPrint2D(
      atom hashing type=Atom,
      feature size=8080,
      Atom environment height=2
      )
)
```

Element2.object

```
Combine(
  UMol2D(
  atom hashing type=Element,
  feature size=5112,
  Atom environment height=2
  )
)
```

RSR.object

```
# Descriptor set described in supplement for
# Mendenhall, Meiler "Advances in Machine Learning Applied to Quantitative Structural
    ↪ Activity Relationship Modeling"
# Unpublished 2015
# 391 columns total
# This descriptor set performed equivalently to the short-range "Minimal" variant, yet has
    ↪  only 1/3 of the total
# descriptors due to careful selection of the atom properties used when computing 2DA and
    ↪ 3DA.
Combine(
  # Max # of bonds between any two atoms in the molecule
  Define(BondGirth=DescriptorSum(2DAMax(steps=96,property=Atom_Identity,substitution_value
      ↪ =nan))),

  # 1 For H, -1 for heavy atoms
  Define(IsHTernary=Add(Constant(-1),Multiply(IsH,Constant(2)))),

  # 1 for H-Bond donors (O or N that have bond to an H), -1 for H-Bond Acceptors (any O or
      ↪  N) that are not donors,
  # 0 for all other atoms
  Define(Atom_IsInAromaticRing=GreaterEqual(lhs=BondTypeCount(property=IsAromatic,value=1)
      ↪ ,rhs=2)),
  Define(Atom_IsInAromaticRingTernary=Add(Constant(-1),Multiply(Atom_IsInAromaticRing,
      ↪ Constant(2)))),

  # Whether an atom is at the intersection of two aromatic rings (commonly due to ring
      ↪ fusion, but rarely spiro too)
  Define(Atom_InAromaticRingIntersection=GreaterEqual(lhs=BondTypeCount(property=
      ↪ IsAromatic,value=1),rhs=3)),
  Define(Atom_InRingIntersection=GreaterEqual(lhs=BondTypeCount(property=IsInRing,value=1)
      ↪ ,rhs=3)),

  # Scalar descriptors (1 number each)
  Weight,
  HbondDonor,
  HbondAcceptor,
  LogP,
  TotalCharge,
  NRotBond,
  NAromaticRings,
```

```
NRings,
TopologicalPolarSurfaceArea,
Girth,
BondGirth,
MaxRingSize,
Limit(MinRingSize,max=8,min=0),
MoleculeSum(Atom_InAromaticRingIntersection),
MoleculeSum(Atom_InRingIntersection),
MoleculeStandardDeviation(Atom_Vcharge),
MoleculeStandardDeviation(Atom_SigmaCharge),
MoleculeMax(Atom_Vcharge),
MoleculeMax(Atom_SigmaCharge),
MoleculeMin(Atom_Vcharge),
MoleculeMin(Atom_SigmaCharge),
MoleculeSum(Abs(Atom_Vcharge)),
MoleculeSum(Abs(Atom_SigmaCharge)),

# Sign-aware 2DA's, out to 11 bonds (36 numbers each)
# Partial is used to exclude the bin at index 2, which corresponds to when atom property
    ↪ ^2 is negative, which does
# not occur since all atom properties return real numbers
Template(
 signature=2DASign11(X),
 Partial(
  2DASign(property=X,steps=11),
  indices(
  0,1,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,
  19,20,21,22,23,24,25,26,27,28,29,30,31,32
  )
 )
),

# Sign-aware 3DA's, out to 6A, beyond which rotomer-dependent effects begin to play a
    ↪ significant role
# The partial is used here to remove the first 1A of data, which is always redundant
    ↪ because the 0A bin is
# identical to the 2DA case, and the remaining bins 0.25, 0.5, and 0.75 are generally 0
Template(
 signature=3DASign24(X),
 Partial(
  3daSmoothSign(property=X,step size=0.25,temperature=100,steps=24,gaussian=False,
```

```
        ↪ interpolate=True),
    indices(
      12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,
      7,28,29,30,31,32,33,34,35,36,37,38,39,40,41,
      42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,
      57,58,59,60,61,62,63,64,65,66,67,68,69,70,71
    )
   )
  ),
  2DASign11(Atom_SigmaCharge),
  2DASign11(Atom_Vcharge),
  2DASign11(IsHTernary),
  2DASign11(Atom_IsInAromaticRingTernary),
  3DASign24(Atom_SigmaCharge),
  3DASign24(Atom_Vcharge),
  3DASign24(IsHTernary),
  3DASign24(Atom_IsInAromaticRingTernary)
)
```

Atom1RSR.object, Element1RSR.object are simply the combination of Atom1 or Element1 and the RSR descriptor set.

The command.config file contains different options for BCL model::Train application to train the machine learning models of interest

```
[bcl]
# specify bcl executable
bcl: /path/to/bcl/build/linux64_release/bin/bcl-apps-static.exe
[main]
[variables]
# activity cutoff used in objective functions
cutoff: 0.5
# parity determines whether values smaller then the cutoff are considered active (0) or
    ↪ inactive (1)
parity: 1
objective-function: 'Bootstrap(repeats=2000, function=AucRocCurve(cutoff=0.5,parity=1,
    ↪ x_axis_log=1,min fpr=0.001,max fpr=0.1),
visdrop: 0.05
hiddrop: 0.25
hiddenneurons: 32
alpha: 0.5
eta: 0.05
```

```
balanceratio: 0.1
droptype: Zero
[learning]
learning-method: 'NeuralNetwork( transfer function = Sigmoid, weight update = Simple(alpha
    ↪ =%(alpha)s,eta=%(eta)s),dropout(%(visdrop)s,%(hiddrop)s),objective function = %(
    ↪ objective-function)s,scaling=AveStd,steps per update=1,hidden architecture(%(
    ↪ hiddenneurons)s), balance=True,balance target ratio=%(balanceratio)s,shuffle=True,
    ↪ input dropout type=%(droptype)s)'
# maximum training iterations of chosen learning-method
max-iterations: 100
max-minutes: 40000
result_averaging_window: 0
# choose one type of training data data assembly
'InformationGainRatio(cutoff=%(cutoff)s,measure='PPV',parity=%(parity)s)'
[score]
# choose one dataset scoring type
scoring-type: InformationGain
[cv]
monitoring-id-range: [0,4]
independent-id-range: [0,4]
cross-validations: 5
cv-repeats: 1
override-memory-multiplier: 1.5
```

The benchmark script qsar_train.sh will train the ANN network on all 9 pubchem datasets and seven different descriptor sets ('Element1RSR', 'Atom1RSR', 'Element2', 'Atom2', 'Element1', 'Atom1', and 'RSR')

```
#!/bin/bash

#number of local cpu cores
cpu_n=10

# train QSAR models
for descriptor in 'Element1RSR' 'Atom1RSR' 'Element2' 'Atom2' 'Element1' 'Atom1' 'RSR'
do

  #run locally
  cat datasets.lst | xargs -n1 -I@ python /path/to/BCL/scripts/machine_learning/launch.py
      ↪ -t cross_validation --config-file qsar.config --datasets data/@_${descriptor}.bin
      ↪ --id @_${descriptor} --features feature_labels/${descriptor}.object --opencl None
```

```
    ↪ --local $cpu_n --override-memory-multiplier 1.5 --cv-repeats 2


 # For running on computer cluster using srun
 #cat datasets.lst | xargs -n1 -I@ python /path/to/BCL/scripts/machine_learning/launch.py
    ↪  -t cross_validation --config-file qsar.config --datasets data/@_${descriptor}.bin
    ↪  --id @_${descriptor} --features feature_labels/${descriptor}.object --slurm --
    ↪ just-submit --no-flock-submit --opencl None --override-memory-multiplier 1.5 --cv-
    ↪ repeats 2
done
```

Run this script in the main directory. The training should take around several hours per ANN model

```
bash qsar_train.sh
```

7.3.4 Sensitivity analysis

In the sensitivity_analysis folder:

Compute the decrement and increment dervatives of different atom environments for each active and inactive compound by running the compute_sensitivity_scores.sh script

```
#!/bin/bash
#This script compute increment and decrement derivatives for each atom environement (AE)
    ↪ in the AE library (see file:AE_library/Atom_AE_1_bond_sorted.txt )
#Author: Oanh Vu
#Date: Jan 2019

#specify path to bcl executable
#bcl=/path/to/BCL/build/linux64_release/bin/bcl-apps-static.exe
# Specify number of cpu cores used
cpu_n=10

# generate bin files
cat sdf.lst| xargs -n1 -I@ -P $cpu_n $bcl descriptor:GenerateDataset -source 'SdfFile(
    ↪ filename=@.sdf)' -feature_labels Atom1.object -result_labels '0' -output @.bin

#compute increment derivatives
cat sdf.lst | xargs -n1 -I@ -P $cpu_n $bcl descriptor:ScoreDataset -source 'Subset(
    ↪ filename=@.bin)' -output @_increment.score -opencl Disable -score '
    ↪ InputSensitivityDiscrete(derivative=Increment,storage=File(directory=qsar_2689,
    ↪ prefix=model), weights=(consistency=0.0,square=0.0,absolute=0.0,utility=0.0,average
    ↪ =1.0,consistency best=0.0,balance=False,categorical=False))' -feature_labels Atom1.
    ↪ object
```

```
#compute decrement derivatives
cat sdf.lst | xargs -n1 -I@ -P $cpu_n $bcl descriptor:ScoreDataset -source 'Subset(
    ↪ filename=@.bin)' -output @_decrement.score -opencl Disable -score '
    ↪ InputSensitivityDiscrete(derivative=Decrement,storage=File(directory=qsar_2689,
    ↪ prefix=model), weights=(consistency=0.0,square=0.0,absolute=0.0,utility=0.0,average
    ↪ =1.0,consistency best=0.0,balance=False,categorical=False))' -feature_labels Atom1.
    ↪ object
```

The github repository https://github.com/vuoanh/BCL_Mol2D_benchmark contains all the related files and scripts mentioned in this protocol capture.

7.4 Protocol capture for chapter 5: Mapping the binding sites of UDP and prostaglandin E2 glyceryl ester in the nucleotide receptor P2Y6

For the protocol of creating homology modeling of P2Y6 and induced-fit docking of UDP and PGE2G to P2Y6, please refer to the corresponding protocols in the protocol capture in the section 7.2.2.

7.4.1 Convention molecular simulation of the output docking models of UDP or PGE2G to P2Y6

Selected docking models were then refined with molecular dynamics simulation. For each ligand, two independent replicates with around two μs in total simulation time were conducted. Downser++(Morozenko and Stuchebrukhov, 2016) were then used to dock waters inside the transmembrane region of P2Y6 in the presence of the ligands. Water molecules are docked into two top docking poses using Dowser++. Then the models are then minimized with implicit solvents and Amber14 forcefield in MOE. The output pdb models are then converted to amber atom names and format with pdb4amber

```
pdb4amber -i pge2g-p2y6-water.pdb --add-missing-atoms -l pdb_processing_p2y6_pge2g.log -o
    ↪ p2Y6_pge2g_amber.pdb --reduce --no-reduce-db
```

Make sure that the name of the ligand is set to P2G in the pdb file. All membrane systems were built with the membrane building tool PackMol-Memgen.

```
packmol-memgen --lipids CHL1:POPC//CHL1:POPC --ratio $1:10//1:10 --dist_wat 25 --plot --
    ↪ log packmol_p2Y6_pge2g.log --output p2Y6_pge2g_packmol.pdb --salt --leaflet 23 --
    ↪ verbose --keeplig --pdb p2Y6_pge2g_amber.pdb --ffprot ff19SB --ffwat TIP3P --
    ↪ tailplane 3
```

The output bi-membrane system contained POPC and Cholesterol with a molecule number ratio of 10:1. Proteins, lipids, TIP3P water, and ions were modeled with the FF19SB(Maier et al., 2015) and Amber Lipid17(Gould et al., 2018) force fields, and the ligands were modeled with the GAFF2 small molecule

force field(He et al., 2020; Wang et al., 2004). A TIP3P water layer of 25 Å was included, and Cl− or K+ ions were added to neutralize the charge of the system.

```
#!/bin/bash
# Input variables
PDB=$1 # Input system PDB file
param=$2 # param file of the ligand
prepi=$3 # prepi file of the ligand
output_dir=`readlink -e $4` #directory of the output files
water=$5
ds_bonds=$6 #file with info of additional covalent bonds

# Derived variables
tag=`basename ${PDB} .pdb`

# Make tleap file
cat > ${output_dir}/tleap_${tag}_${water}.in << EOF
source leaprc.lipid17
source leaprc.protein.ff19SB
source leaprc.water.${water}
source leaprc.gaff2
loadamberparams ${param}
loadAmberPrep ${prepi}
SYS = loadpdb ${PDB}
setBox SYS vdw
addionsrand SYS K+ 0
addionsrand SYS Cl- 0
savepdb SYS ${output_dir}/${tag}_out.pdb
EOF

if [[ $(wc -l < $ds_bonds) -ge 1 ]]
then
   awk '{print "bond SYS."$1".SG SYS."$2".SG"}' $ds_bonds >> ${output_dir}/tleap_${tag}
       ↪ _tip3p.in
fi
echo "saveAmberParm SYS ${output_dir}/${tag}/${tag}.parm7 ${output_dir}/${tag}/${tag}.crd"
    ↪  >> ${output_dir}/tleap_${tag}_tip3p.in
echo "quit" >> ${output_dir}/tleap_${tag}_tip3p.in

# Run tleap
echo output pdb file ${output_dir}/${tag}_out.pdb
```

```
echo param files of the system is ${output_dir}/${tag}.parm7 ${output_dir}/${tag}.crd
tleap -f tleap_${tag}_tip3p.in
```

running the tleap script to solvate and neutralize the system

```
ligand_name=PEG ; tag=p2Y6_pge2g
bash ~vuot2/bin/MD/scripts/tleap_water_to_receptor_ligand_MB_complex.sh ${tag}.pdb PEG_${
    ↪ tag}.frcmod ${ligand_name}_${tag}.prepi . $water_type $disulfile
```

Each bilayer system was first minimized for 5,000 steps using steepest descent followed by 15,000 steps of conjugate gradient minimization.

```
#!/bin/tcsh
# Number of amino acids and number of atoms composing amino acids (excluding water and
    ↪ ions)
set numAA = 302
set numAALip = <Add the number residue + number of lipid +1>
@ lipidStart = $numAA + 1
@ wationStart = $numAALip + 1

# Working directory for input/output files
set tag = p2Y6_pge2g
set dirname = <current directory>
ls -alh $dirname
cd $dirname
set parm = $dirname/${tag}.parm7
set start_crd = $dirname/${tag}.crd
echo minimization for ${tag}

##########################################
##### Minimization #####
##########################################

cat > ${dirname}/${tag}_minlip.in << EOF
Initial minimization
 &cntrl
  ntx = 1, irest = 0, ntrx = 1, ntxo = 1,
  ntpr = 20, ntwx = 0, ntwv = 0, ntwe = 0,
  ntf = 1, ntb = 1,
  es_cutoff = 10.0,
  vdw_cutoff = 10.0,
  ibelly = 0, ntr = 1,
```

```
 imin = 1,
 maxcyc = 15000,
 ncyc = 5000,
 ntmin = 1, dx0 = 0.1, drms = 0.0001,
 ntc = 1, tol = 0.00001,

 &end
Hold protein fixed
1000.0
RES 1 $numAA
END
Hold Water/ions fixed
1000.0
RES $wationStart 999999
END
END
EOF

cat > ${dirname}/${tag}_minsolv.in << EOG
Initial minimization
 &cntrl
 ntx = 1, irest = 0, ntrx = 1, ntxo = 1,
 ntpr = 20, ntwx = 0, ntwv = 0, ntwe = 0,
 ntf = 1, ntb = 1,
 es_cutoff = 10.0,
 vdw_cutoff = 10.0,
 ibelly = 0, ntr = 1,
 imin = 1,
 maxcyc = 15000,
 ncyc = 5000,
 ntmin = 1, dx0 = 0.1, drms = 0.0001,
 ntc = 1, tol = 0.00001,

 &end
Hold protein fixed
1000.0
RES 1 $numAA
END
END
EOG
```

```
# Minimization script with solvent restraints
cat > ${dirname}/${tag}_minpro.in << EOH
Initial minimization
 &cntrl
  ntx = 1, irest = 0, ntrx = 1, ntxo = 1,
  ntpr = 20, ntwx = 0, ntwv = 0, ntwe = 0,
  ntf = 1, ntb = 1,
  cut = 10.0, nsnb = 10,
  ibelly = 0, ntr = 1,

  imin = 1,
  maxcyc = 15000,
  ncyc = 2000,
  ntmin = 1, dx0 = 0.1, drms = 0.0001,
  ntc = 1, tol = 0.000001,
 &end
Hold solvent fixed
1000.0
RES $lipidStart 9999999
END
END
EOH

# Minimization script with no restraints
cat > ${dirname}/${tag}_minall.in << EOI
Initial minimization
 &cntrl
  ntx = 1, irest = 0, ntrx = 1, ntxo = 1,
  ntpr = 20, ntwx = 0, ntwv = 0, ntwe = 0,
  ntf = 1, ntb = 1,
  es_cutoff = 10.0, nsnb = 10,vdw_cutoff = 10.0,
  ibelly = 0, ntr = 0,
  imin = 1,
  maxcyc = 15000,
  ncyc = 500,
  ntmin = 1, dx0 = 0.1, dxm = 0.5, drms = 0.0001,
  ntc = 1, tol = 0.000001,
 &end
EOI

echo minimization for ${tag}
```

```
# Minimize the lipid
# mpirun -np 16 $AMBERHOME/bin.MPI/pmemd.MPI -O \
pmemd.cuda -O -i ${tag}_minlip.in -o ${tag}_minlip.out -p $parm -c ${start_crd} -ref ${
    ↪ start_crd} -r ${tag}_minlip.crd > ${tag}_minlip.log

# Minimize the solvent
# mpirun -np 16 $AMBERHOME/bin.MPI/pmemd.MPI -O \
pmemd.cuda -O -i ${tag}_minsolv.in -o ${tag}_minsolv.out -p $parm -c ${tag}_minlip.crd -
    ↪ ref ${tag}_minlip.crd -r ${tag}_minsolv.crd > ${tag}_minsolv.log

# Minimize the protein
pmemd.cuda -O -i ${tag}_minpro.in -o ${tag}_minpro.out -p $parm -ref ${tag}_minsolv.crd -c
    ↪  ${tag}_minsolv.crd -r ${tag}_minpro.crd > ${tag}_minsolv.log

# Minimize everything
pmemd.cuda -O -i ${tag}_minall.in -o ${tag}_minall.out -p $parm -c ${tag}_minpro.crd -r ${
    ↪ tag}_minall.crd > ${tag}_minall.log
echo DONE >> ${tag}_minall.log
```

During heating, the protein backbone and sidechain atoms, lipid and water were restrained to their starting coordinates with harmonic force constants of 10 and 5, heated to 10 K over 1,000 steps with a step size of 1 fs using constant boundary conditions and Langevin dynamics with a rapid collision frequency of 10,000 ps^{-1}.

The system was then heated to 100 K over 50,000 steps with constant volume dynamics and the collision frequency set to 1000 ps^{-1} and, finally, to 303 K over 100,000 steps with constant pressure dynamics and anisotropic pressure scaling turned on, while the positional restraints on the system were gradually removed.

```
# Number of Amino Acids and number of Amino Acids + Lipids for this system
set numAA = 302
set numAALip = <Add the number residue + number of lipid +1>
@ lipidStart = $numAA + 1
@ wationStart = $numAALip + 1

# Working directory for input/output files
set tag = p2Y6_pge2g
set dirname = <working driectory>
cd $dirname
set parm_file = $dirname/${tag}.parm7
set start_crd = $dirname/${tag}_minall.crd
```

```
echo heating for ${tag}
#################################################################
##### Heat to 10K - 0.1ps (1000 steps 0.1fs) #####
#################################################################

# Restraint forces
set ProBB = 10.0
set ProSC = 5.0
set LH = 2.5
set LT = 2.5

echo heating for ${tag}

cat << EOF > ${tag}_heat10K.in
Careful Lipid heating 10K - bad contacts assumed
&cntrl
imin=0, ! Molecular dynamics (no minimization)
ntx=1, ! Positions read formatted with no initial velocities
irest=0, ! No restart, run a new simulation
ntc=2, ! SHAKE on for bonds with hydrogen
ntf=2, ! No force evaluation for bonds with hydrogen
tol=0.0000001, ! SHAKE tolerance
nstlim=1000, ! Number of MD steps
ntt=3, ! Langevin dynamics
gamma_ln=10000.0, ! Collision frequency for Langevin dynamics
ntr=1, ! Restrain atoms using a harmonic potential (See the GROUP input below)
ig=-1, ! Random seed for Langevin dynamics
ntpr=10, ! Every ntpr steps, information will be printed to mdout and mdinfo
ntwx=10, ! Write to trajectory file every ntwx steps
dt=0.0001, ! Timestep (ps) 0.1 fs
nmropt=1, ! Read the weight change information below
ntb=1,  ! Constant volume
ntp=0,  ! No pressure scaling
cut=10.0, ! classical non-bond cut off
&end
&wt
type='TEMP0', ! Varies the target temperature TEMP0
istep1=1, ! Initial step
istep2=5000, ! Final step
value1=0.0, ! Initial temp0 (K)
value2=10.0, ! final temp0 (K)
```

```
&end
&wt
type='TEMP0', ! Varies the target temperature TEMP0
istep1=5001, ! Initial step
istep2=10000, ! Final step
value1=10.0, ! Initial temp0 (K)
value2=10.0, ! final temp0 (K)
&end
&wt
type='END',
&end ! End of varying conditions
Protein backbone restraints
$ProBB
FIND
* * M *
SEARCH
RES 1 $numAA
END
Protein non-main-chain restraints
$ProSC
FIND
* * S *
* * B *
* * 3 *
* * E *
SEARCH
RES 1 $numAA
END
Lipid head restraints
$LH
FIND
* * * PC
SEARCH
RES $lipidStart $numAALip
END
Lipid tail restraints
$LT
FIND
* * * PA
* * * OL
SEARCH
```

```
RES $lipidStart $numAALip
END
END ! End GROUP input
EOF

pmemd.cuda -O -p $parm_file -c $start_crd -ref $start_crd -i ${tag}_heat10K.in -o ${tag}
    ↪ _heat10K.out -r ${tag}_heat10K.crd -x ${tag}_heat10K.nc > ${tag}_heating.log
cpptraj -p kcne1.prmtop -y heat10K.crd -x heat10K.pdb

#####################################################################
##### Heat to 100K - 5ps (50,000 steps 0.1fs) #####
#####################################################################

# Restraint forces
set ProBB = 10.0
set ProSC = 5.0
set LH = 2.5
set LT = 2.5

cat << EOF > ${tag}_heat100K.in
Careful Lipid heating 100K
&cntrl
imin=0, ! Molecular dynamics
ntx=1, ! Positions read formatted with no initial velocities
irest=0, ! No restart
ntc=2, ! SHAKE on for bonds with hydrogen
ntf=2, ! No force evaluation for bonds with hydrogen
tol=0.0000001, ! SHAKE tolerance
nstlim=50000, ! Number of MD steps
ntt=3, ! Langevin dynamics
gamma_ln=1000.0, ! Collision frequency for Langevin dynamics
ntr=1, ! Restrain atoms using a harmonic potential (See the GROUP input below)
ig=-1, ! Random seed for Langevin dynamics
ntpr=1000,
ntwx=1000, ! Write to trajectory file every ntwx steps
dt=0.0001, ! Timestep (ps) 0.1 fs
nmropt=1, ! Read the weight change information below
ntb=1,
ntp=0,
cut=10.0,
&end
```

```
&wt
type='TEMP0', ! Varies the target temperature TEMP0
istep1=1, ! Initial step
istep2=50000, ! Final step
value1=0.0, ! Initial temp0 (K)
value2=100.0, ! final temp0 (K)
&end
&wt
type='TEMP0', ! Varies the target temperature TEMP0
istep1=50001, ! Initial step
istep2=500000, ! Final step
value1=100.0, ! Initial temp0 (K)
value2=100.0, ! final temp0 (K)
&end
&wt
type='END',
&end ! End of varying conditions
Protein backbone restraints
$ProBB
FIND
* * M *
SEARCH
RES 1 $numAA
END
Protein non-main-chain restraints
$ProSC
FIND
* * S *
* * B *
* * 3 *
* * E *
SEARCH
RES 1 $numAA
END
Lipid head restraints
$LH
FIND
* * * PC
SEARCH
RES $lipidStart $numAALip
END
```

```
Lipid tail restraints
$LT
FIND
* * * PA
* * * OL
SEARCH
RES $lipidStart $numAALip
END
END ! End GROUP input
EOF

pmemd.cuda -O -p $parm_file -c ${tag}_heat10K.crd -ref ${tag}_heat10K.crd -i ${tag}
    ↪ _heat100K.in -o ${tag}_heat100K.out -r ${tag}_heat100K.crd -x ${tag}_heat100K.nc >>
    ↪  ${tag}_heating.log
#cpptraj -p kcne1.prmtop -y heat100K.crd -x heat100K.pdb

######################################################################
##### Initial Heat to 303K - 100s (200,000 steps 0.05fs) #####
######################################################################

# Restraint forces
set ProBB = 10.0
set ProSC = 5.0
set LH = 2.5
set LT = 2.5

cat << EOF > ${tag}_heat303K-1.in
Heating 303K constant pressure dynamics with protein and lipid restraints
&cntrl
imin=0,
ntx=1, ! Positions read formatted with no initial velocities
irest=0, ! No Restart
ntc=2,
ntf=2,
tol=0.0000001,
nstlim=200000, ! Number of MD steps
ntt=3,
gamma_ln=100.0,
ntr=1,
ig=-1,
ntpr=2000,
```

```
ntwx=2000,
dt=0.00005,
nmropt=1, ! Read the weight change information below
ntb=2, ! Constant pressure periodic boundary conditions
ntp=1, ! Anisotropic pressure coupling
taup=2.0, ! Pressure relaxation time (ps)
cut=10.0,
&end
&wt
type='TEMP0',
istep1=1,
istep2=200000,
value1=50.0,
value2=303.0,
&end
&wt
type='TEMP0',
istep1=200001,
istep2=2000000,
value1=303.0,
value2=303.0,
&end
&wt
type='END',
&end
Protein backbone restraints
$ProBB
FIND
* * M *
SEARCH
RES 1 $numAA
END
Protein non-main-chain restraints
$ProSC
FIND
* * S *
* * B *
* * 3 *
* * E *
SEARCH
RES 1 $numAA
```

```
END
Lipid head restraints
$LH
FIND
* * * PC
SEARCH
RES $lipidStart $numAALip
END
Lipid tail restraints
$LT
FIND
* * * PA
* * * OL
SEARCH
RES $lipidStart $numAALip
END
END ! End GROUP input
EOF

pmemd.cuda -O -p $parm_file -c ${tag}_heat100K.crd -ref ${tag}_heat100K.crd -i ${tag}
    ↪ _heat303K-1.in -o ${tag}_heat303K-1.out -r ${tag}_heat303K-1.crd -x ${tag}_heat303K
    ↪ -1.nc >> ${tag}_heating.log
#cpptraj -p kcne1.prmtop -y heat303K-1.crd -x heat303K-1.pdb

#########################################################################
##### Heat to 303K 2 - 100ps (2,000,000 steps 0.05fs) #####
#########################################################################

# Restraint forces
set ProBB = 10.0
set ProSC = 5.0
set LH = 0.0
set LT = 1.0

cat << EOF > ${tag}_heat303K-2.in
Heating 303K constant pressure dynamics with protein and lipid restraints
&cntrl
imin=0,
ntx=1, ! Positions read formatted with no initial velocities
irest=0, ! No Restart
ntc=2,
```

```
ntf=2,
tol=0.0000001,
nstlim=2000000, ! Number of MD steps
ntt=3,
gamma_ln=10.0,
ntr=1,
ig=-1,
ntpr=2000,
ntwx=2000,
dt=0.00005,
nmropt=1, ! Read the weight change information below
ntb=2, ! Constant pressure periodic boundary conditions
ntp=1, ! Anisotropic pressure coupling
taup=2.0, ! Pressure relaxation time (ps)
cut=10.0,
/
&wt
type='TEMP0',
istep1=1,
istep2=200000,
value1=50.0,
value2=303.0,
&end
&wt
type='TEMP0',
istep1=200001,
istep2=2000000,
value1=303.0,
value2=303.0,
&end
&wt
type='END',
&end
Protein backbone restraints
$ProBB
FIND
* * M *
SEARCH
RES 1 $numAA
END
Protein non-main-chain restraints
```

```
$ProSC
FIND
* * S *
* * B *
* * 3 *
* * E *
SEARCH
RES 1 $numAA
END
Lipid head restraints
$LH
FIND
* * * PC
SEARCH
RES $lipidStart $numAALip
END
Lipid tail restraints
$LT
FIND
* * * PA
* * * OL
SEARCH
RES $lipidStart $numAALip
END
END ! End GROUP input
EOF

pmemd.cuda -O -p $parm_file -c ${tag}_heat303K-1.crd -ref ${tag}_heat303K-1.crd -i ${tag}
    ↪ _heat303K-2.in -o ${tag}_heat303K-2.out -r ${tag}_heat303K-2.crd -x ${tag}_heat303K
    ↪ -2.nc >> ${tag}_heating.log
#cpptraj -p kcne1.prmtop -y heat303K-2.crd -x heat303K-2.pdb

######################################################################
##### Heat to 303K 3 - 100ps (2000,000 steps 0.05fs) #####
######################################################################

# Restraint forces
set ProBB = 5.0
set ProSC = 2.5
set LH = 0.0
set LT = 0.5
```

```
cat << EOF > ${tag}_heat303K-3.in
Heating 303K constant pressure dynamics with protein and lipid restraints
&cntrl
imin=0,
ntx=1, ! Positions read formatted with no initial velocities
irest=0, ! No Restart
ntc=2,
ntf=2,
tol=0.0000001,
nstlim=2000000, ! Number of MD steps
ntt=3,
gamma_ln=10.0,
ntr=1,
ig=-1,
ntpr=2000,
ntwx=2000,
dt=0.00005, ! time step 0.05 fs
nmropt=1, ! Read the weight change information below
ntb=2, ! Constant pressure periodic boundary conditions
ntp=1, ! Anisotropic pressure coupling
taup=2.0, ! Pressure relaxation time (ps)
cut=10.0,
&end
&wt
type='TEMP0',
istep1=1,
istep2=400000,
value1=50.0,
value2=303.0,
&end
&wt
type='TEMP0',
istep1=400001,
istep2=2000000,
value1=303.0,
value2=303.0,
&end
&wt
type='END',
&end
```

```
Protein backbone restraints
$ProBB
FIND
* * M *
SEARCH
RES 1 $numAA
END
Protein non-main-chain restraints
$ProSC
FIND
* * S *
* * B *
* * 3 *
* * E *
SEARCH
RES 1 $numAA
END
Lipid head restraints
$LH
FIND
* * * PC
SEARCH
RES $lipidStart $numAALip
END
Lipid tail restraints
$LT
FIND
* * * PA
* * * OL
SEARCH
RES $lipidStart $numAALip
END
END ! End GROUP input
EOF

pmemd.cuda -O -p $parm_file -c ${tag}_heat303K-2.crd -ref ${tag}_heat303K-2.crd -i ${tag}
    ↪ _heat303K-3.in -o ${tag}_heat303K-3.out -r ${tag}_heat303K-3.crd -x ${tag}_heat303K
    ↪ -3.nc >> ${tag}_heating.log
#cpptraj -p kcne1.prmtop -y heat303K-3.crd -x heat303K-3.pdb

#####################################################################
```

```
##### Heat to 303K 4 - 100ps (1,000,000 steps 0.1fs) #####
#########################################################################

# Restraint forces
set ProBB = 2.5
set ProSC = 0.5
set LH = 0.0
set LT = 0.0

cat << EOF > ${tag}_heat303K-4.in
Heating 303K constant pressure dynamics with protein and lipid restraints
&cntrl
imin=0,
ntx=1, ! Positions read formatted with no initial velocities
irest=0, ! No Restart
ntc=2,
ntf=2,
tol=0.0000001,
nstlim=1000000, ! Number of MD steps
ntt=3,
gamma_ln=1.0,
ntr=1,
ig=-1,
ntpr=1000,
ntwx=1000,
dt=0.0001, ! Timestep 0.1 fs
nmropt=1, ! Read the weight change information below
ntb=2, ! Constant pressure periodic boundary conditions
ntp=1, ! Anisotropic pressure coupling
taup=2.0, ! Pressure relaxation time (ps)
cut=10.0,
&end
&wt
type='TEMP0',
istep1=1,
istep2=200000,
value1=50.0,
value2=303.0,
&end
&wt
type='TEMP0',
```

```
istep1=200001,
istep2=1000000,
value1=303.0,
value2=303.0,
&end
&wt
type='END',
&end
Protein backbone restraints
$ProBB
FIND
* * M *
SEARCH
RES 1 $numAA
END
Protein non-main-chain restraints
$ProSC
FIND
* * S *
* * B *
* * 3 *
* * E *
SEARCH
RES 1 $numAA
END
Lipid head restraints
$LH
FIND
* * * PC
SEARCH
RES $lipidStart $numAALip
END
Lipid tail restraints
$LT
FIND
* * * PA
* * * OL
SEARCH
RES $lipidStart $numAALip
END
END ! End GROUP input
```

```
EOF

pmemd.cuda -O -p $parm_file -c ${tag}_heat303K-3.crd -ref ${tag}_heat303K-3.crd -i ${tag}
    ↪ _heat303K-4.in -o ${tag}_heat303K-4.out -r ${tag}_heat303K-4.crd -x ${tag}_heat303K
    ↪ -4.nc >> ${tag}_heating.log
#cpptraj -p kcne1.prmtop -y heat303K-4.crd -x heat303K-4.pdb

######################################################################
##### Heat to 303K 5 - 25ps (50,000 steps 0.5fs) #####
######################################################################

# Restraint forces
set ProBB = 0.5
set ProSC = 0.0
set LH = 0.0
set LT = 0.0

cat << EOF > ${tag}_heat303K-5.in
Heating 303K constant pressure dynamics with protein and lipid restraints
&cntrl
imin=0,
ntx=1, ! Positions read formatted with no initial velocities
irest=0, ! No Restart
ntc=2,
ntf=2,
tol=0.0000001,
nstlim=50000, ! Number of MD steps
ntt=3,
gamma_ln=1.0,
ntr=1,
ig=-1,
ntpr=200,
ntwx=200,
dt=0.0005, ! Timestep 0.5 fs
nmropt=1, ! Read the weight change information below
ntb=2, ! Constant pressure periodic boundary conditions
ntp=1, ! Anisotropic pressure coupling
taup=2.0, ! Pressure relaxation time (ps)
cut=10.0,
&end
&wt
```

```
type='TEMP0',
istep1=1,
istep2=40000,
value1=50.0,
value2=303.0,
&end
&wt
type='TEMP0',
istep1=40001,
istep2=200000,
value1=303.0,
value2=303.0,
&end
&wt
type='END',
&end
Protein backbone restraints
$ProBB
FIND
* * M *
SEARCH
RES 1 $numAA
END
Protein non-main-chain restraints
$ProSC
FIND
* * S *
* * B *
* * 3 *
* * E *
SEARCH
RES 1 $numAA
END
Lipid head restraints
$LH
FIND
* * * PC
SEARCH
RES $lipidStart $numAALip
END
Lipid tail restraints
```

```
$LT
FIND
* * * PA
* * * OL
SEARCH
RES $lipidStart $numAALip
END
END ! End GROUP input
EOF

pmemd.cuda -O -p $parm_file -c ${tag}_heat303K-4.crd -ref ${tag}_heat303K-4.crd -i ${tag}
    ↪ _heat303K-5.in -o ${tag}_heat303K-5.out -r ${tag}_heat303K-5.crd -x ${tag}_heat303K
    ↪ -5.nc >> ${tag}_heating.log
#cpptraj -p kcne1.prmtop -y heat303K-5.crd -x heat303K-5.pdb
echo DONE >> ${tag}_heating.log
```

The system was then run with the protein-complex held fixed for another one ns at 303 K.

```
# Number of Amino Acids and number of Amino Acids + Lipids for this system
set numAA = 302
set numAALip = <Add the number residue + number of lipid +1>
@ lipidStart = $numAA + 1
@ wationStart = $numAALip + 1

# Working directory for input/output files
# set prefix = p2Y6_pge2g
# set simID = 1
set tag = $3
#set basedir = /dors/meilerlab/home/vuot2/Documents/workspace/MD/GPCR_MD/pilot_test/
    ↪ bilayer/P2Y6_pilot/POPC_CHL1
#set dirname = $basedir/${prefix}_"$simID"
set dirname = $4
cd $dirname
set parm_file = $dirname/$tag.parm7
set start_crd = $dirname/${tag}_heat303K-5.crd

echo holding for ${tag}
cat << EOF > ${tag}_hold.in
 Lipid production 303K 1ns
 &cntrl
  imin=0,  ! No minimization
  ntx=5,  ! our inpcrd file has velocities
```

```
 irest=1,   ! This IS a restart of an old MD simulation

 ! SHAKE
 ntc=2,  ! Constrain bonds containing hydrogen
 ntf=2,  ! Do not calculate forces of bonds containing hydrogen
 ntr=1,  ! restraining specified atoms in Cartesian space using a harmonic potential
 tol=0.0000001, ! Relative geometrical tolerance for coordinate resetting in shake
 nstlim=1000000, ! max #steps
 ntt=3,  ! Langevin dynamics
 gamma_ln=1.0,  ! collision frequency for Langevin dynamics
 ig=-1,  ! seed for the pseudo-random number generator
 temp0=303.0,  ! temp0
 ntpr=5000,
 ntwr=5000,
 ntwx=5000,
 ioutfm=1,  ! Write NetCDF format
 ntxo=2,
 dt=0.001,  ! timestep 1fs

 ! Constant pressure control.
 barostat=2,  ! MC barostat... change to 1 for Berendsen
 ntp=3,  ! 1=isotropic, 2=anisotropic, 3=semi-isotropic w/ surften
 pres0=1.0,  ! Target external pressure, in bar

 ! Constant surface tension (needed for semi-isotropic scaling). Uncomment
 ! for this feature. csurften must be nonzero if ntp=3 above
 csurften=3, ! Interfaces in 1=yz plane, 2=xz plane, 3=xy plane
 gamma_ten=0.0, ! Surface tension (dyne/cm). 0 gives pure semi-iso scaling
 ninterface=2, ! Number of interfaces (2 for bilayer)

 ! Wrap coordinates when printing them to the same unit cell
 iwrap=0,
 cut=10.0,
/
/
&ewald
 skinnb=5, ! Increase skinnb to avoid skinnb errors
/
Hold protein fixed
1.0
RES 1 $numAA
```

```
END
END
EOF

pmemd.cuda -O -i ${tag}_hold.in -o ${tag}_hold.1.out -p $parm_file -c $start_crd -r ${tag}
    ↪ _hold.1.crd -ref $start_crd -x ${tag}_hold.1.nc > ${tag}_hold.log
#cpptraj -p $parm_file -y ${tag}_hold.1.crd -x ${tag}_hold.1.pdb

pmemd.cuda -O -i ${tag}_hold.in -o ${tag}_hold.2.out -p $parm_file -c ${tag}_hold.1.crd -r
    ↪  ${tag}_hold.2.crd -ref ${tag}_hold.1.crd -x ${tag}_hold.2.nc >> ${tag}_hold.log
#cpptraj -p $parm_file -y ${tag}_hold.2.crd -x ${tag}_hold.2.pdb

pmemd.cuda -O -i ${tag}_hold.in -o ${tag}_hold.3.out -p $parm_file -c ${tag}_hold.2.crd -r
    ↪  ${tag}_hold.3.crd -ref ${tag}_hold.2.crd -x ${tag}_hold.3.nc >> ${tag}_hold.log
#cpptraj -p $parm_file -y ${tag}_hold.3.crd -x ${tag}_hold.3.pdb

pmemd.cuda -O -i ${tag}_hold.in -o ${tag}_hold.4.out -p $parm_file -c ${tag}_hold.3.crd -r
    ↪  ${tag}_hold.4.crd -ref ${tag}_hold.3.crd -x ${tag}_hold.4.nc >> ${tag}_hold.log
#cpptraj -p $parm_file -y ${tag}_hold.4.crd -x ${tag}_hold.4.pdb

pmemd.cuda -O -i ${tag}_hold.in -o ${tag}_hold.5.out -p $parm_file -c ${tag}_hold.4.crd -r
    ↪  ${tag}_hold.5.crd -ref ${tag}_hold.4.crd -x ${tag}_hold.5.nc >> ${tag}_hold.log
#cpptraj -p $parm_file -y ${tag}_hold.5.crd -x ${tag}_hold.5.pdb
echo DONE >> ${tag}_hold.log
```

Production MD was conducted for 500 ns at 303K using a step size of 4 fs with hydrogen mass repartitioning(Hopkins et al., 2015), constant pressure periodic boundary conditions (NPT system), semi-anisotropic pressure scaling, and Langevin dynamics.

```
#!/bin/tcsh
# Number of Amino Acids and number of Amino Acids + Lipids for this system
set numAA = 302
set numAALip = <Add the number residue + number of lipid +1>
@ lipidStart = $numAA + 1
@ wationStart = $numAALip + 1

# Working directory for input/output files
set tag = p2Y6_pge2g
set dirname = <curent working directory>
set parm_file = $dirname/${tag}.parm7
set start_crd = $dirname/${tag}_hold.5
```

```
cd $dirname
echo production for ${tag}

# Variable for start and end
# set cnt = 0
# set cntmax = 19
set cnt = $5
@ cntmax = $cnt + 9

echo production for ${tag}
# prod. for 100 ns
while ( ${cnt} <= ${cntmax} )
   @ pcnt = ${cnt} - 1
   set prev = `echo $pcnt | awk '{printf( "%04d", $0)}'`
   set run = `echo $cnt | awk '{printf( "%04d", $0)}'`

   #Use different random seeds for Langevin dynamics
cat << EOF > $dirname/${tag}_prod.${run}.in
Production, 303K 10ns
&cntrl
imin=0, ! Molecular dynamics
ntx=5, ! Positions and velocities read NetCDF (or formatted)
irest=1, ! Restart calculation
ntc=2, ! SHAKE on for bonds with hydrogen
ntf=2, ! No force evaluation for bonds with hydrogen
tol=0.0000001, ! SHAKE tolerance
nstlim=2500000, ! Number of MD steps = 10 ns
ntt=3, ! Langevin dynamics
gamma_ln=1.0, ! Collision frequency for Langevin dyn.
temp0=303.0, ! Simulation temperature (K)
ntpr=25000, ! Print to mdout every ntpr steps
ntwr=25000, ! Write a restart file every ntwr steps
ntwx=25000, ! Write to trajectory file every ntwx steps
ioutfm=1, ! Write binary NetCDF trajectory
ntxo=2, ! Write binary restart file
iwrap=1, ! Wrap coordinates when printing them to the same unit cell
dt=0.004, ! Timestep (ps)
ig=-1, ! Random seed for Langevin dynamics
ntb=2, ! Constant pressure periodic boundary conditions

! Constant pressure control.
```

```
barostat=2,  ! MC barostat... change to 1 for Berendsen
ntp=3,  ! 1=isotropic, 2=anisotropic, 3=semi-isotropic w/ surften
pres0=1.0,  ! Target external pressure, in bar

! Constant surface tension (needed for semi-isotropic scaling). Uncomment
! for this feature. csurften must be nonzero if ntp=3 above
csurften=3,  ! Interfaces in 1=yz plane, 2=xz plane, 3=xy plane
gamma_ten=0.0,  ! Surface tension (dyne/cm). 0 gives pure semi-iso scaling
ninterface=2,  ! Number of interfaces (2 for bilayer)
cut=10.0, ! Nonbonded cutoff (Angstroms)
/
EOF

   if ( ${cnt} == 0 ) then
      set pstep = $start_crd
      set istep = ${tag}_prod.${run}
      pmemd.cuda -O -i $dirname/${istep}.in -p $parm_file -c ${pstep}.crd -o ${istep}.
          ↪ mdout -r ${istep}.crd -inf ${istep}.mdinfo -ref ${pstep}.crd -x ${istep}.nc >
          ↪  ${tag}_prod.log
   else
      set pstep = ${tag}_prod.${prev}
      set istep = ${tag}_prod.${run}
      pmemd.cuda -O -i $dirname/${istep}.in -p $parm_file -c ${pstep}.crd -o ${istep}.
          ↪ mdout -r ${istep}.crd -inf ${istep}.mdinfo -ref ${pstep}.crd -x ${istep}.nc
          ↪ >> ${tag}_prod.log
      cpptraj -p $parm_file -y ${istep}.crd -x ${istep}.pdb
   endif
   echo DONE >> ${tag}_prod.log
   @ cnt += 1
end
```

7.4.2 Analysis of the MD trajectories

the MD trajectories were analyzed using CPPTRAJ (version 18.0) and PTRAJ (version 2.0.2.dev0)(Roe and Cheatham, 2013), as well as VMD (visual molecular dynamics; version 1.9)(Humphrey et al., 1996).

This script stripped water and membrane and ions residues from the MD trajectory, and perfomed autoimage and superimposition of all frames on to the first frame.

```
#!/bin/bash
parm_f=p2y6-pge2g_rep1_prod.parm7
prefix=p2y6-pge2g_rep1_prod
```

```
w_dir=<MD output directory>
atom_end=$4 #id of the last atom in the nc files

cd $w_dir
cat << EOF > ${w_dir}/${prefix}_post_MD_prosessing.in
parm $parm_f
parmstrip @$((atom_end+1))-9999999 outprefix debug
run
trajin ${prefix}.00*.nc
autoimage
rmsd all @CA first
run
trajout ${prefix}_out.nc
parmwrite out debug.${prefix}.parm7
run
exit
EOF
cpptraj -i ${w_dir}/${prefix}_post_MD_prosessing.in
```

The first 100 ns of the simulation was removed before we performed the calculation of RMSF, atomic contact, and Molecular Mechanics with a Poisson-Boltzmann/Surface Area solvent (MM-PBSA). Relative contract strength is calculated as the sum of atom pair contact frequency between the agonist and each P2Y6 residue.

```
#!/bin/bash
# This is the script to analyze the trajectory of
PREFIX=p2y6-pge2g_rep1_prod
W_DIR= <output MD directory>

cd $W_DIR
# Read in the variables to perform data cleaning
cpptraj <<EOF
set prefix=$PREFIX
parm $prefix.parm7
trajin $prefix_prod_out.nc 1 last
reference $prefix_prod_out.nc 1 [hl]
rmsd bb :1-301@CA ref [hl] out p2y6_rmsd.dat
rmsd udp :302&!@H nofit ref [hl] out ligand_rmsd.dat
run
clear trajin
trajin $prefix_prod_out.nc 1000 last
```

```
average crdset myave :1-301&@CA
rms ref myave :9-289&@CA
run
hbond contracts :1-302 avgout $prefix_avg.dat series uuseries $prefix_hbond.gnu nointramol
nativecontacts name ligand :1-301&!@H= :302&!@H= savenonnative byresidue map mapout
    ↪ $prefix_contacts.txt
atomicfluct out $prefix_fluct_res.dat :1-301&@CA byres
atomicfluct out $prefix_fluct_ligand.dat :302
run
EOF
```

Trajectories of the last 400 ns of each of three MD replicates were combined and then clustered into 6 clusters, and the representative frame, which is also the centroid, of the largest cluster were chosen as final refined docking models.

```
#!/bin/bash
# This is the script to analyze the trajectory of
PREFIX=p2y6-pge2g_rep1_prod
CLUSTER_NUM=6
W_DIR=< MD output directory>

cd $W_DIR
# Read in the variables to perform data cleaning
cat << EOF > ${PREFIX}_${CLUSTER_NUM}_traj_cluster_analysis.in
parm ${PREFIX}_rep1/strip.mem.${PREFIX}_rep1_prob.parm7
trajin ${PREFIX}_rep1/${PREFIX}_rep1_prob_out.nc 1 last 1
trajin ${PREFIX}_rep1/${PREFIX}_rep1_prob_out.nc 1 last 1
trajin ${PREFIX}_rep1/${PREFIX}_rep1_prob_out.nc 1 last 1
cluster c1 \
kmeans clusters ${CLUSTER_NUM} randompoint maxit 500 \
rms :302&!@H= nofit \
sieve 10 random \
out ${PREFIX}_${CLUSTER_NUM}_time.dat \
summary ${PREFIX}_${CLUSTER_NUM}_summary.dat \
info ${PREFIX}_${CLUSTER_NUM}_info.dat \
cpopvtime ${PREFIX}_${CLUSTER_NUM}_cpopvtime.agr normframe \
repout rep repfmt pdb \
singlerepout ${PREFIX}_${CLUSTER_NUM}_singlerep.nc singlerepfmt netcdf \
avgout avg avgfmt pdb
run
EOF
```

```
cpptraj -i ${PREFIX}_${CLUSTER_NUM}_traj_cluster_analysis.in
```

References

Adrian, T. E., Greenberg, G. R., Besterman, H. S., and Bloom, S. R. (1978). Pharmacokinetics of pancreatic polypeptide in man. *Gut*, 19(10):907–9.

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using polyphen-2. *Current Protocols in Human Genetics*, 76(1):7.20.1–7.20.41.

Alexander, N., Woetzel, N., and Meiler, J. (2011). bcl::cluster : A method for clustering biological molecules coupled with visualization in the pymol molecular graphics system. *IEEE Int Conf Comput Adv Bio Med Sci*, 2011:13–18.

Alhouayek, M. and Muccioli, G. G. (2014). Cox-2-derived endocannabinoid metabolites as novel inflammatory mediators. *Trends Pharmacol Sci*, 35(6):284–92.

Andersen, N. H., Chen, C. P., Marschner, T. M., Krystek, S. R., J., and Bassolino, D. A. (1992). Conformational isomerism of endothelin in acidic aqueous media: a quantitative noesy analysis. *Biochemistry*, 31(5):1280–95.

Andersen, N. H. and Hammen, P. K. (1991). A conformation-preference/potency correlation for gnrh analogs: Nmr evidence. *Bioorganic & Medicinal Chemistry Letters*, 1(5):263–266.

Andersson, A. and Maler, L. (2002). Nmr solution structure and dynamics of motilin in isotropic phospholipid bicellar solution. *J Biomol NMR*, 24(2):103–12.

Asada, H., Horita, S., Hirata, K., Shiroishi, M., Shiimura, Y., Iwanari, H., Hamakubo, T., Shimamura, T., Nomura, N., Kusano-Arai, O., Uemura, T., Suno, C., Kobayashi, T., and Iwata, S. (2018). Crystal structure of the human angiotensin ii type 2 receptor bound to an angiotensin ii analog. *Nat Struct Mol Biol*, 25(7):570–576.

Atkins, A. R., Martin, R. C., and Smith, R. (1995). 1h nmr studies of sarafotoxin srtb, a nonselective endothelin receptor agonist, and irl 1620, an etb receptor-specific agonist. *Biochemistry*, 34(6):2026–33.

Aumelas, A., Chiche, L., Kubo, S., Chino, N., Tamaoki, H., and Kobayashi, Y. (1995). [lys(-2)-arg(-1)]endothelin-1 solution structure by two-dimensional 1h-nmr: possible involvement of electrostatic interactions in native disulfide bridge formation and in biological activity decrease. *Biochemistry*, 34(14):4546–61.

Bader, R., Bettio, A., Beck-Sickinger, A. G., and Zerbe, O. (2001). Structure and dynamics of micelle-bound neuropeptide y: comparison with unligated npy and implications for receptor selection. *J Mol Biol*, 305(2):307–29.

Bajic, G., Yatime, L., Klos, A., and Andersen, G. R. (2013). Human c3a and c3a desarg anaphylatoxins have conserved structures, in contrast to c5a and c5a desarg. *Protein Sci*, 22(2):204–12.

Balasubramaniam, A., Mullins, D. E., Lin, S., Zhai, W., Tao, Z., Dhawan, V. C., Guzzi, M., Knittel, J. J., Slack, K., Herzog, H., and Parker, E. M. (2006). Neuropeptide y (npy) y4 receptor selective agonists based on npy(3236): development of an anorectic y4 receptor selective agonist with picomolar affinity. *Journal of Medicinal Chemistry*, 49(8):2661–2665.

Ballesteros, J. A. and Weinstein, H. (1995). Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in g protein-coupled receptors. *Methods in Neurosciences*, 25:366–428.

Bar, I., Guns, P. J., Metallo, J., Cammarata, D., Wilkin, F., Boeynams, J. M., Bult, H., and Robaye, B. (2008). Knockout mice reveal a role for p2y6 receptor in macrophages, endothelial cells, and vascular smooth muscle cells. *Mol Pharmacol*, 74(3):777–84.

Baskin, I., Ait, A. O., Halberstam, N. M., Palyulin, V. A., and Zefirov, N. S. (2002). An approach to the interpretation of backpropagation neural network models in qsar studies. *SAR QSAR Environ Res*, 13(1):35–41.

Batterham, R. L., Le Roux, C. W., Cohen, M. A., Park, A. J., Ellis, S. M., Patterson, M., Frost, G. S., Ghatei, M. A., and Bloom, S. R. (2003). Pancreatic polypeptide reduces appetite and food intake in humans. *J Clin Endocrinol Metab*, 88(8):3989–92.

Bdioui, S., Verdi, J., Pierre, N., Trinquet, E., Roux, T., and Kenakin, T. (2018). Equilibrium assays are required to accurately characterize the activity profiles of drugs modulating gq-protein-coupled receptors. *Mol Pharmacol*, 94(3):992–1006.

Beck, A., Bussat, M. C., Klinguer-Hamour, C., Goetsch, L., Aubry, J. P., Champion, T., Julien, E., Haeuw, J. F., Bonnefoy, J. Y., and Corvaia, N. (2001). Stability and ctl activity of n-terminal glutamic acid containing peptides. *J Pept Res*, 57(6):528–38.

Beck-Sickinger, A. G. and Jung, G. (1995). Structure-activity relationships of neuropeptide y analogues with respect to y1 and y2 receptors. *Biopolymers*, 37(2):123–42.

Bednarek, M. A., Feighner, S. D., Pong, S. S., McKee, K. K., Hreniuk, D. L., Silva, M. V., Warren, V. A., Howard, A. D., Van Der Ploeg, L. H., and Heck, J. V. (2000). Structure-function studies on the new growth hormone-releasing peptide, ghrelin: minimal sequence of ghrelin necessary for activation of growth hormone secretagogue receptor 1a. *J Med Chem*, 43(23):4370–6.

Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004). Similarity searching of chemical databases using atom environment descriptors (molprint 2d): evaluation of performance. *J Chem Inf Comput Sci*, 44(5):1708–18.

Bender, B. J., Cisneros, A., r., Duran, A. M., Finn, J. A., Fu, D., Lokits, A. D., Mueller, B. K., Sangha, A. K., Sauer, M. F., Sevy, A. M., Sliwoski, G., Sheehan, J. H., DiMaio, F., Meiler, J., and Moretti, R. (2016). Protocols for molecular modeling with rosetta3 and rosettascripts. *Biochemistry*, 55(34):4748–63.

Bender, B. J., Marlow, B., and Meiler, J. (2020). Improving homology modeling from low-sequence identity templates in rosetta: A case study in gpcrs. *PLOS Computational Biology*, 16(10):e1007597.

Bender, B. J., Vortmeier, G., Ernicke, S., Bosse, M., Kaiser, A., Els-Heindl, S., Krug, U., Beck-Sickinger, A., Meiler, J., and Huster, D. (2019). Structural model of ghrelin bound to its g protein-coupled receptor. *Structure*, 27(3):537–544.e4.

Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P., and Wu, C. W. (1981). X-ray analysis (1. 4-a resolution) of avian pancreatic polypeptide: Small globular protein hormone. *Proc Natl Acad Sci U S A*, 78(7):4175–9.

Bockaert, J. and Pin, J. P. (1999). Molecular tinkering of g protein-coupled receptors: an evolutionary success. *EMBO J*, 18(7):1723–9.

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem*, 72:248–54.

Brogden, R. N., Buckley, M. M. T., and Ward, A. (1990). Buserelin. *Drugs*, 39(3):399–437.

Broto, P., Moreau, G., , and Vandycke, C. (1984). Molecular structures: Perception, autocorrelation descriptor and sar studies. autocorrelation descriptor. *European Journal of Medicinal Chemistry*, 19(1):66–70.

Bruser, A., Zimmermann, A., Crews, B. C., Sliwoski, G., Meiler, J., Konig, G. M., Kostenis, E., Lede, V., Marnett, L. J., and Schoneberg, T. (2017). Prostaglandin e2 glyceryl ester is an endogenous agonist of the nucleotide receptor p2y6. *Sci Rep*, 7(1):2380.

Bryksin, A. V. and Matsumura, I. (2010). Overlap extension pcr cloning: a simple and reliable way to create recombinant plasmids. *Biotechniques*, 48(6):463–5.

Burford, N. T., Watson, J., Bertekap, R., and Alt, A. (2011). Strategies for the identification of allosteric modulators of g-protein-coupled receptors. *Biochemical pharmacology*, 81(6):691–702.

Burg, J. S., Ingram, J. R., Venkatakrishnan, A. J., Jude, K. M., Dukkipati, A., Feinberg, E. N., Angelini, A., Waghray, D., Dror, R. O., Ploegh, H. L., and Garcia, K. C. (2015). Structural biology. structural basis for chemokine recognition and activation of a viral g protein-coupled receptor. *Science*, 347(6226):1113–7.

Butkiewicz, M., Lowe, E. W., J., Mueller, R., Mendenhall, J. L., Teixeira, P. L., Weaver, C. D., and Meiler, J. (2013). Benchmarking ligand-based virtual high-throughput screening with the pubchem database. *Molecules*, 18(1):735–56.

Butkiewicz, M., Wang, Y., Bryant, S. H., Lowe, E. W., C, D. W., and Meiler, J. (2017). High-throughput screening assay datasets from the pubchem database. *Chemical Informatics*, 3(1).

Böhme, I., Stichel, J., Walther, C., Mörl, K., and Beck-Sickinger, A. G. (2008). Agonist induced receptor internalization of neuropeptide y receptor subtypes depends on third intracellular loop and c-terminus. *Cell Signal*, 20(10):1740–9.

Cabrele, C. and Beck-Sickinger, A. G. (2000). Molecular characterization of the ligand-receptor interaction of the neuropeptide y family. *J Pept Sci*, 6(3):97–122.

Cao, C., Tan, Q., Xu, C., He, L., Yang, L., Zhou, Y., Zhou, Y., Qiao, A., Lu, M., Yi, C., Han, G. W., Wang, X., Li, X., Yang, H., Rao, Z., Jiang, H., Zhao, Y., Liu, J., Stevens, R. C., Zhao, Q., Zhang, X. C., and Wu, B. (2018). Structural basis for signal recognition and transduction by platelet-activating-factor receptor. *Nature Structural and Molecular Biology*, 25(6):488–495.

Carlsson, L., Helgee, E. A., and Boyer, S. (2009). Interpretation of nonlinear qsar models applied to ames mutagenicity data. *Journal of Chemical Information and Modeling*, 49(11):2551–2558.

Carraway, R. and Leeman, S. E. (1975). The amino acid sequence of a hypothalamic peptide, neurotensin. *J Biol Chem*, 250(5):1907–11.

Chance, R. E. (2003). *Pancreatic Polypeptide*, pages 142–146. Academic Press, New York.

Chang, J. K., Williams, R. H., Humphries, A. J., Johansson, N. G., Folkers, K., and Bowers, C. Y. (1972). Luteinizing releasing hormone, synthesis and arg 8 -analogs, and conformation-sequence-activity relationships. *Biochem Biophys Res Commun*, 47(4):727–32.

Chary, K. V., Srivastava, S., Hosur, R. V., Roy, K. B., and Govil, G. (1986). Molecular conformation of gonadoliberin using two-dimensional nmr spectroscopy. *Eur J Biochem*, 158(2):323–32.

Che, T., Majumdar, S., Zaidi, S. A., Ondachi, P., McCorvy, J. D., Wang, S., Mosier, P. D., Uprety, R., Vardy, E., Krumm, B. E., Han, G. W., Lee, M.-Y., Pardon, E., Steyaert, J., Huang, X.-P., Strachan, R. T., Tribo, A. R., Pasternak, G. W., Carroll, F. I., Stevens, R. C., Cherezov, V., Katritch, V., Wacker, D., and Roth, B. L. (2018). Structure of the nanobody-stabilized active state of the kappa opioid receptor. *Cell*, 172(1):55–67.e15.

Cheng, R. K. Y., Fiez-Vandal, C., Schlenker, O., Edman, K., Aggeler, B., Brown, D. G., Brown, G. A., Cooke, R. M., Dumelin, C. E., Dore, A. S., Geschwindner, S., Grebner, C., Hermansson, N. O., Jazayeri, A., Johansson, P., Leong, L., Prihandoko, R., Rappas, M., Soutter, H., Snijder, A., Sundstrom, L., Tehan, B., Thornton, P., Troast, D., Wiggin, G., Zhukov, A., Marshall, F. H., and Dekker, N. (2017). Structural insight into allosteric modulation of protease-activated receptor 2. *Nature*, 545(7652):112–115.

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Kuhn, P., Weis, W. I., Kobilka, B. K., and Stevens, R. C. (2007). High-resolution crystal structure of an engineered human beta2-adrenergic g protein-coupled receptor. *Science*, 318(5854):1258–65.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., and Tropsha, A. (2014). Qsar modeling: Where have you been? where are you going to? *Journal of Medicinal Chemistry*, 57(12):4977–5010.

Chien, E. Y., Liu, W., Zhao, Q., Katritch, V., Han, G. W., Hanson, M. A., Shi, L., Newman, A. H., Javitch, J. A., Cherezov, V., and Stevens, R. C. (2010). Structure of the human dopamine d3 receptor in complex with a d2/d3 selective antagonist. *Science*, 330(6007):1091–5.

Combs, S. A., Deluca, S. L., Deluca, S. H., Lemmon, G. H., Nannemann, D. P., Nguyen, E. D., Willis, J. R., Sheehan, J. H., and Meiler, J. (2013). Small-molecule ligand docking into comparative models with rosetta. *Nat Protoc*, 8(7):1277–98.

Communi, D., Janssens, R., Suarez-Huerta, N., Robaye, B., and Boeynaems, J. M. (2000). Advances in signalling by extracellular nucleotides. the role and transduction mechanisms of p2y receptors. *Cell Signal*, 12(6):351–60.

Conn, P. J., Lindsley, C. W., Meiler, J., and Niswender, C. M. (2014). Opportunities and challenges in the discovery of allosteric modulators of gpcrs for treating cns disorders. *Nature Reviews Drug Discovery*, 13(9):692–708.

Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E., and Baker, D. (2014). Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein science : a publication of the Protein Society*, 23(1):47–55.

Cook, W. J., Galakatos, N., Boyar, W. C., Walter, R. L., and Ealick, S. E. (2010). Structure of human desarg-c5a. *Acta Crystallogr D Biol Crystallogr*, 66(Pt 2):190–7.

Costanzi, S., Joshi, B. V., Maddileti, S., Mamedova, L., Gonzalez-Moa, M. J., Marquez, V. E., Harden, T. K., and Jacobson, K. A. (2005). Human p2y(6) receptor: molecular modeling leads to the rational design of a novel agonist based on a unique conformational preference. *J Med Chem*, 48(26):8108–11.

Coster, M., Wittkopf, D., Kreuchwig, A., Kleinau, G., Thor, D., Krause, G., and Schoneberg, T. (2012). Using ortholog sequence data to predict the functional relevance of mutations in g-protein-coupled receptors. *FASEB J*, 26(8):3273–81.

Craik, D. J., Fairlie, D. P., Liras, S., and Price, D. (2013). The future of peptide-based drugs. *Chem Biol Drug Des*, 81(1):136–47.

Cramer, R. D. (2012). The inevitable qsar renaissance. *Journal of Computer-Aided Molecular Design*, 26(1):35–38.

Da Costa, G., Bondon, A., Coutant, J., Curmi, P., and Monti, J. P. (2013). Intermolecular interactions between the neurotensin and the third extracellular loop of human neurotensin 1 receptor. *J Biomol Struct Dyn*, 31(12):1381–92.

De Ricco, R., Valensin, D., Gaggelli, E., and Valensin, G. (2013). Conformation propensities of des-acyl-ghrelin as probed by cd and nmr. *Peptides*, 43:62–7.

Deflorian, F. and Jacobson, K. A. (2011). Comparison of three gpcr structural templates for modeling of the p2y12 nucleotide receptor. *Journal of computer-aided molecular design*, 25(4):329–338.

Deluca, S. H., Rathmann, D., Beck-Sickinger, A. G., and Meiler, J. (2013). The activity of prolactin releasing peptide correlates with its helicity. *Biopolymers*, 99(5):314–25.

Denys, L., Bothner-By, A. A., Fisher, G. H., and Ryan, J. W. (1982). Conformational diversity of bradykinin in aqueous solution. *Biochemistry*, 21(25):6531–6.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7.

Fan, X., Zhou, N., Zhang, X., Mukhtar, M., Lu, Z., Fang, J., DuBois, G. C., and Pomerantz, R. J. (2003). Structural and functional study of the apelin-13 peptide, an endogenous ligand of the hiv-1 coreceptor, apj. *Biochemistry*, 42(34):10163–8.

Ferré, G., Louet, M., Saurel, O., Delort, B., Czaplicki, G., M'Kadmi, C., Damian, M., Renault, P., Cantel, S., Gavara, L., Demange, P., Marie, J., Fehrentz, J.-A., Floquet, N., Milon, A., and Banères, J.-L. (2019). Structure and dynamics of g protein-coupled receptor–bound ghrelin reveal the critical role of the octanoyl chain. *Proceedings of the National Academy of Sciences*, 116(35):17525–17530.

Fredriksson, R., Lagerström, M. C., Lundin, L. G., and Schiöth, H. B. (2003). The g-protein-coupled receptors in the human genome form five main families. phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol*, 63(6):1256–72.

Fredslund, F., Laursen, N. S., Roversi, P., Jenner, L., Oliveira, C. L., Pedersen, J. S., Nunn, M. A., Lea, S. M., Discipio, R., Sottrup-Jensen, L., and Andersen, G. R. (2008). Structure of and influence of a tick complement inhibitor on human complement component 5. *Nat Immunol*, 9(7):753–60.

Gao, Z. G. and Jacobson, K. A. (2013). Allosteric modulation and functional selectivity of g protein-coupled receptors. *Drug Discov Today Technol*, 10(2):e237–43.

Garcia, R. A., Yan, M., Search, D., Zhang, R., Carson, N. L., Ryan, C. S., Smith-Monroy, C., Zheng, J., Chen, J., Kong, Y., Tang, H., Hellings, S. E., Wardwell-Swanson, J., Dinchuk, J. E., Psaltis, G. C., Gordon, D. A., Glunz, P. W., and Gargalovic, P. S. (2014). P2y6 receptor potentiates pro-inflammatory responses in macrophages and exhibits differential roles in atherosclerotic lesion development. *PLoS One*, 9(10):e111385.

Gasteiger, J. and Marsili, M. (1980). Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36(22):3219–3228.

Gasteiger, J., Teckentrup, A., Terfloth, L., and Spycher, S. (2003). Neural networks as data mining tools in drug design. *Journal of Physical Organic Chemistry*, 16(4):232–245.

Gates, D. M., Succop, P., Brehm, B. J., Gillespie, G. L., and Sommers, B. D. (2008). Obesity and presenteeism: The impact of body mass index on workplace productivity. *Journal of Occupational and Environmental Medicine*, 50(1):39–45.

Gerald, C., Walker, M. W., Criscione, L., Gustafson, E. L., Batzl-Hartmann, C., Smith, K. E., Vaysse, P., Durkin, M. M., Laz, T. M., Linemeyer, D. L., Schaffhauser, A. O., Whitebread, S., Hofbauer, K. G., Taber, R. I., Branchek, T. A., and Weinshank, R. L. (1996). A receptor subtype involved in neuropeptide-y-induced food intake. *Nature*, 382(6587):168–71.

Giannattasio, G., Ohta, S., Boyce, J. R., Xing, W., Balestrieri, B., and Boyce, J. A. (2011). The purinergic g protein-coupled receptor 6 inhibits effector t cell activation in allergic pulmonary inflammation. *J Immunol*, 187(3):1486–95.

Gould, I., A.A., S., Dickson, C., Madej, B., and Walker, R. (2018). Lipid17: A comprehensive amber force field for the simulation of zwitterionic and anionic lipids [in prep].

Grace, C. R., Durrer, L., Koerber, S. C., Erchegyi, J., Reubi, J. C., Rivier, J. E., and Riek, R. (2005). Somatostatin receptor 1 selective analogues: 4. three-dimensional consensus structure by nmr. *J Med Chem*, 48(2):523–33.

Grace, C. R., Erchegyi, J., Koerber, S. C., Reubi, J. C., Rivier, J., and Riek, R. (2006). Novel sst2-selective somatostatin agonists. three-dimensional consensus structure by nmr. *J Med Chem*, 49(15):4487–96.

Grace, C. R., Erchegyi, J., Reubi, J. C., Rivier, J. E., and Riek, R. (2008). Three-dimensional consensus structure of sst2-selective somatostatin (srif) antagonists by nmr. *Biopolymers*, 89(12):1077–87.

Grace, C. R., Koerber, S. C., Erchegyi, J., Reubi, J. C., Rivier, J., and Riek, R. (2003). Novel sst(4)-selective somatostatin (srif) agonists. 4. three-dimensional consensus structure by nmr. *J Med Chem*, 46(26):5606–18.

Granier, S., Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Weis, W. I., and Kobilka, B. K. (2012). Structure of the delta-opioid receptor bound to naltrindole. *Nature*, 485(7398):400–4.

Grant, G. and Vale, W. (1972). Speculations on structural relationships between the hypothalamic releasing factors of pituitary hormones. *Nat New Biol*, 237(75):182–3.

Gregory, K. J., Sexton, P. M., and Christopoulos, A. (2010). Overview of receptor allosterism. *Curr Protoc Pharmacol*, Chapter 1:Unit 1.21.

Guha, R. and Jurs, P. C. (2005). Interpreting computational neural network qsar models: a measure of descriptor importance. *J Chem Inf Model*, 45(3):800–6.

Haugaard-Kedstrom, L. M., Hossain, M. A., Daly, N. L., Bathgate, R. A., Rinderknecht, E., Wade, J. D., Craik, D. J., and Rosengren, K. J. (2015). Solution structure, aggregation behavior, and flexibility of human relaxin-2. *ACS Chem Biol*, 10(3):891–900.

Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schioth, H. B., and Gloriam, D. E. (2017). Trends in gpcr drug discovery: new agents, targets and indications. *Nat Rev Drug Discov*, 16(12):829–842.

Hauser, A. S., Chavali, S., Masuho, I., Jahn, L. J., Martemyanov, K. A., Gloriam, D. E., and Babu, M. M. (2018). Pharmacogenomics of gpcr drug targets. *Cell*, 172(1):41–54.e19.

He, X., Man, V. H., Yang, W., Lee, T.-S., and Wang, J. (2020). A fast and high-quality charge model for the next generation general amber force field. *The Journal of Chemical Physics*, 153(11):114502.

Heise, H., Luca, S., de Groot, B. L., Grubmuller, H., and Baldus, M. (2005). Probing conformational disorder in neurotensin by two-dimensional solid-state nmr and comparison to molecular dynamics simulations. *Biophys J*, 89(3):2113–20.

Henry, J. A., Horwell, D. C., Meecham, K. G., and Rees, D. C. (1993). A structure-affinity study of the amino acid side-chains in neurotensin : N and c terminal deletions and ala-scan. *Bioorganic & Medicinal Chemistry Letters*, 3(5):949–952.

Hewage, C. M., Jiang, L., Parkinson, J. A., Ramage, R., and Sadler, I. H. (1999). Solution structure of a novel etb receptor selective agonist et1-21 [cys(acm)1,15, aib3,11, leu7] by nuclear magnetic resonance spectroscopy and molecular modelling. *J Pept Res*, 53(3):223–33.

Hilal-Dandan, R., Villegas, S., Gonzalez, A., and Brunton, L. L. (1997). The quasi-irreversible nature of endothelin binding and g protein-linked signaling in cardiac myocytes. *J Pharmacol Exp Ther*, 281(1):267–73.

Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K., and Pease, L. R. (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene*, 77(1):51–9.

Hofmann, S., Frank, R., Hey-Hawkins, E., Beck-Sickinger, A. G., and Schmidt, P. (2013). Manipulating y receptor subtype activation of short neuropeptide y analogs by introducing carbaboranes. *Neuropeptides*, 47(2):59–66.

Hopkins, C. W., Le Grand, S., Walker, R. C., and Roitberg, A. E. (2015). Long-time-step molecular dynamics through hydrogen mass repartitioning. *Journal of Chemical Theory and Computation*, 11(4):1864–1874.

Hu, S. S., Bradshaw, H. B., Chen, J. S., Tan, B., and Walker, J. M. (2008). Prostaglandin e2 glycerol ester, an endogenous cox-2 metabolite of 2-arachidonoylglycerol, induces hyperalgesia and modulates nfkappab activity. *Br J Pharmacol*, 153(7):1538–49.

Huber, R., Scholze, H., Paques, E. P., and Deisenhofer, J. (1980). Crystal structure analysis and molecular model of human c3a anaphylatoxin. *Hoppe Seylers Z Physiol Chem*, 361(9):1389–99.

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

Isberg, V., de Graaf, C., Bortolato, A., Cherezov, V., Katritch, V., Marshall, F. H., Mordalski, S., Pin, J.-P., Stevens, R. C., Vriend, G., and Gloriam, D. E. (2015). Generic gpcr residue numbers - aligning topology maps while minding the gaps. *Trends in pharmacological sciences*, 36(1):22–31.

Isberg, V., Mordalski, S., Munk, C., Rataj, K., Harpsoe, K., Hauser, A. S., Vroling, B., Bojarski, A. J., Vriend, G., and Gloriam, D. E. (2017). Gpcrdb: an information system for g protein-coupled receptors. *Nucleic Acids Res*, 45(5):2936.

Jaakola, V. P., Griffith, M. T., Hanson, M. A., Cherezov, V., Chien, E. Y., Lane, J. R., Ijzerman, A. P., and Stevens, R. C. (2008). The 2.6 angstrom crystal structure of a human a2a adenosine receptor bound to an antagonist. *Science*, 322(5905):1211–7.

Jacob, T. F., Singh, V., Dixit, M., Ginsburg-Shmuel, T., Fonseca, B., Pintor, J., Youdim, M. B. H., Major, D. T., Weinreb, O., and Fischer, B. (2018). A promising drug candidate for the treatment of glaucoma based on a p2y6-receptor agonist. *Purinergic Signal*, 14(3):271–284.

Janes, R. W., Peapus, D. H., and Wallace, B. A. (1994). The crystal structure of human endothelin. *Nat Struct Biol*, 1(5):311–9.

Joedicke, L., Mao, J., Kuenze, G., Reinhart, C., Kalavacherla, T., Jonker, H. R. A., Richter, C., Schwalbe, H., Meiler, J., Preu, J., Michel, H., and Glaubitz, C. (2018). The molecular basis of subtype selectivity of human kinin g-protein-coupled receptors. *Nat Chem Biol*, 14(3):284–290.

Kaiser, A., Muller, P., Zellmann, T., Scheidt, H. A., Thomas, L., Bosse, M., Meier, R., Meiler, J., Huster, D., Beck-Sickinger, A. G., and Schmidt, P. (2015). Unwinding of the c-terminal residues of neuropeptide y is critical for y(2) receptor binding and activation. *Angew Chem Int Ed Engl*, 54(25):7446–9.

Keller, M., Kaske, M., Holzammer, T., Bernhardt, G., and Buschauer, A. (2013). Dimeric argininamide-type neuropeptide y receptor antagonists: chiral discrimination between y1 and y4 receptors. *Bioorg Med Chem*, 21(21):6303–22.

Kenakin, T. P. (2006). *A Pharmacology Primer - Theory, Applications, and Methods*. Elsvier Academic Press, Amsterdam, Boston, 2nd ed. edition.

Kepp, O., Loos, F., Liu, P., and Kroemer, G. (2017). Extracellular nucleosides and nucleotides as immunomodulators. *Immunol Rev*, 280(1):83–92.

Khasabova, I. A., Uhelski, M., Khasabov, S. G., Gupta, K., Seybold, V. S., and Simone, D. A. (2019). Sensitization of nociceptors by prostaglandin e2-glycerol contributes to hyperalgesia in mice with sickle cell disease. *Blood*, 133(18):1989–1998.

Kingsley, P. J., Rouzer, C. A., Morgan, A. J., Patel, S., and Marnett, L. J. (2019). Aspects of prostaglandin glycerol ester biology. *Adv Exp Med Biol*, 1161:77–88.

Koehl, A., Hu, H., Maeda, S., Zhang, Y., Qu, Q., Paggi, J. M., Latorraca, N. R., Hilger, D., Dawson, R., Matile, H., Schertler, G. F. X., Granier, S., Weis, W. I., Dror, R. O., Manglik, A., Skiniotis, G., and Kobilka, B. K. (2018). Structure of the micro-opioid receptor-gi protein complex. *Nature*, 558(7711):547–552.

Koizumi, S., Shigemoto-Mogami, Y., Nasu-Tada, K., Shinozaki, Y., Ohsawa, K., Tsuda, M., Joshi, B. V., Jacobson, K. A., Kohsaka, S., and Inoue, K. (2007). Udp acting at p2y6 receptors is a mediator of microglial phagocytosis. *Nature*, 446(7139):1091–5.

Kojima, M., Hosoda, H., Date, Y., Nakazato, M., Matsuo, H., and Kangawa, K. (1999). Ghrelin is a growth-hormone-releasing acylated peptide from stomach. *Nature*, 402(6762):656–60.

Kooistra, A. J., Kuhne, S., de Esch, I. J., Leurs, R., and de Graaf, C. (2013). A structural chemogenomics analysis of aminergic gpcrs: lessons for histamine receptor ligand design. *Br J Pharmacol*, 170(1):101–26.

Kooistra, A. J., Mordalski, S., Pándy-Szekeres, G., Esguerra, M., Mamyrbekov, A., Munk, C., Keserű, G. M., and Gloriam, D. E. (2021). Gpcrdb in 2021: integrating gpcr sequence, structure and function. *Nucleic Acids Research*, 49(D1):D335–D343.

Kostenis, E., Degtyarev, M. Y., Conklin, B. R., and Wess, J. (1997). The n-terminal extension of galphaq is critical for constraining the selectivity of receptor coupling. *J Biol Chem*, 272(31):19107–10.

Kothiwale, S., Mendenhall, J. L., and Meiler, J. (2015). Bcl::conf: small molecule conformational sampling using a knowledge based rotamer library. *J Cheminform*, 7:47.

Kozak, K. R., Crews, B. C., Ray, J. L., Tai, H. H., Morrow, J. D., and Marnett, L. J. (2001). Metabolism of prostaglandin glycerol esters and prostaglandin ethanolamides in vitro and in vivo. *J Biol Chem*, 276(40):36993–8.

Kuhn, K. K., Ertl, T., Dukorn, S., Keller, M., Bernhardt, G., Reiser, O., and Buschauer, A. (2016). High affinity agonists of the neuropeptide y (npy) y4 receptor derived from the c-terminal pentapeptide of human pancreatic polypeptide (hpp): synthesis, stereochemical discrimination, and radiolabeling. *J Med Chem*, 59(13):6045–58.

Kuhn, K. K., Littmann, T., Dukorn, S., Tanaka, M., Keller, M., Ozawa, T., Bernhardt, G., and Buschauer, A. (2017). In search of npy y4r antagonists: Incorporation of carbamoylated arginine, aza-amino acids, or d-amino acids into oligopeptides derived from the c-termini of the endogenous agonists. *ACS omega*, 2(7):3616–3631.

Kumar, P. and Sharma, A. (2014). Gonadotropin-releasing hormone analogs: Understanding advantages and limitations. *J Hum Reprod Sci*, 7(3):170–4.

Laimou, D. K., Katsara, M., Matsoukas, M. T., Apostolopoulos, V., Troganis, A. N., and Tselios, T. V. (2010). Structural elucidation of leuprolide and its analogues in solution: insight into their bioactive conformation. *Amino Acids*, 39(5):1147–60.

Lange, O. F., Lakomek, N. A., Fares, C., Schroder, G. F., Walter, K. F., Becker, S., Meiler, J., Grubmuller, H., Griesinger, C., and de Groot, B. L. (2008). Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. *Science*, 320(5882):1471–5.

Langelaan, D. N., Bebbington, E. M., Reddy, T., and Rainey, J. K. (2009). Structural insight into g-protein coupled receptor binding by apelin. *Biochemistry*, 48(3):537–48.

Lattig, J., Oksche, A., Beyermann, M., Rosenthal, W., and Krause, G. (2009). Structural determinants for selective recognition of peptide ligands for endothelin receptor subtypes eta and etb. *J Pept Sci*, 15(7):479–91.

Laursen, N. S., Andersen, K. R., Braren, I., Spillner, E., Sottrup-Jensen, L., and Andersen, G. R. (2011). Substrate recognition by complement convertases revealed in the c5-cobra venom factor complex. *EMBO J*, 30(3):606–16.

Laursen, N. S., Gordon, N., Hermans, S., Lorenz, N., Jackson, N., Wines, B., Spillner, E., Christensen, J. B., Jensen, M., Fredslund, F., Bjerre, M., Sottrup-Jensen, L., Fraser, J. D., and Andersen, G. R. (2010). Structural basis for inhibition of complement c5 by the ssl7 protein from staphylococcus aureus. *Proc Natl Acad Sci U S A*, 107(8):3681–6.

Lazarowski, E. R., Boucher, R. C., and Harden, T. K. (2003). Mechanisms of release of nucleotides and integration of their action as p2x- and p2y-receptor activating molecules. *Mol Pharmacol*, 64(4):785–95.

Le Duc, D., Schulz, A., Lede, V., Schulze, A., Thor, D., Bruser, A., and Schoneberg, T. (2017). P2y receptors in immune response and inflammation. *Adv Immunol*, 136:85–121.

Leach, K., Sexton, P. M., and Christopoulos, A. (2007). Allosteric gpcr modulators: taking advantage of permissive receptor pharmacology. *Trends Pharmacol Sci*, 28(8):382–9.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. (2011). Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology*, 487:545–574.

Lemmon, G. and Meiler, J. (2012). *Rosetta Ligand Docking with Flexible XML Protocols*, pages 143–155. Springer New York, New York, NY.

Li, Q., Olesky, M., Palmer, R. K., Harden, T. K., and Nicholas, R. A. (1998). Evidence that the p2y3 receptor is the avian homologue of the mammalian p2y6 receptor. *Mol Pharmacol*, 54(3):541–6.

Li, R., Tan, B., Yan, Y., Ma, X., Zhang, N., Zhang, Z., Liu, M., Qian, M., and Du, B. (2014). Extracellular udp and p2y6 function as a danger signal to protect mice from vesicular stomatitis virus infection through an increase in ifn-beta production. *J Immunol*, 193(9):4515–26.

Li, S., Li, J., Wang, N., Hao, G., and Sun, J. (2018). Characterization of udp-activated purinergic receptor p2y(6) involved in japanese flounder paralichthys olivaceus innate immunity. *Int J Mol Sci*, 19(7).

Liao, Z., Thibaut, L., Jobson, A., and Pommier, Y. (2006). Inhibition of human tyrosyl-dna phosphodiesterase by aminoglycoside antibiotics and ribosome inhibitors. *Molecular Pharmacology*, 70(1):366.

Lindner, D., Stichel, J., and Beck-Sickinger, A. G. (2008a). Molecular recognition of the npy hormone family by their receptors. *Nutrition*, 24(9):907–17.

Lindner, D., Stichel, J., and Beck-Sickinger, A. G. (2008b). Molecular recognition of the npy hormone family by their receptors. *Nutrition*, 24(9):907–17.

Liu, H., Kim, H. R., Deepak, R., Wang, L., Chung, K. Y., Fan, H., Wei, Z., and Zhang, C. (2018). Orthosteric and allosteric action of the c5a receptor antagonists. *Nat Struct Mol Biol*, 25(6):472–481.

Liu, W., Chun, E., Thompson, A. A., Chubukov, P., Xu, F., Katritch, V., Han, G. W., Roth, C. B., Heitman, L. H., IJzerman, A. P., Cherezov, V., and Stevens, R. C. (2012). Structural basis for allosteric regulation of gpcrs by sodium ions. *Science*, 337(6091):232–236.

Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I., and Lomize, A. L. (2012). Opm database and ppm web server: resources for positioning of proteins in membranes. *Nucleic Acids Res*, 40(Database issue):D370–6.

Lopez, J. J., Shukla, A. K., Reinhart, C., Schwalbe, H., Michel, H., and Glaubitz, C. (2008). The structure of the neuropeptide bradykinin bound to the human g-protein coupled receptor bradykinin b2 as determined by solid-state nmr spectroscopy. *Angew Chem Int Ed Engl*, 47(9):1668–71.

Lu, J., Byrne, N., Wang, J., Bricogne, G., Brown, F. K., Chobanian, H. R., Colletti, S. L., Di Salvo, J., Thomas-Fowlkes, B., Guo, Y., Hall, D. L., Hadix, J., Hastings, N. B., Hermes, J. D., Ho, T., Howard, A. D., Josien, H., Kornienko, M., Lumb, K. J., Miller, M. W., Patel, S. B., Pio, B., Plummer, C. W., Sherborne, B. S., Sheth, P., Souza, S., Tummala, S., Vonrhein, C., Webb, M., Allen, S. J., Johnston, J. M., Weinglass, A. B., Sharma, S., and Soisson, S. M. (2017). Structural basis for the cooperative allosteric activation of the free fatty acid receptor gpr40. *Nature Structural and Molecular Biology*, 24(7):570–577.

Lubecka, E. A., Sikorska, E., Sobolewski, D., Prahl, A., Slaninova, J., and Ciarkowski, J. (2015). Arginine-, d-arginine-vasopressin, and their inverso analogues in micellar and liposomic models of cell membrane: Cd, nmr, and molecular dynamics studies. *Eur Biophys J*, 44(8):727–43.

Luca, S., White, J. F., Sohal, A. K., Filippov, D. V., van Boom, J. H., Grisshammer, R., and Baldus, M. (2003). The conformation of neurotensin bound to its g protein-coupled receptor. *Proc Natl Acad Sci U S A*, 100(19):10706–11.

Lundell, I., Blomqvist, A. G., Berglund, M. M., Schober, D. A., Johnson, D., Statnick, M. A., Gadski, R. A., Gehlert, D. R., and Larhammar, D. (1995). Cloning of a human receptor of the npy receptor family with high affinity for pancreatic polypeptide and peptide yy. *J Biol Chem*, 270(49):29123–8.

Ma, Y., Yue, Y., Ma, Y., Zhang, Q., Zhou, Q., Song, Y., Shen, Y., Li, X., Ma, X., Li, C., Hanson, M. A., Han, G. W., Sickmier, E. A., Swaminath, G., Zhao, S., Stevens, R. C., Hu, L. A., Zhong, W., Zhang, M., and Xu, F. (2017). Structural basis for apelin control of the human apelin receptor. *Structure*, 25(6):858–866 e4.

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *J Chem Theory Comput*, 11(8):3696–713.

Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Mathiesen, J. M., Sunahara, R. K., Pardo, L., Weis, W. I., Kobilka, B. K., and Granier, S. (2012). Crystal structure of the micro-opioid receptor bound to a morphinan antagonist. *Nature*, 485(7398):321–6.

Marcou, G., Horvath, D., Solov'ev, V., Arrault, A., Vayer, P., and Varnek, A. (2012). Interpretability of sar/qsar models of any complexity by atomic contributions. *Molecular Informatics*, 31(9):639–642.

Markham, A. and Goa, K. L. (1997). Valsartan. *Drugs*, 54(2):299–311.

Martin-Pastor, M., De Capua, A., Alvarez, C. J., Diaz-Hernandez, M. D., Jimenez-Barbero, J., Casanueva, F. F., and Pazos, Y. (2010). Interaction between ghrelin and the ghrelin receptor (ghs-r1a), a nmr study using living cells. *Bioorg Med Chem*, 18(4):1583–90.

Matsumoto, M., Hosoda, H., Kitajima, Y., Morozumi, N., Minamitake, Y., Tanaka, S., Matsuo, H., Kojima, M., Hayashi, Y., and Kangawa, K. (2001a). Structure-activity relationship of ghrelin: pharmacological study of ghrelin peptides. *Biochem Biophys Res Commun*, 287(1):142–6.

Matsumoto, M., Kitajima, Y., Iwanami, T., Hayashi, Y., Tanaka, S., Minamitake, Y., Hosoda, H., Kojima, M., Matsuo, H., and Kangawa, K. (2001b). Structural similarity of ghrelin derivatives to peptidyl growth hormone secretagogues. *Biochem Biophys Res Commun*, 284(3):655–9.

Medhurst, A. D., Jennings, C. A., Robbins, M. J., Davis, R. P., Ellis, C., Winborn, K. Y., Lawrie, K. W., Hervieu, G., Riley, G., Bolaky, J. E., Herrity, N. C., Murdock, P., and Darker, J. G. (2003). Pharmacological and immunohistochemical characterization of the apj receptor and its endogenous ligand apelin. *J Neurochem*, 84(5):1162–72.

Meiler, J. and Baker, D. (2006). Rosettaligand: protein-small molecule docking with full side-chain flexibility. *Proteins*, 65(3):538–48.

Meiler, J. and Will, M. (2002). Genius: A genetic algorithm for automated structure elucidation from 13c nmr spectra. *Journal of the American Chemical Society*, 124(9):1868–1870.

Mendenhall, J., Brown, B. P., Kothiwale, S., and Meiler, J. (2021). Bcl::conf: Improved open-source knowledge-based conformation sampling using the crystallography open database. *J Chem Inf Model*, 61(1):189–201.

Mendenhall, J. and Meiler, J. (2016). Improving quantitative structure–activity relationship models using artificial neural networks trained with dropout. *Journal of Computer-Aided Molecular Design*, 30(2):177–189.

Merten, N., Lindner, D., Rabe, N., Rompler, H., Morl, K., Schoneberg, T., and Beck-Sickinger, A. G. (2007). Receptor subtype-specific docking of asp6.59 with c-terminal arginine residues in y receptor ligands. *J Biol Chem*, 282(10):7543–51.

Michino, M., Beuming, T., Donthamsetti, P., Newman, A. H., Javitch, J. A., and Shi, L. (2015). What can crystal structures of aminergic receptors tell us about designing subtype-selective ligands? *Pharmacol Rev*, 67(1):198–213.

Millar, R. P. (2005). Gnrhs and gnrh receptors. *Anim Reprod Sci*, 88(1-2):5–28.

Millar, R. P., Flanagan, C. A., Milton, R. C., and King, J. A. (1989). Chimeric analogues of vertebrate gonadotropin-releasing hormones comprising substitutions of the variant amino acids in positions 5, 7, and 8. characterization of requirements for receptor binding and gonadotropin release in mammalian and avian pituitary gonadotropes. *J Biol Chem*, 264(35):21007–13.

Minovski, N., Župerl, , Drgan, V., and Novič, M. (2013). Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum euclidean distance space analysis: A case study. *Analytica Chimica Acta*, 759:28–42.

Monahan, M. W., Amoss, M. S., Anderson, H. A., and Vale, W. (1973). Synthetic analogs of the hypothalamic luteinizing hormone releasing factor with increased agonist or antagonist properties. *Biochemistry*, 12(23):4616–20.

Moretti, R., Bender, B. J., Allison, B., and Meiler, J. (2016). Rosetta and the design of ligand binding sites. *Methods Mol Biol*, 1414:47–62.

Morozenko, A. and Stuchebrukhov, A. A. (2016). Dowser++, a new method of hydrating protein structures. *Proteins*, 84(10):1347–1357.

Murza, A., Parent, A., Besserer-Offroy, E., Tremblay, H., Karadereye, F., Beaudet, N., Leduc, R., Sarret, P., and Marsault, E. (2012). Elucidation of the structure-activity relationships of apelin: influence of unnatural amino acids on binding, signaling, and plasma stability. *ChemMedChem*, 7(2):318–25.

Mysinger, M. M. and Shoichet, B. K. (2010). Rapid context-dependent ligand desolvation in molecular docking. *Journal of Chemical Information and Modeling*, 50(9):1561–1573.

Mäde, V., Babilon, S., Jolly, N., Wanka, L., Bellmann-Sickert, K., Diaz Gimenez, L. E., Mörl, K., Cox, H. M., Gurevich, V. V., and Beck-Sickinger, A. G. (2014). Peptide modifications differentially alter g protein-coupled receptor internalization and signaling bias. *Angew Chem Int Ed Engl*, 53(38):10067–71.

Nakabayashi, K., Matsumi, H., Bhalla, A., Bae, J., Mosselman, S., Hsu, S. Y., and Hsueh, A. J. (2002). Thyrostimulin, a heterodimer of two new human glycoprotein hormone subunits, activates the thyroid-stimulating hormone receptor. *J Clin Invest*, 109(11):1445–52.

Nettesheim, D. G., Edalji, R. P., Mollison, K. W., Greer, J., and Zuiderweg, E. R. (1988). Secondary structure of complement component c3a anaphylatoxin in solution as determined by nmr spectroscopy: differences between crystal and solution conformations. *Proc Natl Acad Sci U S A*, 85(14):5036–40.

Ngo, T., Ilatovskiy, A. V., Stewart, A. G., Coleman, J. L., McRobb, F. M., Riek, R. P., Graham, R. M., Abagyan, R., Kufareva, I., and Smith, N. J. (2017). Orphan receptor ligand discovery by pickpocketing pharmacological neighbors. *Nat Chem Biol*, 13(2):235–242.

Nikiforovich, G. V., Marshall, G. R., and Baranski, T. J. (2008). Modeling molecular mechanisms of binding of the anaphylatoxin c5a to the c5a receptor. *Biochemistry*, 47(10):3117–30.

Nirodi, C. S., Crews, B. C., Kozak, K. R., Morrow, J. D., and Marnett, L. J. (2004). The glyceryl ester of prostaglandin e2 mobilizes calcium and activates signal transduction in raw264.7 cells. *Proc Natl Acad Sci U S A*, 101(7):1840–5.

O'Connor, C., White, K. L., Doncescu, N., Didenko, T., Roth, B. L., Czaplicki, G., Stevens, R. C., Wuthrich, K., and Milon, A. (2015). Nmr structure and dynamics of the agonist dynorphin peptide bound to the human kappa opioid receptor. *Proc Natl Acad Sci U S A*, 112(38):11852–7.

Ogden CL, Carroll MD, F. C. F. K. (2015). Prevalence of obesity among adults and youth: United states, 2011–2014. nchs data brief, no 219. *National Center for Health Statistics*.

Oswald, C., Rappas, M., Kean, J., Dore, A. S., Errey, J. C., Bennett, K., Deflorian, F., Christopher, J. A., Jazayeri, A., Mason, J. S., Congreve, M., Cooke, R. M., and Marshall, F. H. (2016). Intracellular allosteric antagonism of the ccr9 receptor. *Nature*, 540(7633):462–465.

Parker, E. M., Babij, C. K., Balasubramaniam, A., Burrier, R. E., Guzzi, M., Hamud, F., Mukhopadhyay, G., Rudinski, M. S., Tao, Z., Tice, M., Xia, L., Mullins, D. E., and Salisbury, B. G. (1998). Gr231118 (1229u91) and other analogues of the c-terminus of neuropeptide y are potent neuropeptide y y1 receptor antagonists and neuropeptide y y4 receptor agonists. *Eur J Pharmacol*, 349(1):97–105.

Pedragosa-Badia, X., Sliwoski, G. R., Dong Nguyen, E., Lindner, D., Stichel, J., Kaufmann, K. W., Meiler, J., and Beck-Sickinger, A. G. (2014). Pancreatic polypeptide is recognized by two hydrophobic domains of the human y4 receptor binding pocket. *J Biol Chem*, 289(9):5846–59.

Pedragosa-Badia, X., Stichel, J., and Beck-Sickinger, A. G. (2013a). Neuropeptide y receptors: how to get subtype selectivity. *Front Endocrinol (Lausanne)*, 4:5.

Pedragosa-Badia, X., Stichel, J., and Beck-Sickinger, A. G. (2013b). Neuropeptide y receptors: how to get subtype selectivity. *Front Endocrinol*, 4:5.

Portoghese, P. S. (1989). Bivalent ligands and the message-address concept in the design of selective opioid receptor antagonists. *Trends Pharmacol Sci*, 10(6):230–5.

Qin, L., Kufareva, I., Holden, L. G., Wang, C., Zheng, Y., Zhao, C., Fenalti, G., Wu, H., Han, G. W., Cherezov, V., Abagyan, R., Stevens, R. C., and Handel, T. M. (2015). Structural biology. crystal structure of the chemokine receptor cxcr4 in complex with a viral chemokine. *Science*, 347(6226):1117–22.

Rana, S. and Sahoo, A. R. (2015). Model structures of inactive and peptide agonist bound c5ar: Insights into agonist binding, selectivity and activation. *Biochem Biophys Rep*, 1:85–96.

Rathmann, D., Lindner, D., DeLuca, S. H., Kaufmann, K. W., Meiler, J., and Beck-Sickinger, A. G. (2012). Ligand-mimicking receptor variant discloses binding and activation mode of prolactin-releasing peptide. *J Biol Chem*, 287(38):32181–94.

Raveh, B., London, N., Zimmerman, L., and Schueler-Furman, O. (2011). Rosetta flexpepdock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. *PLoS One*, 6(4):e18934.

Reifel Saltzberg, J. M., Garvey, K. A., and Keirstead, S. A. (2003). Pharmacological characterization of p2y receptor subtypes on isolated tiger salamander muller cells. *Glia*, 42(2):149–59.

Richie-Jannetta, R., Nirodi, C. S., Crews, B. C., Woodward, D. F., Wang, J. W., Duff, P. T., and Marnett, L. J. (2010). Structural determinants for calcium mobilization by prostaglandin e2 and prostaglandin f2alpha glyceryl esters in raw 264.7 cells and h1819 cells. *Prostaglandins Other Lipid Mediat*, 92(1-4):19–24.

Robertson, N., Rappas, M., Dore, A. S., Brown, J., Bottegoni, G., Koglin, M., Cansfield, J., Jazayeri, A., Cooke, R. M., and Marshall, F. H. (2018). Structure of the complement c5a receptor bound to the extra-helical antagonist ndt9513727. *Nature*, 553(7686):111–114.

Roe, D. R. and Cheatham, T. E. (2013). Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation*, 9(7):3084–3095.

Rogers, D. J. and Tanimoto, T. T. (1960a). A computer program for classifying plants. *Science*, 132(3434):1115–8.

Rogers, D. J. and Tanimoto, T. T. (1960b). A computer program for classifying plants. *Science*, 132(3434):1115–1118.

Rosengren, K. J., Lin, F., Bathgate, R. A., Tregear, G. W., Daly, N. L., Wade, J. D., and Craik, D. J. (2006). Solution structure and novel insights into the determinants of the receptor specificity of human relaxin-3. *J Biol Chem*, 281(9):5845–51.

Rudolf, K., Eberlein, W., Engel, W., Wieland, H. A., Willim, K. D., Entzeroth, M., Wienen, W., Beck-Sickinger, A. G., and Doods, H. N. (1994). The first highly potent and selective non-peptide neuropeptide y y1 receptor antagonist: Bibp3226. *Eur J Pharmacol*, 271(2-3):R11–3.

Sahoo, A. R., Mishra, R., and Rana, S. (2018). The model structures of the complement component 5a receptor (c5ar) bound to the native and engineered (h)c5a. *Sci Rep*, 8(1):2955.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–25.

Sangkuhl, K., Schulz, A., Schultz, G., and Schoneberg, T. (2002). Structural requirements for mutational lutropin/choriogonadotropin receptor activation. *J Biol Chem*, 277(49):47748–55.

Sastry, M., Lowrie, J. F., Dixon, S. L., and Sherman, W. (2010). Large-scale systematic analysis of 2d fingerprint methods and parameters to improve virtual screening enrichments. *J Chem Inf Model*, 50(5):771–84.

Saudek, V. and Pelton, J. T. (1990). Sequence-specific 1h nmr assignment and secondary structure of neuropeptide y in aqueous solution. *Biochemistry*, 29(19):4509–15.

Schatz-Jakobsen, J. A., Yatime, L., Larsen, C., Petersen, S. V., Klos, A., and Andersen, G. R. (2014). Structural and functional characterization of human and murine c5a anaphylatoxins. *Acta Crystallogr D Biol Crystallogr*, 70(Pt 6):1704–17.

Schoneberg, T., Schulz, A., Biebermann, H., Gruters, A., Grimm, T., Hubschmann, K., Filler, G., Gudermann, T., and Schultz, G. (1998). V2 vasopressin receptor dysfunction in nephrogenic diabetes insipidus caused by different molecular mechanisms. *Hum Mutat*, 12(3):196–205.

Schott-Verdugo, S. and Gohlke, H. (2019). Packmol-memgen: A simple-to-use, generalized workflow for membrane-protein–lipid-bilayer system building. *Journal of Chemical Information and Modeling*, 59(6):2522–2528.

Schubert, M., Stichel, J., Du, Y., Tough, I. R., Sliwoski, G., Meiler, J., Cox, H. M., Weaver, C. D., and Beck-Sickinger, A. G. (2017a). Identification and characterization of the first selective y4 receptor positive allosteric modulator. *J Med Chem*, 60(17):7605–7612.

Schubert, M., Stichel, J., Du, Y., Tough, I. R., Sliwoski, G., Meiler, J., Cox, H. M., Weaver, C. D., and Beck-Sickinger, A. G. (2017b). Identification and characterization of the first selective y4 receptor positive allosteric modulator. *J Med Chem*.

Schwartz, T. W. (1983). Pancreatic polypeptide: a hormone under vagal control. *Gastroenterology*, 85(6):1411–25.

Schüß, C., Vu, O., Schubert, M., Du, Y., Mishra, N. M., Tough, I. R., Stichel, J., Weaver, C. D., Emmitte, K. A., Cox, H. M., Meiler, J., and Beck-Sickinger, A. G. (2021a). Highly selective y4 receptor antagonist binds in an allosteric binding pocket. *Journal of Medicinal Chemistry*, 64(5):2801–2814.

Schüß, C., Vu, O., Schubert, M., Du, Y., Mishra, N. M., Tough, I. R., Stichel, J., Weaver, C. D., Emmitte, K. A., Cox, H. M., Meiler, J., and Beck-Sickinger, A. G. (2021b). Highly selective y(4) receptor antagonist binds in an allosteric binding pocket. *J Med Chem*, 64(5):2801–2814.

Scott, L. J. and McCormack, P. L. (2008). Olmesartan medoxomil. *Drugs*, 68(9):1239–1272.

Sealfon, S. C., Weinstein, H., and Millar, R. P. (1997). Molecular mechanisms of ligand interaction with the gonadotropin-releasing hormone receptor. *Endocr Rev*, 18(2):180–205.

Shaik, M. M., Peng, H., Lu, J., Rits-Volloch, S., Xu, C., Liao, M., and Chen, B. (2019). Structural basis of coreceptor recognition by hiv-1 envelope spike. *Nature*, 565(7739):318–323.

Shihoya, W., Izume, T., Inoue, A., Yamashita, K., Kadji, F. M. N., Hirata, K., Aoki, J., Nishizawa, T., and Nureki, O. (2018). Crystal structures of human etb receptor provide mechanistic insight into receptor activation and partial activation. *Nature Communications*, 9(1):4711.

Shihoya, W., Nishizawa, T., Okuta, A., Tani, K., Dohmae, N., Fujiyoshi, Y., Nureki, O., and Doi, T. (2016). Activation mechanism of endothelin etb receptor by endothelin-1. *Nature*, 537(7620):363–368.

Shinozaki, Y., Kashiwagi, K., Namekata, K., Takeda, A., Ohno, N., Robaye, B., Harada, T., Iwata, T., and Koizumi, S. (2017). Purinergic dysregulation causes hypertensive glaucoma-like optic neuropathy. *JCI Insight*, 2(19).

Siciliano, S. J., Rollins, T. E., DeMartino, J., Konteatis, Z., Malkowitz, L., Van Riper, G., Bondy, S., Rosen, H., and Springer, M. S. (1994). Two-site binding of c5a by its receptor: an alternative binding paradigm for g protein-coupled receptors. *Proc Natl Acad Sci U S A*, 91(4):1214–8.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol Syst Biol*, 7:539.

Siligardi, G. and Drake, A. F. (1995). The importance of extended conformations and, in particular, the pii conformation for the molecular recognition of peptides. *Biopolymers*, 37(4):281–92.

Silva Elipe, M. V., Bednarek, M. A., and Gao, Y. D. (2001). 1h nmr structural analysis of human ghrelin and its six truncated analogs. *Biopolymers*, 59(7):489–501.

Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E. W., J. (2014). Computational methods in drug discovery. *Pharmacol Rev*, 66(1):334–95.

Sliwoski, G., Mendenhall, J., and Meiler, J. (2016a). Autocorrelation descriptor improvements for qsar: 2da_sign and 3da_sign. *J Comput Aided Mol Des*, 30(3):209–17.

Sliwoski, G., Schubert, M., Stichel, J., Weaver, D., Beck-Sickinger, A. G., and Meiler, J. (2016b). Discovery of small-molecule modulators of the human y4 receptor. *PloS one*, 11(6):e0157146.

Smith, C. A. and Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of Molecular Biology*, 380(4):742–756.

Song, Y., DiMaio, F., Wang, R. Y., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013a). High-resolution comparative modeling with rosettacm. *Structure*, 21(10):1735–42.

Song, Y., DiMaio, F., Wang, R. Y.-R., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013b). High-resolution comparative modeling with rosettacm. *Structure (London, England : 1993)*, 21(10):1735–1742.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Stachon, P., Peikert, A., Michel, N. A., Hergeth, S., Marchini, T., Wolf, D., Dufner, B., Hoppe, N., Ayata, C. K., Grimm, M., Cicko, S., Schulte, L., Reinohl, J., von zur Muhlen, C., Bode, C., Idzko, M., and Zirlik, A. (2014). P2y6 deficiency limits vascular inflammation and atherosclerosis in mice. *Arterioscler Thromb Vasc Biol*, 34(10):2237–45.

Staes, E., Absil, P. A., Lins, L., Brasseur, R., Deleu, M., Lecouturier, N., Fievez, V., Rieux, A., Mingeot-Leclercq, M. P., Raussens, V., and Preat, V. (2010). Acylated and unacylated ghrelin binding to membranes and to ghrelin receptor: towards a better understanding of the underlying mechanisms. *Biochim Biophys Acta*, 1798(11):2102–13.

Takashima, H., Tamaoki, H., Teno, N., Nishi, Y., Uchiyama, S., Fukui, K., and Kobayashi, Y. (2004). Hydrophobic core around tyrosine for human endothelin-1 investigated by photochemically induced dynamic nuclear polarization nuclear magnetic resonance and matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Biochemistry*, 43(44):13932–6.

Takasuka, T., Sakurai, T., Goto, K., Furuichi, Y., and Watanabe, T. (1994). Human endothelin receptor etb. amino acid sequence requirements for super stable complex formation with its ligand. *J Biol Chem*, 269(10):7509–13.

Tan, Q., Zhu, Y., Li, J., Chen, Z., Han, G. W., Kufareva, I., Li, T., Ma, L., Fenalti, G., Li, J., Zhang, W., Xie, X., Yang, H., Jiang, H., Cherezov, V., Liu, H., Stevens, R. C., Zhao, Q., and Wu, B. (2013). Structure of the ccr5 chemokine receptor-hiv entry inhibitor maraviroc complex. *Science*, 341(6152):1387–90.

Tanaka, H., Yoshida, T., Miyamoto, N., Motoike, T., Kurosu, H., Shibata, K., Yamanaka, A., Williams, S. C., Richardson, J. A., Tsujino, N., Garry, M. G., Lerner, M. R., King, D. S., O'Dowd, B. F., Sakurai, T., and Yanagisawa, M. (2003). Characterization of a family of endogenous neuropeptide ligands for the g protein-coupled receptors gpr7 and gpr8. *Proceedings of the National Academy of Sciences*, 100(10):6251–6256.

Thomas, L., Scheidt, H. A., Bettio, A., Huster, D., Beck-Sickinger, A. G., Arnold, K., and Zschornig, O. (2005). Membrane interaction of neuropeptide y detected by epr and nmr spectroscopy. *Biochim Biophys Acta*, 1714(2):103–13.

Thompson, A. A., Liu, W., Chun, E., Katritch, V., Wu, H., Vardy, E., Huang, X. P., Trapella, C., Guerrini, R., Calo, G., Roth, B. L., Cherezov, V., and Stevens, R. C. (2012). Structure of the nociceptin/orphanin fq receptor in complex with a peptide mimetic. *Nature*, 485(7398):395–9.

Tikhonova, I. G., Gigoux, V., and Fourmy, D. (2019). Understanding peptide binding in class a g protein-coupled receptors. *Molecular Pharmacology*, page mol.119.115915.

Tough, I. R., Holliday, N. D., and Cox, H. M. (2006). Y(4) receptors mediate the inhibitory responses of pancreatic polypeptide in human and mouse colon mucosa. *J Pharmacol Exp Ther*, 319(1):20–30.

Tyndall, J. D., Nall, T., and Fairlie, D. P. (2005a). Proteases universally recognize beta strands in their active sites. *Chem Rev*, 105(3):973–99.

Tyndall, J. D., Pfeiffer, B., Abbenante, G., and Fairlie, D. P. (2005b). Over one hundred peptide-activated g protein-coupled receptors recognize ligands with turn structure. *Chem Rev*, 105(3):793–826.

Van Craenenbroeck, M., Gregoire, F., De Neef, P., Robberecht, P., and Perret, J. (2004). Ala-scan of ghrelin (1-14): interaction with the recombinant human ghrelin receptor. *Peptides*, 25(6):959–65.

Van Kemmel, F. M., Dubuc, I., Bourdel, E., Fehrentz, J. A., Martinez, J., and Costentin, J. (1996). A c-terminal cyclic 8-13 neurotensin fragment analog appears less exposed to neprilysin when it crosses the blood-brain barrier than the cerebrospinal fluid-brain barrier in mice. *Neurosci Lett*, 217(1):58–60.

Viklund, H. and Elofsson, A. (2008). Octopus: improving topology prediction by two-track ann-based preference scores and an extended topological grammar. *Bioinformatics*, 24(15):1662–8.

Vilar, S., Cozza, G., and Moro, S. (2008). Medicinal chemistry and the molecular operating environment (moe): Application of qsar and molecular docking to drug discovery. *Current Topics in Medicinal Chemistry*, 8(18):1555–1572.

Vortmeier, G., DeLuca, S. H., Els-Heindl, S., Chollet, C., Scheidt, H. A., Beck-Sickinger, A. G., Meiler, J., and Huster, D. (2015). Integrating solid-state nmr and computational modeling to investigate the structure and dynamics of membrane-associated ghrelin. *PLoS One*, 10(3):e0122444.

Wacker, D., Stevens, R. C., and Roth, B. L. (2017). How ligands illuminate gpcr molecular pharmacology. *Cell*, 170(3):414–427.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *J Comput Chem*, 25(9):1157–74.

Wanka, L., Babilon, S., Burkert, K., Morl, K., Gurevich, V. V., and Beck-Sickinger, A. G. (2017). C-terminal motif of human neuropeptide y4 receptor determines internalization and arrestin recruitment. *Cell Signal*, 29:233–239.

Warny, M., Aboudola, S., Robson, S. C., Sevigny, J., Communi, D., Soltoff, S. P., and Kelly, C. P. (2001). P2y(6) nucleotide receptor mediates monocyte interleukin-8 production in response to udp or lipopolysaccharide. *J Biol Chem*, 276(28):26051–6.

Weaver, S. and Gleeson, M. P. (2008). The importance of the domain of applicability in qsar modeling. *Journal of Molecular Graphics and Modelling*, 26(8):1315–1326.

Webb, T. E., Henderson, D., King, B. F., Wang, S., Simon, J., Bateson, A. N., Burnstock, G., and Barnard, E. A. (1996). A novel g protein-coupled p2 purinoceptor (p2y3) activated preferentially by nucleoside diphosphates. *Mol Pharmacol*, 50(2):258–65.

White, J. F., Noinaj, N., Shibata, Y., Love, J., Kloss, B., Xu, F., Gvozdenovic-Jeremic, J., Shah, P., Shiloach, J., Tate, C. G., and Grisshammer, R. (2012). Structure of the agonist-bound neurotensin receptor. *Nature*, 490(7421):508–13.

Wieland, H. A., Engel, W., Eberlein, W., Rudolf, K., and Doods, H. N. (1998). Subtype selectivity of the novel nonpeptide neuropeptide y y1 receptor antagonist bibo 3304 and its effect on feeding in rodents. *Br J Pharmacol*, 125(3):549–55.

Williamson, M. P. and Madison, V. S. (1990). Three-dimensional structure of porcine c5adesarg from 1h nuclear magnetic resonance data. *Biochemistry*, 29(12):2895–905.

Williamson, P. T., Bains, S., Chung, C., Cooke, R., and Watts, A. (2002). Probing the environment of neurotensin whilst bound to the neurotensin receptor by solid state nmr. *FEBS Lett*, 518(1-3):111–5.

Wingler, L. M., McMahon, C., Staus, D. P., Lefkowitz, R. J., and Kruse, A. C. (2019). Distinctive activation mechanism for angiotensin receptor revealed by a synthetic nanobody. *Cell*, 176(3):479–490.e12.

Woodward, D. F., Jones, R. L., and Narumiya, S. (2011). International union of basic and clinical pharmacology. lxxxiii: classification of prostanoid receptors, updating 15 years of progress. *Pharmacol Rev*, 63(3):471–538.

Woodward, D. F., Poloso, N. J., and Wang, J. W. (2016). Prostaglandin e2-glyceryl ester: In vivo evidence for a distinct pharmacological identity from intraocular pressure studies. *J Pharmacol Exp Ther*, 358(2):173–80.

Woolley, M. J. and Conner, A. C. (2017). Understanding the common themes and diverse roles of the second extracellular loop (ecl2) of the gpcr super-family. *Molecular and Cellular Endocrinology*, 449:3–11.

Wootten, D., Christopoulos, A., and Sexton, P. M. (2013). Emerging paradigms in gpcr allostery: implications for drug discovery. *Nat Rev Drug Discov*, 12(8):630–44.

Wren, A. M. and Bloom, S. R. (2007). Gut hormones and appetite control. *Gastroenterology*, 132(6):2116–30.

Wu, B., Chien, E. Y., Mol, C. D., Fenalti, G., Liu, W., Katritch, V., Abagyan, R., Brooun, A., Wells, P., Bi, F. C., Hamel, D. J., Kuhn, P., Handel, T. M., Cherezov, V., and Stevens, R. C. (2010). Structures of the cxcr4 chemokine gpcr with small-molecule and cyclic peptide antagonists. *Science*, 330(6007):1066–71.

Wu, F., Song, G., de Graaf, C., and Stevens, R. C. (2017). Structure and function of peptide-binding g protein-coupled receptors. *J Mol Biol*, 429(17):2726–2745.

Wu, H., Wacker, D., Mileni, M., Katritch, V., Han, G. W., Vardy, E., Liu, W., Thompson, A. A., Huang, X. P., Carroll, F. I., Mascarella, S. W., Westkaemper, R. B., Mosier, P. D., Roth, B. L., Cherezov, V., and Stevens, R. C. (2012). Structure of the human κ-opioid receptor in complex with jdtic. *Nature*, 485(7398):327–32.

Xu, G. Y. and Deber, C. M. (1991). Conformations of neurotensin in solution and in membrane environments studied by 2-d nmr spectroscopy. *Int J Pept Protein Res*, 37(6):528–35.

Yang, Z., Han, S., Keller, M., Kaiser, A., Bender, B. J., Bosse, M., Burkert, K., Kogler, L. M., Wifling, D., Bernhardt, G., Plank, N., Littmann, T., Schmidt, P., Yi, C., Li, B., Ye, S., Zhang, R., Xu, B., Larhammar, D., Stevens, R. C., Huster, D., Meiler, J., Zhao, Q., Beck-Sickinger, A. G., Buschauer, A., and Wu, B. (2018). Structural basis of ligand binding modes at the neuropeptide y y1 receptor. *Nature*, 556(7702):520–524.

Yao, X., Parnot, C., Deupi, X., Ratnala, V. R., Swaminath, G., Farrens, D., and Kobilka, B. (2006). Coupling ligand structure to specific conformational switches in the beta2-adrenoceptor. *Nat Chem Biol*, 2(8):417–22.

Yarov-Yarovoy, V., Schonbrun, J., and Baker, D. (2006). Multipass membrane protein structure prediction using rosetta. *Proteins*, 62(4):1010–1025.

Yin, J., Babaoglu, K., Brautigam, C. A., Clark, L., Shao, Z., Scheuermann, T. H., Harrell, C. M., Gotter, A. L., Roecker, A. J., Winrow, C. J., Renger, J. J., Coleman, P. J., and Rosenbaum, D. M. (2016). Structure and ligand-binding mechanism of the human ox1 and ox2 orexin receptors. *Nat Struct Mol Biol*, 23(4):293–9.

Yin, J., Chapman, K., Clark, L. D., Shao, Z., Borek, D., Xu, Q., Wang, J., and Rosenbaum, D. M. (2018). Crystal structure of the human nk1 tachykinin receptor. *Proc Natl Acad Sci U S A*, 115(52):13264–13269.

Yin, J., Mobarec, J. C., Kolb, P., and Rosenbaum, D. M. (2015). Crystal structure of the human ox2 orexin receptor bound to the insomnia drug suvorexant. *Nature*, 519(7542):247–50.

Yuan, S., Chan, H. C. S., Vogel, H., Filipek, S., Stevens, R. C., and Palczewski, K. (2016). The molecular mechanism of p2y1 receptor activation. *Angewandte Chemie International Edition*, 55(35):10331–10335.

Yulyaningsih, E., Zhang, L., Herzog, H., and Sainsbury, A. (2011). Npy receptors as potential targets for anti-obesity drug development. *Br J Pharmacol*, 163(6):1170–202.

Zhang, C., Srinivasan, Y., Arlow, D. H., Fung, J. J., Palmer, D., Zheng, Y., Green, H. F., Pandey, A., Dror, R. O., Shaw, D. E., Weis, W. I., Coughlin, S. R., and Kobilka, B. K. (2012). High-resolution crystal structure of human protease-activated receptor 1. *Nature*, 492(7429):387–92.

Zhang, D., Gao, Z. G., Zhang, K., Kiselev, E., Crane, S., Wang, J., Paoletta, S., Yi, C., Ma, L., Zhang, W., Han, G. W., Liu, H., Cherezov, V., Katritch, V., Jiang, H., Stevens, R. C., Jacobson, K. A., Zhao, Q., and Wu, B. (2015a). Two disparate ligand-binding sites in the human p2y1 receptor. *Nature*, 520(7547):317–21.

Zhang, H., Han, G. W., Batyuk, A., Ishchenko, A., White, K. L., Patel, N., Sadybekov, A., Zamlynny, B., Rudd, M. T., Hollenstein, K., Tolstikova, A., White, T. A., Hunter, M. S., Weierstall, U., Liu, W., Babaoglu, K., Moore, E. L., Katz, R. D., Shipman, J. M., Garcia-Calvo, M., Sharma, S., Sheth, P., Soisson, S. M., Stevens, R. C., Katritch, V., and Cherezov, V. (2017). Structural basis for selectivity and diversity in angiotensin ii receptors. *Nature*, 544(7650):327–332.

Zhang, H., Unal, H., Desnoyer, R., Han, G. W., Patel, N., Katritch, V., Karnik, S. S., Cherezov, V., and Stevens, R. C. (2015b). Structural basis for ligand recognition and functional selectivity at angiotensin receptor. *J Biol Chem*, 290(49):29127–39.

Zhang, H., Unal, H., Gati, C., Han, G. W., Liu, W., Zatsepin, N. A., James, D., Wang, D., Nelson, G., Weierstall, U., Sawaya, M. R., Xu, Q., Messerschmidt, M., Williams, G. J., Boutet, S., Yefanov, O. M., White, T. A., Wang, C., Ishchenko, A., Tirupula, K. C., Desnoyer, R., Coe, J., Conrad, C. E., Fromme, P., Stevens, R. C., Katritch, V., Karnik, S. S., and Cherezov, V. (2015c). Structure of the angiotensin receptor revealed by serial femtosecond crystallography. *Cell*, 161(4):833–44.

Zhang, J., Zhang, K., Gao, Z. G., Paoletta, S., Zhang, D., Han, G. W., Li, T., Ma, L., Zhang, W., Muller, C. E., Yang, H., Jiang, H., Cherezov, V., Katritch, V., Jacobson, K. A., Stevens, R. C., Wu, B., and Zhao, Q. (2014a). Agonist-bound structure of the human p2y12 receptor. *Nature*, 509(7498):119–22.

Zhang, K., Zhang, J., Gao, Z. G., Zhang, D., Zhu, L., Han, G. W., Moss, S. M., Paoletta, S., Kiselev, E., Lu, W., Fenalti, G., Zhang, W., Muller, C. E., Yang, H., Jiang, H., Cherezov, V., Katritch, V., Jacobson, K. A., Stevens, R. C., Wu, B., and Zhao, Q. (2014b). Structure of the human p2y12 receptor in complex with an antithrombotic drug. *Nature*, 509(7498):115–8.

Zhang, L., Bijker, M. S., and Herzog, H. (2011). The neuropeptide y system: pathophysiological and therapeutic implications in obesity and cancer. *Pharmacol Ther*, 131(1):91–113.

Zhang, L., Riepler, S. J., Turner, N., Enriquez, R. F., Lee, I. C., Baldock, P. A., Herzog, H., and Sainsbury, A. (2010). Y2 and y4 receptor signaling synergistically act on energy expenditure and physical activity. *Am J Physiol Regul Integr Comp Physiol*, 299(6):R1618–28.

Zhang, X., Boyar, W., Toth, M. J., Wennogle, L., and Gonnella, N. C. (1997). Structural definition of the c5a c terminus by two-dimensional nuclear magnetic resonance spectroscopy. *Proteins*, 28(2):261–7.

Zheng, W., Cho, S. J., and Tropsha, A. (1998). Rational combinatorial library design. 1. focus-2d: A new approach to the design of targeted combinatorial chemical libraries. *Journal of Chemical Information and Computer Sciences*, 38(2):251–258.

Zheng, Y., Han, G. W., Abagyan, R., Wu, B., Stevens, R. C., Cherezov, V., Kufareva, I., and Handel, T. M. (2017). Structure of cc chemokine receptor 5 with a potent chemokine antagonist reveals mechanisms of chemokine recognition and molecular mimicry by hiv. *Immunity*, 46(6):1005–1017 e5.

Zheng, Y., Qin, L., Zacarias, N. V., de Vries, H., Han, G. W., Gustavsson, M., Dabros, M., Zhao, C., Cherney, R. J., Carter, P., Stamos, D., Abagyan, R., Cherezov, V., Stevens, R. C., AP, I. J., Heitman, L. H., Tebben, A., Kufareva, I., and Handel, T. M. (2016). Structure of cc chemokine receptor 2 with orthosteric and allosteric antagonists. *Nature*, 540(7633):458–461.

Zhou, Q., Yang, D., Wu, M., Guo, Y., Guo, W., Zhong, L., Cai, X., Dai, A., Jang, W., Shakhnovich, E. I., Liu, Z. J., Stevens, R. C., Lambert, N. A., Babu, M. M., Wang, M. W., and Zhao, S. (2019). Common activation mechanism of class a gpcrs. *Elife*, 8.

Ziemek, R., Schneider, E., Kraus, A., Cabrele, C., Beck-Sickinger, A. G., Bernhardt, G., and Buschauer, A. (2007). Determination of affinity and activity of ligands at the human neuropeptide y y4 receptor by flow cytometry and aequorin luminescence. *J Recept Signal Transduct Res*, 27(4):217–33.

Zuiderweg, E. R., Nettesheim, D. G., Mollison, K. W., and Carter, G. W. (1989). Tertiary structure of human complement component c5a in solution from nuclear magnetic resonance data. *Biochemistry*, 28(1):172–85.