Computer-Aided Elucidation of Structural Determinants for Sphingosine-1-phosphate Receptor Subtype Selectivity and Class A GPCR Activation

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry, Pharmacy

of Freie Universität Berlin

by

Trung Ngoc Nguyen

2022

The presented thesis was prepared from 2017 to 2022 under the supervision of Prof. Dr. Gerhard Wolber at the Institute of Pharmacy of the Freie Universität Berlin.

1st reviewer:	Prof. Dr. Gerhard Wolber
2nd reviewer:	Prof. Dr. Burkhard Kleuser
Date of defense:	20.09.2022

“

Corpora non agunt nisi fixate

A drug can only act when bound to its target

Paul Ehrlich 1891

Acknowledgment

I would first like to thank my thesis advisor Prof. Dr. Gerhard Wolber of the institute of pharmacy at Freie Universität Berlin for giving me the opportunity to write this thesis. He provided me with the workspace, resources, and ideas for this thesis. I am gratefully indebted to him for his valuable comments on this thesis.

I would also like to acknowledge Dr. Marcel Bermudez for his insights throughout this journey and Dr. Szymon Pach for the scientific and emotional support he has provided. I would also like to thank Dr. David Schaller for his aid in scripting a framework for the analysis of dihedral angles. I would also like to thank all the members of the Molecular Design Lab for these challenging, fruitful, and fun years we had together.

Finally, I must express my very profound gratitude to my parents, my partner Lara D. Hoppe, and friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Trung Ngoc Nguyen

Independence Declaration

I hereby declare that the presented work has been independently completed without any unauthorized assistance. No additional aids than these stated in the text were used. Furthermore, this thesis has not been submitted to any other institution.

Berlin, Germany

Juni 25, 2022

Trung Ngoc Nguyen

Table of Contents

Summary

G-protein coupled receptors (GPCRs) represent the biggest family of membrane receptors. The physiological accessibility of drugs, their regulatory roles in a vast amount of physiological and pathophysiological processes, and their prevalence in many tissues are reasons why they are highly targeted in a therapeutic context. To exploit the modulatory capabilities of GPCRs prior knowledge of their mechanism of action on a molecular level is required. This is especially important for the successful application of rational structure-based drug design campaigns. Sphingosine-1-phosphate receptors (S1PR) has shown to be effective drug targets against multiple sclerosis (MS), but non-selective drugs suffer from serious adverse effects. Furthermore, targeting a single S1PR holds a lot of potentials to modulate different inflammatory and autoimmune diseases.

In this thesis, we present *in silico* mechanistic models for the identification of crucial structural determinants to illuminate the molecular basis for Sphingosine-1-phosphate receptor selectivity and the general class A GPCR activation process. We employed extensive molecular dynamic (MD) simulation models of $S1PR_{1\text{-}5}$ and performed a holistic comparative orthosteric protein-ligand interaction analysis of identified different binding modes of a pan-agonist to overcome the challenge of highly similar binding pockets between each subtype and rationally explained the selective behavior of marketed drugs Ozanimod and Siponimod used in multiple sclerosis therapy. Furthermore, we provide insights into the mechanism of class A GPCR activation and how dihedral angles of the protein backbone are involved in this process. By developing a data extraction and machine learning workflow we created predictive models for active and inactive state GPCR conformations and showed possible applications for GPCR homology models and MD simulation predictions in a fast and efficient way. The methods and workflow to apply the prediction models are provided as a python package on GitHub (github.com/TrungNgocNguyen/GPCRml).

The unique and rational combination of state-of-the-art molecular modeling, data extraction and analysis, and machine learning methods demonstrate how explanatory *in silico* models can be developed to elevate the understanding of challenging problems in the field of GPCRs.

Zusammenfassung

G-Protein-gekoppelte Rezeptoren stellen die größte Familie von Membranrezeptoren dar. Die physiologische Zugänglichkeit für Medikamente, ihre regulatorische Rolle in einer Vielzahl von physiologischen und pathophysiologischen Prozessen und die Prävalenz in vielen Geweben sind Gründe, warum sie in einem therapeutischen Kontext gezielt genutzt werden. Um die modulierenden Fähigkeiten von GPCRs voll auszuschöpfen, ist eine vorherige Kenntnis ihres Wirkungsmechanismus auf molekularer Ebene erforderlich. Dies ist besonders wichtig für die erfolgreiche Anwendung rationaler strukturbasierter Arzneimitteldesign-Kampagnen. Sphingosin-1-Phosphat-Rezeptoren (S1PR) haben sich als wirksame Arzneimittelziele gegen Multiple Sklerose (MS) erwiesen, aber nicht-selektive Arzneimittel leiden unter schwerwiegenden Nebenwirkungen. Darüber hinaus birgt das Ansprechen einzelner S1PR ein großes Potenzial zur Modulation verschiedener Entzündungs- und Autoimmunerkrankungen.

In dieser Dissertation präsentieren wir *in silico* mechanistische Modelle zur Identifizierung wichtiger struktureller Determinanten, um die molekularen Grundlagen für die Selektivität der Sphingosin-1-phosphat-Rezeptor Familie und den allgemeinen Aktivierungsprozess von Klasse A GPCR zu beleuchten. Wir verwenden umfangreiche molekulardynamische Modelle von $S1PR_{1-5}$ und führen eine ganzheitliche vergleichende Protein-Ligand-Interaktionsanalyse der orthosterischen Bindetasche mit verschiedenen entdeckten Bindungsmodi eines Pan-agonisten durch, um das Problem der fast einheitlichen Bindetaschen zwischen den einzelnen Subtypen zu überwinden. So konnten wir das selektive Verhalten der zugelassenen Medikamente Ozanimod und Siponimod rational erklären, welche in der Multiple-Sklerose-Therapie eingesetzt werden. Darüber hinaus haben wir Einblicke in den Mechanismus der GPCR-Aktivierung der Klasse A gegeben und wie Änderungen der Torsionswinkel des Proteinrückgrats an diesem Prozess beteiligt sind. Durch die Entwicklung eines Arbeitsablaufs für die Datenextraktion mit maschinellem Lernen haben wir ein Vorhersagemodell für GPCR-Konformationen im aktiven und inaktiven Zustand erstellt und mögliche Anwendungen für die schnelle und effiziente Analyse von GPCR-Homologiemodellen und Molekular Dynamik Simulationen gezeigt. Die Methode zur

Anwendung der Vorhersagemodelle ist als Python-Paket auf GitHub (https://github.com/TrungNgocNguyen/GPCRml) verfügbar.

Die einzigartige und rationale Kombination aus modernster molekularer Modellierung, Datenextraktion und -analyse sowie maschinellen Lernmethoden zeigt, wie aussagekräftige In-silico-Modelle entwickelt werden können, um das Verständnis für herausfordernde Probleme auf dem Gebiet der GPCRs zu verbessern.

1. Introduction

1.1. G Protein-coupled Receptors

G protein-coupled receptors (GPCRs), the biggest family of membrane receptor proteins, play an important role in the mediation of cellular responses through, for instance, small molecules, lipids, and peptides [1]. Different ligands interacting with orthosteric or allosteric binding sites of the receptor can trigger various regulatory effects through effector protein recruitment [2]. On a physiological level GPCRs are responsible e.g. the regulation of sensory [3, 4], cardiovascular [5, 6], and neurotransmission [7, 8] processes. Dysregulated GPCRs contribute to pathophysiological processes, such as cardiovascular diseases, neurological, and metabolic disorders [9]. Due to their prevalence, expression in many tissues with different regulatory properties, and the physiological accessibility of drugs, GPCRs play a major role as targets in clinical practice [10, 11].

1.1.1. The Glutamate, Rhodopsin, Adhesion, Frizzled/Taste2, and Secretin Classification

Fredriksson et al. classified 802 unique human GPCR sequences into five different groups forming the GRAFS classification: glutamate, rhodopsin, adhesion, frizzled/taste2, and secretin (Figure 1) [12]. The classification is based on the phylogenetic distance and specific common structures of GPCRs. As an example, most of the rhodopsin family members have short N-termini not involved in orthosteric ligand binding, while the N-termini in other GPCR families play an important role as ligand-binding domains [13]. Furthermore, highly conserved amino acid motifs such as 'NSxxNPxxY', located in the transmembrane domain seven, and 'E/DRY', located between the transmembrane domain three and the intracellular loop (ICL) two, are found throughout the rhodopsin family [12]. The rhodopsin family (also called class A) represents the biggest cluster with 701 GPCRs of which 460 belong to the olfactory cluster.

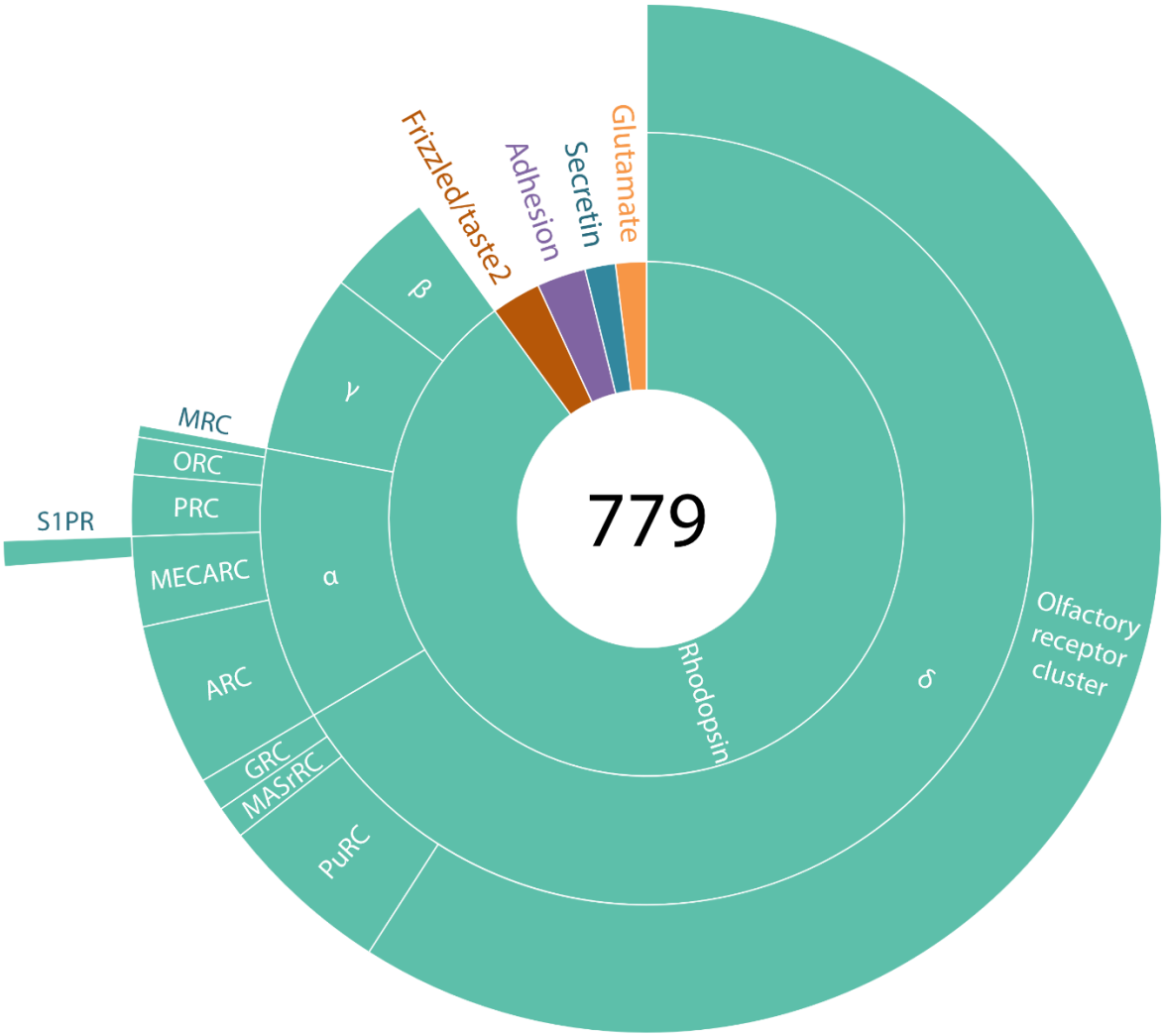


Figure 1. **Graphical representation of the subdivision and distribution of the GPCR data set with 779 GPCRs (23 could not be included in any family) from Fredriksson et al. ordered clockwise from biggest to smallest groups of the GRAFS classification.** The glutamate (15), adhesion (24), frizzled/taste2 (24), and secretin (15) families represent together only ten percent of the GPCRs categorized in the classification. The by far largest group, the rhodopsin family (701), is further divided into four main groups: α (89), β (35), γ (59), and δ (518), which are further subdivided into 13 main branches. ARC – amine receptor cluster, GRC – glycoproteins receptor cluster, MASrRC – MAS related receptor cluster, MECARC – Melanocortin, Endothelial differentiation, Cannabinoid, Adenosine binding receptor cluster, MRC – melatonin receptor cluster, ORC – opsin receptor cluster, PuRC – purine receptor cluster, PRC – prostaglandin receptor cluster

1.1.2. Structural Characteristics of G Protein-coupled Receptors

The advancement in the structural elucidation of GPCRs has been an immense leap forward for research of GPCR allosteric signal transduction [14], activation mechanisms [15], and computer-aided drug design/discovery [10]. The first GPCR crystal structure of the rhodopsin receptor was imaged in 2000 [16] (Figure 2). Based on the high expression rate in the retina, the covalently bound chromophore, and the absence of basal activity due to the physiological role as a light receptor, the rhodopsin receptor was the ideal candidate for the first crystallization attempts of a GPCR family member [17, 18]. The improvements in stabilizing other GPCRs via high-affinity ligands, proper detergents, thermostabilizing mutants, the employment of fusion proteins, and other techniques [19, 20] have led to the first successfully solved structure of a non-rhodopsin GPCR. The inactive state of the β2-adrenergic receptor was reported in 2007 [21] and the active state of the same receptor in 2011 [22]. Kobilka et al. overcame numerous challenges creating a blueprint and catalysator for the revolution of structural GPCR elucidation. Since then over 500 GPCR structures have been solved via X-ray crystallography and cryogenic electron microscopy (Cryo-EM) [23, 24], of which more than 400 are class A GPCR structures. This includes apo structures, complexes with different ligands, intracellular binding partners (IBPs), and ternary complexes. The research community gained structural insights into the range of possible conformations and their dependency on each factor on a molecular level [25].

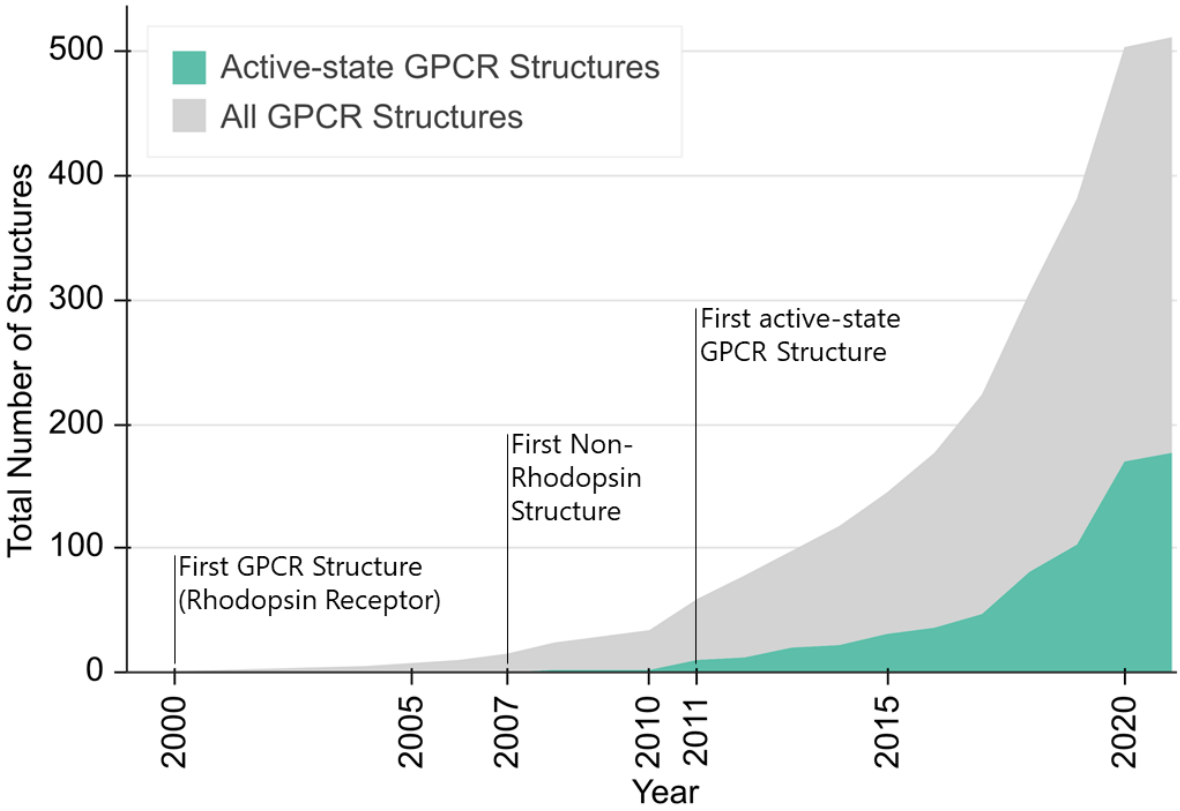


Figure 2. **The total number of GPCR structures available per year.** The revolution in structural G protein-coupled receptors elucidation has led to an explosive growth of structural data. Due to the more challenging nature to stabilize the ternary complex, there is less structural data available for the active-state GPCR structures. Data from the GPCRdb [26]

GPCRs are generally characterized by seven α-helical transmembrane domains (TM1 – 7) located in the phospholipid bilayer of the cells (Figure 3). These α-helices are the most homologous structures of GPCRs and consist of mostly hydrophobic amino acids. Three extracellular loops (ECL1 – 3) and the N-terminus are located outside of the cell (Figure 4). Intracellularly there are also three intracellular loops (ICL1 – 3) and the C-terminus. The ECLs are the most variable parts of GPCRs.

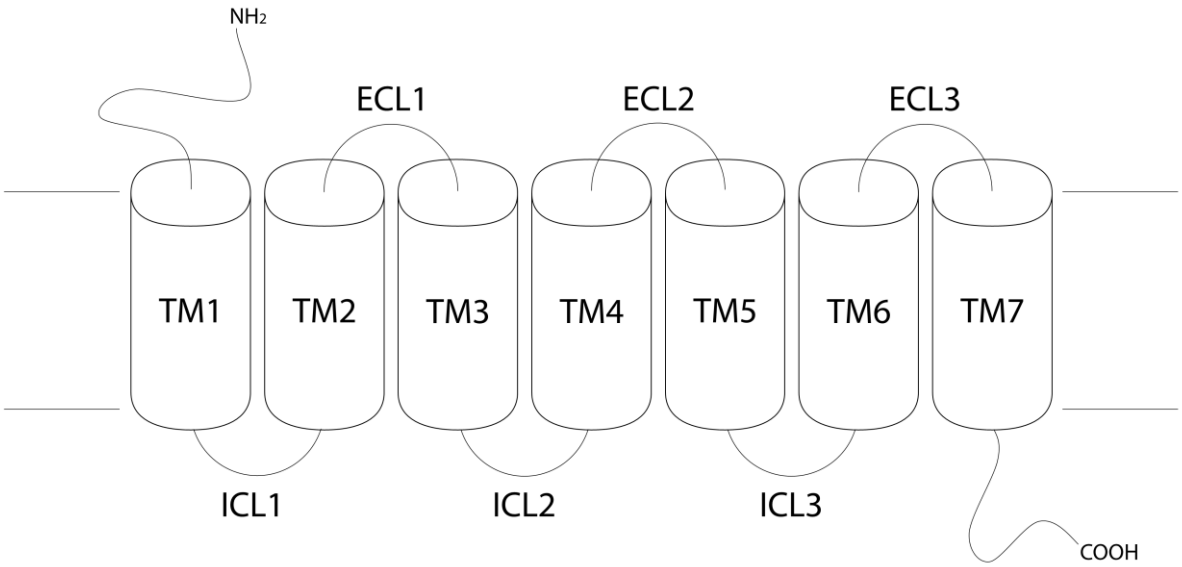


Figure 3. **Two-dimensional topography of GPCRs.** TM – transmembrane domains, ECL – extracellular loops, ICL – intracellular loops

Ballesteros and Weinstein have developed a numbering scheme for the transmembrane domains of class A GPCRs by taking the most conserved residues of all TMs as the central position [27]. The most conserved amino acid in TM1 is numbered $N^{1.50}$, where the letter indicates the amino acid, the first number the transmembrane domain 1, and the number 50 the amino acid position. 50 is used to indicate the most conserved residue in the TM of class A GPCRs (Figure 4). Every other amino acid is numbered accordingly with lower/higher numbers diverging from position 50. The Ballesteros-Weinstein numbers (BWN) will be used throughout this work and be referred to as the Ballesteros Weinstein number.

The short length of ECL1 is highly conserved in the rhodopsin family. This was shown by Peeters et al, taking the two most conserved amino acids in TM2 ($BWN^{2.50}$) and TM3 ($BWN^{3.50}$) as the starting and ending points for measurement. The average number of amino acids between $BWN^{2.50}$ and $BWN^{3.50}$ in the rhodopsin family is 52±2 amino acids. [28] With a distance of 50 amino acids between $BWN^{2.50}$ and $BWN^{3.50}$, the S1PR group lies within the average length. Furthermore, reports postulate that ECL1 has an impact on the shape of the binding site [29, 30].

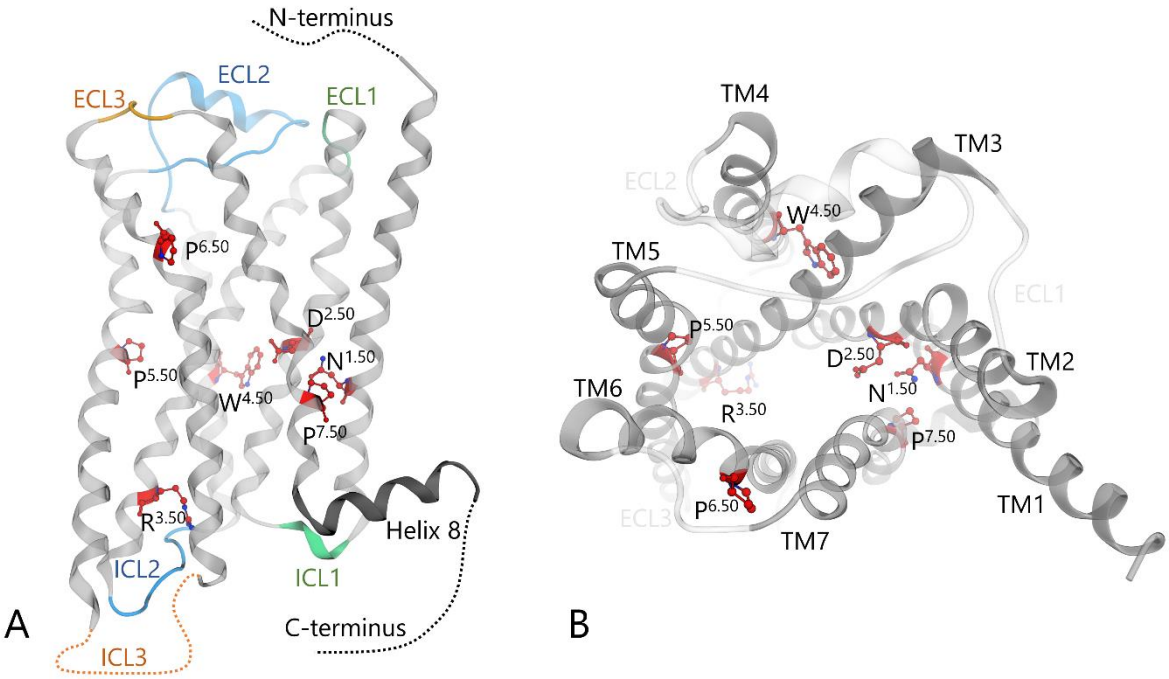


Figure 4. **Three-dimensional depiction of GPCR domains based on the structure of $S1PR_2$.** Transmembrane (A) and extracellular view (B) of the $S1PR_2$ show tertiary structure based on the built homology model in this thesis (chapter 4.1.1). Most conserved residues are highlighted in red and labeled with the Ballesteros-Weinstein numbering scheme. ECL – extracellular loop, TM – transmembrane domain

ECL2 shows the highest diversity in length and sequence of the rhodopsin family [28]. As a result, ECL2 plays an important role in ligand selectivity. Furthermore, binding site shape, ligand binding specificity, allosteric modulation, and biased agonism can be affected by ECL2 [31]. Concerning $S1PR_1$ crystal structure (PDB 3V2Y [32]) the ligand-binding site is swayed by the hydrophobic residues of ECL2 [32]. Lastly, the third extracellular loop (ECL3) is also short in length similar to the ECL1 [28]. Despite the short length, ECL3 can influence receptor functions in terms of activation and allosteric modulation [33–35]. The crystal structure of $S1PR_1$, reveals no ECL3 involvement in binding site molding [32].

Due to the evolutionary demand for GPCRs to recognize a vast number of distinct extracellular stimuli [36], the extracellular domains show the highest sequence diversity [37]. This variety is opposed by the more homologous intracellular regions with only dozens of interacting partners, which transfer incoming extracellular signals [38].

1.1.3. Dihedral Angles of Protein Backbones

Dihedral angles of protein backbones are internal angles, also called torsion angles, of aminoacids consisting of three angles per residue called Φ (phi), ψ (psi), and ω (omega) angles. The phi angle is calculated based on the backbone atoms C'-N-Cα-C and the Φ angle is based on N-Cα-C-N' (Figure 5). Due to the planar nature of the amide (peptide bond), the ω angle is restricted to 180° (trans) or 0° (cis). Ramachandran et al. developed a plot for the visualization of the Φ and ψ angles of proteins [39]. Secondary protein structures are defined in allowed regions around -60° Φ and -50° ψ for right-handed α-helices, 60° Φ and ψ for left-handed α-helices, and -120° Φ and 120° ψ for β-sheets. One exception is proline, which has a fixed Φ angle of approximately -65° due to its cyclic side chain structure [40]. The Ramachandran plot is used for the evaluation of protein structure geometry.

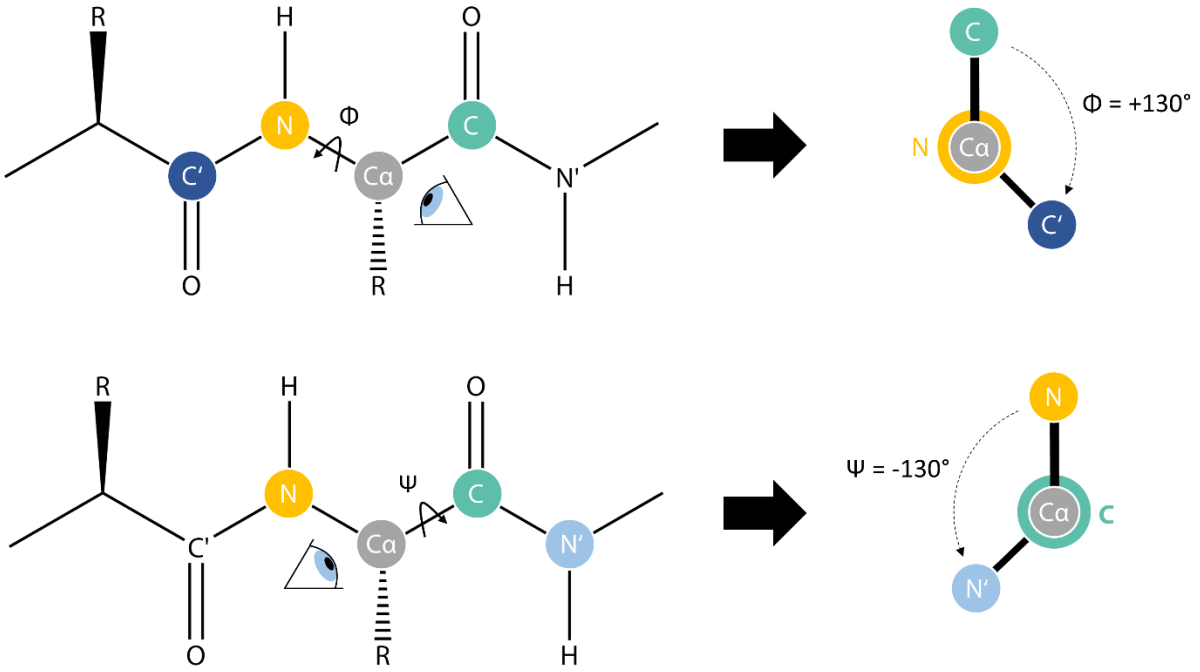


Figure 5. **Protein backbone dihedral angles Φ (phi, top) and ψ (psi, bottom).** Dihedral angles of the protein backbone show specific angles if secondary structures are formed (α-helices or β-sheets).

1.1.4. Molecular Mechanisms of Class A G Protein-coupled Receptor Activation

The concept of GPCR activation has evolved into a complex theorem composed of different receptor pharmacology mechanisms, such as ligand influence [41–43], basal activity [44–46], allosteric coupling mechanisms [14, 42, 47], conserved motifs [48, 43], and more [49]. We are now able to draw a more holistic picture of the evolutionary fine-tuned cell signal reception and translation of this highly dynamic receptor group [25].

In general, GPCR activation describes the process of ligand-induced recruitment of intracellular binding partners. Advances in structural GPCR elucidation have revealed a common class A GPCR activation mechanism. [50, 37] Comparison of structural pairs of active and inactive state GPCRs, e.g. of the β2-adrenergic receptor [21, 22], muscarinic M_2 receptor [51, 52], or adenosine A_{2A} receptor [53, 54], has unveiled common structural rearrangements (Figure 6).

Upon agonist binding to the orthosteric binding site, the extracellular parts of TM5, TM6, and TM7 are moving inward to the longitudinal receptor axis. This subtle contraction of the extracellular domain affects the intracellular binding domain leading to the outward movement of intracellular parts of TM2, TM4, TM5, and TM6 as well as the inward movement of TM7 ultimately enables intracellular binding partner (IBP) recruitment, such as G proteins, β-arrestins, and GPCR kinases, and GPCR stabilization in the active state.

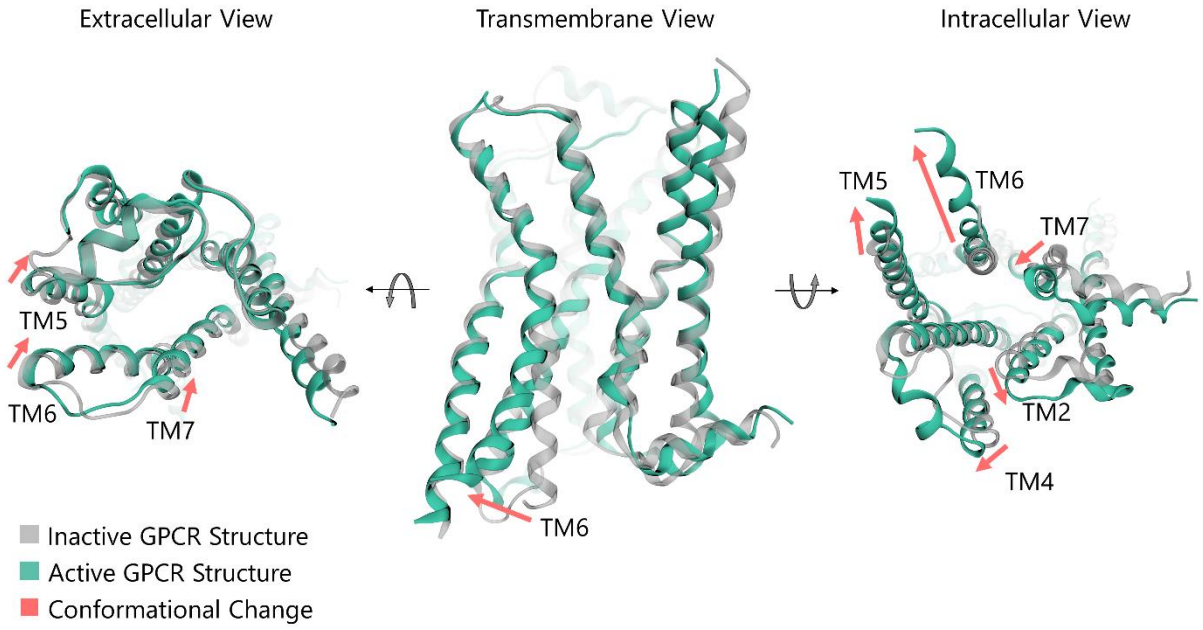


Figure 6. **Conformational changes between the active (PDB 3SN6 [22]) and inactive (PDB 2RH1 [21]) β2-adrenergic GPCR crystal structure.** The β2-adrenergic receptor is shown as a model system to demonstrate conformational changes during activation of class A GPCRs. Figure adapted from [42].

Different mechanisms on an atomistic level are responsible for class A GPCR activation. The first conformational changes after agonist binding are observable under the binding pocket where the conserved $P^{5.50}$–$I^{3.40}$–$F^{6.44}$ and $C^{6.47}$–$W^{6.48}$–x–$P^{6.50}$ motifs are located (Figure 7). The consecutive conformational change of the toggle-switch $W^{6.48}$ plays an important role in the characteristic intracellular outward movement of TM6. Another important microswitch is the ionic lock, or salt bridge, located at the D(E)$^{3.49}$–$R^{3.50}$–$Y^{3.51}$ motif, where $R^{3.50}$ is forming an intrahelical salt bridge with $E^{6.30}$ [42]. This salt bridge stabilizes the inactive state and breaks during activation, which allows subsequent translocation of TM6'. Furthermore, the D(E)$^{3.49}$–$R^{3.50}$–$Y^{3.51}$ motif is directly involved in G protein-binding. Finally, the conserved $N^{7.49}$–$P^{7.50}$–x–x–$Y^{7.53}$ motif's rearrangement stabilizes the active GPCR state and is forming the G protein-binding site.

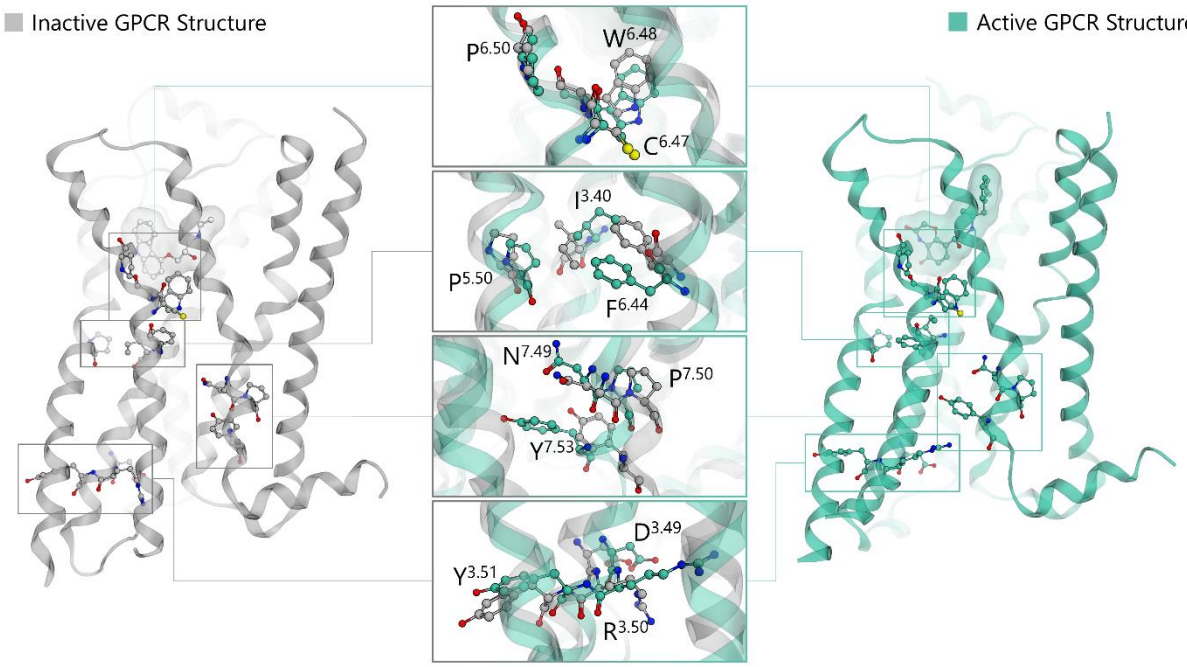


Figure 7. **Comparison of active and inactive β2-adrenergic receptor crystal structures.** Left. Inactive structure (PDB 2RH1 [21]). Right. Active structure (PDB 3SN6 [22]). Middle. From top to bottom: Motif CWxP, PIF, NPxxY, E/DRY

The transfer of extracellular ligand-binding information via the transmembrane domain to the intracellular transducer binding domain is called allosteric coupling. The binding of IBPs, such as G proteins, triggers a reciprocal effect on the ligand-binding domain enhancing ligand affinity to the orthosteric binding pocket through further tightening the extracellular part caging the ligand inside as long as the ternary complex is formed [14]. Due to the highly dynamic properties of GPCRs IBP recruitment can also occur without prior ligand binding. This phenomenon is also known as basal or constitutive activity. Interestingly, the formation of the receptor IBP complex leads to ligand affinity decrease until the complex dissociates [14].

1.1.5. Ligand Influence on G Protein-coupled Receptors

Orthosteric ligand binding can influence, based on the mode of action of the ligand, the receptors' probability to recruit IBPs through ligand-dependent stabilization of distinct

receptor conformations [41]. Full agonists trigger the maximum signaling capacity at a given receptor and pathway, whereas partial agonists elicit activity below that level. Antagonists only bind to the orthosteric binding site but do not affect the basal activity and equilibrium of receptor conformations. Lastly, inverse agonists inhibit basal activity. [55]

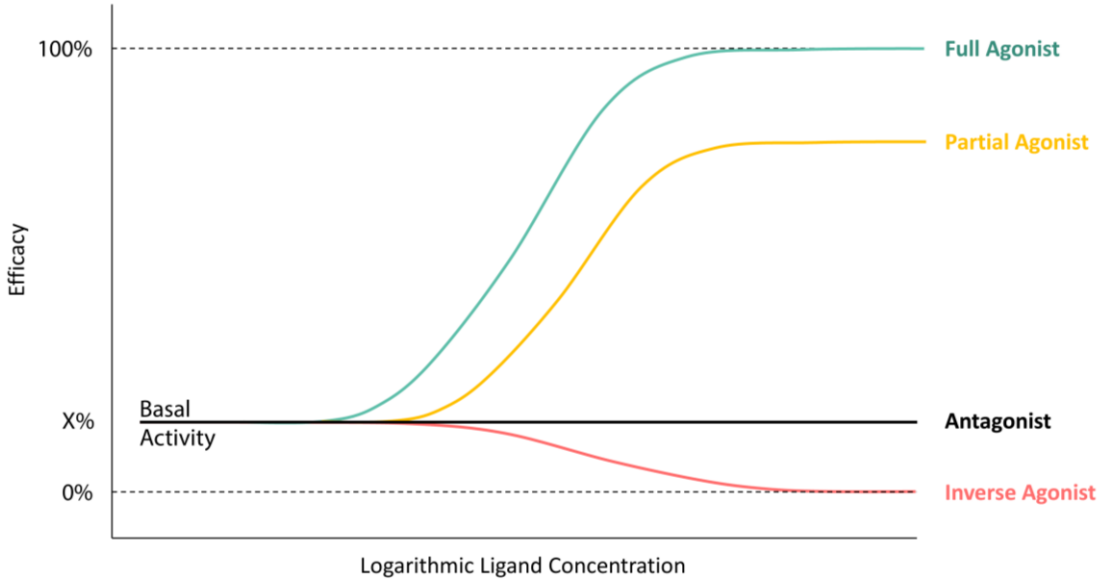


Figure 8. **Ligand influence on receptor signaling.**

Ligands can also bind, apart from the orthosteric site, at distinct allosteric sites [56]. They allosterically enhance or decrease orthosteric ligand binding and are categorized as positive or negative allosteric modulators, respectively. Furthermore, allosteric modulators can also elicit signaling without orthosteric ligands. [57]

1.1.6. G Protein-coupled Receptor Signaling

Upon GPCR activation, various downstream signaling pathways are initiated (Figure 9). The most classical interactions occur with heterotrimeric G proteins. After guanosine diphosphate (GDP) is exchanged with guanosine triphosphate (GTP) on the Gα subunit, the G protein dissociates into the Gα and Gβγ subunits. Gα-GTP activated subunits regulate, via four subunit types ($G\alpha_s$, $G\alpha_i$, $G\alpha_q$, and $G\alpha_{12/13}$), different downstream effectors such as

adenylyl cyclase, Rho GTPase, and phospholipase Cβ, which in turn modulate additional downstream effectors and/or second messengers [58]. Gβγ subunits bind directly to and activate G protein-coupled inwardly-rectifying potassium (GIRK) channels. Other important signaling pathways are conducted through β-arrestins [59]. They initiate internalization and hinder further activation of GPCRs.

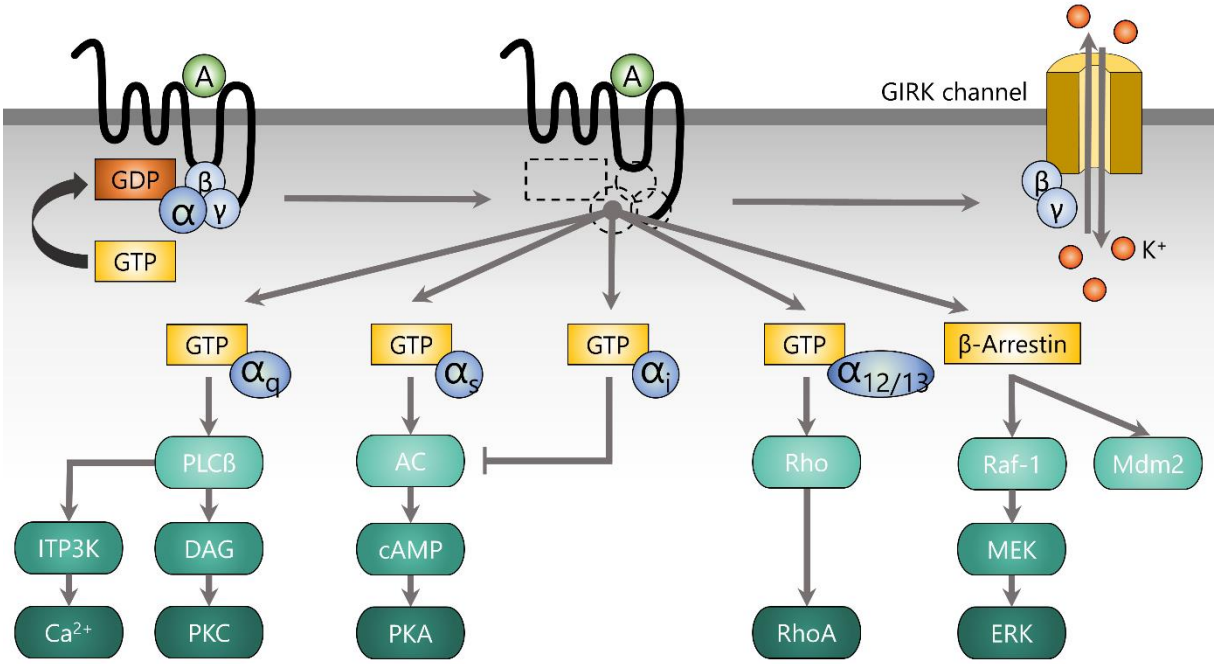


Figure 9. **General signaling of GPCRs.** Different Gα subunits activate different downstream signaling cascades. Abbreviations: A – agonist, AC – adenylyl cyclase, cAMP – cyclic adenosine monophosphate, DAG – diacylglycerol, ERK – extracellular signal-regulated kinase, GDP – guanosine diphosphate, GIRK channel – G protein-coupled inwardly-rectifying potassium channel, GTP – guanosine triphosphate, K^+ – potassium ion, ITP3K – inositol-trisphosphate 3-kinase, MEK – mitogen-activated protein kinase (also known as MAP2K), Mdm2 – mouse double minute 2 homolog, PLCβ – phospholipase Cβ, PKA – protein kinase A, PKC – protein kinase C, Raf-1 – RAF proto-oncogene serine/threonine-protein kinase, RhoA – Ras homolog family member A. Figure adapted from [60].

1.1.7. G Protein-coupled Receptors as Druggable Targets

GPCRs play a major role as targets in clinical medicine. This is due to their prevalence, expression in many tissues with different regulatory properties, and physiological accessibility for drugs [61].

An analysis by Rask-Andersen et al. in 2011 identified 989 drugs from the Drugbank database [62] acting on 435 human targets. Drugs with unknown targets and drugs with non-human targets and non-therapeutic targets were not considered. Only 19% (82) of the identified targets were GPCRs, while 36% (357) of the drugs target GPCRs [63, 64]. A more recent evaluation by Sriram et al. in 2017 combined different public databases (CHEMBL [65], IUPHAR [11], and DRUGBANK [62]) and found ~700 drugs approved in the United States and the European Union (35% of all marketed drugs) targeting 134 GPCRs [66]. (Figure 10) Even though more than one-third of all drugs are acting on GPCRs, the significant amount of untargeted GPCRs provide considerable potential for future clinical use.

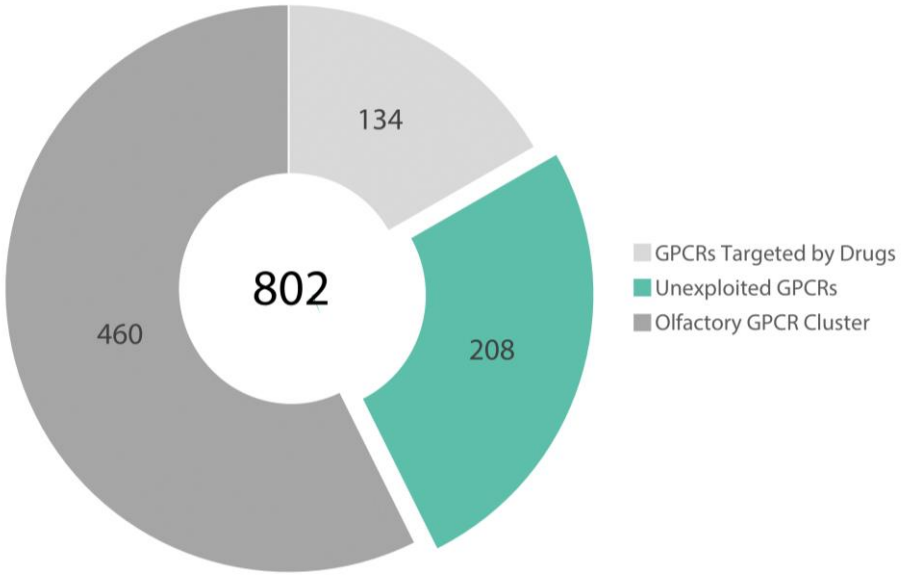


Figure 10. **Untargeted GPCRs carry the potential for future clinical use.** Of 802 GPCRs 460 are categorized as olfactory GPCRs with unknown therapeutic relevance. The unexploited GPCRs (208) are mostly orphan receptors that could become clinically relevant over time.

Marketed drugs targeting GPCRs have proven their importance in many therapeutic fields, such as obstructive lung diseases, inflammatory diseases, cardiovascular diseases, central nervous system disorders, autoimmune system diseases, and more [67, 68]. This is also reflected in the commercial impact calculated from 75 countries between 2011 – 2015. Drugs targeting GPCRs have the highest total market share of 27.42% (889.17 billion USD aggerated sales) compared to other target classes. Furthermore, 7 of the top 20 targets are GPCRs in terms of total drug sales from 2011 to 2015. [69] The best performing drug target is angiotensin II receptor type 1 (AT_1 receptor) with total revenue of 99.98 billion US$.

1.2. Sphingosine-1-phosphate Receptors

The S1PRs belong to the lysophospholipid receptor family, with the phospholipid sphingosine-1-phosphate (S1P) as their endogenous ligand (Figure 13). The receptor group, also called the endothelial differentiation gene (EDG) family, consists of five subtypes, $S1PR_1$ – $S1PR_5$ [70].

The S1PRs belong to the MECA (melanocortin, endothelial differentiation, cannabinoid, adenosine binding) receptor cluster (RC) of the subcategory α of the rhodopsin group. Receptors in this cluster bind structurally different ligands despite their short phylogenetic distance. All S1PRs inherit the motif 'NSxxNPxxY', but only $S1PR_1$ inherits the 'ERY' motif. $S1PR_2$ and $S1PR_3$ bear the motif 'ERH', $S1PR_4$ 'ERF', and $S1PR_5$ 'ERS' instead.

S1PRs couple with $G\alpha_i$, $G\alpha_q$ and $G\alpha_{12/13}$. Most of the subtypes do not couple exclusively with one G-protein subunit but have preferences such as $S1PR_1$ for $G\alpha_i$, $S1PR_2$ for $G\alpha_{12/13,}$ and $S1PR_3$ for $G\alpha_q$ [71–74]. Due to different coupling combinations (mainly discriminated by Gα subunits and tissue expression rate) different downstream signaling mechanisms and therefore different biological effects are regulated by S1PRs (Figure 11).

1.2.1. Physiological Roles of Sphingosine-1-phosphate Receptors

S1PRs fulfill many different regulatory functions in various organ systems with the largest impact on the regulation of the cardiovascular, immune, and nervous systems (Figure 11)

[75]. The maintenance of cardiovascular homeostasis is achieved through attuning the basal permeability of vascular barriers, vascular tone, and lymphocyte trafficking [76].

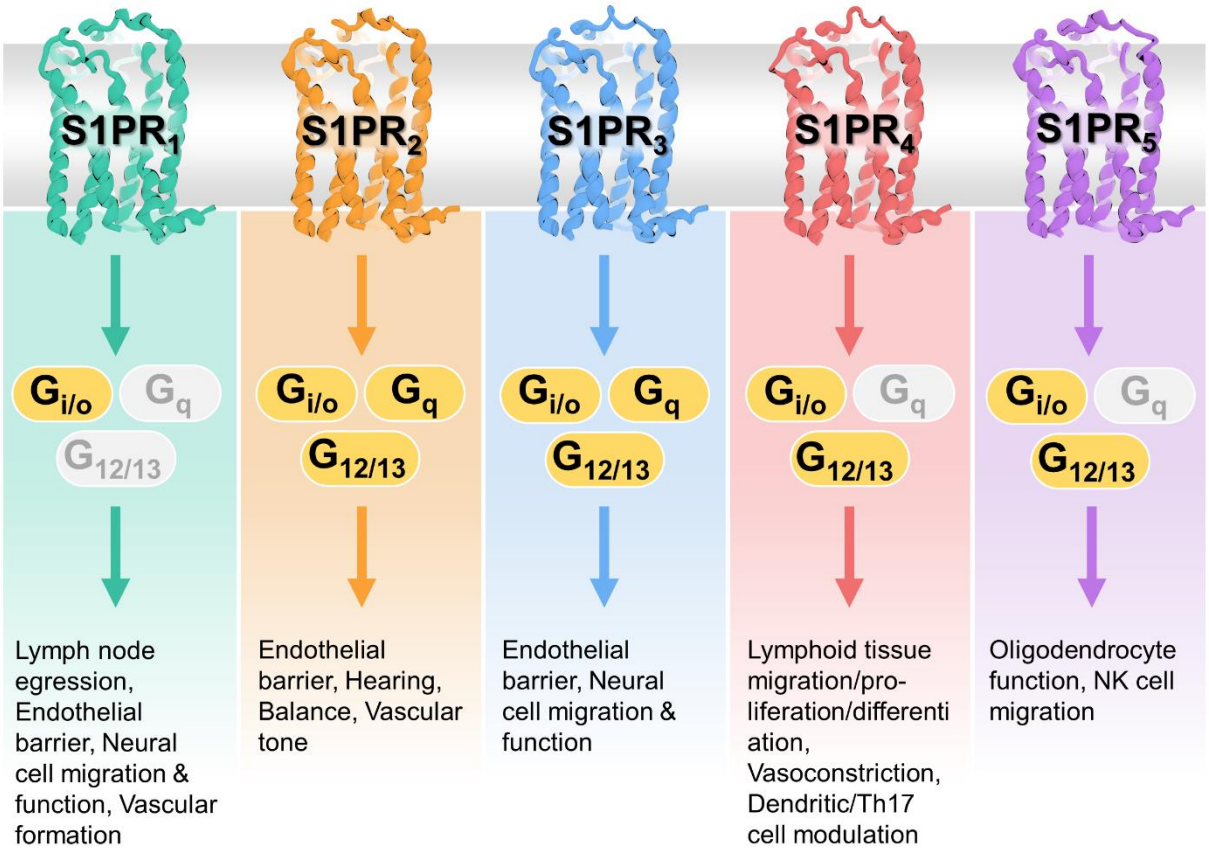


Figure 11. **Sphingosine-1-phosphate receptor coupling and physiological functions.**

$S1PR_1$ regulates important domains of the immune system, promoting the migration of lymphocytes from secondary lymphoid organs into systemic circulation as one of the best-characterized functions [77]. Furthermore, $S1PR_1$ has an impact on the polarization of T cells to Th17 cells [78]. Findings by Li et al. revealed the pro-tumorigenic character of $S1PR_1$ related to nephroblastoma. Upon activation, $S1PR_1$ promotes tumor cell migration and invasion [79]. The formation of endothelial cells by the vascular barrier system is affected by $S1PR_1$ signaling, controlled through the tightening of adherens junctions. Jung et al. showed that this is not only dependent on ligand activation of $S1PR_1$ but also fluid shear stress [80]. Moreover, $S1PR_1$ regulates vascular relaxation and blood pressure homeostasis [81]. Studies

on knockout mice demonstrated that $S1PR_1$ is required for normal embryonic vascular development, especially of large vessels, leading to death by hemorrhage in absence of $S1PR_1$ [82]. $S1PR_1$ signaling is also discussed to target osteoclastogenesis and osteogenesis and therefore play a complex role in the bone remodeling process [83].

$S1PR_2$ antagonizes many functions of $S1PR_1$, which leads to increased paracellular permeability of the vascular systems, due to disruption of adherens junctions [84]. In addition, $S1PR_2$´s antiproliferative and anti-migrative activity inhibits angiogenic sprouting [85] and lessens arterial smooth muscle cell formation [86]. Interestingly, the anti-migratory effect also inhibits the development of distant cancer metastasis cells in mice and nephroblastoma's in vitro [87, 79]. In terms of the immune system, $S1PR_2$ has a negative repercussion on the chemotaxis and migration speed of macrophages [88] but seems to influence phagocytosis positively [89]. Furthermore, hearing impairments [90] and seizures, due to the increased excitability of neocortical pyramidal neurons [91], were observed in knockout mice. $S1PR_2$ knockout mice also showed in lysolecithin-induced demyelination increased remyelinated axons. A similar outcome could have also been achieved by inhibiting $S1PR_2$ with JTE-013 [92]. It should be noted though that JTE-013 is also an antagonist of $S1PR_4$ and several sphingolipid metabolic enzymes, which might have a significant influence through these off-target effects [93]. Other physiological roles of $S1PR_2$ are restraining tissue-resident lymphocyte's egress [94] and regulation of intestinal barrier function [95].

Subtypes three to five of S1PR, unlike $S1PR_1$ and $S1PR_2$, have been studied in a limited fashion. $S1PR_3$ physiological roles involve the maturation of dendritic cells in the immune system [96], recruitment of macrophages during inflammation [97], and controlling the permeability of vessels [98]. $S1PR_3^{-/-}$ knockout mice did not show any phenotypical abnormalities, whereas $S1PR_2^{-/-}$ and $S1PR_3^{-/-}$ knockout mice died during gestation or infancy [99]. $S1PR_{1-3}$ are expressed ubiquitously in the body in contrast to $S1PR_4$ and $S1PR_5$. $S1PR_4$ is only expressed in lung, lymphoid, and hematopoietic tissues [100], while $S1PR_5$ is specifically expressed in natural killer cells [101], dendritic cells [102], the central nervous system (CNS) [103], and endothelial cells [104]. This tissue-specific expression of $S1PR_4$ and $S1PR_5$ suggests specialized functions. $S1PR_4$ deficient mice showed aberrant megakaryocytes and delayed platelet recovery after a deletion with antibodies, but did not show abnormal phenotype

changes during embryonic development otherwise [105]. Moreover, $S1PR_4$ is involved in dendritic cell migration and inflammation by T cell function modulation via activation of myeloid cells [106]. $S1PR_5$ regulates the integrity of the blood-brain barrier, increasing tight junctions and therefore decreasing the permeability [107]. Moreover, natural killer cell migration [108] and circulation of monocytes [109] are dependent on $S1PR_5$.

The ubiquitous expression and involvement in many regulatory functions highlight the importance of S1PRs in the human body.

1.2.2. $S1PR_1$ Crystal Structure

The crystal structure of $S1PR_1$ (PDB 3V2Y [32]) is used in this thesis. In 2012 Hanson et al. fused the receptor to a T4-lysozyme in the intracellular part of TM5 and TM6 and solved this fusion protein in complex with the selective $S1PR_1$ antagonist ML056 ((R)-3-amino-(3-hexylphenylamino)-4-oxobutylphosphonic acid, Figure 12) [110]. The mean resolution of 2.8Å shows the $S1PR_1$ structure on an atomic level. The detailed view of the electron density map reveals a general good resolution of transmembrane domains and the ligand. Unsolved domains include part of the N-terminal domain, single residues in the extracellular domain (e.g. K^{34}, $R^{7.34}$, $E^{7.36}$), and the last intracellular residues of each TM. This indicates the high flexibility of intra- and extracellular domains as opposed to the stable TM cores of GPCRs. Interestingly, $R^{7.34}$, $E^{7.36}$ residues near the binding pocket entrance between TM1 and TM7, are discussed to act as amino acids attracting ligands into the binding pocket [111]. This could explain why residue $R^{7.34}$ and $E^{7.36}$ are not solved as well as their absence in ligand binding and binding pocket shaping.

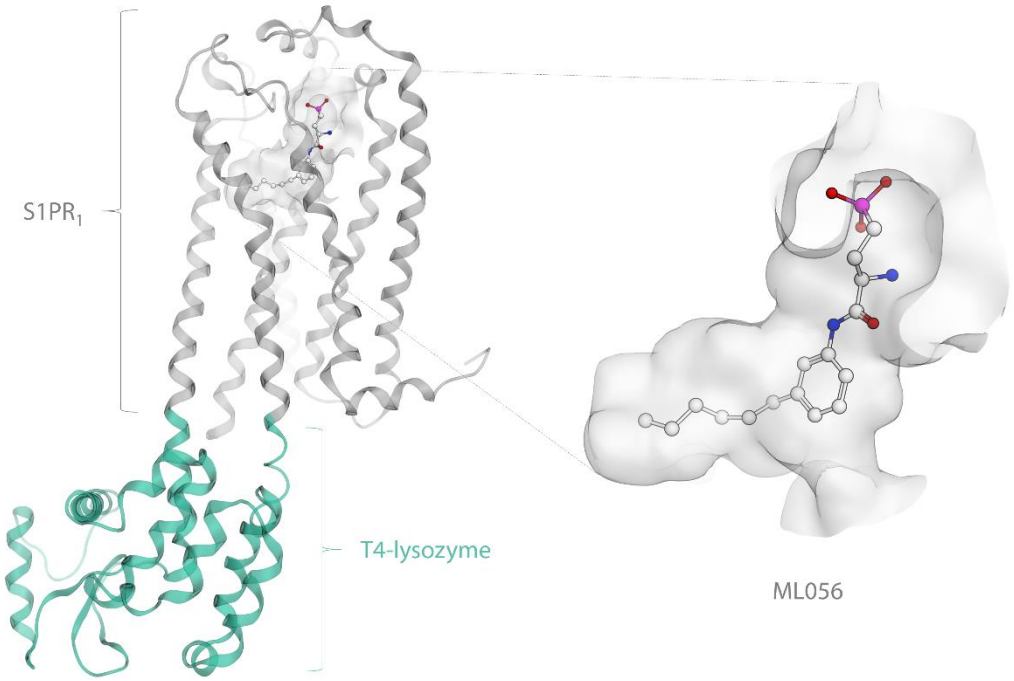


Figure 12. **Crystal structure of $S1PR_1$ (PDB 3V2Y).** The crystal structure of $S1PR_1$ is composed of a T4-lysozyme fused intracellularly to TM5 and TM6 of $S1PR_1$ protein, replacing the ICL3. Furthermore, ML056, a selective sphingolipid-mimic $S1PR_1$ antagonist, is used to stabilize the inactive conformation.

1.2.3. Sphingosine-1-phosphate Receptors as Drug Targets

Targeting S1PRs to modulate the regulatory function of lymphocyte trafficking has shown to be an effective/promising treatment for different autoimmune diseases [112]. The functional antagonist Fingolimod (2-amino-2[2-(4-octylphenyl)ethyl]-1,3-propanediol, FTY720, **1,** Figure 13) is the first-in-class orally bioavailable drug against remitting relapsing multiple sclerosis (RRMS) [113] and was approved 2010 in the USA and 2011 in the EU. It is a structural analog derived from myriocin, a non-proteinogenic immunosuppressive amino acid found in the fungus *Isaria sinclairii*, with improved immunosuppressive activity and reduced toxicity *[114].* Fingolimod is activated *in vivo* via sphingosine kinase 2 (SphK2) to Fingolimod-phosphate (FTY720-P, **2**) [115]. **2** downregulates $S1PR_1$, due to its super-agonist bioactivity, creating an $S1PR_1$-null state in primary and secondary lymphoid organs leading to lymphocyte sequestration. Furthermore, **2** also binds to $S1PR_{3-5}$ [116–118]. The *in vivo* efficacy and safety of Fingolimod (daily oral dose of 1.25 or 0.5 mg) for RRMS treatment has

been assessed in the 24-months FREEDOMS [119], FREEDOMS II [120], and 12-months TRANSFORMS [121] double-blind, randomized clinical studies (RCT). Fingolimod showed a significant decrease in annual relapse rates and disease activity in comparison with either placebo or interferon β-1a. The long-term benefits of Fingolimod treatment in FREEDOMS were sustained during the 2-4 years long FREEDOMS extension. Even though long-term treatment was reported to be well-tolerated [122], serious adverse events (SAE) like neoplasms, hepatobiliary disorders, CNS disorders, and infections/infestations occurred in around 10% of the patients. Commonly reported adverse events (AE) include nasopharyngitis, upper respiratory tract infection, lymphopenia, and headache. Additional safety concerns evolve around Fingolimod's cardiac toxicity covering inter alia transient bradycardia and first-degree atrioventricular block. These adverse events occurred in both FREEDOMS as well as TRANSFORMS trials during the first dose and might also be dose-dependent. Animal studies suggest that $S1PR_3$ is mainly involved in transient heart rate reduction by Fingolimod [123, 124]. It is also hypothesized, that $S1PR_3$ agonism is the cause of the adverse effect of macular edema through a breakdown of the inner blood-retinal barrier [125]. Dusaban et al. uncovered the role of $S1PR_3$ in CNS injury and disease (e.g., multiple sclerosis). S1P and Fingolimod activation of $S1PR_3$ induces inflammatory cytokines (e.g. IL-6) and cytotoxic mediators (e.g. COX-2, VEGFa) in astrocytes, which contributes to neuronal cell death, demyelination, and increased permeability of the blood-brain barrier leading to S1P and pathogenic lymphocyte leakage into the CNS [125]. This could have a negative impact on the progression of CNS inflammatory diseases. To circumvent adverse effects and detrimental mechanisms linked to the non-selective binding of Fingolimod, great effort has been put into developing selective S1PR drugs with better safety profiles and comparable efficacy.

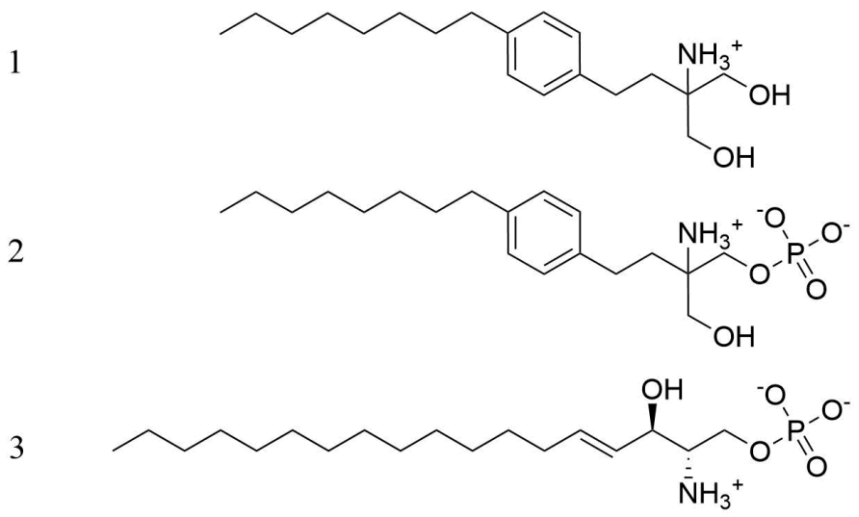


Figure 13. **(1) Fingolimod, (2) Fingolimod-phosphate and (3) sphingosine-1-phosphate.** The endogenous ligand S1P has an amphiphilic structure with a phosphate head group as the hydrophilic part and a long alkyl chain as the lipophilic part. The first-in-class drug Fingolimod inherits analog structural features.

Siponimod (BAF312, (E)-1-(4-[1-({[4-cyclohexyl-3-(trifluoromethyl)benzyl]oxy}imino)ethyl]-2-ethylbenzyl)azetidine-3-carboxylic acid, Figure 13, **4**) was recently approved by the FDA in 2019 and European commission in 2020 for secondary progressive MS (SPMS). Siponimod was discovered via *de novo* design based on Fingolimod. The initial structure-activity relationship study was focused on achieving selectivity against $S1PR_3$ by introducing rigidity into the lipophilic part of the ligand. Furthermore, the *in vivo* elimination half-life was reduced by replacing the phosphate group with carboxylic acid [126]. These successful design strategies have led to Siponimod's $S1PR_1$ and $S1PR_5$ selectivity (Table 1), more favorable pharmacokinetic properties (t_{max} 3-4.5 h, $t_{1/2}$ 30 h) compared to Fingolimod (t_{max} 12-16 h, $t_{1/2}$ 144-216 h) [127] as well as being a non-pro drug. The 6 months long clinical study BOLD showed a significant decrease in combined unique active brain magnetic resonance image lesions for RRMS in comparison with placebo for all Siponimod doses (0,5 mg – 10 mg) [128]. Long-term studies with Siponimod revealed a significant reduction in annual relapse time as well as a relative risk reduction of physical and cognitive impairment development for SPMS compared to placebo [129, 130]. Similar adverse events, like macular oedema and bradycardia, have been observed in comparison to Fingolimod. Initial bradycardia could be mitigated by dose titration. A recent meta-analysis from 2021

compared the relative risk of adverse events in RCTs between different S1PR modulators and the comparator [131]. Siponimod did not show a significant change in relative risk (RR) for occurring adverse events of 1.09 (1.04-1.13 95% CI) compared to Fingolimod (1.05, 1.00-1.10 95% CI). Surprisingly, RR for SAEs is higher for Siponimod (2.07, 0.33-13.05 with 95% CI) vs. Fingolimod (1.21, 0.96-1.53 with 95% CI). Siponimod showed reduced RR for bradycardia (1.88, 1.09-3.22 95% CI vs. 4.78, 2.30-9.93 with 95% CI for Fingolimod) but increased RR for macular oedema (8.93, 1.20-66.75 95% CI vs 1.87, 0.85-4.13 95% CI for Fingolimod). Due to missing clinical studies directly comparing Siponimod to Fingolimod treatment and the scarce study availability for Siponimod, the presented data might not be sufficient for conclusions about the safety profile of Siponimod vs. Fingolimod.

Ozanimod (RPC1063, (S)-5-(3-{1-[(2-hydroxyethyl) amino]-2,3-dihydro-1H-inden4-yl}-1,2,4-oxadiazol-5-yl)-2-(1-methylethyloxy)benzonitrile, Figure 13, **5**) was also approved 2020 by the FDA and European Commission for RRMS treatment. It was initially developed from a hit found by the Scripps Research Institutes [132]. Ozanimod is a selective $S1PR_1$ and $S1PR_5$ agonist (Table 1) with ideal pharmacokinetic properties (t_{max} 6-8 h, $t_{1/2}$ 19 h) for oral daily dose application and fast recovery time (2-3 d) of lymphocyte count after discontinuation (Siponimod 1-5 d, Fingolimod 30-60 d) [133, 134]. The *in vivo* efficacy has been demonstrated in several studies [135–138], revealing significantly fewer cumulative lesions and annual relapse rates vs. either placebo or interferon β-1a. Most common AEs include nasopharyngitis, headache, and urinary tract infections. Interestingly no clinically relevant cardiac AEs or macular edema was observed. The meta-analysis from Lasa et al. also included RCT data of Ozanimod [131]. Comparing the relative risk of Fingolimod to Ozanimod, Ozanimod shows reduced RR for occurring AEs (0.86, 0.77-0.96 95% CI vs. 1.05, 1.00-1.10 95% CI for Fingolimod), SAEs (1.00, 0.59-1.70 95% CI vs. 1.21, 0.96-1.53 95% CI for Fingolimod), bradycardia (2.88, 0.50-16.58 95% CI vs. 4.78, 2.30-9.93 95% CI for Fingolimod) and macular oedema (0.67, 0.15-3.05 95% CI vs. 1.87, 0.85-4.13 95% CI for Fingolimod). Overall Ozanimod seems to have a better safety profile with the same efficacy compared to Fingolimod for RRMS treatment.

There are many other drug candidates in the pipeline, like amiselimod [139], ponesimod [140], and cenerimod [141], for the treatment of RRMS and other autoimmune diseases [142].

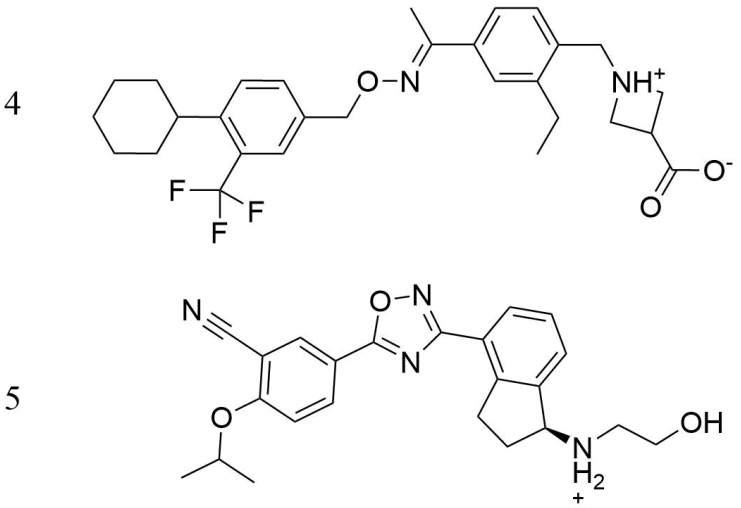


Figure 14. **(4) Siponimod and (5) Ozanimod.** Siponimod and Ozanimod are $S1PR_1$ and $S1PR_5$ selective drugs. They have been successfully marketed in 2019 in the European Union and the United States.

Table 1. **In vitro activity for marketed drugs targeting S1PRs.** Data from Scott et al. [132]

		$S1PR_1$ GTPγS	$S1PR_2$ GTPγS	$S1PR_3$ GTPγS	$S1PR_4$ ß-arrestin	$S1PR_5$ GTPγS
FTY720-P	EC_{50} (nM)	0.27	>10 000	0.90	345	0.50
	E_{max}[a] (nM)	85	22	57	130	61
Siponimod	EC_{50} (nM)	0.39	>10 000	>10 000	920	0.38
	E_{max}[a] (nM)	75	15	97	38	89
Ozanimod	EC_{50} (nM)	0.41	>10 000	>10 000	>7865	11
	E_{max}[a] (nM)	97	70	109	21	83

[a]E_{max} relative to 100% S1P signaling.

2. Aims & Objectives

G-protein coupled receptors are the largest family of membrane receptors involved in a plethora of physiological and pathophysiological processes. They represent the most drug-targeted receptor family with uncovered potential to be exploited as targets for many diseases. One main challenge to specifically targeting a GPCR is to achieve subtype selectivity between closely related receptors of a family with sparse binding pocket differences. Furthermore, finding subtle changes in the activation process could elevate the analysis of molecular dynamics simulations and homology models to support structure-based ligand design.

This thesis aims to elucidate structural determinants for subtype selectivity and general class A GPCR receptor activation. One main focus is to decipher the high sequence identity of the orthosteric binding pocket between Sphingosine-1-phosphate receptor subtypes 1 to 5 with *in silico* models. Moreover, we want to evaluate the role of dihedral angles in class A GPCR activation with a thorough data mining and machine learning approach to all available structural data.

Answering the following questions should enhance our understanding of selected topics and demonstrate the viability of structure-based ligand design:

1. Which structural determinants are important for S1PR subtype selectivity?
2. Can these determinants explain the selectivity of marketed drugs?
3. Are dihedrals angles as structural determinants able to distinguish between activation states of GPCRs?
4. What impact can a machine learning model based on dihedral angles have on structure-based drug design?

To answer these questions and overcome posed challenges we combined state-of-the-art molecular modeling, data analysis, and machine learning methods with suitable workflows.

3. Computational Methods

3.1. Molecular Modeling

3.1.1. Homology Modeling

Homology modeling (HM) represents an established method to develop three-dimensional structural predictions of protein targets. As a starting point, similar proteins are used as templates, which have their structural data experimentally confirmed [143]. This assumes that similar protein sequences have resembling structures. To build a valid HM the target protein and template protein need to have high sequence and/or structural similarities. The first step to compile an HM is to select a template by utilizing sequence alignment algorithms such as Clustal Omega [144], FASTA [145], or BLAST [146]. After the identification of a viable template, different model building solutions are available as web services and software suits (e.g. iTasser [147], SWISS-MODEL [148], and MOE (Chemical Computing Group ULC, Montreal, Canada)). In this work, the MOE homology model builder was used, which constructs HM based on the coordinates of matching residues between template and target as well as a library of high-resolution structures and rotamers for non-matching residues to sample the backbone and sidechain rotamers. The initial models should be refined and validated closely before further application for structure-based design. Validation includes the analysis of protein geometry (Ramachandran outliers [39]), atomic clashes, and secondary/tertiary protein structure. Furthermore, another validation method is the docking of ligands with known activity/binding mode [149]. Due to the availability of $S1PR_1$ as crystal structure (PDB 3V2Y) [32], the shared tertiary conformation (chapter 1.1.2), and the high sequence identity of S1PRs (chapter 4.1.1), they consequently represent auspicious conditions for HM.

3.1.2. Molecular Docking

Molecular docking is a structure-based tool for flexibly fitting small molecules into rigid protein binding pockets [150]. It is primarily used for binding mode hypothesis generation

and virtual screening in drug design. The docking process is separated into two components: (I) generating binding modes with a search algorithm and (II) ranking the results with a scoring function [151]. Search algorithms have different approaches to construct conformations. This includes stochastic optimization techniques like Monte Carlo methods (Autodock [152]) and Tabu searches [153], genetic algorithms (GOLD [154]), incremental reconstruction (FlexX [155]), and systematic searches (GLIDE [156]) [157, 158]. After generating a set of docking poses, the conformations can be ranked with methods like molecular mechanic force fields, empirical free energy functions, or knowledge-based functions [150]. In this thesis, the software suite GOLD, which uses a genetic approach, has been used for docking experiments.

3.1.3. Molecular Dynamics Simulations

All-atom molecular dynamics (MD) simulations have become an important method for structure-based drug design since it uses structural data of a rigid macromolecular complex to sample its different conformations around the rigid state [159]. Three steps are taken to perform all-atom MD simulations: the parametrization of the system, equilibration, and the calculation of the velocities and position of the atoms in the main simulation run from starting velocities based on the Boltzmann distribution.

The parametrization step infers detailed properties to each atom of the system (e.g., atomic mass, partial charge, bonds, and van der Waals radii). After the successful parametrization, the system undergoes several equilibration steps (short MD simulations) to reach the desired temperature and pressure for the main calculation and to relax the protein to remove possible crystallization/cryo-EM artifacts of the underlying structural data used [160]. The calculation of MD simulations for the equilibrium steps as well as the main MD simulations are based on Newtonian mechanics as the underlying principle [161]. Furthermore, detailed parameters of the potential energy functions (e.g. bonds, angles, and torsions of covalently bound atoms) or intermolecular interactions through van der Waals forces are summarized as parameters in the commonly termed 'force-fields' (OPLS [162], CHARMM [163], GAFF [164]). These parameters are based on *ab initio* physics, quantum mechanics, and experimental data [165]. Different software solutions are available to

perform MD simulations (Desmond [166], Amber [167], GROMACS [168]). In this thesis, the force-field OPLS 2005 (OPLS_2005) [169] was used in combination with Desmond [166].

3.1.4. Dynamic 3D Pharmacophore Analysis

In 1998 Camille G. Wermuth defined the concept of a pharmacophore in the IUPAC glossary terms [170]: "A pharmacophore is an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response." In other words, 3D pharmacophore models are comprehensive visualizations of ligand-protein interactions represented by spatial pharmacophoric features such as hydrogen bond donors/acceptors, hydrophobic regions, aromatic rings, and charges. Location, direction, and orientation of the aforementioned features as well as steric hindrance are mandatory for a proper 3D pharmacophore model. They play a critical role as an efficient tool for virtual screening [171].

Different software suites are available to generate 3D pharmacophore models (e.g. LigandScout [172, 173], MOE, PHASE [174], and CATALYST [175], Pharmer [176]). This work uses the pharmacophore feature definitions of LigandScout, which includes hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), positive ionizable (PI) feature, negative ionizable (NI) feature, hydrophobic contacts (H), and aromatic ring π-π stacking (AR) features [173].

The new concept of dynamic pharmacophores (Dynophores) [177, 178] incorporates data from MD simulations into 3D pharmacophore models. This allows the shift from a static representation of ligand-protein interactions to a time-dependent dynophore. A pharmacophore model is generated for each frame of the MD simulation, which gives an insight into conformational changes over time. Recurring pharmacophoric features are represented as density functions of single feature points and collected in so-called dynophore *superfeatures* [179].

3.2. Machine Learning

The field of machine learning and artificial intelligence has made large progress due to the development of new learning algorithms, the availability of data, and the rising computational processing power [180]. Machine learning is a branch of artificial intelligence and, in a nutshell, is a method using algorithms to learn generalizable dependencies from large datasets and improve the performance of a given task over time. This broad discipline is utilized in many different fields like healthcare [181, 182], computational chemistry [183], drug discovery [184], and many other fields [185–188]. Machine learning methods can be divided into three categories: supervised, unsupervised, and reinforced machine learning. While supervised machine learning models learn from labeled data to infer the labels of new data, unsupervised machine learning models learn underlying patterns from unlabeled data to e.g., segment the data into clusters. Reinforced machine learning models learn iteratively based on penalties and rewards to reach the given goal in a sequence of decisions (e.g., winning a chess match).

3.2.1. Supervised Machine Learning

The goal of supervised machine learning is to learn a set of rules from instances of a dataset based on the given label (known correct output) [189]. These generalized hypotheses give insight into dependencies between labeled instances and their features (e.g., single or combined attributes) and can be used to predict the unknown labels of new data. Therefore, every instance of a given dataset is represented by the same set of features. These features can be categorical, binary, or continuous, which is also the case for labels of instances. Based on given labels different machine learning algorithms are used. Categorial or binary labels utilize classification algorithms, while regression algorithms are used when labels are continuous (numerical) values [190, 191].

A general workflow for supervised machine learning can be described as follows (Figure 15): (1) the first step is to collect a raw dataset, e.g., of molecules or structural protein data, depending on the central problem to solve. (2) This dataset set will then be pre-processed. Steps include wrangling of data, handling missing data, feature extraction, splitting the dataset into training (80% of the available dataset) and test dataset (20% of the available dataset), and feature engineering and selection based on the training dataset (chapter 3.2.3). After data preparation, the format should be appropriate for further ML downstream applications. (3) Based on the label, ML algorithms are chosen and trained using the training dataset. Initial models need optimization through hyperparameter settings (algorithm-specific changeable parameters), further feature engineering and/or selection based on the performance measurements and metrics (chapter 3.2.4), or changing to more suitable algorithms for the dataset. (4) The last step involves the validation of the optimized ML model by predicting the labels of unknown data. This test data was split in the beginning from the pre-processed dataset and is now used for the final model evaluation. Due to the labels known to the user but not to the model, we can evaluate the performance in a simulated "real world" case. New published/available data can also be included in the validation process. A further optimization cycle can be implemented after the validation process, but one should be aware that the model might be more prone to over-fitting.

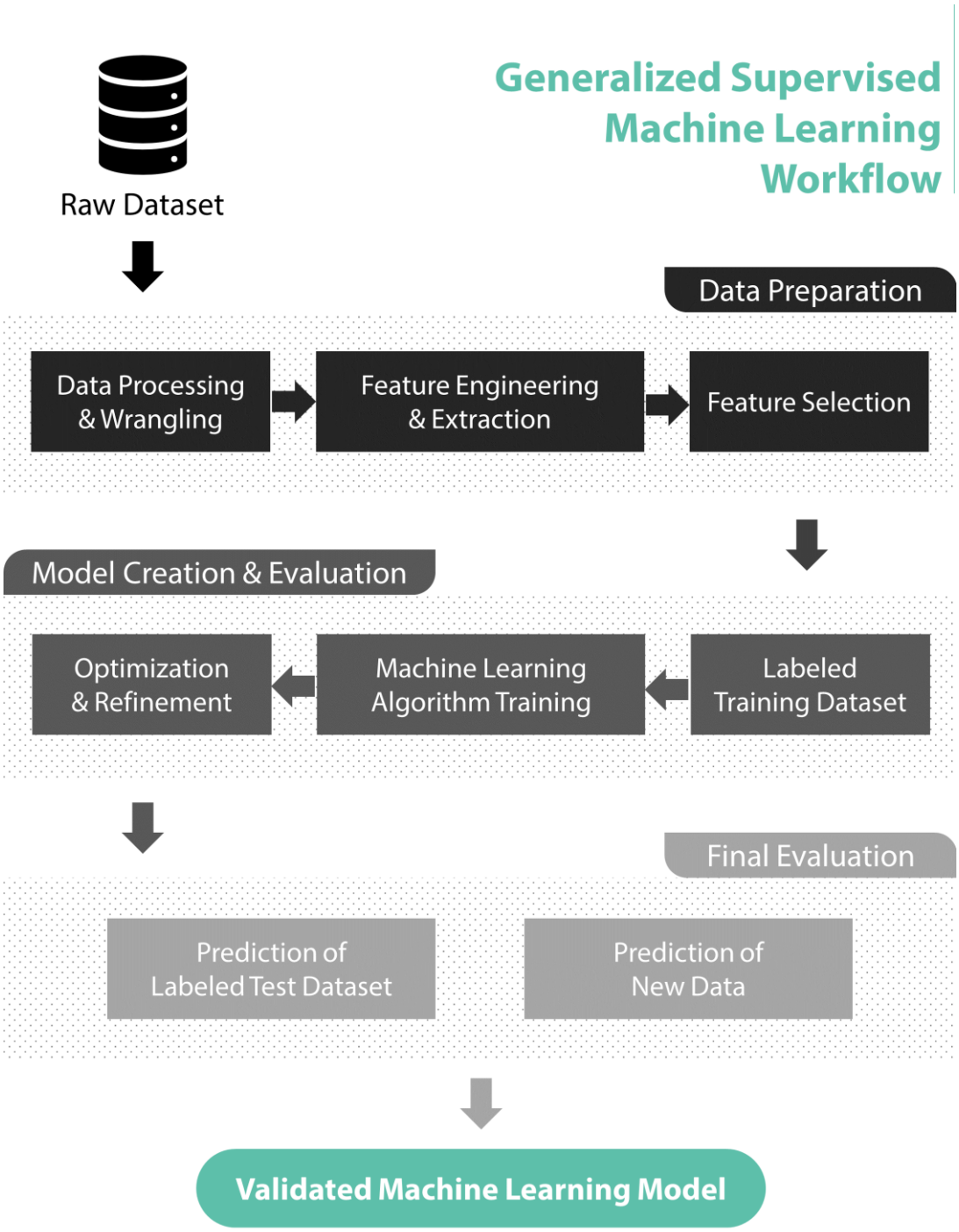


Figure 15. **General workflow for the creation of supervised machine learning models.**

3.2.2. Classification Algorithms

Decision Trees

A family of commonly used methods in supervised machine learning for classification and data mining is non-incremental top-down inductions of decision trees (TDIDT) [192, 193]. In other words, the algorithms are building decision trees (induction) by continuously splitting (top-down) the entire training dataset (non-incremental) based on the features/attributes of the labeled instances. The training dataset induction starts at the topmost decision node, also called root node, represented by the feature that best divides the instance space into its classes based on a feature value acting as a threshold. The resulting data subsets at child nodes are further split (decision node) in the same manner. This process of optimizing the split function of each decision node without the consideration of prior splits is called recursive partitioning [194]. The recursion ends when a successor child nodes subset is purely populated with instances of the same class or further splitting does not add value to the resulting model (e.g., enhancement in predictive power). These terminal nodes are also called leaf nodes and are attributed with a class label based on the subset. The trained decision tree can then predict the label of an instance X by following the learned decision rules of the root node and consecutive child decision nodes until a leaf node is reached. The class label of the leaf node determines the label of X (Figure 16). The clearly defined decision rules at each node and the possibility to translate every root-to-leaf path into a set of rules as a representation of the decision tree offer insight into the underlying dataset. Furthermore, decision trees can handle high-dimensional data without prior domain knowledge or parameter settings.

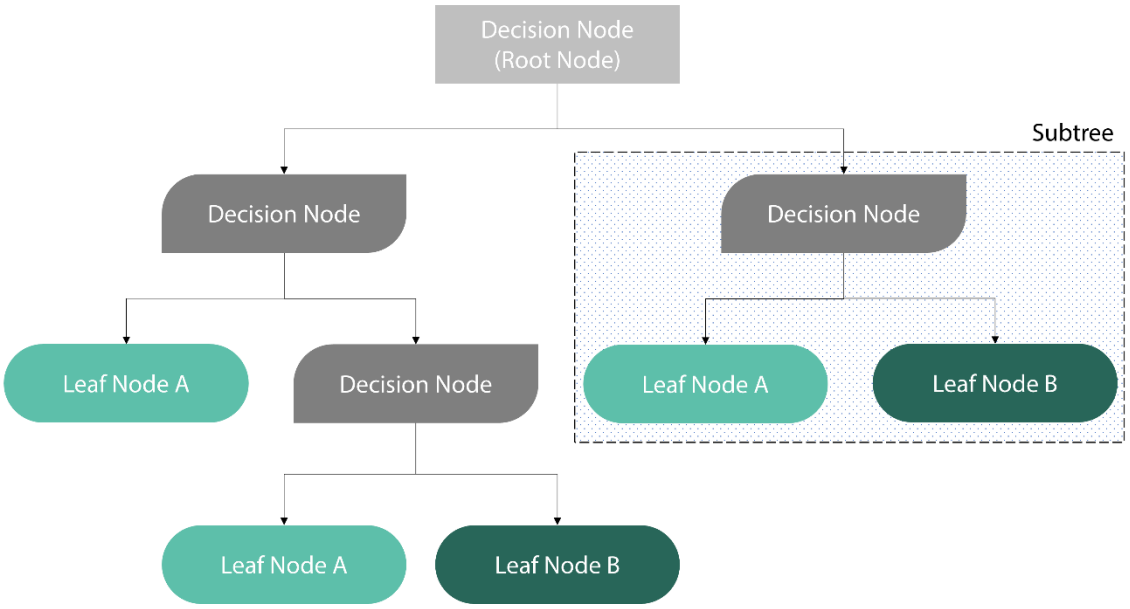


Figure 16. **Generalized Binary Decision Tree.** The class label prediction of an instance X follows a root-to-leaf path. Each decision node on this path sorts X into the successor child node based on learned rules obtained through prior training with a given dataset. Leaf Nodes A and B are representing the binary class labels.

K Nearest Neighbors

The K Nearest Neighbor algorithm is a non-parametric, instance-based classifier and belongs to the group of lazy-learning algorithms [195]. The induction process is performed during the classification of new instances, which leads to less computation time during training, but increased computation time for each classification (in comparison to e.g., decision trees). Furthermore, no generalization of the training data set is carried out leading to a large model which includes all data points. Instances from the training dataset can be seen as points in an n-dimensional instance space with n-features, in which instances with similar properties are in close proximity to each other. By locating the k nearest instances to the new instance, a class label can be determined by a majority vote (Figure 17). The selection of the distance measurement has a significant impact on performance and noise handling [196].

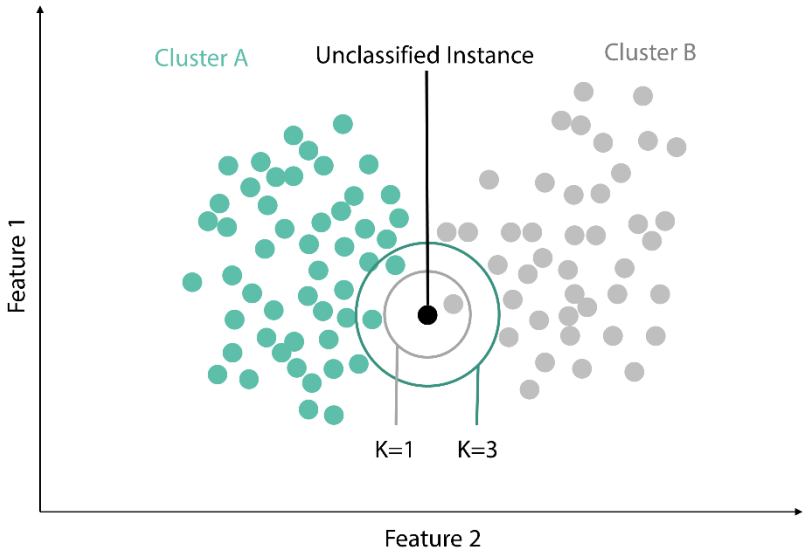


Figure 17. **Classification visualization with K Nearest Neighbor.** By determining the nearest instances to the unclassified instances, a label can be assumed. Based on K different classification outcomes are possible. Therefore, K should be picked properly based on the underlying dataset.

Support Vector Machines

To understand how Support Vector Machines (SVM) induction and classification [197] work we need to understand four concepts of SVMs: (I) The separating hyperplane, (II) the maximum-margin hyperplane, (III) soft-margins, and (IV) the used kernel function [198]. In a one-dimensional dataset with two classes, a separation of the "line" created by one feature can be achieved with a point (comparable to a decision node in a tree-based function), a two-dimensional dataset can be separated by a line (Figure 18), and a three-dimensional dataset by a plane. This can be further upscaled mathematically into higher dimensions. A line in a higher-dimensional space is called a hyperplane. The term is therefore generally referring to a line in an x-dimensional space that separates a dataset with binary classes. By introducing a maximum margin with a distance parameter at each side of the separating hyperplane, the SVM can optimally place the separating hyperplane. The maximization of these margins, also called support vectors, minimizes the upper bound of the generalization error of the model and can therefore maximize the ability to predict labels of unknown data.

Real datasets normally cannot be separated as cleanly as seen in Figure 18 by a hyperplane. Often outliers are found on the opposite cluster of the separating hyperplane. To allow outliers and still place an optimal hyperplane, soft margins have been introduced to SVMs. These soft margins have to be set and optimized (number of allowed outliers, allowed distance to the hyperplane, etc.) so that a large margin can still be achieved for correctly classified instances. The last concept is kernel functions, which are used to mathematically transform low-dimensional data into higher dimensionalities. By choosing the correct kernel function a dataset can become linearly separable in a higher dimension in contrast to other or no kernel function used. One must be aware that the projection into higher dimensionalities with no boundaries will always yield a kernel function that can separate a given dataset. But increasing dimensionality also increases the number of possible solutions leading to a selection problem of the correct function. Moreover, a high-dimensionality kernel function is prone to overfit the training data leading to a poor generalization of the underlying data.

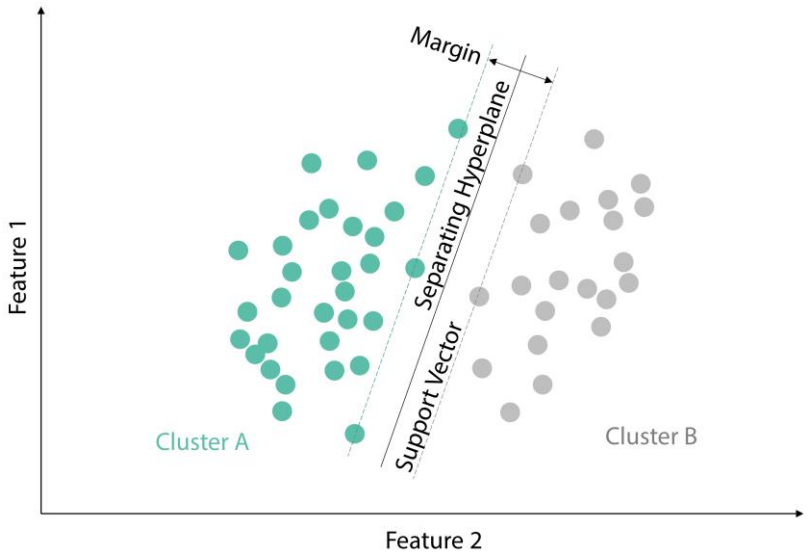


Figure 18. **Support Vector Machine Hyperplane and Maximum Margin using two Features.**

3.2.3. Feature Selection for Supervised Machine Learning

The selection of meaningful features represents an important method before training a supervised machine learning model [199, 200]. Irrelevant features can be described as features that do not correlate with the dataset's class labels. They increase computational cost, introduce the curse of dimensionality (the higher the dimensions, the sparser data becomes), and reduce predictive performance. Besides selecting features based on domain knowledge different methods can be utilized. A baseline method represents the removal of features based on a variance threshold. If a feature shows no or low variance, we can assume that this feature does not contain information to distinguish between different classes from the dataset. Another common approach targets the number of missing values for a feature in a given dataset. The premise here is to remove features with a high number of missing values and therefore remove features with missing information. Other methods can be categorized into filter and wrapper methods. Filter methods are preprocessing steps that utilize statistical tests to rank features independent of the learning algorithms. Features can be examined individually (univariate) or in subsets (multivariate). Examples for filter methods include analysis of variance (ANOVA) [201, 202], mutual information [203, 204] and feature correlation [205, 206]. Wrapper methods on the other hand iteratively rank feature subsets by utilizing predictors and resulting predictive performances until a defined criterion is met (e.g., highest learning performance, number of desired features). Popular wrapper methods include recursive feature elimination with cross-validation (RFECV) [207] or feature selection based on the random forest algorithm [208]. Due to high computational costs, different search strategies have been developed, such as sequential search and genetic algorithms [209, 210].

3.2.4. Model Performance

The performance of a trained machine learning model is important to give an estimate to two central questions: (1) Which of the models performs the best and (2) how will the model perform on "unseen" data [211]. These two aspects are also known as model selection and generalization performance. Based on the underlying data different algorithms can yield performance differences [212]. To evaluate the performance of a binary classification model

we can hold out some of the initial data to validate the model with "unseen" data. The available data is randomly split in the beginning into a stratified (correct representation of each class) training and test set. Performance measure calculations are based on correctly classified positive and negative instances (true positives (TP) and true negatives (TN)) as well as incorrectly classified positive and negative instances (false positives (FP) and false negatives (FN)). Each instance classification can be visualized and counted in a confusion matrix (Table 2) The sum of correctly classified instances (TP + TN) divided by the total number of instances (TP + TN + FP +FN) is defined as the prediction accuracy. By comparing the training and test set accuracy we can evaluate the generalization performance. A high training set accuracy which yields a lower test set accuracy is an indicator of an over-fitting model. In other words, the learned rules of the model are specifically tailored to the training data, leading to bad performance on unseen data.

Table 2. **Confusion matrix for the classification of the test set.**

	Predicted Positive (PP)	**Predicted Negative (PN)**
Actual Positive (P)	True Positive (TP)	False Negative (FN)
Actual Negative (N)	False Positive (FP)	True Negative (TN)

Another popular performance measure is the Mathews correlation coefficient (MCC), which is a measurement for binary classification problems [213]. It not only takes into account the correctly classified instances (TP and TN) but also FP and FN instances leading to a more robust and reliable metric. It is defined as "TP x TN – FP x FN" divided by the square root of "(TP + FP) x (TP + FN) x (TN + FP) x (TN + FN)". The MCC value ranges from -1 to +1, in which +1 indicates a perfect prediction, 0 a random prediction, and -1 no correct classifications.

The discussed hold-out method comes with a crucial caveat: the accuracy calculation is a point estimate of the generalization performance making it sensitive to the particular training and test set split (especially when the dataset shows high variance). A common method used to obtain more robust performance measures is the k-fold cross-validation (Figure 19). Assuming we want to perform 5-fold cross-validation, the first step involves splitting the dataset into five parts (also called folds) of equal size. In an iterative process,

one "fold" is being used as the test set and the remaining folds to train the model. This leads to the generation of five different models with a test opportunity for every instance in the dataset (Figure 19). The average accuracy of the models represents the k-fold cross-validation performance measure.

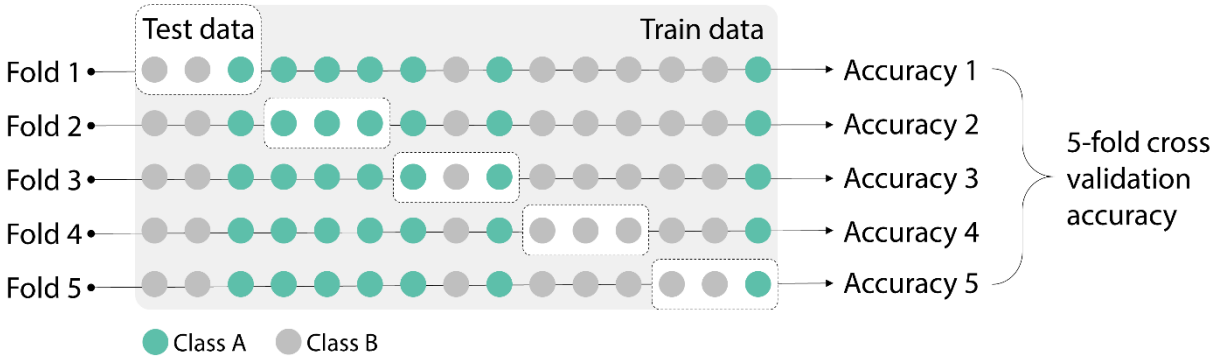


Figure 19. **k-fold cross-validation is shown with 5-folds as an example.**

4. Results

4.1. Structural Determinants for Sphingosine-1-phosphate Receptor Selectivity

4.1.1. $S1PR_{2-5}$ Homology Models Reveal Highly Similar Orthosteric Binding Pockets of Receptor Subtypes

In order to identify structural differences between the binding pockets of the Sphingosine-1-phosphate receptors (S1PR) on an atomistic level, we built the homology models of $S1PR_{2-5}$, which at the time of investigation, lacked experimentally determined coordinates. The crystal structure of $S1PR_1$ (PDB 3V2Y [32]) was used as the template for homology model building because of the high sequence identity (45,3 - 61,5%, depending on the S1PR subtype, Figure 20C), high sequence similarity (65.3 – 78.1%, depending on the S1PR subtype, Figure 20C), and the same receptor classification (chapter 1.1.1) of $S1PR_{1-5}$. Since the template protein was co-crystallized with an antagonist, the built homology models mimic the inactive state. All derived homology models showed under 1 Å overall atomic deviation (measured as root-mean-square deviation, RMSD) of the Cα of the protein backbone in comparison with the template (Figure 20B) suggesting a comparable global fold to the template protein.

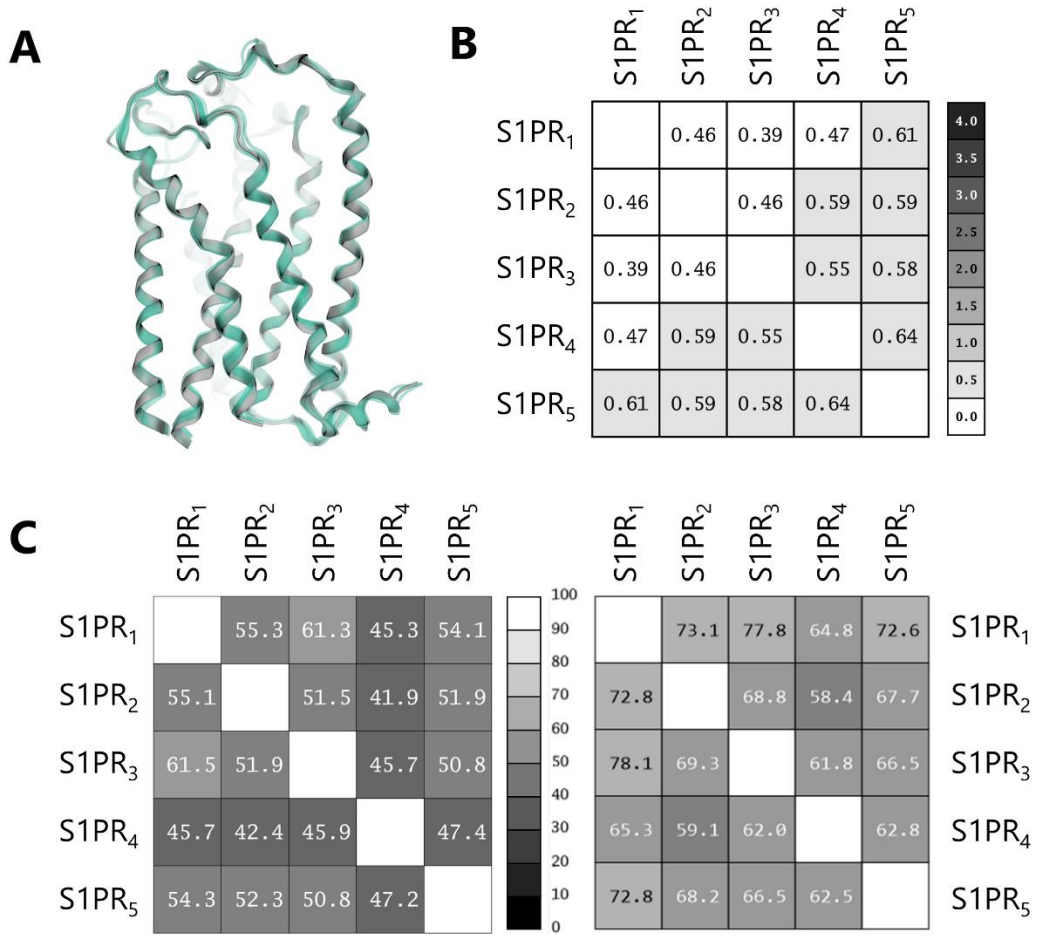


Figure 20. **(A) $S1PR_{2-5}$ homology models (green) superimposed on the $S1PR_1$ template (grey). (B)** RMSD (root-mean-square deviation) of the Cα protein backbone matrix in Å. **(C)** Sequence identity (left) and similarity (right) in percent.

For further model validation, the protein geometry based on dihedral angles (Chapter 1.1.3) was analyzed using the Ramachandran plot [214] shown in Figure 21. Residues with unfavorable backbone geometries are marked as outliers (red cross). There are four outliers, one occurring in $S1PR_2$ (N^{19} in the N-terminus), two in $S1PR_3$ (L^{28} in the N-terminus and N^{68} found in ICL1), and one in $S1PR_4$ ($R^{4.41}$ found in TM4). The outliers do not affect the binding site in any of the homology models. In addition, the outliers are in highly flexible regions and were thus not further investigated (Figure 21). Furthermore, no atom clashes were observed.

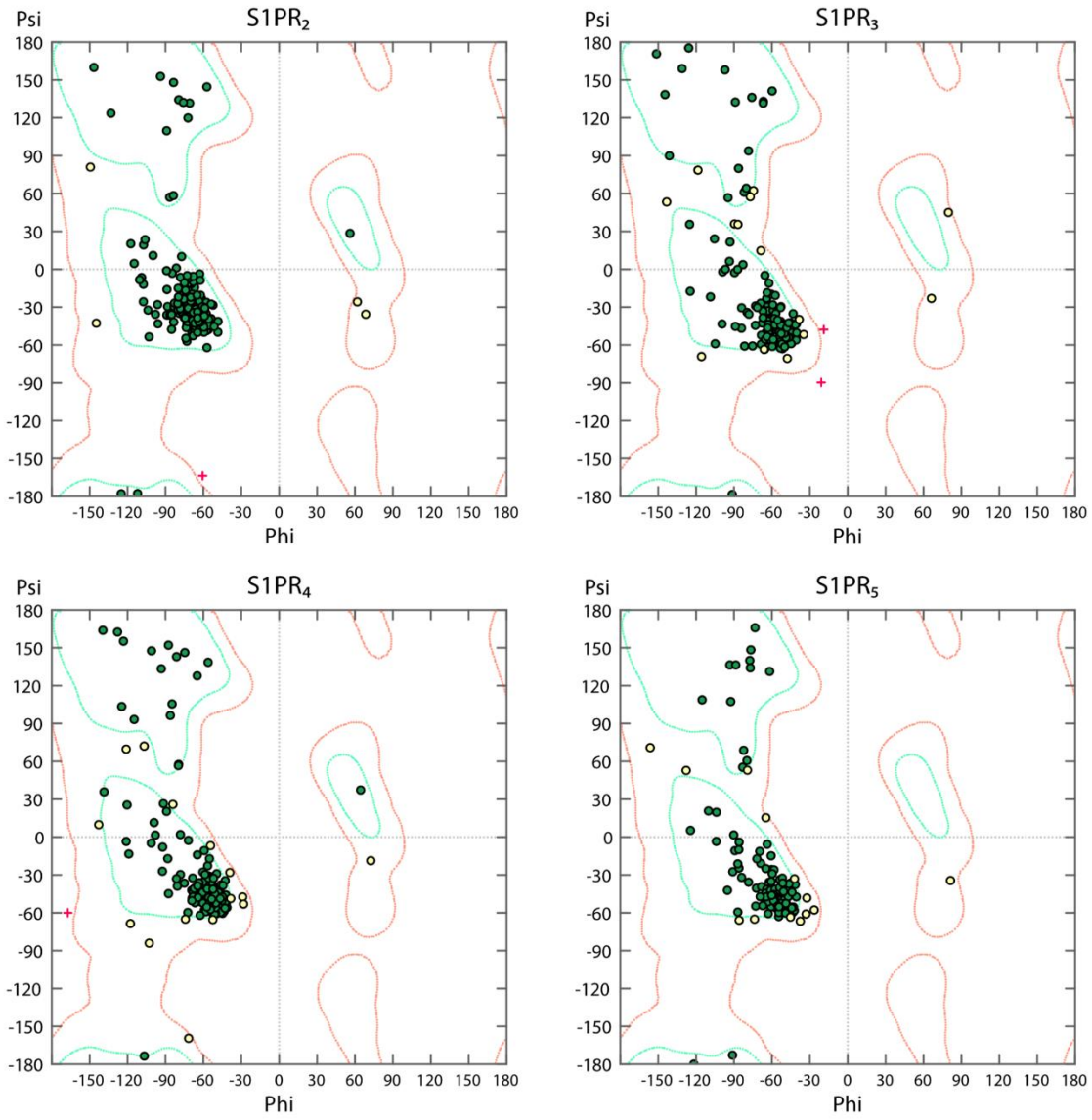
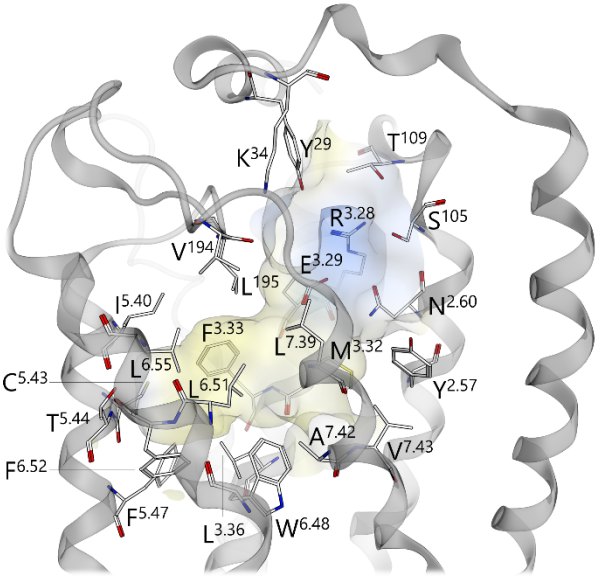


Figure 21. **Ramachandran plots of homology model S1PR$_{2\text{-}5}$.** Green points represent optimal phi and psi angles, yellow circles are still in the allowed region and red crosses indicate angles in non-favorable regions of the Ramachandran plot.

The orthosteric binding pocket of S1PRs can be divided into a hydrophilic and lipophilic region. A sequence alignment and superimposition of the structural data reveals highly conserved residues in the hydrophilic region throughout S1PR$_{1\text{-}5}$ (N$^{2.60}$, S^{105}, T^{109}, R$^{3.28,}$ and E$^{3.29}$, Figure 22). One exception is R^{34} expressed in S1PR$_4$ instead of K^{34} found in S1PR$_{1,2,3,5}$. Seven residues in the lipophilic core of orthosteric binding pocket are also conserved, such as Y^{29,18,22,29,19}, L^{195,183,189,197,186} (S1PR$_{1\text{-}5}$, respectively), F$^{3.33}$, L$^{3.36}$, F$^{5.47}$, W$^{6.48}$ and F$^{6.52}$ (Figure 22). The differences in the binding pocket are present in the lipophilic region at BWN residues 2.57,

5.40, 6.55, 7.39, and 7.42 (Figure 22, right). These residues share strongly similar properties (calculated with Clustal Omega [144, 215]). Based on these findings, achieving subtype selectivity via the orthosteric binding pocket is only feasible by exploiting the lipophilic residue differences. To gain more insights between receptor subtypes a dynamic comparative analysis was utilized to unveil structural determinants important for S1PR subtype selectivity.



	N-Term	N-Term	2.57	2.60	ECL1	ECL1	3.28	3.29	3.32	3.33	3.36	ECL2
$S1PR_1$	Y^{29}	K^{34}	Y	N	S^{105}	T^{109}	R	E	M	F	L	V^{194}
$S1PR_2$	Y^{18}	K^{22}	F	N	S^{93}	T^{97}	R	E	A	F	L	V^{182}
$S1PR_3$	Y^{22}	K^{27}	Y	N	S^{99}	T^{103}	R	E	M	F	L	I^{188}
$S1PR_4$	Y^{29}	R^{34}	Y	N	S^{106}	T^{110}	R	E	L	F	L	L^{196}
$S1PR_5$	Y^{19}	K^{24}	Y	N	S^{96}	T^{100}	R	E	V	F	L	V^{185}
	*	:	:	*	*	*	*	*		*	*	:

	ECL2	5.40	5.43	5.44	5.47	6.48	6.51	6.52	6.55	7.39	7.42	7.43
$S1PR_1$	L^{195}	I	C	T	F	W	L	F	L	L	A	V
$S1PR_2$	L^{183}	V	V	V	F	W	A	F	L	F	S	T
$S1PR_3$	L^{189}	I	C	I	F	W	L	F	F	I	A	V
$S1PR_4$	L^{197}	I	C	L	F	W	L	F	L	L	A	V
$S1PR_5$	L^{186}	V	C	V	F	W	L	F	L	L	A	M
	*	:			*	*		*	:	:	:	

Figure 22. **Orthosteric binding site residue comparison between $S1PR_{1-5}$ structures. Left:** $S1PR_1$ binding pocket residues. Blue surface – hydrophilic environment, yellow surface – lipophilic environment. **Right:** Binding pocket residue comparison between $S1PR_{1-5}$. "*" - single, fully conserved residue, ":" - conservation between groups of strongly similar properties, ECL1/2 - extracellular loop 1/2, top row numbers indicate Ballesteros-Weinstein nomenclature, Yellow - apolar amino acids, Blue - polar amino acids, Green – cysteine

4.1.2. Knowledge-based Constrained Docking of a Sphingosine-1-phosphate Receptor Pan-agonist

The main goal of this docking study is to prepare valid binding modes for a dynamic interaction analysis between $S1PR_{1-5}$. Docking studies were performed using the pan-agonist 2-ammonio-3-hydroxy-2-(5-(4-octylphenyl)-1H-imidazol-2-yl)propylphosphate (**6**, Figure 23) [216]. Compound **6** was discovered in 2005 by Clemens et al. during the development of 4(5)-phenylimidazole-based analogs for $S1PR_1$. Its chemical difference to the endogenous ligand S1P and the measured agonistic activity for all five subtypes (EC_{50} $S1PR_1$-$S1PR_5$ in nM: 7.9, 18, 630, 160, 17; E_{max} based on 100% S1P for $S1PR_{1-5}$: 0.91, 0.95, 0.21, 0.87, 0.66) predestine **6'** utilization for the development of comparative dynamic models.

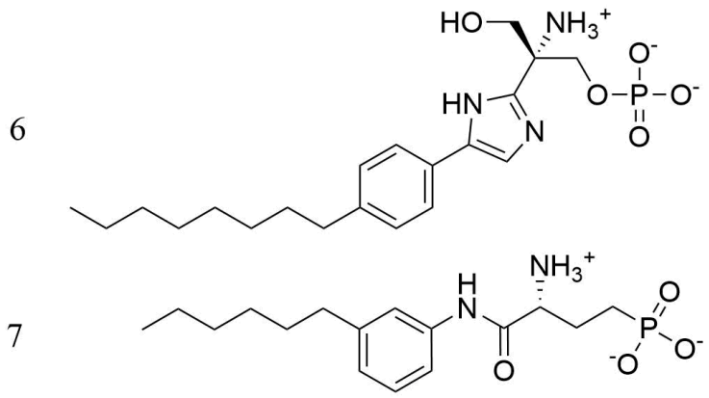


Figure 23. **Structure of docked compound (6) and ML056 (7). (6)** shows agonistic activity on all five S1PR substructures (EC_{50} $S1PR_1$-$S1PR_5$ in nM: 7.9, 18, 630, 160, 17; E_{max} based on 100% S1P for $S1PR_{1-5}$: 0.91, 0.95, 0.21, 0.87, 0.66). **(7)** ML056 is a selective $S1PR_1$ antagonist co-crystallized with $S1PR_1$ (PDB 3V2Y).

The constrained docking of compound **6** was performed based on the position and interactions of the ligand ML056 in crystal structure 3V2Y (see Chapter 1.2.2) and S1P loss-of-function alanine mutation studies of key residues $R^{3.28}$A and $E^{3.29}$A in the binding pocket of $S1PR_1$ [217] (Table 3).

Table 3. **Constraints used for docking compound 6 into S1PR subtypes.**

Constraint Type	*Compound 6*	*Residue*	*Rationale*
Hydrogen Bond	Deprotonated Oxygen of Phosphate Group	$S1PR_1$ K^{34} (HZ2 atom) $S1PR_2$ K^{22} (HZ2 atom) $S1PR_3$ K^{27} (HZ2 atom) $S1PR_4$ R^{34} (HH21 atom) $S1PR_5$ K^{24} (HZ2 atom)	Observed in $S1PR_1$ crystal structure (PDB 3V2Y) [32]
Hydrogen Bond	Deprotonated Oxygen of Phosphate Group	$S1PR_{1-5}$ $R^{3.28}$ (HH11 atom)	Observed in $S1PR_1$ crystal structure (PDB 3V2Y) [32], Mutation study [217]
Hydrogen Bond	Protonated Nitrogen of Primary Amine	$S1PR_{1-5}$ $E^{3.29}$ (OE2 atom)	Observed in $S1PR_1$ crystal structure (PDB 3V2Y) [32], Mutation study [217]
Distance	Terminal C-atom (CH_3-) of alkane substituent	$S1PR_{1-5}$ $F^{5.47}$ (Cz atom)	Observed in $S1PR_1$ crystal structure (PDB 3V2Y) [32], Deepest point of the lipophilic pocket

The resulting docking poses of **6** show overall homogenous conformations in all five subtypes with the expected fitting of the lipophilic tail in the lipophilic pocket and the head group of compound **6** in the hydrophilic pocket (Figure 24). While $S1PR_1$ and $S1PR_5$ show the most comparable conformations between each docking pose, $S1PR_{2-4}$ docking poses of compound **6** reveal more deviations in the position of the phosphate group.

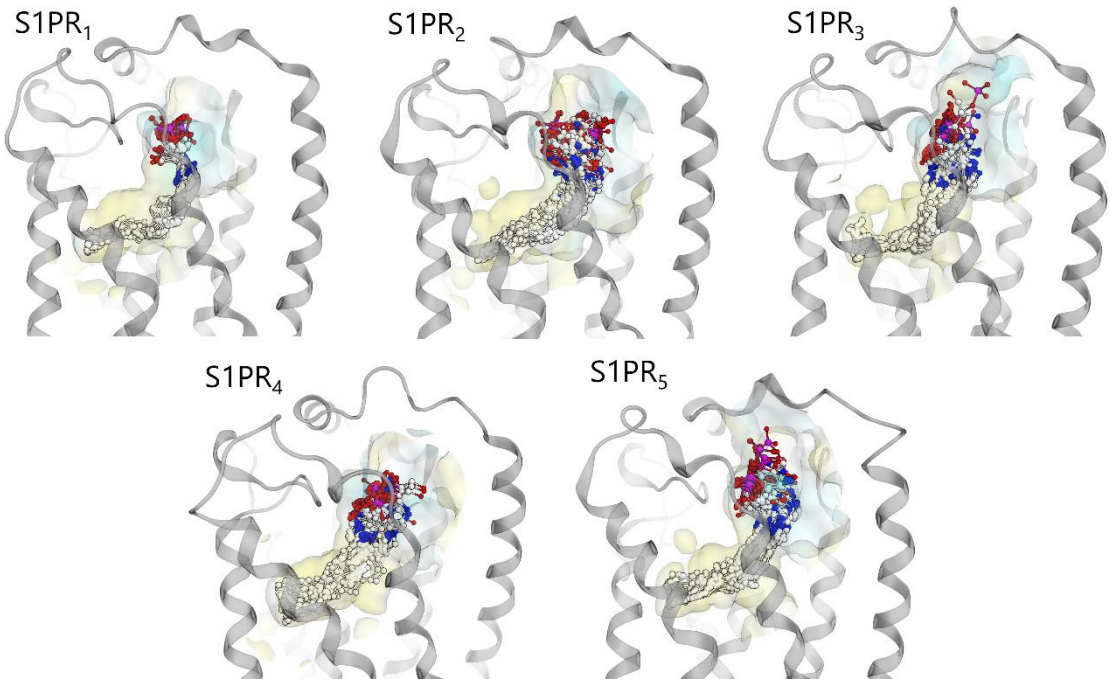


Figure 24. **Docking solutions of compound 6 in $S1PR_{1-5}$.** Yellow surface – lipophilic, blue surface – hydrophilic.

The selection of the final docking pose of **6** was based on the number of fulfilled constraints. If multiple poses fulfill the same set of constraints the scoring function "ChemPLP Fitness Score" [218] was used as an additional selection criterion. $S1PR_1$ fulfilled all defined constraints (Figure 25). For $S1PR_{2-5}$ the interaction between the phosphate group of **6** and K^{22}, K^{27}, K^{24} ($S1PR_2$, $S1PR_3$, $S1PR_5$, respectively) or R^{34} of $S1PR_4$ could not be established without losing hydrogen bonds to crucial residues $R^{3.28}$ or $E^{3.29}$. Due to additional mutational studies on residues $R^{3.28}$ and $E^{3.29}$, hydrogen bonds to these residues were prioritized over K^{22}, K^{27}, K^{24} ($S1PR_2$, $S1PR_3$, $S1PR_5$, respectively) or R^{34} of $S1PR_4$ in the final docking pose selection. The five subtypes in complex with **6** have been used for preparations of molecular dynamics simulations.

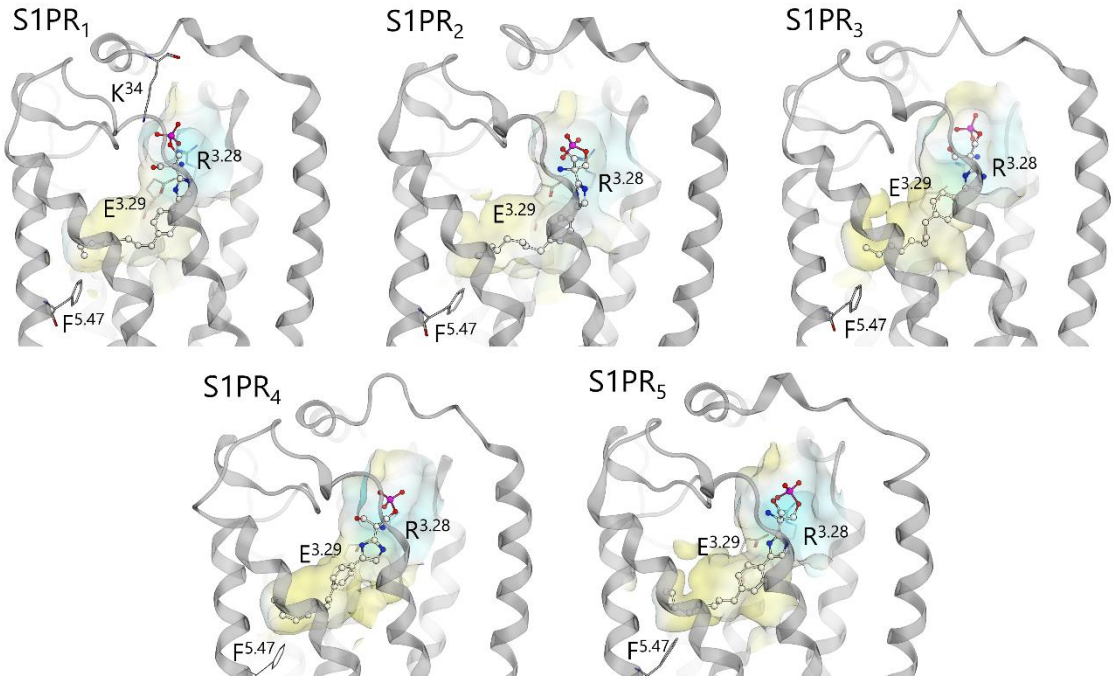


Figure 25. **Selected docking pose of compound 6 for S1PR$_{1\text{-}5}$.** Yellow surface – lipophilic, blue surface – hydrophilic.

4.1.3. Quality Assessment of Molecular Dynamics Simulations of S1PR$_{1\text{-}5}$ in Complex with Compound 6

Extensive MD simulations for each subtype in complex with the chosen docking pose of compound **6** have been carried out (10 simulations à 100 nanoseconds per subtype). MD simulations were prepared as described in the methods section.

To quantify the conformational changes of the ligand in the binding pocket and the protein, we calculated the RMSD of the Cα atoms of the protein backbones and the RMSD of the heavy atoms of compound **6**. Protein RMSD plots for S1PR$_{1\text{-}5}$ reveal comparable low distance deviations between replicas and subtypes (Figure 26) with mean RMSD values for each subtype range between 2.83 - 3.31 Å (Table 4) with a maximum standard deviation of 0.20 Å. Compound **6'** mean RMSD values (2.44 – 3.15 Å), as well as the plots, show higher fluctuations between replicas leading to higher standard deviations (0.36 – 0.64 Å). In addition to the RMSD based assessment further evaluation criteria are applied. This includes the visual inspection for simulation artifacts (unwinding of α-helices, unnatural kinks in

transmembrane domains, and displacement of transmembrane domains), the recurring formation (at least 75% during any MD trajectory) of the negative ionizable feature between the phosphate moiety of compound **6** and the highly $S1PR_{1-5}$ conserved residue $R^{3.28}$, and the recurring formation (at least 75% during any MD trajectory) of the positive ionizable feature between compound **6'** primary amine and the highly conserved $S1PR_{1-5}$ residue $E^{3.29}$. 7 out of 50 MD simulations did not meet one or more of these criteria: $S1PR_1$ replica 0, $S1PR_2$ replica 1 and 2, $S1PR_3$ replica 4, 7, and 8, and $S1PR_4$ replica 4. These replicas are outliers. All outlier replicas are due to simulation artifacts, except for replica 4 of $S1PR_3$, which only formed 52% of the MD simulations trajectory the positive ionizable feature between compound **6'** primary amine and residue $E^{3.29}$ of $S1PR_3$. The remaining MD simulations have been used for dynamic receptor-ligand interaction analysis. Based on the protein RMSD plots the first 10 ns from each MD simulation have been discarded to account for the protein and ligand equilibration.

Table 4. **Mean RMSD in Å for each MD simulation with overall mean and standard deviation in Å for S1PR subtype.**

	Replica	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Ø*	*SD*
$S1PR_1$	Protein	2.75	2.72	2.89	2.87	2.60	2.77	2.94	3.04	2.70	3.03	2.83	0.15
	Ligand	3.17	2.92	3.00	4.04	2.97	2.98	2.97	2.94	2.90	3.56	3.15	0.37
$S1PR_2$	Protein	3.13	3.01	3.40	3.26	2.95	2.95	2.90	3.31	3.03	3.25	3.12	0.18
	Ligand	2.42	2.80	2.51	2.21	2.86	2.68	3.59	3.55	2.88	3.00	2.85	0.45
$S1PR_3$	Protein	3.20	3.30	3.42	3.04	3.64	3.30	3.43	3.21	3.48	3.07	3.31	0.19
	Ligand	3.48	3.86	2.08	2.73	3.49	2.58	2.29	2.97	2.45	2.83	2.88	0.58
$S1PR_4$	Protein	3.26	3.11	3.08	3.28	3.45	3.23	3.06	3.67	3.05	3.13	3.23	0.20
	Ligand	1.90	2.31	2.19	2.67	2.05	2.56	2.21	2.85	2.72	2.95	2.44	0.36
$S1PR_5$	Protein	2.96	3.23	3.17	3.22	3.38	3.27	3.27	3.29	3.45	3.38	3.26	0.14
	Ligand	2.60	3.61	2.67	4.21	2.79	3.62	2.65	2.58	2.50	3.82	3.11	0.64

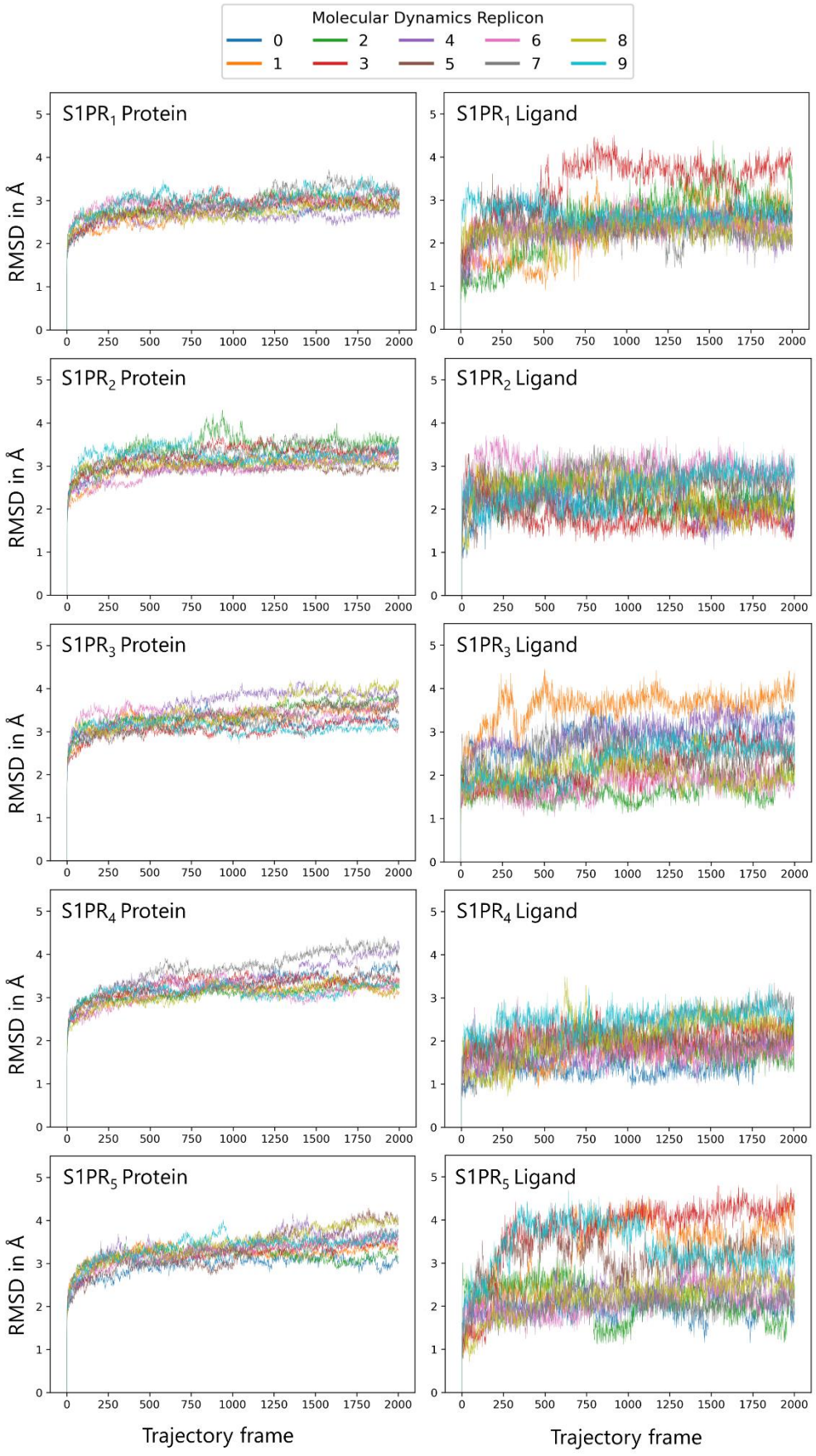
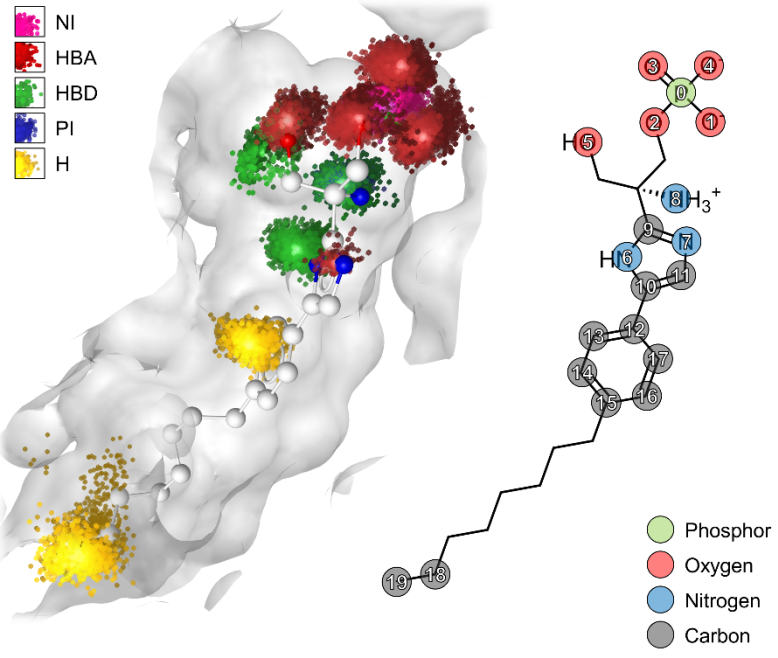


Figure 26. **Protein and ligand RMSD plots for S1PR subtypes.**

4.1.4. Frequent Hydrophilic Contacts Cannot Explain Subtype Selectivity

The dynamic receptor-ligand interactions have been calculated with the *Dynophore* app (chapter 3.1.4). Interaction occurrences are analyzed as the mean percent of replicas per subtype. Ionic interactions consist of a negative ionizable (NI) feature formed with the phosphate group of **6** and a positive ionizable (PI) feature formed with the primary amine of **6**. Both features occur for S1PR subtypes 100% of the simulation time, except for PI of $S1PR_1$ with 99.3% occurrence (Figure 27, Appendix Table 1). The occurrence of hydrogen bond acceptors (HBA) formed by the phosphate head group of compound **6** varies between subtypes due to the fact that feature occurrences are strictly counted for one functional group (or atom) and that the phosphate group exhibits a rotatable bond with a single degree of freedom. The calculation of the mean HBA feature of the phosphate moiety reveals comparable total interaction occurrences ($S1PR_1$ 69.3%, $S1PR_2$ 70.5%, $S1PR_3$ 70.5%, $S1PR_4$ 80.6%, $S1PR_5$ 71.0%). Only $S1PR_4$ shows a 10% higher mean interaction occurrence. The reason for this is the mutation of $R^{3.4}$ expressed in $S1PR_4$ instead of $K^{3.4}$ for the other subtypes. The chemical property of arginine's guanine moiety enables the side chain to form two hydrogen bonds simultaneously to the phosphate moiety in comparison to the primary amine of lysine. The primary amine moiety of **6** forms added to the PI feature a hydrogen bond donor (HBD) feature. In all 5 subtypes, the occurrence ranges from 96.1 to 99.9%, which indicates frequent interaction formation. All aforementioned features are established to one of the two key residues $R^{3.28}$ and $E^{3.29}$ important for S1P binding [216]. This suggests a plausible binding mode of compound **6** in the orthosteric binding pocket subtype wide with low mobility within the binding pocket over the course of MD simulations. Furthermore, the lipophilic contacts of $S1PR_{1\text{-}5}$ formed by the alkyl chain and phenyl moiety occurred 100% of the time, except for $S1PR_5$' alkyl contact with 99.8% occurrence.



	S1PR$_1$	S1PR$_2$	S1PR$_3$	S1PR$_4$	S1PR$_5$
AR [6,7,9,11,12]	0	6,0	0	0	0
AR [12,13,14, 15,16,17]	0	8,0	0	0	0,1
HBA [1]	83,7	89,4	79,8	99,6	84,3
HBA [2]	44,5	18,6	38,6	86,8	29,8
HBA [3]	85,4	84,5	76,6	51,3	73,5
HBA [4]	63,5	89,5	86,9	84,7	96,5
HBA [5]	34,7	22,3	19,5	52,1	15,5
HBA [7]	5,3	0,6	13,0	0,4	1,1
HBD [5]	43,1	56,6	34,0	54,9	57,1
HBD [6]	80,9	53,9	99,0	99,1	72,5
HBD [8]	96,3	99,6	96,1	99,9	96,2
H [12,13,14, 15,16,17]	100	100	100	100	100
H [18,19]	100	100	100	100	99,8
NI [0,1,2,3,4]	100	100	100	100	100
PI [8]	99,3	100	100	100	100

Figure 27. **S1PR$_{1\text{-}5}$ mean interaction occurrence of pan agonist. Left:** 3D view of S1PR$_1$ in complex with pan agonist showing a representative dynamic pharmacophore. **Middle:** 2D representation of pan agonist. Atoms involved in an interaction are numbered. **Right:** Mean interaction occurrences of S1PR$_{1\text{-}5}$. Numbers in square brackets represent the according to atoms involved in a given interaction. AR – aromatic ring interaction, HBA – hydrogen bond acceptor, HBD – hydrogen bond donor, H – lipophilic contact, NI – negative ionizable feature, PI – positive ionizable feature

Interactions with occurrence fluctuations between S1PR subtypes include HBA features formed by the hydroxyl (15.5 - 52.1%), imidazole moiety (0.6 – 13.0%), and phosphate moiety (34.0 – 57.1% calculated for separate oxygen atoms) as well as one HBD feature formed by imidazole moiety (53,9 – 99.1%). Interestingly, S1PR$_2$ is the only subtype that forms aromatic ring (AR) interactions. Both phenyl and imidazole moieties form aromatic π-π (AR) interactions with F$^{7.39}$ (6.0 and 8.0%, respectively). Furthermore, AR interaction switches between the phenyl and imidazole moiety of compound **6** have been observed in MD simulations. Since S1PR$_2$ exclusively expresses F$^{7.39}$ (S1PR$_{1,4,5}$ L$^{7.39}$, S1PR$_3$ I$^{7.39}$, Figure 28) this interaction could contribute to the higher binding affinity of compound **6** to S1PR$_2$ in comparison to the S1PR$_{3,4}$ and similar binding affinity to S1PR$_5$ (EC_{50} S1PR$_1$-S1PR$_5$ in nM: 7.9, 18, 630, 160, 17; E_{max} based on 100% S1P for S1PR$_{1\text{-}5}$: 0.91, 0.95, 0.21, 0.87, 0.66 [216]).

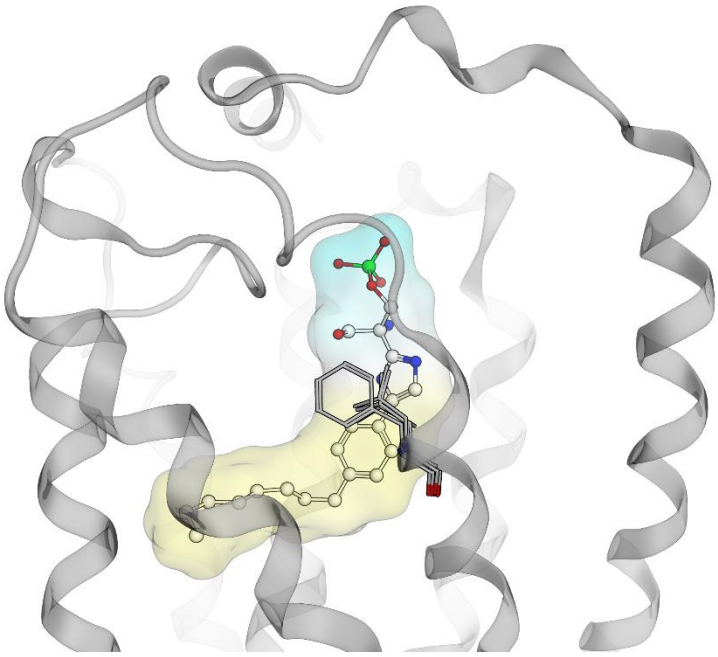


Figure 28. **S1PR$_2$ expresses F$^{7.39}$, while other subtypes express I/L$^{7.39}$.** The phenylalanine forms aromatic ring interactions with the imidazole or phenyl moiety of compound 6 during MD simulations. Blue surface – hydrophilic, yellow surface – lipophilic

Due to the possible formation of multiple HBA interactions with the phosphate moiety, or in extreme cases the formation of none, an analysis of total formed hydrogen bonds for each time step of the MD simulations trajectory has been conducted based on the dynophore data (Figure 29, Appendix Table 2). The formation of two or more hydrogen bonds per frame between S1PR subtypes ranges between 95.8 – 99.6%, with S1PR$_3$ showing the lowest total count which in turn also leads to S1PR$_3$ expressing the highest count of one or fewer hydrogen bonds during MD simulations (4.2%). This sugge sts less energetically favorable phosphate group interactions with key residues important for binding and activity of S1PR$_3$. This result is also in line with compound **6** showing the weakest potency for S1PR$_3$ [216]. Due to highly conserved hydrophilic residues between all subtypes the hydrogen bond network analysis of the phosphate moiety represents an indicator correlating with the subtype-selective behavior of compound **6**.

In summary, the presented data and analysis of dynamic hydrophilic interactions and their occurrences are in line with the subtype-selective behavior of compound **6**. The unveiled unique aromatic interactions of S1PR$_2$ residue F$^{7.39}$ with the imidazole and phenyl moiety of

compound **6** represents a $S1PR_2$ selectivity determinant which supports the high agonistic potency of compound **6** for $S1PR_2$ binding affinity expressed by compound **6.** The next chapters will focus on an in-depth analysis of lipophilic interactions.

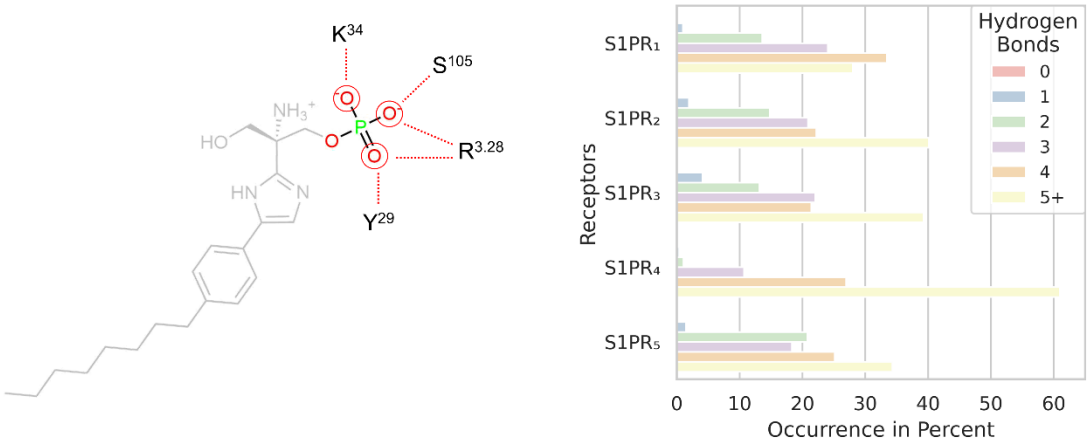


Figure 29. **Hydrogen bond count per frame between the phosphate group of compound 6 and S1PR subtypes. Left:** 2D representation of compound **6'** phosphate group forming 5 hydrogen bond interactions with $S1PR_1$ residues during one frame. **Right:** Total occurrence of hydrogen bond count per frame during all MD simulations for each subtype.

4.1.5. Pan-agonist Shows Three Distinct Conformations of the Lipophilic Tail During Molecular Dynamics Simulations

Visual inspections of $S1PR_{1\text{-}5}$ MD simulations revealed recurring stable binding modes distinct from the initial docking pose conformation. The shared characteristics of these binding modes are (I) comparable types and the number of hydrophilic interactions, (II) highly frequent phosphate and primary amine moiety ionic interactions, and (III) comparable frequency of the lipophilic contacts of the phenyl moiety of compound **6.** Differences have only been observed for the alkyl moiety expressing three distinct conformations. While we have observed all three binding modes in $S1PR_2$ and $S1PR_5$, the other subtypes $S1PR_{1,3,4}$ only show two binding modes consistently. The analysis in this chapter will therefore use $S1PR_5$ as a reference point because compound **6** exhibits all three binding modes in $S1PR_5$.

The initial docking pose of compound **6** resembles the pose of antagonist ML056 co-crystallized with $S1PR_1$ (PDB 3V2Y [32]). This state will be further referred to as "binding mode 1" (BM1). BM1 was observed in all subtypes. Lipophilic contacts of the BM1 alkyl moiety are characterized by interactions to the side chains of $C^{5.43}$, $V^{5.44}$, $W^{6.48}$, $L^{6.51}$, $F^{6.52}$, and $L^{6.55}$ for $S1PR_5$ (Figure 31). Residue $W^{6.48}$ and $F^{6.52}$ are conserved for all subtypes. $C^{5.43}$ is expressed in $S1PR_1$, $S1PR_3$, $S1PR_4$, and $S1PR_5$, while $S1PR_2$ contains $V^{5.43}$. For BWN-residue 5.44 $S1PR_{2\text{-}5}$ express amino acids with lipophilic properties ($V^{5.44}$, $I^{5.44}$, $L^{5.44}$, and $V^{5.44}$, respectively), while $S1PR_1$ shows less lipophilic $T^{5.44}$ with a polar hydroxyl group. The residue $L^{6.51}$ is conserved for $S1PR_1$, $S1PR_3$, $S1PR_4$, and $S1PR_5$. $S1PR_2$ expresses $A^{6.51}$ instead. The last transmembrane position 6.55 expresses in $S1PR_{1\text{-}5}$ residues with lipophilic properties ($L^{6.55}$, $L^{6.55}$, $F^{6.55}$, $L^{6.55}$, $L^{6.55}$, respectively).

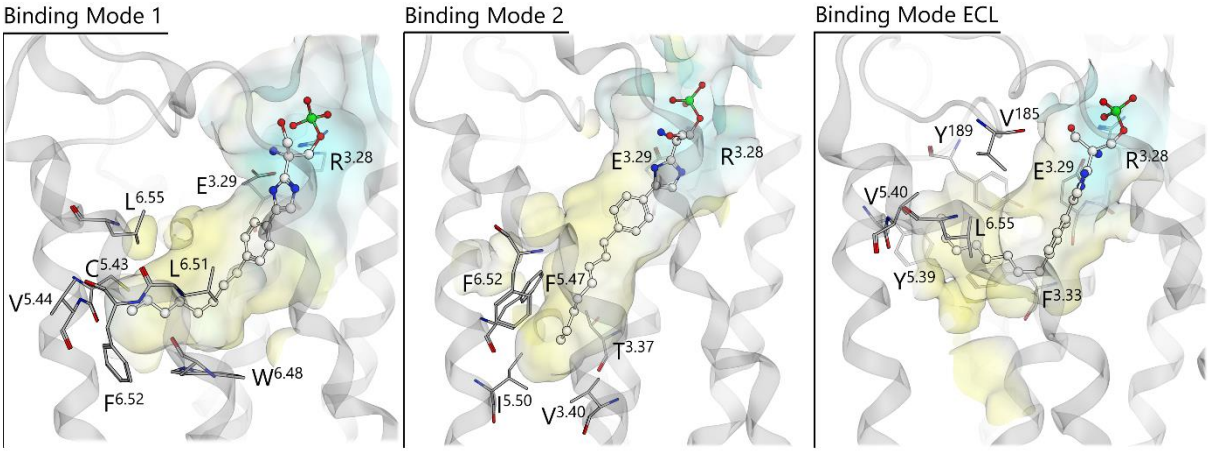


Figure 30. **Static representation of Binding Mode 1, 2, and ECL with compound 6 shown in complex with $S1PR_5$.** Residues forming a lipophilic contact with the alkane moiety in this lipophilic sub-pocket are shown for all three binding modes. Furthermore, the residues $R^{3.28}$ and $E^{3.29}$ are shown due to their importance for anchoring the phosphate and amine moiety of compound **6** into the binding pocket. Blue surface – hydrophilic, yellow surface – lipophilic

In "binding mode 2" (BM2) of compound **6**, the alkyl moiety binds a sub-pocket near the transmembrane receptor core, orienting the pan agonist in parallel to the transmembrane domains. BM2 was observed in all S1PR subtypes. The BM2 sub-pocket consists of $T^{3.37}$, $V^{3.40}$, $I^{5.50}$, $F^{5.47}$, and $F^{6.52}$ for $S1PR_5$ (Figure 31). $S1PR_1$, $S1PR_2$, and $S1PR_5$ express at BWN position 3.37

residues with neutral-polar characteristics ($S^{3.37}$, $S^{3.37}$, and $T^{3.37}$, respectively), while $S1PR_3$ expresses $G^{3.37}$ and $S1PR_4$ $A^{3.37}$. Position 3.40 expresses $V^{3.40}$ for $S1PR_1$, $S1PR_2$, and $S1PR_5$ and $T^{3.40}$ for $S1PR_3$ and $S1PR_4$. $S1PR_{1-5}$ share residues with strongly similar lipophilic properties at residue position 5.50 ($L^{5.50}$, $I^{5.50}$, $I^{5.50}$, $V^{5.50}$, and $I^{5.50}$, respectively). $F^{5.47}$ and $F^{6.52}$ are fully conserved in $S1PR_{1-5}$.

The last observed conformation will be referred to as "binding mode ECL" (BM ECL). BM ECL shows lipophilic contacts between compound **6'** alkyl moiety and ECL2 residues $Y^{198/186/192/200/185}$ and $V^{194/182/198/196/185}$ ($S1PR_{1-5}$, respectively, Figure 31). This leads to a "bent" of the flexible alkyl moiety adjacent to the phenyl moiety of compound **6**. BM ECL has been only observed in $S1PR_2$, $S1PR_5$, and marginally in $S1PR_1$. Therefore, further analysis will focus on $S1PR_2$ and $S1PR_5$. The BM ECL sub-pocket is fully conserved between $S1PR_2$ and $S1PR_5$, consisting of $F^{3.33}$, $Y^{5.39}$, $V^{5.40}$, $L^{6.55}$, Y^{189}, and V^{185}. The residues $F^{3.33}$, $Y^{5.39}$, $Y^{198/186/192/200/185}$, and $V^{194/182/198/196/185}$ ($S1PR_{1-5}$, respectively) are fully conserved between subtypes (Figure 22). $S1PR_2$ and $S1PR_5$ express $V^{5.40}$ while the other subtypes express the bulkier amino acid $I^{5.40}$. The residue at BWN 5.40 is surrounded during the unbound ligand state by the other residues of the BM ECL sub-pocket. To enable ligand binding, the residue 5.40 needs to unbind from the energetically favorable lipophilic environment opening up the sub-pocket. The larger lipophilic contact area of the side chain of $I^{5.40}$ compared to $V^{5.40}$ within the sub-pockets lipophilic environment could explain why BM ECL states are only observed in $S1PR_2$ and $S1PR_5$.

4.1.6. Binding Mode Occurrences in Simulations of Compound 6-$S1PR_{1-5}$ Complexes

To quantify the total occurrence of each binding mode while excluding transitional states between each binding mode a set of rules have been established: BM1, BM2, and BM ECL are determined by lipophilic contacts formed by the alkyl moiety of compound **6** to specific residues of each sub-pocket detected by Dynophore. The comparability between subtypes is ensured by selecting residue positions between subtypes based on the same BWN. To be classified as a BM1 conformation a lipophilic contact has to be established to either $C^{5.43}$ and/or $T^{5.44}$ for $S1PR_1$ ($S1PR_2$ $V^{5.43}/V^{5.44}$, S1PR3 $C^{5.43}/I^{5.44}$, $S1PR_4$ $C^{5.43}/L^{5.44}$, $S1PR_5$ $C^{5.43}/V^{5.44}$). BM2

classification requires a lipophilic contact to $V^{3.40}$ for $S1PR_1$ ($S1PR_2$ $V^{3.40}$, $S1PR_3$ $T^{3.40}$, $S1PR_4$ $T^{3.40}$, $S1PR_5$ $V^{3.40}$). A lipophilic contact to Y^{198} for $S1PR_1$ ($S1PR_2$ Y^{186}, $S1PR_3$ Y^{192}, $S1PR_4$ Y^{200}, $S1PR_5$ Y^{189}) determines the conformation as BM ECL. Transitional conformations of BM1, BM2, and BM ECL are excluded from the total occurrence if lipophilic contacts are formed simultaneously to the specified key residues of any other binding mode.

BM1 and BM2 are found consistently during MD simulations of all S1PR subtypes, while BM ECL conformations are only formed by $S1PR_2$ and $S1PR_5$ (Table 5). $S1PR_3$ exhibits with 67.9% the highest total BM1 occurrence, followed by $S1PR_2$ with 46.9%, $S1PR_4$ with 41.4%, $S1PR_5$ with 39.1%, and $S1PR_1$ with 22.0%. The BM1 occurrence relation between subtypes $S1PR_1$, $S1PR_3$, $S1PR_4$, and $S1PR_5$ is correlated to the measured EC_{50}/E_{max} values of compound **6** (EC_{50} $S1PR_1$-$S1PR_5$ in nM: 7.9, 18, 630, 160, 17; E_{max} based on 100% S1P for $S1PR_{1-5}$: 0.91, 0.95, 0.21, 0.87, 0.66 [216]). This could be explained with the assumption that BM1 represents an inactive conformation backed by four arguments: (I) BM1 resembles the binding mode of ML056 co-crystallized with $S1PR_1$ (PDB 3V2Y), (II) ML056 is an antagonist, (III) the crystal structure of $S1PR_1$ represents an inactive receptor state, and therefore (IV) the homology models also represent an inactive state.

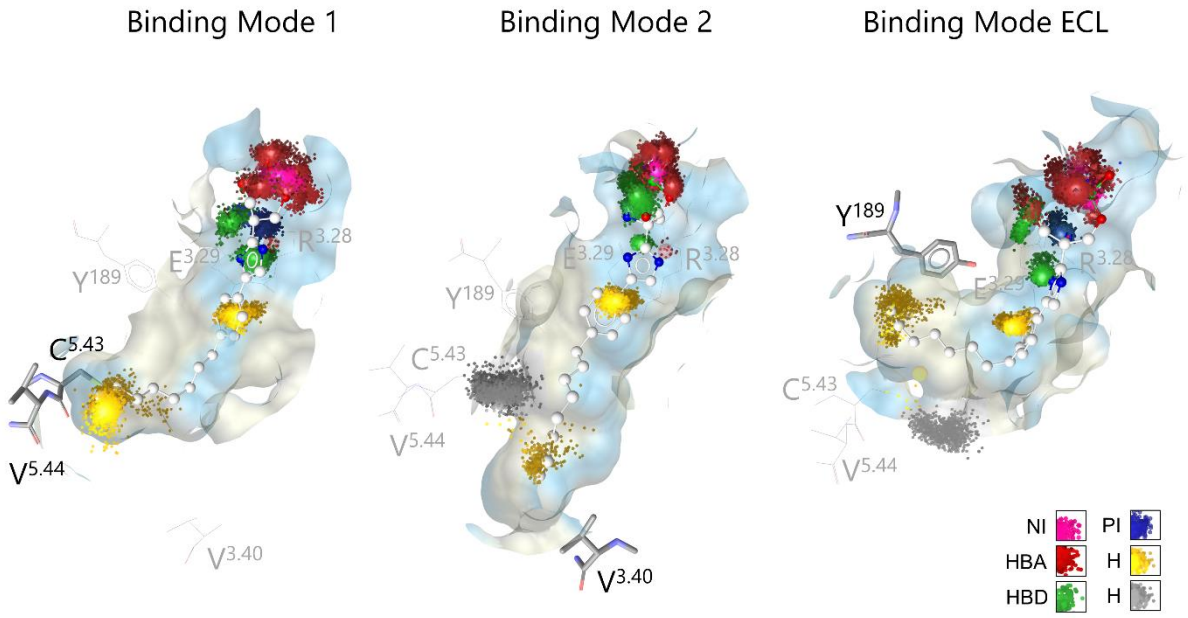


Figure 31. **Representative $S1PR_5$ Dynophores representing three binding modes observed during MD simulations of $S1PR_{1-5}$. Binding Mode 1** occurrence is determined via lipophilic contacts to $S1PR_5$ $C^{5.43}/V^{5.44}$ ($S1PR_1$ $C^{5.43}/T^{5.44}$, $S1PR_2$ $V^{5.43}/V^{5.44}$, $S1PR_3$ $C^{5.43}/I^{5.44}$, $S1PR_4$ $C^{5.43}/L^{5.44}$), **Binding Mode 2** via contacts to $S1PR_5$ $V^{3.40}$ ($S1PR_1$ $V^{3.40}$, $S1PR_2$ $V^{3.40}$, $S1PR_3$ $T^{3.40}$, $S1PR_4$ $T^{3.40}$) and **Binding Mode ECL** via contacts to $S1PR_5$ Y^{189} ($S1PR_1$ Y^{198}, $S1PR_2$ Y^{186}, $S1PR_3$ Y^{192}, $S1PR_4$ Y^{200}). HBA – hydrogen bond acceptor, HBD – hydrogen bond donor, H – lipophilic contact, NI – negative ionizable feature, PI – positive ionizable feature, yellow cloud – lipophilic contact of binding mode, grey cloud – lipophilic contact of other binding mode, blue surface - hydrophilic environment, yellow surface - lipophilic environment, Black residue names – Residue contact which determined the binding modes.

While BM1 represents the known inactive conformation of compound **6**, the implication of the newly observed BM2 conformation of compound **6** is not established yet. The highest BM2 occurrence exhibits compound **6** in $S1PR_1$ with 36.3%, followed by $S1PR_4$ with 23.1%, $S1PR_2$ with 18.9%, $S1PR_3$ with 15.2%, and $S1PR_5$ with 4.3%. The BM2 occurrences of $S1PR_1$, $S1PR_3$, and $S1PR_4$ are inversely correlated with the measured activity of compound **6**. Compound **6** in complex with $S1PR_5$ shows a low total occurrence of BM2 even though compound **6** has similar EC_{50} values for $S1PR_2$ and $S1PR_5$ (18 nM and 17 nM, respectively). Moreover, BM2 occurrences of $S1PR_4$ are higher than $S1PR_2$, which does not support the

measured activity of compound **6** (18 nM and 160 nM, respectively). A closer look at the complex conformations with simultaneously occurring key residue interactions of BM1 ($V^{5.43}$/$V^{5.44}$) and BM2 ($V^{3.40}$) for $S1PR_2$ revealed that in $S1PR_2$ compound **6** alkyl moiety preferably establishes lipophilic contacts with $V^{5.43}$ instead of $V^{5.44}$, due to BWN position 5.43 protruding farther into the lipophilic sub pocket of BM1. This results in conformations with shared key residue interactions of BM1 and BM2 (5.7%) leading to no categorization to any of the two binding modes. This phenomenon is unique to $S1PR_2$, which is the only subtype expressing $V^{5.43}$ instead of the less lipophilic $C^{5.43}$ and could explain the difference in BM2 occurrences between $S1PR_2$ and $S1PR_4$. Moreover, the additional consideration of $S1PR_2$ and $S1PR_5$ exclusive BM ECL occurrences (2.1 and 16.9%, respectively) as another possible active conformation of compound **6** could explain the difference of $S1PR_2$ and $S1PR_5$ BM2 occurrences compared to higher $S1PR_4$ BM2 occurrences. This leads to a better alignment of the presented *in silico* data with the experimental data.

Table 5. **Occurrence of distinct compound 6 binding modes during MD simulations in percent.**

	Binding Mode 1	*Binding Mode 2*	*Binding Mode ECL*
$S1PR_1$	22.0%	36.3%	0.2%
$S1PR_2$	46.9%	18.9%	2.1%
$S1PR_3$	67.9%	15.2%	0.0%
$S1PR_4$	41.4%	23.1%	0.0%
$S1PR_5$	39.1%	4.3%	16.9%

ECL - Extracellular Loop

4.1.7. Markov Chain Models Elucidate S1PR Binding Mode Occurrences of Compound 6

To explore the sequential relation between BM1, BM2 and BM ECL for $S1PR_{1-5}$ a stochastic model based on discrete-time Markov chains has been chosen. The subtypes initial data set consists of multiple sequences of MD simulation frames labeled with BM1, BM2, or BM ECL (binding mode definitions are described in chapter 4.1.6, page 52). If frames do not satisfy any of the three binding modes classification rules they are labeled as "Transitional Binding Mode" (T-BM). The T-BM can be defined as the "negative space" between BM1, BM2, and BM ECL conformations. One key component of Markov chains is the Markov property, which states that the probability of the next state is only dependent on the current state. Therefore, the state transition probability from e.g., BM1 to BM2 is calculated by dividing the number of observed BM1 to BM2 transitions by the total number of BM1 occurrences. All observed state transitions with corresponding probabilities are summarized in state transition diagrams for each subtype. The combined transition probability of one state is always 100%.

The state transition diagram of $S1PR_1$ reveals neglectable (<0.1%) inter-state transition probabilities between BM1, BM2, and BM ECL (Figure 32). Probabilities for recurrence of the same state are the highest for BM2 (90.8%) followed by BM1 (85.5%) and BM ECL (72.4%). Consequently, BM ECL has the highest probability to transition to T-BM followed by BM1 and BM2 (27.6%, 14.5%, and 9.2%, respectively). The transition probability from T-BM to BM2, BM1, or BM ECL (8.1%, 7.7%, and 0.1%, respectively) shows the same trend as the probabilities for recurring states. These results explain the binding mode occurrences of BM1, BM2, and BM ECL (22.0%, 36.3%, and 0.2%). Due to neglectable transition probabilities to BM ECL and high transition probabilities from BM ECL, the total BM ECL occurrence can be interpreted as neglectable in $S1PR_1$. High recurrence and transition probabilities to BM2 lead to the highest occurrence between the three binding modes.

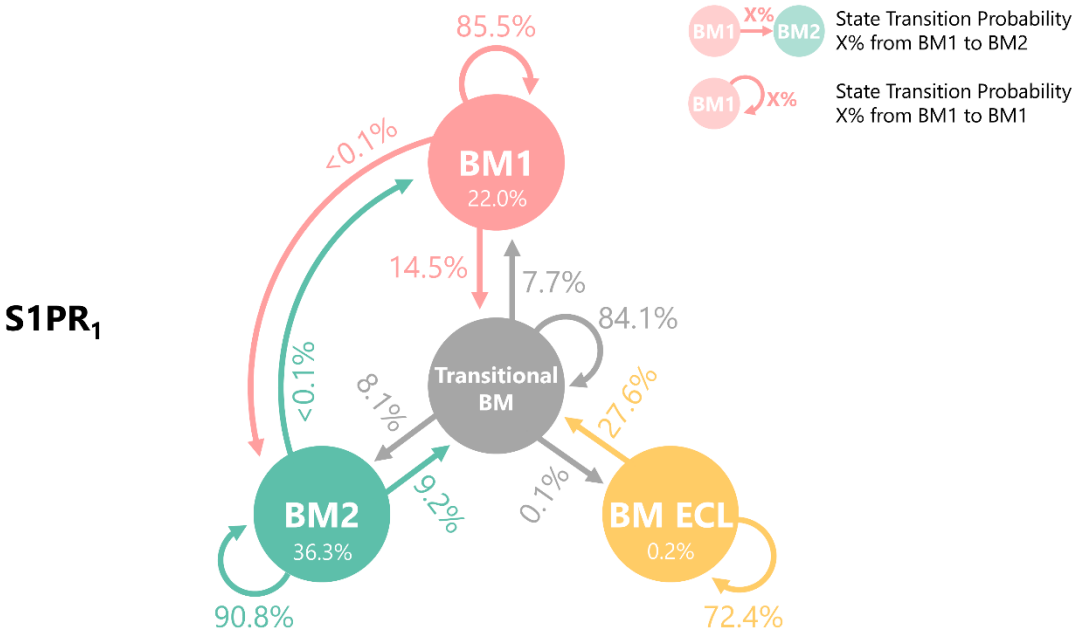


Figure 32. **Markov Chain State Transition Diagram of S1PR$_1$.** Based on transition probabilities from Appendix Table 3. Occurrences for each binding mode are shown in percent (white). BM1 – Binding Mode 1, BM2 – Binding Mode 2, BM ECL – Binding Mode Extracellular Loop, Transitional BM – Transitional Binding Mode

The BM1 to BM2 or BM ECL inter-state transition probabilities of S1PR$_2$ range between 0.3% and 9.1% (Figure 33), which are the highest among all subtypes. The reason for this is the BM1 categorization of S1PR$_2$ frames based on mostly V$^{5.43}$ instead of 5.44 for all other subtypes. The lipophilic residue V$^{5.43}$ is found only in S1RP$_2$ (C$^{5.43}$ for other subtypes) leading to interactions with BM2 or BM ECL key residues simultaneously with BM1'. This results in direct transitions between the binding modes. Recurrence probabilities are the highest for BM1 followed by BM2 and BM ECL (80.9%, 62.1%, and 41.2%, respectively). Binding mode transition probabilities to T-BM are complementary to the recurrence probabilities with BM ECL showing the highest followed by BM2 and BM1 (53.6%, 28.8%, and 15.0%, respectively). Interestingly, the transition probabilities from T-BM to BM1 are higher compared *versa* (22.2% vs. 15.0%), while T-BM transition probabilities to BM2 and BM ECL are lower than *versa* (16.6% vs. 28.8% and 3.5% vs. 53.6%, respectively). These results are in line with the highest occurrences is BM1 in S1PR$_2$.

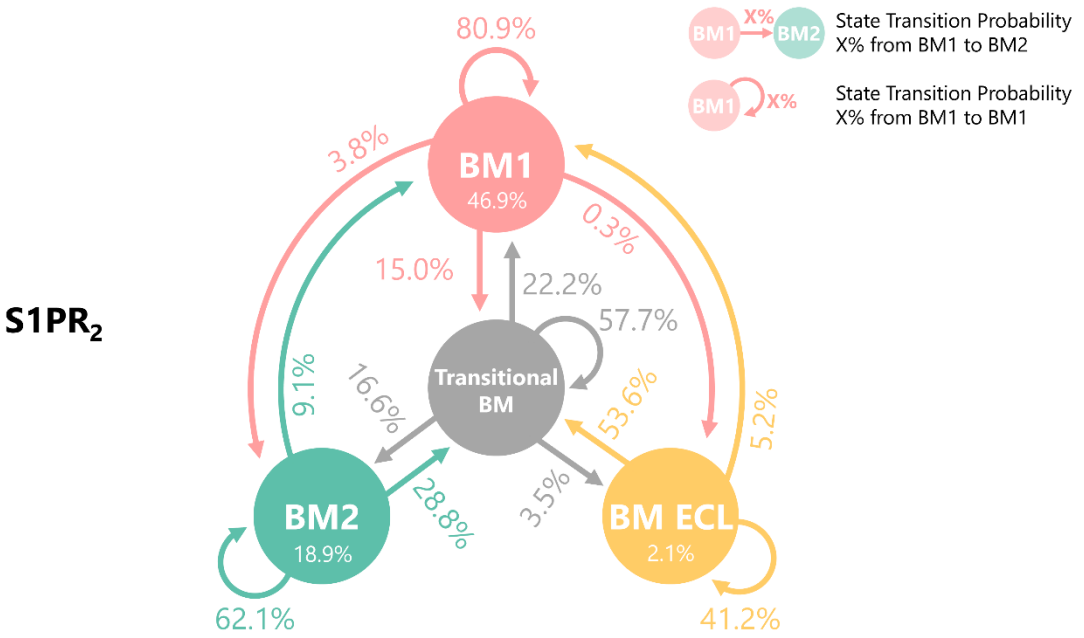


Figure 33. **Markov Chain State Transition Diagram of $S1PR_2$.** Based on transition probabilities from Appendix Table 4. Occurrences for each binding mode are shown in percent (white). BM1 – Binding Mode 1, BM2 – Binding Mode 2, BM ECL – Binding Mode Extracellular Loop, Transitional BM – Transitional Binding Mode

$S1PR_3$ does not form any BM ECL conformations, therefore no transition probabilities from or to BM ECL exist (Figure 34). Inter-state transition probabilities from BM1 to BM2 or *versa* are neglectable (0.1% and 0.5%, respectively). Even though the recurring state transition probabilities of BM1 are only slightly higher than BM2 (95.7% and 91.3%, respectively) the transition probability of T-BM to BM1 is more than double the transition probability of T-BM to BM2 (17.0% and 7.3%, respectively). Together with higher transition probabilities from BM2 to T-BM compared to BM1 to T-BM (8.2% and 4.2%, respectively), the presented data is in line with $S1PR_3$ binding mode occurrences, which show that BM1 is more frequent with 67.9% than BM2 with 15.2%.

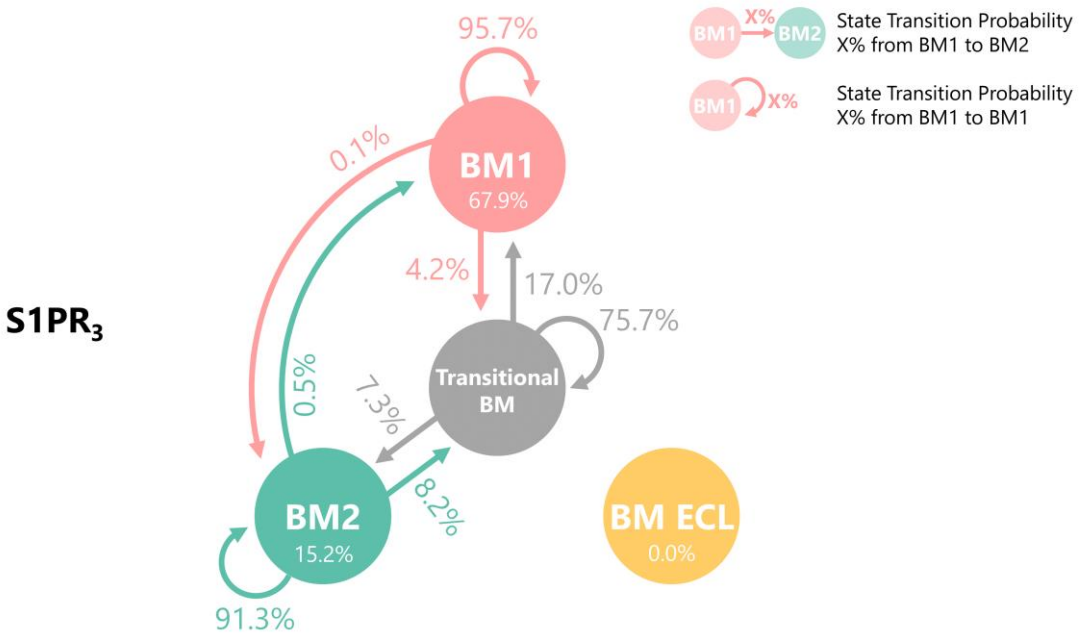


Figure 34. **Markov Chain State Transition Diagram of $S1PR_3$.** Based on transition probabilities from Appendix Table 5. Occurrences for each binding mode are shown in percent (white). BM1 – Binding Mode 1, BM2 – Binding Mode 2, BM ECL – Binding Mode Extracellular Loop, Transitional BM – Transitional Binding Mode

$S1PR_4$ expresses low neglectable inter-state transition probabilities between BM1 and BM2 (0.3% and 0.6%, respectively, Figure 35). Furthermore, no BM ECL conformations have been observed. Recurring state probabilities are higher for BM1 compared to BM2 (87.3% vs. 81.4%). Transition probabilities to T-BM are therefore higher from BM2 compared to BM1 (18.0% vs. 12.4%). Adding the higher T-BM to BM1 probability compared to BM2', the overall probability for BM1 is higher than BM2, which is also reflected by the occurrences.

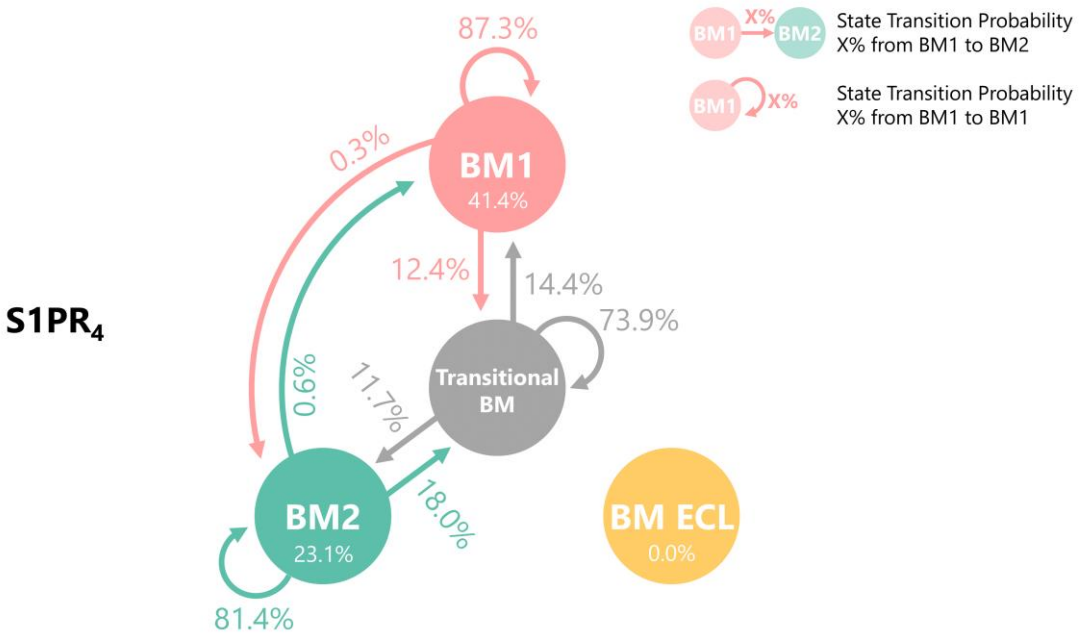


Figure 35. **Markov Chain State Transition Diagram of $S1PR_4$.** Based on transition probabilities from Appendix Table 6. Occurrences for each binding mode are shown in percent (white). BM1 – Binding Mode 1, BM2 – Binding Mode 2, BM ECL – Binding Mode Extracellular Loop, Transitional BM – Transitional Binding Mode

Lastly, in $S1PR_5$ all binding modes are observed with low inter-state transition probabilities between BM1, BM2, and BM ECL (<0.1% - 1.4%, Figure 36). Recurring state transition probabilities are the highest for BM ECL followed by BM1 and BM2 (97.7%, 79.7%, and 64.9%, respectively). Consequential, the transition probabilities to T-BM are the highest from BM2 followed by BM1 and BM ECL (33.7%, 20.1%, and 2.2%, respectively). Interestingly, BM ECL has a total occurrence during MD simulations of 16.9%. Even though only low transition probabilities to BM ECL are observed, a very high recurring state transition probability suggests very stable BM ECL conformations. BM2 on the other hand has low recurring transition probabilities and high transition probabilities to T-BM, which is reflected by the low total BM2 occurrence rate of 4.3%.

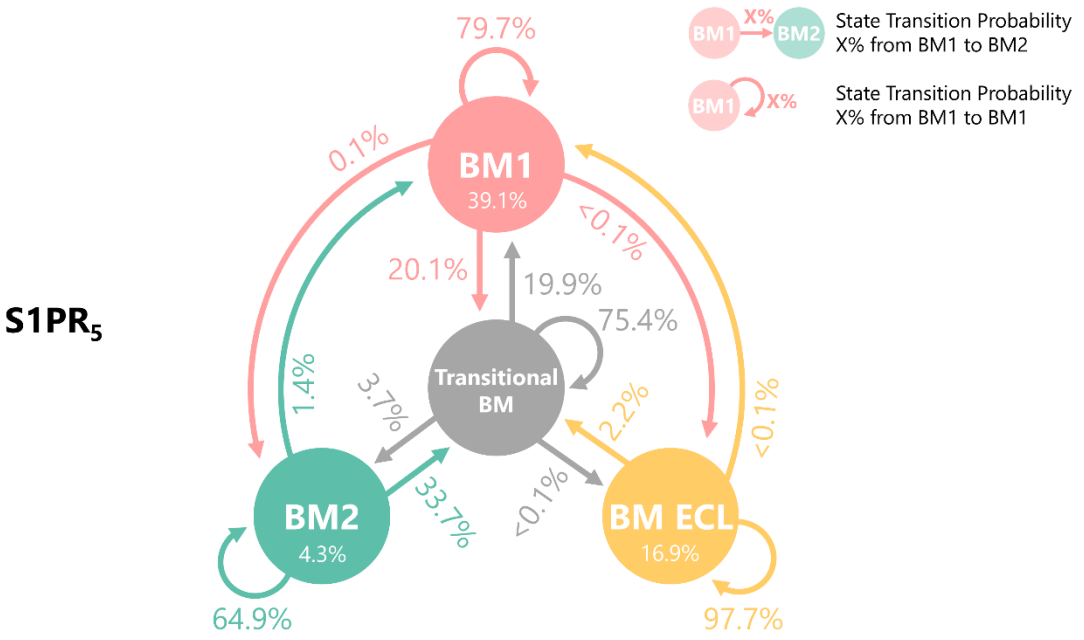


Figure 36. **Markov Chain State Transition Diagram of S1PR$_5$.** Based on transition probabilities from Appendix Table 7. Occurrences for each binding mode are shown in percent (white). BM1 – Binding Mode 1, BM2 – Binding Mode 2, BM ECL – Binding Mode Extracellular Loop, Transitional BM – Transitional Binding Mode

In general, no S1PR subtype expresses inter-state transitions from BM2 to BM ECL. This can be explained by the localization of BM ECL's lipophilic sub-pocket near the extracellular membrane part of the receptor, while BM2 is located in the core part of the receptor. Therefore, transitions to BM ECL are only possible either with a low percentage from BM1 or through T-BM. The Markov chain state transition diagrams of S1PR$_3$ and S1PR$_4$ do not show state transition probabilities to BM ECL, whereas S1PR$_1$, S1PR$_2$, and S1PR$_5$ express state transitions to BM ECL. The state transition probabilities to BM ECL are comparably low between S1PR$_1$ and S1PR$_5$, but S1PR$_1$ has a twelve times higher state transition probability from BM ECL to T-BM. S1PR$_2$ expresses higher transition probabilities to BM ECL (3.8% in total) but shows the lowest recurring state transition probability (41.2%) of all S1PR$_{1\text{-}5}$ BM states leading to only 2.1% BM ECL occurrences in MD simulations. The starting conformation of all S1PR$_{1\text{-}5}$ in complex with compound **6** is BM1 due to the initial docking conformations with consecutive MD simulations. Based on this starting conformation we can calculate the transition probability of the pathway sequence from BM1 to BM ECL with

a recurring BM ECL state (BM1 → T-BM → BM ECL → BM ECL), which is the product of the transition probabilities of said sequence. $S1PR_1$ has a probability of 0.01%, $S1PR_2$ 0.22%, and $S1PR_5$ 17.67%. This illustrates the much higher probability of $S1PR_5$ compared to $S1PR_1$ and $S1PR_2$ to reach BM ECL from BM1 and maintain this state, explaining the different BM ECL occurrences. The same method can be applied for sequences from BM1 to BM2 with a recurring BM2 state (i: BM1 → T-BM → BM2 → BM2, ii: BM1 → BM2 → BM2) in which we can compare all five subtypes with one another. $S1PR_1$ shows the highest combined probability of 10.69% (i: 10.66%, ii: 0.03%) followed by $S1PR_2$ with 3.91% (i: 1.55%, ii: 2.36%), $S1PR_4$ with 1.43% (i: 1.18%, ii: 0.25%), $S1PR_5$ with 0.49% (i: 0.48%, ii: 0.09%), and $S1PR_3$ with 0.41% (i: 0.28%, ii: 0.13%). Even though $S1PR_4$ shows higher BM2 occurrences than $S1PR_2$, the state transition probability of the BM1 to BM2 pathway sequences is higher for $S1PR_2$ due to direct transitions from BM1 to BM2. These results support the high potency of compound **6** in $S1PR_2$.

4.1.8. Dynamic Binding Mode 1 Comparison between $S1PR_{1-5}$ Subtypes Reveals Subtype-specific Interaction Patterns of Compound 6 Alkyl Moiety

The visual inspection of BM1 dynophores reveals in $S1PR_{1-5}$ comparable dynamic polar superfeatures and one lipophilic superfeature formed by the phenyl moiety of compound **6** (Figure 37). These results are in line with the discussed stable occurrences of dynamic interactions in chapter 4.1.3. Major differences are found between the $S1PR_{1-5}$ lipophilic superfeature formed by the alkyl moiety of compound **6**, revealing subtype-specific spatiotemporal shapes of the dynamic interaction patterns. $S1PR_1$ and $S1PR_3$ form lipophilic superfeatures with spherical shapes. While lipophilic interaction in $S1PR_5$ is "bean" shaped, $S1PR_2$'s shape can be described as a "half-moon". Finally, $S1PR_4$' lipophilic superfeature is "crescent moon" shaped from the site but shows a wide diameter circular form when turned 90° upwards. It can therefore be described as "plate" shaped. Due to the other S1PR superfeatures exhibiting stable and dense spherical shapes, the subtype-specific patterns of the alkyl moiety displayed can be directly correlated to the lipophilic binding pocket differences between $S1PR_{1-5}$. To uncover S1PR subtype-specific residue interaction patterns

of compound **6** during BM1 we analyzed occurring BM1 residue interaction clusters (simultaneously occurring lipophilic contacts between the lipophilic tail and residues based on calculated dynophore raw data).

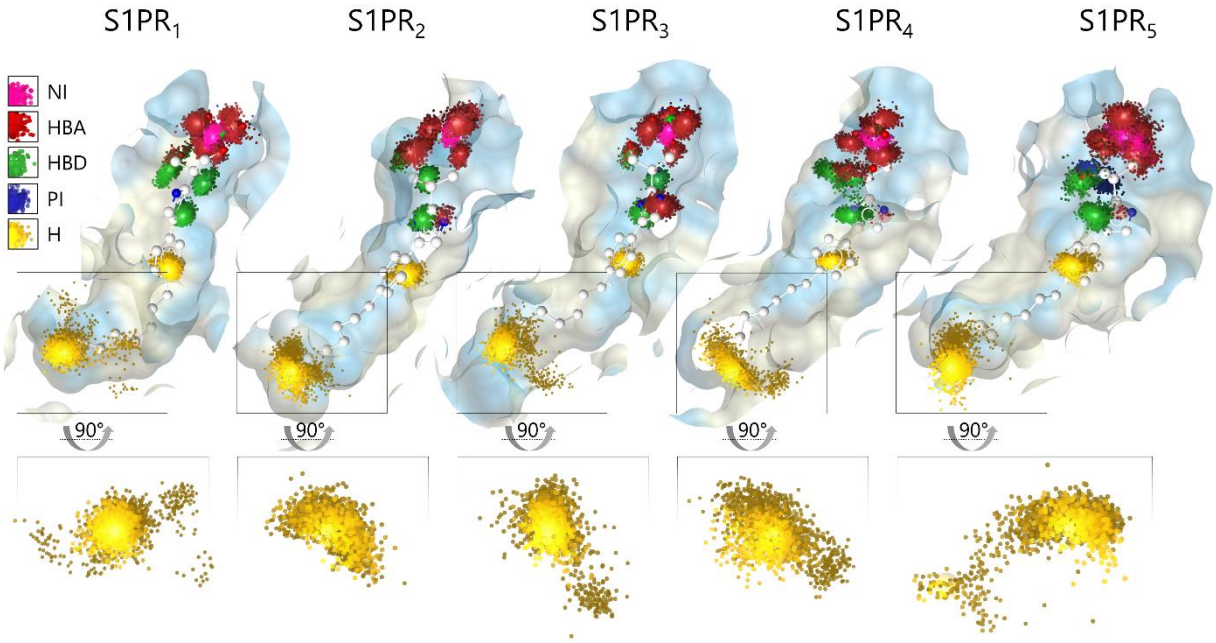


Figure 37. **Binding mode 1 dynophores reveal distinct lipophilic superfeature shapes between S1PR$_{1\text{-}5}$.** HBA – hydrogen bond acceptor, HBD – hydrogen bond donor, H – lipophilic contact, NI – negative ionizable feature, PI – positive ionizable feature, Blue surface - hydrophilic environment, Yellow surface - lipophilic environment.

Compound **6** alkyl moiety in complex with S1PR$_1$ in BM1 mainly interacts with TM5 and TM6 residues T$^{5.44}$, F$^{6.52}$, L$^{6.55}$, F$^{5.47}$, T$^{5.48}$, W$^{6.48}$, and L$^{6.51}$ (100.0%, 82.0%, 66.2%, 44.5%, 44.2%, 43.6%, and 28.3%, respectively). Furthermore, minor interactions with TM3 residues F$^{3.33}$ and L$^{3.36}$ are occurring (9.4% and 5.1%, respectively, Figure 38). The most frequent BM1 residue interaction cluster consists of T$^{5.44}$, F$^{5.47}$, F$^{6.52}$, and L$^{6.55}$ occurring 14.3% of the total BM1 count followed by the same cluster with an additional interaction to W$^{6.48}$ occurring 6.7% of the time. The third most occurring cluster consists of T$^{5.44}$, T$^{5.48}$, W$^{6.48}$, and F$^{6.52}$ (5.2% of total BM1 count). The highest occurring cluster which includes F$^{3.33}$ also additionally consists of the highest occurring clusters T$^{5.44}$, F$^{5.47}$, F$^{6.52}$, and L$^{6.55}$ (3.4% of total BM1 count), while the highest

cluster including $L^{3.36}$ consists of $T^{5.44}$ and $T^{5.48}$ (2.2% of total BM1 count). Interaction clusters with $T^{5.44}$, $F^{5.47}$, $F^{6.52}$, and $L^{6.55}$ as their core interactions with alternating additional residue interactions occur 33.4% of total BM1 count.

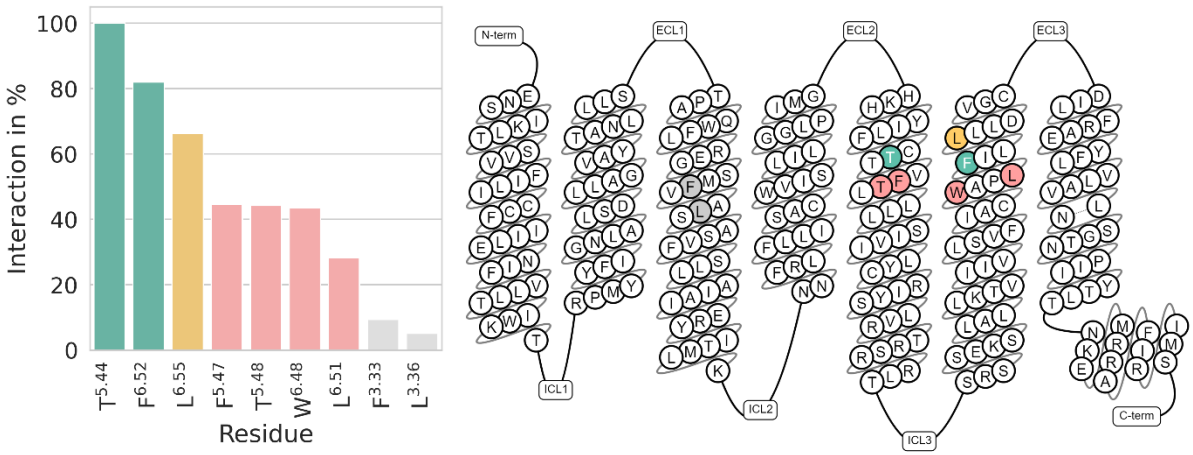


Figure 38. **BM1 $S1PR_1$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 8. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of $S1PR_1$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

Compound **6** alkyl moiety in complex with $S1PR_2$ in BM1 mainly interacts with TM3, TM4, and TM5 residues $V^{5.43}$, $L^{4.56}$, $I^{5.46}$, $F^{5.47}$, $F^{3.33}$, $I^{5.50}$, and $L^{3.36}$ (98.9%, 87.1%, 50.0%, 48.9%, 41.4%, 17.3%, and 14.8%, respectively). The most frequent BM1 residue interaction cluster consists of $L^{4.56}$, $V^{5.43}$, and $I^{5.46}$ occurring 16.2% of the total BM1 count followed by the $L^{4.56}$, $V^{5.43}$, and $F^{5.47}$ cluster occurring 11.3% of the time. The third most occurring cluster consists also of $L^{4.56}$, $V^{5.43}$, and $F^{5.47}$ with additional interactions to $F^{3.33}$ (10.2% of total BM1 count). The highest occurring cluster which includes $I^{5.50}$ additionally consists of $F^{3.33}$, $L^{4.56}$, $V^{5.43}$, and $I^{5.46}$ (3.5% of total BM1 count), while the highest cluster which includes $L^{3.36}$ consists of $F^{3.33}$, $L^{4.56}$, $V^{5.43}$, and $F^{5.47}$ (3.7% of total BM1 count). Interaction clusters with $L^{4.56}$, $V^{5.43}$, and $I^{5.46}$ as their core interactions with alternating additional residue interactions occur 46.0%. Interestingly, $V^{5.43}$

is the main interaction occurring during BM1, but 1.1% of total BM1 occurrence is via $V^{5.44}$ interactions and 1.2% of the time via $V^{5.43}$ and $V^{5.44}$ interactions.

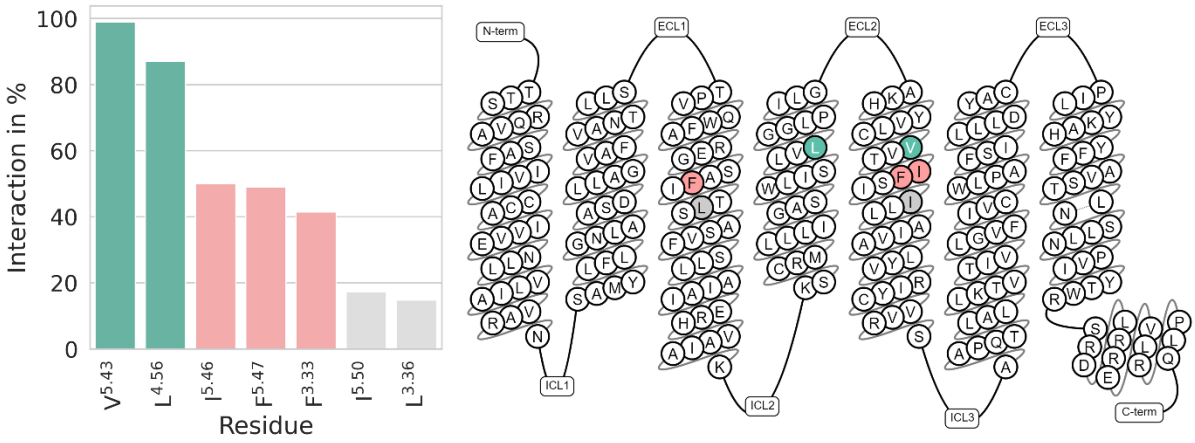


Figure 39. **BM1 $S1PR_2$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 8. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of $S1PR_2$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

Compound **6** alkyl moiety in complex with $S1PR_2$ in binding mode 1 mainly interacts with TM5 and TM6 residue $I^{5.44}$, $F^{6.55}$, $F^{6.52}$, $F^{5.47}$, $T^{5.48}$, $I^{5.40}$, $L^{6.56}$, and $L^{6.51}$ (100.0%, 66.5%, 66.2%, 40.5%, 28.6%, 21.2%, 12.5%, and 28.3%, respectively). Furthermore, interactions with TM3 residue $F^{3.33}$ is frequently occurring (21.5%, Figure 40). The most frequent BM1 residue interaction cluster consists of $F^{3.33}$, $I^{5.40}$, $I^{5.44}$, and $F^{6.55}$ occurring 17.2% of the total BM1 count followed by the interaction cluster $I^{5.44}$, $T^{5.48}$, $F^{6.52}$, and $F^{6.55}$ occurring 11.4% of the time. The third most occurring cluster consists of $I^{5.44}$, $F^{5.47}$, $F^{6.52}$, and $F^{6.55}$ also occurring 11.4% of the total BM1 count. The highest occurring cluster which includes $L^{6.56}$ additionally consists of $I^{5.44}$, $F^{5.47}$, and $F^{6.52}$ (4.9% of total BM1 count), while the highest cluster which includes $L^{6.51}$ also consists of $I^{5.44}$, $F^{5.47}$, and $F^{6.52}$ (3.2% of total BM1 count). Interaction clusters with $F^{3.33}$, $I^{5.40}$, $I^{5.44}$, and $F^{6.55}$ as their core interactions with alternating additional residue interactions occur with 19.7% frequency of total BM1 count.

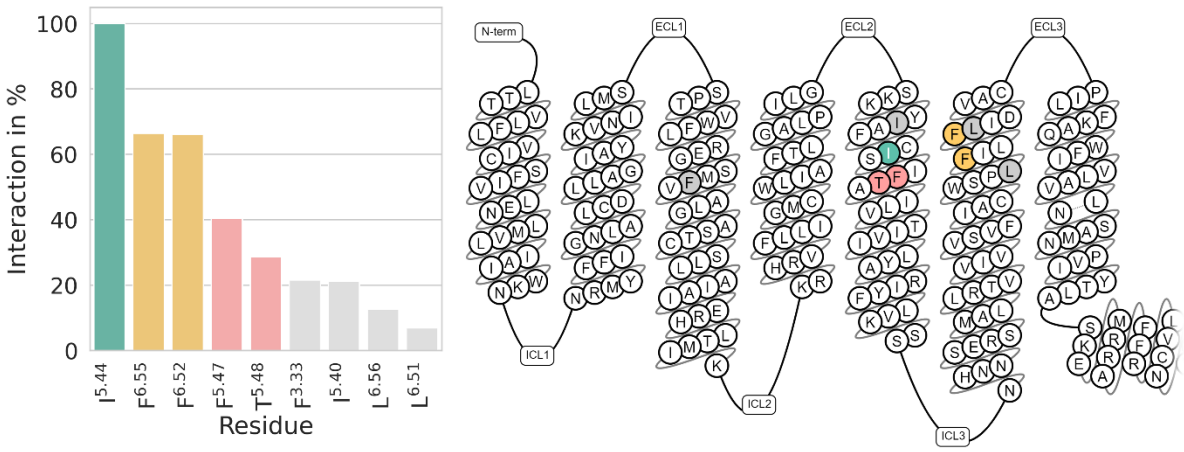


Figure 40. **BM1 S1PR$_3$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 8. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of S1PR$_3$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

Compound **6** alkyl moiety in complex with S1PR$_4$ in binding mode 1 mainly interacts with TM3, TM5, and TM6 residues L$^{5.44}$, F$^{5.47}$, L$^{6.55}$, F$^{6.52}$, F$^{3.33}$, I$^{5.40}$, L$^{3.36}$, and W$^{6.48}$ (100.0%, 88.9%, 66.7%, 63.3%, 30.7%, 26.7%, 21.6, and 13.7%, respectively, Figure 41). The most frequent BM1 residue interaction cluster consists of L$^{5.44}$, F$^{5.47}$, F$^{6.52}$, and L$^{6.55}$ occurring 15.0% of total BM1 count followed by the F$^{3.33}$, L$^{5.44}$, F$^{5.47}$, and F$^{6.52}$ cluster occurring 7.5% of the time. The third most occurring cluster consists of L$^{5.44}$, F$^{5.47}$, and F$^{6.52}$ (7.0% of total BM1 count). The highest occurring cluster which includes interactions to L$^{3.36}$ additionally consists of F$^{3.33}$, L$^{4.56}$, V$^{5.43}$, and I$^{5.46}$ (3.5% of total BM1 count), while the highest cluster to include W$^{6.48}$ consists of L$^{5.44}$, F$^{5.47}$, and L$^{6.55}$ (1.9% of total BM1 count). Interaction clusters with L$^{5.44}$, F$^{5.47}$, and F$^{6.52}$ as their core interactions with alternating additional residue interactions occur 56.7% of total BM1 count.

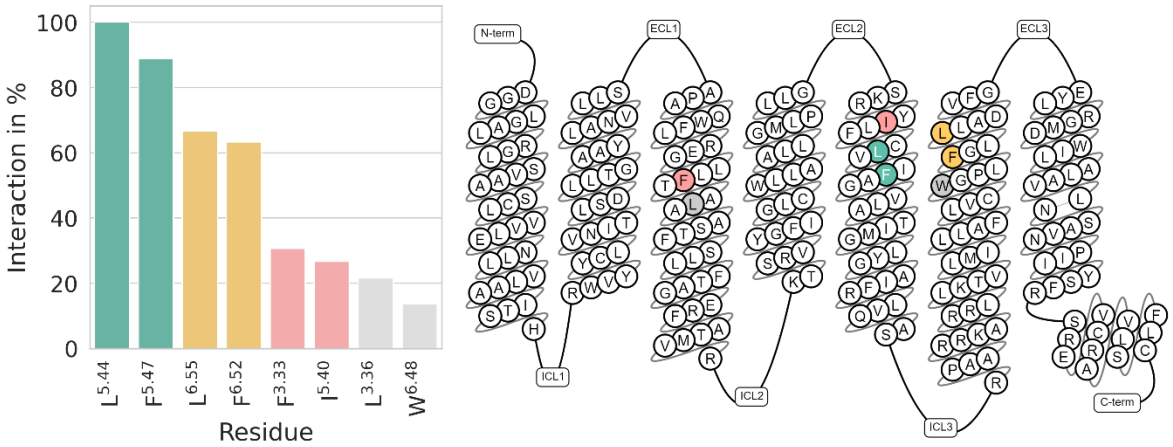


Figure 41. **BM1 $S1PR_4$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 8. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of $S1PR_4$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

Compound **6** alkyl moiety in complex with $S1PR_5$ in BM1 mainly interacts with TM3, TM5, and TM6 residues $V^{5.44}$, $F^{5.47}$, $L^{6.55}$, $F^{6.52}$, $F^{3.33}$, $L^{3.36}$, $W^{6.48}$, and $Y^{5.39}$ (100.0%, 82.8%, 59.1%, 48.0%, 29.4%, 23.5%, 6.7%, and 6.1%, respectively, Figure 42). The most frequent BM1 residue interaction cluster consists of $V^{5.44}$, $F^{5.47}$, and $L^{6.55}$ occurring 16.0% of the total BM1 count followed by the same cluster with an additional interaction to $F^{6.52}$ occurring 15.0% of the time. The third most occurring cluster consists of $F^{3.33}$, $V^{5.44}$, $F^{5.47}$, and $L^{6.55}$ (6.5% of total BM1 count). The highest occurring cluster which includes interactions to $L^{3.36}$ additionally consists of $V^{5.44}$ (5.2% of total BM1 count), while the highest cluster to include $W^{6.48}$ consists of $V^{5.44}$, $F^{5.47}$, and $F^{6.52}$ (1.3% of total BM1 count). Interactions with $Y^{5.39}$, $V^{5.44}$, and $F^{5.47}$ also occur 1.3% of the time. Interaction clusters with $V^{5.44}$, $F^{5.47}$, and $L^{6.55}$ as their core interactions with alternating additional residue interactions occur 49.5% of total BM1 count.

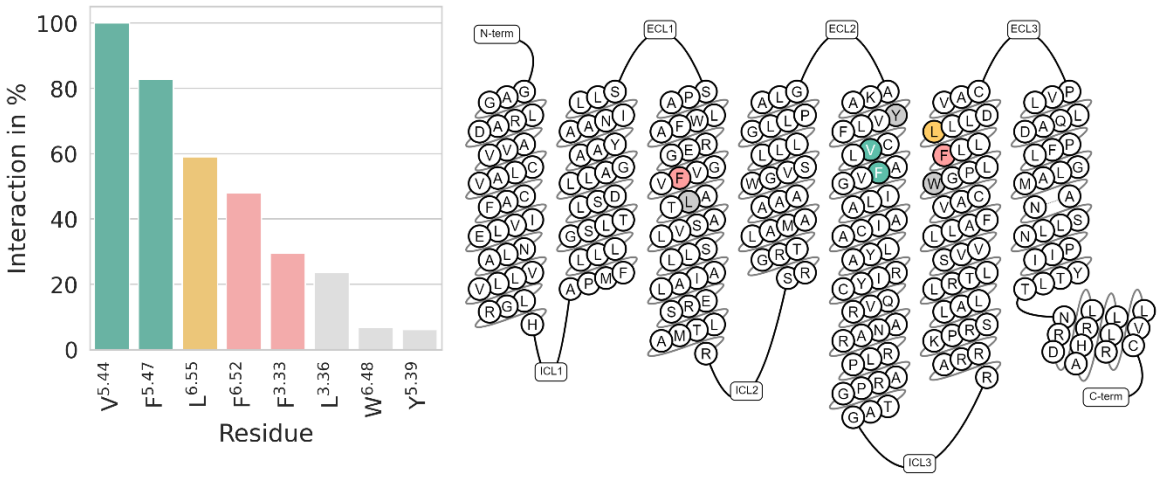


Figure 42. **BM1 S1PR$_5$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 8. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of S1PR$_5$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

Comparing all S1PR$_{1\text{-}5}$ BM1 interaction patterns reveal S1PR$_2$ BM1 with the most distinct set of lipophilic residues, showing interactions to V$^{5.43}$, I$^{5.46}$, and I$^{5.50}$ exclusively observed in S1PR$_2$ (Figure 43). Furthermore, interactions to residue L$^{4.56}$ are found in more than 80% during BM1 occurrences in S1PR$_2$ with minor interaction occurrences in S1PR$_5$ (under 5%). This is also reflected by the main interaction cluster consisting of L$^{4.56}$, V$^{5.43}$, and I$^{5.46}$ observed in S1PR$_2$ as shown in this chapter. The other subtypes reveal a comparable set of residues interacting with the lipophilic tail of compound **6**. This includes similar interaction occurrences to the conserved residue F$^{6.52}$ as well as non-conserved residues for S1PR$_1$, S1PR$_3$, S1PR$_4$, and S1PR$_5$ at BWN position 5.44 (T$^{5.44}$, I$^{5.44}$, L$^{5.44}$, and V$^{5.44}$, respectively) and 6.55 (L$^{6.55}$, F$^{6.55}$, L$^{6.55}$, and L$^{6.55}$, respectively). Interactions to conserved residue F$^{5.47}$ are made for S1PR$_{1\text{-}3}$ with around 40% but more than 80% for S1PR$_4$ and S1PR$_5$, while interactions to BWN positions 5.48, 6.48, and 6.51 are mainly observed in S1PR$_1$. Interactions to position 5.40 are only found in S1PR$_1$, S1PR$_3$, and S1PR$_4$ due to the substitution of V$^{5.40}$ with I$^{5.40}$ in S1PR$_2$ and S1PR$_5$.

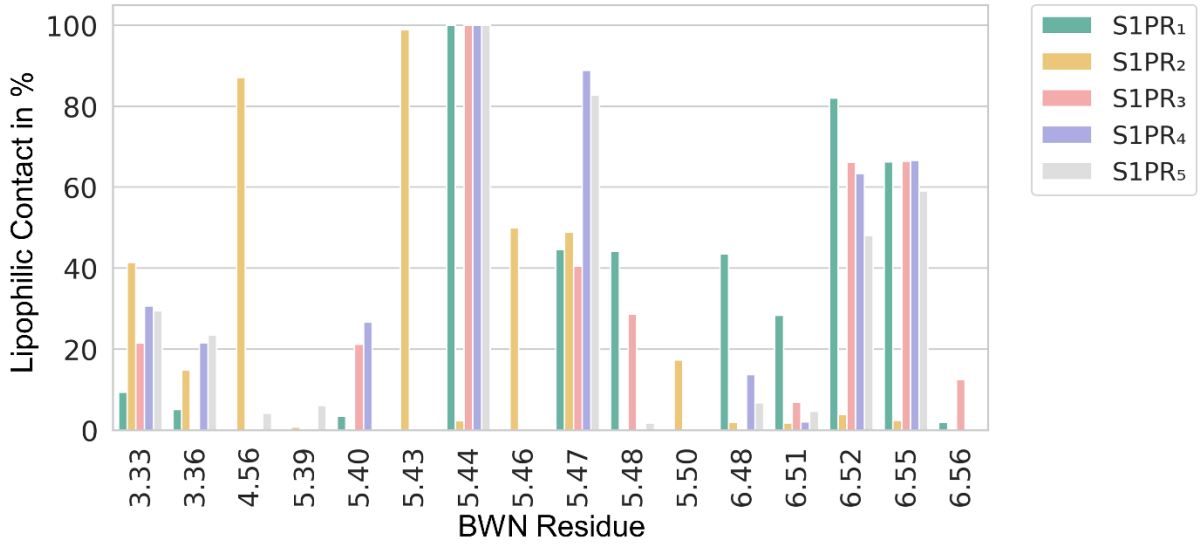


Figure 43. **Occurrence comparison of lipophilic residue interactions in percent for binding mode 1 of S1PR$_{1\text{-}5}$.** The corresponding data can be found in Appendix Table 8.

4.1.9. Dynamic Binding Mode 2 Comparison between S1PR$_{1\text{-}5}$ Subtypes Reveal Subtype-specific Interaction Patterns of Compound 6 Alkyl Moiety

The visual inspection of BM2 dynophores also reveals S1PR$_{1\text{-}5}$ comparable dynamic polar superfeatures and one lipophilic superfeature formed by the phenyl moiety of compound **6** (Figure 44) as in BM1. These results are in line with the discussed stable occurrences of dynamic interactions in chapter 4.1.3, page 47. The differences found between the S1PR$_{1\text{-}5}$ lipophilic superfeature formed by the alkyl moiety of compound **6** reveal subtype-specific dynamic interaction patterns. Comparable to S1PR$_{1\text{-}5}$ BM1's lipophilic feature, BM2's lipophilic superfeatures spatiotemporal shapes are more alike. S1PR$_1$ and S1PR$_3$ form dense lipophilic superfeatures with spherical shapes. Even though S1PR$_2$ also shows a spherical superfeature, it shows a less dense distribution in comparison to S1PR$_1$ and S1PR$_3$. S1PR$_5$'s lipophilic interaction is oval-shaped. Finally, S1PR$_4$'s lipophilic superfeature is triangular shaped. Due to the other S1PR superfeatures exhibiting durable and frequent spherical shapes over the course of the simulations, the subtype-specific patterns of the alkyl moiety displayed can be directly correlated to the lipophilic BM2 differences between S1PR$_{1\text{-}5}$. We have also analyzed the simultaneously occurring lipophilic contacts of BM2 between the

lipophilic tail of compound **6** and the $S1PR_{1-5}$ residues to reveal subtype-specific interaction patterns based on the calculated dynophore raw data.

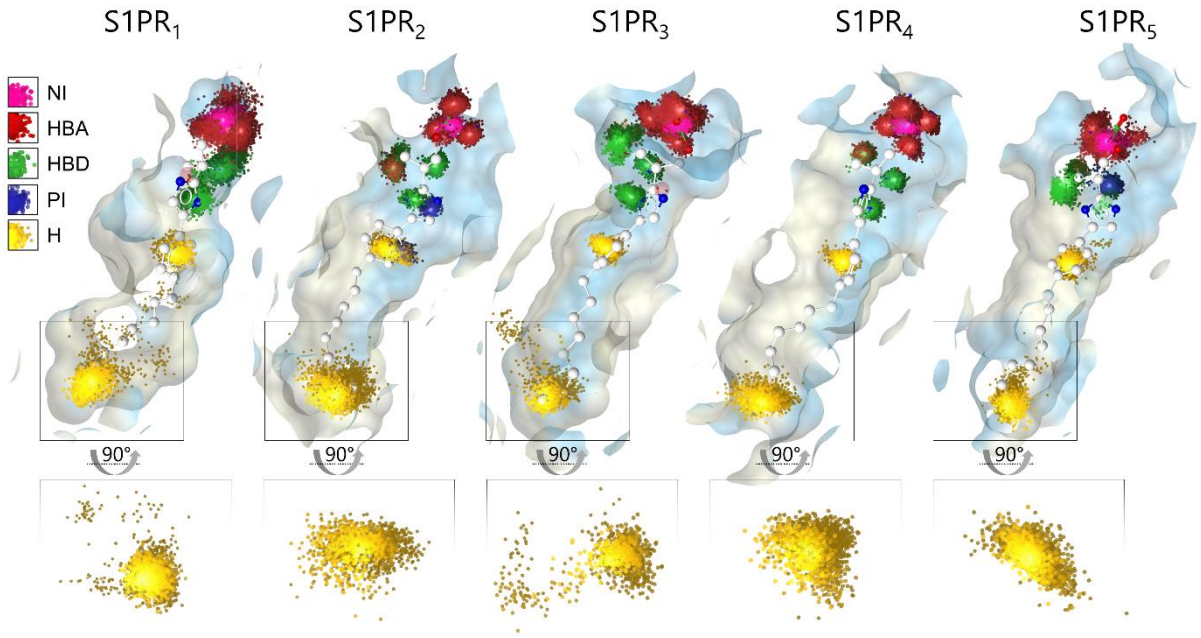


Figure 44. **Binding mode 2 dynophores reveal distinct lipophilic feature shapes between $S1PR_{1-5}$.** HBA – hydrogen bond acceptor, HBD – hydrogen bond donor, H – lipophilic contact, NI – negative ionizable feature, PI – positive ionizable feature, Blue surface - hydrophilic environment, Yellow surface - lipophilic environment.

Compound **6** alkyl moiety in complex with $S1PR_1$ in BM2 mainly interacts with TM3, TM5, and TM6 residues $V^{3.40}$, $F^{5.47}$, $F^{6.52}$, $W^{6.48}$, $L^{3.36}$, $L^{5.51}$, and $L^{6.51}$ (100.0%, 91.2%, 84.2%, 71.9%, 42.7%, 24.4%, and 11.6%, respectively, Figure 45). The most frequent BM2 residue interaction cluster consists of $L^{3.36}$, $V^{3.40}$, $F^{5.47}$, $W^{6.48}$, and $F^{6.52}$ occurring 17.0% of total BM2 count followed by the same cluster, but without interactions to $L^{3.36}$, occurring 16.8% of the time. The third most occurring cluster consists also of $V^{3.40}$, $F^{5.47}$, $L^{5.51}$, $W^{6.48}$, $F^{6.52}$ (10.2% of total BM2 count). The highest occurring cluster which includes $L^{6.51}$ additionally consists of $V^{3.40}$, $F^{5.47}$, $W^{6.48}$, and $F^{6.52}$ (4.1% of total BM2 count). Interaction clusters with $V^{3.40}$, $F^{5.47}$, $W^{6.48}$, and $F^{6.52}$ as their core interactions with alternating additional residue interactions which occur 54.6% of total BM2 count.

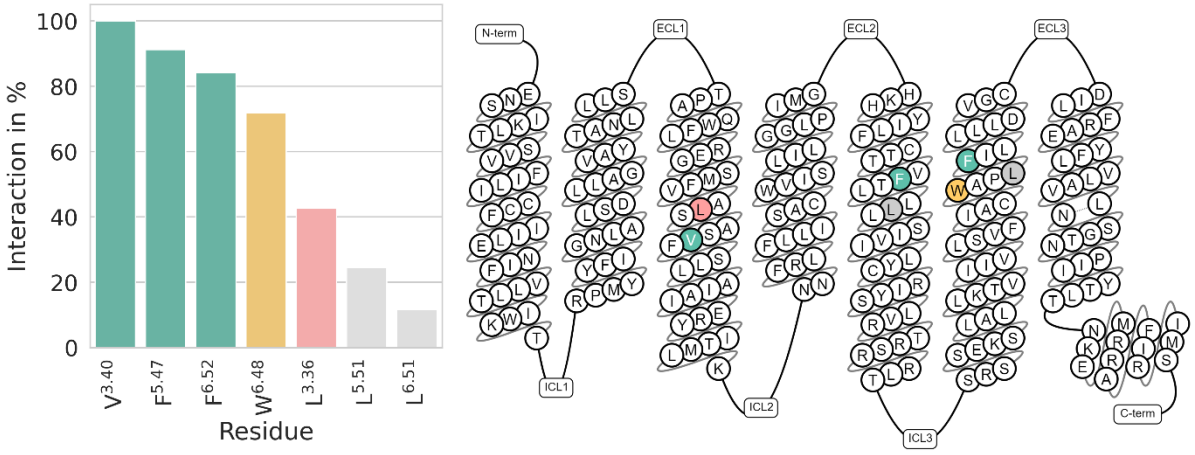


Figure 45. **BM2 $S1PR_1$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 9. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of $S1PR_1$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

Compound **6** alkyl moiety in complex with $S1PR_2$ in BM2 mainly interacts with TM3, TM4, and TM5 residues $V^{3.40}$, $F^{5.47}$, $I^{5.46}$, $L^{5.51}$, $L^{3.36}$, $L^{4.56}$, $I^{5.55}$, $F^{3.41}$, and $F^{3.33}$ (100.0%, 52.5%, 37.5%, 35.4%, 34.4%, 29.7%, 12.1%, 12.0%, and 6.3%, respectively). Moreover, interactions with TM6 residues $F^{6.44}$, $I^{6.45}$, $F^{6.52}$, and $W^{6.48}$ occur (23.5%, 22.5%, 15.0%, and 11.4%, respectively, Figure 46). The most frequent BM2 residue interaction cluster consists of $V^{3.40}$, $L^{4.56}$, and $I^{5.46}$ occurring 8.3% of the total BM2 count followed by the same cluster with an additional interaction to $F^{5.47}$ occurring 4.7% of the time. The third most occurring cluster consists of $V^{3.40}$ and $L^{5.51}$ (4.6% of total BM2 count). Residue interaction clusters containing TM6 residue interactions are under 2.5% of the total BM2 count each. Interaction clusters with $V^{3.40}$, $L^{4.56}$, and $I^{5.46}$ as their core interactions with alternating additional residue interactions occur 28.1% of total BM2 count.

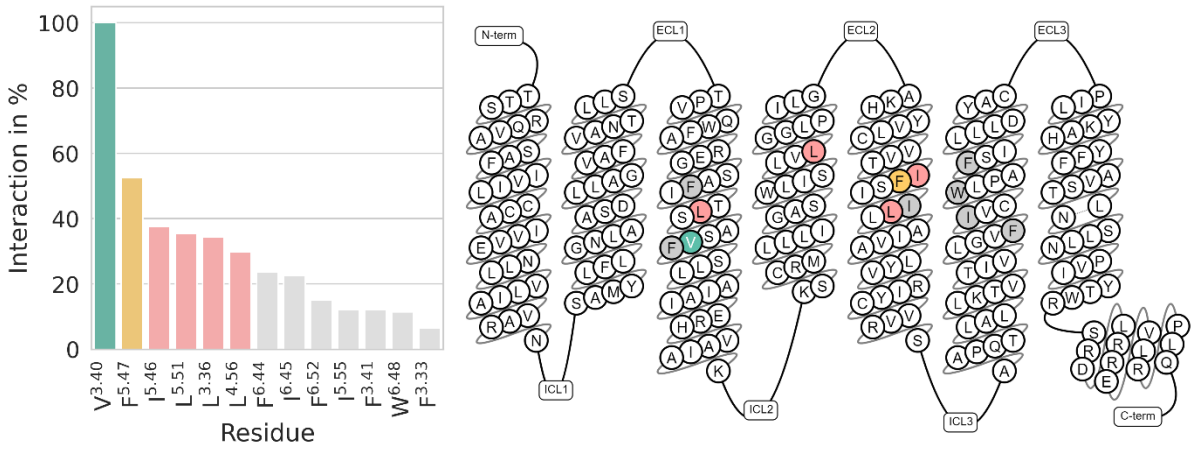


Figure 46. **BM2 S1PR$_2$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 9. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of S1PR$_2$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

Compound **6** alkyl moiety in complex with S1PR$_3$ in BM2 mainly interacts with TM3, TM5, and TM6 residues T$^{3.40}$, F$^{5.47}$, L$^{5.51}$, F$^{6.52}$, W$^{6.48}$, I$^{5.46}$, and L$^{3.36}$ (100.0%, 80.5%, 58.8%, 57.1%, 36.1%, 17.0%, and 11.6%, respectively). Furthermore, interactions with TM4 residue L$^{4.56}$ have been observed (19.2%, Figure 47). The most frequent BM2 residue interaction cluster consists of T$^{3.40}$, F$^{5.47}$, L$^{5.51}$, W$^{6.48}$, and F$^{6.52}$ occurring 21.3% of total BM2 count followed by the same cluster, but without interactions to W$^{6.48}$, occurring 21.2% of the time. The third most occurring cluster consists of lipophilic interactions to V$^{3.40}$ and F$^{5.47}$ (8.4% of total BM2 count). The highest occurring cluster which includes I$^{5.46}$ additionally consists of V$^{3.40}$ and L$^{4.56}$ (6.1% of total BM2 count), while the largest cluster including L$^{3.36}$ also consists of T$^{3.40}$, F$^{5.47}$, L$^{5.51}$, W$^{6.48}$, and F$^{6.52}$ (3.4% of total BM2 count). Interaction clusters with T$^{3.40}$, F$^{5.47}$, L$^{5.51}$, and F$^{6.52}$ as their core interactions with alternating additional residue interactions occur 45.9% of total BM2 occurrences.

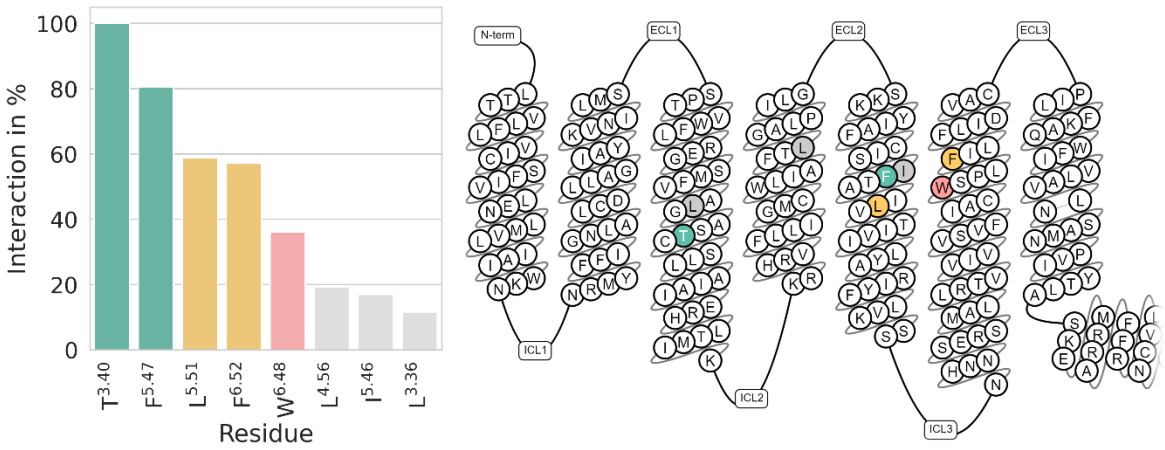


Figure 47. **BM2 S1PR$_3$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 9. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of S1PR$_3$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

Compound **6** alkyl moiety in complex with S1PR$_4$ in BM2 mainly interacts with TM3, TM5, and TM6 residues T$^{3.40}$, F$^{5.47}$, L$^{3.36}$, A$^{3.37}$, I$^{5.46}$, F$^{6.52}$, W$^{6.48}$, L$^{6.51}$, L$^{6.55}$, and F$^{3.41}$ (100.0%, 82.5%, 41.4%, 36.6%, 33.0%, 27.4%, 18.3%, 16.6%, 8.8%, 7.4%, 5.3%, respectively). Moreover, minor interactions with TM4 residue L$^{4.56}$ occur (7.4%, Figure 48). The most frequent BM2 residue interaction cluster consists of A$^{3.37}$, T$^{3.40}$, I$^{5.46}$, and F$^{5.47}$ occurring 12.1% of the total BM2 count followed by the cluster with interaction to L$^{3.36}$, T$^{3.40}$, and F$^{5.47}$ occurring 11.3% of the time. The third most occurring cluster consists of T$^{3.40}$, F$^{5.47}$, and F$^{6.52}$ (5.1% of total BM2 count). Residue interaction clusters containing TM4 L$^{4.56}$ residue interactions also consist of A$^{3.37}$, T$^{3.40}$, and I$^{5.46}$ (3.4% total BM2 count). The biggest cluster with W$^{6.48}$ is accompanied by interactions with T$^{3.40}$ and F$^{5.47}$ occurring 3.6% percent of total BM2 count, while the interaction cluster with L$^{6.51}$ occurs with interactions to L$^{3.36}$, T$^{3.40}$, and F$^{5.47}$ 4.0% of the time. The highest occurrence of interaction clusters with L$^{6.55}$ or F$^{3.41}$ is around 2%. Interactions with A$^{3.37}$, T$^{3.40}$, I$^{5.46}$, and F$^{5.47}$ as their core interactions with alternating additional residue interactions occur 16.5% of total BM2 count.

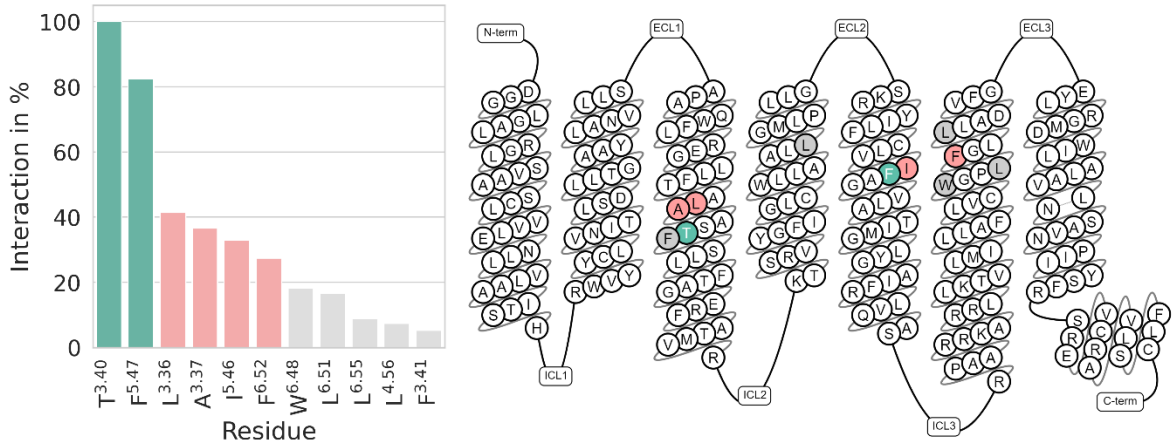


Figure 48. **BM2 S1PR$_4$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 9. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of S1PR$_4$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

Compound **6** alkyl moiety in complex with S1PR$_5$ in BM2 mainly interacts with TM3, TM5, and TM6 residues V$^{3.40}$, T$^{3.37}$, F$^{5.47}$, A$^{5.46}$, I$^{5.50}$, W$^{6.48}$, and F$^{6.52}$ (100.0%, 91.5%, 59.9%, 52.4%, 51.5%, 32.0%, and 19.4%, respectively). Furthermore, interactions with TM4 residue L$^{4.56}$ have been observed (32.3%, Figure 49). The most frequent BM2 residue interaction cluster consists of T$^{3.37}$, V$^{3.40}$, L$^{4.56}$, A$^{5.46}$, and I$^{5.50}$ occurring 24.8% of total BM2 count followed by the same cluster, but without interactions to L$^{4.56}$, occurring 12.2% of the time. The third most occurring cluster consists of lipophilic interactions to T$^{3.37}$, V$^{3.40}$, F$^{5.47}$, W$^{6.48}$, and F$^{6.52}$ (11.2% of total BM2 count). Interaction clusters with T$^{3.37}$, V$^{3.40}$, A$^{5.46}$, and I$^{5.50}$ as their core interactions with alternating additional residue interactions occur 50.5% of total BM2 occurrences.

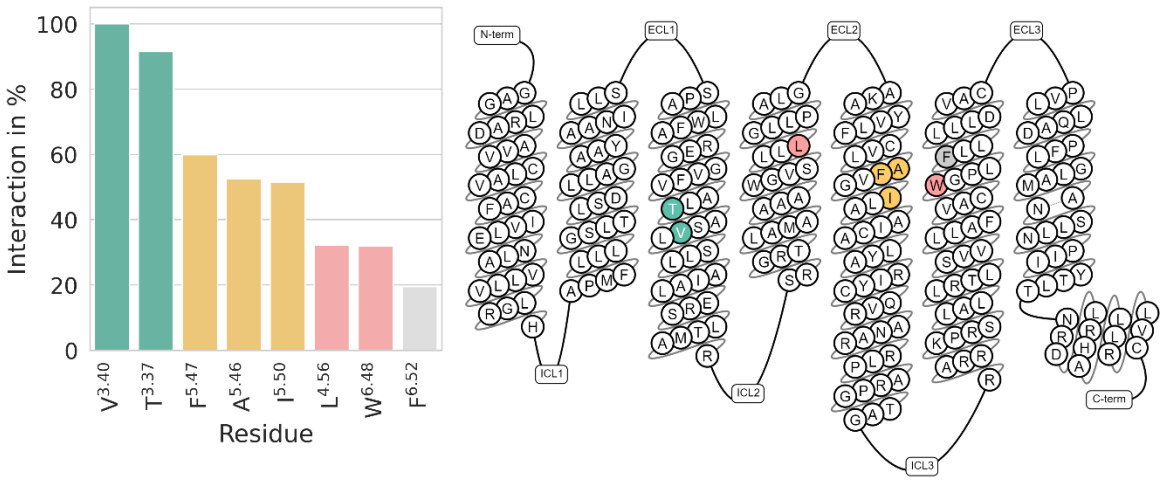


Figure 49. **BM2 $S1PR_5$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 9. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of $S1PR_5$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

By comparing the different BM2 interaction patterns of $S1PR_{1\text{-}5}$ we can deduce that $S1PR_2$ shows the most distinct lipophilic residue contacts (Figure 50), which interestingly was also the case for BM1 (chapter 4.1.8, Figure 43). Compound **6** lipophilic tail interactions in $S1PR_2$ BM2 express exclusive interactions to $F^{3.33}$, $F^{6.44}$, and $I^{6.45}$. The residues $F^{6.44}$ and $I^{6.45}$ are located farther to the intracellular domain of the receptor than other interactions observed with TM6 residues, leading to BM2 conformations of compound **6** in $S1PR_2$ reaching the deepest into the receptor core compared to other subtypes. On the other hand, interaction occurrences to residues, besides BWN position 3.40, which is used to define BM2 conformations, are only observed at around 50% maximal during BM2 conformations in $S1PR_2$. Moreover, the number of different residue interactions over 5% during BM2 is the highest for $S1PR_2$ with 13 different residues, leading to diverse interaction clusters with the biggest combined cluster amounting to 28.1% (core interactions to $V^{3.40}$, $L^{4.56}$, and $I^{5.46}$). This is also reflected in the lipophilic superfeature cloud of $S1PR_2$ showing the largest lipophilic point distribution. $S1PR_4$ follows with 11 different residue interactions during BM2, but without expressing exclusive interactions as seen in $S1PR_2$. $S1PR_4$ shows an even lower recurrence of interaction

clusters amounting to only 16.5% (core interactions to $A^{3.37}$, $T^{3.40}$, $I^{5.46}$, and $F^{5.47}$). The other subtypes $S1PR_1$, $S1PR_3$, and $S1PR_5$ show a smaller and more comparable set of interactions with recurring clusters ranging between 45.9% and 54.6%. The most comparable interaction poses the conserved residue $F^{5.47}$ between all subtypes. $F^{5.47}$ is located directly between BM1 and BM2, therefore acting as a common denominator between the two binding modes. Interestingly, interactions to conserved residue $W^{6.48}$ and $F^{6.52}$ are observed the most for $S1PR_1$ followed by $S1PR_3$.

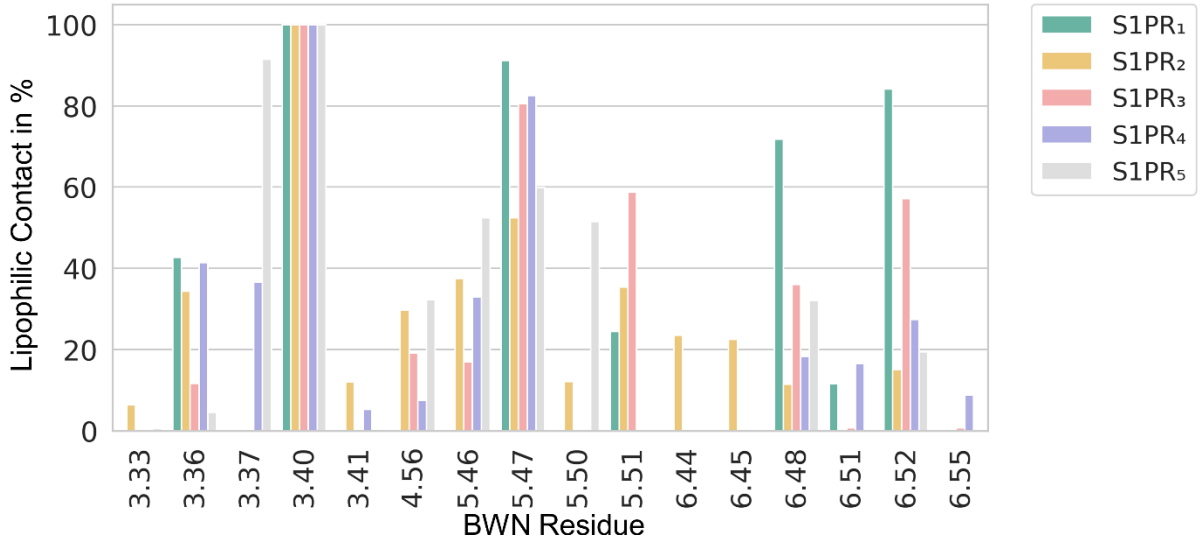


Figure 50. **Occurrence of lipophilic residue interactions in percent for BM2 of $S1PR_{1\text{-}5}$.** Corresponding data can be found in Appendix Table 9.

4.1.10. Dynamic Binding Mode ECL Comparison between $S1PR_2$ and $S1PR_5$ in Complex with Compound 6

The visual inspection of BM ECL dynophores reveals in $S1PR_2$ and $S1PR_5$ comparable dynamic polar superfeatures and one lipophilic superfeature formed by the phenyl moiety of compound **6** (Figure 51). These results are in line with the discussed frequent occurrences of dynamic interactions in chapter 4.1.3, page 47. Major differences are the aromatic superfeatures only observed between the phenyl and imidazole moiety and $F^{7.39}$ of $S1PR_2$. Surprisingly, the formed lipophilic interaction by the alkyl moiety of compound **6** reveals

subtype-specific dynamic interaction patterns, even though the binding pocket of BM ECL is fully conserved between $S1PR_2$ and $S1PR_5$. $S1PR_2$ forms a lipophilic superfeature with a "bean" shape from the side. When turned 90° upwards the superfeature shows a circular shape with a wide diameter. $S1PR_5$ on the other hand forms a "banana" shape.

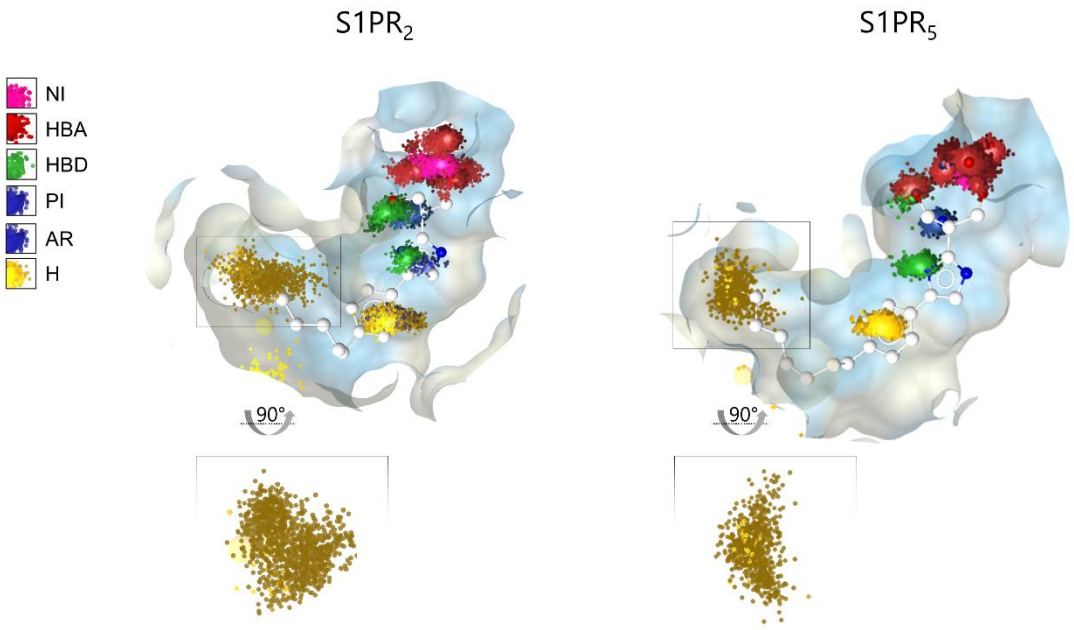


Figure 51. **Binding mode ECL dynophores reveal distinct lipophilic feature shapes between $S1PR_2$ and $S1PR_5$.** Aromatic interactions are only observed for $S1PR_2$ shown in blue dots at the imidazole and phenyl moiety. AR – aromatic interaction, HBA – hydrogen bond acceptor, HBD – hydrogen bond donor, H – lipophilic contact, NI – negative ionizable feature, PI – positive ionizable feature, Blue surface - hydrophilic environment, Yellow surface - lipophilic environment.

Compound **6** alkyl moiety in complex with $S1PR_2$ in BM ECL mainly interacts with TM3, ECL2, and TM5 residues Y^{186}, L^{183}, $F^{3.33}$, $Y^{5.39}$, and $A^{3.32}$ (100.0%, 95.8%, 59.8%, 38.2%, and 8.2%, respectively). Moreover, minor interactions with TM6 residue $L^{6.55}$ occur 16.3% of the total BM ECL occurrence count (Figure 52). The most frequent BM ECL residue interaction cluster consists of L^{183}, Y^{186}, $F^{3.33}$, and $Y^{5.39}$ occurring 25.8% of the total BM ECL count followed by the same cluster without interactions to $Y^{5.39}$ occurring 22.9% of the time. The third most

occurring cluster consists of L^{183} and Y^{186} (16.0% of total BM ECL count). The most observed residue interaction cluster containing TM6 residue $L^{6.55}$ or $A^{3.32}$ also consists of L^{183} and Y^{186} occurring 9.8% with $L^{6.55}$ and 4.2% with $A^{3.32}$ of the total BM ECL count. Interaction clusters with L^{183}, Y^{186}, and $F^{3.33}$ as their core interactions with alternating additional residue interactions occur 55.6% of the total BM ECL count.

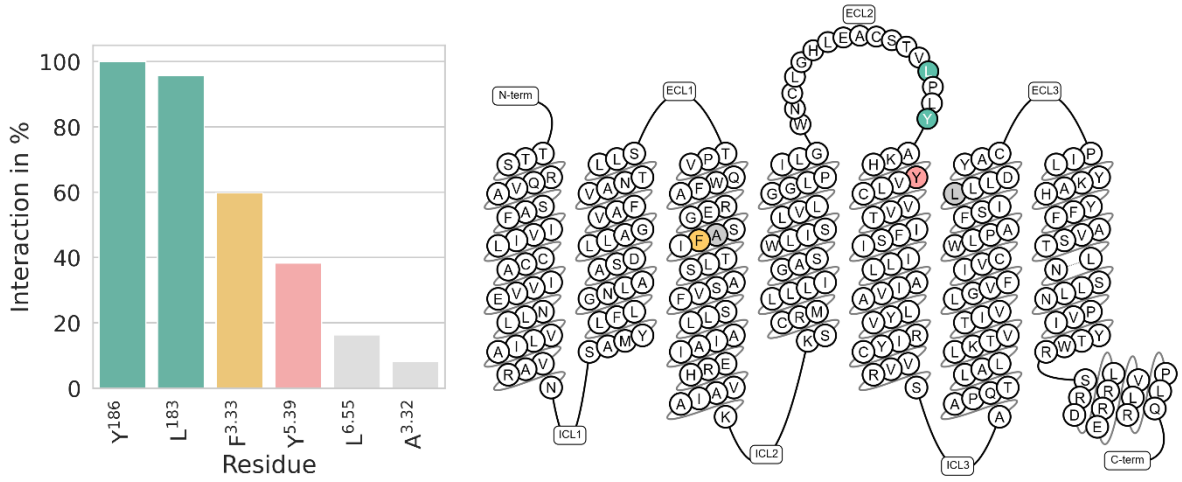


Figure 52. **BM ECL $S1PR_2$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 10. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of $S1PR_2$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

Compound **6** alkyl moiety in complex with $S1PR_5$ in BM ECL mainly interacts TM3, ECL2, and TM5 residues Y^{189}, $Y^{5.39}$, $F^{3.33}$, $V^{5.40}$, L^{186}, and V^{185} (100.0%, 92.8%, 69.8%, 56.7%, 24.4%, and 8.3%, respectively). Furthermore, interactions with TM6 residue $L^{6.55}$ have been observed (23.0%, Figure 53). The most frequent BM ECL residue interaction cluster consists of $F^{3.33}$, Y^{189}, and $Y^{5.39}$ occurring 38.1% of the total BM ECL count followed by the same cluster including interactions with $V^{5.40}$, occurring 20.8% of the time. The third most occurring cluster consists of lipophilic interactions to $F^{3.33}$, Y^{189}, $Y^{5.39}$, and $V^{5.40}$ (5.3% of total BM ECL count). Interaction

clusters with $F^{3.33}$, Y^{189}, and $Y^{5.39}$ as their core interactions with alternating additional residue interactions occur 68.9% of total BM ECL occurrences.

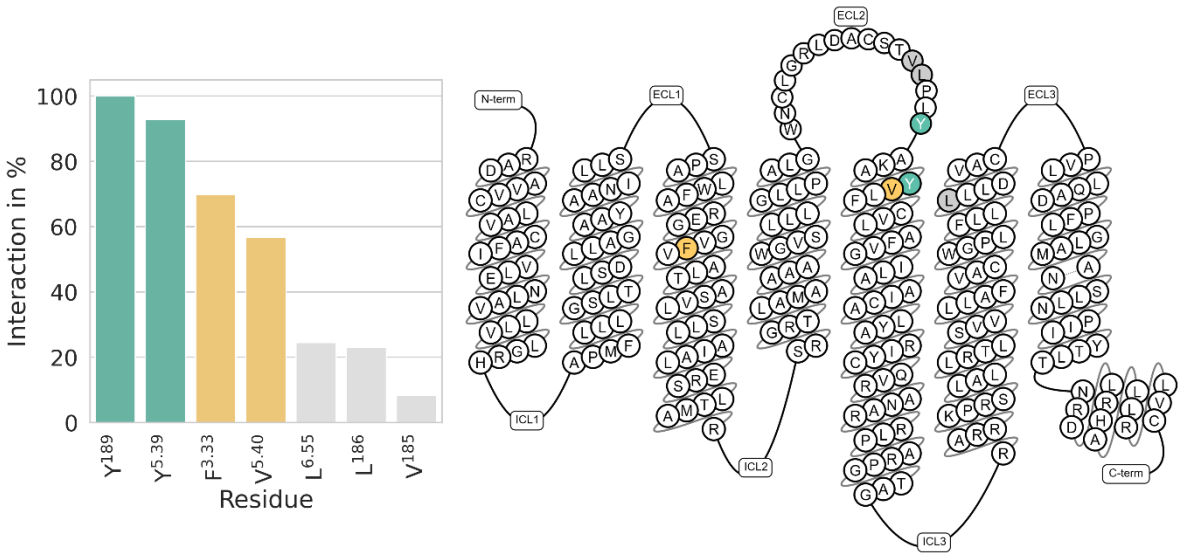


Figure 53. **BM ECL $S1PR_5$ residue interaction occurrence of compound 6' lipophilic tail.** Interaction occurrences under 5% have not been considered. The corresponding data can be found in Appendix Table 10. **Left:** Barplot of residue interactions in percent. **Right:** Snakeplot derived from gpcrdb.com of $S1PR_2$ with colored residues. Green – 75% to 100% occurrence, yellow – 50% to under 75% occurrence, red – 25% to under 50% occurrence, grey – under 25% occurrence.

Comparisons between $S1PR_2$ and $S1PR_5$ compound **6** BM ECL conformations reveal for both subtypes exclusive residue interactions. While $S1PR_2$ expresses minor exclusive interactions to $A^{3.32}$, $S1PR_5$ exclusively interacts with V^{185} and $V^{5.40}$ (Figure 54). Major interaction occurrences between both subtypes are also observed for $L^{183/186}$ ($S1PR_2$ with 95.8% and $S1PR_5$ with 23.0% of total BM ECL count, respectively) and $Y^{5.39}$ ($S1PR_2$ with 38.2% and $S1PR_5$ with 92.8% of total BM ECL count). Even though the residues of the lipophilic sub pocket of BM ECL are fully conserved between $S1PR_2$ and $S1PR_5$ we observed major differences in shaped lipophilic superfeatures and different interaction patterns. This might be due to additional aromatic interactions of compound **6'** phenyl moiety with residue $F^{7.39}$ of $S1PR_2$.

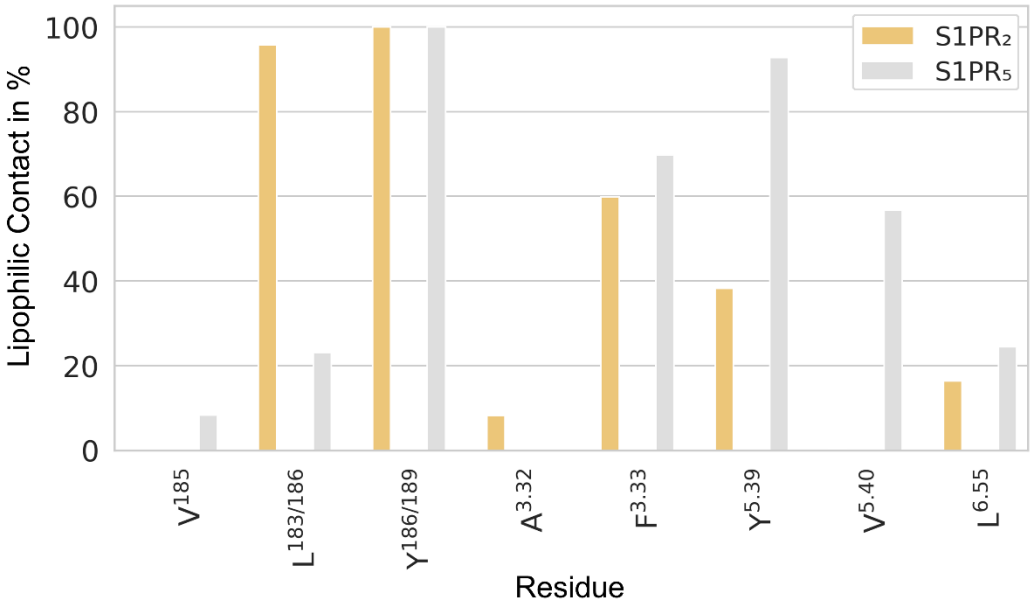


Figure 54. **Occurrence of lipophilic residue interactions in percent for BM3 of $S1PR_2$ and $S1PR_5$.** Corresponding data can be found in Appendix Table 10.

4.1.11. Comparison Between Experimentally Solved Binding Mode of Ozanimod/Siponimod and BM1/BM2 of Pan-agonist 6 Lipophilic Interactions

During the time of our investigation new cryo-EM structures of active state $S1PR_1$ in complex with the superagonists Ozanimod and Siponimod as well as $S1PR_5$ in complex with Siponimod has been published (PDB 7EW0, 7EVY, and 7EW1, respectively) [219]. In this chapter we want to evaluate the lipophilic binding pockets in $S1PR_1$ and $S1PR_5$, compare them to our binding modes 1 and 2, and explain the selectivity of Ozanimod and Siponimod based on our models and observations.

$S1PR_1$ in complex with Ozanimod shows lipophilic interactions of the isopropyl moiety with $V^{3.40}$, $F^{3.41}$, $L^{4.56}$, $V^{5.46}$, $F^{5.47}$, $L^{5.50}$, and $W^{6.48}$ (Figure 55 left, $L^{4.56}$ for the sake of clarity not shown). The nitrile moiety is interacting with $T^{5.44}$ through a hydrogen bond. Siponimod shows similar interactions of the cyclohexyl moiety with $S1PR_1$ as Ozanimod's isopropyl moiety ($V^{3.40}$, $F^{3.41}$, $L^{4.56}$, $V^{5.46}$, $F^{5.47}$, $L^{5.50}$, and $W^{6.48}$, Figure 55 middle, $L^{4.56}$ for the sake of clarity not shown). The trifluoromethyl moiety shows lipophilic contacts to the side chains of $T^{5.44}$, $F^{5.47}$,

$F^{6.52}$, and $L^{6.55}$. Interestingly, $T^{5.44}$ adopts a different side chain rotamer over the chi1 angle compared to the Ozanimod complex leading to different positions of the hydroxyl and methyl group to accommodate the different moieties of Ozanimod and Siponimod in $S1PR_1$. In $S1PR_5$ Siponimods cyclohexyl moiety shows lipophilic contacts to $T^{3.37}$, $V^{3.40}$, $L^{4.56}$, $A^{5.46}$, $F^{5.47}$, $I^{5.50}$, and $W^{6.48}$ and the trifluoromethyl moiety to $V^{5.44}$, $F^{5.47}$, $F^{6.52}$, and $L^{6.55}$ (Figure 55 right, $L^{4.56}$ for the sake of clarity not shown).

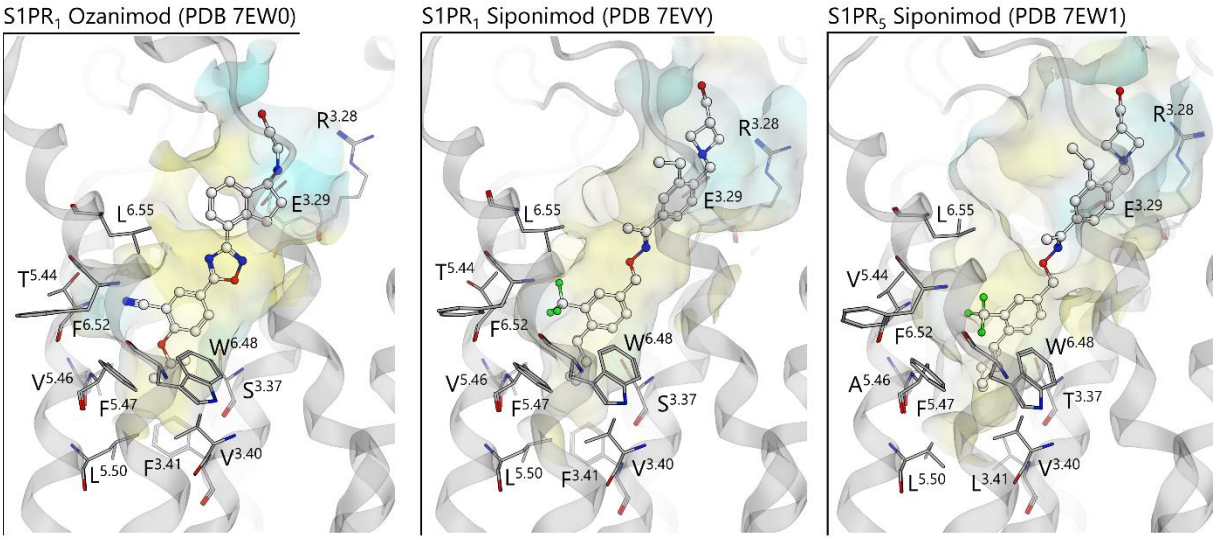


Figure 55. **$S1PR_1$ in complex with Ozanimod and $S1PR_{1,5}$ in complex with Siponimod (PDB 7EW0, 7EVY, and 7EW1 [219], respectively).** Yellow surface color – lipophilic, Blue surface color - hydrophilic

The superposition of representative frames for BM1 and BM2 (chapter 4.1.5) taken from our MD simulations of pan agonist **6** in complex with $S1PR_1$ with the structurally solved Ozanimod in complex with $S1PR_1$ revealed that the nitrile moiety occupies the cavity of BM1 occupied by the lipophilic tail of compound **6** while the isopropyl moiety occupies the same cavity compound **6** lipophilic tail occupies in BM2 (Figure 56A). The lipophilic tail of pan-agonist **6** perfectly aligns in each binding mode with the corresponding moiety of Ozanimod in complex with $S1PR_1$. Superposing the BM1 and BM2 conformations of pan-agonist **6** in $S1PR_1$ with the structurally solved Siponimod in complex with $S1PR_1$ also revealed matching binding pocket accommodations of the trifluoromethyl group with pan-agonist **6** lipophilic tail BM1 conformation and the cyclohexyl moiety with pan agonist 6

BM2 conformation (Figure 56B). The most frequent compound **6** BM1 lipophilic contacts in $S1PR_1$ ($T^{5.44}$, $F^{5.47}$, $F^{6.52}$, and $L^{6.55}$, chapter 4.1.8) are also observed for the trifluoromethyl group of Siponimod in complex with $S1PR_1$, while the nitrile moiety is surrounded by these residues in $S1PR_1$. Compound **6** BM2's most frequent lipophilic interactions in $S1PR_1$ are $V^{3.40}$, $F^{5.47}$, $W^{6.48}$, and $F^{6.52}$ (chapter 4.1.9). Due to the rigidity of Ozanimod and Siponimod as well as the large isopropyl moiety or cyclohexyl moiety compared to the n-alkyl chain in **6**, we observe no lipophilic contacts of both moieties to $F^{6.52}$ but with additional contacts to $L^{4.56}$ and $V^{5.46}$.

We also observed compound **6** BM1 and BM2 alkyl moiety conformations for Siponimod lipophilic moieties (trifluoromethyl and cyclohexyl, respectively) in complex with $S1PR_5$ as observed for **6** in complex with $S1PR_5$ (Figure 56C). For **6'** BM1 in complex with $S1PR_5$ $V^{5.44}$, $F^{5.47}$, $F^{6.52}$, and $L^{6.55}$ are the most frequent interactions (chapter 4.1.8), which are also observed for the trifluoromethyl moiety of Siponimod in complex with $S1PR_5$. Compound **6'** BM2 most recurring interactions in $S1PR_5$ consist of $T^{3.37}$, $V^{3.40}$, $A^{5.46}$, $F^{5.47,}$ and $I^{5.50}$ with less frequent interactions to $L^{4.56}$ and $W^{6.48}$. These are also observed for the cyclohexyl moiety of Siponimod in complex with $S1PR_5$.

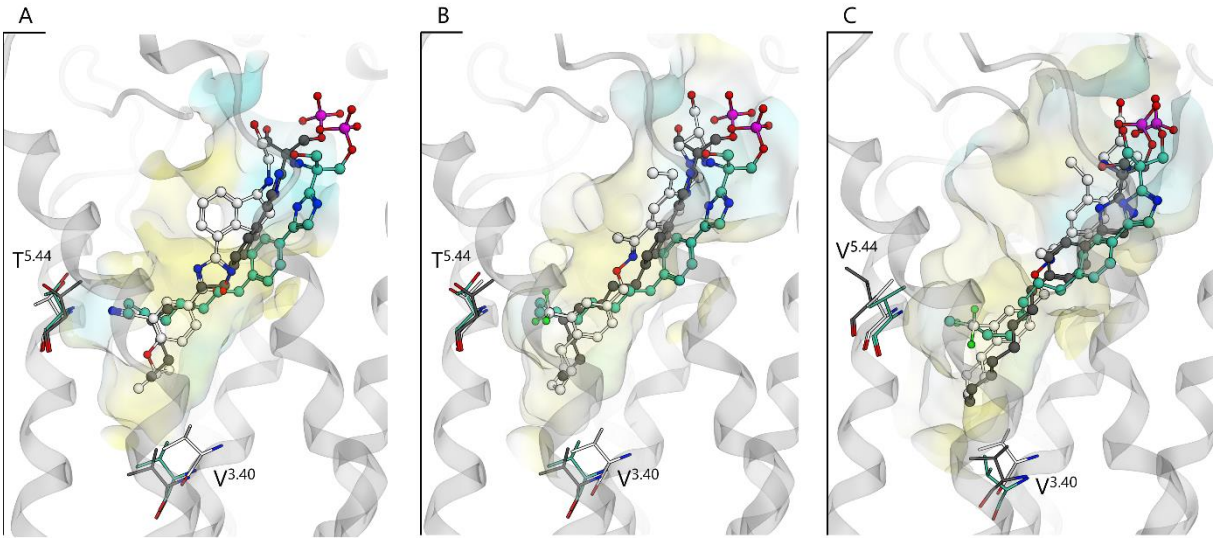


Figure 56. **Superpositions of representative frames for BM1 and BM2 of pan agonist 6 in complex with $S1PR_1$ (A and B, respectively) or $S1PR_5$ (C) with structurally solved $S1PR_1$ in complex with Ozanimod (A) and $S1PR_{1,5}$ in complex with Siponimod (B and C, respectively).** Residues used to define BM1 (BWN 5.44) and BM2 (BWN 3.40) are shown. White atoms – structurally solved ligand and protein, Green atoms – BM1 of pan agonist 6, Dark grey atoms – BM2 of pan agonist 6, Yellow surface color – lipophilic, Blue surface color - hydrophilic

Our identification and definition of BM1 and BM2 from MD simulations (chapter 4.1.6) is therefore in line with the structurally solved binding of Ozanimod to $S1PR_1$ and Siponimod to $S1PR_{1,5}$. From that, we can conclude that the higher potency of Ozanimod to $S1PR_1$ against $S1PR_5$ (EC_{50} 0.41nm and 11nm, respectively) comes from the hydrogen bond formed between the hydroxyl group in $T^{5.44}$ and the nitrile moiety, which is $V^{5.44}$ in $S1PR_5$. The residue $T^{5.44}$ or $V^{5.44}$ does not impact the binding of the trifluoromethyl moiety of Siponimod since both can establish a lipophilic contact. Even though in $S1PR_1$ BM2 residue $S^{3.37}$ is less favorable for binding the lipophilic cyclohexyl moiety in comparison to $S1PR_5$ BM2 residue $T^{3.37}$, $S^{3.37}$ is less bulky and enables more space in BM2. Therefore, Siponimod shows similar potency to $S1PR_1$ and $S1PR_5$ (EC_{50} 0.39nm and 0.38nm, respectively).

Neither Ozanimod nor Siponimod binds to $S1PR_2$. This can be explained by following selectivity determinants. First, $S1PR_2$ expresses $F^{7.39}$ while the subtypes $S1PR_{1,4,5}$ express $L^{7.39}$ and $S1PR_3$ $I^{7.39}$ (chapter 4.1.3, Figure 28). The larger residue located between the lipophilic

and hydrophilic binding pockets prevents the large and rigid propoxybenznitrile moiety of Ozanimod and cyclohexyltrifluoromethylphenyl of Siponimod from entering the lipophilic binding pockets of BM1 and BM2. Furthermore, $V^{5.44}$ in BM1 is less favorable for binding the nitrile moiety of Ozanimod compared to $S1PR_1$ with $T^{5.44}$. While $V^{5.43}$ is expressed only in $S1PR_2$ (vs $S1PR_{1,3,4,5}$ $C^{5.43}$) and is favorable for binding the trifluoromethyl group of Siponimod, the size of $S1PR_2$ BM2 side chain $I^{5.46}$ in comparison to $V^{5.46}$ $S1PR_1$ and $A^{5.46}$ $S1PR_5$ is less favorable for the accommodation of the cyclohexyl moiety of Siponimod. We have also observed interactions in $S1PR_2$ BM1 which are predominantly seen in BM2 ($L^{4.56}$ and $I^{5.50}$), as well as interaction partners in BM2 ($F^{6.44}$ and $F^{6.45}$) only observed in $S1PR_2$, which are located closer to the intracellular region. This might indicate a deeper location of BM1 and BM2 for $S1PR_2$ in comparison to the other subtypes.

Ozanimod and Siponimod also shows no potency at $S1PR_3$. In $S1PR_3$ BM1 $F^{6.55}$ (vs $S1PR_{1,2,4,5}$ $L^{6.55}$) might act like a lid to the binding pocket, which prevents the large and rigid propoxybenznitrile moiety of Ozanimod and cyclohexyltrifluoromethylphenyl of Siponimod from entering the lipophilic binding pockets of BM1 and BM2. Furthermore, the side chain of $S1PR_3$ BM1 $I^{5.44}$ is larger than $S1PR_1$ $T^{5.44}$ and $S1PR_5$ $V^{5.44}$, which is for both the nitrile moiety of Ozanimod and trifluoromethyl moiety of Siponimod less favorable. The larger residue of $S1PR_3$ $I^{5.46}$ from BM2 in comparison to $V^{5.46}$ $S1PR_1$ and $A^{5.46}$ $S1PR_5$ is also less favorable for the accommodation of the cyclohexyl moiety of Siponimod. On the other hand, $S1PR_3$ $G^{3.37}$ (vs $S^{3.37}$, $S^{3.37}$, $A^{3.37}$, and $T^{3.37}$ for $S1PR_{1,2,4,5}$, respectively) smaller size could allow binding, even though no lipophilic contact would stabilize the lipophilic moiety.

Lastly, Ozanimod shows no potency to $S1PR_4$, while Siponimod exhibits weak potency at $S1PR_4$ (EC_{50} 920 nM). $S1PR_4$ BM1 residue $L^{5.44}$ does not allow hydrogen bonding to the nitrile moiety of Ozanimod compared to $S1PR_1$ $T^{5.44}$ and is larger than $S1PR_5$ $V^{5.44}$. This is also less favorable for binding the trifluoromethyl moiety of Siponimod. For BM2 only $S1PR_4$ and $S1PR_5$ show interactions to residue $A^{3.37}$ / $T^{3.37}$ during BM2. Especially $A^{3.37}$ with the small size and possible lipophilic contact to the large cyclohexyl moiety of Siponimod is in favor of binding of Siponimod. On the other hand, the larger $I^{5.46}$ in comparison to $S1PR_1$ $V^{5.46}$ and $S1PR_5$ $A^{5.46}$ is less favorable to accommodate the large cyclohexyl moiety of Siponimod. In conclusion, the stated points explain the weak potency of Siponimod to $S1PR_4$.

4.2. Dihedral Angle Dynamics of Class A G Protein-coupled Receptor Activation Hotspots

While the elucidation of S1PR subtype selectivity determinants is important for the rational design of ligands for modulating disease relevant pathophysiological processes and to reduce adverse effects triggered through off-targets, we want to also understand the underlying mechanisms of general class A GPCR activation. The following chapters will show how dihedral angles are involved in the class A GPCR activation process. Furthermore, we present a fast and reliable method which uses dihedral angles as determinants to predict the activation states of homology models (HM) and molecular dynamics (MD) simulations of class A GPCRs utilized in structure-based drug design.

4.2.1. Structural Dataset Assessment & Curation

The underlying dataset used for machine learning models initially consists of all available class A GPCR structures (as of 29.09.2021). In total 487 structural data points consisting of 384 crystal structures and 103 cryo-EM structures (Figure 57) has been collected from the PDB [220]. 261 structures are defined as an inactive state, 188 as active, 37 as intermediate, and 1 as undefined. These states are adopted from the GPCRdb [26, 221] and are based on intracellular Cα distances of ternary structure complexes as active state references and highly closed intracellular TM structures as references for inactive states. The structural dataset spans 38 different GPCR families with 85 unique receptors. To evaluate the quality of the structures five indicators have been extracted from the RCSB PDB database [222, 223]: overall resolution, clashscore [224], average B factor of protein atoms [225], Rfree [226, 227], and Δ Rfree/Rwork. While the resolution and clashscore were available for all structural data, Rfree, Δ Rfree/Rwork, and average B factor were only found for crystal structures obtained via x-ray diffraction (487, 487, 384, 384, and 256 structures, respectively). The initial outlier evaluation was carried out with boxplots (Figure 58). Resolution-wise, we observed 24 outliers past the upper whisker (4.03 Å) of the boxplot and five under the lower whisker of 1.82 Å. Even though 3 Å is generally defined as the minimum resolution to observe secondary structures of the protein [228], the less flexible TM domains show higher resolutions compared to the overall, intracellular, and extracellular resolutions [229].

Therefore, 4 Å has been chosen as the maximum resolution tolerated for our curated datasets. The clashscore boxplot revealed 51 outliers above the upper whisker of 15.32, the average B factor ten outliers above the upper whisker of 153.25, Rfree eight outliers above the upper whisker of 0.34, and Δ Rfree/Rwork 26 outliers above the upper whisker of 0.06 and 24 outliers below the lower whisker of 0.01.

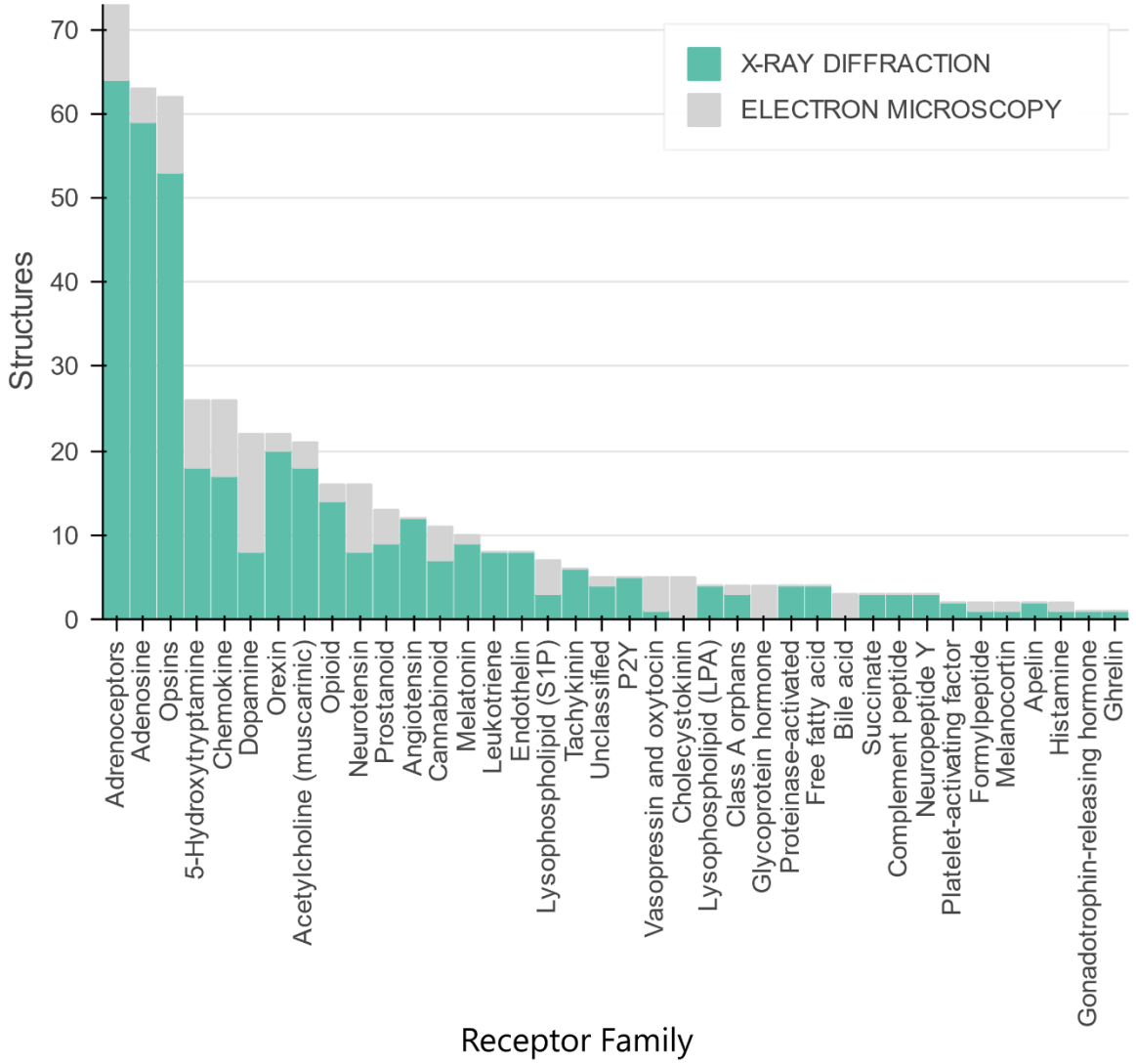


Figure 57. **Distribution of all structural GPCR data between receptor families.** 487 structural data points in total (as of 29.09.2021).

All indicators showed a problematic analysis with boxplots due to the nature of the data itself being either heavy skewed (asymmetric distribution), right-hand oriented (0 as natural

lower limit, outliers only for high values), or showing long-tails in the data distribution. An example of that is the identification of outliers below the lower whisker of the resolution and Δ Rfree/Rwork, even though optimum values for the indicators are 0. Another example is the long-tailed distribution of the clashscore. Shao et al. has addressed these problems in 2018 with a probability-density-ranking (PDR) approach to analyze PDB data outliers [230].

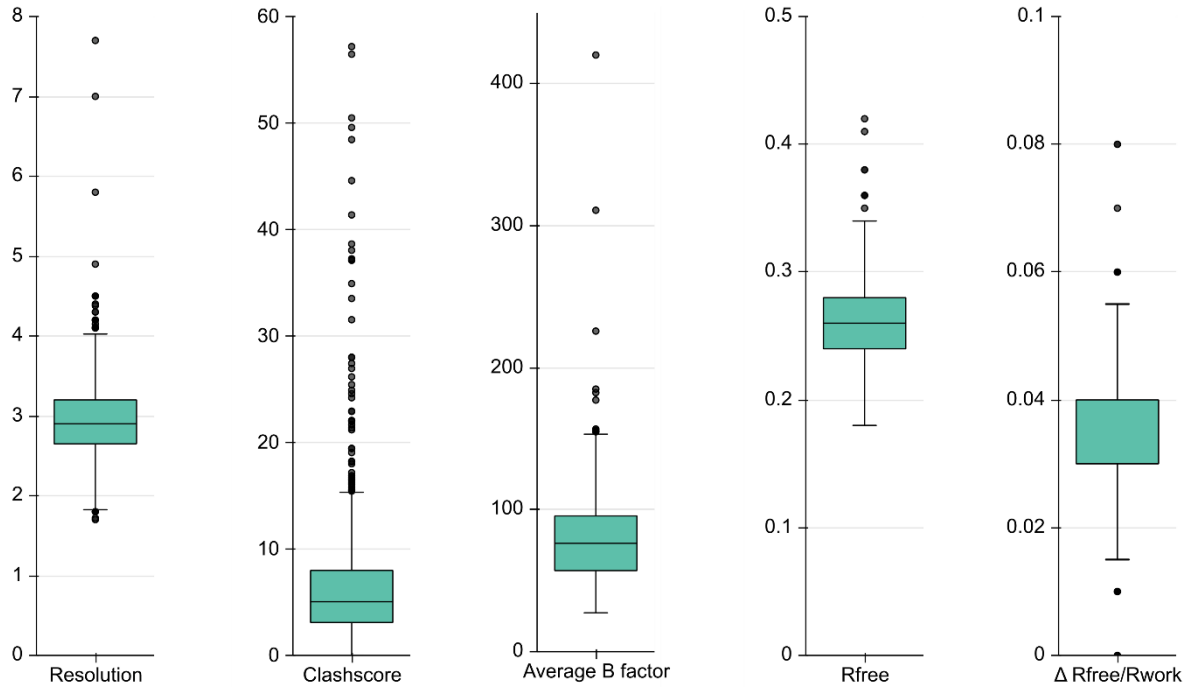


Figure 58. **Boxplots of different measures of structural data quality for all data.** 487 structural data points in total (as of 29.09.2021).

Finally, our curated dataset used 4 Å for resolution as defined from the boxplots and 5% PDR boundaries for the clashscore (34.0), average B factor (87.105), Rfree (0.312), and Δ Rfree/Rwork (0.076) to eliminate outliers as recommended by Shao et al. The 5% PDR boundary datasets also only consist of data with active/inactive labels. The 5% PDR boundary dataset included 324 structures with 242 crystal structures and 82 cryo-EM structures. 176 structures are labeled as inactive, while 148 are labeled as active. These structures span 33 GPCR families. In comparison to the boxplot of the full data set, the 5% PDR boundary dataset fully eliminates upper outliers for the resolution, average B factor, and Rfree values, 23 clashscore outliers, and 10 upper whiskers Δ Rfree/Rwork outliers. The

different amount of structures between the 5% PDR boundary and full datasets (448) consisting of active and inactive structures results in a more favorably balanced dataset between active and inactive structures for the 5% PDR boundary dataset of around 1:1.2 active/inactive state ratio, while the full dataset shows a more imbalanced ratio around 1:1.4 (188 active and 260 inactive structures). The 5% PDR boundary dataset has been used for the creation of all following machine learning models.

4.2.2. Dihedral Angles as Features for the Machine Learning Models

Features for the machine learning model consist of GPCR intercomparable Φ (phi) and ψ (psi) dihedral angles of the backbone (chapter 1.1.3) in the transmembrane domain. To ensure comparability between each feature we aligned all structures from the 5% PDR structural dataset (324 structures) based on the adjusted BWN numbering scheme of the GPCRdb. Due to possible differences in length, insertions, or deletions between each GPCR subtype the extracted dihedral angles of each feature can contain non-defined values. Initial feature extraction resulted in 621 dihedral angle features. All features with more than 1% undefined values were removed from the data set. Feature values with undefined values were replaced with an average value of the corresponding feature leading to a total of 286 features (145 Φ and 141 ψ angle features) from 159 residues (Figure 59).

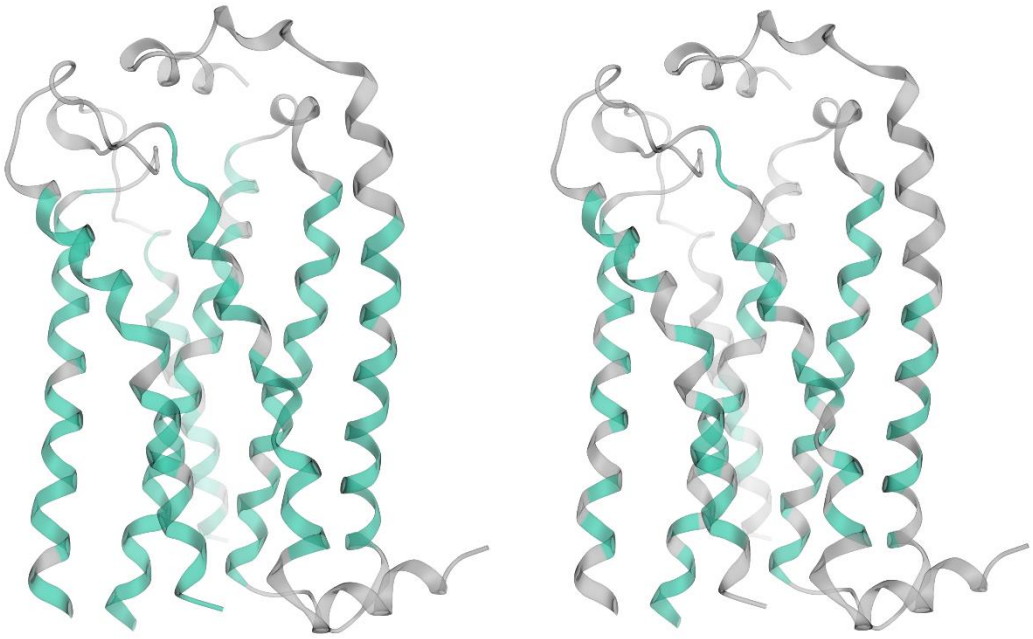


Figure 59. **Extracted dihedral angle residue positions (green). Left:** Under 1% undefined angle values are handled with the mean value of the corresponding feature resulting in 159 residue positions with 286 dihedral angle features. **Right:** Removal of every feature with undefined values results in 94 residues positions with 141 dihedral angle features.

The removal of every feature containing an undefined value leads to a total of 141 features (71 Φ and 70 ψ angle features) from 94 residues (nonan dataset). Handling the non-defined values with a 1% threshold provides a larger coverage (Figure 59 left) of the TM domains in comparison with the removal of every feature if non-defined values occur while maintaining a high information density of the dataset provided by the structural data (1% nan dataset).

4.2.3. Model Building and Evaluation

1% nan dataset with variance-based feature reduction yielded the best model performances

For both the 1% nan dataset and nonan dataset, the number of features leads to very high dimensionality. We, therefore, employed the following feature reduction methods: recursive feature elimination with cross-validation (RFECV), random forest feature selection (RFSEL), and variance-based statistical selection (ANOVA f-value) to reduce the model dimensionality

and to remove features that do not contribute to the prediction accuracy between active and inactive states. These are paired with three different machine learning algorithms with optimized hyperparameter settings for each model (Figure 60): decision tree classifier (DTC), support vector machine (SVM), and k-nearest neighbor (KNN).

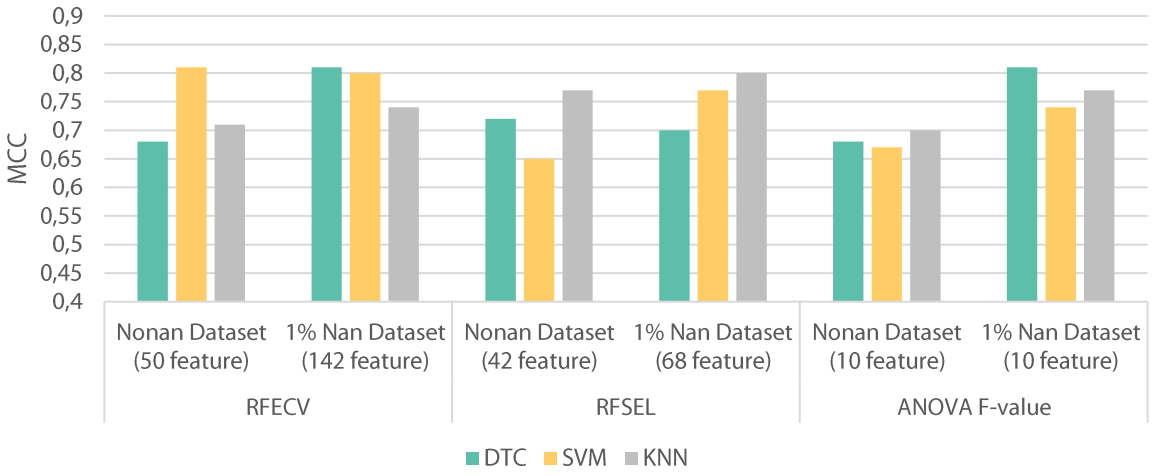


Figure 60. **Feature reduction methods with optimized hyperparameter settings for the nonan and 1% nan datasets.** The X-axis states the dataset and resulting feature numbers in brackets from the applied feature reduction methods. The Mathews correlation coefficient (MCC) values range from -1 to 1 (1 being the perfect prediction score and 0 a random prediction). DTC - decision tree classifier, SVM - support vector machine, KNN - k-nearest neighbor

Both datasets have been split into a training and test set (80:20) with even distribution of active and inactive state data. The initial RFECV and RFSEL models yielded a good prediction performance measure of 0.65 to 0.81 MCC with features still ranging between 42 to 142. Using a variance-based method (ANOVA f-value) to select the ten best features yielded with the 1% nan dataset an MCC of 0.74 paired with SVM, 0.77 with KNN, and 0.81 with DTC. In comparison with the nonan dataset, the 1% nan dataset using the ANOVA f-value for feature selection performed better with the DTC model (MCC 0.68 and 0.81, respectively), the SVM model (MCC 0.67 and 0.74, respectively), and the KNN model (MCC 0.70 and 0.77, respectively). The ten features of the 1% nan dataset (with ANOVA f-value) consist of residue

positions 1.56, 2.39, 4.51, 4.55, 6.44, 7.47, 7.48, 7.52, 7.53, and 7.54 (Figure 61). Position 6.44 is part of the $P^{5.50}$–$I^{3.40}$–$F^{6.44}$ motif and 7.52 – 7.53 is part of the $N^{7.49}$–$P^{7.50}$–x–x–$Y^{7.53}$ motif both highly conserved and important for activation of class A GPCRs (chapter 1.1.4) [42]. Due to the introspective possibility and highest MCC, we further evaluated and analyzed the ANOVA/DTC model of the 1% nan dataset.

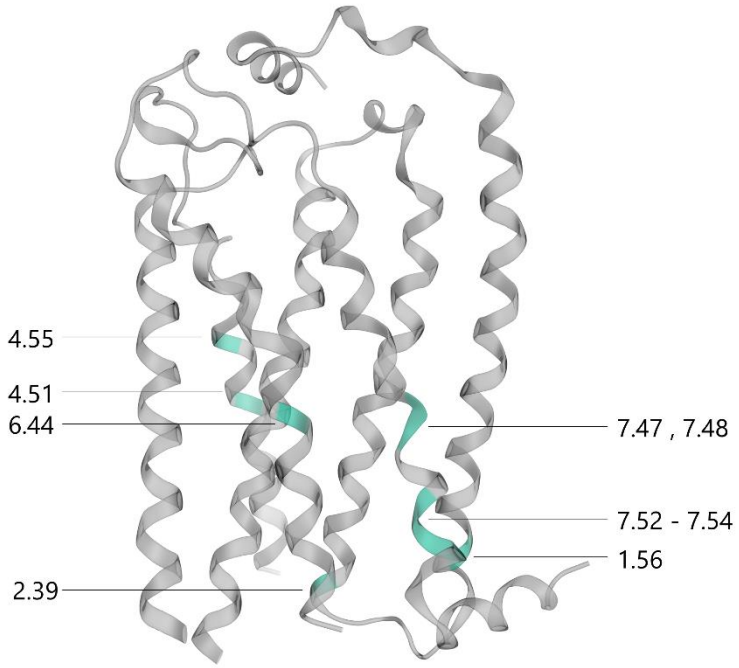


Figure 61. **Feature positions of the models after feature selection through ANOVA f-value of the 1% nan dataset.** Positions are marked in green and numbered via the BWN scheme. The shown structure is $S1PR_1$ (PDB 3V2Y).

4.2.4. Decision Tree Model and Dihedral Angle Analysis

The ANOVA/DTC model of the 1% nan dataset performance is the highest with 0.81 MCC. The training set accuracy is 91.52% ± 5.71% and the test set accuracy is 90.77%, which suggests that the model is not overfitting. Structure wise the DTC model has a depth of two with one root decision node at the top branching into one successor decision node and one terminal successor child node at depth one and two terminal successor child nodes (leaf node) on depth two branching from the decision node (Figure 62A). The root decision node

contains 136 inactive state structures and 123 active state structures. These are split based on the feature value -69.98° Φ of residue 7.53. If the feature value of a given sample structure in the root node is smaller or equal (-69.98°) it will be categorized in the second decision node. Else it will be categorized into the terminal successor child node with the label inactive at the decision tree depth one. 118 inactive state structures and 10 active state structures have been categorized into the said terminal node. The remaining samples of 18 inactive state and 113 active state structures are further split by the second decision node with a feature value of -76.33° Φ of residue 7.45. Sample features are split into the left terminal successor child node with an active state if the value is smaller or equal as the decision node (-76.33°). 15 inactive state structures and 9 active state structures are found in the inactive labeled right terminal node, while 3 inactive and 104 active state structures are found in the active labeled left terminal node.

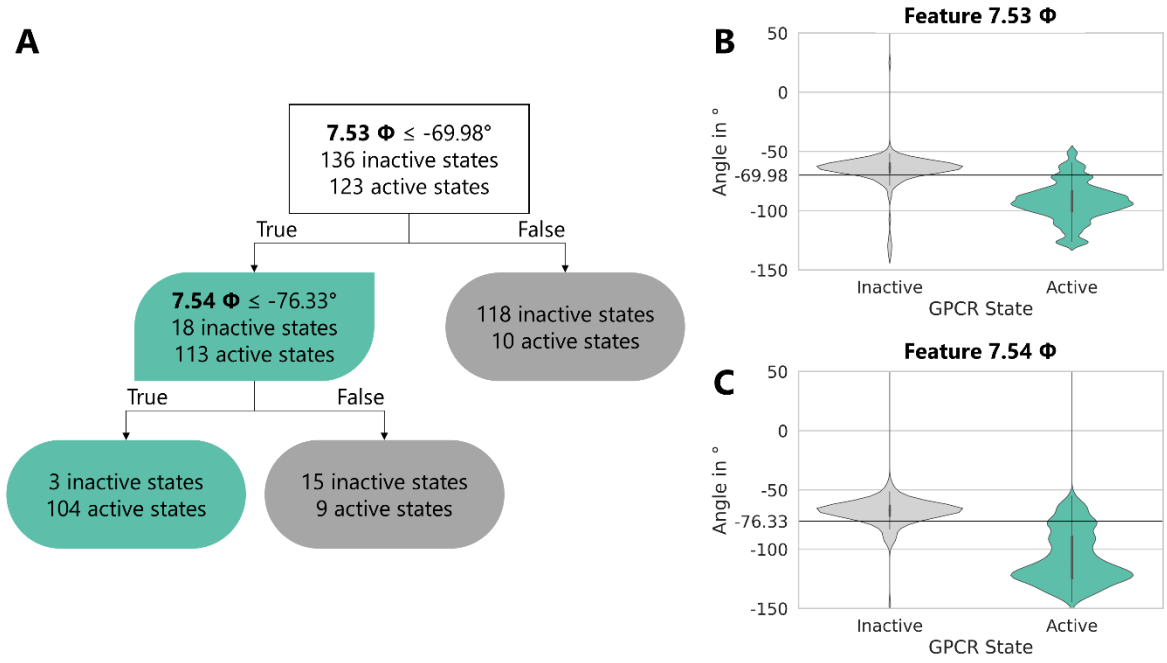


Figure 62. **(A) DTC model and (B, C) corresponding violin plots for both features.**

To gain insights into the decision nodes, we plotted the angles of feature 7.53 Φ and 7.54 Φ with violin plots of the full data set (Figure 62B and Figure 62C, respectively). Both reveal that most inactive state structures show an Φ angle over the decision nodes feature value of -69.98° Φ for the residue 7.53 and -76.33° Φ for the residue 7.54. These two positions are also

in line with the general activation mechanism of GPCR, in which the change from inactive to active state structure leads to a final rearrangement of the conserved $N^{7.49}$–$P^{7.50}$–x–x–$Y^{7.53}$ motif, which forms the binding site for intracellular binding partners (Figure 63). This rearrangement seems to correlate with the change in the Φ angle of both positions 7.53 and 7.54. The typical degree for α-helices according to the values sampled by Ramachandran et al. [39] lies in the range of -60° Φ angle, which means during GPCR activation we observe an unwinding of these two residue positions out of the normative Φ angles observed for α-helices moving the intracellular TM7 part inwards.

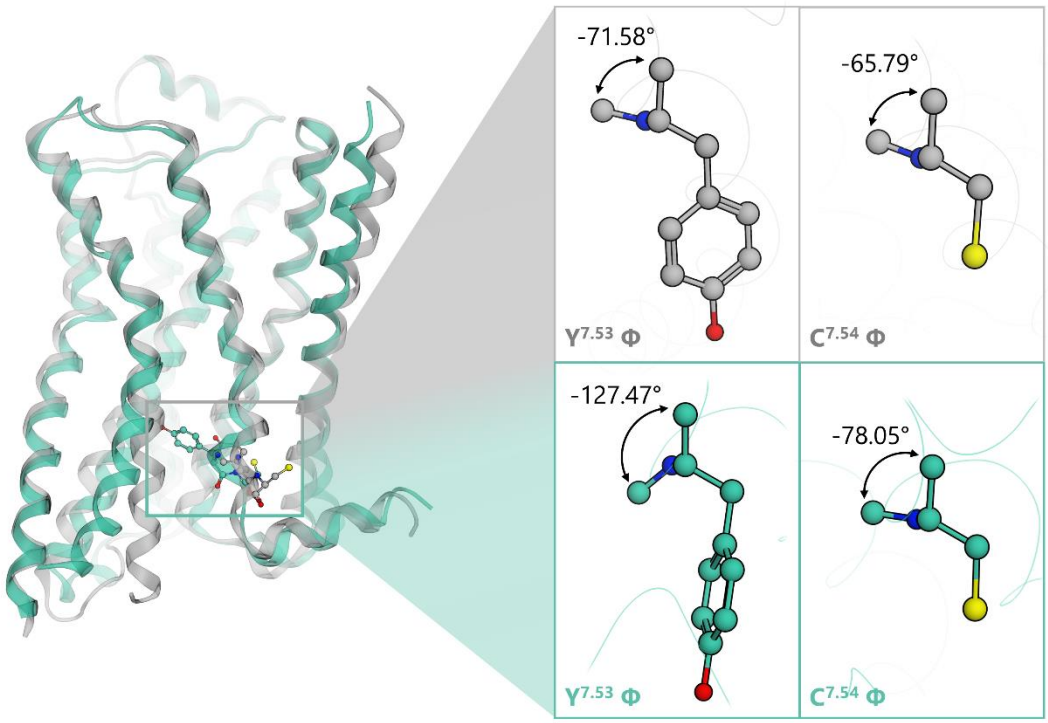


Figure 63. **Dihedral angle change of $Y^{7.53}$ Φ and $C^{7.54}$ Φ between an inactive (grey, PDB 2RH1 [231]) and active state (green, PDB 3SN6 [22]) β_2-adrenoceptor.**

Due to the wider distribution of active state angles seen in the violin plots for 7.53 Φ and 7.54 Φ (Figure 62B and Figure 62C, respectively), we further investigated the angle distributions of active state GPCRs in complex with G_s, $G_{i/o}$, and $G_{q/11}$. To ensure a correct G protein coupling representation of the structural data, we only selected active state structures solved with a G protein. Furthermore, we discarded GPCR structures that are not

solved with their main intracellular binding partner (IBP) or have several main IBPs (based on the gproteindb.org [26, 221]). The kernel density estimation plot revealed a narrower distribution of 7.53 Φ and 7.54 Φ angles for $G_{q/11}$ < G_s < $G_{i/o}$ coupled GPCRs in a kernel density estimation plot (Figure 64), although it should be noted that only five $G_{q/11}$ coupled GPCR structures are included into the plot (vs. 31 $G_{i/o}$ and 29 G_s-coupled GPCRs). Even though GPCRs binding $G_{i/o}$ proteins seem to allow a larger distribution of both angles in comparison to G_s-coupled GPCRs the distribution is still comparable and is in line with residue position 7.53 representing a universal mediator of GPCR activation.

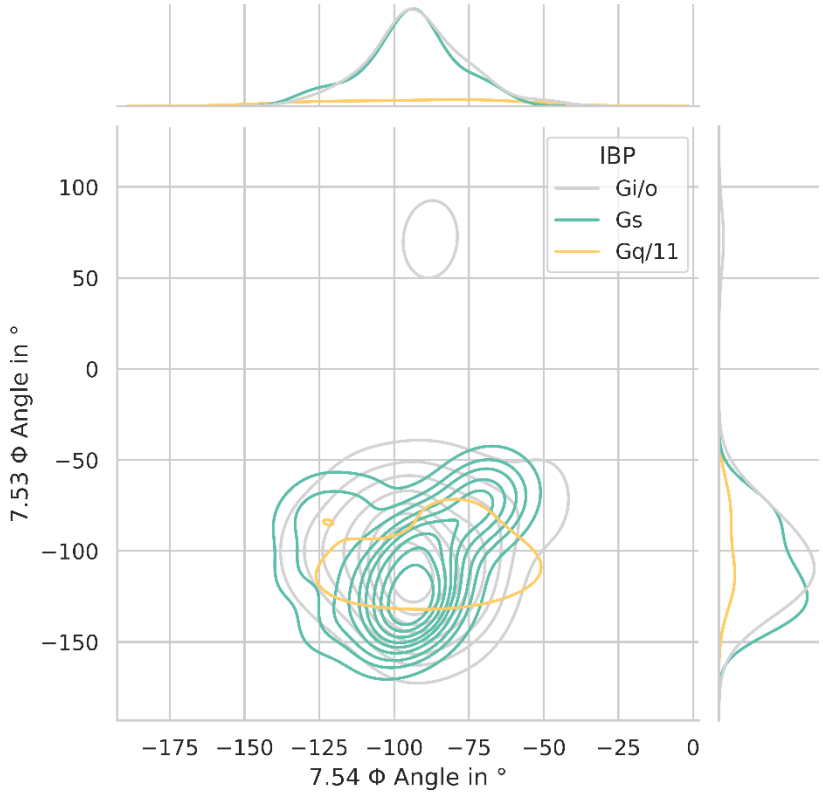


Figure 64. **Kernel density estimation plot of 7.53 Φ and 7.54 Φ active state structure angles coupled to different G proteins.**

4.2.5. Prediction of Homology Model Activation States

GPCRdb Homology Models

The GPCRdb integrated an automated pipeline for the construction of active, intermediate, and inactive state homology models (HM) [26, 232]. We have downloaded all available homology models with active or inactive labels and predicted with our model the state of each homology model to check against the given labels. In total 217 inactive and 230 active GPCRdb labeled homology models from 59 different class A GPCR families have been analyzed (Figure 65).

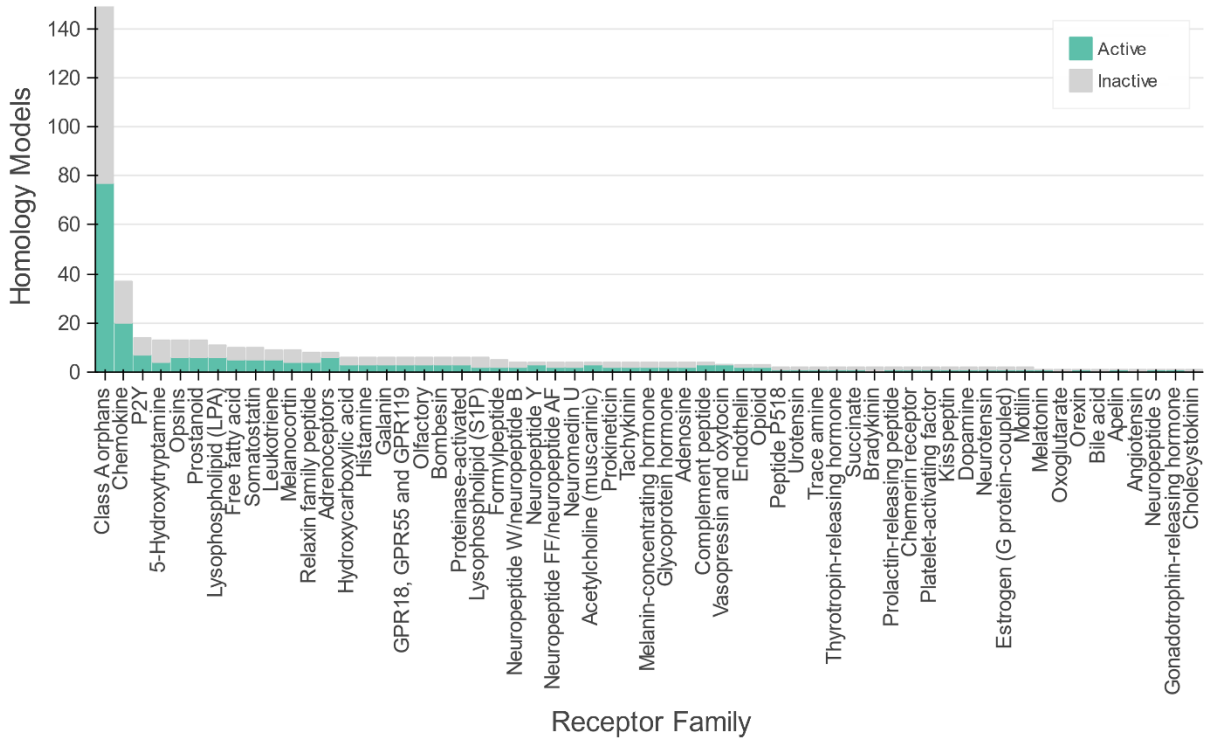


Figure 65. **Homology models from different class A families with active/inactive labels from the GPCRdb.**

From 217 inactive state homology models, 210 (96.8%) have been also predicted as an inactive state from our DTC model, while seven has been labeled as active. A closer look at the wrongly categorized HM reveals that two homology models (FFAR2 and NPY4R) have been built with structural templates (PDB 7F8Y [233] and 5ZBH [234], respectively) that were

not included in our model since they did not satisfy our outlier and quality assessment (chapter 4.2.1). The structural templates of three miscategorized originally as inactive labeled homology models (LPAR6, S1PR4, and TAAR1) were identified correctly as inactive (PDB 4MBS [235], 3V2Y [32], and 7BVQ [236], respectively). We have aligned and superposed each of the three models with the corresponding template to investigate possible reasons for the different categorizations between the GPCRdb and our model. The homology model of LPAR6 shows a high mean RMSD of 3.3 Å to its template (PDB 4MBS [235]) with especially high RMSD values at residue positions 7.53 (5.3 Å) and 7.54 (6.6 Å). The $S1PR_4$ GPCRdb HM also shows a high mean RMSD of 3.9 Å and high RMSD values for positions 7.53 (3.6 Å) and 7.54 (4.6 Å) against its template (PDB 3V2Y [32]). Both high RMSD between HM and template leads to a different angle and therefore different categorization of our model vs. the GPCRdb. Superposing the homology model TAAR1 with its structural template (PDB 7BVQ [236]) showed a good mean RMSD value of 1.4 Å with values at positions 7.53 (0.25 Å) and 7.54 (0.22 Å). The TAAR1 homology models phi angle at position 7.53 with -73.7° and 7.54 with -85.9°, which is very close to the first and second decision node value of our model (-69.98° and -76.34°, respectively) leading to the categorization as active state. Interestingly, the phi angle of the structural template of TAAR1 at position 7.53 is -68.9° slightly larger than the decision nodes value and therefore labeled as inactive. Lastly, two of the homology models (IGR4 and IGR5) are both built based on PDB structure 7FIJ [237], which was also incorrectly classified by our model. Therefore, these homology models adopt the angles of the incorrectly classified PDB structures and are outliers of our model.

The prediction of 230 active state GPCRdb homology models revealed 158 (68.7%) also predicted as active states and 72 (31.3%) as inactive states by our DTC model. We have further analyzed 34 of the 72 homology models built upon structural data, which has also been used for our model meeting the requirements of our outlier and quality assessment (chapter 4.2.1). Of 34 homology models, 13 are based on PDB structures (AA2BR based on PDB 4UHR [238]; NTR2 and GPR39 based on PDB 4XES [239]; PAR2, P2Y11, GPR55, GPR4, GPR174, GPR151, GPR15, and GPR132 based on PDB 5UNG [240]; TAR2 and PF2R based on PDB 6AK3 [241]), which are solved with an agonist and without an IBP. These structures are not fully active and therefore the homology models resemble these structures. 16 homology models were built upon six ternary structures (OPRM based on PDB 6DDE [242]; GPR21 based on PDB 6LI3 [243]; GPR153 and GPR84 based on 6WHA [244]; GPR146, ACKR4, CCR10,

CCR7, CCR9, CXCR4, GPR34, and CXCR6 based on PDB 6WWZ [245]; GPR22, HRH3, and GPR160 based on PDB 7DFL [246]; TAAR6 based on PDB 7EXD [247]) with atypical angles (-60° to -70°) for active state structures (Figure 62) and therefore inherit the inactive label and atypical angles from the corresponding templates. The last 9 homology models were built upon correctly classified ternary structures (G37L1 based on PDB 7DB6 [248]; GPR148 based on 7E32 [249]; GPR63 and GPR162 based on PDB 7JVR [250]; V1AR, OXYR, GNRHR, GP150, and V1BR based on PDB 7KH0 [251]). The classification as inactive is due to the error margin of our model.

The prediction and analysis of provided homology models revealed, that not all GPCRdb given labels represent the state of the homology model. Our model provides a second measure to identify mislabeled homology models, especially for high average RMSD values between PDB structure templates and created models and homology models built with templates that are not in full active state due to a missing IBP.

Sphingosine-1-phosphate receptor 2-5 Homology Models

We have built HM of $S1PR_{2-5}$ based on the $S1PR_1$ crystal structure (PDB 3V2Y [32]) to elucidate structural determinants for subtype selectivity of $S1PR_{1-5}$ (chapter 4.1.1, page 37). The HM of $S1PR_{2-5}$ had low RMSD of the Cα protein backbone <1 Å compared to its template $S1PR_1$. All four HM were predicted to be inactive with our model, which is in line with the inactive state of the template $S1PR_1$.

4.2.6. β_2-adrenoceptors as a Model System for Dynamic Predictions

The effect of intracellular binding partners on the activation state of β_2-adrenoceptor

We have chosen the β_2-adrenoceptor as a test system because of the well-known mechanistic, ligands, and structural data available. To test our predictive model we prepared four dynamic model systems, which include (i) an β_2-adrenoceptor crystallized in complex with the active state stabilizing nanobody Nb80 and the full agonist BI-167107 (PDB 3P0G [252]), (ii) the same β_2-adrenoceptor without Nb80 but in complex with BI-167107, (iii) an β_2-adrenoceptor in complex with the inactive state stabilizing nanobody Nb60 and the inverse

agonist carazolol (PDB 5JQH [253]), (iv) and the same structure without Nb60 but in complex with carazolol (Figure 66). The active state crystal structure (PDB 3P0G [252]) has been categorized as active and the inactive state crystal structure (PDB 5JQH [253]) as inactive by our DTC model. Every model system has been simulated in five replicastes à 100 ns (Appendix Figure 3). For each frame of the MD simulations, we have predicted the state with our DTC machine learning model.

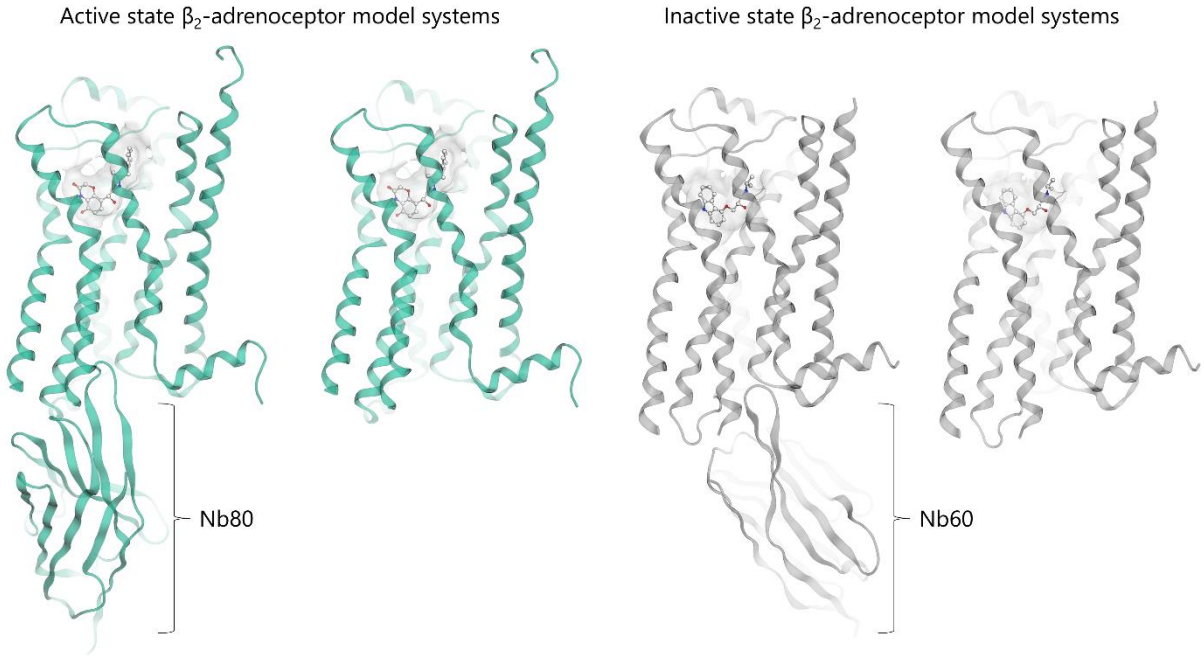


Figure 66. **Model systems to evaluate the effect of intracellular stabilizing nanobodies for active (PDB 3P0G [252]) and inactive (PDB 5JQH [253]) structures in a dynamic system.** Green – active state, grey -inactive state

As expected, the prediction of the active state for the ternary complex (i) resulted in mostly predicted active state frames of 77.6% of all simulations, while the inactive ternary complex (iii) resulted in the lowest amount of active states predicted at around 18.2% (Table 6). Removing the active state stabilizing nanobody lead to a reduction of 42.1% active frames to a mean of 35.5% for the active state β_2-adrenoceptor without a nanobody (ii), showing a high correlation between bound and non-bound active state stabilizing nanobody Nb80. The removal of the inactive state stabilizing nanobody resulted only in a slightly higher amount of predictive active states overall of 21.9%, which is in comparison to the active state complexes resulted only in a difference of 3.7%. Though this is in line with the known

activation process and stabilized conformational change intracellular through an IBP (chapter 1.1.4, page 8).

Table 6. **Prediction of MD simulations of β_2-adrenoceptor model systems with and without intracellular binding partners.** Green (100%) to white (50%) to grey (0%) indicate the number of active state frames predicted by the DTC model. Values are shown in %.

β_2-adrenoceptor models	Replica 1	Replica 2	Replica 3	Replica 4	Replica 5	Ø
Inactive + Carazolol + Nb60	38.8	5.5	17.7	13.3	15.8	18.2
Inactive + Carazolol	34.4	7.0	53.3	5.6	9.2	21.9
Active + BI167107	72.3	12.2	15.3	65.1	12.9	35.5
Active + BI167107 + Nb80	94.5	76.8	82.7	46.0	88.0	77.6

Different modes of ligands in complex with active state β_2-adrenoceptor

To evaluate possible influences of different ligand protein complexes on the activation state we have docked full agonists (adrenaline, hydroxybenzylisoproterenol, and isoprenaline), partial agonists (salbutamol and salmeterol), inverse agonists (carazolol, ICI-118,551, and timolol), and antagonists (alprenolol and propranolol) ligands (chapter 1.1.5) [254–256] to the orthosteric binding site of the active state β_2-adrenoceptor without Nb80 (PDB 3P0G [252], Figure 66). Docking of adrenaline, hydroxybenzylisoproterenol, salmeterol, carazolol, ICI-118,551, timolol, and alprenolol has been carried out based on co-crystallized ligand conformations with β_2-adrenoceptor (PDB 4LDO [257], 4LDL [257], 6MXT [258], 5JQH [253], 3NY9 [256], 3D4S [259], and 3NYA [256], respectively, Figure 67). Docking conformations of isoprenaline, salbutamol, and propranolol have been superposed and compared with new available crystal/cryo-EM structures (PDB 7DHR [260], 7DHI [260], and 6PS5 [261], respectively, Figure 67). The docking poses of all full and partial agonists match the crystallized pose, while the docking poses of the inverse agonists and antagonists mimic the orientation in the binding pocket with a minor offset to the extracellular part of the receptor. Only alprenolol showed a major offset to the extracellular part of the receptor after energy minimization of the docking conformation.

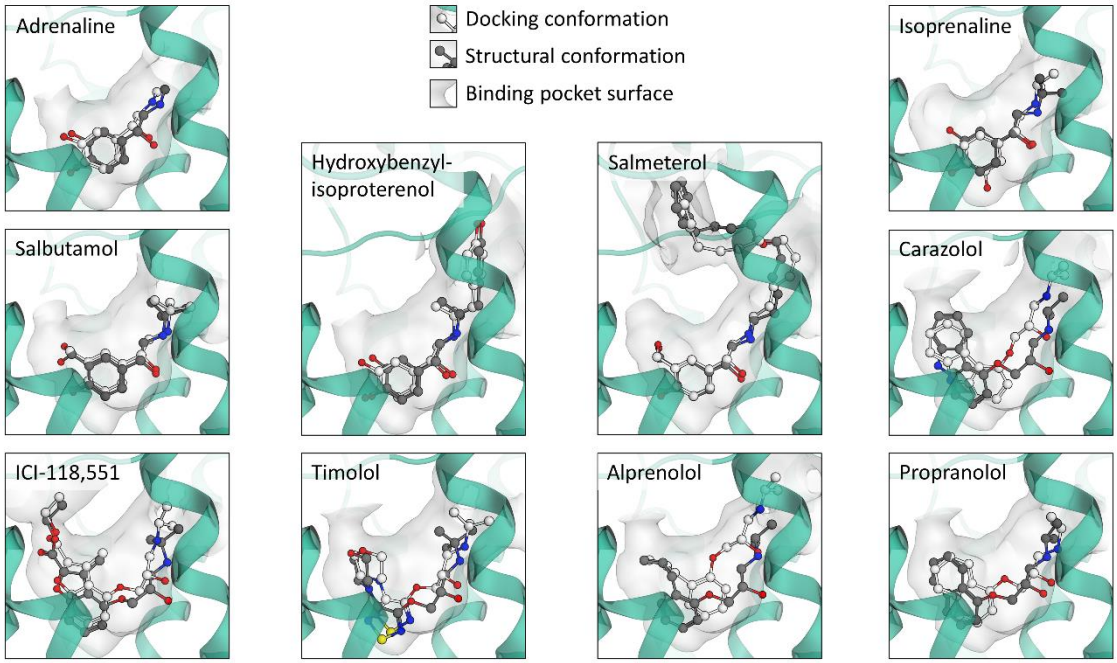


Figure 67. **Comparison between docking conformation and ligand conformations from crystal or cryo-EM structural data.** Green ribbon – active state β_2-adrenoceptor (PDB 3P0G)

The active/inactive state prediction of each frame of the MD simulation trajectories of full agonists in complex with the active state β_2-adrenoceptor revealed that the adrenaline and hydroxylisoproterenol models show a high amount of active state frames (52.7% and 64.6%, respectively), while the other full agonist complexes with isoprenaline and BI-167107 show a lower amount of predicted active state frames (44.1% and 35.5%, respectively). In comparison to the two partial agonist complexes with salbutamol (34.0% active state) and salmeterol (24.9% active state), the full agonist systems reveal more predicted active state frames. There seem to be no correlations to inverse agonist (carazolol, ICI-118,551, and timolol) and antagonist complexes (alprenolol and propranolol) showing high amounts of predicted active state frames as seen in the full agonist complexes (69.5%, 60.0%, 43.5%, 39.1%, and 54.0%, respectively). The lowest mean predicted active state is therefore salmeterol in complex with β2-adrenoceptor (24.9%) and the highest carazolol in complex with β2-adrenoceptor (69.5%). Mean active state predicted frames through all simulations reveal a similar amount as seen in the removal of the nanobody from active state β2-adrenoceptor in complex with the reference ligand BI-167107 (47.4% and 35.5%, respectively).

Table 7. **Prediction of MD simulations of active state β_2-adrenoceptor model in complex with different ligands.** Green (100%) to white (50%) to grey (0%) indicate the number of active state frames predicted by the DTC model. Values are shown in %.

Type	Ligand	Replica 1	Replica 2	Replica 3	Replica 4	Replica 5	Ø
Agonist	*Adrenaline*	60.9	11.7	77.5	83.2	30.1	52.7
	Hydroxybenzyl-isoproterenol	59.8	72.8	74.3	54.9	61.1	64.6
	Isoprenaline	17.5	48.3	17.6	74.6	62.6	44.1
	BI-167107	72.3	12.2	15.3	65.1	12.9	35.5
Partial agonist	*Salbutamol*	18.8	53.0	59.7	32.9	5.7	34.0
	Salmeterol	56.3	24.7	18.4	3.9	21.4	24.9
Inverse agonist	*Carazolol*	89.1	43.9	79.8	58.9	76.0	69.5
	ICI-118,551	52.7	91.7	88.2	15.9	51.7	60.0
	Timolol	90.8	18.9	81.6	8.2	17.9	43.5
Antagonist	*Alprenolol*	5.2	73.1	75.6	10.3	31.4	39.1
	Propranolol	61.6	67.1	44.9	69.9	26.7	54.0

5. Discussion

Even though the last decade of GPCR research has led to a tremendous amount of new structural data, the three-dimensional coordinates for less than 150 unique GPCRs have been experimentally solved [221]. The development of many mechanistic structure-based GPCR models is therefore still confronted with a lack of structural data. The utilization of molecular modeling methods can close this gap and lead to the development of significantly more precise mechanistic models. Combined with an in-depth data analysis we have developed dynamic mechanistic models of $S1PR_{1-5}$ and unveiled molecular determinants able to explain S1PR subtype selectivity. Furthermore, we report a novel approach for the analysis of GPCR equilibrium simulations regarding receptor activation and conformational dynamics. We have utilized all available class A GPCR structures to develop predictive models which can classify active and inactive receptor conformations.

5.1. Structural Determinants for Sphingosine-1-phosphate Selectivity

Targeting S1PRs is an effective treatment for multiple sclerosis. This has been demonstrated with the first-in-class drug Fingolimod. Due to the non-selective binding to $S1PR_{1,3,4,5}$, therapy with Fingolimod bears a risk of serious adverse events, such as bradycardia and macular edema. The development of subtype-selective S1PR drugs can enhance the drug safety profile while retaining the effectiveness against multiple sclerosis compared to Fingolimod (chapter 1.2.3). S1PR subtype-selective ligands also have considerable potential to modulate pathophysiological processes involved in autoimmune and inflammatory diseases. Hence, understanding the molecular mechanistics of S1PR subtype selectivity can facilitate the rational ligand design of S1PR subtype-selective modulators.

The $S1PR_{2-5}$ derived homology models based on the $S1PR_1$ crystal structure (PDB 3V2Y [32]) revealed the expected secondary and tertiary structure with under 1 Å RMSD of the Cα of the protein backbone between each $S1PR_{1-5}$ structure. Initial structural comparisons of the orthosteric binding pocket of $S1PR_{1-5}$ revealed a general S1PR family-wide comparable binding pocket divided into a hydrophilic and lipophilic sub pocket, which is in line with the amphiphilic character of the endogenous ligand S1P. Binding pocket differences are predominantly located in the lipophilic pocket (chapter 4.1.1). Due to the high conservation

of the residues forming the lipophilic sub pocket in all S1PRs, a clear differentiation between the binding pockets with static 3D pharmacophore models would not adequately render the important spots for selectivity. To address this challenge we developed mechanistic dynamic pharmacophore models for every S1PR subtype and performed a comparative analysis that can protrude selectivity determinants more clearly. We have docked the experimentally validated $S1PR_{1-5}$ pan agonist **6** [216] to the orthosteric binding pocket of all five S1PR subtypes. Three reasons make compound **6** the ideal ligand for this comparative study. Firstly, compound **6** shows different agonistic activity for $S1PR_{1-5}$. Secondly, the agonistic activity was measured with a GTPγS binding assay, which is less subject to amplification or regulation through cellular processes because the monitored GTP exchange occurs early in the signal transduction cascade and is proximal to the receptor activation. Lastly, **6** is chemically different from the endogenous ligand S1P while retaining the overall amphiphilic properties important for S1PR binding.

The dynamic pharmacophore interaction analysis revealed expected recurring and comparable ionic interactions of the phosphate and amine moiety of compound **6** to $S1PR_{1-5}$ key residues $R^{3.28}$ and $E^{3.29}$, respectively. Furthermore, we analyzed the simultaneously occurring hydrogen bonds of compound **6'** phosphate moiety to the conserved residues $S1PR_{1-5}$ polar sub pocket to the phosphate moiety of compound **6** revealing that compound **6** in complex with $S1PR_3$ exhibits the weakest hydrogen bond network from all subtypes. While these results correlate with compound **6** lowest activity in $S1PR_3$ (EC_{50} 630 nM), it should be interpreted as an indicator of weak potency linked to other unfavorable protein ligand interactions. In general, these results can be used as a reference point for the analysis of hydrogen bond networks of structure-based dynamic S1PR models in complex with novel ligands with a phosphate moiety or bioisosteric moiety. This novel S1PR tailored analysis can therefore be utilized to support the rational assessment and selection of potent S1PR agonists from large ligand datasets. Another selectivity determinant we have revealed is π-π aromatic ring (AR) interactions of the phenyl and imidazole moiety of compound **6** exclusively observed in $S1PR_2$. This is explained by the expression of residue $F^{7.39}$ of $S1PR_2$ in comparison to $L^{7.39}$ of $S1PR_1$, $I^{7.39}$ of $S1PR_3$, $L^{7.39}$ of $S1PR_4$, and $L^{7.39}$ of $S1PR_5$. The $F^{7.39}$ determinant supports the high potency of compound **6** in $S1PR_2$ (EC_{50} 18 nM). Residue $F^{7.39}$ switches AR interactions between both aromatic moieties. Exploiting this selectivity determinant could be achieved in two ways. First, an aromatic moiety could be placed more

strategically in a ligand to achieve stable AR interactions for $S1PR_2$ selectivity. Another way could be the addition of bulky lipophilic moieties to a compound to exclude $S1PR_2$ binding due to the constrained access of the $S1PR_2$ lipophilic sub pocket through $F^{7.39}$. The second approach is in line with Ozanimod and Siponimods bulky lipophilic moieties that would cause a clash with $F^{7.39}$ upon binding to $S1PR_2$ leading to a complete affinity loss on this receptor subtype.

The dynamic pharmacophore analysis of compound **6** lipophilic tail moiety in $S1PR_{1\text{-}5}$ revealed two new binding modes distinct from the initial binding mode similar to the $S1PR_1$ crystal structure in complex with ML056 (PDB 3V2Y, binding mode 1). We surmised that binding mode 1 (BM1) represents an inactive conformation based on the antagonist ML056 conformation. The second binding mode (BM2) is proximal to the receptor core, while the third binding mode (BM ECL) is proximal to the extracellular domain. Both BM2 and BM ECL might play an important role in receptor activation. BM1 and BM2 have been observed in all S1PR subtypes, while BM ECL was only present in complex with $S1PR_{2,5}$. Occurrences of $S1PR_{1,3,4,5}$ BM1 correlate with the measured activity of compound **6**, while BM2 occurrences of $S1PR_{1,2,3,4}$ are inversely correlated with the measured activity of **6**. One possible reason why compound **6**' BM1 occurrences of $S1PR_5$ and BM2 occurrences of $S1PR_2$ are not correlating with the measured potency is the inactive S1PR-compound **6** structure models as the basis for conducted MD simulations. This might introduce a bias through stable BM1 conformations trapping the pan-agonist **6** in the energy well of the inactive conformation [262]. By conducting ten replicas of each S1PR MD system we minimized the risk of nonrepresentative mechanistic models and ensured reproducibility and convergence [263]. Interestingly, $S1PR_{2,5}$ BM ECL occurrences of compound **6** could explain the lower BM2 occurrence of $S1PR_{2,5}$. Furthermore, the BM ECL conformation might be exploitable to achieve $S1PR_{2,5}$ selective ligands by introducing a lipophilic substituent in a rigid sharp angle to simultaneously accommodate the hydrophilic binding pocket as well as the BM ECL lipophilic sub pocket proximal to the extracellular domain of $S1PR_{2,5}$. Due to no available data on the importance or (activation) role of BM ECL on $S1PR_{2,5,}$ this result needs further careful evaluation and in vitro testing of novel optimized S1PR ligands to confirm the stated hypothesis.

To evaluate the sequential relation and transition probabilities between the three binding modes we created Markov-chain models for each S1PR subtype. The use of Markov state models has become a popular method for the analysis of conformational changes of proteins and protein ligand binding [264–266]. While the popular python package PyEMMA [267] provides a full framework for handling MD simulation data, incorporation of the systems topology, guide in feature selection, cluster the data into distinct states, estimate Markov state model, and methods to analyze the model [267], we have developed our own workflow to create the Markov-chain models. The workflow is tailored towards the dynamic protein ligand interactions calculated by the Dynophore app with defined states based on distinct protein ligand interactions and labeled the MD simulation frames. The combination of dynamic pharmacophores and Markov chain models revealed transition probabilities is in line with the binding mode occurrences.

New structural data of Ozanimod in complex with $S1PR_1$ and Siponimod in complex with $S1PR_{1,5}$ (PDB 7EW0, 7EVY, and 7EW1, respectively) [219] has been published during the time of this investigation. The comparison of the structural data with our binding modes revealed that Ozanimod and Siponimods lipophilic moieties both interact with the lipophilic sub pockets stabilizing BM1 and BM2 of $S1PR_{1,5}$ confirming our observed and defined BM1 and BM2. Based on the $S1PR_{1\text{-}5}$ subtype selective interaction pattern of compound **6** alkyl moiety with the lipophilic sub pockets targeted within the BM1 and BM2 we were able to explain the selectivity and potency of Ozanimod and Siponimod for all S1PR subtypes.

The main limitation of this study is the inability of our mechanistic models to explain the binding and selectivity of S1PR antagonists due to the subtype selectivity determinants inferred based on the used pan-agonist. Furthermore, the observed $S1PR_{1\text{-}5}$ subtype selective lipophilic interaction patterns of compound **6'** alkyl moiety with the lipophilic sub pockets detected in BM1, BM2, and BM ECL are dependent on stable ionic interactions of the phosphate and amine moiety of compound **6** to $S1PR_{1\text{-}5}$ residue $R^{3.28}$ and $E^{3.29}$, respectively. This should be considered when our selectivity determinants are applied to known S1PR agonists, which do not inherit the amphiphilic structure of S1P or compound **6**. BM ECL has not been experimentally validated. Hence, mutational studies based on our hypothesis for BM ECL $S1PR_{2,5}$ exclusive binding would elucidate BM ECL. Specifically, mutation of $V^{5.40}$ to $I^{5.40}$ in $S1PR_{2,5}$ should diminish compounds **6** $S1PR_{2,5}$ activity.

5.2. Dihedral Angle Dynamics of Class A G Protein-coupled Receptor Activation Hotspots

During the activation of GPCRs different important conformational rearrangements occur, such as an outward tilt of the transmembrane (TM) domain 6. Since the full transition from one GPCR activation state to another is not achievable in all-atom molecular dynamics (MD) simulations, most approaches follow equilibrium simulations starting from a distinct receptor-ligand complex. Subtle changes linked to receptor activation or deactivation are hard to grasp and thus represent a major challenge for the use of MD simulations in the GPCR field. Thus, there is a need to identify new determinants able to grasp these subtle changes. This study explores dihedral angles as possible determinants for machine learning models which can classify static and dynamic structural models as active or inactive conformations. Furthermore, we developed a new method to efficiently predict any given structural class A GPCR model.

The underlying dataset consisted of all active and inactive state experimentally validated class A GPCR structures. By performing a thorough dataset assessment and curation we created a tailored qualitative dataset to reduce noise from the dataset and therefore enhance the predictive power of our created machine learning models. Due to the great amount of unique class A GPCR structures, one of the main challenges was the alignment between GPCRs to create intercomparable features. This was addressed with the use of GPCRdb adjusted generic Ballesteros-Weinstein number (BWN) for TM residues of class A GPCRs. Due to the possible different lengths, insertions, and deletions between class A GPCR TMs some of our dihedral angle extracted features had gaps (missing data point). All features are required to have the same amount of data points to be used for machine learning model development. We considered two ways of handling these gaps. Either remove any feature with gaps or fill the gaps with the mean value of the feature. Both options come with advantages and caveats. By removing all features with gaps the initial feature set (621 dihedral angle features) would shrink to only 22.7% (141 dihedral angle features) and therefore lose a lot of possibly valuable data. On the other hand, filling missing data points with mean values might introduce a significant amount of noise to the data set. We, therefore, tried to find a balance between these two options and only filled in missing values of features if the number of missing values does not exceed 1% of a given feature. This led

to the dataset of 286 dihedral angle features and a balance between retainment of higher information density and noise introduction. To evaluate the impact on the predictive performance of trained machine learning models we used both datasets for model building.

An important part of model building is the selection of meaningful features. We applied three feature selection methods (RFECV, random forest-based selection, and ANOVA F-value) and evaluated the predictive performance with three different machine learning algorithms (DTC, SVM, and KNN). In general, the dataset with filled mean values for 1% missing data performed better than the dataset which removed any feature with missing values. Interestingly one of the top-performing models was the DTC model with applied ANOVA F-value, a variance-based statistical selection method, with only two features selected. This posed an ideal outcome due to the low dimensionality of the dataset (2 dimensional), introspective possibility of the DTC model, and highest possible prediction performance. A detailed analysis of the two features ($BWN^{7.53}$ Φ and $BWN^{7.54}$ Φ) revealed that the positions are in line with the known $N^{7.49}$–$P^{7.50}$–x–x–$Y^{7.53}$ motif important for activation and binding site molding when an IBP is binding. Furthermore, we discovered that the mechanism of structural rearrangement of the $N^{7.49}$–$P^{7.50}$–x–x–$Y^{7.53}$ motif goes hand in hand with an untangling of the two feature positions $BWN^{7.53}$ and $BWN^{7.54}$. An angle analysis with only G-protein activated complexes revealed no difference in angular distribution between Gi/o, Gs, and Gq/11. We have found a universal angle difference of the $N^{7.49}$–$P^{7.50}$–x–x–$Y^{7.53}$ motif for class A GPCR activation, which is not dependent on the type of IBP and is in line with the literature.

The prediction of homology models from the GPCRdb revealed partly mislabeled homology models. Prediction of inactive state labeled homology models was very robust and identified three high RMSD homology models in comparison to their templates. Only three structures were correctly labeled by the GPCRdb but misinterpreted by our model. This was due to the angle values of the inactive state structures being very close to the decision node values and lies in the error margin of the model. Prediction of active state homology models revealed the identification of 13 structures based on active-like structures (protein-ligand complex). 16 Structures that were correctly labeled by the GPCRdb but misinterpreted by our model were based on ternary structures with angles normally found in inactive structures. More surprising were nine homology models which used correctly classified

templates by our model but had different feature angle values and were therefore misclassified by our model. The general architecture of these structures was active state. This might be an indicator of mishandled construction of the protein backbone at the $N^{7.49}$–$P^{7.50}$–x–x–$Y^{7.53}$ motif.

The prediction of four dynamic β_2-adrenoceptor systems with and without active state stabilizing nanobodies and with and without inactive state stabilizing nanobodies revealed the expected shift in active states. Interestingly, the highest difference of the four systems was observed between the active state ternary structure and the active state structure only in complex with the agonist (77.6% active state to 35.5% active state), while the decline in active state frames is less than 10% between the active and inactive state protein-ligand complex and only around 3% between the inactive state protein-ligand complex and inactive state ternary complex. This result shows that without an active state stabilizing IBP the receptor complex probability to reach an inactive state is very high and to maintain a fully active state in MD simulations an IBP is mandatory. The prediction of different active state protein-ligand complexes is in line with these results. We have seen no correlation between the different modes of action of a ligand in complex with an active state structure during MD simulations.

The main limitation of our model is the identification of only fully active and inactive state structures dependent on IBP binding. The analysis and use of dihedral angles as a predictor are therefore limited and could not show active state-like stabilizing properties of agonists. Furthermore, even though our model is easy to understand and is in line with the scientific literature this feature is also one of the main caveats of the model due to the reliance on just one endpoint of the activation process of class A GPCRs. Future applications of new machine learning models should benefit from our found dihedral angle features if combined with more features or used in a sequential prediction model with our angular features as the endpoint for the evaluation of fully active state structures.

6. Conclusion

Understanding GPCRs on a molecular level in both, a smaller subtype-specific and greater class-wide context is essential for the future development of structure-based ligands with potential therapeutic value.

Selectively targeting S1PR subtypes hold a lot of potential to modulate different diseases through their omnipresence in different tissues and regulatory functions as described in chapter 1.2. Using homology models, docking experiments, extensive molecular dynamics simulations, dynamic pharmacophores (Dynophores), and subsequent rigorous data analysis three distinct binding modes in $S1PR_{1-5}$ with subtype-specific residue interaction patterns were discovered. Comparison with new structural data of Ozanimod and Siponimod confirmed the hypothesis of subtype selectivity by targeting sub pockets seen in binding modes 1 and 2 and occurring in the molecular dynamics simulations of S1PR in respective complexes with a pan-agonist. Furthermore, based on our models, we could explain the subtype-selective behavior of both marketed drugs. Hence, we created an S1PR wide holistic model with high explanatory capabilities and provide the structural basis for subtype selectivity in S1PR.

Understanding the general class A GPCR activation mechanisms can lead to generalized methods for fast and accurate analysis of structural (dynamic) models. We have explored dihedral angles as possible structural determinants for GPCR activation and discovered specific residue positions which reveal a change of Φ angles at residue positions 7.53 and 7.54 and subsequent untangling of the α-helix of TM7 in the intracellular part of the receptor. These positions are also in line with the scientific literature and posed useful as machine learning features to distinguish between active and inactive states. We developed a framework to use our models efficiently and showed successful application for batch analysis of homology models as well as MD simulations.

Hence, in this thesis, we have shown that a structural basis through an in-depth analysis of S1PR subtype selectivity and class A GPCR activation determinants can be created to support and deepen our understanding of GPCR research. By applying our developed unique workflows we exploited the scarce data availability for S1PR at the time of investigation and large structural datasets to extract dihedral angle data as determinants for class A GPCR

activation. Employing these workflows to other current and pressing topics for other GPCRs or targets beyond GPCRs can further provide insights in activation and selectivity mechanisms.

7. Experimental Section

Homology Modeling

The homology models of $S1PR_{2-5}$ were built based on the $S1PR_1$ crystal structure (PDB 3V2Y [32]). Protein sequences for the homology models (HM) were obtained from UniProtKB ($S1PR_{1-5}$ accession numbers: P21453, O95136, Q99500, O95977, Q9H228, respectively) [268]. Sequence similarity was calculated with MOE 2013.08 (Chemical Computing Group ULC, Montreal, Canada). Before building the HM, $S1PR_1$ crystal structure involved the excision of the T4-lysozyme fused to TM5 and TM6 at the intracellular part of the receptor. The HM $S1PR_{2-5}$ in this study was built using MOE 2013.08. The force field OPLS-AA [162] with the temperature set on 300 K was used to build ten main chain HM and scored with the integrated GB/VI [269] scoring.

Molecular Docking

All docking experiments in this work were carried out with Cambridge Crystallographic Data Centre's (CCDC) GOLD v5.7.0 [154]. Generated docking conformations have been energy minimized by using the MMFF94 [270] force field.

Compound **6** was docked into the orthosteric binding site of $S1PR_{1-5}$ with constraints based on known interactions important for S1P binding and ML056 binding to $S1PR_1$ (PDB 3V2Y [32], Table 3). The orthosteric binding pocket has been defined in GOLD as 10 Å radius around $S1PR_{1-5}$ residue $BWN^{3.32}$. 100 conformations have been generated without early termination and with diverse solutions (Cluster size 1, RMSD 1.5 Å). The selected scoring function was ChemPLP Fitness Score [218]. The selected docking poses for $S1PR_{1-5}$ used for molecular dynamics (MD) simulations were based on the number of fulfilled constraints and the scores given from the ChemPLP Fitness Score.

The ligands isoprenaline, salbutamol, and propranolol were docked into the orthosteric binding pocket of β_2-adrenoceptor (PDB 3P0G [252]). The orthosteric binding pocket has been defined as 12 Å radius of a defined point (62.229, 19.461, 15.333) visually estimated as the binding pocket center. 100 conformations have been generated without early termination and with diverse solutions (Cluster size 1, RMSD 1.5 Å). The selected scoring

function was Goldscore [218]. Docking pose selection was initially guided based on binding conformations from available structural data at the time as well as the score given by the Goldscore function. Docking of adrenaline, hydroxybenzylisoproterenol, salmeterol, carazolol, ICI-118,551, timolol, and alprenolol was guided by experimentally confirmed ligand conformations from available structural data (PDB 4LDO [257], 4LDL [257], 6MXT [258], 5JQH [253], 3NY9 [256], 3D4S [259], 3NYA [256], respectively) and inserted into the binding pocket of β_2-adrenoceptor (PDB 3P0G [252]) via superposition.

Molecular Dynamics Simulations

Protein preparation was carried out with Maestro v11.7 (Schrödinger, LLC: New York, USA) and involved assigning bond orders, adding hydrogen's, creating disulfide bonds, cap termini, deleting waters beyond 5 Å from the protein surface, and optimization of the hydrogen bond network based on predicted pKa values (PROPKA [271]). Model systems were prepared in an orthotrombic box where the GPCRs are solved with SPC water [272] and 0.15 M sodium chloride and embedded in a POPC-membrane (palmitoyl-oleoyl-phosphatidyl-choline bilayer). Membrane placement has been adjusted to the transmembrane regions defined by the OPM-database [273]. The netto systems charge was neutralized with sodium ions. The systems were parametrized using the OPLS2005 force field [162]. Desmond v2018-3 [274] was used to perform the equilibration of the systems (default Desmond protocol) and MD simulations under periodic boundary conditions as an NPT (constant number of particles, pressure, and temperature) ensemble (300 K temperature and 1.01325 bar pressure). Each system was simulated in ten (S1PRs) or five (β_2-adrenoceptor) replicas for 100 ns with 2000 recorded frames per simulation. The trajectories were wrapped and aligned based on the heavy atoms of the protein backbone of the first frame using VMD 1.9.3 [275]. All RMSD plots were created with mdanalysis 1.0.0 [276, 277].

Dynamic Pharmacophore analysis

The MD simulations of $S1PR_{1\text{-}5}$ in their respective complexes with compound **6** have been analyzed using the in-house developed Dynophore software [179, 177]. After the first 10% of the trajectory, the RMSD has reached a stable plateau for the protein and ligand. All

analysis has been carried out after the first 10% of each trajectory. For every frame of a trajectory, the protein-ligand interactions are extracted as pharmacophores and gathered into spatiotemporal interaction clouds. We developed a workflow for in-depth data analysis within this work. The raw data of the dynophores have been extracted, processed, and analyzed with Python 3.7 [278] and pandas 1.0.1 [279]. Data visualization and plots were created with seaborn 0.11.0 [280, 281]. Three distinct binding modes of compound **6** have been defined based on unique interactions to label the frames of all MD simulations. This enabled us to perform a rigorous comparative analysis of three binding modes between $S1PR_{1\text{-}5}$.

Machine Learning

The data assessment and curation were carried out with Python 3.7 and pandas 1.0.1. Feature extraction of Φ (phi) and ψ (psi) dihedral angles of the protein backbone was done using mdanalysis 1.0.0. Dataframe preparation, feature selection/reduction methods, machine learning algorithms, machine learning model creation, and evaluation have been carried out with Scikit-learn 0.22.1 [282]. To ensure reproducible and deterministic behavior of machine learning algorithms or if randomization is part of a Scikit-learn algorithm (e.g. dataset split into train and test set) the "random state" parameter was set to 5. The feature selection method based on the random forest algorithm was performed with 1000 estimators. The recursive feature elimination with cross-validation with the estimator linear support vector classification eliminates one feature at a time using accuracy as scoring and the stratified k-fold as cross-validation. Hyperparameter search for DTC, KNN, and SVM was done via the brute force method GridSearchCV (accuracy as scoring, five times cross-validation). Hyperparameters used for GridSearchCV of the DTC model: Max_depth - 3, 4, 5, 6, 7, 8; Criterion - gini; Min_samples_split - 2, 3, 4, 5, 6, 7, 8, 9, 10; Min_samples_leaf - 3, 4, 5, 6, 7, 8, 9, 10; Max_leaf_nodes - 3, 4, 5, 6, 7, 8, 9, 10. The final hyperparameter used for the DTC model is 3 for max_depth, 3 for max_leaf_nodes, 3 for min_samples_leaf, and 2 for min_samples_split. KNN hyperparameter: N_neighbours - 3, 4, 5, 6, 7, 8, 9, 10. The final KNN model had an N-neighbours value of 3. SVM hyperparameter: C - 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5, 2.0, 3.0, 4.0; class_weight – balanced. The final SVM model had a C value of 0.2.

We have created a python package that allows for a fast and reliable application of our machine learning model. We provide on GitHub (github.com/TrungNgocNguyen/GPCRml) methods for the automatic assignment of the adjusted BWN from the GPCRdb to a given topology file based on its PDB code or UniProt ID. Furthermore, methods to extract dihedral angles from the provided topology file as well as from MD simulation are implemented to provide the features for using our machine learning model. All three trained models (DTC, KNN, SVM) are provided in the package to predict active and inactive states of class A GPCRs.

Markov chain modeling

The occurring sequential binding mode pairs have been extracted from the dynophore data to calculate a transition probability matrix [283] for each S1PR subtype using Python 3.7 [278] and pandas 1.0.1 [279]. Binding mode labels are given based on the rules defined in chapter 4.1.6. The state transition probability matrices are represented through state transition diagrams [284].

8. References

1. Hilger, D.; Masureel, M.; Kobilka, B. K. Structure and dynamics of GPCR signaling complexes. *Nat. Struct. Mol. Biol.* **2018,** *25* (1), 4–12.
2. Gurevich, V. V.; Gurevich, E. V. Molecular Mechanisms of GPCR Signaling: A Structural Perspective. *Int. J. Mol. Sci.* **2017,** *18* (12), 2519.
3. Pisani, D.; Rota-Stabelli, O.; Feuda, R. Sensory Neuroscience: A Taste for Light and the Origin of Animal Vision. *Curr. Biol.* **2020,** *30* (13), R773-R775.
4. Ricart-Ortega, M.; Font, J.; Llebaria, A. GPCR photopharmacology. *Mol. Cell. Endocrinol.* **2019,** *488*, 36–51.
5. Aldossary, H. S.; Alzahrani, A. A.; Nathanael, D.; Alhuthail, E. A.; Ray, C. J.; Batis, N.; Kumar, P.; Coney, A. M.; Holmes, A. P. G-Protein-Coupled Receptor (GPCR) Signaling in the Carotid Body: Roles in Hypoxia and Cardiovascular and Respiratory Disease. *Int. J. Mol. Sci.* **2020,** *21* (17).
6. Ali, D. C.; Naveed, M.; Gordon, A.; Majeed, F.; Saeed, M.; Ogbuke, M. I.; Atif, M.; Zubair, H. M.; Changxing, L. β-Adrenergic receptor, an essential target in cardiovascular diseases. *Heart Fail. Rev.* **2020,** *25* (2), 343–354.
7. Kendall, D. A.; Yudowski, G. A. Cannabinoid Receptors in the Central Nervous System: Their Signaling and Roles in Disease. *Front. Cell. Neurosci.* **2016,** *10*, 294.
8. Huang, Y.; Todd, N.; Thathiah, A. The role of GPCRs in neurodegenerative diseases: avenues for therapeutic intervention. *Curr. Opin. Pharmacol.* **2017,** *32*, 96–110.
9. Hauser, A. S.; Attwood, M. M.; Rask-Andersen, M.; Schiöth, H. B.; Gloriam, D. E. Trends in GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discov.* **2017,** *16* (12), 829–842.
10. Congreve, M.; Graaf, C. de; Swain, N. A.; Tate, C. G. Impact of GPCR Structures on Drug Discovery. *Cell* **2020,** *181* (1), 81–91.
11. Armstrong, J. F.; Faccenda, E.; Harding, S. D.; Pawson, A. J.; Southan, C.; Sharman, J. L., et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. *Nucleic Acids Res.* **2020,** *48* (D1), D1006-D1021.

12. Fredriksson, R.; Lagerstrom, M. C.; Lundin, L.-G.; Schioth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **2003,** *63* (6), 1256–1272.
13. Schioth, H. B.; Fredriksson, R. The GRAFS classification system of G-protein coupled receptors in comparative perspective. *Gen. Comp. Endocrinol.* **2005,** *142* (1-2), 94–101.
14. DeVree, B. T.; Mahoney, J. P.; Vélez-Ruiz, G. A.; Rasmussen, S. G. F.; Kuszak, A. J.; Edwald, E., et al. Allosteric coupling from G protein to the agonist-binding pocket in GPCRs. *Nature* **2016,** *535* (7610), 182–186.
15. Zhou, Q.; Yang, D.; Wu, M.; Guo, Y.; Guo, W.; Zhong, L., et al. Common activation mechanism of class A GPCRs. *eLife* **2019,** *8*.
16. Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A., et al. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **2000,** *289* (5480), 739–745.
17. Palczewski, K. G protein-coupled receptor rhodopsin. *Annu. Rev. Biochem.* **2006,** *75*, 743–767.
18. Costanzi, S.; Siegel, J.; Tikhonova, I. G.; Jacobson, K. A. Rhodopsin and the others: a historical perspective on structural studies of G protein-coupled receptors. *Curr. Pharm. Des.* **2009,** *15* (35), 3994–4002.
19. Ghosh, E.; Kumari, P.; Jaiman, D.; Shukla, A. K. Methodological advances: the unsung heroes of the GPCR structural revolution. *Nat. Rev. Mol. Cell Biol.* **2015,** *16* (2), 69–81.
20. Thal, D. M.; Vuckovic, Z.; Draper-Joyce, C. J.; Liang, Y.-L.; Glukhova, A.; Christopoulos, A.; Sexton, P. M. Recent advances in the determination of G protein-coupled receptor structures. *Curr. Opin. Struct. Biol.* **2018,** *51*, 28–34.
21. Rasmussen, S. G. F.; Choi, H.-J.; Rosenbaum, D. M.; Kobilka, T. S.; Thian, F. S.; Edwards, P. C., et al. Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **2007,** *450* (7168), 383–387.
22. Rasmussen, S. G. F.; DeVree, B. T.; Zou, Y.; Kruse, A. C.; Chung, K. Y.; Kobilka, T. S., et al. Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature* **2011,** *477* (7366), 549–555.
23. García-Nafría, J.; Tate, C. G. Cryo-Electron Microscopy: Moving Beyond X-Ray Crystal Structures for Drug Receptors and Drug Development. *Annu. Rev. Pharmacol. Toxicol.* **2020,** *60*, 51–71.

24. Safdari, H. A.; Pandey, S.; Shukla, A. K.; Dutta, S. Illuminating GPCR Signaling by Cryo-EM. *Trends Cell Biol.* **2018,** *28* (8), 591–594.
25. Hauser, A. S.; Kooistra, A. J.; Munk, C.; Heydenreich, F. M.; Veprintsev, D. B.; Bouvier, M.; Babu, M. M.; Gloriam, D. E. GPCR activation mechanisms across classes and macro/microscales. *Nat. Struct. Mol. Biol.* **2021,** *28* (11), 879–888.
26. Kooistra, A. J.; Mordalski, S.; Pándy-Szekeres, G.; Esguerra, M.; Mamyrbekov, A.; Munk, C.; Keserű, G. M.; Gloriam, D. E. GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Res.* **2021,** *49* (D1), D335-D343.
27. Ballesteros, J. A.; Weinstein, H. [19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. In *Receptor molecular biology*; Sealfon, S. C., Ed.; Methods in Neurosciences v. 25; Academic Press, **1995**, pp 366–428.
28. Peeters, M. C.; van Westen, G. J. P.; Li, Q.; IJzerman, A. P. Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation. *Trends Pharmacol. Sci.* **2011,** *32* (1), 35–42.
29. Clark, S. D.; Tran, H. T.; Zeng, J.; Reinscheid, R. K. Importance of extracellular loop one of the neuropeptide S receptor for biogenesis and function. *Peptides* **2010,** *31* (1), 130–138.
30. Harterich, S.; Koschatzky, S.; Einsiedel, J.; Gmeiner, P. Novel insights into GPCR-peptide interactions: mutations in extracellular loop 1, ligand backbone methylations and molecular modeling of neurotensin receptor 1. *Bioorg. Med. Chem.* **2008,** *16* (20), 9359–9368.
31. Woolley, M. J.; Conner, A. C. Understanding the common themes and diverse roles of the second extracellular loop (ECL2) of the GPCR super-family. *Mol. Cell. Endocrinol.* **2017,** *449*, 3–11.
32. Hanson, M. A.; Roth, C. B.; Jo, E.; Griffith, M. T.; Scott, F. L.; Reinhart, G., et al. Crystal structure of a lipid G protein-coupled receptor. *Science* **2012,** *335* (6070), 851–855.
33. Claus, M.; Jaeschke, H.; Kleinau, G.; Neumann, S.; Krause, G.; Paschke, R. A hydrophobic cluster in the center of the third extracellular loop is important for thyrotropin receptor signaling. *Endocrinology* **2005,** *146* (12), 5197–5203.

34. Chen, Y.; Green, S. R.; Almazan, F.; Quehenberger, O. The amino terminus and the third extracellular loop of CX3CR1 contain determinants critical for distinct receptor functions. *Mol. Pharmacol.* **2006,** *69* (3), 857–865.
35. Peeters, M. C.; Wisse, L. E.; Dinaj, A.; Vroling, B.; Vriend, G.; IJzerman, A. P. The role of the second and third extracellular loops of the adenosine A1 receptor in activation and allosteric modulation. *Biochem. Pharmacol.* **2012,** *84* (1), 76–87.
36. Bockaert, J.; Pin, J. P. Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J.* **1999,** *18* (7), 1723–1729.
37. Venkatakrishnan, A. J.; Deupi, X.; Lebon, G.; Heydenreich, F. M.; Flock, T.; Miljus, T., et al. Diverse activation pathways in class A GPCRs converge near the G-protein-coupling region. *Nature* **2016,** *536* (7617), 484–487.
38. Stevens, R. C.; Cherezov, V.; Katritch, V.; Abagyan, R.; Kuhn, P.; Rosen, H.; Wuthrich, K. The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. *Nat. Rev. Drug Discov.* **2013,** *12* (1), 25–34.
39. Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963,** *7* (1), 95–99.
40. Morris, A. L.; MacArthur, M. W.; Hutchinson, E. G.; Thornton, J. M. Stereochemical quality of protein structure coordinates. *Proteins* **1992,** *12* (4), 345–364.
41. Wacker, D.; Stevens, R. C.; Roth, B. L. How Ligands Illuminate GPCR Molecular Pharmacology. *Cell* **2017,** *170* (3), 414–427.
42. Bock, A.; Bermudez, M. Allosteric coupling and biased agonism in G protein-coupled receptors. *FEBS J.* **2021.**
43. Nygaard, R.; Frimurer, T. M.; Holst, B.; Rosenkilde, M. M.; Schwartz, T. W. Ligand binding and micro-switches in 7TM receptor structures. *Trends Pharmacol. Sci.* **2009,** *30* (5), 249–259.
44. Lamichhane, R.; Liu, J. J.; Pljevaljcic, G.; White, K. L.; van der Schans, E.; Katritch, V.; Stevens, R. C.; Wüthrich, K.; Millar, D. P. Single-molecule view of basal activity and activation mechanisms of the G protein-coupled receptor β2AR. *Proc. Natl. Acad. Sci. U.S.A* **2015,** *112* (46), 14254–14259.
45. Gavriilidou, A. F. M.; Hunziker, H.; Mayer, D.; Vuckovic, Z.; Veprintsev, D. B.; Zenobi, R. Insights into the Basal Activity and Activation Mechanism of the β1 Adrenergic Receptor Using Native Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2019,** *30* (3), 529–537.

46. Bond, R. A.; Ijzerman, A. P. Recent developments in constitutive receptor activity and inverse agonism, and their potential for GPCR drug discovery. *Trends Pharmacol. Sci.* **2006,** *27* (2), 92–96.
47. Hilger, D.; Masureel, M.; Kobilka, B. K. Structure and dynamics of GPCR signaling complexes. *Nat. Struct. Mol. Biol.* **2018,** *25* (1), 4–12.
48. Tehan, B. G.; Bortolato, A.; Blaney, F. E.; Weir, M. P.; Mason, J. S. Unifying family A GPCR theories of activation. *Pharmacol. Ther.* **2014,** *143* (1), 51–60.
49. Wang, W.; Qiao, Y.; Li, Z. New Insights into Modes of GPCR Activation. *Trends Pharmacol. Sci.* **2018,** *39* (4), 367–386.
50. Hauser, A. S.; Kooistra, A. J.; Munk, C.; Heydenreich, F. M.; Veprintsev, D. B.; Bouvier, M.; Babu, M. M.; Gloriam, D. E. GPCR activation mechanisms across classes and macro/microscales. *Nat. Struct. Mol. Biol.* **2021,** *28* (11), 879–888.
51. Kruse, A. C.; Ring, A. M.; Manglik, A.; Hu, J.; Hu, K.; Eitel, K., et al. Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature* **2013,** *504* (7478), 101–106.
52. Haga, K.; Kruse, A. C.; Asada, H.; Yurugi-Kobayashi, T.; Shiroishi, M.; Zhang, C., et al. Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* **2012,** *482* (7386), 547–551.
53. Xu, F.; Wu, H.; Katritch, V.; Han, G. W.; Jacobson, K. A.; Gao, Z.-G.; Cherezov, V.; Stevens, R. C. Structure of an agonist-bound human A2A adenosine receptor. *Science* **2011,** *332* (6027), 322–327.
54. Jaakola, V.-P.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y. T.; Lane, J. R.; Ijzerman, A. P.; Stevens, R. C. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* **2008,** *322* (5905), 1211–1217.
55. Kenakin, T. Theoretical Aspects of GPCR-Ligand Complex Pharmacology. *Chem. Rev.* **2017,** *117* (1), 4–20.
56. Wang, Y.; Yu, Z.; Xiao, W.; Lu, S.; Zhang, J. Allosteric binding sites at the receptor-lipid bilayer interface: novel targets for GPCR drug discovery. *Drug Discov. Today* **2021,** *26* (3), 690–703.
57. Wu, Y.; Tong, J.; Ding, K.; Zhou, Q.; Zhao, S. GPCR Allosteric Modulator Discovery. *Adv. Exp. Med. Biol.* **2019,** *1163*, 225–251.
58. Ritter, S. L.; Hall, R. A. Fine-tuning of GPCR activity by receptor-interacting proteins. *Nat. Rev. Mol. Cell Biol.* **2009,** *10* (12), 819–830.

59. Pierce, K. L.; Lefkowitz, R. J. Classical and new roles of beta-arrestins in the regulation of G-protein-coupled receptors. *Nat. Rev. Neurol* **2001,** *2* (10), 727–733.
60. Bermudez, M.; Nguyen, T. N.; Omieczynski, C.; Wolber, G. Strategies for the discovery of biased GPCR ligands. *Drug Discov. Today* **2019,** *24* (4), 1031–1037.
61. Alexander, S. P.; Davenport, A. P.; Kelly, E.; Marrion, N.; Peters, J. A.; Benson, H. E., et al. The Concise Guide to PHARMACOLOGY 2015/16: G protein-coupled receptors. *Br. J. Pharmacol.* **2015,** *172* (24), 5744–5869.
62. Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008,** *36* (Database issue), D901-6.
63. Rask-Andersen, M.; Almen, M. S.; Schioth, H. B. Trends in the exploitation of novel drug targets. *Nat. Rev. Drug Discov.* **2011,** *10* (8), 579–590.
64. Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **2006,** *5* (12), 993–996.
65. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D., et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017,** *45* (D1), D945-D954.
66. Sriram, K.; Insel, P. A. G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? *Mol. Pharmacol.* **2018,** *93* (4), 251–258.
67. Insel, P. A.; Tang, C.-M.; Hahntow, I.; Michel, M. C. Impact of GPCRs in clinical medicine: monogenic diseases, genetic variants and drug targets. *Biochim. Biophys. Acta* **2007,** *1768* (4), 994–1005.
68. Shonberg, J.; Kling, R. C.; Gmeiner, P.; Lober, S. GPCR crystal structures: Medicinal chemistry in the pocket. *Bioorg. Med. Chem.* **2015,** *23* (14), 3880–3906.
69. The IDG Knowledge Management Center. Unexplored opportunities in the druggable human genome. *Nat. Rev.* **2016.**
70. Chun, J.; Hla, T.; Lynch, K. R.; Spiegel, S.; Moolenaar, W. H. International Union of Basic and Clinical Pharmacology. LXXVIII. Lysophospholipid receptor nomenclature. *Pharmacol. Rev.* **2010,** *62* (4), 579–587.
71. PYNE, S.; PYNE, N. J. Sphingosine 1-phosphate signalling in mammalian cells. *Biochem. J.* **2000,** *349* (2), 385–402.
72. Takuwa, N.; Okamoto, Y.; Yoshioka, K.; Takuwa, Y. Sphingosine-1-phosphate signaling and cardiac fibrosis. *Inflamm. Regen.* **2013,** *33* (2), 96–108.

73. Takabe, K.; Paugh, S. W.; Milstien, S.; Spiegel, S. "Inside-out" signaling of sphingosine-1-phosphate: therapeutic targets. *Pharmacol. Rev.* **2008,** *60* (2), 181–195.
74. Arish, M.; Husein, A.; Kashif, M.; Saleem, M.; Akhter, Y.; Rub, A. Sphingosine-1-phosphate signaling: unraveling its role as a drug target against infectious diseases. *Drug Discov. Today* **2016,** *21* (1), 133–142.
75. Jozefczuk, E.; Guzik, T. J.; Siedlinski, M. Significance of sphingosine-1-phosphate in cardiovascular physiology and pathology. *Pharmacol. Res.* **2020,** *156*, 104793.
76. Blaho, V. A.; Hla, T. An update on the biology of sphingosine 1-phosphate receptors. *J. Lipid Res.* **2014,** *55* (8), 1596–1608.
77. Marsolais, D.; Rosen, H. Chemical modulators of sphingosine-1-phosphate receptors as barrier-oriented therapeutic molecules. *Nat. Rev. Drug Discov.* **2009,** *8* (4), 297–307.
78. Garris, C. S.; Wu, L.; Acharya, S.; Arac, A.; Blaho, V. A.; Huang, Y., et al. Defective sphingosine 1-phosphate receptor 1 (S1P1) phosphorylation exacerbates TH17-mediated autoimmune neuroinflammation. *Nat. Immunol.* **2013,** *14* (11), 1166–1172.
79. Li, M.-H.; Sanchez, T.; Yamase, H.; Hla, T.; Oo, M. L.; Pappalardo, A.; Lynch, K. R.; Lin, C.-Y.; Ferrer, F. S1P/S1P1 signaling stimulates cell migration and invasion in Wilms tumor. *Cancer Lett.* **2009,** *276* (2), 171–179.
80. Jung, B.; Obinata, H.; Galvani, S.; Mendelson, K.; Ding, B.; Skoura, A., et al. Flow-regulated endothelial S1P receptor-1 signaling sustains vascular development. *Dev. Cell* **2012,** *23* (3), 600–610.
81. Cantalupo, A.; Gargiulo, A.; Dautaj, E.; Liu, C.; Zhang, Y.; Hla, T.; Di Lorenzo, A. S1PR1 (Sphingosine-1-Phosphate Receptor 1) Signaling Regulates Blood Flow and Pressure. *Hypertension* **2017,** *70* (2), 426–434.
82. Skoura, A.; Hla, T. Lysophospholipid receptors in vertebrate development, physiology, and pathology. *J. Lipid Res.* **2009,** *50 Suppl*, S293-8.
83. Xiao, L.; Zhou, Y.; Friis, T.; Beagley, K.; Xiao, Y. S1P-S1PR1 Signaling: the "Sphinx" in Osteoimmunology. *Front. Immunol.* **2019,** *10*, 1409.
84. Sanchez, T.; Skoura, A.; Wu, M. T.; Casserly, B.; Harrington, E. O.; Hla, T. Induction of vascular permeability by the sphingosine-1-phosphate receptor-2 (S1P2R) and its downstream effectors ROCK and PTEN. *Arterioscler. Thromb. Vasc. Biol.* **2007,** *27* (6), 1312–1318.

85. Del Galdo, S.; Vettel, C.; Heringdorf, D. M. Z.; Wieland, T. The activation of RhoC in vascular endothelial cells is required for the S1P receptor type 2-induced inhibition of angiogenesis. *Cell. Signal.* **2013,** *25* (12), 2478–2484.
86. Takashima, S.-I.; Sugimoto, N.; Takuwa, N.; Okamoto, Y.; Yoshioka, K.; Takamura, M.; Takata, S.; Kaneko, S.; Takuwa, Y. G12/13 and Gq mediate S1P2-induced inhibition of Rac and migration in vascular smooth muscle in a manner dependent on Rho but not Rho kinase. *Cardiovasc. Res.* **2008,** *79* (4), 689–697.
87. Adada, M.; Canals, D.; Hannun, Y. A.; Obeid, L. M. Sphingosine-1-phosphate receptor 2. *FEBS J.* **2013,** *280* (24), 6354–6366.
88. Michaud, J.; Im, D.-S.; Hla, T. Inhibitory role of sphingosine 1-phosphate receptor 2 in macrophage recruitment during inflammation. *J. Immunol.* **2010,** *184* (3), 1475–1483.
89. McQuiston, T.; Luberto, C.; Del Poeta, M. Role of sphingosine-1-phosphate (S1P) and S1P receptor 2 in the phagocytosis of Cryptococcus neoformans by alveolar macrophages. *Microbiology (Reading, Engl.)* **2011,** *157* (Pt 5), 1416–1427.
90. Kono, M.; Belyantseva, I. A.; Skoura, A.; Frolenkov, G. I.; Starost, M. F.; Dreier, J. L., et al. Deafness and stria vascularis defects in S1P2 receptor-null mice. *J. Biol. Chem.* **2007,** *282* (14), 10690–10696.
91. MacLennan, A. J.; Carney, P. R.; Zhu, W. J.; Chaves, A. H.; Garcia, J.; Grimes, J. R.; Anderson, K. J.; Roper, S. N.; Lee, N. An essential role for the H218/AGR16/Edg-5/LP B2 sphingosine 1-phosphate receptor in neuronal excitability. *Eur. J. Neurosci.* **2001,** *14* (2), 203–209.
92. Seyedsadr, M. S.; Weinmann, O.; Amorim, A.; Ineichen, B. V.; Egger, M.; Mirnajafi-Zadeh, J.; Becher, B.; Javan, M.; Schwab, M. E. Inactivation of sphingosine-1-phosphate receptor 2 (S1PR2) decreases demyelination and enhances remyelination in animal models of multiple sclerosis. *Neurobiol. Dis.* **2019,** *124*, 189–201.
93. Pitman, M. R.; Lewis, A. C.; Davies, L. T.; Moretti, P. A. B.; Anderson, D.; Creek, D. J.; Powell, J. A.; Pitson, S. M. The sphingosine 1-phosphate receptor 2/4 antagonist JTE-013 elicits off-target effects on sphingolipid metabolism. *Sci. Rep.* **2022,** *12* (1), 454.
94. Laidlaw, B. J.; Gray, E. E.; Zhang, Y.; Ramírez-Valle, F.; Cyster, J. G. Sphingosine-1-phosphate receptor 2 restrains egress of γδ T cells from the skin. *J. Exp. Med.* **2019,** *216* (7), 1487–1496.
95. Chen, T.; Lin, R.; Jin, S.; Chen, R.; Xue, H.; Ye, H.; Huang, Z. The Sphingosine-1-Phosphate/Sphingosine-1-Phosphate Receptor 2 Axis in Intestinal Epithelial Cells

Regulates Intestinal Barrier Function During Intestinal Epithelial Cells-CD4+T-Cell Interactions. *Cell Physiol. Biochem.* **2018,** *48* (3), 1188–1200.

96. Bajwa, A.; Huang, L.; Ye, H.; Dondeti, K.; Song, S.; Rosin, D. L.; Lynch, K. R.; Lobo, P. I.; Li, L.; Okusa, M. D. Dendritic cell sphingosine 1-phosphate receptor-3 regulates Th1-Th2 polarity in kidney ischemia-reperfusion injury. *J. Immunol.* **2012,** *189* (5), 2584–2596.
97. Keul, P.; Lucke, S.; Wnuck Lipinski, K. von; Bode, C.; Graler, M.; Heusch, G.; Levkau, B. Sphingosine-1-phosphate receptor 3 promotes recruitment of monocyte/macrophages in inflammation and atherosclerosis. *Circ. Res.* **2011,** *108* (3), 314–323.
98. Sun, X.; Singleton, P. A.; Letsiou, E.; Zhao, J.; Belvitch, P.; Sammani, S., et al. Sphingosine-1-phosphate receptor-3 is a novel biomarker in acute lung injury. *Am. J. Respir. Cell Mol. Biol.* **2012,** *47* (5), 628–636.
99. Ishii, I.; Ye, X.; Friedman, B.; Kawamura, S.; Contos, J. J. A.; Kingsbury, M. A.; Yang, A. H.; Zhang, G.; Brown, J. H.; Chun, J. Marked perinatal lethality and cellular signaling deficits in mice null for the two sphingosine 1-phosphate (S1P) receptors, S1P(2)/LP(B2)/EDG-5 and S1P(3)/LP(B3)/EDG-3. *J. Biol. Chem.* **2002,** *277* (28), 25152–25159.
100. Gräler, M. H.; Bernhardt, G.; Lipp, M. EDG6, a novel G-protein-coupled receptor related to receptors for bioactive lysophospholipids, is specifically expressed in lymphoid tissue. *Genomics* **1998,** *53* (2), 164–169.
101. Walzer, T.; Chiossone, L.; Chaix, J.; Calver, A.; Carozzo, C.; Garrigue-Antar, L.; Jacques, Y.; Baratin, M.; Tomasello, E.; Vivier, E. Natural killer cell trafficking in vivo requires a dedicated sphingosine 1-phosphate receptor. *Nat. Immunol.* **2007,** *8* (12), 1337–1344.
102. Idzko, M.; Hammad, H.; van Nimwegen, M.; Kool, M.; Müller, T.; Soullié, T.; Willart, M. A. M.; Hijdra, D.; Hoogsteden, H. C.; Lambrecht, B. N. Local application of FTY720 to the lung abrogates experimental asthma by altering dendritic cell function. *J. Clin. Invest.* **2006,** *116* (11), 2935–2944.
103. Terai, K.; Soga, T.; Takahashi, M.; Kamohara, M.; Ohno, K.; Yatsugi, S.; Okada, M.; Yamaguchi, T. Edg-8 receptors are preferentially expressed in oligodendrocyte lineage cells of the rat CNS. *Neuroscience* **2003,** *116* (4), 1053–1062.
104. van Doorn, R.; Lopes Pinheiro, M. A.; Kooij, G.; Lakeman, K.; van het Hof, B.; van der Pol, S. M. A., et al. Sphingosine 1-phosphate receptor 5 mediates the immune quiescence of the human brain endothelial barrier. *J. Neuroinflammation* **2012,** *9* (1), 133.

105. Golfier, S.; Kondo, S.; Schulze, T.; Takeuchi, T.; Vassileva, G.; Achtman, A. H., et al. Shaping of terminal megakaryocyte differentiation and proplatelet development by sphingosine-1-phosphate receptor S1P4. *FASEB J.* **2010,** *24* (12), 4701–4710.

106. Olesch, C.; Ringel, C.; Brüne, B.; Weigert, A. Beyond Immune Cell Migration: The Emerging Role of the Sphingosine-1-phosphate Receptor S1PR4 as a Modulator of Innate Immune Cell Activation. *Mediators Inflamm.* **2017,** *2017*, 6059203.

107. van Doorn, R.; Lopes Pinheiro, M. A.; Kooij, G.; Lakeman, K.; van het Hof, B.; van der Pol, S. M. A., et al. Sphingosine 1-phosphate receptor 5 mediates the immune quiescence of the human brain endothelial barrier. *J. Neuroinflammation* **2012,** *9*, 133.

108. Drouillard, A.; Mathieu, A.-L.; Marçais, A.; Belot, A.; Viel, S.; Mingueneau, M.; Guckian, K.; Walzer, T. S1PR5 is essential for human natural killer cell migration toward sphingosine-1 phosphate. *J. Allergy Clin. Immunol.* **2018,** *141* (6), 2265-2268.e1.

109. Debien, E.; Mayol, K.; Biajoux, V.; Daussy, C.; Aguero, M. G. de; Taillardet, M., et al. S1PR5 is pivotal for the homeostasis of patrolling monocytes. *Eur. J. Immunol.* **2013,** *43* (6), 1667–1675.

110. Sanna, M. G.; Wang, S.-K.; Gonzalez-Cabrera, P. J.; Don, A.; Marsolais, D.; Matheu, M. P., et al. Enhancement of capillary leakage and restoration of lymphocyte egress by a chiral S1P1 antagonist in vivo. *Nat. Chem. Biol.* **2006,** *2* (8), 434–441.

111. Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995,** *268* (5214), 1144–1149.

112. Stepanovska, B.; Huwiler, A. Targeting the S1P receptor signaling pathways as a promising approach for treatment of autoimmune and inflammatory diseases. *Pharmacol. Res.* **2020,** *154*, 104170.

113. Sanford, M. Fingolimod: a review of its use in relapsing-remitting multiple sclerosis. *Drugs* **2014,** *74* (12), 1411–1433.

114. Brinkmann, V.; Pinschewer, D.; Chiba, K.; Feng, L. FTY720: a novel transplantation drug that modulates lymphocyte traffic rather than activation. *Trends Pharmacol. Sci.* **2000,** *21* (2), 49–52.

115. Billich, A.; Bornancin, F.; Dévay, P.; Mechtcheriakova, D.; Urtz, N.; Baumruker, T. Phosphorylation of the immunomodulatory drug FTY720 by sphingosine kinases. *J. Biol. Chem.* **2003,** *278* (48), 47408–47415.

116. Chun, J.; Hartung, H.-P. Mechanism of action of oral fingolimod (FTY720) in multiple sclerosis. *Clin. Neuropharmacol.* **2010,** *33* (2), 91–101.

117. Matloubian, M.; Lo, C. G.; Cinamon, G.; Lesneski, M. J.; Xu, Y.; Brinkmann, V.; Allende, M. L.; Proia, R. L.; Cyster, J. G. Lymphocyte egress from thymus and peripheral lymphoid organs is dependent on S1P receptor 1. *Nature* **2004,** *427* (6972), 355–360.

118. Brinkmann, V.; Davis, M. D.; Heise, C. E.; Albert, R.; Cottens, S.; Hof, R., et al. The immune modulator FTY720 targets sphingosine 1-phosphate receptors. *J. Biol. Chem.* **2002,** *277* (24), 21453–21457.

119. Kappos, L.; Radue, E.-W.; O'Connor, P.; Polman, C.; Hohlfeld, R.; Calabresi, P., et al. A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *N. Engl. J. Med.* **2010,** *362* (5), 387–401.

120. Calabresi, P. A.; Radue, E.-W.; Goodin, D.; Jeffery, D.; Rammohan, K. W.; Reder, A. T., et al. Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Neurol.* **2014,** *13* (6), 545–556.

121. Cohen, J. A.; Barkhof, F.; Comi, G.; Hartung, H.-P.; Khatri, B. O.; Montalban, X., et al. Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis. *N. Engl. J. Med.* **2010,** *362* (5), 402–415.

122. Kappos, L.; O'Connor, P.; Radue, E.-W.; Polman, C.; Hohlfeld, R.; Selmaj, K., et al. Long-term effects of fingolimod in multiple sclerosis: the randomized FREEDOMS extension trial. *Neurology* **2015,** *84* (15), 1582–1591.

123. Camm, J.; Hla, T.; Bakshi, R.; Brinkmann, V. Cardiac and vascular effects of fingolimod: mechanistic basis and clinical implications. *Am. Heart J.* **2014,** *168* (5), 632–644.

124. Koyrakh, L.; Roman, M. I.; Brinkmann, V.; Wickman, K. The heart rate decrease caused by acute FTY720 administration is mediated by the G protein-gated potassium channel I. *Am. J. Transplant.* **2005,** *5* (3), 529–536.

125. Dusaban, S. S.; Chun, J.; Rosen, H.; Purcell, N. H.; Brown, J. H. Sphingosine 1-phosphate receptor 3 and RhoA signaling mediate inflammatory gene expression in astrocytes. *J. Neuroinflammation* **2017,** *14* (1), 111.

126. Pan, S.; Gray, N. S.; Gao, W.; Mi, Y.; Fan, Y.; Wang, X., et al. Discovery of BAF312 (Siponimod), a Potent and Selective S1P Receptor Modulator. *ACS Med. Chem. Lett.* **2013,** *4* (3), 333–337.

127. Kovarik, J. M.; Schmouder, R.; Barilla, D.; Wang, Y.; Kraus, G. Single-dose FTY720 pharmacokinetics, food effect, and pharmacological responses in healthy subjects. *Br. J. Clin. Pharmacol.* **2004,** *57* (5), 586–591.

128. Selmaj, K.; Li, D. K. B.; Hartung, H.-P.; Hemmer, B.; Kappos, L.; Freedman, M. S., et al. Siponimod for patients with relapsing-remitting multiple sclerosis (BOLD): an adaptive, dose-ranging, randomised, phase 2 study. *Lancet Neurol.* **2013,** *12* (8), 756–767.

129. Ludwig Kappos; Gavin Giovannoni; Ralf Gold; Robert J. Fox; Patrick Vermersch; Ralph H.B. Benedict, et al. Long-term Efficacy and Safety of Siponimod in Patients with SPMS: EXPAND Extension Analysis Up to 5 Years (4128). *Neurology* **2020,** *94* (15 Supplement).

130. Kappos, L.; Bar-Or, A.; Cree, B. A. C.; Fox, R. J.; Giovannoni, G.; Gold, R., et al. Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. *Lancet* **2018,** *391* (10127), 1263–1273.

131. Lasa, J. S.; Olivera, P. A.; Bonovas, S.; Danese, S.; Peyrin-Biroulet, L. Safety of S1P Modulators in Patients with Immune-Mediated Diseases: A Systematic Review and Meta-Analysis. *Drug Saf.* **2021.**

132. Scott, F. L.; Clemons, B.; Brooks, J.; Brahmachary, E.; Powell, R.; Dedman, H., et al. Ozanimod (RPC1063) is a potent sphingosine-1-phosphate receptor-1 (S1P1) and receptor-5 (S1P5) agonist with autoimmune disease-modifying activity. *Br. J. Pharmacol.* **2016,** *173* (11), 1778–1792.

133. Olson, A.; Timony, G.; Hartung, J. P.; Smith, H.; Peach, R. J.; Rosen, H.; Pan, C.; Brooks, J.; Boehm, M.; Gujrathi, S. Sa1222 Results of a Safety and Tolerability Study of Rpc1063, a Novel Orally Administered Sphingosine-1-Phosphate Receptor 1 (S1p1r) Agonist, in Healthy Adult Volunteers. *Gastroenterology* **2013,** *144* (5), S-234.

134. Cohen, J. A.; Arnold, D. L.; Comi, G.; Bar-Or, A.; Gujrathi, S.; Hartung, J. P.; Cravets, M.; Olson, A.; Frohna, P. A.; Selmaj, K. W. Safety and efficacy of the selective sphingosine 1-phosphate receptor modulator ozanimod in relapsing multiple sclerosis (RADIANCE): a randomised, placebo-controlled, phase 2 trial. *Lancet Neurol.* **2016,** *15* (4), 373–381.

135. Cohen, J. A.; Comi, G.; Arnold, D. L.; Bar-Or, A.; Selmaj, K. W.; Steinman, L., et al. Efficacy and safety of ozanimod in multiple sclerosis: Dose-blinded extension of a randomized phase II study. *Mult. Scler.* **2019,** *25* (9), 1255–1262.

136. Comi, G.; Kappos, L.; Selmaj, K. W.; Bar-Or, A.; Arnold, D. L.; Steinman, L., et al. Safety and efficacy of ozanimod versus interferon beta-1a in relapsing multiple sclerosis

(SUNBEAM): a multicentre, randomised, minimum 12-month, phase 3 trial. *Lancet Neurol.* **2019,** *18* (11), 1009–1020.

137. Tran, J. Q.; Hartung, J. P.; Peach, R. J.; Boehm, M. F.; Rosen, H.; Smith, H., et al. Results From the First-in-Human Study With Ozanimod, a Novel, Selective Sphingosine-1-Phosphate Receptor Modulator. *J. Clin. Pharmacol.* **2017,** *57* (8), 988–996.

138. Cohen, J. A.; Comi, G.; Selmaj, K. W.; Bar-Or, A.; Arnold, D. L.; Steinman, L., et al. Safety and efficacy of ozanimod versus interferon beta-1a in relapsing multiple sclerosis (RADIANCE): a multicentre, randomised, 24-month, phase 3 trial. *Lancet Neurol.* **2019,** *18* (11), 1021–1033.

139. Sugahara, K.; Maeda, Y.; Shimano, K.; Mogami, A.; Kataoka, H.; Ogawa, K., et al. Amiselimod, a novel sphingosine 1-phosphate receptor-1 modulator, has potent therapeutic efficacy for autoimmune diseases, with low bradycardia risk. *Br. J. Pharmacol.* **2017,** *174* (1), 15–27.

140. Bolli, M. H.; Abele, S.; Binkert, C.; Bravo, R.; Buchmann, S.; Bur, D., et al. 2-imino-thiazolidin-4-one derivatives as potent, orally active S1P1 receptor agonists. *J. Med. Chem.* **2010,** *53* (10), 4198–4211.

141. Piali, L.; Birker-Robaczewska, M.; Lescop, C.; Froidevaux, S.; Schmitz, N.; Morrison, K., et al. Cenerimod, a novel selective S1P1 receptor modulator with unique signaling properties. *Pharmacol. Res. Perspect.* **2017,** *5* (6).

142. Dyckman, A. J. Modulators of Sphingosine-1-phosphate Pathway Biology: Recent Advances of Sphingosine-1-phosphate Receptor 1 (S1P1) Agonists and Future Perspectives. *J. Med. Chem.* **2017,** *60* (13), 5267–5289.

143. Cavasotto, C. N.; Phatak, S. S. Homology modeling in drug discovery: current trends and applications. *Drug Discov. Today* **2009,** *14* (13-14), 676–683.

144. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W., et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011,** *7*, 539.

145. Lipman, D. J.; Pearson, W. R. Rapid and sensitive protein similarity searches. *Science* **1985,** *227* (4693), 1435–1441.

146. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990,** *215* (3), 403–410.

147. Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **2010,** *5* (4), 725–738.

148. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R., et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018,** *46* (W1), W296-W303.

149. Schaller, D.; Hagenow, S.; Stark, H.; Wolber, G. Ligand-guided homology modeling drives identification of novel histamine H3 receptor ligands. *PLoS ONE* **2019,** *14* (6), e0218820.

150. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* **2002,** *16* (3), 151–166.

151. Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003,** *32*, 335–373.

152. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009,** *30* (16), 2785–2791.

153. Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **1998,** *33* (3), 367–382.

154. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997,** *267* (3), 727–748.

155. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996,** *261* (3), 470–489.

156. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T., et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004,** *47* (7), 1739–1749.

157. Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002,** *47* (4), 409–443.

158. Dias, R.; Azevedo Jr., W. de. Molecular Docking Algorithms. *Curr. Drug Targets* **2008,** *9* (12), 1040–1047.

159. Mortier, J.; Rakers, C.; Bermudez, M.; Murgueitio, M. S.; Riniker, S.; Wolber, G. The impact of molecular dynamics on drug design: applications for the characterization of ligand-macromolecule complexes. *Drug Discov. Today* **2015,** *20* (6), 686–702.

160. Walton, E. B.; Vanvliet, K. J. Equilibration of experimentally determined protein structures for molecular dynamics simulation. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2006,** *74* (6 Pt 1), 61901.

161. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., JR. Computational methods in drug discovery. *Pharmacol. Rev.* **2014,** *66* (1), 334–395.

162. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001,** *105* (28), 6474–6487.

163. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J., et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010,** *31* (4), 671–690.

164. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004,** *25* (9), 1157–1174.

165. Bermudez, M.; Mortier, J.; Rakers, C.; Sydow, D.; Wolber, G. More than a look into a crystal ball: Protein structure elucidation guided by molecular dynamics simulations. *Drug Discov. Today* **2016,** *21* (11), 1799–1805.

166. K. J. Bowers; D. E. Chow; H. Xu; R. O. Dror; M. P. Eastwood; B. A. Gregersen, et al. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, **2006**, p 43.

167. Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An overview of the Amber biomolecular simulation package. *WIREs Comput. Mol. Sci.* **2013,** *3* (2), 198–210.

168. Berendsen, H.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995,** *91* (1-3), 43–56.

169. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996,** *118* (45), 11225–11236.

170. Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.* **1998,** *70* (5).

171. Thomas Seidel; Gökhan Ibis; Fabian Bendix; Gerhard Wolber. Strategies for 3D pharmacophore-based virtual screening. *Drug Discov. Today Technol.* **2010,** *7* (4), e203-70.

172. Wolber, G.; Dornhofer, A. A.; Langer, T. Efficient overlay of small organic molecules using 3D pharmacophores. *J. Comput. Aided Mol. Des.* **2006,** *20* (12), 773–788.

173. Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005,** *45* (1), 160–169.

174. Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. PHASE: a novel approach to pharmacophore modeling and 3D database searching. *Chem. Biol. Drug Des.* **2006,** *67* (5), 370–372.

175. Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **1996,** *36* (3), 563–571.

176. Koes, D. R.; Camacho, C. J. Pharmer: efficient and exact pharmacophore search. *J. Chem. Inf. Model.* **2011,** *51* (6), 1307–1314.

177. Bock, A.; Bermudez, M.; Krebs, F.; Matera, C.; Chirinda, B.; Sydow, D., et al. Ligand Binding Ensembles Determine Graded Agonist Efficacies at a G Protein-coupled Receptor. *J. Biol. Chem.* **2016,** *291* (31), 16375–16389.

178. Schaller, D.; Šribar, D.; Noonan, T.; Deng, L.; Nguyen, T. N.; Pach, S.; Machalz, D.; Bermudez, M.; Wolber, G. Next generation 3D pharmacophore modeling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2020,** *10* (4).

179. Sydow, D. Dynophores: Novel Dynamic Pharmacophores - Implementation of Pharmacophore Generation Based on Molecular Dynamics Trajectories and Their Graphical Representation.

180. Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **2015,** *349* (6245), 255–260.

181. Panch, T.; Szolovits, P.; Atun, R. Artificial intelligence, machine learning and health systems. *J. Glob. Health* **2018,** *8* (2), 20303.

182. Jones, L. D.; Golan, D.; Hanna, S. A.; Ramachandran, M. Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern? *Bone Jt. Res.* **2018,** *7* (3), 223–225.

183. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018,** *559* (7715), 547–555.

184. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G., et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019,** *18* (6), 463–477.

185. Nichols, J. A.; Herbert Chan, H. W.; Baker, M. A. B. Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophys. Rev.* **2019,** *11* (1), 111–118.

186. Kan, A. Machine learning applications in cell image analysis. *Immunol. Cell Biol.* **2017,** *95* (6), 525–530.

187. Kourou, K.; Exarchos, T. P.; Exarchos, K. P.; Karamouzis, M. V.; Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015,** *13*, 8–17.

188. Libbrecht, M. W.; Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015,** *16* (6), 321–332.

189. Muhammad, I.; Yan, Z. SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *IJSC* **2015,** *05* (03), 946–952.

190. F.Y, O.; J.E.T, A.; O, A.; J. O, H.; O, O.; J, A. *Supervised Machine Learning Algorithms: Classification and Comparison*; Vol. 48, **2017.**

191. Praveena, M.; Jaiganesh, V. A Literature Review on Supervised Machine Learning Algorithms and Boosting Process. *Int. J. Comput. Appl.* **2017,** *169* (8), 32–35.

192. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1986,** *1* (1), 81–106.

193. Blockeel, H.; Raedt, L. de. Top-down induction of first-order logical decision trees. *Artif. Intell.* **1998,** *101* (1-2), 285–297.

194. Strobl, C.; Malley, J.; Tutz, G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* **2009,** *14* (4), 323–348.

195. Fix, E.; Hodges, J. L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *Int. Stat. Rev.* **1989,** *57* (3), 238.

196. Abu Alfeilat, H. A.; Hassanat, A. B. A.; Lasassmeh, O.; Tarawneh, A. S.; Alhasanat, M. B.; Eyal Salman, H. S.; Prasath, V. B. S. Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data* **2019,** *7* (4), 221–248.

197. Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational*

Learning Theory, July 27-29, 1992, Pittsburgh, Pennsylvania; Association for Computing Machinery: New York, N.Y., **1992.**

198. Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* **2006,** *24* (12), 1565–1567.

199. M.A., A.; Thomas, P. A. Comparative Review of Feature Selection and Classification modeling. In *International Conference on Advances in Computing, Communication and Control - 2019: December 20 & 21, 2019 : venue: Samvaad Autitorium, Fr. Conceicao Rodrigues College of Engineering, Bandstand, Bandra (W.), Mumbai 4000 050, Maharashtra, India*; IEEE: [Piscataway, NJ], **2019**, pp 1–9.

200. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; Liu, H. Feature Selection. *Comput. Surv.* **2018,** *50* (6), 1–45.

201. Abdulsalam, S. O.; Mohammed, A. A.; Ajao, J. F.; Babatunde, R. S.; Ogundokun, R. O.; Nnodim, C. T.; Arowolo, M. O. Performance Evaluation of ANOVA and RFE Algorithms for Classifying Microarray Dataset Using SVM. In *Information Systems: 17th European, Mediterranean, and Middle Eastern Conference, EMCIS 2020, Dubai, United Arab Emirates, November 25-26, 2020, Proceedings*; Themistocleous, M., Papadaki, M., Kamal, M. M., Eds.; Springer International Publishing; Springer: Cham, **2020**, pp 480–492.

202. Héberger, K.; Kollár-Hunek, K. Comparison of validation variants by sum of ranking differences and ANOVA. *J. Chemom.* **2019,** *33* (6).

203. Zhou, H.; Zhang, Y.; Zhang, Y.; Liu, H. Feature selection based on conditional mutual information: minimum conditional relevance and minimum conditional redundancy. *Appl. Intell.* **2019,** *49* (3), 883–896.

204. Mielniczuk, J.; Teisseyre, P. Stopping rules for mutual information-based feature selection. *Neurocomputing* **2019,** *358*, 255–274.

205. Sharma, S.; Jain, A. An Empirical Evaluation of Correlation Based Feature Selection for Tweet Sentiment Classification. In *Advances in cybernetics, cognition, and machine learning for communication technologies,* 1st ed. 2020; Gunjan, V. K., Senatore, S., Kumar, A., Gao, X.-Z., Merugu, S., Eds.; Lecture Notes in Electrical Engineering Vol. 643; Springer Singapore, **2020**, pp 199–208.

206. Che, X.; Chen, D.; Mi, J. A novel approach for learning label correlation with application to feature selection of multi-label data. *Inf. Sci.* **2020,** *512*, 795–812.

207. Misra, P. and Yadav, A. S. Improving the classification accuracy using recursive feature elimination with cross-validation. *Int. J. Emerg. Technol. Learn.* **2020** (11(3)), 659–665.

208. Bolón-Canedo, V.; Alonso-Betanzos, A. Ensembles for feature selection: A review and future trends. *Inf. Fusion* **2019,** *52*, 1–12.

209. Kewat, A.; Srivastava, P. N.; Kumhar, D. Performance Evaluation of Wrapper-Based Feature Selection Techniques for Medical Datasets. In *Advances in Computing and Intelligent Systems*: *Proceedings of ICACM 2019*; Sharma, H., Govindan, K., Poonia, R. C., Kumar, S., El-Medany, W. M., Eds.; Springer Singapore; Imprint Springer: Singapore, **2020,** pp 619–633.

210. Sharma, H., Govindan, K., Poonia, R. C., Kumar, S., El-Medany, W. M., Eds. *Advances in Computing and Intelligent Systems*: *Proceedings of ICACM 2019*, 1st ed. 2020; Algorithms for Intelligent Systems; Springer Singapore; Imprint Springer, **2020.**

211. Raschka, S. *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning.* https://arxiv.org/pdf/1811.12808.

212. Merghadi, A.; Yunus, A. P.; Dou, J.; Whiteley, J.; ThaiPham, B.; Bui, D. T.; Avtar, R.; Abderrahmane, B. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Sci. Rev.* **2020,** *207*, 103225.

213. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020,** *21* (1), 6.

214. Ramakrishnan, C.; Ramachandran, G. N. Stereochemical Criteria for Polypeptide and Protein Chain Conformations. *Biophys. J.* **1965,** *5* (6), 909–933.

215. Goujon, M.; McWilliam, H.; Li, W.; Valentin, F.; Squizzato, S.; Paern, J.; Lopez, R. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* **2010,** *38* (Web Server issue), W695-9.

216. Clemens, J. J.; Davis, M. D.; Lynch, K. R.; Macdonald, T. L. Synthesis of 4(5)-phenylimidazole-based analogues of sphingosine-1-phosphate and FTY720: discovery of potent S1P1 receptor agonists. *Bioorganic Med. Chem. Lett.* **2005,** *15* (15), 3568–3572.

217. Parrill, A. L.; Wang, D.; Bautista, D. L.; van Brocklyn, J. R.; Lorincz, Z.; Fischer, D. J.; Baker, D. L.; Liliom, K.; Spiegel, S.; Tigyi, G. Identification of Edg1 receptor residues that recognize sphingosine 1-phosphate. *J. Biol. Chem.* **2000,** *275* (50), 39379–39384.

218. Korb, O.; Stützle, T.; Exner, T. E. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009,** *49* (1), 84–96.

219. Yuan, Y.; Jia, G.; Wu, C.; Wang, W.; Cheng, L.; Li, Q., et al. Structures of signaling complexes of lipid receptors S1PR1 and S1PR5 reveal mechanisms of activation and drug recognition. *Cell Res.* **2021,** *31* (12), 1263–1274.

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

221. Pándy-Szekeres, G.; Esguerra, M.; Hauser, A. S.; Caroli, J.; Munk, C.; Pilger, S.; Keserű, G. M.; Kooistra, A. J.; Gloriam, D. E. The G protein database, GproteinDb. *Nucleic Acids Res.* **2022,** *50* (D1), D518-D525.

222. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28* (1), 235–242.

223. Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G. V., et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **2021,** *49* (D1), D437-D451.

224. Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **2010,** *66* (Pt 1), 12–21.

225. Smith, D. K.; Radivojac, P.; Obradovic, Z.; Dunker, A. K.; Zhu, G. Improved amino acid flexibility parameters. *Protein Sci.* **2003,** *12* (5), 1060–1072.

226. Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **1992,** *355* (6359), 472–475.

227. Yang, H.; Peisach, E.; Westbrook, J. D.; Young, J.; Berman, H. M.; Burley, S. K. DCC: a Swiss army knife for structure factor analysis and validation. *J. Appl. Crystallogr.* **2016,** *49* (Pt 3), 1081–1084.

228. Blow, D. *Outline of crystallography for biologists*, Reprint. with corr; Oxford University Press, **2002, 2010.**

229. Lamb, A. L.; Kappock, T. J.; Silvaggi, N. R. You are lost without a map: Navigating the sea of protein structures. *Biochim. Biophys. Acta* **2015,** *1854* (4), 258–268.

230. Shao, C.; Liu, Z.; Yang, H.; Wang, S.; Burley, S. K. Outlier analyses of the Protein Data Bank archive using a probability-density-ranking approach. *Sci. Data* **2018,** *5* (1), 180293.

231. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S., et al. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007,** *318* (5854), 1258–1265.

232. Pándy-Szekeres, G.; Munk, C.; Tsonkov, T. M.; Mordalski, S.; Harpsøe, K.; Hauser, A. S.; Bojarski, A. J.; Gloriam, D. E. GPCRdb in 2018: adding GPCR structure models and ligands. *Nucleic Acids Res.* **2018,** *46* (D1), D440-D446.

233. Zhang, X.; He, C.; Wang, M.; Zhou, Q.; Yang, D.; Zhu, Y., et al. Structures of the human cholecystokinin receptors bound to agonists and antagonists. *Nat. Chem. Biol.* **2021,** *17* (12), 1230–1237.

234. Yang, Z.; Han, S.; Keller, M.; Kaiser, A.; Bender, B. J.; Bosse, M., et al. Structural basis of ligand binding modes at the neuropeptide Y Y1 receptor. *Nature* **2018,** *556* (7702), 520–524.

235. Tan, Q.; Zhu, Y.; Li, J.; Chen, Z.; Han, G. W.; Kufareva, I., et al. Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science* **2013,** *341* (6152), 1387–1390.

236. Xu, X.; Kaindl, J.; Clark, M. J.; Hübner, H.; Hirata, K.; Sunahara, R. K.; Gmeiner, P.; Kobilka, B. K.; Liu, X. Binding pathway determines norepinephrine selectivity for the human β1AR over β2AR. *Cell Res.* **2021,** *31* (5), 569–579.

237. Duan, J.; Xu, P.; Cheng, X.; Mao, C.; Croll, T.; He, X., et al. Structures of full-length glycoprotein hormone receptor signalling complexes. *Nature* **2021,** *598* (7882), 688–692.

238. Lebon, G.; Edwards, P. C.; Leslie, A. G. W.; Tate, C. G. Molecular Determinants of CGS21680 Binding to the Human Adenosine A2A Receptor. *Mol. Pharmacol.* **2015,** *87* (6), 907–915.

239. Krumm, B. E.; White, J. F.; Shah, P.; Grisshammer, R. Structural prerequisites for G-protein activation by the neurotensin receptor. *Nat. Commun.* **2015,** *6*, 7895.

240. Zhang, H.; Han, G. W.; Batyuk, A.; Ishchenko, A.; White, K. L.; Patel, N., et al. Structural basis for selectivity and diversity in angiotensin II receptors. *Nature* **2017,** *544* (7650), 327–332.

241. Morimoto, K.; Suno, R.; Hotta, Y.; Yamashita, K.; Hirata, K.; Yamamoto, M.; Narumiya, S.; Iwata, S.; Kobayashi, T. Crystal structure of the endogenous agonist-bound prostanoid receptor EP3. *Nat. Chem. Biol.* **2019,** *15* (1), 8–10.

242. Koehl, A.; Hu, H.; Maeda, S.; Zhang, Y.; Qu, Q.; Paggi, J. M., et al. Structure of the μ-opioid receptor-Gi protein complex. *Nature* **2018,** *558* (7711), 547–552.

243. Lin, X.; Li, M.; Wang, N.; Wu, Y.; Luo, Z.; Guo, S., et al. Structural basis of ligand recognition and self-activation of orphan GPR52. *Nature* **2020,** *579* (7797), 152–157.

244. Kim, K.; Che, T.; Panova, O.; DiBerto, J. F.; Lyu, J.; Krumm, B. E., et al. Structure of a Hallucinogen-Activated Gq-Coupled 5-HT2A Serotonin Receptor. *Cell* **2020,** *182* (6), 1574-1588.e19.

245. Wasilko, D. J.; Johnson, Z. L.; Ammirati, M.; Che, Y.; Griffor, M. C.; Han, S.; Wu, H. Structural basis for chemokine receptor CCR6 activation by the endogenous protein ligand CCL20. *Nat. Commun.* **2020,** *11* (1), 3031.

246. Xia, R.; Wang, N.; Xu, Z.; Lu, Y.; Song, J.; Zhang, A.; Guo, C.; He, Y. Cryo-EM structure of the human histamine H1 receptor/Gq complex. *Nat. Commun.* **2021,** *12* (1), 2086.

247. Huang, S.; Xu, P.; Tan, Y.; You, C.; Zhang, Y.; Jiang, Y.; Xu, H. E. Structural basis for recognition of anti-migraine drug lasmiditan by the serotonin receptor 5-HT1F-G protein complex. *Cell Res.* **2021,** *31* (9), 1036–1038.

248. Okamoto, H. H.; Miyauchi, H.; Inoue, A.; Raimondi, F.; Tsujimoto, H.; Kusakizako, T., et al. Cryo-EM structure of the human MT1-Gi signaling complex. *Nat. Struct. Mol. Biol.* **2021,** *28* (8), 694–701.

249. Xu, P.; Huang, S.; Zhang, H.; Mao, C.; Zhou, X. E.; Cheng, X., et al. Structural insights into the lipid and ligand regulation of serotonin receptors. *Nature* **2021,** *592* (7854), 469–473.

250. Zhuang, Y.; Xu, P.; Mao, C.; Wang, L.; Krumm, B.; Zhou, X. E., et al. Structural insights into the human D1 and D2 dopamine receptor signaling complexes. *Cell* **2021,** *184* (4), 931-942.e18.

251. Wang, L.; Xu, J.; Cao, S.; Sun, D.; Liu, H.; Lu, Q.; Liu, Z.; Du, Y.; Zhang, C. Cryo-EM structure of the AVP-vasopressin receptor 2-Gs signaling complex. *Cell Res.* **2021,** *31* (8), 932–934.

252. Rasmussen, S. G. F.; Choi, H.-J.; Fung, J. J.; Pardon, E.; Casarosa, P.; Chae, P. S., et al. Structure of a nanobody-stabilized active state of the β(2) adrenoceptor. *Nature* **2011,** *469* (7329), 175–180.

253. Staus, D. P.; Strachan, R. T.; Manglik, A.; Pani, B.; Kahsai, A. W.; Kim, T. H., et al. Allosteric nanobodies reveal the dynamic range and diverse mechanisms of G-protein-coupled receptor activation. *Nature* **2016,** *535* (7612), 448–452.

254. Velmurugan, B. K.; Baskaran, R.; Huang, C.-Y. Detailed insight on β-adrenoceptors as therapeutic targets. *Biomed. Pharmacother.* **2019,** *117*, 109039.

255. Oliver, E.; Mayor, F.; D'Ocon, P. Beta-blockers: Historical Perspective and Mechanisms of Action. *Rev. Esp. Cardiol. (Engl. Ed.)* **2019,** *72* (10), 853–862.

256. Wacker, D.; Fenalti, G.; Brown, M. A.; Katritch, V.; Abagyan, R.; Cherezov, V.; Stevens, R. C. Conserved binding mode of human beta2 adrenergic receptor inverse agonists and antagonist revealed by X-ray crystallography. *J. Am. Chem. Soc.* **2010,** *132* (33), 11443–11445.

257. Ring, A. M.; Manglik, A.; Kruse, A. C.; Enos, M. D.; Weis, W. I.; Garcia, K. C.; Kobilka, B. K. Adrenaline-activated structure of β2-adrenoceptor stabilized by an engineered nanobody. *Nature* **2013,** *502* (7472), 575–579.

258. Masureel, M.; Zou, Y.; Picard, L.-P.; van der Westhuizen, E.; Mahoney, J. P.; Rodrigues, J. P. G. L. M., et al. Structural insights into binding specificity, efficacy and bias of a β2AR partial agonist. *Nat. Chem. Biol.* **2018,** *14* (11), 1059–1066.

259. Hanson, M. A.; Cherezov, V.; Griffith, M. T.; Roth, C. B.; Jaakola, V.-P.; Chien, E. Y. T.; Velasquez, J.; Kuhn, P.; Stevens, R. C. A specific cholesterol binding site is established by the 2.8 A structure of the human beta2-adrenergic receptor. *Structure* **2008,** *16* (6), 897–905.

260. Yang, F.; Ling, S.; Zhou, Y.; Zhang, Y.; Lv, P.; Liu, S., et al. Different conformational responses of the β2-adrenergic receptor-Gs complex upon binding of the partial agonist salbutamol or the full agonist isoprenaline. *Natl. Sci. Rev.* **2021,** *8* (9).

261. Ishchenko, A.; Stauch, B.; Han, G. W.; Batyuk, A.; Shiriaeva, A.; Li, C., et al. Toward G protein-coupled receptor structure-based drug design using X-ray lasers. *IUCrJ* **2019,** *6* (Pt 6), 1106–1119.

262. Deupi, X.; Kobilka, B. K. Energy landscapes as a tool to integrate GPCR structure, dynamics, and function. *Physiology (Bethesda)* **2010,** *25* (5), 293–303.

263. Gapsys, V.; Groot, B. L. de. On the importance of statistics in molecular simulations for thermodynamics, kinetics and simulation box size. *Elife* **2020,** *9*.

264. Ge, Y.; Voelz, V. A. Markov State Models to Elucidate Ligand Binding Mechanism. *Methods Mol. Biol.* **2021,** *2266*, 239–259.

265. Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018,** *140* (7), 2386–2396.

266. Plattner, N.; Noé, F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat Commun* **2015,** *6* (1), 7653.

267. Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015,** *11* (11), 5525–5542.

268. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021,** *49* (D1), D480-D489.

269. Labute, P. The generalized Born/volume integral implicit solvent model: estimation of the free energy of hydration using London dispersion instead of atomic surface area. *J. Comput. Chem.* **2008,** *29* (10), 1693–1698.

270. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996,** *17* (5-6), 490–519.

271. Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011,** *7* (2), 525–537.

272. Mark, P.; Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **2001,** *105* (43), 9954–9960.

273. Lomize, M. A.; Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I. OPM: orientations of proteins in membranes database. *Bioinformatics* **2006,** *22* (5), 623–625.

274. Horner-Miller, B., Ed. *Proceedings of the 2006 ACM/IEEE conference on Supercomputing - SC '06;* ACM Press, **2006**.

275. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996,** *14* (1), 33–38.

276. Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **2011,** *32* (10), 2319–2327.

277. Gowers, R.; Linke, M.; Barnoud, J.; Reddy, T.; Melo, M.; Seyler, S., et al. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In *Proceedings of the 15th Python in Science Conference*; SciPy, **2016**, pp 98–105.

278. G. van Rossum. Python tutorial: Technical Report CS-R9526 **1995.**

279. McKinney. Data structures for statistical computing in python. *SciPy* **2010,** *445*, 51.

280. Waskom, M. seaborn: statistical data visualization. *JOSS* **2021,** *6* (60), 3021.

281. Thomas A Caswell; Michael Droettboom; Antony Lee; John Hunter; Elliott Sales de Andrade; Eric Firing, et al. *matplotlib/matplotlib: REL: v3.3.3;* Zenodo, **2020.**

282. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O., et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011.**

283. Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018,** *140* (7), 2386–2396.

284. Kachapova. Representing Markov chains with transition diagrams. *J. Math. Stat.* **2013,** *9* (3), 149–154.

9. Publications

Peer-reviewed articles

1. Pach, S.*; Nguyen, T.N.*; Trimpert, J.; Kunec, D.; Osterrieder, N.; Wolber, G. ACE2-Variants Indicate Potential SARS-CoV-2-Susceptibility in Animals: An Extensive Molecular Dynamics Study. *Mol. Inf.* **2021**, 40 (9).
2. Ottanà, R.; Paoli, P.; Cappiello, M.; Nguyen, T.N.; Adornato, I.; Del Corso, A.; Genovese, M.; Nesi, I.; Moschini, R.; Naß, A.; Wolber, G.; Maccari, R. In Search for Multi-Target Ligands as Potential Agents for Diabetes Mellitus and Its Complications—A Structure-Activity Relationship Study on Inhibitors of Aldose Reductase and Protein Tyrosine Phosphatase 1B. *Molecules* **2021**, 26, 330.
3. Schaller, D.*; Šribar, D.*; Noonan, T.*; Deng, L.; Nguyen, T.N.; Pach, S.; Machalz, D.; Bermudez, M.; Wolber, G. Next generation 3D pharmacophore modeling. WIREs Computational *Mol. Sci.* **2020**, 10 (4).
4. Denzinger, K.; Nguyen, T.N.; Noonan, T.; Wolber, G.; Bermudez, M. Biased Ligands Differentially Shape the Conformation of the Extracellular Loop Region in 5-HT_{2B} Receptors. *Int. J. Mol. Sci.* **2020**, 21, 9728.
5. Bermudez, M.; Nguyen, T.N.; Omieczynski, C.; Wolber, G. Strategies for the discovery of biased GPCR ligands. *Drug Discov. Today* **2019**, 24, 1031–1037.

* Shared authorship.

Pre-print articles

6. Omieczynski, C.*; Nguyen, T.N.*; Sribar, D.; Deng, L.; Stepanov, D.; Schaller, D.; Wolber, G.; Bermudez, M. Biasdb: A comprehensive database for biased gpcr ligands. bioRxiv **2019**, 742643.
7. Noonan, T.C.; Denzinger, K.; Talagayev, V. ; Chen, Y.; Puls, K.; Wolf, C.A.; Liu S.; Nguyen, T.N.; Wolber, G. Mind the Gap – Deciphering GPCR Pharmacology Using 3D Pharmacophores and Artificial Intelligence. *Preprints* **2022**, 2022090201

* Shared authorship.

Publications

Oral Presentations at Conferences

1. ACE2-variants indicate potential SARS-CoV-2 susceptibility in animals: data analysis. *EUROPIN Summer School of Drug Design*, Vienna, Austria **2021**.
2. Sphingosine-1-phosphate receptors: Determinants for subtype selectivity. *EUROPIN Retreat*, online **2020**.
3. Dihedral angle dynamics of GPCR activation hotspots. *EUROPIN Summer School of Drug Design*, Vienna, Austria **2019**.
4. Rationalizing signaling of GPCRs with mechanistic models. *EUROPIN Summer School of Drug Design*, Vienna, Austria **2017**.

Poster Presentations at Conferences

1. Discriminating GPCR activation states using machine learning. *Frontiers in Medicinal Chemistry*, Würzburg, Germany **2019**
2. Machine learning methods to predict different activation states of GPCRs. *11th International Conference on Chemical Structures,* Noordwijkerhout, Netherlands **2018**.
3. Structural determinants of Sphingosine-1-phosphate receptor subtype selectivity. *1st Berlin Symposium for Interdisciplinary GPCR research*, Berlin, Germany **2018**
4. A structure-based workflow for novel Sphingosine-1-phosphate receptor 2 ligands. *EUROPIN Retreat*, Vienna, Austria **2017**.

10. Appendix

10.1. List of Abbrevations

AR Aromatic ring interaction

BM Binding mode

BWN Ballesteros-Weinstein number

Cryo-EM cryogenic electron microscopy

DTC Decision tree classification

ECL Extracellular loop

GPCR G protein-coupled receptor

HBA Hydrogen bond acceptor

HBD Hydrogen bond donor

HM Homology model

IBP Intracellular binding partner

ICL Intracellular loop

KNN k-nearest neighbour

MCC Mathews correlation coefficient

MD Molecular dynamics

ML Machine Learning

NI Negative ionizable

PDB Protein data bank

PI Positive ionizable

RFSEL Random forest based selection

RMSD Root-mean-square deviation

S1P Sphingosine-1-phosphate

S1PR Sphingosine-1-phosphate receptor

SVM Support vector machine

TM Transmembrane domain

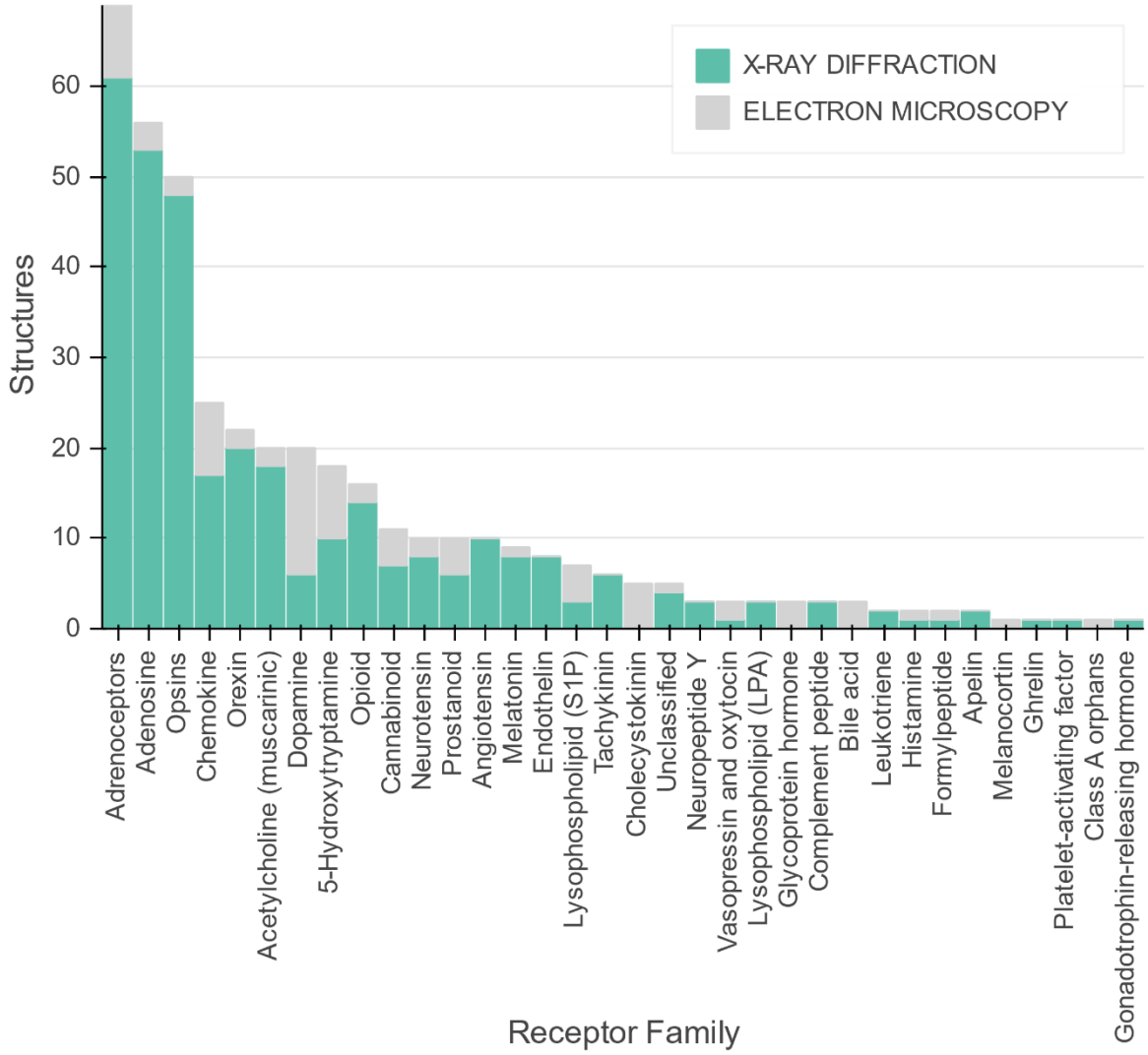
10.2. List of Figures

Appendix

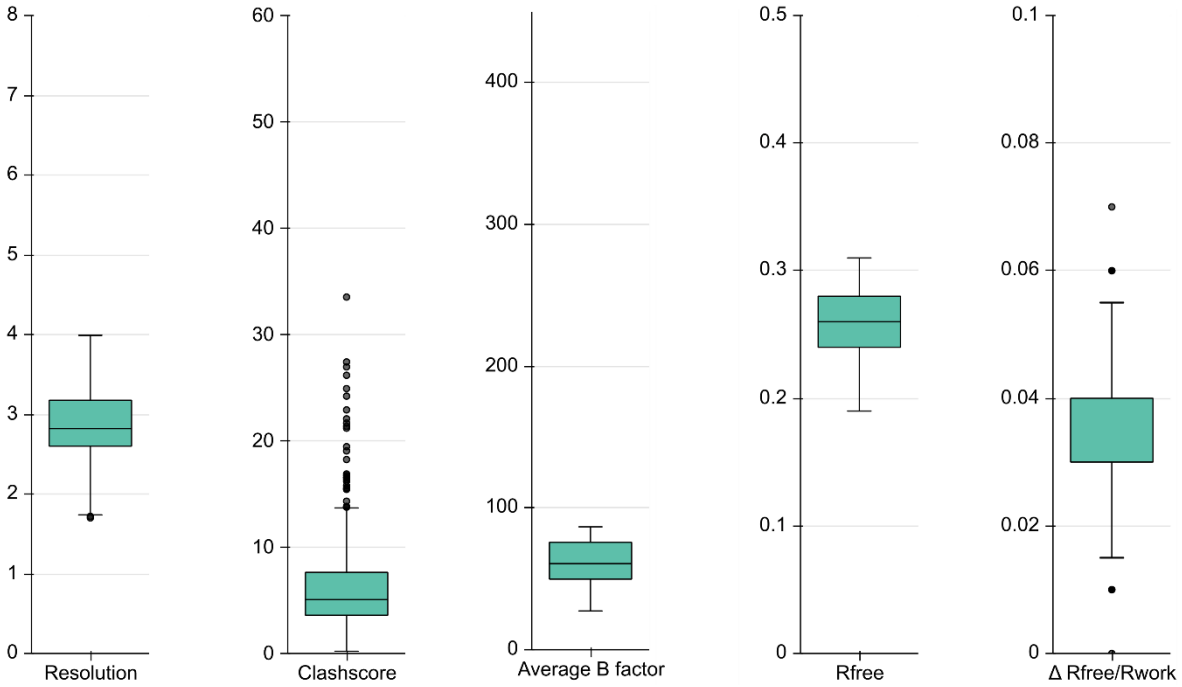
Appendix

10.3. List of Tables

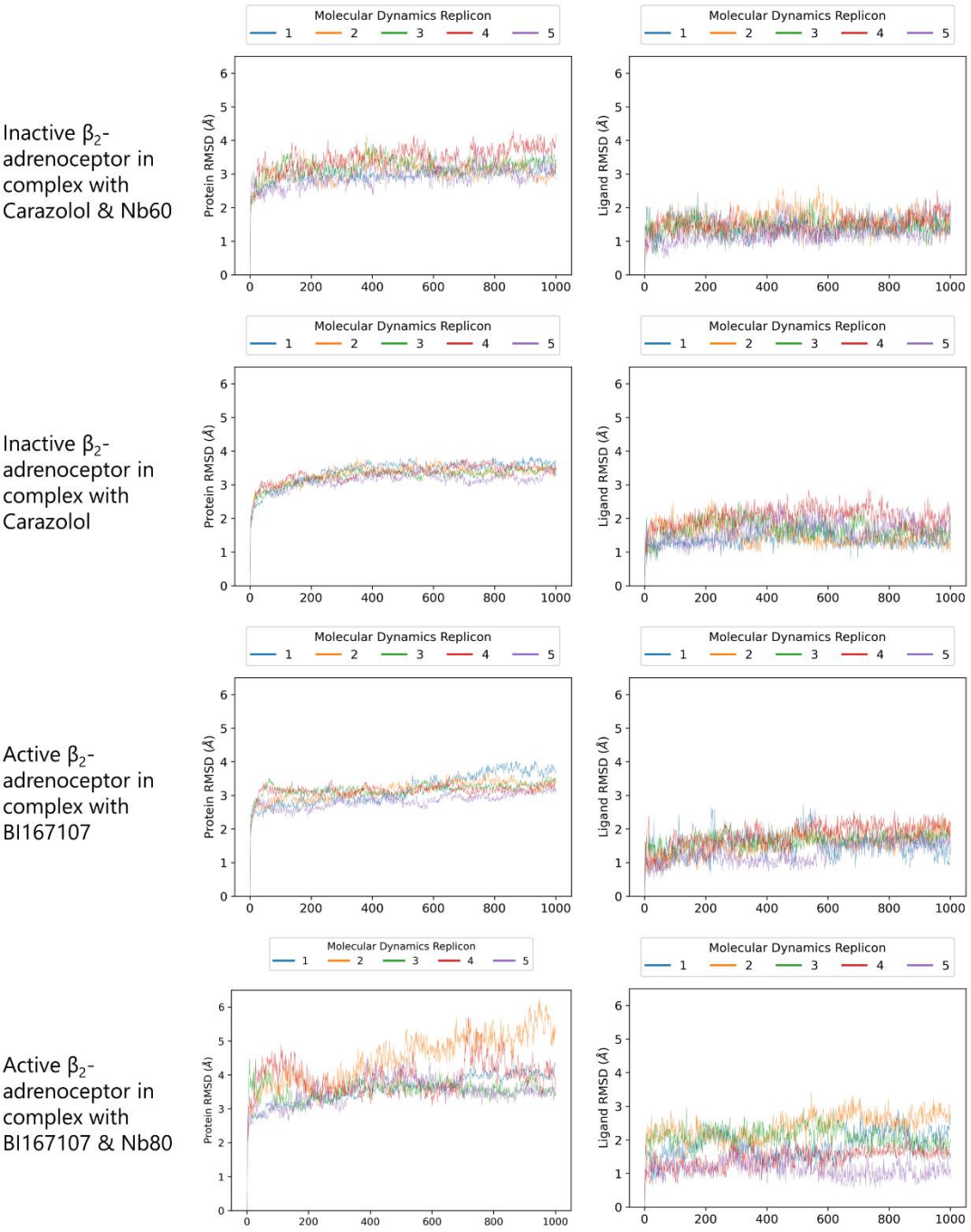
10.4. Appendix Figures



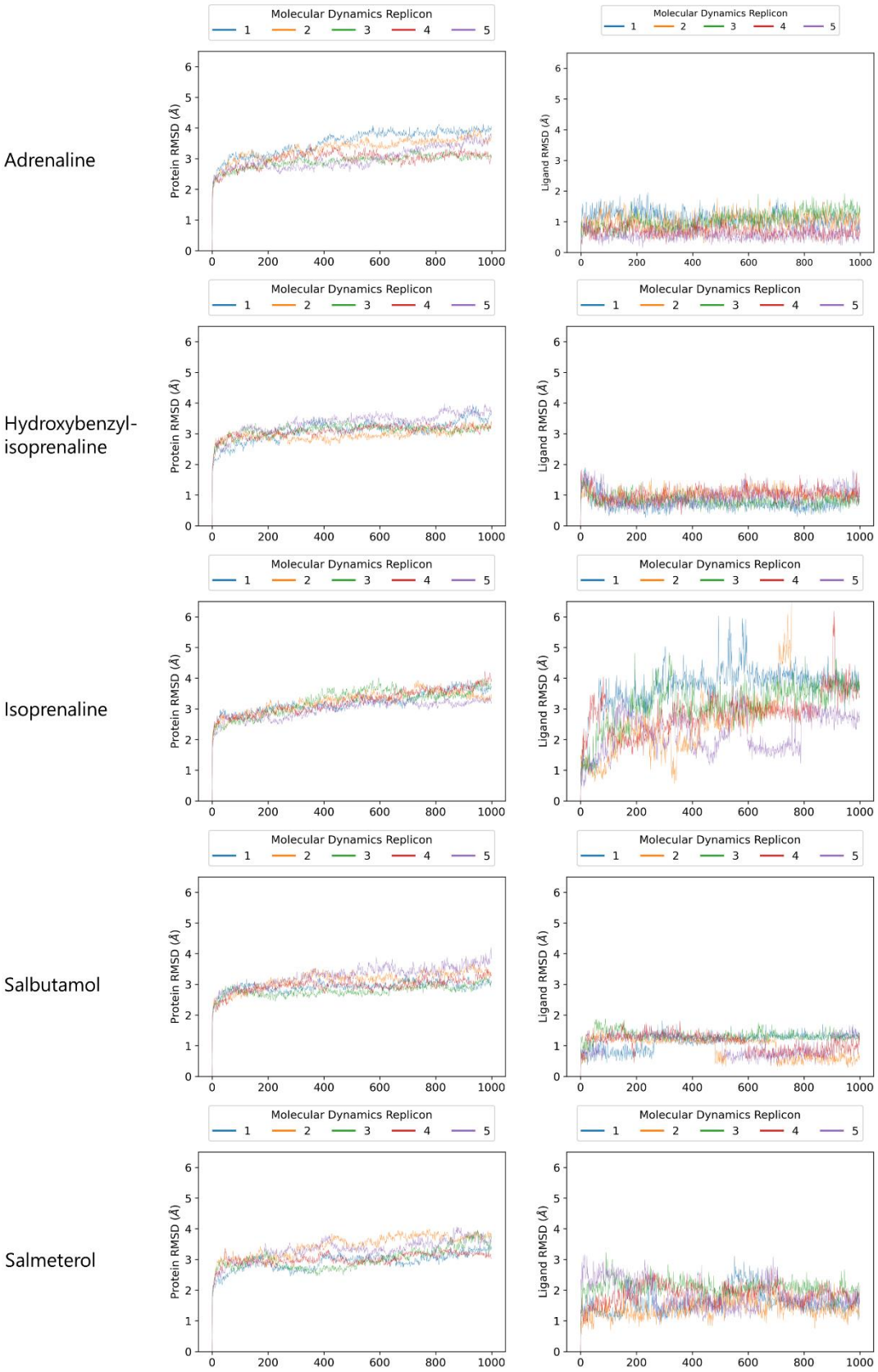
Appendix Figure 1. **Distribution of structural GPCR data between receptor families after removing outliers in the 5% probability density boundary.** 324 structural data points (as of 29.09.2021).



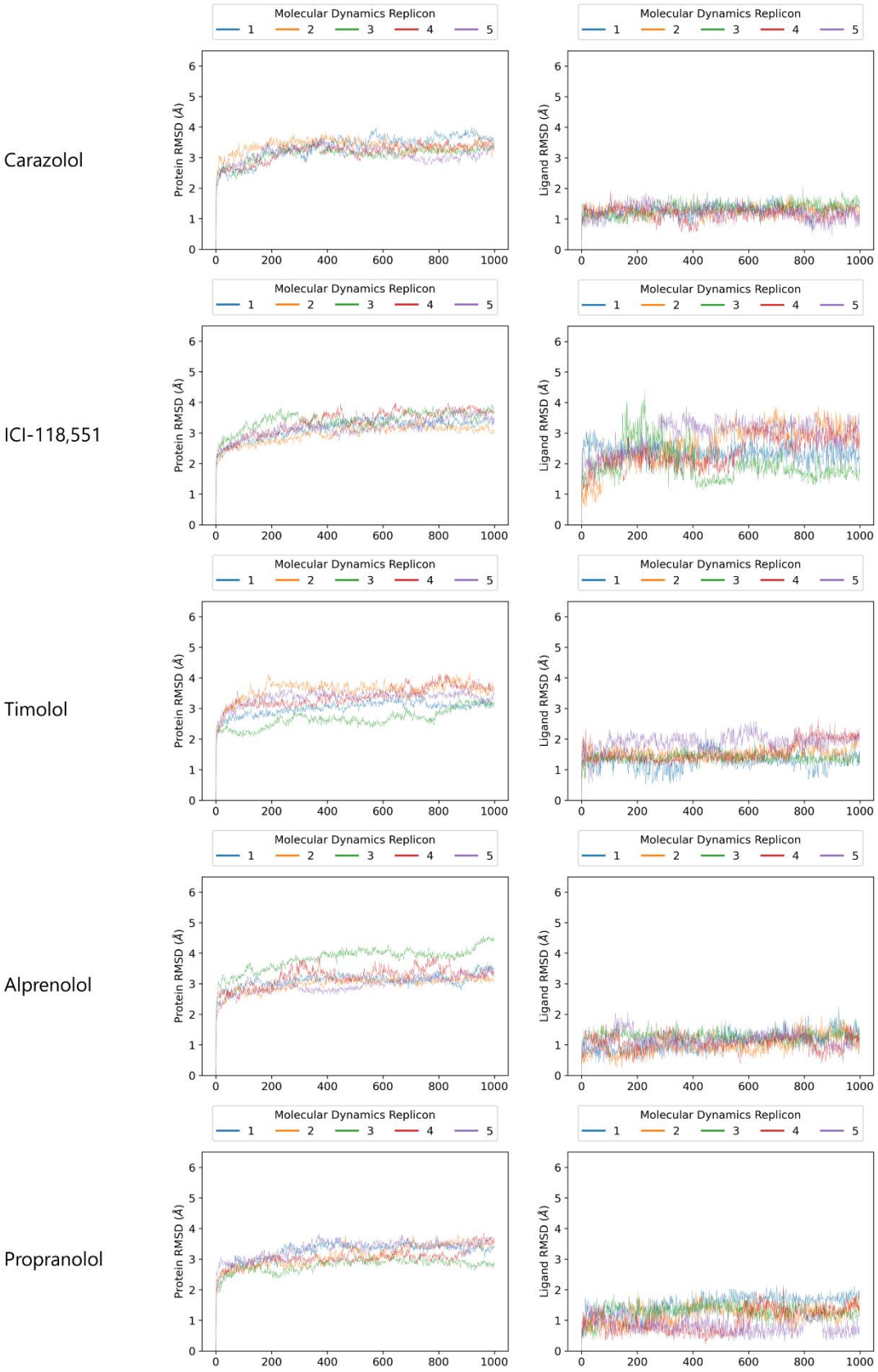
Appendix Figure 2. **Boxplots of different measures of structural data quality after removing outliers in the 5% probability density boundary.** 324 structural data points (as of 29.09.2021).



Appendix Figure 3. **Protein and ligand RMSD plots from MD simulations of active (PDB 3P0G) and inactive (PDB 5JQH) β_2-adrenoceptor systems.**



Appendix Figure 4. **Protein and ligand RMSD plots from MD simulations of active (PDB 3P0G) β_2-adrenoceptor systems in complex with full and partial agonist.**



Appendix Figure 5. **Protein and ligand RMSD plots from MD simulations of active (PDB 3P0G) β_2-adrenoceptor systems in complex with inverse agonists and antagonists.**

10.5. Appendix Tables

Appendix Table 1. **Detailed occurrences of superfeatures for each MD simulation.** Grey background – discarded MD simulations

	AR[6,7,9,11,12]	AR[12,13,14,15,16,17]	HBA[1]	HBA[2]	HBA[3]	HBA[4]	HBA[5]	HBA[7]	HBD[5]	HBD[6]	HBD[8]	H[12,13,14,15,16,17]	H[18,19]	NI[0,1,2,3,4]	PI[8]
s1pr1_0	0	0	33	94	98	21	16	0	92	4	100	100	99	100	100
s1pr1_1	0	0	91	47	97	70	59	26	45	92	96	100	100	100	100
s1pr1_2	0	0	71	37	88	83	34	0	81	9	93	100	100	100	94
s1pr1_3	0	0	58	77	39	65	19	0	55	97	98	100	100	100	100
s1pr1_4	0	0	100	35	100	25	41	0	45	93	87	100	100	100	100
s1pr1_5	0	0	85	14	98	95	8	0	57	100	100	100	100	100	100
s1pr1_6	0	0	72	22	55	90	30	0	25	94	100	100	100	100	100
s1pr1_7	0	0	83	71	95	56	66	9	18	92	100	100	100	100	100
s1pr1_8	0	0	98	27	99	3	6	4	33	64	99	100	100	100	100
s1pr1_9	0	0	94	71	99	84	50	7	30	88	94	100	100	100	100
s1pr2_0	0	0	100	39	100	98	40	4	29	100	100	100	100	100	100
s1pr2_1	2	0	98	50	100	99	18	0	38	78	100	100	100	100	100
s1pr2_2	0	2	96	98	100	100	0	6	36	47	94	100	100	100	100
s1pr2_3	0	0	94	8	18	97	50	0	35	100	100	100	100	100	100
s1pr2_4	4	20	98	23	100	96	0	0	33	24	99	100	100	100	100
s1pr2_5	10	0	85	1	94	73	0	0	59	9	100	100	100	100	100
s1pr2_6	1	3	77	38	95	100	4	0	84	28	100	100	100	100	100
s1pr2_7	23	16	92	5	100	100	69	0	93	34	100	100	100	100	100
s1pr2_8	5	5	94	30	100	93	0	0	67	82	100	100	100	100	100
s1pr2_9	5	19	75	4	70	60	14	0	53	54	99	100	100	100	100
s1pr3_0	0	0	98	41	73	87	4	0	66	100	99	100	100	100	100
s1pr3_1	0	0	79	37	33	99	25	0	30	97	85	100	100	100	100
s1pr3_2	0	0	97	24	100	100	11	81	13	99	100	100	100	100	100
s1pr3_3	0	0	50	21	66	77	61	0	33	99	92	100	100	100	100
s1pr3_4	0	0	99	1	99	2	38	0	70	92	1	100	100	100	52
s1pr3_5	0	0	94	28	92	100	33	5	36	100	100	100	100	100	100
s1pr3_6	0	0	85	77	95	100	1	0	28	100	99	100	100	100	100
s1pr3_7	0	0	75	3	42	69	5	1	23	98	98	100	100	100	100
s1pr3_8	0	0	95	67	91	100	73	0	35	99	100	100	100	100	100
s1pr3_9	0	0	56	43	77	46	2	5	32	99	98	100	100	100	100
s1pr4_0	0	0	100	60	89	99	33	3	65	99	100	100	100	100	100
s1pr4_1	0	0	100	99	96	70	60	0	51	99	100	100	100	100	100
s1pr4_2	0	0	100	100	15	86	68	0	59	100	100	100	100	100	100
s1pr4_3	0	0	100	89	100	100	45	0	51	100	100	100	100	100	100
s1pr4_4	0	0	100	99	0	98	99	0	21	99	100	100	100	100	100
s1pr4_5	0	0	100	99	97	83	43	0	63	99	100	100	100	100	100
s1pr4_6	0	0	100	91	11	54	14	1	53	98	100	100	100	100	100
s1pr4_7	0	0	97	50	52	94	27	0	30	100	100	100	100	100	100
s1pr4_8	0	0	100	96	1	87	84	0	57	98	100	100	100	100	100
s1pr4_9	0	0	100	99	1	89	96	0	65	100	100	100	100	100	100
s1pr5_0	0	0	100	8	98	83	50	0	31	99	100	100	100	100	100
s1pr5_1	0	0	90	23	55	98	2	2	36	20	84	100	100	100	100
s1pr5_2	0	0	99	40	87	95	49	4	27	99	100	100	98	100	100
s1pr5_3	0	0	13	11	95	98	13	0	79	2	100	100	100	100	100
s1pr5_4	0	0	100	33	73	99	5	1	32	84	94	100	100	100	100
s1pr5_5	0	0	50	9	99	96	7	0	59	99	100	100	100	100	100
s1pr5_6	0	0	100	87	98	99	15	1	93	29	99	100	100	100	100
s1pr5_7	0	0	100	74	100	99	1	3	74	96	86	100	100	100	100
s1pr5_8	0	0	100	8	4	100	1	0	97	98	100	100	100	100	100
s1pr5_9	0	0	92	4	28	100	13	0	43	100	100	100	100	100	100
All MDs	**AR[6,7,9,11,12]**	**AR[12,13,14,15,16,17]**	**HBA[1]**	**HBA[2]**	**HBA[3]**	**HBA[4]**	**HBA[5]**	**HBA[7]**	**HBD[5]**	**HBD[6]**	**HBD[8]**	**H[12,13,14,15,16,17]**	**H[18,19]**	**NI[0,1,2,3,4]**	**PI[8]**
s1pr1_mean	0,0	0,0	78,6	49,5	86,7	59,2	32,8	4,8	48,0	73,2	96,7	100,0	99,9	100,0	99,4
s1pr2_mean	5,0	6,6	90,9	29,7	87,6	91,5	19,7	1,0	52,7	55,6	99,1	100,0	100,0	100,0	100,0
s1pr3_mean	0,0	0,0	82,7	34,1	76,9	77,9	25,3	9,3	36,6	98,2	87,1	100,0	100,0	100,0	95,2
s1pr4_mean	0,0	0,0	99,6	88,1	46,1	86,1	56,8	0,3	51,5	99,1	99,9	100,0	100,0	100,0	100,0
s1pr5_mean	0,0	0,1	84,3	29,8	73,5	96,5	15,5	1,1	57,1	72,5	96,2	100,0	99,8	100,0	100,0
s1pr1_std	0,0	0,0	20,8	27,1	21,7	32,3	21,1	8,3	23,9	36,6	4,2	0,0	0,4	0,0	1,9
s1pr2_std	7,0	8,4	8,9	29,4	26,3	13,7	25,0	2,1	22,7	32,5	1,7	0,0	0,1	0,0	0,0
s1pr3_std	0,0	0,0	17,7	24,4	23,7	32,3	25,9	25,3	18,0	2,5	30,8	0,0	0,0	0,0	15,1
s1pr4_std	0,0	0,0	1,1	18,1	45,0	14,4	29,7	0,9	15,0	0,7	0,1	0,0	0,0	0,0	0,0
s1pr5_std	0,0	0,1	29,3	29,4	34,1	5,0	18,8	1,3	27,0	39,0	6,1	0,0	0,5	0,0	0,0
Selected MDs	**AR[6,7,9,11,12]**	**AR[12,13,14,15,16,17]**	**HBA[1]**	**HBA[2]**	**HBA[3]**	**HBA[4]**	**HBA[5]**	**HBA[7]**	**HBD[5]**	**HBD[6]**	**HBD[8]**	**H[12,13,14,15,16,17]**	**H[18,19]**	**NI[0,1,2,3,4]**	**PI[8]**
s1pr1_mean	0,0	0,0	83,7	44,5	85,4	63,5	34,7	5,3	43,1	80,9	96,3	100,0	100,0	100,0	99,3
s1pr2_mean	6,0	8,0	89,4	18,6	84,5	89,5	22,3	0,6	56,6	53,9	99,6	100,0	100,0	100,0	100,0
s1pr3_mean	0,0	0,0	79,8	38,6	76,6	86,9	19,5	13,0	34,0	99,0	96,1	100,0	100,0	100,0	100,0
s1pr4_mean	0,0	0,0	99,6	86,8	51,3	84,7	52,1	0,4	54,9	99,1	99,9	100,0	100,0	100,0	100,0
s1pr5_mean	0,0	0,1	84,3	29,8	73,5	96,5	15,5	1,1	57,1	72,5	96,2	100,0	99,8	100,0	100,0
s1pr1_std	0,0	0,0	13,2	22,1	21,3	29,3	20,2	8,1	18,2	27,2	4,0	0,0	0,0	0,0	1,9
s1pr2_std	7,1	8,3	8,8	14,8	27,0	13,8	25,5	1,4	22,4	33,3	0,3	0,0	0,1	0,0	0,0
s1pr3_std	0,0	0,0	18,0	17,4	21,2	18,7	20,4	27,8	14,8	0,8	5,4	0,0	0,0	0,0	0,0
s1pr4_std	0,0	0,0	1,1	17,7	42,0	13,7	25,7	0,8	10,3	0,6	0,1	0,0	0,0	0,0	0,0
s1pr5_std	0,0	0,1	27,8	27,8	32,4	4,8	17,8	1,2	25,6	37,0	5,8	0,0	0,5	0,0	0,0

Appendix Table 2. **Hydrogen bond acceptor interaction with phosphate moiety of compound 6 per frame of each MD simulation in %.**

	$S1PR_1$	*$S1PR_2$*	*$S1PR_3$*	*$S1PR_4$*	*$S1PR_5$*
0 HBA	0.04	0.13	0.11	0.01	0.01
1 HBA	0.95	1.92	4.06	0.37	1.44
2 HBA	13.59	14.75	13.14	1.05	20.78
3 HBA	24.02	20.84	22.04	10.69	18.32
4 HBA	33.40	22.21	21.40	26.93	25.14
5+ HBA	28.00	40.14	39.26	60.96	34.31

Appendix Table 3. **Transition probability between different binding modes of $S1PR_1$.**

	BM1	*BM2*	*BM ECL*	*T-BM*
BM1	85.48%	0.03%	0.00%	14.49%
BM2	0.02%	90.76%	0.00%	9.22%
BM ECL	0.00%	0.00%	72.41%	27.59%
T-BM	7.70%	8.05%	0.12%	84.13%

Appendix Table 4. **Transition probability between different binding modes of $S1PR_2$.**

	BM1	*BM2*	*BM ECL*	*T-BM*
BM1	80.85%	3.81%	0.27%	14.98%
BM2	9.12%	62.13%	0.00%	28.75%
BM ECL	5.23%	0.00%	41.18%	53.59%
T-BM	22.18%	16.62%	3.50%	57.70%

Appendix Table 5. **Transition probability between different binding modes of S1PR$_3$.**

	BM1	*BM2*	*BM ECL*	*T-BM*
BM1	95.66%	0.14%	0.00%	4.20%
BM2	0.47%	91.35%	0.00%	8.18%
BM ECL	0.00%	0.00%	0.00%	0.00%
T-BM	16.97%	7.36%	0.00%	75.67%

Appendix Table 6. **Transition probability between different binding modes of S1PR$_4$.**

	BM1	*BM2*	*BM ECL*	*T-BM*
BM1	87.28%	0.31%	0.00%	12.41%
BM2	0.62%	81.38%	0.00%	18.00%
BM ECL	0.00%	0.00%	0.00%	0.00%
T-BM	14.35%	11.74%	0.00%	73.91%

Appendix Table 7. **Transition probability between different binding modes of S1PR$_5$.**

	BM1	*BM2*	*BM ECL*	*T-BM*
BM1	79.75%	0.14%	0.01%	20.10%
BM2	1.42%	64.86%	0.00%	33.72%
BM ECL	0.07%	0.00%	97.73%	2.20%
T-BM	19.91%	3.70%	0.90%	75.49%

Appendix Table 8. **BM1 lipophilic tail interaction occurrence in percent for $S1PR_{1-5}$.**

	$S1PR_1$	$S1PR_2$	$S1PR_3$	$S1PR_4$	$S1PR_5$
3.33	9,37	41,42	21,51	30,67	29,44
3.36	5,11	14,79	0,21	21,59	23,51
4.56	0,00	87,11	0,29	0,00	4,13
5.39	0,00	0,70	0,00	0,10	6,07
5.40	3,37	0,09	21,18	26,68	0,21
5.43	0,00	98,92	0,00	0,00	0,00
5.44	100,00	2,31	100,00	100,00	100,00
5.46	0,00	49,98	0,01	0,00	0,00
5.47	44,54	48,94	40,49	88,85	82,78
5.48	44,20	0,00	28,58	0,03	1,74
5.50	0,00	17,26	0,00	0,00	0,00
6.48	43,56	1,88	0,12	13,65	6,74
6.51	28,29	1,76	6,85	1,97	4,58
6.52	81,98	3,83	66,18	63,34	47,99
6.55	66,24	2,37	66,46	66,66	59,07
6.56	1,91	0,00	12,50	0,00	0,00

Appendix Table 9. **BM2 lipophilic tail interaction occurrence in percent for $S1PR_{1-5}$.**

	$S1PR_1$	*$S1PR_2$*	*$S1PR_3$*	*$S1PR_4$*	*$S1PR_5$*
3.33	0,24	6,36	0,26	0,37	0,51
3.36	42,68	34,38	11,60	41,41	4,50
3.37	0,00	0,00	0,00	36,59	91,52
3.40	100,00	100,00	100,00	100,00	100,00
3.41	0,00	12,02	0,00	5,27	0,00
4.56	0,32	29,74	19,20	7,43	32,26
5.46	0,00	37,54	16,96	33,01	52,44
5.47	91,17	52,46	80,54	82,54	59,90
5.50	0,00	12,13	0,00	0,00	51,54
5.51	24,43	35,37	58,79	0,00	0,00
6.44	0,00	23,49	0,00	0,00	0,00
6.45	0,00	22,50	0,00	0,00	0,00
6.48	71,86	11,43	36,06	18,28	32,01
6.51	11,59	0,18	0,68	16,57	0,00
6.52	84,20	15,04	57,13	27,40	19,41
6.55	0,00	0,00	0,73	8,77	0,00

Appendix Table 10. **BM ECL lipophilic tail interaction occurrence in percent for $S1PR_{2,5}$.**

	$S1PR_2$	*$S1PR_5$*
3.32	8,17	0,00
3.33	59,80	69,79
182/185	0,00	8,26
183/186	95,75	23,03
186/189	100,00	100,00
5.39	38,24	92,79
5.40	0,00	56,70
5.59	0,00	0,00
6.55	16,34	24,45

Appendix Table 11. **Boxplot data for Figure 58.**

		All data	*5% PDR Boundary*
Resolution	Outliers	29	2
	Upper Whisker	4.03	3.99
	Lower Whisker	1.82	1.74
	Q1	2.65	2.60
	Q2	2.90	2.82
	Q3	3.20	3.18
	IQR	0.55	0.57
	Structures	487	324
Clashscore	Outliers	51	28
	Upper Whisker	15.32	13.69
	Lower Whisker	0	0.20
	Q1	3.09	3.59
	Q2	5.04	5.07
	Q3	7.99	7.63
	IQR	4.89	4.04
	Structures	487	324
Average B Factor	Outliers	10	0
	Upper Whisker	153.25	86.73
	Lower Whisker	27.40	27.40
	Q1	57.14	49.80
	Q2	76.37	60.77
	Q3	95.59	75.80
	IQR	38.44	26.00
	Structures	256	131
Rfree	Outliers	8	0
	Upper Whisker	0.34	0.31
	Lower Whisker	0.18	0.19
	Q1	0.24	0.24
	Q2	0.26	0.26
	Q3	0.28	0.28
	IQR	0.04	0.04
	Structures	384	242
Δ Rfree/Rwork	Outliers	50	32
	Upper Whisker	0.06	0.06
	Lower Whisker	0.01	0.01
	Q1	0.03	0.03
	Q2	0.03	0.03
	Q3	0.04	0.04
	IQR	0.01	0.01
	Structures	384	242