UNIVERSITÀ DEGLI STUDI DI URBINO CARLO BO

Department of Pure and Applied Sciences

Ph.D. PROGRAMME IN: Research Methods in Science and Technology

XXXVIII CYCLE

Development of an automatic tool to study the protein-membrane interface

ACADEMIC DISCIPLINE: CHEM-07/A

Coordinator: Prof. Luca Lanci

Supervisor: Prof. Giovanni Bottegoni

Ph.D. student: David Sotillo Núñez

ACADEMIC YEAR
2024/2025

The work presented in this manuscript was carried out within the framework of the ALLODD consortium (see Appendix I) and supported by the funding provided by the EU Horizon 2020 MSCA Program under grant agreement 956314.

This thesis is dedicated to my family and friends, especially my grandpa, mi avi, the best and kindest person I have ever met. I aim to be as loved and kind as you were and still are.

Abstract

Biological membranes are not passive scaffolds for membrane proteins but active regulators of their structure, dynamics, and function. In mammalian membranes, cholesterol plays a central role in modulating protein behaviour through both non-specific mechanical effects on bilayer properties and specific interactions at the protein–membrane interface. Despite extensive experimental and computational efforts, the systematic identification of cholesterol binding sites remains challenging, as many experimental techniques provide static or ensemble-averaged views that obscure the dynamics and mechanistic details of lipid–protein interactions.

Coarse-grained molecular dynamics (CG-MD) simulations offer a powerful means to overcome these limitations by enabling extended simulations of large protein–membrane systems over hundreds of microseconds, thereby capturing persistent and functionally relevant cholesterol interactions. However, the lack of standardized and automated workflows has hindered large-scale and comparative studies across membrane protein families.

Here, we present CHAMP, an automated and scalable CG-MD–based protocol for the construction, simulation, and analysis of protein–membrane systems, designed to identify cholesterol interaction hot spots at the protein–membrane interface. The method integrates long-lived protein–cholesterol contact analysis with spatial density mapping to robustly distinguish persistent binding regions from transient encounters. We validate the protocol using two well-characterized systems, namely the cannabinoid 1 receptor and the serotonin transporter, and demonstrate its ability to recover experimentally observed cholesterol binding sites as well as previously unreported interaction regions [1].

We further apply CHAMP to a representative benchmark of 56 class A G protein-coupled receptor structures, enabling a systematic, family-wide characterization of cholesterol interaction landscapes. This analysis reveals recurrent membrane-interface hot spots that are not reliably captured by sequence-based motifs, underscoring the importance of structure- and dynamics-driven approaches. Both the CHAMP software and the complete set of results are made publicly available as a Python package and an interactive online platform, providing a general framework for large-scale studies of lipid–protein interactions and supporting the discovery of membrane-facing allosteric sites relevant to drug design.

Table of Contents

Chapter 1. Introduction

1.1. Biological membranes

The lipid bilayers act as far more complex elements than merely as barriers. They are heterogeneous ecosystems where lipids and proteins interact and are involved in a diversity of vital biological processes, from protein stability and conformational equilibria to cellular trafficking. The concept of biological membranes as a simple, passive solvent has evolved to give place to a more fitting description, where they act as active participants in shaping the structure of the proteins embedded in them, as well as their function [2,3] .

The interaction of membrane proteins with their surrounding lipids can be broadly classified into two categories: i) non-specific, where membrane components modulate protein behaviour through the bulk physical properties of the bilayer, or ii) specific, involving discrete binding sites at which lipids act in a manner analogous to allosteric ligands [4,5]. Identifying such lipid-binding hot spots is therefore a powerful strategy for uncovering regulatory sites at the membrane–protein interface, with direct relevance for drug discovery, as these sites represent potential targets for the design of novel allosteric modulators. However, the dynamic and heterogeneous nature of biological membranes makes the systematic identification and characterization of these interactions challenging, particularly when extended across multiple targets and membrane environments.

1.2. The role of cholesterol as an allosteric modulator

Among the many diverse membrane lipid, cholesterol stands out as a key component of mammalian membranes, accounting at times for up to 40% of the total lipids [6]. The rigid structure of cholesterol allows it to insert itself between the phospholipid tails, thus influencing the properties of the membrane, from the bilayer compressibility to the fluidity of the membrane, sometimes forming liquid-ordered (Lo) domains known as “rafts” [3,7]. There are plenty of reports in the literature that demonstrates the existence of specific cholesterol binding to membrane proteins and affecting their function at some level. Representative examples include the tyrosine kinase receptor TrkB, where cholesterol binding to a CARC motif in the transmembrane region

facilitates brain-derived neurotrophic factor signalling [8], and the T-cell receptor–CD3 complex, in which cholesterol association at the transmembrane interface inhibits complex activity [9].

1.3. G Protein-coupled receptors and cholesterol

G protein-coupled receptors (GPCRs) are one of the most structurally and functionally conserved protein families in the human genome. Their involvement in many essential signalling processes makes them a very suitable target for drug design, to the point where around 36% of the drugs currently approved by the FDA are aimed at these proteins [10]. Beyond the well-known seven-transmembrane (7TM) fold, GPCRs share key features such as common activation mechanisms, conserved microswitches, and similar arrangements of lipid-exposed residues. GPCRs populate multiple conformational ensembles that underpin their pleiotropic signalling behaviour [11,12]. This intrinsic conformational plasticity makes allosteric modulation a particularly attractive therapeutic strategy for fine-tuning receptor activation and downstream signalling [13]. Cholesterol has been shown to directly influence ligand-binding properties and functional activity of GPCRs [14–16], acting as either a positive or negative allosteric modulator depending on the receptor and context, and in some cases modulating receptor oligomerization [17]. However, since the effect of cholesterol are receptor-dependent, it is implied that insights derived from individual receptors may not always generalize across a protein family. For GPCRs in particular, comparative analyses across multiple family members are especially valuable, as they can leverage the conserved structural fold and shared activation mechanisms to identify common patterns of lipid interaction that may not be apparent when studying receptors in isolation.

There have been attempts to predict cholesterol binding regions relying purely on aminoacidic sequence patterns, or motifs, such as the CRAC and CARC motifs [18]. While these domains matched actual cholesterol binding sites in some receptors, they fail to predict these binding sites in many others [18–20], suggesting that the nature of these interactions might not be as obvious and have deeper underlying subtleties.

1.4. Studying cholesterol binding

Due to the “diffusive” nature of cholesterol in biological membranes, researchers usually resort to mixed experimental and computational approaches to assess the presence, binding location or effect it has in the function of membrane proteins.

High resolution X-ray crystallography allows to capture snapshots of proteins and, in some cases even retain cholesterol molecules bound to a specific pocket, helping to demonstrate how cholesterol can contribute to the stability of a protein [21,22]. The introduction of cryo-electron microscopy on the past decade has permitted to observe these interactions in environments that better mimic a realistic cellular membrane, such as nanodiscs [23]. While crystals provide us with a static view, solid-state NMR (ssNMR) and Electron Paramagnetic Resonance (EPR) allow us to study cholesterol in a fluid membrane at physiological temperatures. These techniques can be useful for assessing how cholesterol can affect and modify the properties of the bilayer, such as their thickness and "stiffness" around a protein [24]. The use of “molecular sponges”, which are compounds capable of stripping cholesterol from membranes, allow to study the effect of cholesterol depletion on the protein-membrane systems [25].

Despite the power of these experimental approaches, even when supported by state-of-the-art instrumentation and recent methodological advances, the resulting data often lack key mechanistic details. High-resolution crystal structures provide precise atomic coordinates but represent static snapshots, offering no direct insight into the dynamics of cholesterol binding. Conversely, experimental techniques that probe cholesterol effects over time typically report ensemble-averaged signals across thousands of protein copies, thereby obscuring the molecular determinants of individual binding events.

Molecular dynamics (MD) simulations provide a powerful means of bridging this gap by computationally replicating the physicochemical environment of membrane proteins and offering an atomistic description of binding processes over time [26,27]. In MD simulations, atoms are treated as classical particles whose trajectories are propagated by numerically integrating Newton’s equations of motion. These equations are commonly integrated using finite-differences, such as in the Verlet algorithm, with timesteps that typically range from the magnitude orders of picoseconds to femtoseconds [28]. The forces acting on each atom are defined by an empirical force field, which encodes the potential energy of the system through parameterized terms describing bonded interactions, electrostatic forces, and van der Waals interactions. Some of the most popular force fields during the past decades are AMBER [29,30], CHARMM [31,32] or OPLS [33]. These potential energy functions are usually decomposed into the bonded (bond vibration, angle bending

and dihedral rotations) and the non-bonded (electrostatic and van der Waals interactions) terms, where the latter are commonly described through Coulombic electrostatics and Lennard-Jones potentials [26]. This framework enables the explicit simulation of protein and lipid dynamics across timescales ranging from picoseconds to milliseconds. In order to approximate a bulk behaviour and to avoid finite-sized effects, MD simulations are typically performed under periodic boundary conditions, where the simulation box is infinitely replicated across all spatial directions. The long-range electrostatic interactions are commonly addressed using specific algorithms, such as the Particle Mesh Ewald, which allows for accurate computation of Coulombic interactions when dealing with periodic systems [34]. Access to these short and intermediate timescales allows MD simulations to resolve fast molecular processes that are typically inaccessible to experimental techniques, including conformational fluctuations and transient interaction events [35,36]. Furhermore, all data generated via MD simulations can be analysed in the framework of statistical mechanics, which allows the calculation of thermodynamic and kinetic properties as ensemble averages. Of particular relevance is the ability of MD simulations to directly observe ligand binding and unbinding events at atomic resolution.

Indeed, MD simulations have been used to study cholesterol binding in GPCRs. In 2013, Cang and colleagues performed unbiased MD simulations in the range of microseconds with the aim of uncovering the cholesterol binding mechanism at the β2-adrenergic receptor. They assembled multiple receptor-membrane systems consisting of a complete set of monomers and dimers embedded in pure POPC and POPC:cholesterol(33%) membranes. They also performed three additional simulations where two cholesterol molecules started in the groove corresponding to the Cholesterol Consensus Motif (CCM). Using the Spatial Density Function of the POPC lipids and cholesterol molecules around the receptor and comparing the results across the systems they reported several potential binding sites. They linked cholesterol binding in two of those sites to conformational stabilization, which favours ligand binding (TM1-2-7), and to receptor dimerization (TM1-H8) [37]. In that same year, Lee and colleagues focused on the CCM reported for the A_{2A} adenosine receptor, which matched a cholesterol binding site in the β2-adrenergic receptor (TM1-3-4). They used four crystal structures of the A_{2A} adenosine receptor, which were solved in complex to different ligands in the presence of stabilizing mutations. They built a total of 10 systems, including wild type and mutated structures, where the receptors were embedded in a pure POPC membrane and two cholesterol molecules placed in the β2-AR CCM region. The total simulated time was approximately 11 μs. They followed the protein-cholesterol contacts during the simulations and analysed the results, addressing also the effect of the mutations and of the different ligands bound to the receptors. They found that the simulations with the receptor bound to an agonist (UK432097) behave differently than the rest. In most of their simulations one

of the two cholesterols (labelled as Q) formed persistent interactions with the SER47 and TRP129 (which share an H-bond) located in the CCM, while the other cholesterol (R) unbound relatively quickly. In the agonist-bound receptor, however, the interactions with cholesterol Q were weakened due to the SER-TRP H-bond breaking, interfering with the geometry of the pocket and causing the protein to interact strongly with cholesterol R, located farther from the CCM [38]. Knowing the reported stabilizing effect of cholesterol in the A_{2A} adenosine receptor, in 2017, Guixà-González and colleagues performed a series of classical MD simulations to assess the effect of cholesterol depletion in this receptor. They were interested in the possibility of cholesterol spontaneously entering the internal cavity of the receptor via the transmembrane helices. For this reason, they ran several 1 μs replicas using a multicomponent membrane composed of cholesterol, saturated phospholipids, monosaturated and polyunsaturated phospholipids and sphingomyelin. From these trajectories they selected four snapshots, which were used, in turn, to start new short trajectories. To better observe the effect of the membrane on the cholesterol entrance, the four snapshots (with the relevant cholesterol) were embedded in pure POPC membranes and simulated for 100 ns. From there, they selected the three replicas which presented the best progression in accessing the interior of the receptor. The volumetric analysis of their simulations revealed a potential double mechanism through which cholesterol is able to modulate ligand binding. Besides the already known allosteric modulation at the protein-membrane interface, their computational studies uncovered an accessing gateway located between transmembrane helices TM5-6, which allowed cholesterol to enter the orthosteric binding site of the receptor [39]. Recently, in 2025, Vanegas and colleagues carried out unbiased MD simulations on the Cannabinoid receptor 1 (CB_1) aiming to identify cholesterol binding sites. They prepared 6 systems starting from the same crystal structure, a set of instances that combined the presence of the fusion protein or of membrane cholesterol and the Intracellular Loop 3 (ICL3) being modelled. They simulated each system in two 500 ns replicas and analysed the protein-cholesterol interactions on the last 400 ns using a dedicated Python tool to identify binding sites. They then performed QM/MM analyses on the detected sites to determine the strength of the interaction with cholesterol molecules. They reported preferential cholesterol binding regions on the CB_1 receptor, divided in 5 distinctive binding sites. Notably, the most persistent binding site (BS1, TM2-3-4) aligns with the crystallographic cholesterol of the structure used in their analysis [40]. Studies on cholesterol interaction with membrane proteins have not only been extensively studied in G protein-coupled receptors, as exemplified above, but there have been efforts towards studying these events in other families. In 2014, Weiser used the results of their docking experiments in cholesterol binding on the Mitochondrial Voltage-Dependent Anion Channel 1 (VDAC) membrane protein. They built two types of systems, both with the same DOPC membrane. One system type included 5 cholesterol molecules explicitly bound and 11 additional

randomly placed, while the other system had all 16 cholesterol molecules located randomly in the membrane. They performed 22 individual simulations, adding for a total aggregated time of 1.4 μs, which were used to investigate several properties. The authors measured the protein flexibility by calculating RMSF, the pore size and the hydrogen bonds to understand the effects of cholesterol in the structures. Furthermore, they quantified the ion permeation and the pore geometry and built electrostatic potential maps and ion density distributions to determine if cholesterol affects the ion distribution in the channel. Finally, they computed the PMF profile along the pore to compare the cholesterol-bound channel with the apo one. Their studies showed that the cholesterol binding in the docked positions was likely not random, as cholesterol molecules remained persistently bound during the simulations. Moreover, the cholesterol binding in the relevant sites was shown to affect the dynamics of VDAC, even if not all the sites were occupied, partial occupation on these regions already had a measurable effect [41].

In addition to MD simulations where native membrane components are involved, other strategies have been developed in order to identify binding sites by using molecular dynamics simulations with mixed solvents (MDmix). These methods involve introducing small drug-like organic compounds into the simulation environment to act as probes. These molecules freely diffuse during the simulation and compete for interactions on the protein surface. Their occupancy is then measured and quantified via density maps, which are used to identify potential binding sites [42].

While at the time of writing this thesis there are no reports of the use of MDmix simulations specifically targeted towards identifying cholesterol binding sites, it has been reported as a successful technique to identify allosteric binding sites. A relevant example comes from Chan and colleagues in 2023, where they described and applied an MDmix procedure aimed to study membrane proteins. Their protocol involved building protein-membrane systems where a small concentration of the organic probe was introduced to surround the protein but ultimately only allowed to exist in the aqueous phase and consequently removed from the lipidic phase to avoid membrane disruptions during the first stages of the simulations. They built several systems consisting of GPCRs which presented allosteric modulators with known binding sites embedded in POPC:cholesterol membranes. Their retrospective analysis allowed them to identify the previously reported sites in the cannabinoid receptor type 1, C-C chemokine receptor type 2, M2 muscarinic receptor, P2Y purinoceptor 1, and protease-activated receptor 2. Their novel application to the μ-opioid receptor, for which the binding sites of its known allosteric modulators remained undiscovered, revealed several potential binding sites [43].

1.5. Coarse-grained Molecular Dynamics Simulations for the study of Cholesterol

A common limitation of these atomistic simulations is that the accessible sampling timescales are typically limited to the microsecond regime, limiting their ability to capture slower lipid reorganization and binding processes. A widely adopted solution to overcome these sampling limitations is the use of force fields based on simplified molecular representations, commonly referred to as coarse-grained (CG) models. By grouping multiple atoms into effective interaction sites, CG simulations reduce the complexity of the system and smooth the underlying energy landscape, allowing the exploration of membrane organization and lipid–protein interactions over extended timescales. One of the most widely adopted among these force fields is the Martini CG, currently at its third major release (Martini 3 [44]). As exemplified in **Figure 1** with a phenylalanine residue, in this representation, heavy atoms are grouped into coarse-grained beads that act as effective substitutes for entire chemical moieties, rather than treating each atom as an individual particle. This mapping substantially reduces the total number of interaction sites in the system, thereby decreasing the number of force calculations required at each integration step. As a result, coarse-grained molecular dynamics simulations can achieve speedups of up to two to three orders of magnitude relative to their fully atomistic counterparts.

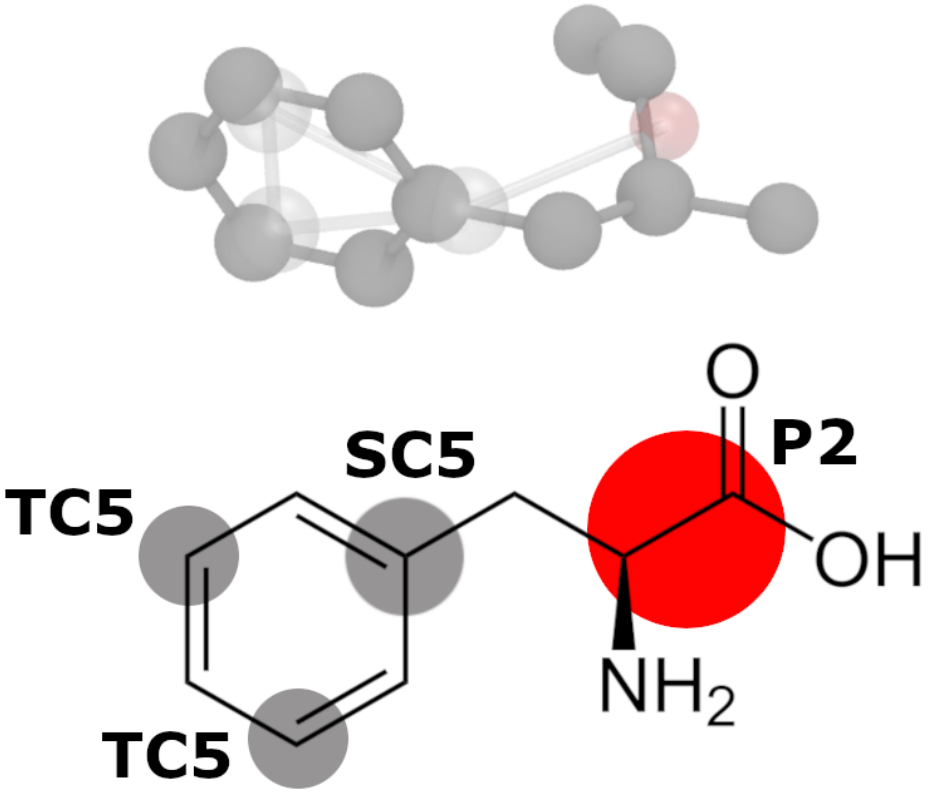


Figure 1. *Comparison between atomistic phenylalanine and its Martini 3 representation. (Top) Three-dimensional visualization of the molecule, where the position of each Martini 3 bead shows the approximation of the original geometry. (Bottom) Schematic drawing of the overlap between representations. The aromatic ring is represented by two tiny non-polar beads (TC5) and a small non-polar bead (SC5) coloured in grey, while the backbone of the residue is substituted by a regular polar bead (P2) coloured in red.*

As previously mentioned, the MARTINI 3 force field is built around a modular "building-block" philosophy, in which spherical beads with distinct interaction properties are used to represent groups of atoms. To balance the inherent loss of fine-grained structural detail, such as secondary structure information and specific intramolecular interactions, with the need to preserve realistic molecular geometries, bead assignment is performed carefully based on the physicochemical characteristics of the chemical groups being represented. **Table 1** provides a concise summary of the different bead classifications, defined according to bead size, chemical class, and polarity scale. Using phenylalanine as an illustrative example, several bead types are employed, including TC5 (tiny, apolar carbon bead with polarity level 5), SC5 (small, apolar carbon bead with polarity level 5), and P2 (regular-sized polar bead with polarity level 2).

Table 1. *Summary of the Martini 3 beads classification. According to their size, beads can be regular, small (S) or tiny (T). Depending of the group a bead is substituting it can be classified as Polar (P), Intermediate or semi-polar (N), apolar (C), halogen (X), +1 charged (Q), +2 charged or divalent (D) or water (W). Finally, within the same chemical*

class Martini 3 makes an additional distinction, assigning a polarity scale ranging 1 to 6, from the lowest polarity (highest hydrophobicity) to the highest polarity (lowest hydrophobicity), respectively.

Resolution (size)	**Chemical class**	**Polarity scale**
Regular	Polar (P)	1 (less polar)
	Intermediate (N)	2
Small (S)	Apolar (C)	3
	Halogen (X)	4
	Charged +1 (Q)	5
Tiny (T)	Divalent / +2 charge (D)	6 (most polar)

There are several studies using CG-MD simulations to study cholesterol that take advantage of this increased sampling, even identifying interaction sites in some cases. In 2017 Genheden and colleagues used the MARTINI force field to study how cholesterol interacts with the β_2-adrenergic receptor (β_2AR) and the A_{2A} adenosine receptor by embedding these receptors in POPC:cholesterol(30%) membranes. They, then, simulated these systems for an extended 50 μs production run. The authors mapped density distribution of cholesterol polar heads around the receptors and tracked the cholesterol occupancy by using a 5Å cutoff distance to determine whether a polar head bead was contacting the protein. Their analyses produced several clusters of residues with high cholesterol occupancy for both receptors, with the highest occupied cluster reaching an occupancy of more than 70% in both cases. The cluster with the highest occupancy in the β_2AR (72%) was located at the TM1-7 region, while in the adenosine A_{2A}R (74%) it was located at the TM4-5 area. Their CG-MD simulations suggested that regular atomistic MD usually cannot identify certain interaction areas because unable to reach full equilibration [45]. In the same year, Rouvier and colleagues used MARTINI MD simulations in combination with atomistic simulations. They used a high-resolution structure of the adenosine A_{2A}R to build a system where the protein was embedded in a POPC:cholesterol(30%) membrane. The system was simulated for an extended period of 200 μs to obtain an increased sampling of cholesterol distribution around the protein, allowing to predict preferred areas of interaction. While volumetric mapping of cholesterol density is effective for identifying preferentially occupied regions when stringent cutoffs are applied, it is not well suited for predicting true binding sites. To address this limitation, the authors introduced a scoring scheme that rewards persistent interaction events while discarding transient contacts shorter than 1 μs. Scores are accumulated for residues involved in

long-lived interactions that are repeatedly revisited over time. Mapping the highest-scoring residues onto the protein surface revealed several putative cholesterol-binding sites, the most prominent of which is located at the TM5–TM6 interface. This site coincides with the crystallographic cholesterol position observed in structures bound to the ZM241385 inverse agonist and was subsequently validated through atomistic simulations [46].

In 2019, Hedger and colleagues investigated cholesterol interactions with the class F GPCR Smoothened. To achieve extended sampling and enable the simulation of large systems, they employed coarse-grained molecular dynamics (CG-MD) simulations of Smoothened embedded in POPC membranes containing 20% cholesterol. Eight independent systems were constructed and simulated for 10 μs each. Analysis of these trajectories revealed specific cholesterol interaction regions located at the extracellular tips of the TM2–TM3 helices. To further assess the robustness of these interactions, the simulations were repeated using a five-component membrane composed of PC, PE, PS, PIP2, and cholesterol. The absence of significant differences between the two membrane compositions supported a genuine affinity of cholesterol for this region, which was subsequently confirmed by atomistic simulations [47]. In the same year, Song and colleagues examined lipid interactions of the $A_{2A}R$, including cholesterol. To assess how receptor activation influences lipid binding, they performed coarse-grained molecular dynamics simulations on three receptor conformations: inactive, active, and active bound to a mini-G_s protein. Each structure was embedded in a complex, asymmetric membrane containing ten lipid species, including cholesterol. For each activation state, ten independent systems were constructed with randomized initial lipid distributions, yielding a total simulation time of 8 μs. Lipid density maps and residue-level residence times were then computed, revealing seven lipid-binding sites with varying degrees of activation-state dependence. In particular, binding sites located at the TM1–TM2–TM4 and TM3–ECL2–TM4 interfaces exhibited increased persistence in both active conformations. This behavior was attributed to conformational shifts of TM4 upon activation and to tight packing of PIP2 around these regions, which restricts cholesterol dissociation. Notably, the authors also reported evidence of competition between cholesterol and PIP2 at the TM5–TM6 site [48].

Chapter 2. Aim of the Thesis

Despite the advances in the field, large-scale computational studies spanning many GPCRs remain relatively scarce. While insightful, existing efforts have often focused on limited subsets of receptors [49] or have required substantial manual effort in system preparation, simulation management, and data curation [50]. Moreover, many studies have not been explicitly designed to systematically interrogate the protein–membrane interface, leaving GPCR–cholesterol interactions underexplored from a comparative, family-wide perspective. These challenges are compounded by the fact that, despite extensive evidence for cholesterol binding, consensus sequence motifs such as the aforementioned CRAC and CARC do not consistently capture cholesterol interaction sites across GPCR, highlighting the need for structure- and dynamics-based approaches. To facilitate large-scale and systematic analyses of GPCRs, as well as other membrane protein families, existing workflows for identifying cholesterol interaction hot spots require adaptation into more general and automated frameworks.

A notable contribution in this direction was provided by Ferraro and colleagues, who introduced a CG MD–based protocol to identify cholesterol interaction sites at the protein–membrane interface of the serotonin transporter (SERT) [51]. Their work leveraged the extended sampling and scalability afforded by MARTINI simulations to map cholesterol-binding regions in a system known to be strongly modulated by membrane cholesterol. SERT operates within cholesterol-rich lipid rafts in neuronal membranes, and experimental depletion of cholesterol has been shown to significantly impair its transport activity, underscoring a functional link between the protein and its lipid environment.

At the time of their study, no high-resolution structure of SERT was available, and simulations were therefore performed using a homology model. The receptor was embedded as a homodimer in three raft-like membrane compositions composed of DSPC, POPE, and cholesterol at varying ratios (50:35:15, 50:30:20, and 50:25:25), designed to mimic the fluid lipid environment surrounding SERT. Two replicas were simulated for each membrane composition, yielding a total of six systems. Each system was simulated for 30 μs, resulting in an aggregated sampling time of 360 μs. As each simulation contained two SERT monomers, this setup produced twelve independent datasets of protein–cholesterol contacts, later expanded to fourteen datasets through additional simulations to assess sampling convergence. Protein–cholesterol contacts were defined using a 6 Å cutoff between protein beads and cholesterol particles, consistent with prior MARTINI-based studies and corresponding to the first minimum of the radial distribution function. For each residue, the maximum occupancy time (t_{max}), defined as the longest continuous

cholesterol contact observed during the simulation, was computed. This metric enabled discrimination between transient encounters and persistent interactions. A subsequent statistical analysis of the t_{max} distribution, based on the interquartile range (IQR), allowed the identification of outlier residues deemed particularly relevant for cholesterol binding. In parallel, Ferraro and colleagues computed the spatial density function (SDF), a three-dimensional volumetric map describing the probability of cholesterol occupancy around the protein. Given sufficient sampling, persistent contact regions and high-density regions are expected to converge. Accordingly, the overlap between statistically significant residues and reproducible cholesterol density was used to define cholesterol-binding sites. Using this combined approach, six putative cholesterol-binding sites were identified, primarily located at the protein–membrane interface between transmembrane helices. Several sites overlapped with canonical CARC and CRAC motifs, notably in TM10 and TM4. The analysis further revealed a dynamic binding behavior, with cholesterol molecules transitioning between adjacent sites or between sub-regions of the same site, suggesting a flexible rather than rigid binding mode.

Building on this foundational work, the aim of the present study is to extend and generalize this strategy into a robust, automated, and scalable framework. While Ferraro's protocol proved highly effective for a single, well-characterized system, its application to larger protein families requires increased automation, standardized workflows, and systematic analysis. Accordingly, this work aims to develop, test, and validate a fully automated protocol capable of identifying cholesterol interaction hot spots using only the input protein structure. The method is designed for consistent application across multiple targets and membrane environments and is here systematically applied to class A G protein–coupled receptors (GPCRs), the largest and most extensively studied receptor family. Through comparative analysis, this study seeks to construct an atlas of GPCR–cholesterol interaction regions, identifying recurrent and functionally relevant hot spots in an organized and reproducible manner. Such a framework may ultimately facilitate the rational design of GPCR allosteric modulators and deepen our understanding of cholesterol-mediated regulation in membrane proteins.

Chapter 3. CHAMP, an automatic tool to detect cholesterol hot spots

3.1. Introduction

As discussed in previous chapters, molecular dynamics (MD) simulations provide a powerful framework for probing the mechanistic details of cholesterol binding by explicitly modeling the physicochemical environment of membrane proteins while retaining time-resolved structural detail. However, classical atomistic simulations, in which each atom is represented explicitly, are often constrained by system size and accessible timescales. These limitations restrict sampling of rare but functionally relevant binding events, making the identification of cholesterol-binding sites computationally demanding and frequently impractical. Coarse-grained molecular dynamics (CG-MD) simulations offer an effective alternative, enabling the simulation of large protein–membrane assemblies over extended timescales, in some cases reaching hundreds of microseconds. Numerous studies, including those discussed above, have demonstrated that contact- and density-based analyses of CG-MD trajectories can reliably identify lipid-binding regions.

Motivated by these advances, we developed an automated framework that leverages CG-MD simulations to identify cholesterol interaction hot spots by integrating protein–cholesterol contact analysis with volumetric cholesterol density mapping. In this chapter, we present the resulting tool and its validation on publicly available structures of two membrane proteins: the cannabinoid 1 (CB_1) receptor and the serotonin transporter (SERT).

3.2. Methods

3.2.1. Input structures

Each crystal structure was processed according to the same protocol. The structure of the serotonin transporter (SERT) was co-crystallized with one cholesterol (PDB ID 7LIA) [52], while

the structure of the cannabinoid 1 receptor (CB1) was co-crystallized with two cholesterol molecules (PDB ID 7FEE) [53]. The pre-oriented protein structure was retrieved from the OPM database [54]. Any non-protein atom, such as ligands and other small molecules, were removed and missing heavy atoms and explicit hydrogens were added using PDBFixer [55].

3.2.2. CG conversion and system build

To enable extensive contact sampling, the protocol relied on the Martini 3 CG representation. The atomistic input structure was converted to its CG form using *martinize2*, in the version of the script distributed within the Vermouth package [56]. Default conversion settings were applied whenever possible to maintain robustness and minimize failures due to PDB formatting inconsistencies. However, several parameters were customized and explicitly defined to ensure consistent and reliable output. The -p backbone option created a section in the topology that allowed positional restraints to be applied to backbone beads during CG simulations. Elastic network terms were introduced with -elastic, which generated elastic bonds (700 kJ mol^{-1} nm^{-2}) between backbone beads within a 9 Å cutoff and separated by at least three residues. This configuration stabilized tertiary structure without restricting side-chain motion. Initial residue numbering was retained to simplify downstream analysis. Hydrogens were ignored (-ignh) to prevent atom name conflicts. The -merge all flag forced *martinize2* to write a single topology, avoiding unwanted fragmentation around unresolved regions. As a final safety step, the -eunit system option ensured that the elastic network was built for the complete protein before merging, preventing inconsistencies when gaps were present. A Martini 3 CG PDB file corresponding to the original structure together with its topology were generated.

After CG conversion, the protein–membrane system was assembled using *insane.py* [57]. This grid-based tool generated a Martini 3 membrane of defined composition. We used it to build a cholesterol-containing bilayer around the CG protein obtained in the previous step. The system consisted of the protein placed at the center of a 14.5 × 14.5 × 10.5 nm orthorhombic box. The membrane was constructed around it, and the box was subsequently filled with water beads and ions to neutralize the total charge. After the conversion, the CG representations employed 1334 (SERT) and 705 (CB1) beads, respectively. Embedding the protein in a 75:25 POPC:cholesterol membrane and adding counterions, led to the following protein-membrane systems: i) SERT: 5832 POPC, 1449 cholesterol, 10280 water, 4 chlorine beads, for a total of 18899 beads including the protein; ii) CB1: 6096 POPC, 1521 cholesterol, 10388 water beads, 9 chlorine beads, for a total of 18719 beads including the protein. For SERT, another system was assembled using a

DSPC (50%):POPE (25%):CHOL(25%) membrane composition. This ternary membrane system encompassed: 3888 DSPC, 1932 POPE, 1449 cholesterol, 10287 water, 4 chlorine beads, for a total of 18894 beads including the protein.

3.2.3. Equilibration and Molecular Dynamics Simulations.

Because grid-based membrane construction can introduce unfavourable contacts or rigid conformations, the system underwent a standardized equilibration protocol before production simulations. According to best practices, from this point onward, each step of the protocol was performed N times, with N defined by the number of replicas specified in the run settings. The MD engine adopted for all simulations was GROMACS 2021.5 [58]. System charges are handled by the Martini 3 topology, which was assembled during the system building process. The reaction-field model was used for short-ranged electrostatics, no long-range PME was used. The v-rescale thermostat [59] was used through all equilibration steps and in the production run, while the barostat switched from Berendsen [60] to Parrinello-Rahman [61] at production. Solute (protein and membrane) and solvent (water and ions) were kept into separate temperature coupling groups. Equilibration began with 5000 steps of steepest-descent energy minimization, followed by a 20 ps NVT simulation to stabilize the temperature at 300 K. Three consecutive NPT simulations were then performed. In the first, backbone restraints were applied (1000 kJ mol^{-1} nm^{-2}) while the membrane relaxed during a 100 ns run. In the second, restraints were removed, and the system was equilibrated for 1 ns using a 1 fs timestep to limit abrupt motions of newly unrestrained atoms. The final NPT step consisted of a 1 ns simulation using the recommended 20 fs Martini timestep.

After equilibration and before production, membrane quality was evaluated. Membrane thickness, area per lipid, and lipid order parameters were calculated with MDAnalysis [62,63] and LiPyphilic [64], providing quantitative descriptors of bilayer structure and organization. Partial-density profiles along the Z-axis were generated to assess leaflet continuity.

Production simulations were then carried out for 30 μs to provide sufficient sampling across systems and replicas.

3.2.4. Contacts analysis

Cholesterol–protein contacts were quantified to identify long-lived interaction events along the simulations. Contact detection was performed with MDAnalysis by iterating over all frames of the trajectory and identifying residues within 6 Å of any cholesterol bead. This cutoff was selected based on previously published work [51]. Contacts were evaluated only after the first 500 ns of the production simulation to ensure that measurements were taken within the fully equilibrated regime. For each residue, consecutive frames in which cholesterol remained within the cutoff were grouped into continuous contact events. This approach was adopted to distinguish sustained interactions from intermittent binding and unbinding. Although two residues may exhibit the same total number of contact frames, their biological relevance differs if one forms a single long-lived event while the other engages in multiple short, transient encounters. For this reason, the analysis relied on continuous-contact durations rather than simple frame occupancy. Only residues exhibiting at least one contact were considered for further analysis to avoid inflating the dataset with residues that could not physically interact with the membrane (e.g., residues facing the extracellular, intracellular, or internal cavity regions). From the resulting list of events, the longest contact per residue (t_{max}) was extracted. To identify statistically significant long-lived contacts, the interquartile range (IQR) method was applied using equation 1

$$T = 1.5 \times IQR + Q3 \quad (1)$$

where T defines the upper threshold for outlier detection. This approach flagged residues with contact durations that substantially exceeded the central distribution. The output of this step consisted of a comma-separated table listing the residues classified as statistical outliers in each replica.

3.2.5. Density analysis

The spatial density of cholesterol around the receptor was quantified by computing the Spatial Density Function (SDF), which describes the probability of cholesterol occupying different regions of the simulation box. The 3D space was discretized into voxels, each assigned a value proportional to the frequency with which cholesterol beads occupied that region over the course of the trajectory. Retaining only the voxels with the highest values enabled identification of the spatial regions most frequently sampled by cholesterol. Prior to SDF calculation, the trajectory underwent two preprocessing steps using the GROMACS trjconv tool. The first corrected for

periodic boundary conditions and centered the protein in every frame. The second removed protein rotation and translation, fixing the receptor at the center of the box while allowing the remaining components of the system to diffuse freely. The SDF was then computed with the GROMACS spatial tool, which produced a Gaussian cube file containing the voxel-based volumetric data. This file was parsed, and voxel coordinates were converted from Bohr to Ångström units. The top 0.01% of voxels, which correspond to the regions with the highest occupation probability, were extracted, converted to Cartesian coordinates, and saved as dummy atoms in an .xyz file.

3.2.6. Replica merging

The results of the contact and density analyses across multiple replicas were merged to generate a final file representing the system's hot spots. For contact-based outliers, the t_{max} values from all replicas were compared, and the union of outlier residues was stored. Duplicate residues were common, as the most relevant contacts typically appeared as outliers in multiple replicas; in these cases, the longest contact observed across replicas was retained. For the density analysis, the voxel coordinates from all replicas were combined through a union merge, removing duplicate points.

3.2.7. Output

For each system, four output files were generated. The merged contact outliers were written as a CSV file containing the residue ID, residue name, and t_{max}. Because one of the aims of the pipeline was to enable large-scale GPCR analyses, an optional column with GPCRdb [65]. generic residue numbers (class A) could be added for comparative studies. The cholesterol spatial distribution was stored as a set of dummy atoms in an .xyz file, preserving the element label and Cartesian coordinates for direct visualization. A PDB file of the starting structure, aligned to the last frame of the trajectory used for SDF calculation, was also generated to facilitate inspection of density regions relative to the receptor. Finally, a reduced PDB file containing only the outlier residues and the reproducible density points within 5 Å of those residues was produced. This file was generated from the aligned structure so that it could be directly visualized alongside the density cloud.

3.2.8. Pocket Scoring

A scoring procedure was developed to rank membrane-interface pockets according to their predicted propensity to accommodate cholesterol, using the information generated during hot-spot detection. Pockets were defined as sets of residues identified by a pocket-detection algorithm. Each pocket was evaluated based on (i) the number of residues classified as outliers in the contact analysis and (ii) the number of SDF density points located near the pocket. Density points associated with a given pocket were defined as all density points lying within 5 Å of any atom of any residue forming that pocket. Scoring weights were assigned as 1000 for each outlier residue and 1 for each associated density point. The score was applied to all detected pockets, which were then ranked accordingly. Although the procedure is in principle compatible with the output of any pocket-detection protocol, the examples shown in this study were based on NanoShaper, a probe-based method for cavity detection [66]. NanoShaper was run with a small probe of 1.4 Å and a large probe of 6.5 Å, together with a minimum water count of 2. The size of the larger probe was required to appropriately capture the shallow and planar cavities commonly found at membrane-protein interfaces. Pockets detected by NanoShaper were filtered to retain only membrane-interface sites compatible with cholesterol binding. Internal or extra-/intracellular pockets were removed by discarding any cavity whose centroid lay within 5 Å of the protein's vertical axis or outside the transmembrane region defined by the OPM pseudo-layers. The remaining pockets were then evaluated using the scoring procedure described above.

3.2.9. Parameterization of the Ligands

Org27569 topology was retrieved through manual atom-to-bead mapping and definition of bonded terms. Briefly, Org27569 CG bonded terms were compared with full-atom MD distributions. 12 beads were used to represent the allosteric ligand. Conversely, THC topology (8 beads) is derived from the work reported by Bartocci and colleagues [67].

3.2.10. CB1 receptor - Org27569 – THC system

To construct the system including tetrahydrocannabinol (THC) in the orthosteric binding site, we started from the crystal complex between the THC synthetic analog HU-210 and the CB1 receptor

(PDB ID: 9ERX [68]). With minor modifications to HU-210 structure, we converted this ligand into THC and the resulting complex was minimized with MacroModel [69,70], restraining the position of protein atoms. We then used the same procedure described above to insert the protein – ligand complex in a 95:5 POPC:Org27569 ratio membrane. The CG representation encompassed 686 beads for the protein, 7692 for POPC, 384 for NAM, 8 for THC, 10585 for water, 13 for chlorine, for a total of 19397 beads including ligands.

3.2.11. Org27569-containing Systems Production Runs

Our standard procedure was employed for equilibrating both systems. In the production stage, 10 runs lasting 10 μs each were performed for the system with Org27569 in the membrane, and 5 runs of 10 μs for the system with Org27569 in the membrane and THC in the orthosteric site.

3.3. Results and discussion

3.3.1. Simulations and Analysis Pipeline.

An automated protocol was developed to identify binding hot spots of cholesterol and, with minimal adaptation, other membrane components at the protein–membrane interface, requiring minimal user input beyond specifying run parameters. The protocol is here described in detail and the overall approach summarized in **Figure 2.**

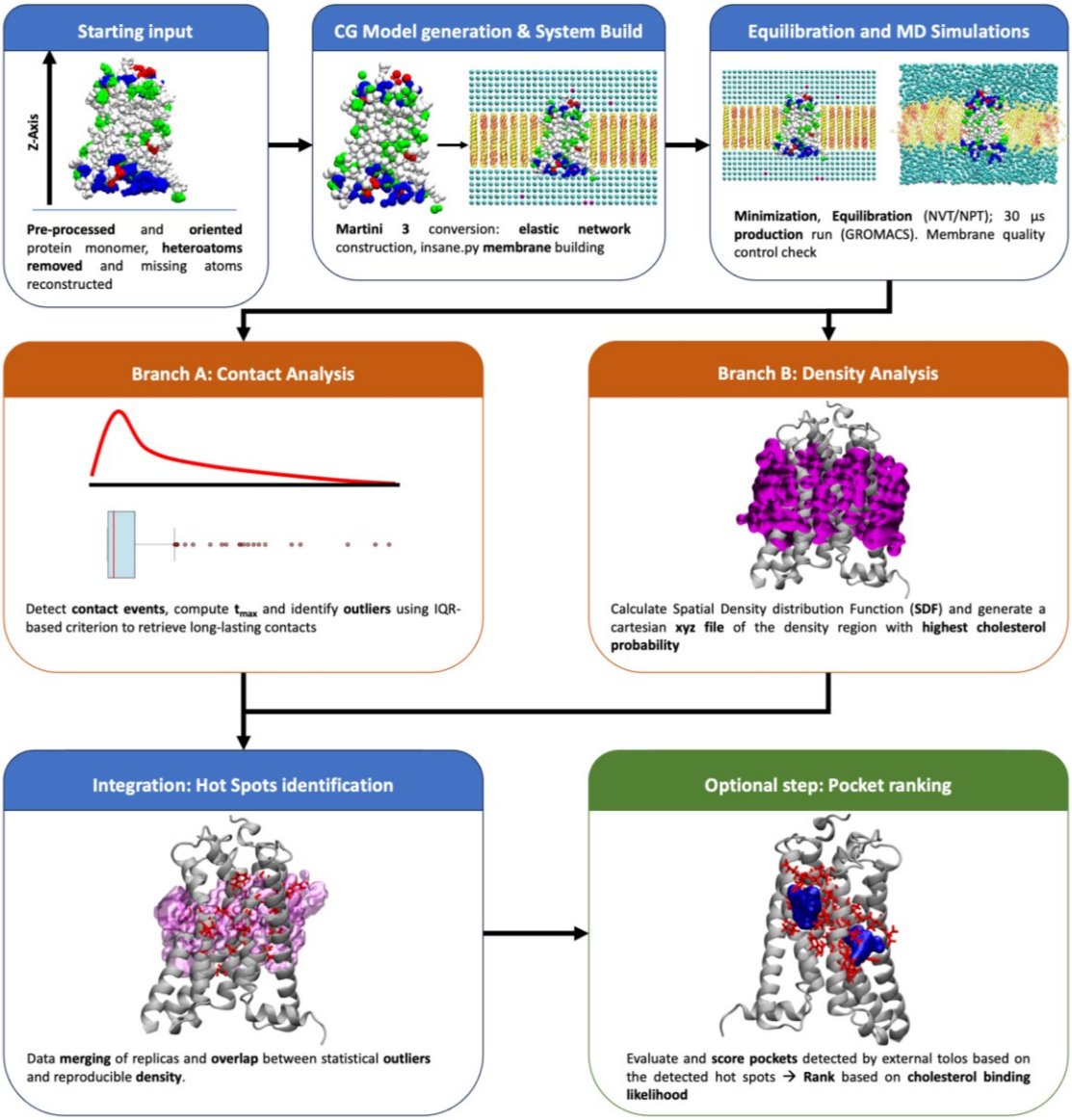


Figure 2. *Standard flowchart of the protocol. The starting input, consisting of a pre-oriented is converted to the Martini 3 representation and embedded in a membrane of the desired composition. The system is then equilibrated and simulated. The trajectories are analysed by the two approaches: contact-based analysis (Branch A) and density distribution computation (Branch B). The agreement regions from both analyses are labelled as hot spots. Optionally, a pocket file can be provided and the pockets will be scored and ranked based on their cholesterol binding propensity.*

3.3.2. Starting Input

The protocol requires a pre-processed, membrane-oriented PDB structure. Structures obtained from the Orientations of Proteins in Membranes (OPM) database [54] are directly compatible, as they are already positioned within the membrane (typically aligned along the Z-axis) and include dummy atoms marking the upper and lower bilayer boundaries. All ligands and non-protein

components must be removed and missing atoms restored. The final input consists of a single oriented monomer with complete receptor residues and no heteroatoms.

3.3.3. CG Model Generation

The CG conversion procedure yields a single, consistent Martini 3 representation for each receptor. Fragmentation issues near unresolved regions, including ICL3 in many GPCR crystal structures, are pre-emptively addressed, residue numbering is preserved, and each output includes a complete elastic network and unified topology. Application of *insane.py* on these standardized models resulted in a complete protein–membrane system with uniform box dimensions, the intended membrane composition, and neutralizing solvent. Each system had the protein correctly centered and embedded in the bilayer, providing consistent inputs for downstream equilibration and sampling stages.

3.3.4. Equilibration and MD Simulations

Within the pipeline, the equilibration stage generates fully prepared membrane–protein systems ready for coarse-grained simulations. After completing minimization, NVT, and NPT steps, the pipeline produces quality-control outputs, including membrane structural descriptors and density profiles along the Z-axis. These files allow users to verify that the system has been assembled and processed correctly before production sampling. At the end of this stage, the pipeline provides a standardized, simulation-ready input for the 30 μs production run performed with the adopted MD engine, GROMACS.

3.3.5. Contact analysis

The contact-analysis stage yields a curated dataset that captures uninterrupted cholesterol–protein interactions. Grouping frames into continuous contact events enables the pipeline to differentiate biologically meaningful, long-lived cholesterol interactions from rapidly exchanging contacts, even when residues exhibit identical total contact occupancies. After filtering non-interacting residues and computing t_{max} values, the IQR-based criterion is applied, and residues classified as

long-lived contact outliers for each replica can be identified. This standardized output provides a consistent summary of meaningful cholesterol-binding events for downstream comparison.

3.3.6. Density Analysis

The density-analysis stage generates a volumetric map describing the regions that are most frequently occupied by cholesterol around the receptor. After trajectory preprocessing and SDF calculation, a Gaussian cube file and a corresponding .xyz file are produced, containing the Cartesian coordinates of the top-density voxels. These points represent the highest-probability cholesterol locations. SDF analyses can be directly related to free energy landscapes in such a manner where areas with enriched cholesterol density correspond to a negative difference in free energy with respect of being located in the bulk of the membrane, suggesting that these positions are energetically favourable interaction sites [71]. All frames of the MD trajectories were retained for analysis, since trajectory striding can influence the resolution of cholesterol density maps.

The overlap of persistent cholesterol–residue contacts and high-density regions across all replicas defines the actual interaction "hot spots," highlighting residues and spatial regions that consistently show the most relevant cholesterol binding.

3.3.7. Output

The pipeline outputs a unified set of cholesterol–interaction features for each system. Contact outliers from all replicas are consolidated into a single list, while density clouds from individual replicas are combined into a consensus spatial distribution. Together, these outputs define reproducible cholesterol–binding hot spots, enabling direct visual and quantitative assessment of the most relevant interaction regions on the receptor.

3.3.8. Pocket rank

The final, optional step of the pipeline uses the information derived from hot-spot detection to evaluate membrane-interface pockets - whether obtained from any pocket-prediction tool or defined directly by the user - for their compatibility with cholesterol binding. By combining

outlier residues from the contact analysis with high-probability regions from the SDF, pockets supported by both types of evidence are assigned greater relevance, thus separated from pockets that lack predicted affinity for cholesterol.

3.3.9. Python package

To support the use and integration of this protocol, we developed a Python package available on PyPI under the name CHAMP (Cholesterol Hot spot Automated Mapping Protocol). The package includes all required external tools and modules, except for GROMACS, which must be compiled separately. To ensure flexibility and scalability, the workflow exposes a wide set of configurable parameters rather than relying on hardcoded values. A full tutorial is provided in the repository, but key options include specifying membrane composition, selecting the number of replicas, or providing a predefined pocket list to directly obtain their scoring and ranking at the end of a run. When using CHAMP on a GPCR, users can also supply a GPCRdb-annotated file to maintain annotation in the output. The package also includes a toolbox to assist with input preparation. Two scripts are provided: one generates a template parameter file needed to launch the protocol, while the other can simply align a structure along the Z-axis when no membrane-oriented model is available. The protocol requires a pre-aligned structure, such as those obtained from the OPM server; when unavailable the alignment script can be applied should be used with care. This procedure yields an approximate Z-axis orientation along the longest protein direction (usually the vertical axis in membrane proteins). As some proteins may have more complex orientation and tilt which is not automatically captured the resulting data should be interpreted accordingly. An additional third script provides an automatic GROMACS benchmarking routine to help users determine the optimal configuration for Martini 3 CG simulations. In the Martini 3 cholesterol topology, a virtual site is used to reproduce the sterol ring geometry. Currently, while most tasks, such as short-range nonbonded interactions, PME, and bonded forces, can still be off-loaded to the GPU, the presence of virtual sites prevents the MD update step (position and velocity integration and virtual-site reconstruction) from running on the GPU [72]. This forces the integrator to run on the CPU, making overall performance sensitive to CPU–GPU communication and strongly dependent on the number of CPU cores available. Given only the number of available cores, the provided script selects suitable test configurations, runs the benchmarks, and produces a concise performance summary.

A final point concerns the size of the data generated. For a standard run consisting of two 30 μs replicas, the total output can reach up to 800 GB. Of this, around 780 GB arise from the MD trajectories and the corresponding trjconv iterations used to produce the density maps, each

trajectory contributing around 130 GB. The remaining space is occupied by the volumetric density maps generated by the spatial analysis tool (roughly 3 GB per replica) and the associated structures, topologies, intermediate files, and outputs.

3.3.10. Serotonin Transporter

We first applied CHAMP to SERT to enable direct comparison with the results reported by Ferraro et al [51]. The transporter was analysed in its outward-facing conformation, taken from a cryo-EM structure of human SERT bound to serotonin, which also features a cholesterol molecule resolved at the membrane-exposed surface (PDB ID: 6LIA [52]). To assess the effect that a realistic membrane could have in comparison with a simpler composition, we applied the protocol twice, each time using a different membrane. The first system used a simple POPC(75%):CHOL(25%) bilayer. Although modern simulations increasingly favor complex lipid mixtures, POPC:CHOL membranes remain a robust and widely used benchmark, capable of reproducing liquid-ordered and liquid-disordered behavior depending on sterol content [73]. This composition provides a generalizable reference point that facilitates comparison across systems. The second membrane reproduced the lipid environment used by Ferraro et al., a ternary DSPC (50%):POPE (25%):CHOL(25%) mixture designed to better approximate raft-like membrane domains [74]. From each run, two independent replicas of 30 μs were generated.

Table 2. *Outlier residues identified by Ferraro et al. and in the two membrane compositions analysed in this study. Residues classified as outliers in all three simulations are shown in bold.*

Source	Detected outliers (residue IDs)
Ferraro et al. 2016	121, 125, **289**, 337, 362, **363**, 365, **366**, **367**, **369**, **495**, 531, 561, 572, 575, 579
POPC:CHOL	93, 122, 123, 126, 127, 166, 170, 171, 174, 191, 256, 264, 265, 268, 280, 283, 287, **289**, 293, 326, **363**, **366**, **367**, **369**, 370, 373, 429, 436, 454, 474, 475, 478, 479, **495**, 514, 577, 581, 584, 597
DSPC:POPE:CHOL	122, 123, 127, 163, 167, 170, 171, 174, 191, 253, 256, 265, 269, 280, 283, **289**, 313, 320, 326, **363**, **366**, **367**, **369**, 370, 373, 374, 436, 468, 474, 478, 479, **495**, 514, 584, 590, 597

Membrane composition had only a modest influence on the residues classified as statistically significant long-lived contacts (**Table 2**), with t_{max} values of the outliers ranging from 73.72 ns (POPC:CHOL) and 82.66 ns (DSPC:POPE:CHOL) to 422.46 ns (POPC:CHOL) and 471.46 ns (DSPC:POPE:CHOL). The POPC:CHOL system yielded 39 outliers, whereas 36 were detected in the DSPC:POPE:CHOL system, with substantial agreement: 27 residues were shared between the two distributions. This level of agreement indicates that, in this case, the protocol consistently recovers the principal features of cholesterol recognition, even when applied across different membrane environments. Both sets also showed notable convergence with the 16 outliers reported by Ferraro et al. Six residues appeared in all three datasets, three of which belong to the cholesterol binding pocket resolved in the experimental structure, considering those residues within a shell of 5Å of the cholesterol molecule (labelled and coloured in red in **Figure 3A**).

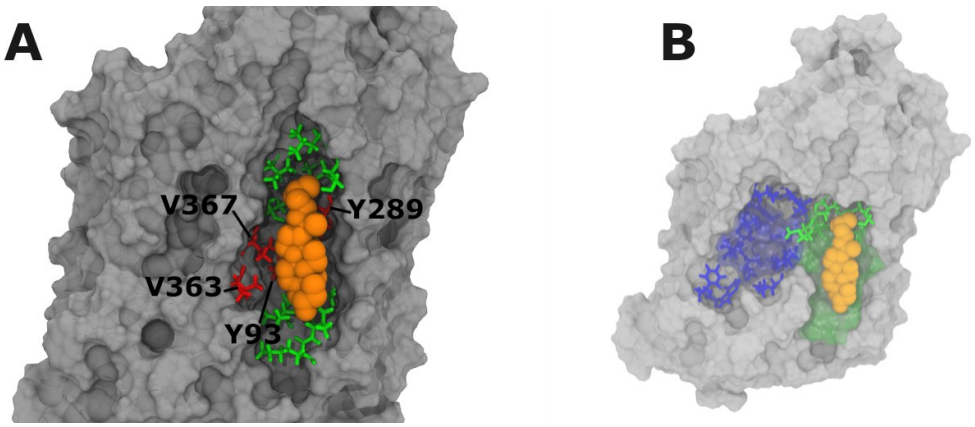


Figure 3. *Results of the CHAMP protocol on the serotonin transporter. (A) Crystallographic pocket (green licorice) of the cholesterol molecule (orange) in SERT. The residues consistently detected as outliers across all three datasets are highlighted in red and explicitly labelled. (B) Spatial overlap between previously reported cholesterol interaction sites (blue and green surfaces), and our top and fourth scoring (licorice blue and green, respectively) NanoShaper-detected pockets.*

Some differences between our results and those of Ferraro et al. were expected. Their study employed longer simulations, more replicas, and a customized workflow, all of which can increase the separation between long-lived contacts and the central distribution, resulting in fewer residues being flagged as outliers. Additionally, previous work was based on a homology model, obtained from a bacterial member of the neurotransmitter sodium symporters (NSS) family, the Leucine transporter (LeuT), and used a previous version of the Martini force field, both of which introduce inherent structural and methodological differences. Nonetheless, given that the goal of the present pipeline is to provide an automated and computationally efficient approach, the close correspondence with Ferraro's more intensive setup supports both the robustness of the method and its suitability for large-scale analyses. Although our methodology inherently represents a

compromise between speed and accuracy, the agreement observed here indicates that this trade-off remains well balanced for practical applications.

Analysing the static structure of the transporter, NanoShaper identified 36 pockets, of which 16 met the geometric criteria for membrane-interface accessibility and were retained as cholesterol-binding candidates. These pockets were then ranked using the hot-spot information derived from the contact outliers and SDF density points. The analysis was performed independently for both membrane compositions (**Table 3**).

Table 3. *Top five highest-scoring NanoShaper pockets identified in SERT. The top five pockets are reported for the POPC:CHOL and DSPC:POPE:CHOL systems along with the corresponding number of density points and scores.*

Rank	Pocket id	ResIDs	System					
			POPC:CHOL			DSPC:POPE:CHOL		
			Density Points	Outlier Score	Total Score	Density Points	Outlier Score	Total Score
1	27	105 118 121 122 125 126 130 362 365 366 369 370 372 373 536 537	952	5000	5952	1504	6000	7504
2	33	184 187 261 262 265 280 283 284 287 429 432 433 436 440	1223	4000	5223	534	6000	6534
3	26	170 171 174 245 248 476 479 488 491 492 573 577 581	796	4000	4796	216	6000	6216
4	31	289 292 293 370 371 374	793	3000	3793	601	3000	3601
5	30	148 163 166 167 169 170 443 446 450 454 465 469 472 473	494	3000	3494	303	3000	3303

Comparison of the two membrane conditions showed that the top five ranked pockets were identical, indicating that membrane composition did not meaningfully alter the pocket ranking at the cholesterol ratio tested. The DSPC:POPE:CHOL system exhibited slightly higher scores for the top three pockets, but the differences were small and likely not sufficient to suggest a composition-dependent effect, particularly given that only four replicas were analysed. Notably, the pocket ranked fourth overlapped with the crystallographic cholesterol site. We next compared our highest-ranked pockets with the cholesterol-binding sites reported by Ferraro et al. Including the crystallographic site, two of the previously reported pockets were recovered. As shown in **Figure 3B**, our first-ranked pocket (licorice blue) and fourth-ranked pocket (licorice green) overlapped with the previous study's K (blue surface) and C (green surface) sites, respectively.

To complement the static analysis with a dynamic characterization of pockets along the trajectory, NanoShaper was also applied through Pocketron as implemented in the BiKi Life Sciences software [66,75,76]. Pocketron evaluates pockets on a frame-by-frame basis along the MD trajectory, enabling the monitoring of pocket appearance, persistence, and geometric evolution over time. The resulting pocket ranking was largely consistent with that obtained from the static NanoShaper analysis. Although the coarse-grained simulations employ an elastic network and therefore do not sample large-scale conformational rearrangements, the close agreement between static and dynamic pocket analyses indicates that the approach is robust to small local fluctuations, such as side-chain motions. This robustness comes at the explicit cost of restricting global protein flexibility, a trade-off that is inherent to the efficiency gains afforded by coarse-grained molecular dynamics. In light of this consistency, and to avoid unnecessary methodological redundancy, subsequent analyses were carried out using the static NanoShaper-based pocket identification.

In this context, it is notable that several crystal structures of the transporter consistently report a cholesterol molecule bound to a region defined by residues 573, 576, 577, and 580 [77,78]. While the pocket analysis does indeed identify a putative binding pocket in this region, this is not ranked among the highest sites in terms of cholesterol-binding propensity, suggesting that the limited conformational sampling intrinsic to the coarse-grained model may influence the relative prioritization of such sites. In order to assess this hypothesis, the full protocol and the NanoShaper analysis were also ran on one of the crystal structures where the cholesterol appears in the aforementioned position, nearing residues 573-580 (PDB ID: 5I6X). Interestingly, the top scoring pocket corresponded to the region closest to the crystallographic cholesterol, formed by the residues 170, 171, 174, 245, 248, 476, 479, 488, 491, 492, 577 and 581, equivalent to that scoring third in the previous structure. Additionally, within the top 5 scoring pockets, equivalent cavities to those in the previous structure were identified. This analysis further strengthens the idea that while the starting structure of the same protein might have little influence in pocket detection

itself, it is a key factor for the hot spot detection, as the analyses showed a certain level of preference in prioritizing the original pocket.

3.3.11. Cannabinoid 1 receptor

We next examined the CB1 receptor, as representative of class A GPCR characterized by allosteric modulators with pharmacological interest and identified binding pockets. We selected the crystal structure of CB1 receptor bound to the orthosteric agonist CP55940 and to the positive allosteric modulator ZCZ011 [53], which also contains two resolved cholesterol molecules. This system is a suitable test case because both experimental studies and recent atomistic simulations have highlighted cholesterol as a modulator of CB1 receptor function, reporting specific interaction regions and long-lived binding events [40,79]. Using CB_1 receptor therefore allowed us to evaluate whether the proposed workflow can recover cholesterol-accessible sites in line with experimental observations and with results obtained from more computationally demanding methodologies.

Following the protocol, the structure was converted to CG using the Martini 3 topology, embedded in a POPC(75%):CHOL(25%) membrane and simulated and analysed on two replicas. After contact analysis, we obtained 27 outliers, with t_{max} ranging from 96.72 ns to 867.10 ns, as shown in **Table 4**. In total, we identified three cholesterol-interaction hot spots at the protein–membrane interface, defined as regions where long-residence-time outlier residues and high SDF cholesterol density co-occur. Notably, two of these hot spots coincide with the cholesterol positions resolved in the crystal structure (coloured in red and blue in **Figure 4**). In addition to these two expected regions, the analysis revealed a third area (green in **Figure 4**) showing a clear consensus between persistent contacts and enriched cholesterol density, suggesting a possible, previously unreported, interaction region. Although candidate site identification relies on a consensus criterion that provides a conservative and robust basis for highlighting cholesterol-interaction regions, whether this third site represents a genuine binding preference remains to be established experimentally.

Table 4. *Detected outlier residues identified in CB_1 simulations after the replica merging.*

Membrane	Detected outliers (residue IDs)
POPC:CHOL (75:25)	$142^{1.58}$, $161^{2.48}$, $165^{2.52}$, $168^{2.55}$, $169^{2.56}$, $172^{2.59}$, $191^{3.27}$, $195^{3.31}$, $198^{3.34}$, $201^{3.37}$, $202^{3.38}$, $208^{3.44}$, $209^{3.45}$, $212^{3.48}$, $232^{4.41}$, $236^{4.45}$,

	$240^{4.49}$, $241^{4.50}$, $243^{4.52}$, $245^{4.54}$, $282^{5.46}$, $287^{5.51}$, $291^{5.55}$, $350^{6.42}$, $353^{6.45}$, $360^{6.52}$, $361^{6.53}$

One of the identified hot spots overlaps very well with the binding site reported as BS1 in a study by Vanegas and co-workers [40]. This region includes $TRP241^{4.50}$ (Ballesteros-Weistein numbering superscript [80]). This residue is the most well conserved in TM4 across class A GPCRs and has been consistently reported to participate in cholesterol interactions [81] and not only in the CB1 receptor, but also in other specific cases, such as the β2-adrenergic receptor [21,37] or the nociceptin opioid peptide receptor [82]. $TRP241^{4.50}$ plays a key role in stabilizing the receptor's transmembrane structure through hydrophobic contacts. Its side chain is capable of forming π-π interactions and hydrogen bonds, which help preserve the proper conformation of the receptor. Importantly, $TRP241^{4.50}$ is involved in the recognition of allosteric modulators. The negative allosteric modulator Org27569 binds to a site outside the transmembrane core that includes $TRP241^{4.50}$, with the indole group of Org27569 engaging in aromatic stacking with the indole side chain of $TRP241^{4.50}$, underscoring the residue's relevance in allostery-related methods. Furthermore, two more hot spots among those identified here partially overlap with areas, namely BS2 and BS5.

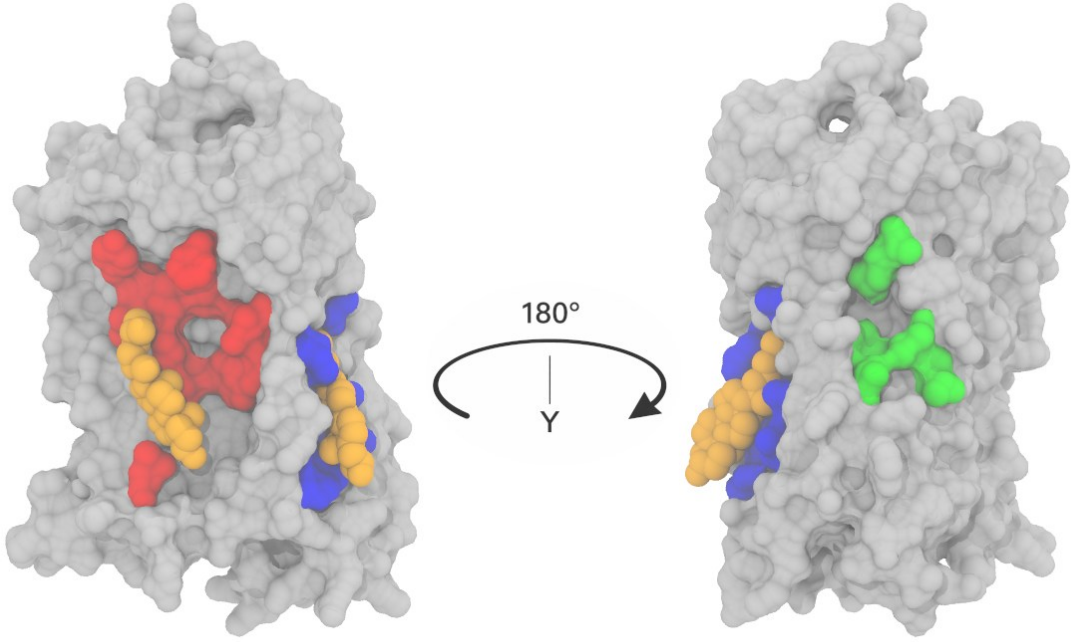


Figure 4. *Hot spots detected on the CB_1 receptor based on the agreement of outliers with reproducible density. For visual clarity, only the residues are shown, while the density cloud is omitted. Three distinct interaction areas are identified: two holding cholesterol molecules (red and blue pockets), and a third representing a putative, unreported, novel interaction site (in green).*

Table 5. *Top five high scoring pockets detected by NanoShaper.*

Rank	Pocket id	ResIDs (outliers bolded)	System		
			Cannabinoid receptor 1		
			Density Points	Outlier Score	Total Score
1	11	**165$^{2.52}$ 169$^{2.56}$ 172$^{2.59}$** 190$^{3.26}$ **191$^{3.27}$** 192$^{3.28}$ 194$^{3.30}$ **195$^{3.31}$ 198$^{3.34}$** 199$^{3.35}$ **241$^{4.50}$ 245$^{4.54}$** 248$^{4.57}$ 249$^{4.58}$ 252$^{4.61}$	1761	8000	9761
2	15	**201$^{3.37}$** 204$^{3.40}$ 205$^{3.41}$ **208$^{3.44}$ 209$^{3.45}$ 240$^{4.49}$ 243$^{4.52}$** 244$^{4.53}$ 247$^{4.56}$ 279$^{5.43}$ **282$^{5.46}$** 283$^{5.47}$ 286$^{5.50}$	1634	6000	7634
3	12	283$^{5.47}$ 284$^{5.48}$ **287$^{5.51}$** 288$^{5.52}$ **291$^{5.55}$** 352$^{6.44}$ **353$^{6.45}$** 356$^{6.48}$ 357$^{6.49}$ 360$^{6.52}$	1986	4000	5986
4	14	**291$^{5.55}$** 294$^{5.58}$ 295$^{5.59}$ 346$^{6.38}$ 349$^{6.41}$ **350$^{6.42}$ 353$^{6.45}$**	1232	3000	4232
5	9	138$^{1.54}$ 141$^{1.57}$ 142$^{1.58}$ 148$^{98.83}$ 153$^{2.40}$ 154$^{2.41}$ 157$^{2.44}$ 158$^{2.45}$	287	0	1287

We next characterized putative pockets at the protein–membrane interface and ranked them according to cholesterol-binding propensity. NanoShaper initially identified 16 geometrically defined cavities, which were reduced to 7 after applying the filtering criteria described in the Methods section. Importantly, the highest-ranked pockets closely correspond to the interaction regions identified by the hot-spot analysis, providing independent confirmation that the detected hot spots map onto *bona fide*, well-defined geometric pockets rather than diffuse regions of lipid enrichment. Specifically, the top-ranked (red in **Figure 5**) and fifth-ranked (yellow in **Figure 5**) pockets together correspond to the previously proposed CRAC binding site and encompass the conserved W241$^{4.50}$ residue. Consistent with crystallographic evidence, one of the co-crystallized cholesterol molecules occupies this site. The second cholesterol molecule is located within the second-ranked pocket (blue in **Figure 5**), partially overlapping with the BS3 site reported by Vanegas and coworkers. Notably, the third-ranked pocket corresponds to the previously discussed hot spot that suggests a potential cholesterol interaction site that has not been previously reported.

The correspondence between dynamically identified hot spots and independently detected geometric pockets renders these sites directly actionable, as they define structurally resolved cavities that can be immediately exploited in structure-based drug design efforts, including docking, pharmacophore modelling, and rational optimization of allosteric modulators at the membrane–protein interface.

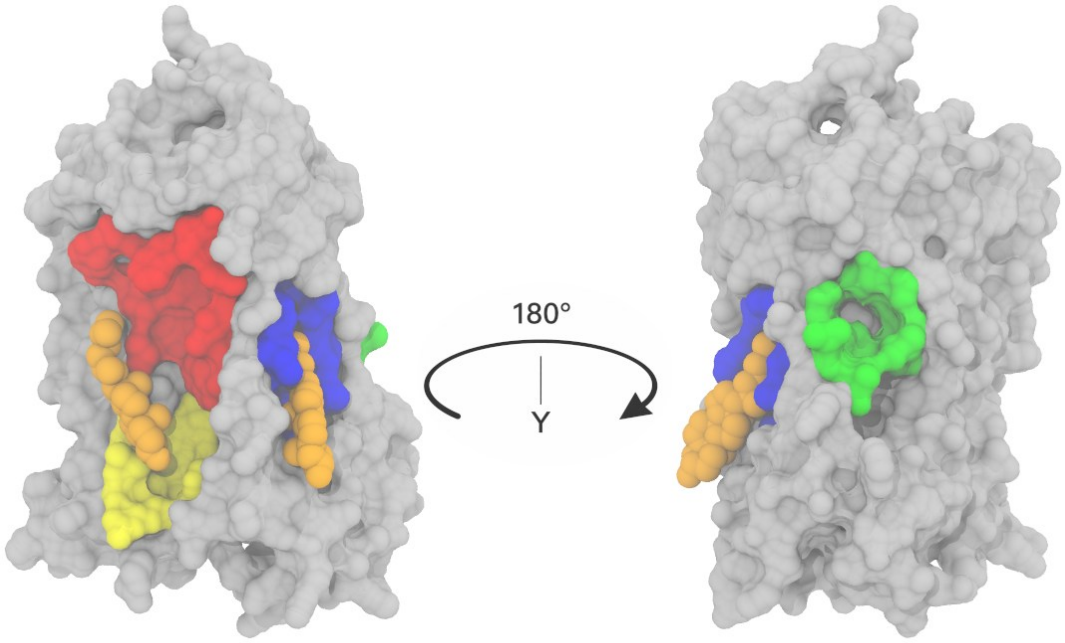


Figure 5. *Spatial distribution of the top-scoring NanoShaper pockets in the CB_1 receptor. The identified pockets are ranked according to their CHAMP cholesterol binding propensity and color-coded as follows: 1st (red), 2nd (blue), 3rd (green), and 5th (yellow).*

3.3.12. Cannabinoid 1 receptor negative allosteric modulator

Our fully automated protocol was designed with cholesterol in mind, aiming to identify critical hot spots at the membrane–protein interface with potential allosteric implications. However, in principle, our framework can be readily extended to other allosteric modulators acting at the membrane–protein interface providing the framework for the identification of privileged interaction sites of small molecules and drug-like compounds. To provide a preliminary proof of concept in this direction, we applied the protocol to the Org27569 ligand. This molecule is a negative allosteric modulator of the CB_1 receptor G-protein signalling [83]. Additionally, a crystal structure of their complex is available and can be used as a reference (PDB ID: 6KQI [84]).

For system construction, we build upon a previous work by Bartocci and colleagues [67], where the authors study allosteric cannabinoid binding at the glycine receptor α1. Specifically, we embedded the CB_1 receptor in a POPC lipid membrane with 5% of Org27569 (i.e. 95:5 POPC:ligand ratio), corresponding to a ligand concentration of about 25 mM. Our goal was to assess the capability of recapitulating the binding site observed in the crystal among detected binding hotspots. Such a semi-quantitative validation would underscore the actual potential for prospective use.

Here, in analogy with Bartocci *et al.* and to increase the statistical robustness of the observation for this molecule, we performed 10 replicas of 10 μs each. By analysing the generated trajectories through our CHAMP protocol, we identified the most persistent contacts between the CB1 receptor and the ligand Org27569. The main findings are summarized in **Table 6** and **Table 7**.

Table 6. *Residues from the CB1 receptor establishing the most persistent contact in each of the 10 replicas of CG MD.*

Replica index	Main contact	Duration (ns)
1	TRP241$^{4.50}$	419
2	PHE379$^{7.35}$	1592
3	PHE379$^{7.35}$	2892
4	TRP241$^{4.50}$	462
5	GLY195$^{3.31}$	447
6	TRP241$^{4.50}$	577
7	TRP241$^{4.50}$	421
8	ALA202$^{3.38}$	316
9	SER284$^{5.48}$	328
10	VAL282$^{5.46}$	718

In particular, **Table 6** reports, for each of the 10 CG-MD replicas, the most persistent contact observed as well as its duration. Interestingly, in six out of ten runs the longest contact is established by residues belonging to (TRP241$^{4.50}$) or in close proximity of (GLY195$^{3.31}$ and ALA202$^{3.38}$) the allosteric site observed in the crystal complex.

Table 7. *Ranking of the most persistent contacts established by residues from the CB_1 receptor with the Org27569 ligand. The top ten residues are reported, plus additional residues that displayed highest persistency in at least one of the simulations.*

Residue	Frequency	Average duration (ns)	Shortest duration (ns)	Longest duration (ns)
TRP241$^{4.50}$	10	396	238	577
ALA202$^{3.38}$	10	286	147	419
ALA198$^{3.34}$	10	247	133	438
TYR172$^{2.59}$	10	227	183	275
GLY195$^{3.31}$	10	226	137	447
VAL161$^{2.48}$	9	201	143	286
PHE208$^{3.44}$	8	213	123	329
ILE243$^{4.52}$	7	296	169	486
THR201$^{3.37}$	7	242	142	337
PHE191$^{3.27}$	7	144	100	215
VAL282$^{5.46}$	6	344	145	718
PHE379$^{7.35}$	3	1582	261	2892
SER284$^{5.48}$	1	328	320	328

In **Table 7**, residues from the CB1 receptor are ranked according to their frequency of appearance across replicas and the related timing information. Notably, all residues engaging persistent contacts in 90% of the simulations belong to the allosteric binding site. The only exception is TYR172$^{2.59}$, which is still relatively close to the allosteric site. It is worth noting that TRP241$^{4.50}$, a central residue in the allosteric site ranks first in **Table 7**, being observed in all ten CG-MD replicas while also displaying longest average duration. PHE379$^{7.35}$ emerges as the residue with the longest average contact, although observed in only three runs. Its duration in replicas 2 and 3 is particularly prolonged with 1592 and 2892 ns (**Table 6**), respectively, suggesting a stable interaction. This residue is located within the transmembrane helical bundle and forms part of the orthosteric binding site. Notably, an interaction between Org27569 and the orthosteric site has

not been previously reported. However, the frequency of this interaction is substantially lower than that observed for residues within the allosteric site, indicating a clear preference for allosteric engagement. By aggregating the results from the ten replicas, ligand density maps projected on the atomistic structure of the CB_1 receptor provide a straightforward visualization of prominent binding hotspots (**Figure 6A**).

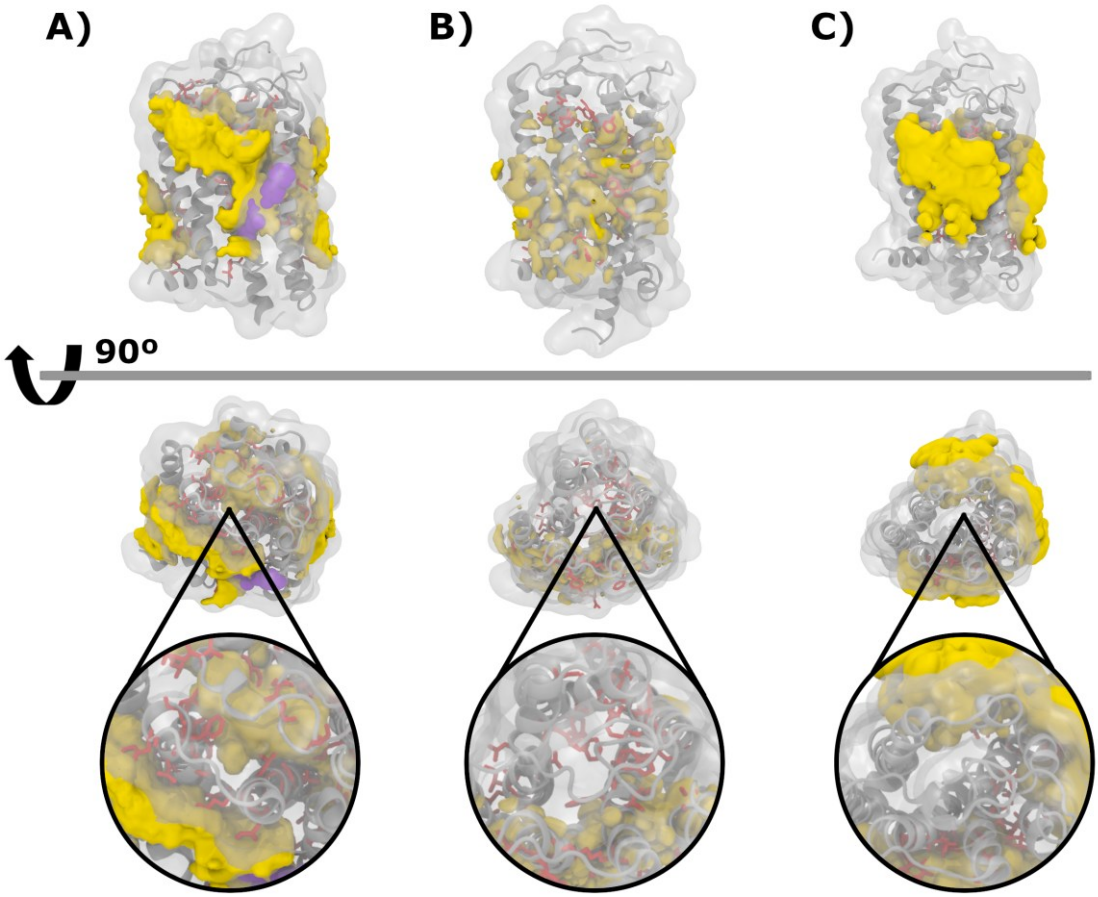


***Figure 6**. Ligand density over the CB_1 receptor surface for A) Org27569, B) Org27569 in presence of THC in the orthosteric site, and C) cholesterol. For all systems, front and upper views are shown. For the upper view, additional focus on the orthosteric binding site within the transmembrane helical bundle is given. The ligand densities are shown in yellow on the CB_1 receptor crystal structures used to construct the systems (PDB IDs: 6KQI [84], 9ERX [68], and 7FEE [53] for panels A, B and C, respectively). In panel A, the Org27569 ligand in its crystallographic pose is shown in purple to highlight the corresponding allosteric binding site. All the residues deemed relevant (statistical outliers) in contacting the ligand are highlighted in red.*

The results for Org27569 can be compared with those from applying the standard CHAMP protocol with cholesterol on the CB_1 receptor (**Figure 6C**). The overall picture is relatively consistent, with cholesterol demonstrating a lower engagement for the binding site of Org27569 and, most importantly, none within the orthosteric cavity. On the one hand, this observation underscores the ability of the pipeline to discriminate between different ligands. On the other

hand, the comparison highlights a peculiar tropism of the allosteric modulator for the orthosteric binding site. To further assess the tendency of Org27569 to occupy the orthosteric site in realistic conditions, we repeated the simulations in presence of the natural ligand tetrahydrocannabinol (THC). To this end, we used the crystal complex between THC and the CB_1 receptor, and embedded it in the same membrane environment containing POPC and Org27569 as in the original simulations. As expected, from five exploratory simulations lasting 10 μs, we can already see that no contacts are established by Org27569 with residues from the orthosteric site of the CB_1 receptor. This underscores the importance of performing simulations in realistic conditions to circumvent, and possibly discriminate, potential simulation artifacts. Additionally, despite the relatively more limited sampling, residues TRP241$^{4.50}$ and VAL161$^{2.48}$ from the allosteric site still emerge with high frequency and prolonged contact duration. A graphical representation of these findings is reported in **Figure 7**, where we highlight the location of the residues within the 3D structure of the CB_1 receptor.

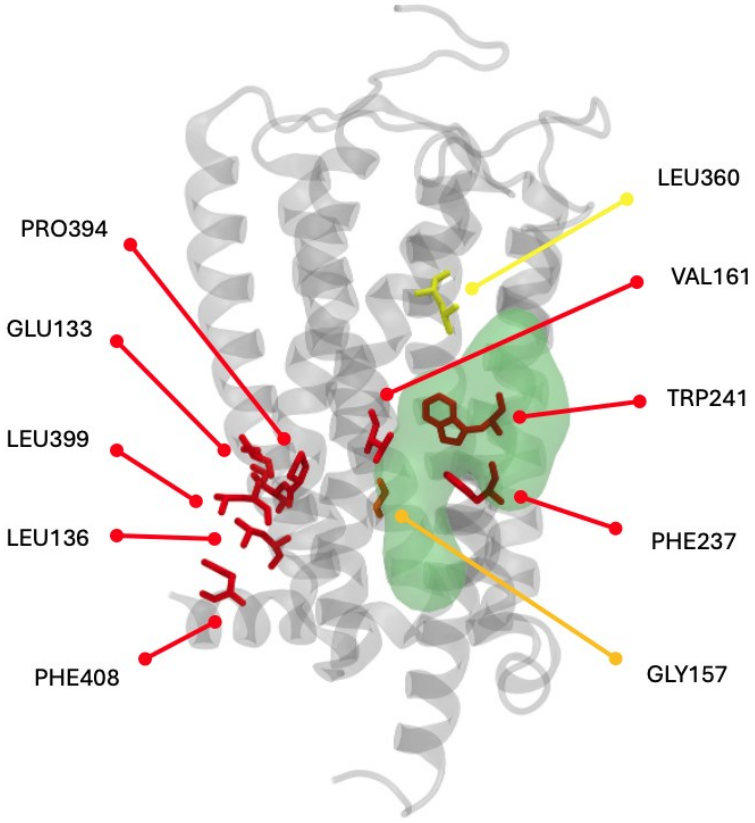


Figure 7. *CB_1 receptor residues involved in the most persistent contacts with the allosteric modulator Org27569. The residues are labelled and highlighted in licorice on the cartoon representation of the receptor. Residue colour scale ranging from yellow to red indicates increasing persistence of the contact. To orient the reader, the allosteric binding site is highlighted as green transparent surface.*

As a final application, and to showcase the flexibility of our CHAMP protocol, we performed the same analysis on externally detected pockets. Therefore, the two CB_1 structures used in this section for the simulations without and with the presence of THC (PDB IDs: 6KQI and 9ERX

respectively) were evaluated using NanoShaper and those pockets ranked based on the data obtained through the analysis of the simulations performed.

As expected, in Figure 7 the highest scoring pocket belongs to the allosteric site of the NAM, which overlaps with the binding site for cholesterol. Interestingly, the other cholesterol binding site in CB_1 (blue surface in Figure 7) is also highly ranked, appearing in the second position for the system without THC and in fourth position when THC is present. Additionally, the rest of the pockets in the top 5 also show significant agreement with those identified for cholesterol. This overlap between cholesterol and NAM hot spots could indicate similar behaviour of both molecules around the CB_1 receptor.

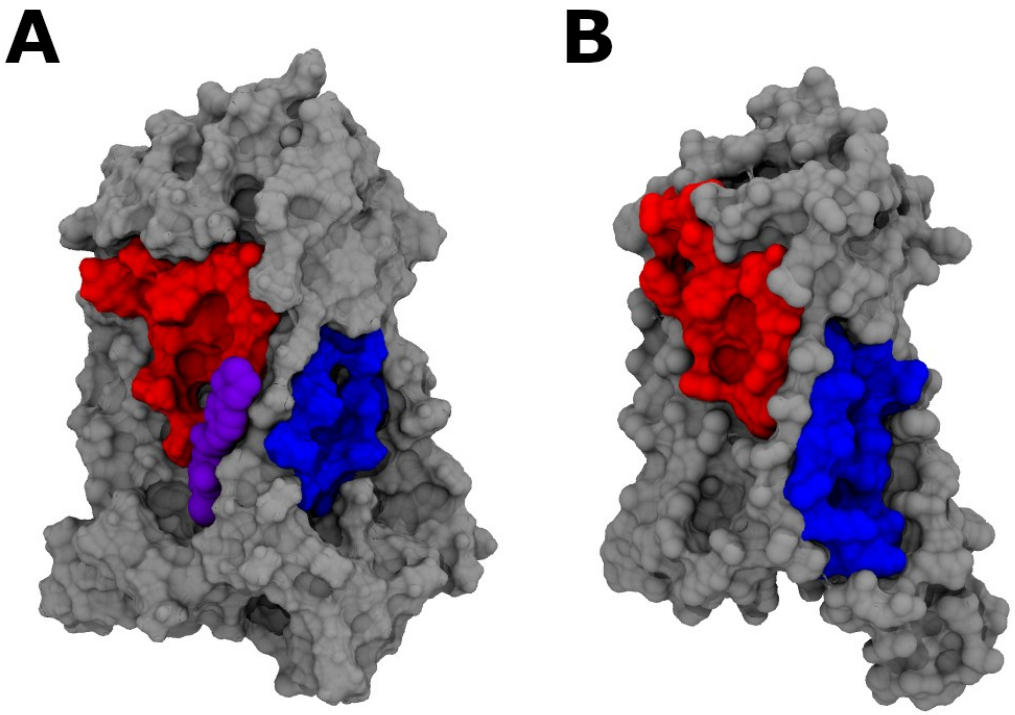


Figure 8. *Comparison of overlapping cholesterol and NAM binding hotspots in the CB_1 receptor. NanoShaper-detected pockets are ranked to identify spatial consensus between cholesterol and the NAM ligand. Results are shown for the CB_1 receptor simulated in the absence (A, PDB ID: 6KQI) and the presence of THC (B, PDB ID: 9ERX).*

All the work discussed in this chapter will be included within a publication, with the manuscript currently in preparation.

Chapter 4. Protocol Development - Technical Details

4.1. Introduction

The design and development of the automated protocol presented in this work required a sustained and systematic effort to optimize each individual step, with the overarching goal of minimizing user intervention in the final workflow. From the outset, our primary objective was full automation. Accordingly, the protocol was conceived as a tool that takes a well-defined input and produces a corresponding output without requiring manual interaction during execution.

Achieving this goal required the careful design and extensive testing of a sequence of processing steps that are robust and broadly applicable, irrespective of the specific membrane protein provided as input. In addition, the protocol had to handle all necessary manipulations of intermediate files generated throughout the workflow. Given that our analyses involve long molecular dynamics trajectories and other large data files, and that most of the pipeline operates automatically, particular attention was devoted to efficient file management. Specifically, operations such as copying, deleting, and modifying files were optimized to balance overall runtime performance with the need to retain sufficient intermediate data. This ensures that all relevant information remains available should the user need to inspect, debug, or revert to any intermediate stage of the process.

The outcome of these efforts is reflected in the final automated pipeline, which is designed to avoid unnecessary file duplication and to prevent the re-execution of steps whose outputs have already been successfully generated. At the same time, the protocol preserves all molecular dynamics trajectories, energy files, and other auxiliary data required for in-depth analysis or troubleshooting of individual systems.

In this section, we describe the development and validation of the protocol in detail, focusing on the rationale behind the implemented actions, the strategies adopted to overcome technical challenges, and the approaches used for error detection and debugging.

4.2 Molecular Dynamics Simulations

The first major challenge in adopting a fully automated approach was the development of a robust molecular dynamics (MD) protocol capable of equilibrating systems and producing reliable coarse-grained (CG) trajectories for any G protein–coupled receptor (GPCR), or more generally for membrane proteins. Initial tests were performed using the Martini 2 force field (reference), which at the time was the most widely adopted version for CG-MD simulations including cholesterol. Moreover, the cholesterol model and topology within Martini 2 were still undergoing active development and refinement [85,86].

The initial systems were constructed using the CHARMM-GUI Martini Bilayer Maker [87,88], which also generates the corresponding GROMACS input files required to perform MD simulations. The builder provides a total of seven GROMACS parameter files: two for energy minimization, four for NPT equilibration with progressively decreasing positional restraints on protein beads and lipid headgroups, and one for the final production run. These input files were used to equilibrate the system using GROMACS 2022.2, with the cannabinoid receptor 1 (CB_1; PDB ID: 5XR8 [89]) embedded in a POPC membrane containing 25% cholesterol. Although the energy minimization and equilibration stages yielded a stable ensemble, as assessed using standard structural and energetic metrics (**Figure 9**), the production simulations consistently failed after a few nanoseconds. In all cases, the simulations terminated with errors associated with the Linear Constraint Solver (LINCS), the algorithm responsible for maintaining bond-length constraints and, consequently, for ensuring the numerical stability of the system during the simulation.

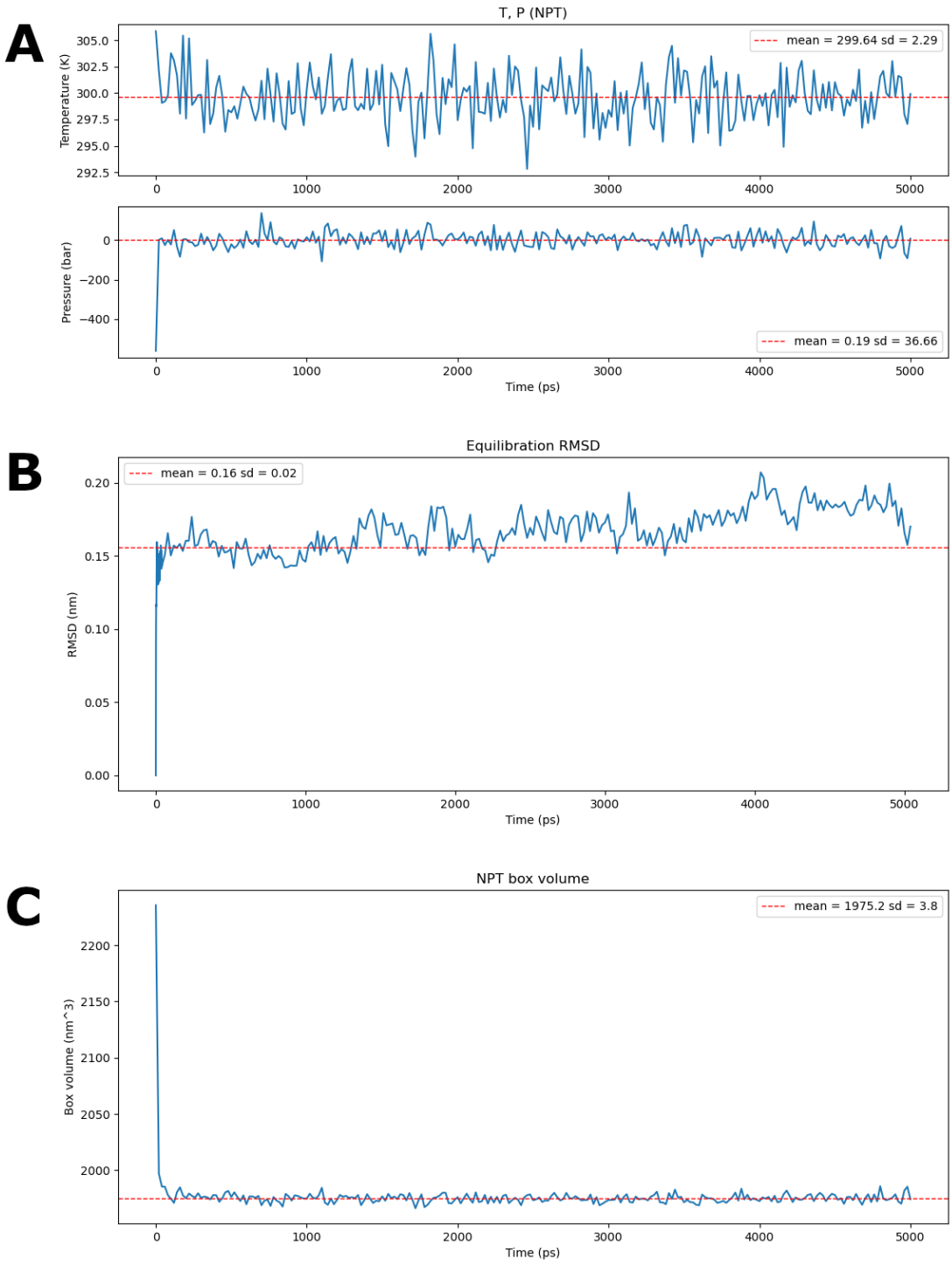


Figure 9. *Tracking of parameters during the equilibration as provided by CHARMM-GUI. The plots show the evolution of temperature and pressure (A), protein backbone RMSD (B) and the simulation box volume (C).*

We systematically tested the MD protocol by modifying a range of simulation parameters, introducing additional equilibration stages, and evaluating alternative barostat algorithms. However, none of these adjustments resolved the observed instabilities. We therefore extended

our testing to different membrane proteins and lipid compositions in order to identify the origin of the issue. These tests revealed that the simulation failures were specifically associated with the presence of cholesterol: systems lacking cholesterol consistently exhibited stable behaviour throughout the production runs. To further isolate the source of the problem, we performed control simulations of a single cholesterol molecule in vacuum, which also resulted in simulation crashes. This behaviour clearly indicated that the instability originated from the cholesterol model itself. We subsequently incorporated a recently published, optimized cholesterol topology [86], but this modification did not resolve the issue. Ultimately, the problem was addressed by reverting to an earlier version of the GROMACS engine, specifically version 2021.5, which restored stable simulation behaviour. The Martini cholesterol topology makes use of virtual sites, which are massless pseudo-atoms that do not possess independent dynamics or experience forces directly. Instead, their positions are calculated from the coordinates of other atoms and are therefore not propagated by the equations of motion. In the case of cholesterol, the use of virtual sites improves molecular stability by enforcing the correct geometry and preventing unphysical distortions. Notably, release notes for later GROMACS versions (≥2022) reported bug fixes related to inaccuracies in the treatment of virtual sites under pressure coupling, which were described as having only minor effects [90]. Nevertheless, based on our observations, we strongly suspect that these inaccuracies were sufficient to destabilize our coarse-grained MD simulations, ultimately leading to the recurrent simulation crashes observed when cholesterol was present.

Given the recurrent instabilities observed during the production runs with Martini 2, we transitioned to the Martini 3 force field, which was explicitly designed to address several of the known limitations of earlier Martini versions. Martini 3 introduces a reparameterized interaction scheme, improved balance between protein–lipid and protein–protein interactions, and a revised treatment of bonded terms, all of which contribute to enhanced numerical stability and physical realism in coarse-grained simulations. In particular, the updated lipid and cholesterol models, together with refined protein backbone representations, substantially reduce the occurrence of excessive forces and constraint violations that can trigger LINCS-related failures. Adopting Martini 3 therefore provided a more robust foundation for an automated MD protocol, as it increased the reproducibility and stability of simulations across different membrane proteins without requiring extensive system-specific tuning.

Once stable simulations could be reliably achieved, we focused on optimizing the MD protocol to better suit the requirements of our automated pipeline. The extensive debugging performed to identify and resolve the simulation crashes informed several modifications to the default CHARMM-GUI input settings, ultimately leading to the final protocol implemented in this work.

First, the two energy minimization stages provided by CHARMM-GUI were consolidated into a single step. This was achieved by reducing the minimization step size from the default 0.01 nm to 0.005 nm and increasing the convergence tolerance (i.e., the maximum force criterion) from the default value of 10 kJ mol^{-1} nm^{-1} to 500 kJ mol^{-1} nm^{-1}. This adjustment reduced the likelihood of minimization failure, at the cost of a potentially higher initial potential energy, which was subsequently alleviated during equilibration.

In addition, we introduced a short NVT equilibration phase (20 ps), which is not included in the default CHARMM-GUI workflow, to ensure proper thermalization of the system at 300 K. The original six NPT equilibration steps were further reduced to three. During the first NPT stage, positional restraints were applied exclusively to the protein backbone beads, while restraints on lipid headgroups were omitted, and the system was equilibrated using a 2 fs time step. In the second NPT stage, all positional restraints were removed and the time step was reduced to 1 fs to prevent abrupt structural rearrangements. In the final equilibration stage, the time step was gradually increased to the production value of 20 fs to ensure the stability of the equilibrated ensemble.

Following the same principles as **Figure 9**, **Figure 10** illustrates the system behaviour during the final equilibration step, showing stable fluctuations in temperature, pressure, and volume, as well as convergence of the protein backbone root-mean-square deviation (RMSD). Following equilibration, the production simulation was successfully performed for 30 μs using the appropriate barostat.

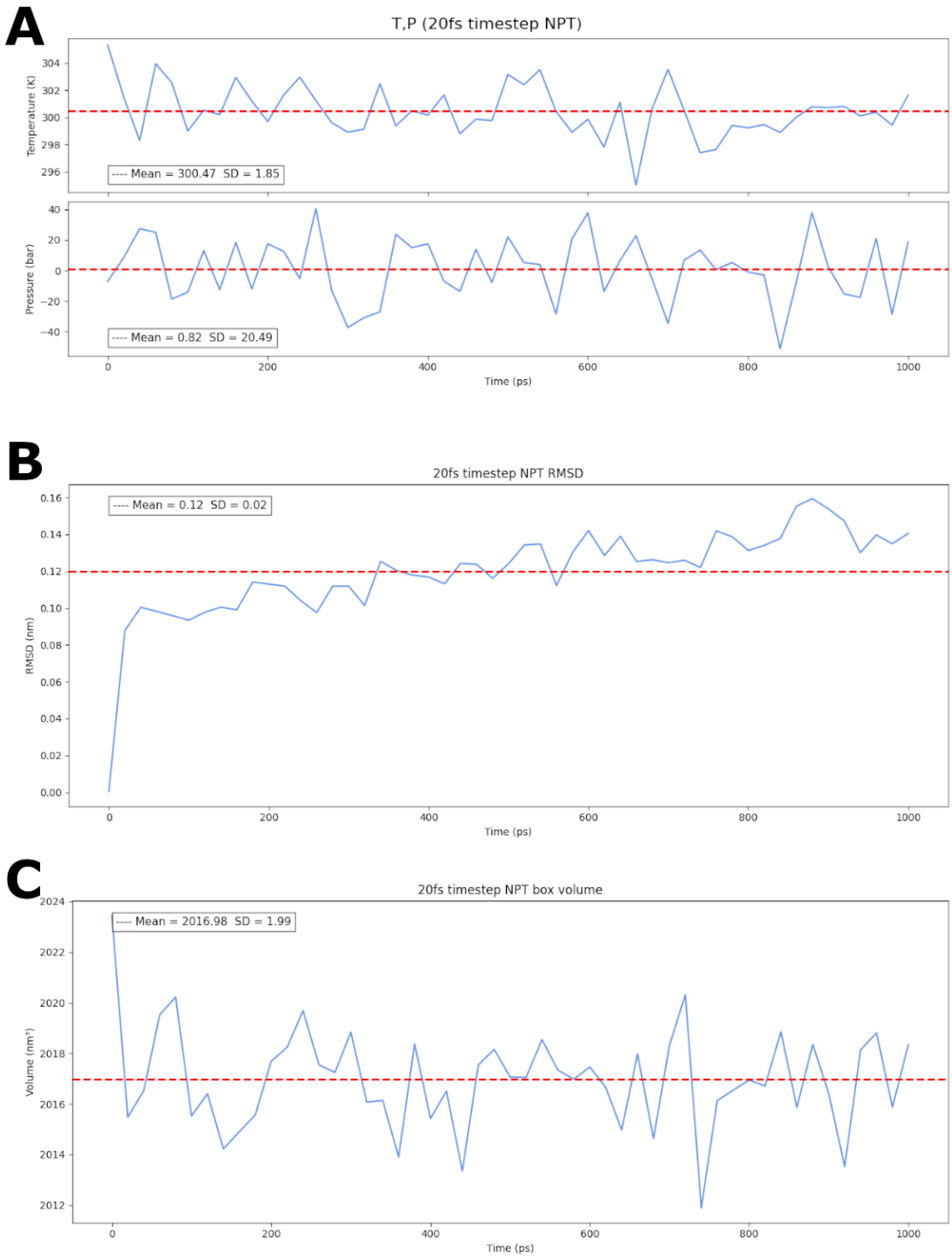


Figure 10. *Tracking of parameters during the equilibration as provided by CHARMM-GUI. The plots show the evolution of temperature and pressure (A), protein backbone RMSD (B) and the simulation box volume (C).*

4.3. System build

Given the overarching objective of maximizing automation, the use of the CHARMM-GUI builder was deemed unsuitable for the final pipeline. In particular, CHARMM-GUI operates as a web-based service that relies on remote server requests, introducing additional execution time and an undesirable dependency on external computational resources. For these reasons, we adopted a fully local, network-independent approach for atomistic-to-coarse-grained (CG) conversion and for the construction of membrane–protein systems, while leaving the preparation of the initial atomistic structure to the user.

Throughout the development of the protocol, we adapted our workflow in response to the continuous evolution of tools within the Martini community. Two key components of this workflow are the *martinize* and *insane.py* scripts [57], which are used for converting atomistic structures into the Martini representation and for building CG protein–membrane systems, respectively. The *martinize* script has evolved from a standalone Python tool distributed via the Martini website [91] to *martinize2*, which is now integrated into the Vermouth framework and provides full support for Martini 3 [56,92].

In contrast, early versions of *insane.py* were not compatible with Python 3, which posed a challenge given that our development environment and external dependencies were exclusively Python 3-based. To address this limitation, we adapted the script to ensure Python 3 compatibility using a dedicated conversion package. During the period in which Martini 2.2 was employed, we also incorporated parameters for the improved cholesterol topology published by Fábián et al. (2023), which were provided directly by the author. At the time of writing, the *insane.py* repository has since been updated to include native Python 3 support and to provide the latest cholesterol topology compatible with Martini 3 [93].

Since *martinize2* is thoroughly documented and discussed in the software documentation [56], it is here only briefly summarised. The basic workflow of *martinize2* is (only options relevant to the way this tool was employed here are discussed):

1) Read input structure: Parse the PDB file, define bonds. The default charged capping to N- and C-terminals is applied in this step.

2) Repair input graph (description of atoms as nodes and bonds as edges): The software will identify missing atoms and residue mutations and try to fix atom names to prepare for mapping.

3) Resolution transformation: Fragment detection and mapping to the corresponding CG beads.

4) Apply links: Defines interactions between residues.

5) Post processing: Elastic network is added.

6) Write output: GROMACS topology file (.itp) and structure (.pdb) are written.

Despite running mostly using the default parameters, we use specific flags for the martinize2 run: "-merge all", makes sure only one protein chain is written, "-p backbone" writes positional restraints that we use during the system equilibration, "-elastic" introduces harmonic restraints between backbone beads around the proper distance (harmonic constant equal to 700 kJ mol^{-1} nm^{-2}), "-residue input" keeps the residue numbering from the input, facilitating the analysis after the simulation, "-maxwarn 100" is introduced to prevent the tool from crashing due to a high number of alerts, which usually come from inconsistent or non-canonical atom naming. Additionally, we make sure to set "-eunit system", as the merging of possible detected protein "fragments" is the last step and affects only to the written output and the elastic network is built on each of the detected fragments. Hence, by setting this option we ensure the elastic network is constructed accounting for the whole system which, in this case only comprises the protein of interest.

Same as with *martinize2*, we describe here the relevant functionality of *insane.py* and its integration into our protocol, as all the details on the tool are disclosed in detail in the original publication [57]. The core concept involves defining a regular grid within the user-specified simulation box, identifying the grid cells occupied by the protein, and filling the remaining space with lipids according to a given composition. In this protocol, we used a box size of 14.5 × 14.5 × 10.5 nm, which is sufficient to avoid artifacts from periodic boundary conditions. The protocol is optimized around GPCRs, and the box size might have to be adjusted for different proteins. As for the membrane composition, we used a binary system of 75% 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) and 25% cholesterol in line with previous studies. In principle, any ratio and/or composition, as long as selected lipids are properly parameterised in the adopted force field, can be used. The script places lipids randomly. Hence, according to best practices in running MD simulations, the protocol creates two instances of the system with different random seeds, ensuring a different starting configuration of the membrane components in the two assembled systems.

4.4. Contact analysis

The trajectories generated by the MD simulations are both long, in terms of the number of frames, and large, in terms of disk usage. To efficiently analyses protein–cholesterol contacts, we devised the following strategy: (i) the first 25,000 frames (corresponding to 500 ns) are discarded to ensure

that the system is fully in the production regime and no longer influenced by the equilibration ensemble; (ii) the remaining trajectory is iterated frame by frame to determine, on a per-residue basis, whether a given residue is in contact with cholesterol. Following on the work of Ferraro[51], we establish a residue is contacting a cholesterol molecule if any of the residue beads fall within a 6Å cut-off distance of any cholesterol bead; (iii) consecutive contacts persisting across multiple sequential frames are identified for each residue, and the longest continuous contact is retained; (iv) residues that never form contacts are excluded from the analysis, as these typically correspond to regions located at the intracellular or extracellular extremities of the protein or within internal cavities; and (v) statistical outliers are identified based on the distribution of the maximum contact occupancy time, defined as the longest continuous contact duration (t_{max}) for each residue.

Trajectory iteration was performed using MDAnalysis [62,63], the Python library employed for simulation analysis. This operation is computationally demanding: a full pass through a trajectory requires, on average, approximately 3.5 hours on local workstations. Consequently, to optimize performance, it was essential to avoid repeating this operation multiple times. To this end, all relevant contact information is extracted during a single traversal of the trajectory and stored in a plain-text output file. Although this file occupies a substantial amount of disk space (approximately 100 MB), it enables significantly faster downstream analyses.

The analysis script is structured around two dictionaries that use residue identifiers as keys. One dictionary tracks contact duration by maintaining a counter that increments for residues that remain in contact with cholesterol across consecutive frames. The second dictionary records the corresponding frame intervals, storing pairs of starting and ending frame indices for each contact event. Together, these data structures allow each continuous contact interval to be associated with its duration in frames. Using this information, the script generates an output file that records all detected contact events for each residue. For clarity, a mock example illustrating the structure of both dictionaries and a representative line of the output file is provided below.

To illustrate the logic of the contact-tracking algorithm, we consider a representative example in which residue LEU104 is monitored over the first 10 trajectory frames. In this hypothetical scenario, the residue is in contact with cholesterol in frames 3, 4, and 5, and again in frames 8 and 9, resulting in two distinct contact events.

Prior to the analysis of the first frame, the system is initialized in a state with no active contacts, and the frame-tracking dictionary is empty:

res_contacts_dict[104] = [0]

res_timeframe_dict[104] = [[]]

When frame 3 is processed, a new contact is detected, marking the beginning of the first interaction:

res_contacts_dict[104] = [1]

res_timeframe_dict[104] = [[3]]

Frames 4 and 5 correspond to the continuation of this contact. Since the residue remains in contact with cholesterol across consecutive frames, the contact counter is incremented for each frame (+2 in total), while the frame dictionary retains the initial frame of the interaction:

res_contacts_dict[104] = [3]

res_timeframe_dict[104] = [[3]]

At frame 6, the residue is no longer in contact with cholesterol, whereas it was in contact in the previous frame. This condition signals the termination of the first contact event. At this point, a new counter initialized to zero is appended to the contacts dictionary in preparation for the next event, and the current frame index is stored as the ending boundary of the contact (defined as the first frame after the contact has ended). An empty entry is also added to the frame dictionary to accommodate future contacts:

res_contacts_dict[104] = [3, 0]

res_timeframe_dict[104] = [[3, 6], []]

The same procedure is applied to the second contact event. After processing frame 9, the dictionaries take the following form:

res_contacts_dict[104] = [3, 2]

res_timeframe_dict[104] = [[3, 6], [8]]

Following frame 10, which marks the end of the second interaction, the final state of the dictionaries is:

res_contacts_dict[104] = [3, 2, 0]

res_timeframe_dict[104] = [[3, 6], [8, 10], []]

The trailing zero values and empty elements are removed during a cleanup step. The script then writes the complete set of contact events for each residue to an output file, with each line containing a list of contact intervals and their corresponding durations. For residue 104 in this example, the output line would be:

104, LEU, ['[3,3,6]', '[2,8,10]']

where each entry encodes the contact duration (in frames), the first frame in contact, and the first frame after contact, respectively.

Subsequent analysis is performed using the data stored in this output file. For each residue, the longest contact duration, expressed as the number of consecutive frames in contact with cholesterol, is identified and defined as the maximum occupancy time (t_{max}). Residues that never form direct contacts with cholesterol over the entire simulation are excluded from this analysis, as they are typically located in regions inaccessible to cholesterol and would otherwise bias the contact-time distribution.

The final output is a comma-separated values (.csv) file reporting t_{max} for each residue, using the format:

ResID, ResName, Longest contact (frames), Longest contact (ps), First frame in contact, First frame after contact

For the example discussed above, the corresponding entry would be:

104, LEU, 3, 60, 3, 6

As shown in **Figure 11A**, the use of extended simulations combined with a 20 fs time step results in contact-time distributions that are strongly right-skewed. In these distributions, the majority of residues exhibit short-lived contacts, which predominantly arise from the intrinsic lateral diffusion of cholesterol molecules within the membrane and their transient interactions with the protein surface. In contrast, functionally relevant contact events are characterized by substantially longer lifetimes. These persistent interactions become increasingly apparent as the lower threshold applied to t_{max} is progressively raised, as illustrated in **Figure 11B**.

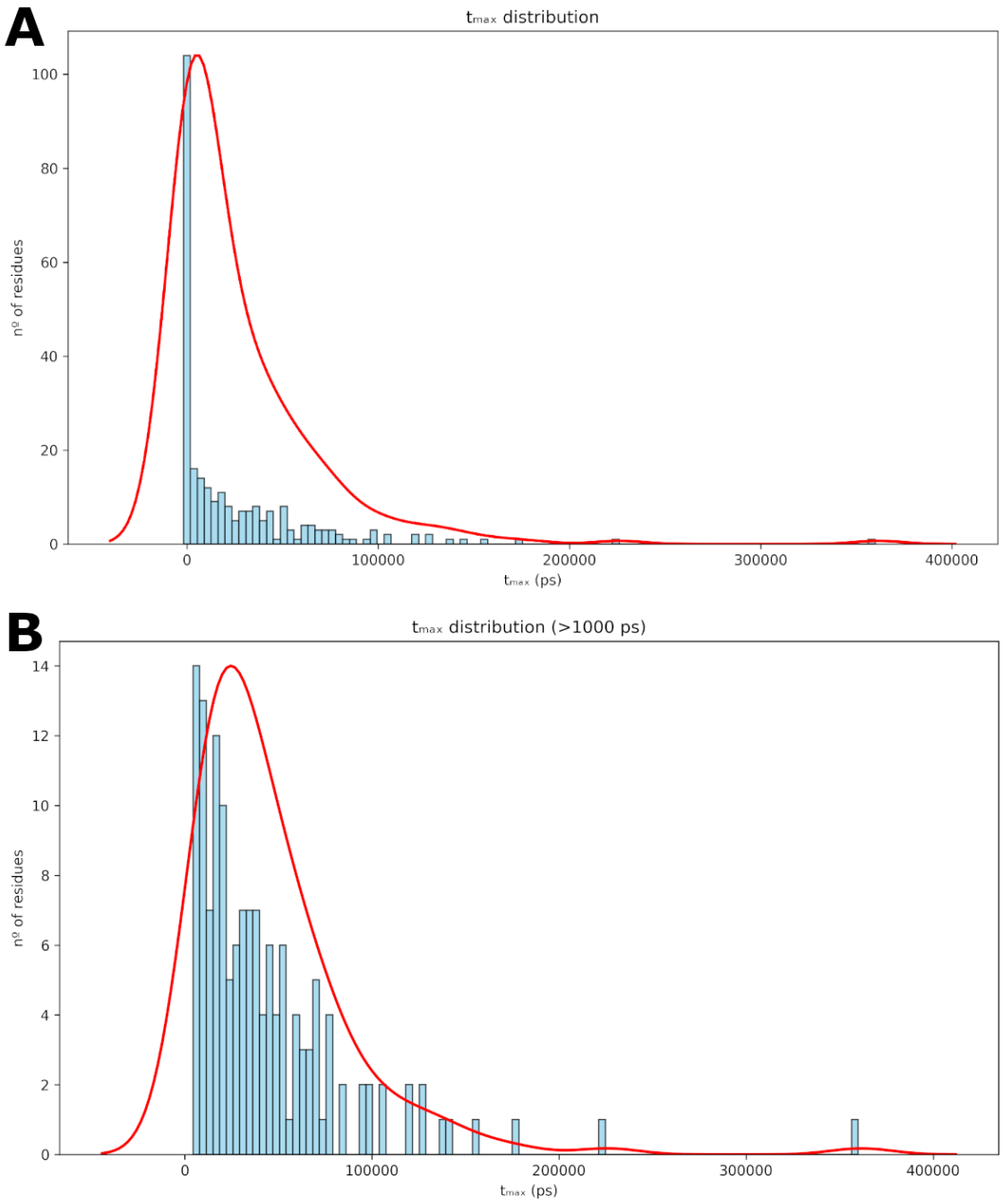


Figure 11. *Distribution of the t_{max} values obtained in one of the replicas of one of the CB1 simulations (PDB id: 7FEE). (A) Full distribution and approximation. (B) Showcase of the same distribution after removing tmax values lower than 1 nanosecond. As expected, most of the residues present a maximum occupancy time corresponding to random contacting events rather than meaningful interactions. This highlights the requirement of a method to distinguish these two types of events. While in some cases it becomes obvious that a contact is relevant (as in the two rightmost values, easily appreciated in the bottom distribution), the closer it gets to the left the less evident the relevant values become.*

Given the non-normal nature of the data distribution, we adopted the same statistical approach described by Ferraro and colleagues [51], in which the sorted data are divided into quartiles to

characterize the spread of the distribution. The interquartile range (IQR), defined as the difference between the third quartile (75th percentile) and the first quartile (25th percentile), encompasses the central 50% of the data and provides a robust, distribution-independent measure of variability. Values that lie substantially outside this range can therefore be reliably identified as statistical outliers. Therefore, following:

$$T = 1.5 \times IQR + 3\text{rd quartile}$$

Where T is the established threshold, t_{max} values above it are considered a statistical outlier and consequently the corresponding residue labelled as relevant.

In order to assess the effect of the simulation time on the residues detected as outliers, we studied their evolution by calculating the statistical outliers on different timestamps (1, 5, 10, 15, 20, 25 and 30 microseconds) on the two replicas of our CB_1 simulations. The results of this study are reported in **Figure 12**. From this test, it is clear that the most relevant and persistent contacts already appear around 10-15 μs. However, other contacts still considered as relevant appear in later stages of the simulation. Additionally, the starting configuration of the cholesterol distribution randomly created by the *insane.py* tool [57] appears to have an effect, making some contacts appear later in the second replica (e.g. residue 201 being labelled an outlier at 15 μs in the first replica and not appearing until 25 μs in the second). Consequently, considering the late-appearing outliers and the potential effect of the starting configuration, which gets mitigated the longer the simulation goes on, we decided to not shorten the simulation time and keep the originally intended 30 μs per replica.

Replica 1								
ResID	Timestamps (ns)							Longest contact (ps)
	1	5	10	15	20	25	30	
165	■	■	■	■	■	■	■	867100
241	■	■	■	■	■	■	■	663000
202		■	■	■	■	■	■	527420
245	■	■	■	■	■	■	■	499240
172	■	■	■	■	■	■	■	344700
198	■	■	■	■	■	■	■	274380
161		■	■	■	■	■	■	234120
201				■	■	■	■	213240
287	■		■	■	■	■	■	207480
240				■	■	■	■	202340
291	■		■	■	■	■	■	197760
353	■		■	■	■	■	■	197500
191	■	■	■	■	■	■	■	190060
169		■	■	■	■	■	■	137320
243	■	■		■	■	■	■	133140
282		■	■	■	■	■	■	131480
236					■	■	■	120340
168		■	■	■	■	■	■	120160
209					■	■	■	117780
360			■	■	■	■	■	115580
212					■	■	■	106180
350					■	■	■	102020
195	■					■	■	101340
232					■	■	■	97800
361			■	■	■	■	■	96720
139	■				■	■		93280
192				■	■	■		92300
208				■	■			92160
247				■				89640
124	■	■						73520
288	■							63000
354	■							55100
292	■							47180
142	■							35600
357	■							27720
135	■							27160

Replica 2								
ResID	Timestamps (ns)							Longest contact (ps)
	1	5	10	15	20	25	30	
241	■	■	■	■	■	■	■	727340
165		■	■	■	■	■	■	609020
245	■	■	■	■	■	■	■	410000
202		■	■	■	■	■	■	377980
161		■	■	■	■	■	■	299760
198		■	■	■	■	■	■	288700
240				■	■	■	■	267000
201						■	■	152200
191			■	■	■	■	■	149200
243	■	■	■	■	■	■	■	144040
282	■	■	■	■	■	■	■	129660
172	■	■	■	■	■	■	■	125240
142				■	■	■	■	118780
208	■					■	■	114120
291	■			■	■			99260
199				■	■			96720
195				■	■			96700
162				■	■			93120
351			■					69100
249			■					64940
353			■					63020
192		■	■					62880
278		■						60680
350	■	■						58920
255		■						57120
247	■							35320
139	■							31860
129	■							29340
381	■							26980
168	■							24140
135	■							23680

Figure 12. *Evolution of detected outliers over time in both CB_1 replicas. While a similar tendency is followed for both distributions, the timelines at which outliers appear do not exactly match, emphasizing the effect of the starting lipid distribution and the need of multiple replicas for better sampling.*

4.5. Density analysis

The calculation of the cholesterol spatial density around the receptor constitutes a complementary approach to the protein–cholesterol contact analysis presented above. In keeping with our objective of minimizing the number of external tools required by the pipeline, we employed the *spatial* module from GROMACS, which is already used for the MD simulations. This tool requires a reference structure (the protein, in our case) and a target molecule for which the spatial density is computed (cholesterol). The output is a three-dimensional density map that represents the probability of finding cholesterol atoms in the vicinity of the receptor. When a trajectory is provided, the *spatial* tool uses the first frame to estimate the dimensions of the simulation box and subdivides this volume into a three-dimensional grid of voxels. Each voxel stores a numerical value corresponding to the average density, and can be considered the three-dimensional analogue

of a two-dimensional pixel. The trajectory is then processed frame by frame to determine how frequently cholesterol atoms occupy each voxel, with regions of higher occupancy yielding higher density values.

During repeated executions of the protocol, we observed that this analysis step occasionally failed due to memory allocation errors. These failures were traced to situations in which the *spatial* tool attempted to write density values to voxels that had not been allocated in memory. As noted above, the voxel grid is defined based solely on the dimensions of the simulation box in the first trajectory frame. In some cases, subsequent frames contain atoms that move beyond or too close to the boundaries of this predefined grid, resulting in an insufficient number of allocated voxels. Under these conditions, the program attempts to access memory locations outside the allocated array, leading to runtime errors. Although the *spatial* tool provides a small default safety margin by allocating 16 additional voxels in each spatial dimension, this proved to be insufficient for our simulations. Incremental increases of this margin had no observable effect on the occurrence of memory errors. Increasing the allocation to 300 extra voxels per dimension resolved the issue in the majority of cases; however, sporadic failures still occurred. Only when the margin was increased to 400 extra voxels in each direction were all memory-related errors consistently eliminated across all simulations. Allocating additional voxels directly translates into increased memory consumption, and therefore higher values must be chosen judiciously, taking into account the computational resources available on the target system. Each voxel stores a floating-point value corresponding to the local density. In the most common scenario, where GROMACS is compiled in single precision, these values are represented as 32-bit floating-point numbers, corresponding to 4 bytes per voxel. Consequently, allocating an additional 400 voxels in each spatial dimension results in an increase of approximately:

$$\text{Extra memory} = 4 \text{ bytes} \times 400^3 \text{ voxels} = 256{,}000{,}000 \text{ bytes} \approx 244 \text{ MB}$$

This memory overhead is modest and well within the capabilities of systems already suitable for running MD simulations. Even in the case of a double-precision GROMACS build, where each voxel occupies 8 bytes (64-bit floating-point representation), the additional memory requirement would be approximately 488 MB. Such memory usage remains manageable on most modern computing systems, particularly when the analysis is executed on a machine dedicated to this task.

The next challenge arose from the nature of the output generated by the *spatial* analysis. This step produces a Gaussian cube file, which encodes a three-dimensional grid of voxel positions together with their associated density values. To enable the automated identification of meaningful cholesterol density regions, it was necessary to determine which voxels exhibited the highest density values and to relate these regions to specific residues in the reference protein.

A straightforward strategy would be to compute distances between voxels with high cholesterol probability and residues identified as outliers in the contact analysis. However, volumetric data cannot be directly used for distance-based calculations, as voxels do not represent physical entities in Cartesian space and lack intrinsic properties such as mass or radius, unlike atoms. To overcome this limitation, we converted selected regions of the volumetric density map into physical representations that coexist with the molecular system and allow for distance measurements.

To this end, we developed a custom script to parse the Gaussian cube file and convert voxel centers into pseudo-atoms (dummy particles). The cube file stores voxel coordinates as a nested array of Cartesian positions, with the Z dimension varying fastest, followed by Y and then X. By correctly traversing this data structure, a pseudo-atom could be placed at the geometric center of each voxel, thereby providing a physical proxy for that region of space. During initial tests, however, the resulting particle-based density maps appeared distorted, exhibiting an artificial stretching along one spatial dimension. This discrepancy was traced to the units used in the Gaussian cube format: as reported by Bourke, when voxel values are positive, spatial dimensions are expressed in Bohr rather than in ångströms [94]. This was the case for our data, and therefore both voxel positions and dimensions were converted using the appropriate factor (1 Bohr = 0.529177249 Å) to accurately reproduce the original density map geometry. Although it was technically feasible to convert the entire density map into pseudo-atoms, doing so would not serve the purpose of identifying the most relevant cholesterol density regions, as such an approach would obscure any meaningful signal with diffuse background noise. Consequently, a filtering strategy was applied during traversal of the cube file to retain only the most significant voxels for conversion. Because cholesterol density distributions are strongly right-skewed and dominated by low-probability values arising from transient diffusive events, the use of an absolute density cutoff was found to be unsuitable and highly system-dependent. Instead, a percentile-based criterion was adopted to ensure robustness and scalability across different systems and simulation conditions. Specifically, only the voxels corresponding to the top 0.01% of density values were retained. This threshold isolates the extreme tail of the density distribution and effectively selects regions whose occupancy is orders of magnitude higher than the background density expected from diffusion alone. In practice, this choice consistently preserved well-defined, spatially localized density maxima corresponding to stable cholesterol interaction sites, while excluding diffuse low-density regions that do not convey meaningful spatial information. The selected voxels were subsequently converted into pseudo-atoms and written to a coordinate file in XYZ format, enabling downstream distance-based analyses with the protein residues.

4.6. Establishing a pocket-based baseline for cholesterol binding at the protein–membrane interface of class A GPCRs

To establish a pocket-based baseline, we considered an initial subset of 30 simulated systems. Across these systems, 97 cholesterol molecules were co-crystallized. For each cholesterol molecule, we defined a reference "ideal pocket" as the set of residues in the corresponding crystal structure containing at least one non-hydrogen atom within 5 Å of the cholesterol molecule. It is important to note that this definition is intentionally permissive, as it considers spatial proximity only and does not require the presence of a well-defined surface concavity. This distinction is critical, since cholesterol frequently associates with relatively flat protein surfaces through transient or collisional interactions, which do not necessarily correspond to classical pocket-like binding sites.

Conceptually, this highlights two complementary but non-equivalent strategies for identifying cholesterol-interacting regions. A hot-spot–based approach identifies regions of the protein surface that repeatedly engage cholesterol, manifested as outliers and persistent density in molecular dynamics simulations, without any explicit assessment of pocket geometry or druggability. In contrast, a pocket-based strategy explicitly focuses on detecting geometric cavities on the protein surface. In this work, NanoShaper [66] was used for this purpose, although the overall approach is agnostic to the specific pocket-detection algorithm employed. This strategy yields a set of *bona fide* pockets but does not, by itself, indicate whether these cavities are relevant for cholesterol binding. Importantly, if cholesterol binds to a flat surface region, no corresponding pocket will be detected, and such an interaction cannot be recovered within a pocket-based framework. Consequently, when establishing a baseline using crystallographic cholesterol positions, the objective is not to assess NanoShaper's ability to reproduce all cholesterol contacts, but rather to estimate how many co-crystallized cholesterol molecules occupy genuine surface cavities. When cholesterol binds within a concavity, NanoShaper is expected to detect the corresponding pocket, which, if relevant, should be enriched among the top-ranked sites identified by our scoring scheme. While the possibility that a true pocket is missed by the detection algorithm cannot be entirely excluded, such cases are expected to be rare and to have a negligible impact on the present analysis.

Leveraging the flexibility of NanoShaper parameterization, we performed preliminary tests to optimize pocket detection for cholesterol, whose binding sites are often extended and relatively shallow. A probe radius of 1.4 Å for the small probe and 6 Å for the large probe provided the most suitable parameter combination and was therefore selected, while all other parameters were

kept at their default values. Using these settings, NanoShaper was applied to all structures in the benchmark, yielding a total of 940 detected pockets. Summary statistics are reported in **Table 8**.

***Table 8**. NanoShaper detected pockets.*

Total nº of pockets	**940**
Mean N. of pockets per system	31.33
Standard deviation	4.47
Max N. of pockets in a system	40
Min N. of pockets in a system	15
Range	25
Median	32
Mode	29, 32, 34

Additionally, we computed the volumes of the detected pockets and associated statistical parameters to assess the variability in pocket size across the dataset. These results are summarized in **Table 9**. In some cases, we observed pockets with extremely large volumes, typically corresponding to internal cavities within the helical bundle or to solvent-exposed clefts in contact with the extracellular or intracellular domains of the receptors. Since this study is focused on interaction at the protein-membrane interface, these pockets should not be particularly relevant for our analysis. Notably, in certain instances, individual pockets extended across nearly the entire transmembrane region, following the surface contours of multiple transmembrane helices.

***Table 9.** Volume análisis of NanoShaper detected pockets.*

Pocket Volume	**($Å^3$)**
Mean	236.10
Median	87.62
Mode	25.25
Standard Deviation	396.00

Variance	156819.71
Range	3569.75

By comparing NS-detected pockets with crystallographic cholesterol coordinates, we identified 74 predicted pockets that overlapped with ideal cholesterol-binding sites. An overlap was defined as the presence of at least 50% of residues in common between a predicted pocket and a reference site. In some instances, multiple cholesterol molecules in the crystal structure were in close proximity, leading to the alignment of their ideal binding sites with the same NS-detected pocket; we refer to these as shared pockets. In contrast, unique pockets are defined as those associated with a single cholesterol molecule. Overall, we identified 58 unique pockets and 16 shared ones. As summarized in **Table 10**, this analysis indicates that a substantial fraction of co-crystallized cholesterol molecules occupies well-defined surface cavities.

Table 10. *Cholesterol ideal pockets and NanoShaper performance on the benchmark.*

id	Receptor	Cholesterol Molecules	Ideal pockets (5Å cut-off)	NS Pockets	Overlapping Pockets
7E2Y	Serotonin 1A	4	4	31	2
7E32	Serotonin 1D	1	1	31	1
7WC4	Serotonin 2A	3	3	28	1
6DRZ	Serotonin 2B	1	1	28	1
8DPF	Serotonin 2C	1	1	27	1
8CU7	Adenosine 2AR	3	3	29	1
8DU3	Adenosine 2AR	3	3	24	1
8HDP	Adenosine 2BR	4	4	34	0
8HDO	Adenosine 2BR	7	7	35	3
7BU6	β_1AR	1	1	15	0

6PS4	β_2AR	1	1	30	1
7FEE	CB_1R	2	2	29	1
6PT0	CB_2R	4	4	37	1
7VL8	CCR1	1	1	33	1
7VLA	CCR1	1	1	40	1
7X2C	Dopamine D1R	1	1	24	1
7JVQ	Dopamine D1R	5	5	33	0
7WQ3	Galanin 1R	5	5	35	2
7WQ4	Galanin 2R	2	2	32	1
7W2Z	Ghrelin R	2	2	30	2
7F9Y	Ghrelin R	2	2	36	1
7VGY	Melatonin 1AR	1	1	36	1
7VGZ	Melatonin 1AR	1	1	32	1
6PT2	Opioid-δ R	1	1	29	0
6B73	Opioid-κ R	1	1	32	0
8EF6	Opioid-μ R	5	5	32	4
7YAE	Somatostatin 2R	2	2	34	1
7Y27	Somatostatin 2R	1	1	34	1
7XW6	Thyrotropin R	14	13	34	3
7XW5	Thyrotropin R	17	16	29	8

In the 11 structures containing a single cholesterol molecule, a corresponding pocket was identified in all cases. Among the 15 structures containing multiple bound cholesterol molecules,

overlapping pockets were detected for several, but not all, cholesterol positions. In a number of instances, no pocket was identified because cholesterol was bound to an essentially flat protein surface. While such regions may still represent genuine interaction hot spots, the absence of a geometrically defined cavity precludes their detection within a pocket-based framework.

Notably, in four structures no overlapping pocket was identified for any of the bound cholesterol molecules, underscoring that cholesterol binding at the protein–membrane interface frequently occurs on flat or shallow surfaces rather than within classical pockets.

4.7. Performance of the MD simulations

The core of the present method consists of coarse-grained molecular dynamics (CG-MD) simulations, which represent the most computationally demanding component of the protocol. Although the workflow is largely automated and requires minimal user intervention, it assumes access to appropriate hardware resources and a properly configured software environment. From its inception, the protocol was designed to exploit state-of-the-art computational tools, including MD acceleration through graphics processing units (GPUs).

We extensively tested our CG-MD simulations using recent versions of GROMACS (v2021–v2025), which support GPU acceleration for NVIDIA™ graphics cards via CUDA, a parallel computing platform [95]. We therefore recommend the use of GPU-enabled architectures, which are now commonly available in both consumer-grade workstations and high-performance computing (HPC) environments.

Despite the availability of suitable hardware, certain features of the simulated systems can limit achievable performance. In particular, the Martini 3 topology for cholesterol employs virtual sites - massless interaction particles introduced to accurately represent the sterol ring and maintain the correct molecular geometry [44,72] .While most force calculations can be efficiently offloaded to the GPU, the presence of virtual sites prevents the offloading of the integration of positions and velocities (the “update” step) to the GPU, as this operation is not currently supported. Consequently, this step must be executed on the central processing unit (CPU), making overall simulation performance sensitive to CPU–GPU communication efficiency and, by extension, to the underlying hardware architecture.

As a result, the CPU resources allocated to the simulation - particularly the number of cores used - have a substantial impact on performance. We therefore recommend that, prior to running the full protocol, users characterize the performance of their computing environment by benchmarking CG-MD simulations with varying numbers of CPU cores using a representative

system. An example of such a benchmark performed on one of our local workstations is shown in **Figure 13**. As illustrated, performance generally improves with increasing core count up to a system-dependent optimum, beyond which additional cores provide diminishing returns. Identifying this "sweet spot" allows users to select the number of CPU cores that yields optimal performance for production MD runs.

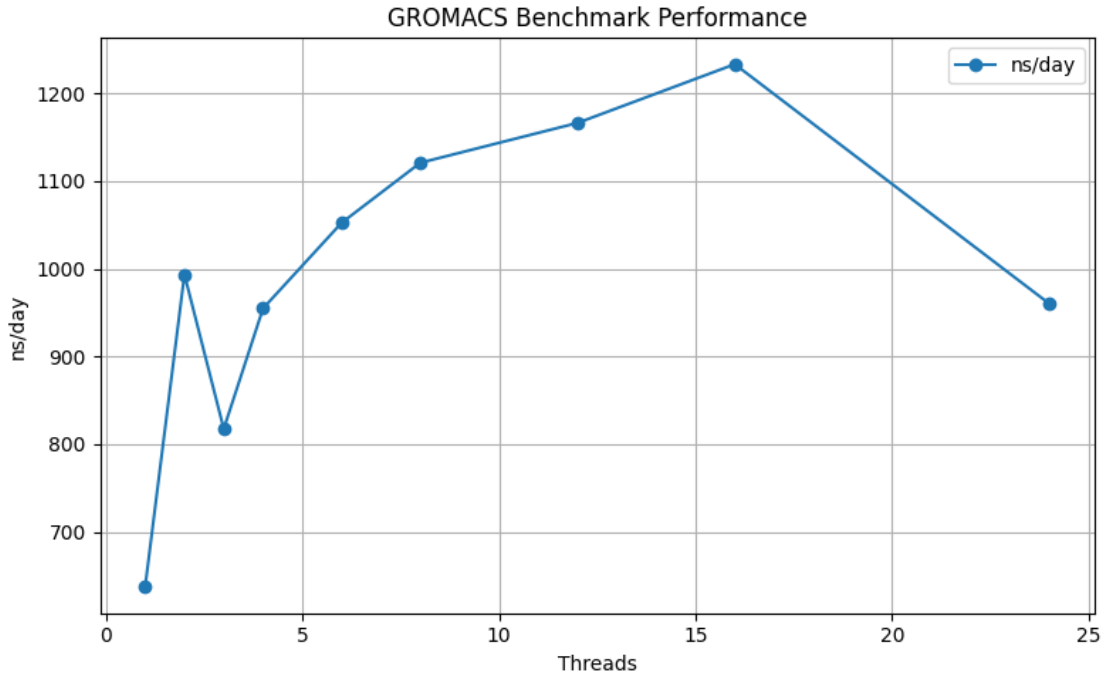


Figure 13. *Benchmark performance in local workstation equipped with an RTX 3080 NVIDIA™ GPU with a 24 cores CPU. This benchmark was obtained through a short 500 ps run using the restrained NPT step from the equilibration protocol.*

To further maximize simulation throughput and facilitate the application of the protocol to dozens of protein–membrane systems, we implemented parallel execution using the NVIDIA™ Multi-Process Service (MPS), an alternative CUDA execution mode that enables controlled sharing of GPU resources among multiple processes [96]. When independent MD simulations are launched concurrently without such coordination, each process competes for GPU resources, often resulting in substantial performance degradation due to frequent context switching and inefficient resource allocation. MPS mitigates this issue by enabling a more orderly and efficient distribution of GPU resources across concurrent tasks. Nevertheless, the limitations discussed above regarding CPU–GPU communication remain applicable. In particular, the presence of virtual sites requires frequent synchronization between CPU and GPU, and assigning multiple CPU cores to each simulation increases the communication overhead, thereby reducing overall efficiency. Following extensive benchmarking across multiple configurations, we found that the most

efficient parallel execution strategy consisted of running each simulation using a single CPU core with MPS enabled.

Table 11 summarizes the most relevant benchmark results obtained on one of our local workstations. As a baseline, we first measured the performance of a single simulation using the workstation's optimal configuration of 20 CPU cores. We then performed 24 parallel simulations, each assigned a single CPU core, and computed the aggregated throughput as the total number of nanoseconds simulated per day across all runs. Although the wall-clock time required for each individual simulation increased under this parallel setup, the combined throughput was substantially higher than that of the single-run baseline. To confirm that allocating additional cores per simulation was detrimental, we repeated the benchmark using two CPU cores per simulation, which resulted in an almost twofold reduction in total throughput. We further tested different numbers of concurrent simulations, up to the 30 runs reported in **Table 11**, and observed that, provided each simulation was restricted to a single CPU core, per-run performance remained largely unaffected. This indicates that, as long as sufficient CPU cores and GPU memory are available, additional parallel simulations can be launched without compromising overall efficiency.

Table 11. *MPS performance benchmark using several test scenarios in a local workstation equipped with an RTX 5080 NVIDIA™ GPU and a 120 cores CPU.*

Nº parallel runs	**Nº cores**	**Wall t (s)**	**Total ns/day**
1	20	10	948,45
24	1	38	6.398,79
24	2	70	3.590,37
30	1	42	7.512,91

Nevertheless, when running multiple parallel runs, one must account for the size of the data produced. For the sake of consistency, easy back-track in case of unexpected behaviour, and good access to all the data produced, we decided to keep intermediate trajectories and other useful files in the final output of the protocol. Considering a default run of two 30 μs replicas, the amount of data produced can add up to 790 GB of disk memory. From that, around 780 GB belong to the two MD trajectories plus their respective iterations of *trjconv* to obtain the density map, each accounting for around 130 GB. The remainder of the size is occupied by the volumetric density maps of each replica generated by the *spatial* tool (around 3 GB each) and the rest of the structures, topologies, intermediate and output files. For easier visualization, **Table 12**

summarizes the heaviest, most disk-consuming files for one default run of the protocol. As already mentioned in previous sections of this text, we systematically retained every intermediate file produced for debugging and development purposes. However, future releases will include the option of automatically delete them.

Table 12. *Heavy data produced by a single, default run of the protocol and its approximate size in disk per replica (column 2) and per run (column 3). The expected size of the directory produced by these types of runs is approximately 786 GB.*

File	Disk usage (GB)	Total disk usage (2 replicas)
MD trajectory	130	260
MD trajectory centered + PBC conditions	130	260
MD trajectory no translation + no rotation	130	260
Gaussian cube	3	6

4.8. Introduction of ligands in the protocol

As discussed in previous chapters of this manuscript, our protocol was designed to identify specifically cholesterol hot spots around membrane proteins. However, and as we presented in Chapter 3, with the appropriate adjustments its use can be extended to other small molecules, such as allosteric ligands that bind in the protein-membrane interface.

When the Martini 3 topology for a ligand is already described and available the process is straightforward and requires minor coding knowledge. The available Python package ships with the topology for Martini 3 cholesterol and will automatically use it. Consequently, the use of cholesterol and its Martini name “CHOL” is hard-coded all along the protocol functions. By identifying all these mentions with those of the desired ligand and providing valid file paths to the topology files, the protocol will promptly process the system correctly.

An important caveat to this workflow is the use of insane.py to build the system [57]. While the recent versions come with cholesterol as one of the available membrane components, ligands are not considered by default. An additional entry with instructions for how to properly place the ligand to start as part of the membrane are necessary. This would most likely require significant knowledge of the topology of the ligand, its beads and bonds and several rounds of trial and error to achieve an appropriate starting positioning of the beads.

However, the use of ligands with non-available topologies would require to undergo the full process of parametrization on top of the steps mentioned above.

Parametrization is a complex and usually long process deeply discussed in the tutorials section Martini CG website. However, and for the sake of completeness, we provide a summary in the following lines. There are currently three main approaches when it comes to parametrizing a small molecule, we describe them here from less to most automatic.

The classic method involves a fully manual approach, where the user defines every aspect of the final CG model, including the spatial mapping scheme, the bead type choosing or bonded parameters. Atomistic simulations of the ligand are used to generate a reference distribution of bonds, angles and dihedrals, which are fitted to represent the same behaviour in the CG model. This is the most thorough and accurate method, which ensures the most consistency with the atomistic model. However, it is also computationally demanding and requires significant knowledge and expertise.

The semi-automatic method uses the computational tool Fast_Forward [97] to guess the best parameters for a given CG topology from a starting input topology and a mapped atomistic trajectory. While this method still requires atomistic simulations from which to extract data, it can save a substantial amount of time and still allows for a high degree of control in the parameters, as the user is still responsible for defining the mapping scheme and bead types.

The third and most automated approach is the use of Bartender, a computational tool aimed to streamline the process by obtaining the bonded parameters from quantum mechanical data (no atomistic MD simulations required) [98]. The user is still in charge of mapping and bead choosing, but this method offers less margin for fine-tuning, given the automatic nature of extracting the bonded parameters and the removal of the user manual iterations.

Chapter 5. Cholesterol Interaction Landscape and Allosteric Pocket Discovery in Class A GPCRs

5.1. Introduction

CHAMP, discussed in detail in the previous chapter, proved to be a reliable method to detect relevant cholesterol-contacting protein regions at the protein-membrane interface. As demonstrated in the previous examples using the Cannabinoid receptor 1 (CB_1) and the serotonin transporter, some of the areas detected correspond to previously characterized cholesterol binding sites, while others remain currently unreported and present opportunities to find novel interactions.

In previous chapters of the thesis, we already discussed how, in GPCRs, consensus motifs and other sequence-based patterns fail to provide a reliable method to predict cholesterol binding [18]. Furthermore, as illustrated in **Figure 14**, inspection of available crystal structures shows that cholesterol frequently occupies multiple adjacent positions. While some structures capture cholesterol in bona fide binding sites, others were obtained from membranes with high cholesterol content, under which conditions excess cholesterol molecules partition into the protein's hydrophobic surface and stack against one another through largely transient and collisional contacts. Therefore, cholesterol molecules appear clustered in the vicinity of the protein, obscuring the distinction between specific, functionally relevant interactions and fortuitous associations, even when high-resolution structural data are available.

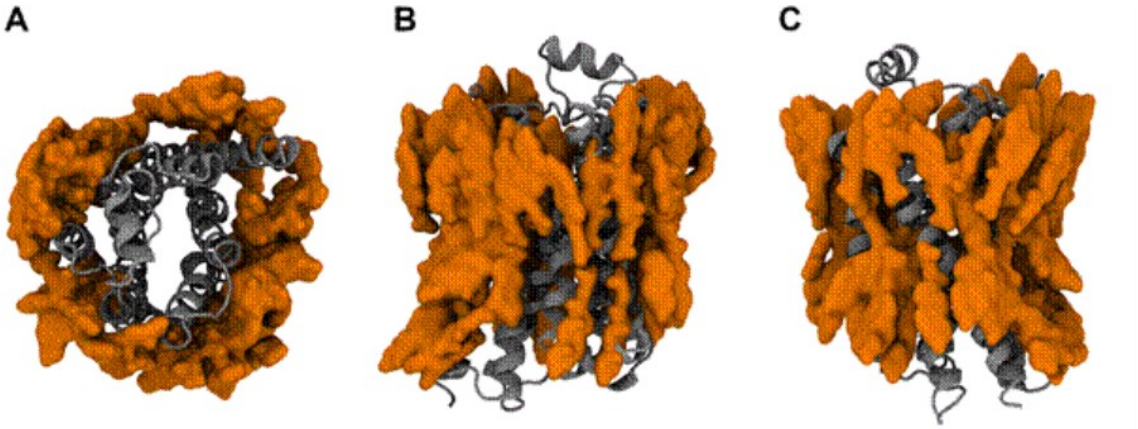


Figure 14. *Spatial distribution of co-crystallized cholesterol molecules (orange, surface representation) across a representative benchmark of 30 GPCR crystal structures (grey, cartoon representation). Panels show (A) top, (B) front, and (C) back views. For clarity, only one of the aligned receptor structures is displayed.*

5.2. Systematic application of CHAMP on a representative benchmark of class A GPCR structures

With this challenge in mind, we addressed the problem by assembling a representative benchmark of available structures, which was then used to systematically apply the protocol we developed. As discussed in the Introduction, comparative analyses across receptors can provide critical insights into the mechanisms of allosteric binding and the resulting protein behaviour. Given that GPCRs - and in particular the class A subfamily - constitute one of the largest and most pharmacologically relevant protein families, and that cholesterol is among the most prevalent membrane components interacting with these receptors, we therefore designed our benchmark to focus on this system [99,100].

At the time the dataset was assembled, 570 class A GPCR structures were publicly available, of which 134 contained at least one co-crystallized cholesterol molecule. Given that the objective of this study is a comparative analysis, we restricted selection to receptor subfamilies represented by at least two structures, or alternatively by multiple structures from closely related subtypes (e.g., two or more CB_1 structures, or a combination of CB_1 and CB_2). The final benchmark comprises 56 structures, whose relevant properties are summarized in **Table 13**.

Table 13. *Representative benchmark of class A GPCR structures used in this study. Structures were selected based on the following criteria: (i) classification as class A GPCRs, (ii) presence of one or more co-crystallized cholesterol molecules, and (iii) availability of multiple structures representing the same receptor or of closely related subtype.*

Receptor	Method	PDB	Resolution	State	Ligand	Reference	PDB Date
5HT1A	Cryo-EM	7E2Y	3	Active	Agonist	10.1038/S41586-021-03376-8	4/14/2021
5HT1D	Cryo-EM	7E32	2.9	Active	Agonist	10.1038/S41586-021-03376-8	4/21/2021
5HT2A	X-ray	7WC4	3.2	Inactive	Agonist	10.1126/SCIENCE.ABL8615	1/26/2022
5HT2B	X-ray	6DRZ	3.1	Intermediate	Antagonist	10.1038/s41594-018-0116-7	8/29/2018
5HT2C	Cryo-EM	8DPF	2.8	Active	Agonist	10.1016/J.CELREP.2022.111211	8/24/2022
5HT4R	Cryo-EM	7XT8	3.1	Active	Agonist	10.1016/J.MOLCEL.2022.05.031	7/27/2022
5HT7R	Cryo-EM	7XTC	3.2	Active	Agonist	10.1016/J.MOLCEL.2022.05.031	7/27/2022
AA2AR	X-ray	8CU7	2.1	Inactive	Antagonist	10.1021/acs.jmedchem.2c00462	8/31/2022
AA2AR	X-ray	8DU3	2.5	Inactive	Antagonist	10.1016/j.ejmech.2022.114620	8/10/2022
AA2BR	Cryo-EM	8HDO	2.9	Active	Agonist	10.1038/S41421-022-00503-1	1/18/2023
AA2BR	Cryo-EM	8HDP	3.2	Active	Agonist	10.1038/S41421-022-00503-1	1/18/2023
ACKR3	Cryo-EM	7SK3	3.8	Active	Agonist	10.1126/SCIADV.ABN8063	7/27/2022
ACKR3	Cryo-EM	7SK8	3.3	Active	Agonist Agonist (partial)	10.1126/SCIADV.ABN8063	7/27/2022
ADRB1	X-ray	7BU6	2.7	Active	Agonist	10.1038/S41422-020-00424-2	12/2/2020
ADRB1	X-ray	7BVQ	2.5	Inactive	Inverse agonist	10.1038/S41422-020-00424-2	12/2/2020

ADRB2	X-ray	5JQH	3.2	Inactive	Inverse agonist	10.1038/NATURE18636	7/13/2016
ADRB2	X-ray	6PS4	2.6	Inactive	Antagonist	10.1107/S2052252519013137	11/13/2019
BKRB2	Cryo-EM	7F6H	2.9	Active	Agonist	10.1038/S41467-022-28399-1	1/5/2022
BKRB2	Cryo-EM	7F6I	2.8	Active	Agonist	10.1038/S41467-022-28399-1	1/5/2022
CCR1	Cryo-EM	7VL8	2.9	Active	Apo (no ligand)	10.1038/S41589-021-00918-Z	3/23/2022
CCR1	Cryo-EM	7VLA	2.7	Active	Agonist	10.1038/S41589-021-00918-Z	3/23/2022
CNR1	X-ray	5XR8	3	Active	Agonist	10.1038/NATURE23272	7/12/2017
CNR1	X-ray	7FEE	2.7	Inactive	Agonist PAM	10.1038/S41589-022-01038-Y	6/15/2022
CNR1	X-ray	7V3Z	3.3	Active	Agonist	10.1021/JACS.1C06847	11/24/2021
CNR2	Cryo-EM	6PT0	3.2	Active	Agonist	10.1016/J.CELL.2020.01.007	2/12/2020
CX3C1	Cryo-EM	7XBW	2.8	Active	Apo (no ligand)	10.1126/SCIADV.ABN8048	7/13/2022
CX3C1	Cryo-EM	7XBX	3.4	Active	Agonist	-	7/13/2022
DRD1	Cryo-EM	7CKZ	3.1	Active	Agonist PAM	10.1016/J.CELL.2021.01.028	3/3/2021
DRD1	Cryo-EM	7CKW	3,2	Active	Agonist	10.1016/J.CELL.2021.01.028	3/3/2021
DRD1	Cryo-EM	7JVQ	3	Active	Agonist	10.1016/J.CELL.2021.01.027	2/24/2021
DRD1	Cryo-EM	7X2C	3.2	Active	Agonist	10.1038/S41467-022-30929-W	6/29/2022

FPR1	Cryo-EM	7EUO	2.9	Active	Agonist	10.1038/S41467-022-32822-Y	5/25/2022
FPR2	X-ray	6LW5	2.8	Active	Agonist	10.1038/S41467-020-15009-1	3/25/2020
GALR1	Cryo-EM	7WQ3	2.7	Active	Agonist	10.1038/S41467-022-29072-3	4/20/2022
GALR2	Cryo-EM	7WQ4	2.6	Active	Agonist	10.1038/S41467-022-29072-3	4/20/2022
GHSR	Cryo-EM	7F9Y	2.9	Active	Agonist	10.1038/S41467-021-25364-2	8/18/2021
GHSR	Cryo-EM	7W2Z	2.8	Active	Agonist	10.1038/S41467-022-27975-9	1/19/2022
GPBAR	Cryo-EM	7CFM	3	Active	Agonist	10.1038/S41586-020-2569-1	9/9/2020
GPBAR	Cryo-EM	7CFN	3	Active	Agonist PAM	10.1038/S41586-020-2569-1	9/9/2020
GPR35	Cryo-EM	8H8J	3.2	Active	Agonist	10.1038/S41421-022-00499-8	2/8/2023
GPR88	Cryo-EM	7EJX	2.4	Active	Allosteric ligand	10.1038/S41467-022-30081-5	13/4/2022
MRGX2	Cryo-EM	7VV4	3	Active	Agonist	10.1038/S41586-021-04077-Y	12/1/2021
MRGX2	Cryo-EM	7VV5	2.8	Active	Agonist	10.1038/S41586-021-04077-Y	12/1/2021
MTR1A	Cryo-EM	7VGY	3.1	Active	Agonist	10.1038/S41467-022-28111-3	3/2/2022
MTR1A	Cryo-EM	7VGZ	3.3	Active	Agonist	10.1038/S41467-022-28111-3	3/2/2022
OPRD	X-ray	6PT2	2.8	Active	Agonist	10.1126/SCIADV.AAX9115	12/11/2019
OPRK	X-ray	6B73	3.1	Active	Agonist	10.1016/J.CELL.2017.12.011	1/17/2018
OPRK	X-ray	6VI4	3.3	Inactive	Antagonist	10.1038/S41467-020-14889-7	3/18/2020

OPRM	Cryo-EM	8EF5	3.3	Active	Agonist	10.1016/J.CELL.2022.09.041	11/9/2022
OPRM	Cryo-EM	8EF6	3.2	Active	Agonist	10.1016/J.CELL.2022.09.041	11/9/2022
P2RY1	X-ray	4XNV	2.2	Active	NAM	10.1038/NATURE14287	4/1/2015
P2Y12	X-ray	4PXZ	2.5	Active	Agonist	10.1038/NATURE13288	4/30/2014
SSR2	Cryo-EM	7Y27	3.5	Active	Agonist	10.1038/S41589-022-01130-3	10/19/2022
SSR2	Cryo-EM	7YAE	3.4	Active	Agonist	10.1038/S41467-023-36673-Z	4/19/2023
TSHR	Cryo-EM	7XW5	3	Active	Agonist	10.1038/S41586-022-05173-3	8/24/2022
TSHR	Cryo-EM	7XW6	2.8	Active	Agonist	10.1038/S41586-022-05173-3	8/17/2022

The selected structures were manually pre-processed as explained in the previous chapter, by making sure they presented the correct orientation, removing all non-receptor residues and adding any missing atoms. To enable meaningful comparisons across class A GPCRs, residues were mapped onto the Ballesteros–Weinstein numbering scheme using GPCRdb annotations [80,101]. This standardized residue framework provides a common reference system across structurally and sequence-diverse receptors, allowing cholesterol interactions to be analysed and compared at equivalent topological positions throughout the family [65,101]. All simulations were performed using the default protocol settings, consisting of two independent 30 μs molecular dynamics replicas per system in a POPC:cholesterol (75:25) membrane. Following analysis and data merging, the pipeline generates a PDB file containing residues identified as outliers, along with their associated density, using a 5 Å cutoff to assign density points to residues. These files define the cholesterol interaction hot spots and were used for the initial comparison. Hot spot files from the 56 simulations were mapped onto GPCRdb-annotated structures, with residue annotations encoded in the temperature-factor field. Using the GPCRdb generic numbering, we then quantified the frequency of each residue across all runs to identify receptor regions with a higher propensity for cholesterol binding. The results of this analysis are reported in **Table 14**.

Table 14. *Residues most frequently identified as relevant for cholesterol binding across 56 protocol runs. Only residues observed in 20 or more runs are reported.*

Residue	Nº occurrences
4.50	54
2.52	50
3.34	45
4.54	36
3.41	33
3.31	32
5.46	32
2.48	31
4.52	30
2.55	29
1.45	25
4.45	23
7.43	22
5.49	21
7.51	20

These findings are visualized in **Figure 15** by mapping residue frequencies onto the β_2-adrenergic receptor structure (PDB ID: 2RH1 [102]), which is commonly used as a canonical class A GPCR reference. Residue frequencies were encoded in the occupancy field and color-coded according to four occurrence intervals (1–10, 11–20, 21–30, and ≥30 systems). This representation reveals two dominant regions of frequent cholesterol binding (**Figure 15A** and **Figure 15B**), located at

the TM2–TM3–TM4 and TM3–TM4–TM5 interfaces, respectively. Both regions coincide with cholesterol-binding sites previously reported in the literature and spatially align with cholesterol molecules resolved in the CB_1 receptor crystal structure (PDB ID: 7FEE [53]). Two additional regions of frequent cholesterol binding are detected with lower prevalence, with their most frequently identified residues occurring in 11–30 of the 56 systems analysed. These regions, located at the TM5–TM6 and TM1–TM7 interfaces (**Figure 15C**, left and right panels, respectively), are nevertheless consistent with reported cholesterol-binding sites. Indeed, multiple published GPCR structures show cholesterol molecules bound at these locations, including examples represented in the present dataset (e.g., PDB IDs: 7E2Y and 6PT0)[103–105].

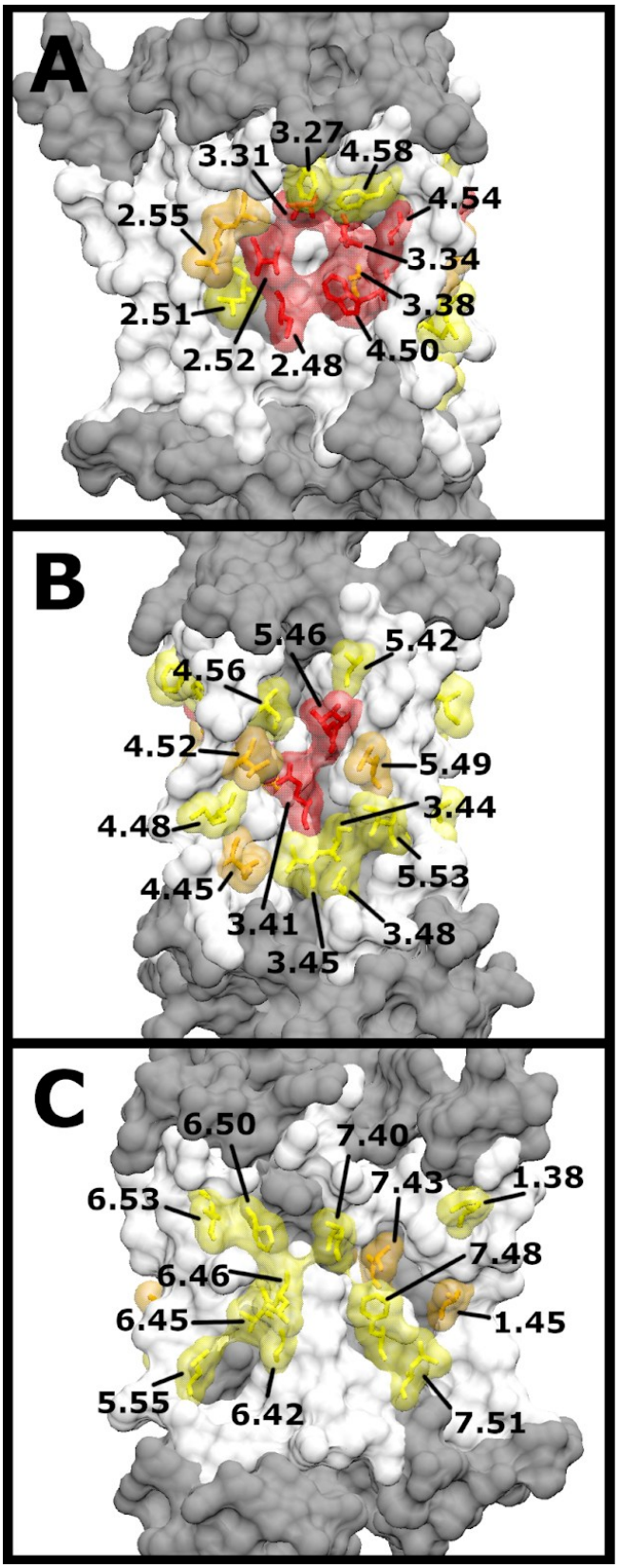


Figure 15. *Mapping of cholesterol interaction hot spots across class A GPCRs. Residue frequencies derived from 56 simulations are projected onto the β2-adrenergic receptor structure (PDB ID: 2RH1). Residues are coloured by absolute occurrence (1–10, white; 11–20, yellow; 21–30, orange; ≥30, red), while residues never detected are shown in grey. Residues with ≥21 occurrences are labelled. Four recurrent hot spot regions emerge from this mapping: two dominant interfaces formed by TM2–TM3–TM4 (A) and TM3–TM4–TM5 (B), and two less frequent but consistent sites at the TM5–TM6 and TM1–TM7 interfaces (C, left and right, respectively).*

Beyond their correspondence with previously reported cholesterol-binding sites, the spatial distribution of these frequently identified positions also raises the question of how lipid interactions at these locations may relate to conserved features of GPCR structure and activation. The identified frequent cholesterol-binding residues cluster on membrane-facing surfaces of TM2/3/4 and TM5/7, delineating regions that are structurally connected to conserved elements of the class A GPCR activation machinery. Positions on TM2 (2.48, 2.52, 2.55), together with those on TM4 (4.45, 4.50, 4.52, 4.54), define a membrane-exposed TM2/TM4 flank (coloured in red in **Figure 16**) adjacent to Asp2.50, a highly conserved residue that coordinates a sodium ion and an associated water network known to stabilize inactive receptor conformations. Perturbations in this region have been shown to influence the energetic balance between inactive and active states by modulating helix packing around the sodium-binding site. Notably, Trp4.50, observed in 54 of the 56 systems and representing the most conserved residue within transmembrane helix 4, has previously been implicated in cholesterol binding [53,102,106], consistent with a conserved role in mediating lipid–protein interactions at the membrane interface. Frequent positions on TM3 (3.31, 3.34, 3.41, coloured in cyan in **Figure 16**) map to a structurally conserved region of the helix that links membrane-facing surfaces to the intracellular portion of TM3, which contains the DRY motif (3.49–3.51), a central determinant of G-protein coupling and receptor activation. Although the discussed positions themselves are distal from the DRY motif, their location along the same helix suggests a potential sensitivity of this transmission pathway to lipid-induced changes in helix orientation or packing, consistent with a role for TM3 as a conduit for long-range conformational coupling. Finally, positions identified at the TM5/TM7 interface (5.46, 5.49, 7.43, 7.51, coloured in yellow in **Figure 16**) delineate a membrane-accessible corridor adjacent to conserved activation motifs on these helices. TM5 harbours Pro5.50, a key component of the PIF motif that contributes to the outward movement of TM6 during activation, while TM7 contains the NPxxY motif (7.49–7.53), which undergoes a characteristic rearrangement associated with formation of the intracellular signalling cavity. Although the identified residues do not directly overlap these motifs, their proximity and helix-facing orientation suggest that lipid interactions in this region could indirectly influence the relative packing of TM5 and TM7 and thereby modulate the conformational transitions associated with activation. Collectively, the distribution of frequent cholesterol-binding positions delineates a set of membrane-accessible regions that are geometrically and mechanically positioned to influence receptor conformational equilibria through indirect coupling to conserved activation elements.

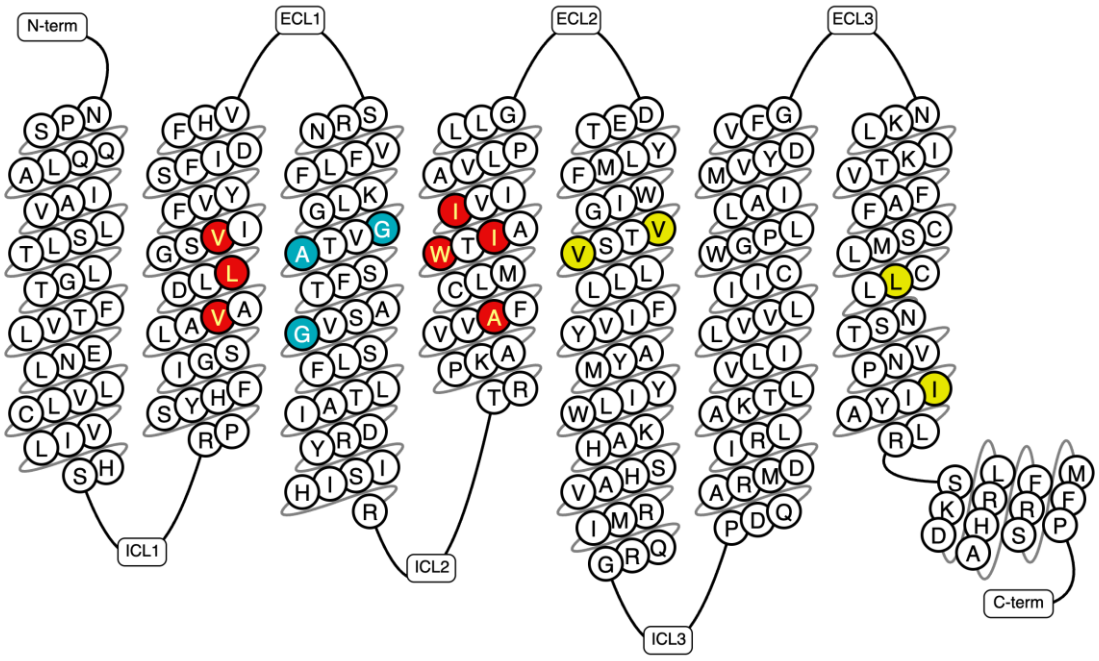


Figure 16. *Snake plot of the CB_1 receptor. The coloured residues match with the TM2/4 interface (red), the TM3 helix containing the DRY motif (cyan) and the TM5/7 interface (yellow).*

While these results provide a robust overview of cholesterol interaction patterns across GPCRs, they are derived from contact and density analyses of molecular dynamics trajectories. As such, they identify regions that frequently engage cholesterol over extended periods of time, but they do not, by themselves, establish whether these regions form well-defined binding pockets. From a drug-discovery perspective, this distinction is critical. Hot spots located on largely flat or solvent-exposed surfaces, where cholesterol binds transiently or due to collisional effects, are unlikely to be actionable. Instead, relevance lies in identifying sites where cholesterol occupies a structured cavity that can stably accommodate a ligand, consistent with a classical pocket-based binding mode. Numerous computational methods are available for binding-pocket prediction, relying on distinct underlying principles. Geometry-based approaches identify cavities from the three-dimensional shape of the protein surface (e.g., Fpocket, CASTp) [107,108], whereas energy- or probe-based methods detect regions with favourable interaction potentials (e.g., SiteMap, FTMap)[109110]. More recently, machine-learning approaches have also been introduced (e.g., DeepSite, P2Rank)[111,112]. The performance of several of these methods has been specifically evaluated for the retrospective detection of allosteric pockets in membrane proteins [113]. Regardless of the underlying methodology, pocket-detection tools typically return multiple candidate sites per structure. While the number, shape, and precise location of predicted pockets may vary across methods, the output is generally a set of potential binding regions. When a ranking is provided, it is usually based on a generic assessment of ligandability rather than on system-specific interaction patterns.

As described in Chapter 3, we incorporated a pocket-detection step into our workflow and leveraged the identified cholesterol hot spots to score candidate sites based on their propensity to accommodate cholesterol. Among the available tools, NanoShaper [66] was particularly well suited to this task due to its computational efficiency and flexible parameterization. NanoShaper computes the solvent-excluded surface using a ray-tracing algorithm combined with a dual-probe approach, making it especially appropriate for detecting pockets at the protein–membrane interface. As reported in Chapter 4, we established a pocket-based baseline for cholesterol binding at the protein–membrane interface of class A GPCRs by comparing co-crystallized cholesterol positions with geometrically defined surface cavities. Using a subset of 30 systems, we distinguished cholesterol interactions occurring within bona fide pockets from those occurring on flat or shallow surfaces, which cannot be recovered by pocket-detection approaches. This analysis provided a quantitative framework to estimate the fraction of cholesterol molecules occupying genuine surface cavities and to contextualize subsequent pocket-ranking results.

Following CHAMP analysis of the 56 systems, pocket detection was performed on the corresponding starting structures used in the simulations, ensuring consistency between NanoShaper predictions and the regions characterized by our protocol. Pocket filtering was then applied according to the following criterion. Pseudo-atoms provided by OPM-oriented structures [54] were used to define the upper and lower boundaries of the membrane, thereby delimiting the transmembrane region along the z-axis. In addition, a cylindrical exclusion region with a diameter of 5 Å and its principal axis aligned with the protein center was defined to remove predominantly internal cavities. This step was intended to eliminate internal cavities - such as the orthosteric binding site, whose presence is expected - that are not relevant to cholesterol-mediated modulation. In particular, this filtering reduces the inclusion of pockets whose scores are driven by outliers that fortuitously contact cholesterol on one face while contributing to an internal cavity on the other. Pockets whose centroids fell outside the membrane boundaries or within the central cylinder were excluded prior to ranking. After filtering, the number of retained pockets varied across structures, with most systems yielding 7–10 ranked pockets and others retaining five or fewer, with a minimum of three pockets per structure. In some cases, pockets passing the geometric filters nevertheless exhibited low scores, primarily due to limited spatial overlap with cholesterol density. To minimize the impact of such pockets in the aggregated analysis, only the top five scoring pockets per structure were retained when available, and only if their score exceeded 1000, a threshold typically requiring the inclusion of at least one outlier residue.

Overall, application of NanoShaper across the 56 receptors yielded 272 pockets that passed all geometric and scoring filters. For clarity, **Table 15** reports the 5 highest-scoring pockets according to our cholesterol-based score, and **Figure 17** shows structural overlays comparing the corresponding protein regions with the original co-crystallized cholesterol positions for the top

five pockets. For brevity, the remaining 262 pockets are not listed individually, although they are included in all aggregate analyses. All pockets were labelled following the notation introduced by Peter et al. for GPCR pocket reporting [113], which adopts the format class–location–TM–location2, where the last term indicates the vertical position within the transmembrane region. Because our dataset is restricted to class A GPCRs and our analysis is limited to the protein–membrane interface, all pockets are labelled as A–EH–TMXX–(ext/mid/int). This nomenclature not only allows rapid identification of pocket location but also facilitates future cross-referencing with published GPCR pocket annotations.

Table 15. *Top 5 scoring pockets across the 56 studied systems. As expected, these pockets are all ranked in the first position of their systems.*

PDB ID	Residues	Total Score	Pocket label	Rank position in the system
7vv5	3.44 3.45 3.48 3.49 3.52 4.41 4.45 4.49 5.46 5.49 5.50 5.53	11133	A-EH-TM345-mid	1
7fee	2.52 2.56 2.59 3.26 3.27 3.28 3.30 3.31 3.34 3.35 4.50 4.54 4.57 4.58 4.61	9761	A-EH-TM234-ext	1
7sk3	2.42 3.41 3.44 3.45 3.48 3.49 3.52 35.95 4.41 4.42 4.45 4.48 4.49 4.52 5.45 5.49 5.50 5.53	9752	A-EH-TM2345-mid	1
5xr8	2.52 2.56 3.27 3.28 3.30 3.31 3.34 3.35 4.50 4.54 4.58 4.61	9191	A-EH-TM234-ext	1
8ef5	1.41 1.42 1.45 1.48 1.49 7.44 7.47 7.48 7.51 7.55	8263	A-EH-TM17-mid	1

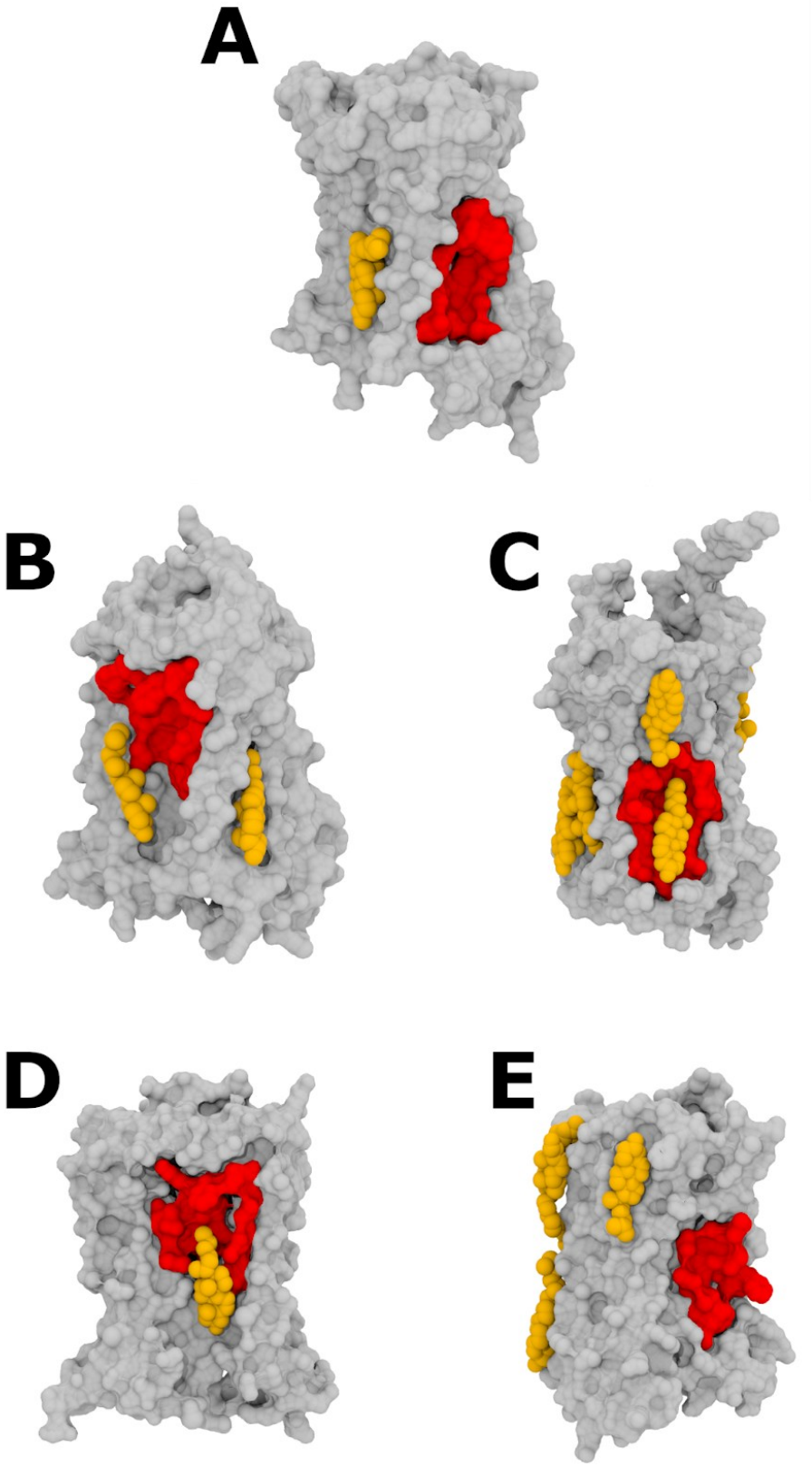


Figure 17. *Top 5 scoring pockets, coloured in red, represented in their corresponding structures: MRGPRX2 (A), CB1 (B and D), ACKR3 (C) and μ-opioid-R. The cholesterol molecules originally present are represented in orange.*

In three of the five cases, the highest-scoring pocket coincides with one of the experimentally observed cholesterol locations (B–D), with both CB1 structures mapping to the same pocket. In

contrast, the MRGPRX2 does not present an overlap between the cholesterol position in the crystal and the highest ranked pocket. We found that the combination of NS plus our scoring does not detect a pocket with at least 1 outlier in the region where the crystallographic cholesterol is, despite our analysis finding relevant outliers and density in that region. Regarding the μ-opioid receptor, although the original structure contains eight cholesterol molecules in the vicinity of the receptor, none are located within the top-ranked pocket. However, the second and third ranked pockets within the system are the ones corresponding to the original cholesterol position, despite scoring much lower in our analysis (1466 and 1449, respectively).

Notably, two of the five highest-scoring pockets originate from CB1 structures, highlighting a pronounced propensity of this receptor to accommodate cholesterol.

Finally, we further examined the ranked pockets according to the transmembrane helices contributing to their formation. We summarized these results in **Table 16**. As expected, the most frequently occurring pockets are formed by the TM3–TM4–TM5, TM5–TM6, TM1–TM6, and TM2–TM3–TM4 interfaces. These regions coincide with the highest-ranked pockets and align closely with the four hot spot regions previously identified in our analysis.

Table 16. *Frequency of helix bundles participating in the scored pockets.*

TM helices	**Nº occurrences**
TM3-4-5	35
TM5-6	26
TM1-7	25
TM2-3-4	18
TM3-4	16
TM2-3	15
TM1-7-8	14
TM6-7	13
TM2-4	10
TM4-5	10

As shown in **Figure 17A–D**, the highest-scoring pockets generally coincide with previously reported cholesterol-binding sites, despite the analysis being entirely agnostic of the original cholesterol positions. This correspondence is not universal, however, as some high-scoring

pockets lack co-crystallized cholesterol in the reference structures, as illustrated in **Figure 17E**. In addition to the previous assessment, we overlapped the top ten scoring pockets with the aim of identifying the most prominent cholesterol-binding locations. As before, we used a representative structure of the β_2-adrenergic receptor to visualize residues appearing in the top ten pockets, which are shown as a transparent red surface in **Figure 18**. Moreover, residues exhibiting an absolute frequency of at least 20 occurrences in the entire 272 pocket set were highlighted in licorice representation and coloured red (20–29 occurrences) or blue (≥30 occurrences). This analysis identifies three main regions around the receptor, as shown in **Figure 18**. These three regions display a clear overlap with those identified in our initial assessment based solely on hot spot positions. Specifically, the relevant pockets encompass the TM2–TM3–TM4 region (**Figure 18A**), the TM3–TM4–TM5 region (**Figure 18B**), and the TM1–TM7 region (**Figure 18C**).

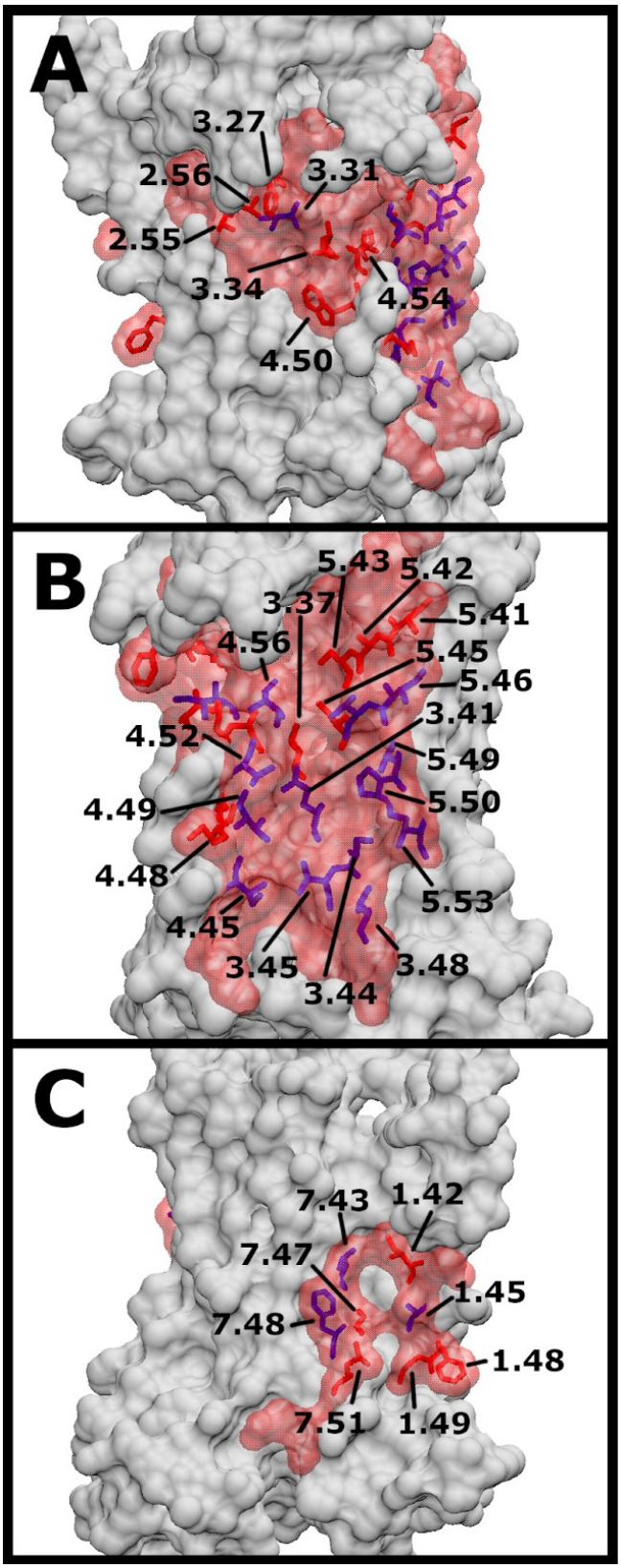


Figure 18. *Overlap of the top 10 scoring NanoShaper pockets (red transparent surface). Three clearly distinct areas, delimited by TM2-3-4 (A), TM3-4-5 (B) and TM1-7 (C) were identified. The residues with an absolute frequency of at least 20 are also represented in licorice, coloured in red (20-29) and blue (>29).*

We next turned to the literature to identify reported allosteric binding events that overlap with the consensus cholesterol-binding regions identified in our analysis. Examples of allosteric modulators binding class A GPCRs at the protein–membrane interface remain relatively scarce, as most known allosteric ligands bind in extracellular, intracellular, or intramembrane vestibule regions with limited direct membrane contact. We therefore compared the regions identified by our approach with the binding sites of reported membrane-facing allosteric modulators (**Table 17**), with the aim of assessing overlap and evaluating the potential of these regions as candidate sites for novel ligand binding. Several crystal structures of class A GPCRs bound to allosteric modulators reveal binding sites at the protein–membrane interface that overlap with regions identified in our analysis. In $P2Y_1R$, the antagonist BPTU[114] binds at the interface between TMs 1–2–3; all residues involved in this interaction are detected as outliers in at least one of the 56 systems analysed, with some appearing in up to 29 systems. In PAR2, the negative allosteric modulator AZ3451[115] occupies a pocket formed by TMs 2–3–4, corresponding to one of the three main regions consistently identified by both the frequency-based analysis and the scored pocket detection; notably, this pocket is composed of residues with particularly high occurrence frequencies across the dataset. In GPR40, binding of the ago-PAM AP8[116] occurs between TMs 3–4–5, in agreement with one of the common cholesterol-interaction regions observed in our systems. A comparable region is involved in binding of the modulator W54011 [117] to C5aR, where both frequent positions and cholesterol density are observed. For the β_2-adrenergic receptor, the modulator Cmpd-6FA [118] binds near the intracellular side at the interface between TMs 2–3–4 and ICL2; although the overlap with outlier positions is less pronounced, several residues in this region display elevated frequencies across multiple systems. The CB1 negative allosteric modulator ORG27569 [119] binds between TMs 2 and 4, a region enriched in both frequent positions and cholesterol density and previously described as a conserved cholesterol-binding site. Another β_2AR modulator, AS408 [120], binds between TMs 3 and 5, overlapping with crystallographically resolved cholesterol molecules in several structures. Finally, the positive allosteric modulator of the A_1 adenosine receptor [121] binds at the TM1–TM6–TM7 interface, again corresponding to a region enriched in frequent positions and cholesterol density.

Although the degree of overlap varies among these examples, all the reported allosteric pockets contain at least two residues with relatively high occurrence frequencies in our dataset. This observation suggests that, even when the correspondence is not exact, likely reflecting the specificity of protein–ligand interactions that cannot be fully recapitulated by cholesterol alone, there could be a meaningful relationship between cholesterol binding and allosteric modulation. In this light, a structure-based approach combining in-depth simulation-driven identification of cholesterol-binding regions (CHAMP) with consensus analysis across aggregated datasets can

provide a rational starting point for drug discovery efforts aimed at identifying membrane-facing allosteric modulators.

Table 17. *Summary of some reported allosteric modulators for class A GPCRs that bind at the protein-membrane interface. Each residue participating in the pockets is accompanied by a bolded number between parentheses, showing its frequency on our 56 systems.*

Target	**Allosteric modulator**	**Type**	**Pocket (outlier frequency)**
P2YR	BPTU	NAM	1.43 **(4)**, 1.47 **(8)**, 2.55 **(29)**, 2.56 **(2)**, 2.58 **(8)**, 2.59 **(7)**, 3.24 **(2)**, 3.27 **(17)**, 3.28 **(1)**
PAR2	AZ3451	NAM	2.52 **(50)**, 3.31 **(32)**, 3.34 **(45)**, 3.38 **(17)**, 4.50 **(54)**, 4.54 **(36)**, 4.61 **(1)**
GPR40	AP8	Ago-PAM	2.42 **(1)**, 3.37 **(2)**, 3.41 **(33)**, 3.44 **(16)**, 3.45 **(16)**, 3.48 **(18)**, 3.49 **(0)**, 3.52 **(2)**, 4.42 **(1)**, 4.45 **(23)**, 4.49 **(9)**, 4.52 **(30)**, 4.56 **(11)**, 5.42 **(11)**, 5.45 **(5)**, 5.46 **(32)**, 5.49 **(21)**, 5.53 **(18)**, 5.57 **(1)**
C5aR	W54011	NAM	3.40 **(0)**, 3.41 **(33)**, 4.48 **(18)**, 4.52 **(30)**, 4.56 **(11)**, 5.45 **(5)**, 5.49 **(21)**, 5.50 **(1)**, 5.54 **(0)**, 5.57 **(1)**
β2AR	Cmpd-6FA	PAM	2.42 **(1)**, 3.45 **(16)**, 3.48 **(18)**, 3.49 **(0)**, 3.52 **(2)**, 3.56 **(0)**, 4.41 **(7)**, 4.44 **(2)**, 4.45 **(23)**, 4.48 **(15)**
CB1	ORG27569	NAM	2.41 **(3)**, 2.44 **(2)**, 2.45 **(8)**, 2.48 **(31)**, 4.46 **(2)**, 4.47 **(2)**, 4.50 **(54)**, 4.51 **(1)**, 4.54 **(36)**
β2AR	AS408	NAM	3.41 **(33)**, 3.44 **(16)**, 3.45 **(16)**, 3.48 **(18)**, 5.45 **(5)**, 5.46 **(32)**, 5.49 **(21)**, 5.50 **(1)**, 5.53 **(18)**
A1AR	MIPS521	PAM	1.41 **(10)**, 1.42 **(3)**, 1.45 **(25)**, 6.43 **(5)**, 6.46 **(17)**, 6.47 **(4)**, 7.40 **(17)**, 7.41 **(0)**, 7.44 **(0)**, 7.48 **(19)**

Chapter 6: Interactive online platform

The data generated in this study comprise a heterogeneous set of files, including protein structures (PDB format), molecular dynamics trajectories (compressed XTC format), and tabular outputs storing key analysis results (comma-separated values, CSV). While direct access to these data is, in principle, possible for third parties, limitations in data-transfer rates and security considerations render public dissemination impractical with currently available means. To address these challenges, we developed a publicly accessible web platform designed to facilitate the visualization, exploration, and download of the most relevant data produced by the CHAMP workflow. To this end, all data were reorganized into a lightweight and structured format suitable for delivery through a standard web interface. While full-resolution molecular dynamics trajectories are not yet publicly available due to network and bandwidth constraints, we aimed to provide at least a representative overview of the simulations. For this purpose, reduced versions of the original trajectories were generated by retaining one frame every 10,000 simulation steps. This approach allows the overall progression of each MD simulation to be visualized from start to end, at the cost of temporal smoothness resulting from larger frame-to-frame jumps. The resulting trajectory files are approximately 1.4 MB in size, enabling their direct hosting as static assets without the need for a dedicated backend server or streaming infrastructure. This design choice allows all data files to be co-located with the remaining web content, such that they can be served directly by the hosting provider without additional server-side communication or processing. The standard workflow of the web platform is illustrated in **Figure 19**. The platform is organized around three main HTML pages: index.html, system.html, and aggregated.html. The landing page (index.html) provides entry points to the available datasets, while system.html enables access to results for individual receptor structures and aggregated.html presents results grouped by receptor families and subfamilies, where applicable. The detailed layout and functionality of each page are described below. Dynamic behaviour across the platform is implemented through a collection of JavaScript (JS) files, which control interactive elements such as the GPCR family tree, molecular viewers, and the on-demand loading of structures and data tables. The underlying data are organized in two complementary ways. First, individual directories are provided for each receptor structure, containing all structure-specific files along with a dedicated subdirectory for data associated with detected pockets. Second, an aggregated data directory contains separate subdirectories for each receptor family and subfamily, each storing the corresponding summary files (e.g., frequently interacting residues and highest-scoring pockets) used for collective analyses and visualizations. At the root of the aggregated data directory, we include a PDB structure of the β_2-adrenergic receptor, which serves as a representative class A GPCR scaffold for displaying aggregated results across the family.

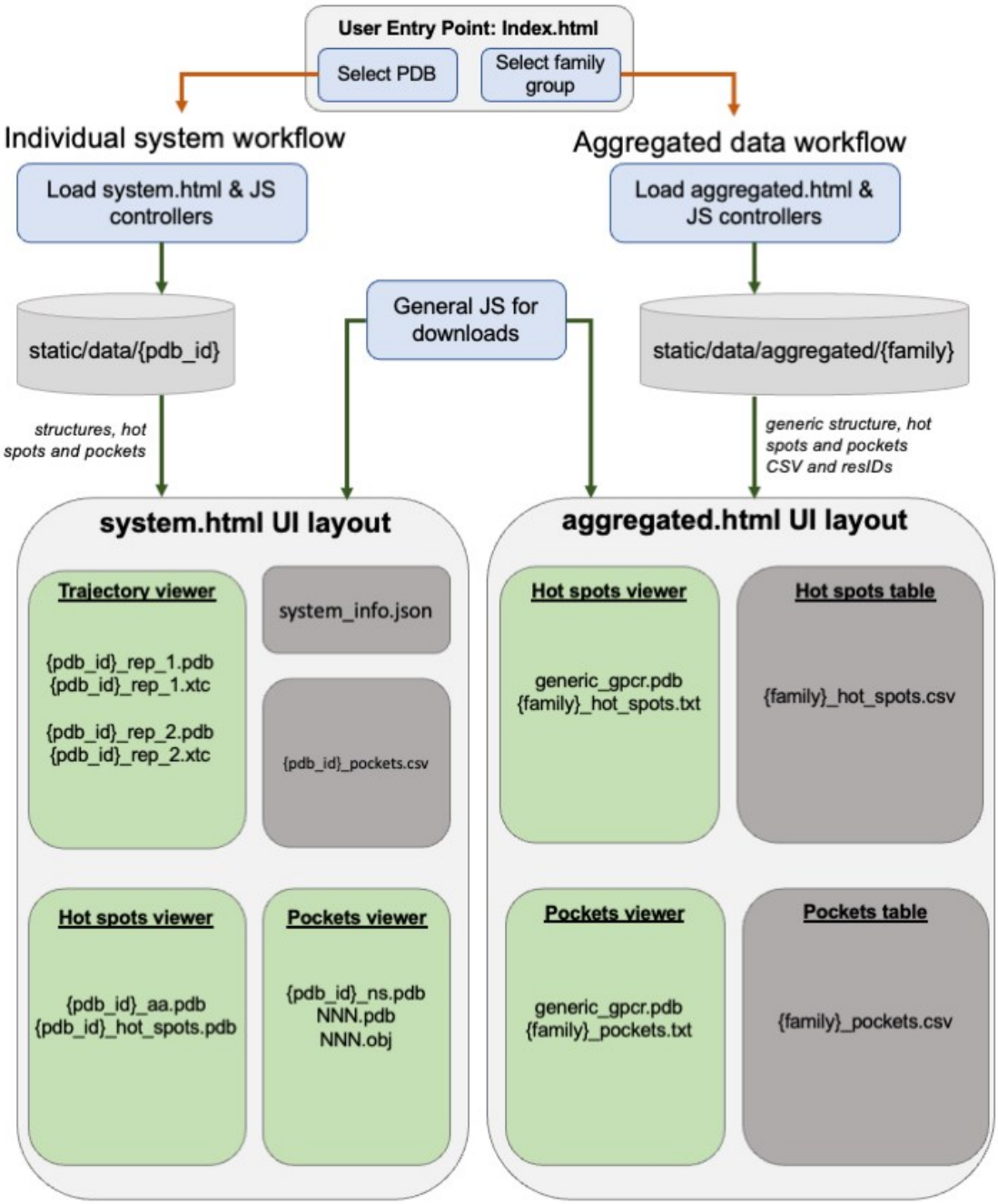


Figure 19. *Standard workflow of the web platform. The backbone of the platform consists of three main HTML pages (index.html, system.html, and aggregated.html). Dynamic elements and user interactions are controlled by the corresponding JavaScript (JS) files, while data are loaded on demand depending on the elements selected through the available navigation menus.*

The landing page of the platform, shown in **Figure 20**, presents the title, authorship information, and a brief description of the platform. Below this introductory section, two fly-out menus provide access to the available data. The first menu allows users to navigate the class A GPCR hierarchy down to a specific receptor and to select an associated PDB structure. The same selection can alternatively be made by directly clicking on the corresponding PDB identifier within the interactive GPCR tree displayed below the menus. The second menu provides access to

aggregated data, organized by receptor families and subfamilies. To ensure that only statistically meaningful summaries are presented, aggregated data were computed exclusively for families and subfamilies represented by at least four PDB structures. As illustrated in **Figure 20**, menu entries that do not meet this criterion are greyed out and disabled, preventing user interaction.

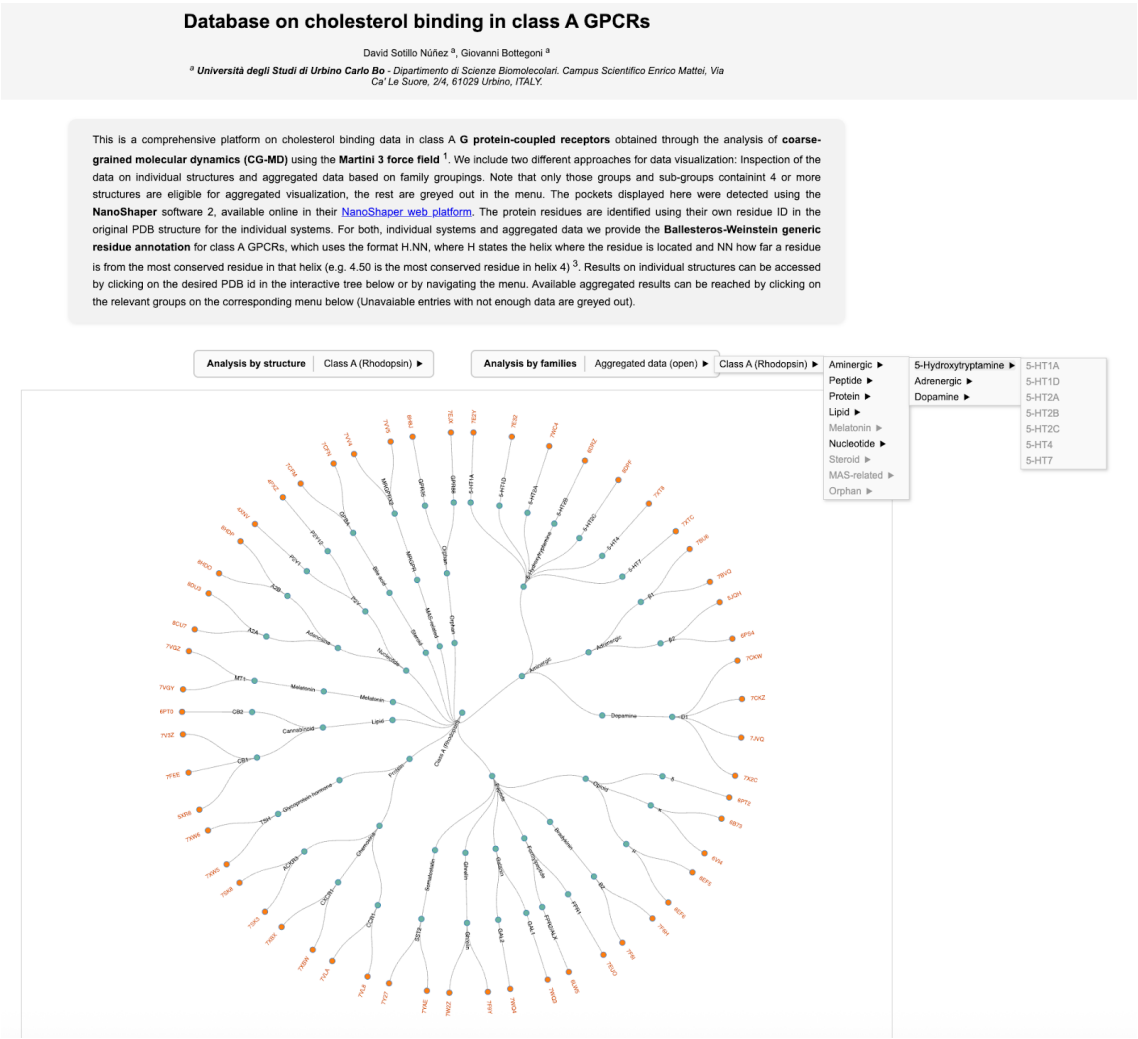


Figure 20. *Landing page of the platform. It includes title, authors, affiliations and a brief description of the platform. It contains both menus for accessing the data and an interactive GPCR tree to directly access PDBs of interest.*

Upon selecting a PDB identifier, either from the navigation menu or directly from the interactive GPCR tree, the corresponding system page opens in a new browser tab. **Figure 21** and **Figure 22** illustrate the elements available on this page, using the CB1 structure PDB 5XR8 [89] as a representative example. The left column of the page contains two molecular viewers. The upper viewer (**Figure 21**, top) allows visualization of the strided coarse-grained MD simulations, with playback controls enabling navigation through the stored frames. The display of Martini 3 beads and cholesterol particles can be independently toggled using the corresponding buttons. Users may switch between the two simulation replicas at any time by selecting the appropriate tab, and

the camera orientation can be reset to predefined views, including a side view (default starting position) and a top view.

The lower viewer (**Figure 21**, bottom) displays the cholesterol-interaction hot spots identified during the CG-MD simulations, including outlier residues and the associated reproducible cholesterol density. These features are rendered on top of an aligned atomistic representation of the receptor structure. Both the strided trajectory and the hotspot structure files can be downloaded directly using the corresponding buttons provided below each viewer.

The right-hand side of the system page is dedicated to displaying system-specific information (top panel in **Figure 22**) including the PDB identifier, membrane composition, total simulated time, and the NanoShaper parameters used for pocket identification. Below this information panel, a table lists the five highest-scoring pockets identified for the selected system, reporting the residues contributing to each pocket together with their associated scores. Residues are labelled using both the original residue numbering of the PDB structure and the generic Ballesteros–Weinstein notation for class A GPCRs [80]. Checkboxes in the first column of the table allow users to toggle the visualization of individual pockets in the molecular viewer located immediately below the table, facilitating inspection of each pocket in isolation. Pockets are rendered using a volumetric representation of the cavity, together with the participating residues, shown in blue and red, respectively (**Figure 22**, bottom). As for the cholesterol-interaction hot spots, the complete table containing all ranked pockets can be downloaded using the corresponding button. Selecting an entry from the second navigation menu opens a new browser tab displaying the corresponding aggregated results page (*aggregated.html*). A snapshot of this page is shown in **Figure 23**, using the aggregated data for the full class A GPCR set as an illustrative example. This page is divided into two main sections. The upper section presents residues most frequently identified as outliers across multiple systems. In the molecular viewer on the left, these residues are displayed on the representative β_2-adrenergic receptor structure, while the adjacent table reports up to the 15 most frequent positions. The lower section is dedicated to detected and scored pockets and follows an analogous layout. The table on the right lists the ten highest-scoring pockets across the analysed systems within the selected group, while the molecular viewer on the left displays the combined spatial distribution of residues participating in these pockets. Both summary tables, frequent residues and detected pockets, can be downloaded in CSV format using the buttons located at the top of the page.

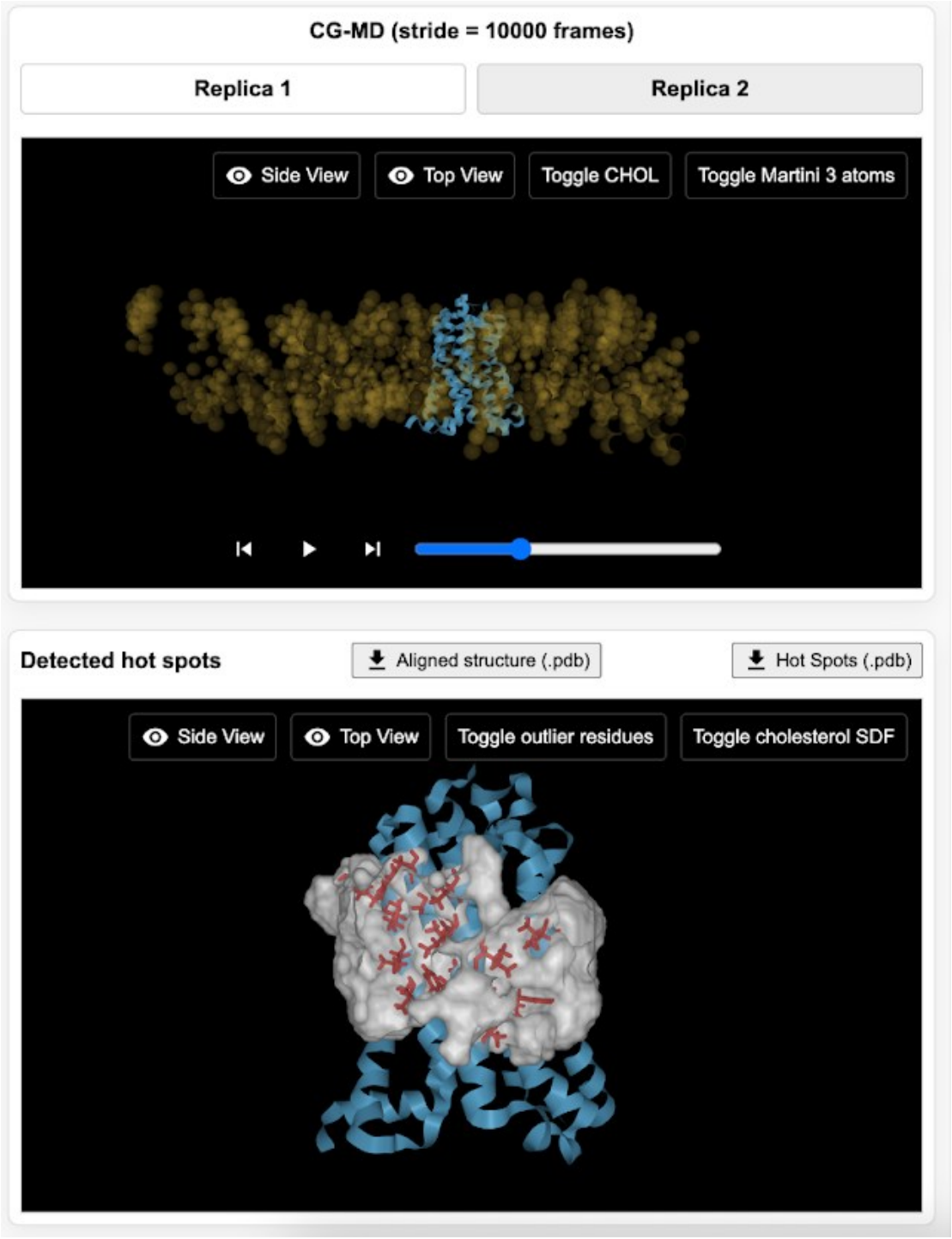


Figure 21. *Snapshot of the left column of the system.html page, which contains two molecular viewers. The upper viewer enables visualization of a reduced version of the two replicas of the original molecular dynamics trajectory from which the results were derived. The lower viewer displays an atomistic structure aligned with the cholesterol-interaction hot spots identified from the coarse-grained MD simulations, including both outlier residues and associated cholesterol density. In both viewers, interface elements allow individual components to be toggled on and off, and the aligned atomistic structure as well as the hotspot data can be downloaded. Controls for adjusting the orientation of the system are also provided.*

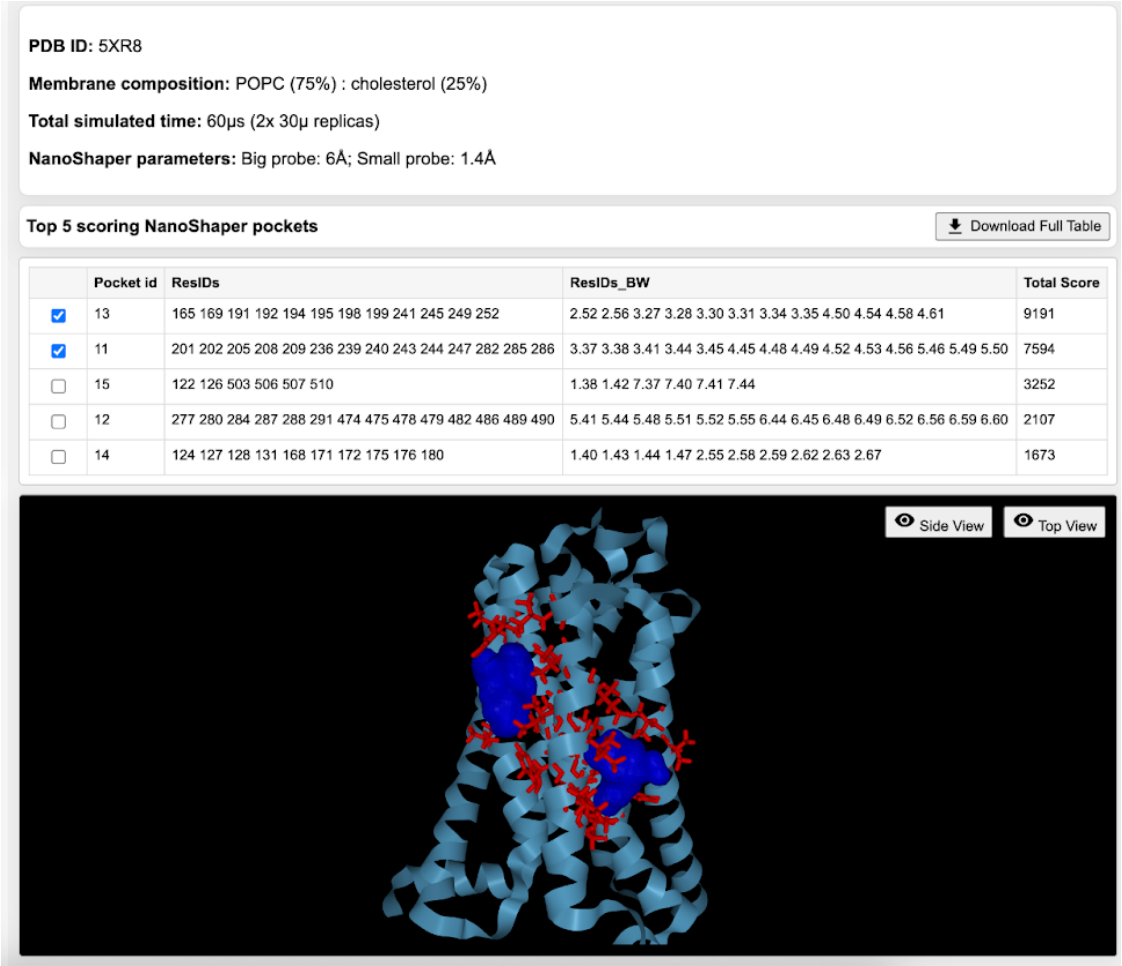


Figure 22. *Snapshot of the right column of the system.html page, showing the system information panel, the table of identified pockets, and the associated molecular viewer. Selecting or deselecting the checkboxes in the table toggles the visualization of the corresponding pockets in the viewer.*

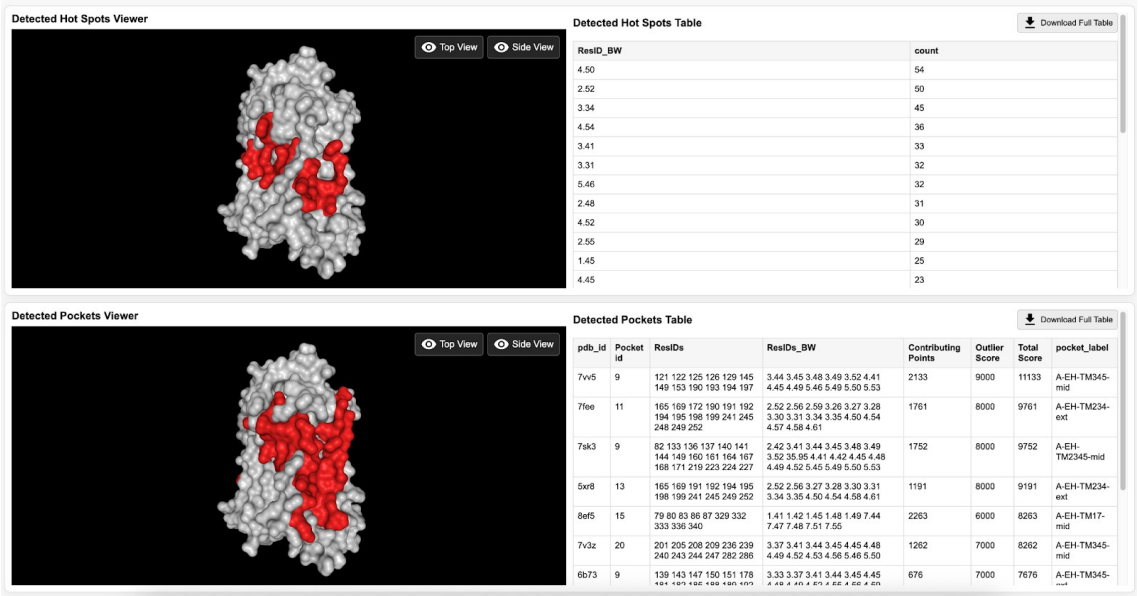


Figure 23. *Snapshot of the aggregated.html page. The upper section displays the most frequent residues identified as statistical outliers (up to the top 15), together with their total occurrence counts across the systems analyzed for the selected group (table on the upper right). These residues are visualized in the molecular viewer located in the upper left. The lower section presents the highest-scoring pockets (up to the top 10), with the combined spatial distribution of residues contributing to these pockets displayed in the molecular viewer in the lower left.*

The platform was implemented as a purely front-end web application, with no back-end component, serving only static files and relying on JavaScript to control all dynamic page behaviour. These design choices make the platform particularly well suited for deployment via GitHub Pages. Accordingly, the complete source code and associated data were uploaded to a dedicated repository within our group's GitHub organization, where they are publicly available [122]. The platform was subsequently deployed as a standalone web application and can currently be accessed at the following address: https://urbino-camd-lab.github.io/GPCRCholesterolAtlas.

For molecular visualization, we employed the NGL molecular viewer [123], which ensures broad compatibility across modern web browsers. As a result, any browser supporting NGL should be able to render the full contents of the platform. The application was tested extensively and verified to function correctly on multiple Chromium-based browsers.

Chapter 7. Conclusions

In this work, we briefly reviewed the current state in the literature of cholesterol-mediated GPCR modulation and the tools and techniques available, both computational and experimental, to study cholesterol binding. Moreover, we discussed the limitations of such methods and highlighted the potential of using of coarse-grained molecular dynamics and a combination of protein-cholesterol contacts analysis and a cholesterol distribution mapping in detecting cholesterol binding sites. Following such trend, we presented here a robust and largely automated protocol for building, simulating, and identifying cholesterol interaction hot spots in CG protein–membrane systems, starting from a pre-processed atomistic protein structure. Through the systematic integration of several publicly available tools, the protocol reliably reproduces results that are consistent with previous computational and experimental studies.

However, it is important to note that the application of restraints (elastic networks) in CG-MD simulations to preserve protein structural integrity may restrict conformational sampling and hinder the identification of binding sites. To adequately capture the structural heterogeneity of certain receptors, the protocol may therefore require the use of multiple input structures for the same system, enabling the exploration of interactions across distinct functional states and hidden binding cavities. This can be achieved either by leveraging multiple experimental structures or by integrating CG-MD with atomistic and, eventually, enhanced-sampling simulations. Such an approach is likely to be essential for addressing protein plasticity and for the comprehensive identification of allosteric sites.

We further demonstrate that, with minimal modifications, the pipeline can be readily adapted to investigate molecules other than cholesterol or membrane lipids. To this end, the CB1 receptor negative allosteric modulator Org27569 was used as a retrospective validation case, focusing on the identification of interaction hot spots. The protocol successfully prioritized the region surrounding Trp241$^{4.50}$, corresponding to the known binding site of Org27569, thus supporting its general applicability beyond lipid ligands.

The availability of the pipeline as an installable Python package makes it an accessible and user-friendly tool that is readily deployable for novel applications involving ligand interactions at the protein–membrane interface. At the time of writing this work a docker implementation of this protocol, aimed to facilitate runs in HPC environments, is under development [124]. Furthermore, we are working on identifying sections of the protocol that could be optimized for parallel runs.

At present, more than 500 class A GPCR structures are publicly available, including over 100 that contain at least one resolved cholesterol molecule. A systematic application of the present

protocol to a representative subset of these structures would enable not only a robust retrospective assessment of known cholesterol binding sites, but also the identification of previously unrecognized sites and their comparative classification across receptor families. Such an analysis could provide a unifying structural framework for GPCR–cholesterol interactions, ultimately leading to an annotated "cholesterolome" of class A GPCRs. On this basis, we directed our efforts on extending the use of our protocol to a larger representative dataset of 56 class A GPCR structures, from which we extracted relevant insights regarding frequent cholesterol-populated regions, matching previously reported motifs.

Finally, on an effort towards sharing the data we gathered in a user-friendly fashion, we developed a publicly accessible platform which includes both results from all the 56 individual systems simulated and the aggregated data grouped by GPCR families and subfamilies. This platform allows the user to visualize on-site and to download all the available information.

As already discussed, this work generated a significant amount of data in the form of MD trajectories that can be analysed through a plethora of different methods. Because of this, we are planning on future analyses, including better ways of sharing the trajectories, possibly by clustering-based methods.

Despite the amount of data produced and analysed in this work, we have barely scratched the surface of what a systematic application of this protocol can achieve. By expanding the current dataset with more receptors and with more structures of each, we would be able to reveal more statistically significant data. Furthermore, our dataset was heavily biased towards active structures due to the restricted limitation of presenting a co-crystalised cholesterol. Now that the method has been extensively tested, removing this limitation and processing structures in intermediate and inactive states in future studies could reveal interesting insights about activation state-dependent cholesterol binding.

References

(1) *Urbino-CAMD-Lab/CHAMP: This project is aimed to detect and analyze cholesterol hot-spots in the protein-membrane interface. Mostly tested in class A G Protein-Coupled Receptors.* https://github.com/Urbino-CAMD-Lab/CHAMP (accessed 2026-04-07).

(2) Van Meer, G.; Voelker, D. R.; Feigenson, G. W. Membrane Lipids: Where They Are and How They Behave. *Nat. Rev. Mol. Cell Biol.* **2008**, *9* (2), 112–124. https://doi.org/10.1038/NRM2330.

(3) Levental, I.; Lyman, E. Regulation of Membrane Protein Structure and Function by Their Lipid Nano-Environment. *Nature Reviews Molecular Cell Biology 2022 24:2* **2022**, *24* (2), 107–122. https://doi.org/10.1038/s41580-022-00524-4.

(4) Andersen, O. S.; Koeppe, R. E. Bilayer Thickness and Membrane Protein Function: An Energetic Perspective. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36* (Volume 36, 2007), 107–130. https://doi.org/10.1146/annurev.biophys.36.040306.132643.

(5) Enkavi, G.; Javanainen, M.; Kulig, W.; Róg, T.; Vattulainen, I. Multiscale Simulations of Biological Membranes: The Challenge To Understand Biological Phenomena in a Living Substance. *Chem. Rev.* **2019**, *119* (9), 5607–5774. https://doi.org/10.1021/acs.chemrev.8b00538.

(6) Yeagle, P. L. Cholesterol and the Cell Membrane. *Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes* **1985**, *822* (3–4), 267–287. https://doi.org/10.1016/0304-4157(85)90011-5.

(7) Marsh, D. Cholesterol-Induced Fluid Membrane Domains: A Compendium of Lipid-Raft Ternary Phase Diagrams. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2009**, *1788* (10), 2114–2123. https://doi.org/10.1016/J.BBAMEM.2009.08.004.

(8) Casarotto, P. C.; Girych, M.; Fred, S. M.; Kovaleva, V.; Moliner, R.; Enkavi, G.; Biojone, C.; Cannarozzo, C.; Sahu, M. P.; Kaurinkoski, K.; Brunello, C. A.; Steinzeig, A.; Winkel, F.; Patil, S.; Vestring, S.; Serchov, T.; Diniz, C. R. A. F.; Laukkanen, L.; Cardon, I.; Antila, H.; Rog, T.; Piepponen, T. P.; Bramham, C. R.; Normann, C.; Lauri, S. E.; Saarma, M.; Vattulainen, I.; Castrén, E. Antidepressant Drugs Act by Directly Binding to TRKB Neurotrophin Receptors. *Cell* **2021**, *184* (5), 1299-1313.e19. https://doi.org/10.1016/J.CELL.2021.01.034.

(9) Chen, Y.; Zhu, Y.; Li, X.; Gao, W.; Zhen, Z.; Dong, D.; Huang, B.; Ma, Z.; Zhang, A.; Song, X.; Ma, Y.; Guo, C.; Zhang, F.; Huang, Z. Cholesterol Inhibits TCR Signaling by Directly Restricting TCR-CD3 Core Tunnel Motility. *Mol. Cell* **2022**, *82* (7), 1278-1287.e5. https://doi.org/10.1016/j.molcel.2022.02.017.

(10) Lorente, J. S.; Sokolov, A. V.; Ferguson, G.; Schiöth, H. B.; Hauser, A. S.; Gloriam, D. E. GPCR Drug Discovery: New Agents, Targets and Indications. *Nature Reviews Drug Discovery 2025 24:6* **2025**, *24* (6), 458–479. https://doi.org/10.1038/s41573-025-01139-y.

(11) Latorraca, N. R.; Venkatakrishnan, A. J.; Dror, R. O. GPCR Dynamics: Structures in Motion. *Chem. Rev.* **2017**, *117* (1), 139–155. https://doi.org/10.1021/ACS.CHEMREV.6B00177.

(12) Elisi, G. M.; Bottegoni, G. Impact of G Protein-Coupled Receptor Conformation on Signaling Bias: Integrating Simulations and Biophysical Experiments. *Pharmacol. Ther.* **2025**, *274*, 108905. https://doi.org/10.1016/J.PHARMTHERA.2025.108905.

(13) Slosky, L. M.; Caron, M. G.; Barak, L. S. Biased Allosteric Modulators: New Frontiers in GPCR Drug Discovery. *Trends Pharmacol. Sci.* **2021**, *42* (4), 283–299. https://doi.org/10.1016/J.TIPS.2020.12.005.

(14) Pucadyil, T. J.; Chattopadhyay, A. Effect of Cholesterol on Lateral Diffusion of Fluorescent Lipid Probes in Native Hippocampal Membranes. *Chem. Phys. Lipids* **2006**, *143* (1–2), 11–21. https://doi.org/10.1016/J.CHEMPHYSLIP.2006.04.003.

(15) Oates, J.; Watts, A. Uncovering the Intimate Relationship between Lipids, Cholesterol and GPCR Activation. *Curr. Opin. Struct. Biol.* **2011**, *21* (6), 802–807. https://doi.org/10.1016/J.SBI.2011.09.007.

(16) Kiriakidi, S.; Kolocouris, A.; Liapakis, G.; Ikram, S.; Durdagi, S.; Mavromoustakos, T. Effects of Cholesterol on GPCR Function: Insights from Computational and Experimental Studies. *Adv. Exp. Med. Biol.* **2019**, *1135*, 89–103. https://doi.org/10.1007/978-3-030-14265-0_5.

(17) Prasanna, X.; Sengupta, D.; Chattopadhyay, A. Cholesterol-Dependent Conformational Plasticity in GPCR Dimers. *Sci. Rep.* **2016**, *6* (1), 31858-. https://doi.org/10.1038/SREP31858;SUBJMETA.

(18) Fantini, J.; Epand, R. M.; Barrantes, F. J. Cholesterol-Recognition Motifs in Membrane Proteins. In *Advances in Experimental Medicine and Biology*; Springer New York LLC, 2019; Vol. 1135, pp 3–25. https://doi.org/10.1007/978-3-030-14265-0_1.

(19) Taghon, G. J.; Rowe, J. B.; Kapolka, N. J.; Isom, D. G. Predictable Cholesterol Binding Sites in GPCRs Lack Consensus Motifs. *Structure* **2021**, *29* (5), 499-506.e3. https://doi.org/10.1016/j.str.2021.01.004.

(20) Grouleff, J.; Irudayam, S. J.; Skeby, K. K.; Schiøtt, B. The Influence of Cholesterol on Membrane Protein Structure, Function, and Dynamics Studied by Molecular Dynamics Simulations. *Biochimica et Biophysica Acta - Biomembranes*. Elsevier B.V. September 1, 2015, pp 1783–1795. https://doi.org/10.1016/j.bbamem.2015.03.029.

(21) 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 Å Structure of the Human B2-Adrenergic Receptor. *Structure* **2008**, *16* (6), 897–905. https://doi.org/10.1016/j.str.2008.05.001.

(22) Gimpl, G. Interaction of G Protein Coupled Receptors and Cholesterol. *Chem. Phys. Lipids* **2016**, *199*, 61–73. https://doi.org/10.1016/J.CHEMPHYSLIP.2016.04.006.

(23) Zhang, M.; Gui, M.; Wang, Z. F.; Gorgulla, C.; Yu, J. J.; Wu, H.; Sun, Z. yu J.; Klenk, C.; Merklinger, L.; Morstein, L.; Hagn, F.; Plückthun, A.; Brown, A.; Nasr, M. L.; Wagner, G. Cryo-EM Structure of an Activated GPCR-G Protein Complex in Lipid Nanodiscs. *Nat. Struct. Mol. Biol.* **2021**, *28* (3), 258–267. https://doi.org/10.1038/S41594-020-00554-6.

(24) Brown, L. S.; Ladizhansky, V. Membrane Proteins in Their Native Habitat as Seen by Solid-State NMR Spectroscopy. *Protein Science* **2015**, *24* (9), 1333–1346. https://doi.org/10.1002/PRO.2700;CTYPE:STRING:JOURNAL.

(25) Zidovetzki, R.; Levitan, I. Use of Cyclodextrins to Manipulate Plasma Membrane Cholesterol Content: Evidence, Misconceptions and Control Strategies. *Biochim. Biophys. Acta Biomembr.* **2007**, *1768* (6), 1311–1324. https://doi.org/10.1016/j.bbamem.2007.03.026.

(26) Understanding Molecular Simulation. *Understanding Molecular Simulation* **2002**. https://doi.org/10.1016/b978-0-12-267351-1.x5000-7.

(27) Allen, M. P.; Tildesley, D. J. Computer Simulation of Liquids: Second Edition. *Computer Simulation of Liquids: Second Edition* **2017**, 1–626. https://doi.org/10.1093/oso/9780198803195.001.0001.

(28) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* **1967**, *159* (1), 98. https://doi.org/10.1103/PhysRev.159.98.

(29) Bayly, C. I.; Merz, K. M.; Ferguson, D. M.; Cornell, W. D.; Fox, T.; Caldwell, J. W.; Kollman, P. A.; Cieplak, P.; Gould, I. R.; Spellmeyer, D. C. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **2002**, *117* (19), 5179–5197. https://doi.org/10.1021/ja00124a002.

(30) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(31) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4* (2), 187–217. https://doi.org/10.1002/jcc.540040211.

(32) Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, A. D.; Pastor, R. W. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *Journal of Physical Chemistry B* **2010**, *114* (23), 7830–7843. https://doi.org/10.1021/jp101759q.

(33) 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. https://doi.org/10.1021/ja9621760.

(34) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. https://doi.org/10.1063/1.464397.

(35) Karplus, M. Molecular Dynamics Simulations of Biomolecules. *Acc. Chem. Res.* **2002**, *35* (6), 321–323. https://doi.org/10.1021/ar020082r.

(36) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41* (1), 429–452. https://doi.org/10.1146/annurev-biophys-042910-155245.

(37) Cang, X.; Du, Y.; Mao, Y.; Wang, Y.; Yang, H.; Jiang, H. Mapping the Functional Binding Sites of Cholesterol in B2-Adrenergic Receptor by Long-Time Molecular Dynamics Simulations. *J. Phys. Chem. B* **2013**, *117* (4), 1085–1094. https://doi.org/10.1021/JP3118192.

(38) Lee, J. Y.; Patel, R.; Lyman, E. Ligand-Dependent Cholesterol Interactions with the Human A2A Adenosine Receptor. *Chem. Phys. Lipids* **2013**, *169*, 39–45. https://doi.org/10.1016/J.CHEMPHYSLIP.2013.02.002.

(39) Guixà-González, R.; Albasanz, J. L.; Rodriguez-Espigares, I.; Pastor, M.; Sanz, F.; Martí-Solano, M.; Manna, M.; Martinez-Seara, H.; Hildebrand, P. W.; Martín, M.; Selent, J. Membrane Cholesterol Access into a G-Protein-Coupled Receptor. *Nat. Commun.* **2017**, *8*, 14505. https://doi.org/10.1038/NCOMMS14505.

(40) Vanegas, M. J.; Gómez, S.; Cappelli, C.; Miscione, G. Pietro. Exploring Membrane Cholesterol Binding to the CB1 Receptor: A Computational Perspective. *Journal of Physical Chemistry B* **2025**, *129* (18), 4350–4365. https://doi.org/10.1021/acs.jpcb.4c08076.

(41) Weiser, B. P.; Salari, R.; Eckenhoff, R. G.; Brannigan, G. Computational Investigation of Cholesterol Binding Sites on Mitochondrial VDAC. *Journal of Physical Chemistry B* **2014**, *118* (33), 9852–9860. https://doi.org/10.1021/JP504516A/SUPPL_FILE/JP504516A_SI_005.MPG.

(42) Alvarez-Garcia, D.; Barril, X.; Xxiii, J.; Spain, B. Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. *J. Med. Chem.* **2014**, *57* (20), 8530–8539. https://doi.org/10.1021/jm5010418.

(43) Chan, W. K. B.; Carlson, H. A.; Traynor, J. R. Application of Mixed-Solvent Molecular Dynamics Simulations for Prediction of Allosteric Sites on G Protein-Coupled Receptors. *Mol. Pharmacol.* **2023**, *103* (5), 274–285. https://doi.org/10.1124/molpharm.122.000612.

(44) Borges-Araújo, L.; Borges-Araújo, A. C.; Ozturk, T. N.; Ramirez-Echemendia, D. P.; Fábián, B.; Carpenter, T. S.; Thallmair, S.; Barnoud, J.; Ingólfsson, H. I.; Hummer, G.; Tieleman, D. P.; Marrink, S. J.; Souza, P. C. T.; Melo, M. N. Martini 3 Coarse-Grained Force Field for Cholesterol. *J. Chem. Theory Comput.* **2023**, *19* (20), 7387–7404. https://doi.org/10.1021/acs.jctc.3c00547.

(45) Genheden, S.; Essex, J. W.; Lee, A. G. G Protein Coupled Receptor Interactions with Cholesterol Deep in the Membrane. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2017**, *1859* (2), 268–281. https://doi.org/10.1016/J.BBAMEM.2016.12.001.

(46) Rouviere, E.; Arnarez, C.; Yang, L.; Lyman, E. Identification of Two New Cholesterol Interaction Sites on the A2A Adenosine Receptor. *Biophys. J.* **2017**, *113* (11), 2415–2424. https://doi.org/10.1016/J.BPJ.2017.09.027.

(47) Hedger, G.; Koldsø, H.; Chavent, M.; Siebold, C.; Rohatgi, R.; Sansom, M. S. P. Cholesterol Interaction Sites on the Transmembrane Domain of the Hedgehog Signal Transducer and Class F G Protein-Coupled Receptor Smoothened. *Structure* **2019**, *27* (3), 549-559.e2. https://doi.org/10.1016/J.STR.2018.11.003.

(48) Song, W.; Yen, H. Y.; Robinson, C. V.; Sansom, M. S. P. State-Dependent Lipid Interactions with the A2a Receptor Revealed by MD Simulations Using In Vivo-Mimetic Membranes. *Structure(London, England:1993)* **2019**, *27* (2), 392. https://doi.org/10.1016/J.STR.2018.10.024.

(49) Kolan, D.; Fonar, G.; Samson, A. O. Elastic Network Normal Mode Dynamics Reveal the GPCR Activation Mechanism. *Proteins: Structure, Function and Bioinformatics* **2014**, *82* (4), 579–586. https://doi.org/10.1002/prot.24426.

(50) Aranda-García, D.; Stepniewski, T. M.; Torrens-Fontanals, M.; García-Recio, A.; Lopez-Balastegui, M.; Medel-Lacruz, B.; Morales-Pastor, A.; Peralta-García, A.; Dieguez-Eceolaza, M.; Sotillo-Nuñez, D.; Ding, T.; Drabek, M.; Jacquemard, C.; Jakowiecki, J.; Jespers, W.; Jiménez-Rosés, M.; Jun-Yu-Lim, V.; Nicoli, A.; Orzel, U.; Shahraki, A.; Tiemann, J. K. S.; Ledesma-Martin, V.; Nerín-Fonz, F.; Suárez-Dou, S.; Canal, O.; Pándy-Szekeres, G.; Mao, J.; Gloriam, D. E.; Kellenberger, E.; Latek, D.; Guixà-González, R.; Gutiérrez-de-Terán, H.; Tikhonova, I. G.; Hildebrand, P. W.; Filizola, M.; Babu, M. M.; Di Pizio, A.; Filipek, S.; Kolb, P.; Cordomi, A.; Giorgino, T.; Marti-Solano, M.; Selent, J. Large

Scale Investigation of GPCR Molecular Dynamics Data Uncovers Allosteric Sites and Lateral Gateways. *Nature Communications* **2025**, *16* (1). https://doi.org/10.1038/s41467-025-57034-y.

(51) Ferraro, M.; Masetti, M.; Recanatini, M.; Cavalli, A.; Bottegoni, G. Mapping Cholesterol Interaction Sites on Serotonin Transporter through Coarse-Grained Molecular Dynamics. *PLoS One* **2016**, *11* (12). https://doi.org/10.1371/journal.pone.0166196.

(52) Yang, D.; Gouaux, E. *Illumination of Serotonin Transporter Mechanism and Role of the Allosteric Site*; 2021; Vol. 7. https://www.science.org.

(53) Yang, X.; Wang, X.; Xu, Z.; Wu, C.; Zhou, Y.; Wang, Y.; Lin, G.; Li, K.; Wu, M.; Xia, A.; Liu, J.; Cheng, L.; Zou, J.; Yan, W.; Shao, Z.; Yang, S. Molecular Mechanism of Allosteric Modulation for the Cannabinoid Receptor CB1. *Nat. Chem. Biol.* **2022**, *18* (8). https://doi.org/10.1038/s41589-022-01038-y.

(54) 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. https://doi.org/10.1093/BIOINFORMATICS/BTK023.

(55) *openmm/pdbfixer: PDBFixer fixes problems in PDB files*. https://github.com/openmm/pdbfixer (accessed 2026-01-28).

(56) Kroon, P. C.; Grunewald, F.; Barnoud, J.; van Tilburg, M.; Brasnett, C.; de Souza, P. C. T.; Wassenaar, T. A.; Marrink, S.-J. J. Martinize2 and Vermouth: Unified Framework for Topology Generation. June 23, 2025. https://doi.org/10.7554/eLife.90627.3.

(57) Wassenaar, T. A.; Ingólfsson, H. I.; Böckmann, R. A.; Tieleman, D. P.; Marrink, S. J. Computational Lipidomics with Insane: A Versatile Tool for Generating Custom Membranes for Molecular Simulations. *J. Chem. Theory Comput.* **2015**, *11* (5), 2144–2155. https://doi.org/10.1021/acs.jctc.5b00209.

(58) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindah, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/J.SOFTX.2015.06.001.

(59) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *Journal of Chemical Physics* **2007**, *126* (1). https://doi.org/10.1063/1.2408420/186581.

(60) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690. https://doi.org/10.1063/1.448118.

(61) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190. https://doi.org/10.1063/1.328693.

(62) Gowers, R.; Linke, M.; Barnoud, J.; Reddy, T.; Melo, M.; Seyler, S.; Domański, J.; Dotson, D.; Buchoux, S.; Kenney, I.; Beckstein, O. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proceedings of the 15th Python in Science Conference* **2016**, 98–105. https://doi.org/10.25080/MAJORA-629E541A-00E.

(63) 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. https://doi.org/10.1002/JCC.21787/FULL.

(64) Smith, P.; Lorenz, C. D. LiPyphilic: A Python Toolkit for the Analysis of Lipid Membrane Simulations. *J. Chem. Theory Comput.* **2021**, *17* (9), 5907–5919. https://doi.org/10.1021/ACS.JCTC.1C00447.

(65) 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. https://doi.org/10.1093/NAR/GKX1109.

(66) Decherchi, S.; Spitaleri, A.; Stone, J.; Rocchia, W. NanoShaper–VMD Interface: Computing and Visualizing Surfaces, Pockets and Channels in Molecular Systems. *Bioinformatics* **2019**, *35* (7), 1241–1243. https://doi.org/10.1093/BIOINFORMATICS/BTY761.

(67) Bartocci, A.; Grazzi, A.; Awad, N.; Corringer, P. J.; Souza, P. C. T.; Cecchini, M. A Millisecond Coarse-Grained Simulation Approach to Decipher Allosteric Cannabinoid Binding at the Glycine Receptor A1. *Nature Communications 2024 15:1* **2024**, *15* (1), 9040-. https://doi.org/10.1038/s41467-024-53098-4.

(68) Thorsen, T.; Kulkarni, Y.; Sykes, D.; Bøggild, A.; Drace, T.; Hompluem, P.; Iliopoulos-Tsoutsouvas, C.; Nikas, S.; Daver, H.; Makriyannis, A.; Nissen, P.; Gajhede, M.; Veprintsev, D.; Boesen, T.; Kastrup, J.; Gloriam, D. Structural Basis of Δ9-THC Analog Activity at the Cannabinoid 1 Receptor. **2024**. https://doi.org/10.21203/RS.3.RS-4277209/V1.

(69) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. Macromodel—an Integrated Software System for Modeling Organic and Bioorganic Molecules Using Molecular Mechanics. *J. Comput. Chem.* **1990**, *11* (4), 440–467. https://doi.org/10.1002/JCC.540110405.

(70) Watts, K. S.; Dalal, P.; Tebben, A. J.; Cheney, D. L.; Shelley, J. C. Macrocycle Conformational Sampling with Macromodel. *J. Chem. Inf. Model.* **2014**, *54* (10), 2680–2696. https://doi.org/10.1021/CI5001696.

(71) Ansell, T. B.; Curran, L.; Horrell, M. R.; Pipatpolkai, T.; Letham, S. C.; Song, W.; Siebold, C.; Stansfeld, P. J.; Sansom, M. S. P.; Corey, R. A. Relative Affinities of Protein–Cholesterol Interactions from Equilibrium Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2021**, *17* (10), 6548–6558. https://doi.org/10.1021/acs.jctc.1c00547.

(72) *Welcome to the GROMACS documentation! — GROMACS documentation*. https://manual.gromacs.org/ (accessed 2026-01-28).

(73) Marrink, S. J.; Corradi, V.; Souza, P. C. T.; Ingólfsson, H. I.; Tieleman, D. P.; Sansom, M. S. P. Computational Modeling of Realistic Cell Membranes. *Chemical Reviews*. American Chemical Society May 8, 2019, pp 6184–6226. https://doi.org/10.1021/acs.chemrev.8b00460.

(74) Ferraro, M.; Masetti, M.; Recanatini, M.; Cavalli, A.; Bottegoni, G. Modeling Lipid Raft Domains Containing a Mono-Unsaturated Phosphatidylethanolamine Species. *RSC Adv.* **2015**, *5* (47), 37102–37111. https://doi.org/10.1039/c5ra02196k.

(75) Decherchi, S.; Bottegoni, G.; Spitaleri, A.; Rocchia, W.; Cavalli, A. BiKi Life Sciences: A New Suite for Molecular Dynamics and Related Methods in Drug Discovery. *J. Chem. Inf. Model.* **2018**, *58* (2), 219–224. https://doi.org/10.1021/ACS.JCIM.7B00680.

(76) La Sala, G.; Decherchi, S.; De Vivo, M.; Rocchia, W. Allosteric Communication Networks in Proteins Revealed through Pocket Crosstalk Analysis. *ACS Cent. Sci.* **2017**, *3* (9), 949–960. https://doi.org/10.1021/ACSCENTSCI.7B00211.

(77) Coleman, J. A.; Green, E. M.; Gouaux, E. X-Ray Structures and Mechanism of the Human Serotonin Transporter. *Nature* **2016**, *532* (7599), 334–339. https://doi.org/10.1038/nature17629.

(78) Coleman, J. A.; Gouaux, E. Structural Basis for Recognition of Diverse Antidepressants by the Human Serotonin Transporter. *Nature Structural & Molecular Biology 2018 25:2* **2018**, *25* (2), 170–175. https://doi.org/10.1038/s41594-018-0026-8.

(79) Oddi, S.; Dainese, E.; Fezza, F.; Lanuti, M.; Barcaroli, D.; De Laurenzi, V.; Centonze, D.; MacCarrone, M. Functional Characterization of Putative Cholesterol Binding Sequence (CRAC) in Human Type-1 Cannabinoid Receptor. *J. Neurochem.* **2011**, *116* (5), 858–865. https://doi.org/10.1111/j.1471-4159.2010.07041.x.

(80) 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. *Methods in Neurosciences* **1995**, *25* (C), 366–428. https://doi.org/10.1016/S1043-9471(05)80049-7.

(81) Sanchez-Reyes, O. B.; Cooke, A. L. G.; Tranter, D. B.; Rashid, D.; Eilers, M.; Reeves, P. J.; Smith, S. O. G Protein-Coupled Receptors Contain Two Conserved Packing Clusters. *Biophys. J.* **2017**, *112* (11), 2315–2326. https://doi.org/10.1016/j.bpj.2017.04.051.

(82) Thompson, A. A.; Liu, J. J.; Chun, E.; Wacker, D.; Wu, H.; Cherezov, V.; Stevens, R. C. GPCR Stabilization Using the Bicelle-like Architecture of Mixed Sterol-Detergent Micelles. *Methods* **2011**, *55* (4), 310–317. https://doi.org/10.1016/j.ymeth.2011.10.011.

(83) Ahn, K. H.; Mahmoud, M. M.; Kendall, D. A. Allosteric Modulator ORG27569 Induces CB1 Cannabinoid Receptor High Affinity Agonist Binding State, Receptor Internalization, and Gi Protein-Independent ERK1/2 Kinase Activation. *Journal of Biological Chemistry* **2012**, *287* (15), 12070–12082. https://doi.org/10.1074/jbc.M111.316463.

(84) Shao, Z.; Yan, W.; Chapman, K.; Ramesh, K.; Ferrell, A. J.; Yin, J.; Wang, X.; Xu, Q.; Rosenbaum, D. M. Structure of an Allosteric Modulator Bound to the CB1 Cannabinoid Receptor. *Nat. Chem. Biol.* **2019**, *15* (12), 1199–1205. https://doi.org/10.1038/S41589-019-0387-2.

(85) Melo, M. N.; Ingólfsson, H. I.; Marrink, S. J. Parameters for Martini Sterols and Hopanoids Based on a Virtual-Site Description. *J. Chem. Phys.* **2015**, *143* (24). https://doi.org/10.1063/1.4937783.

(86) Fábián, B.; Thallmair, S.; Hummer, G. Optimal Bond Constraint Topology for Molecular Dynamics Simulations of Cholesterol. *J. Chem. Theory Comput.* **2023**, *19* (5), 1592–1601. https://doi.org/10.1021/ACS.JCTC.2C01032.

(87) Qi, Y.; Ingólfsson, H. I.; Cheng, X.; Lee, J.; Marrink, S. J.; Im, W. CHARMM-GUI Martini Maker for Coarse-Grained Simulations with the Martini Force Field. *J. Chem. Theory Comput.* **2015**, *11* (9), 4486–4494. https://doi.org/10.1021/ACS.JCTC.5B00513.

(88) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.* **2008**, *29* (11), 1859–1865. https://doi.org/10.1002/jcc.20945.

(89) Hua, T.; Vemuri, K.; Nikas, S. P.; Laprairie, R. B.; Wu, Y.; Qu, L.; Pu, M.; Korde, A.; Jiang, S.; Ho, J. H.; Han, G. W.; Ding, K.; Li, X.; Liu, H.; Hanson, M. A.; Zhao, S.; Bohn, L. M.; Makriyannis, A.; Stevens, R. C.; Liu, Z. J. Crystal Structures of Agonist-Bound Human Cannabinoid Receptor CB 1. *Nature* **2017**, *547* (7664), 468–471. https://doi.org/10.1038/NATURE23272;TECHMETA.

(90) *Bugs fixed — GROMACS 2022-rc1 documentation*. https://manual.gromacs.org/2022-rc1/release-notes/2022/major/bugs-fixed.html#fixed-slight-inaccuracies-when-using-virtual-sites-with-pressure-coupling (accessed 2026-01-28).

(91) *Topology/Structure Generation – Martini Force Field Initiative*. https://cgmartini.nl/docs/downloads/tools/topology-structure-generation.html (accessed 2026-01-28).

(92) *marrink-lab/vermouth-martinize: Describe and apply transformation on molecular structures and topologies*. https://github.com/marrink-lab/vermouth-martinize (accessed 2026-01-28).

(93) *Tsjerk/Insane: INSert membrANE - A simple, versatile tool for building coarse-grained simulation systems*. https://github.com/Tsjerk/Insane (accessed 2026-01-28).

(94) *Gaussian Cube File Format*. https://paulbourke.net/dataformats/cube/ (accessed 2026-01-28).

(95) Nickolls, J.; Buck, I.; Garland, M.; Nvidia; Skadron, K. Scalable Parallel Programming. *ACM SIGGRAPH 2008 Classes* **2008**, 16. https://doi.org/10.1145/1401132.1401152.

(96) *Multi-Process Service — Multi-Process Service*. https://docs.nvidia.com/deploy/mps/index.html (accessed 2026-01-28).

(97) Brasnett, C.; Duve, T.; Grunewald, Dr. F.; Fidlin, M.; Melo, M. N.; Tsjerk; lgruenewald. Martini-Force-Field-Initiative/Fast_forward: V0.0.3. https://doi.org/10.5281/ZENODO.19353057.

(98) Pereira, G. P.; Alessandri, R.; Domínguez, M.; Araya-Osorio, R.; Grünewald, L.; Borges-Araújo, L.; Wu, S.; Marrink, S. J.; Souza, P. C. T.; Mera-Adasme, R. Bartender: Martini 3 Bonded Terms via Quantum Mechanics-Based Molecular Dynamics. *J. Chem. Theory Comput.* **2024**, *20* (13), 5763–5773. https://doi.org/10.1021/acs.jctc.4c00275.

(99) Hu, G. M.; Mai, T. L.; Chen, C. M. Visualizing the GPCR Network: Classification and Evolution. *Scientific Reports 2017 7:1* **2017**, *7* (1), 15495-. https://doi.org/10.1038/s41598-017-15707-9.

(100) Zhang, M.; Chen, T.; Lu, X.; Lan, X.; Chen, Z.; Lu, S. G Protein-Coupled Receptors (GPCRs): Advances in Structures, Mechanisms and Drug Discovery. *Signal Transduction and Targeted Therapy 2024 9:1* **2024**, *9* (1), 88-. https://doi.org/10.1038/s41392-024-01803-6.

(101) Isberg, V.; De Graaf, C.; Bortolato, A.; Cherezov, V.; Katritch, V.; Marshall, F. H.; Mordalski, S.; Pin, J. P.; Stevens, R. C.; Vriend, G.; Gloriam, D. E. Generic GPCR Residue Numbers - Aligning Topology Maps While Minding the Gaps. *Trends Pharmacol. Sci.* **2015**, *36* (1), 22–31. https://doi.org/10.1016/j.tips.2014.11.001.

(102) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Foon, S. T.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-Resolution Crystal Structure of an Engineered Human B2-Adrenergic G Protein-Coupled Receptor. *Science (1979).* **2007**, *318* (5854), 1258–1265. https://doi.org/10.1126/SCIENCE.1150577;PAGE:STRING:ARTICLE/CHAPTER.

(103) Zhang, K.; Zhang, J.; Gao, Z. G.; Zhang, D.; Zhu, L.; Han, G. W.; Moss, S. M.; Paoletta, S.; Kiselev, E.; Lu, W.; Fenalti, G.; Zhang, W.; Müller, C. E.; Yang, H.; Jiang, H.; Cherezov, V.; Katritch, V.; Jacobson, K. A.; Stevens, R. C.; Wu, B.; Zhao, Q. Structure of the Human P2Y12 Receptor in Complex with an Antithrombotic Drug. *Nature 2014 509:7498* **2014**, *509* (7498), 115–118. https://doi.org/10.1038/nature13083.

(104) Xing, C.; Zhuang, Y.; Xu, T. H.; Feng, Z.; Zhou, X. E.; Chen, M.; Wang, L.; Meng, X.; Xue, Y.; Wang, J.; Liu, H.; McGuire, T. F.; Zhao, G.; Melcher, K.; Zhang, C.; Xu, H. E.; Xie, X. Q. Cryo-EM Structure of the Human Cannabinoid Receptor CB2-Gi Signaling Complex. *Cell* **2020**, *180* (4), 645-654.e13. https://doi.org/10.1016/j.cell.2020.01.007.

(105) Xu, P.; Huang, S.; Zhang, H.; Mao, C.; Zhou, X. E.; Cheng, X.; Simon, I. A.; Shen, D. D.; Yen, H. Y.; Robinson, C. V.; Harpsøe, K.; Svensson, B.; Guo, J.; Jiang, H.; Gloriam, D. E.; Melcher, K.; Jiang, Y.; Zhang, Y.; Xu, H. E. Structural Insights into the Lipid and Ligand Regulation of Serotonin Receptors. *Nature 2021 592:7854* **2021**, *592* (7854), 469–473. https://doi.org/10.1038/s41586-021-03376-8.

(106) Christofidi, M.; Tzortzini, E.; Mavromoustakos, T.; Kolocouris, A. Effects of Membrane Cholesterol on the Structure and Function of Selected Class A GPCRs–Challenges and Future Perspectives. *Biochemistry* **2025**, *64* (19), 4011–4049. https://doi.org/10.1021/ACS.BIOCHEM.5C00145.

(107) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics 2009 10:1* **2009**, *10* (1), 168-. https://doi.org/10.1186/1471-2105-10-168.

(108) Binkowski, T. A.; Naghibzadeh, S.; Liang, J. CASTp: Computed Atlas of Surface Topography of Proteins. *Nucleic Acids Res.* **2003**, *31* (13), 3352. https://doi.org/10.1093/NAR/GKG512.

(109) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap Family of Web Servers for Determining and Characterizing Ligand Binding Hot Spots of Proteins. *Nat. Protoc.* **2015**, *10* (5), 733. https://doi.org/10.1038/NPROT.2015.043.

(110) Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, *49* (2), 377–389. https://doi.org/10.1021/CI800324M.

(111) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. DeepSite: Protein-Binding Site Predictor Using 3D-Convolutional Neural Networks. *Bioinformatics* **2017**, *33* (19), 3036–3042. https://doi.org/10.1093/BIOINFORMATICS/BTX350.

(112) Krivák, R.; Hoksza, D. P2Rank: Machine Learning Based Tool for Rapid and Accurate Prediction of Ligand Binding Sites from Protein Structure. *Journal of Cheminformatics 2018 10:1* **2018**, *10* (1), 39-. https://doi.org/10.1186/S13321-018-0285-8.

(113) Peter, S.; Siragusa, L.; Thomas, M.; Palomba, T.; Cross, S.; O'Boyle, N. M.; Bajusz, D.; Ferenczy, G. G.; Keserű, G. M.; Bottegoni, G.; Bender, B.; Chen, I.; De Graaf, C. Comparative Study of Allosteric GPCR Binding Sites and Their Ligandability Potential. *J. Chem. Inf. Model.* **2024**, *64* (21), 8176–8192. https://doi.org/10.1021/ACS.JCIM.4C00819.

(114) Zhang, D.; Gao, Z. G.; Zhang, K.; Kiselev, E.; Crane, S.; Wang, J.; Paoletta, S.; Yi, C.; Ma, L.; Zhang, W.; Han, G. W.; Liu, H.; Cherezov, V.; Katritch, V.; Jiang, H.; Stevens, R. C.; Jacobson, K. A.; Zhao, Q.; Wu, B. Two Disparate Ligand-Binding Sites in the Human P2Y1 Receptor. *Nature 2015 520:7547* **2015**, *520* (7547), 317–321. https://doi.org/10.1038/nature14287.

(115) Cheng, R. K. Y.; Fiez-Vandal, C.; Schlenker, O.; Edman, K.; Aggeler, B.; Brown, D. G.; Brown, G. A.; Cooke, R. M.; Dumelin, C. E.; Doré, A. S.; Geschwindner, S.; Grebner, C.; Hermansson, N. O.; Jazayeri, A.; Johansson, P.; Leong, L.; Prihandoko, R.; Rappas, M.; Soutter, H.; Snijder, A.; Sundström, L.; Tehan, B.; Thornton, P.; Troast, D.; Wiggin, G.; Zhukov, A.; Marshall, F. H.; Dekker, N. Structural Insight into Allosteric Modulation of

Protease-Activated Receptor 2. *Nature 2017 545:7652* **2017**, *545* (7652), 112–115. https://doi.org/10.1038/nature22309.

(116) Lu, J.; Byrne, N.; Wang, J.; Bricogne, G.; Brown, F. K.; Chobanian, H. R.; Colletti, S. L.; Di Salvo, J.; Thomas-Fowlkes, B.; Guo, Y.; Hall, D. L.; Hadix, J.; Hastings, N. B.; Hermes, J. D.; Ho, T.; Howard, A. D.; Josien, H.; Kornienko, M.; Lumb, K. J.; Miller, M. W.; Patel, S. B.; Pio, B.; Plummer, C. W.; Sherborne, B. S.; Sheth, P.; Souza, S.; Tummala, S.; Vonrhein, C.; Webb, M.; Allen, S. J.; Johnston, J. M.; Weinglass, A. B.; Sharma, S.; Soisson, S. M. Structural Basis for the Cooperative Allosteric Activation of the Free Fatty Acid Receptor GPR40. *Nature Structural & Molecular Biology 2017 24:7* **2017**, *24* (7), 570–577. https://doi.org/10.1038/nsmb.3417.

(117) Sumichika, H.; Sakata, K.; Sato, N.; Takeshita, S.; Ishibuchi, S.; Nakamura, M.; Kamahori, T.; Ehara, S.; Itoh, K.; Ohtsuka, T.; Ohbora, T.; Mishina, T.; Komatsu, H.; Naka, Y. Identification of a Potent and Orally Active Non-Peptide C5a Receptor Antagonist. *Journal of Biological Chemistry* **2002**, *277* (51), 49403–49407. https://doi.org/10.1074/jbc.M209672200.

(118) Liu, X.; Masoudi, A.; Kahsai, A. W.; Huang, L. Y.; Pani, B.; Staus, D. P.; Shim, P. J.; Hirata, K.; Simhal, R. K.; Schwalb, A. M.; Rambarat, P. K.; Ahn, S.; Lefkowitz, R. J.; Kobilka, B. Mechanism of B2AR Regulation by an Intracellular Positive Allosteric Modulator. *Science* **2019**, *364* (6447), 1283–1287. https://doi.org/10.1126/SCIENCE.AAW8981.

(119) Shen, S.; Zhao, C.; Wu, C.; Sun, S.; Li, Z.; Yan, W.; Shao, Z. Allosteric Modulation of G Protein-Coupled Receptor Signaling. *Front. Endocrinol. (Lausanne).* **2023**, *14*, 1137604. https://doi.org/10.3389/FENDO.2023.1137604.

(120) Liu, X.; Kaindl, J.; Korczynska, M.; Stößel, A.; Dengler, D.; Stanek, M.; Hübner, H.; Clark, M. J.; Mahoney, J.; Matt, R. A.; Xu, X.; Hirata, K.; Shoichet, B. K.; Sunahara, R. K.; Kobilka, B. K.; Gmeiner, P. An Allosteric Modulator Binds to a Conformational Hub in the B2 Adrenergic Receptor. *Nature Chemical Biology 2020 16:7* **2020**, *16* (7), 749–755. https://doi.org/10.1038/s41589-020-0549-2.

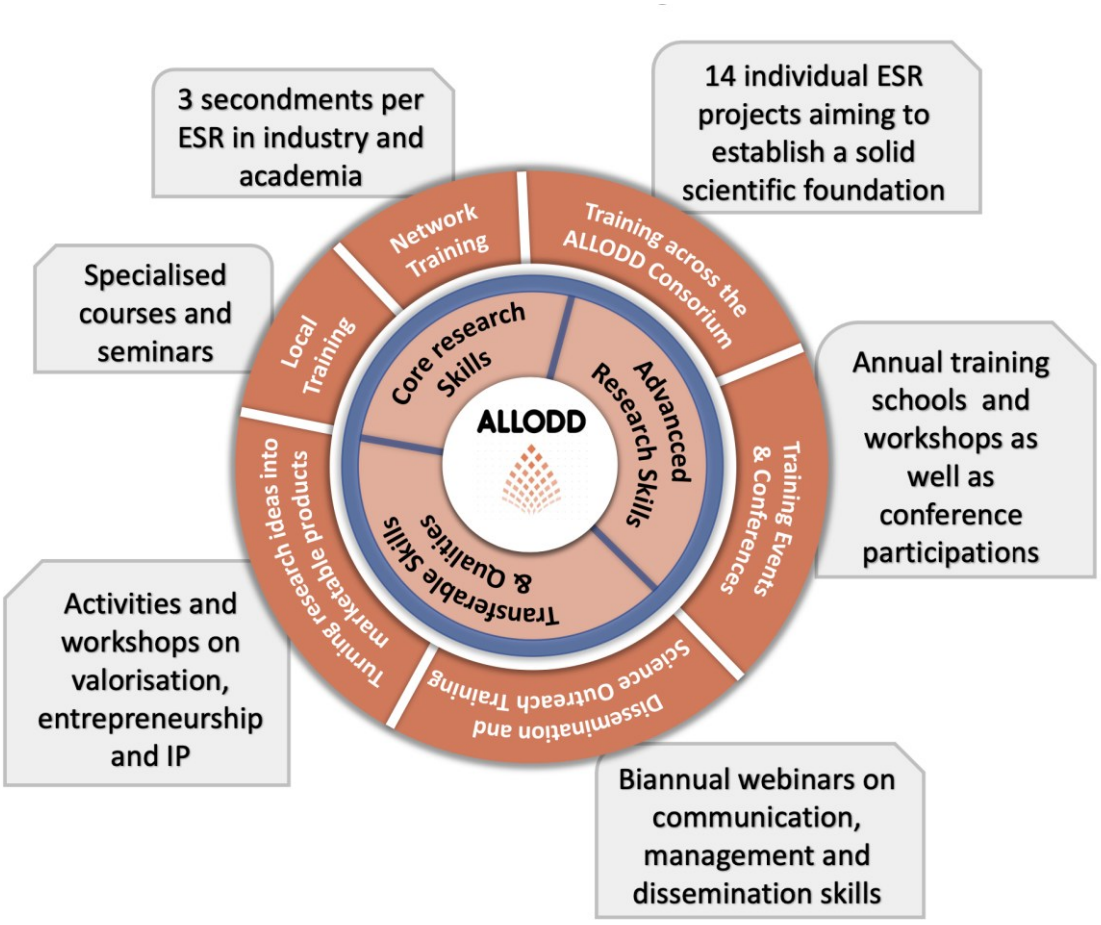
(121) Draper-Joyce, C. J.; Bhola, R.; Wang, J.; Bhattarai, A.; Nguyen, A. T. N.; Cowie-Kent, I.; O'Sullivan, K.; Chia, L. Y.; Venugopal, H.; Valant, C.; Thal, D. M.; Wootten, D.; Panel, N.; Carlsson, J.; Christie, M. J.; White, P. J.; Scammells, P.; May, L. T.; Sexton, P. M.; Danev, R.; Miao, Y.; Glukhova, A.; Imlach, W. L.; Christopoulos, A. Positive Allosteric Mechanisms of Adenosine A1 Receptor-Mediated Analgesia. *Nature 2021 597:7877* **2021**, *597* (7877), 571–576. https://doi.org/10.1038/s41586-021-03897-2.

(122) *Urbino CAMD Lab*. https://github.com/Urbino-CAMD-Lab (accessed 2026-01-29).

(123) Rose, A. S.; Bradley, A. R.; Valasatava, Y.; Duarte, J. M.; Prlic, A.; Rose, P. W. NGL Viewer: Web-Based Molecular Graphics for Large Complexes. *Bioinformatics* **2018**, *34* (21), 3755–3758. https://doi.org/10.1093/BIOINFORMATICS/BTY419.

(124) MerkelDirk. Docker. *Linux Journal* **2014**. https://doi.org/10.5555/2600239.2600241.

Appendix I. Allostery in Drug Discovery (ALLODD)



The **ALLO**stery in **D**rug **Di**scovery (ALLODD) project is a collaboration between 13 academic and industrial organizations with 14 ESR/PhD students available positions. The aim of ALLODD is to train a new generation of scientists to exploit the concept of allostery in drug design, putting together a whole array of technologies to identify and characterize allosteric modulators of protein function that will be applied to therapeutically relevant systems.

Appendix II. CHAMP repository and manual

Description

CHAMP (Cholesterol Hot spot Automated Mapping Protocol) is a python module aimed at detecting potential cholesterol binding sites around membrane proteins. Currently, it is mainly tailored to process class A G Protein-Coupled Receptors, but it can work with other membrane proteins with minimal adaptation. This tool is designed to be executed with minimal user intervention. It only requires a pre-processed, membrane-oriented coordinate file of a membrane protein in PDB format and an input file specifying the parameters for the run. CHAMP relies on the coarse-grained Martini 3 representation to perform molecular dynamics (MD) simulations of the input protein after embedding it in a membrane model which includes cholesterol. Cholesterol hot spots around the protein are detected by analysing the simulated trajectory using two complementary approaches: i) pinpointing statistically relevant cholesterol-contacting residues and ii) extracting cholesterol distribution from the MD simulations. The consensus regions where there is an overlap between statistical outliers and relevant density are labelled as cholesterol hot spots. The protocol returns the identified hot spots in PDB format, with the residues considered statistical outliers, as well as the relevant density points associated with them.

Installation

This package can be installed using the command

pip install CHAMP

Inputs

Pre-processed PDB file:

For this tool to work efficiently and to minimise potential errors that could arise from PDB formatting or the quality of the starting structure, it is mandatory to use a PDB file that has been previously pre-processed. The input structure should be properly aligned within the membrane, all residues should be complete with no missing atoms and only atoms corresponding to the protein – *no ligands or other HETATM* – should be included.

Databases such as **OPM** (https://opm.phar.umich.edu/) have a wide variety of already oriented structures that can be used after removing unnecessary atoms.

Parameters input file:

To run the protocol, several parameters must be defined in advance and provided to the tool. Below is an example .prm file that includes all available parameters that can be configured:

```
run_mode = full
step = system_build/MD_simulation/analysis/merge_data/rank_pockets
pdb_id = 1ABC
starting_structure = /path/to/structure
n_reps = 2
mb_comp = POPC:75,CHOL:25
martinize2_path = martinize2
gmx_path = gmx
gpu = 0
gpcrdb_file_path = /path/to/annotated/structure
```

rank_pockets = True/False
pockets_file = /path/to/pockets/file
bw_notation = True/False
MD_Mode = sequential
output_dir = /output/path (Default ./CHAMP_results)

MANDATORY PARAMETERS

· **run mode** - The protocol can be run in full or individual steps can be selected. Possible values are: full or standalone.

· **step** - Required only if *run_mode=standalone*. Possible values are: system_build, MD_simulation, analysis, merge_data or rank_pockets.

· **pdb_id** - PDB standard 4 character code of the starting structure, e.g. 7fee. This is used for the naming of the output, so the use of upper or lowercase is left to preference.

· **mb_comp** - Desired membrane composition. The tool will use insane.py [1] to create the membrane. The format is very specific and CHAMP will either complain or crash if written incorrectly. Correct format is as shown in the example: LIP1:[ratio],LIP2:[ratio], etc. As in the example POPC:75,CHOL:25. If cholesterol is not selected in the membrane the protocol will not run. On the current version of this software the following phospholipids are supported: POPC, DOPC, DPPC, POPE.

· **gpu** - Required only if a MD simulation is going to be performed. Id of the GPU to be used for MD simulations (available GPUs can be checked via nvidia-smi as long as the CUDA toolkit and NVIDIA(R) drivers are installed). NOTE: This protocol relies on the CUDA acceleration implemented in GROMACS. If no GPU is available the source code in MDsimulation.py corresponding to GROMACS runs have to be properly modified.

· **starting_structure** - Path to the input structure. Either the relative path – *relative to where the script will be executed from* – or the absolute path can be used. We recommend the use of absolute paths to avoid confusions.

OPTIONAL PARAMETERS

· **n_reps** - Nº of replica to create and process (Default: 2).

· **martinize2_path** - Martinize 2, included in the Vermouth package [2] is included in the dependencies of this tool and the executable should already be on the PATH. Therfore, its default value is martinize2.

· **gmx_path** - Path to the executable of GROMACS. Default is gmx, assuming the executable is already in PATH. Note that if GROMACS was compiled with MPI support the executable (unless instructed differently during compilation) is created as gmx_mpi, and therefore this parameter needs to be set as such.

· **bw_notation** - If working with GPCRs there is the possibility of annotating the outliers file with the Ballesteros-Weinstein notation adapted from *GPCRdb* [3]. The annotated structure can be downloaded from their database. Possible values: True or False (Default is False).

· **gpcrdb_file_path** - Required only if bw_notation=True. Path to the GPCRdb annotated structure. It is assumed that the annotation is on the B-factor column of the PDB *–as GPCRdb does–* and passing a structure with a different format will lead to errors or unexpected behaviour.

· **MD_mode** - Decides how the MD simulations will be performed. This version of the protocol only supports the value sequential, in which each replica will be performed after the previous finishes. A future version will include the possibility for advanced systems with multiple GPUs to run replicas in parallel.

· **rank_pockets** - CHAMP offers the possibility of ranking a list of pockets detected by an external software based on their propensity to lodge cholesterol by scoring them using the detected outliers and density points. Possible values: True or False (Default: False).

· **pockets_file** - Required only if rank_pockets=True. Path to the file containing the pockets. It will read pockets in a specific format, so depending on the software utilised the file might need to be adapted. The correct format is a text file, with one pocket per line. Each line should contain the residue IDs conforming the pocket, separated by a single white space. Find an example below, obtained from the Cannabinoid 1 Receptor (PDB:7FEE):

201 202 205 208 209 240 243 244 247 282
165 169 191 192 195 198 199 245 248 249 252
138 141 142 148 154 157 158 161 230 234 237 238
280 284 287 288 291 353 356 357 360 361 364
351 354 355 388

· **output_dir** - Path to the directory where the output will be saved. If the target directory does not exist it will be created. Default: ./CHAMP_results.

Protocol Steps

A full run of CHAMP will run all the steps detailed below in sequential order. Standalone runs of selected steps can be performed as well. The same **.prm** file can be used for different standalone steps just by adjusting the parameter **step**, as the parameters not used in that step will simply be ignored. However, in order to run individual steps, the same directory structure and file naming as used by CHAMP should be used. It is therefore recommended that individual steps are run on output directories created by CHAMP or adapt the structure and file naming of your data to match what the software expects. An example of the directory structure and the output of a mock run on 7FEE on a POPC:75,CHOL:25 membrane can be found in the example_run/ directory on this repository.

1. **SYSTEM BUILD**

This step of the protocol creates the main structure of the output directories and processes the starting structure. The starting PDB is converted to a coarse-grain representation using martinize2 and it is embedded in a membrane of the defined composition with insane.py. CHAMP takes care of copying the necessary inputs and preparing topology files for the simulation step.

2. **MD SIMULATION**

The prepared system undergoes several steps of equilibration before proceeding to the production run.

a) Energy minimization: Using a steepest descent minimizer the system is moved toward a local energy minimum.

b) Equilibration 1 - NVT: Short NVT to thermalize the system to the target temperature (300 K)

c) Equilibration 2 - NPT (constrained): NPT step with protein backbone atoms restrained. Lipids will adjust and equilibrate.

d) Equilibration 3 - NPT (1fs timestep): NPT step simulated using a short timestep to prevent sudden unexpected atom movement when releasing protein restraints.

e) Equilibration 3 - NPT (20fs timestep): NPT step simulated at the target 20fs timestep, same as in the production run.

After the equilibration is completed, CHAMP performs a quality check on the membrane. This allows the user to identify any potential issues arising from the system building and equilibration. To this aim, the membrane thickness, area per lipid and order parameters are calculated. Additionally, a plot showing the partial densities of the membrane components is produced. By comparing obtained values with equivalent ones in the literature, the user can ensure that the produced system is realistic.

The production run last 30 μs, ensuring proper sampling of protein-cholesterol contacts.

3. **ANALYSIS**

CHAMP uses two approaches to detect cholesterol hot spots.

a) **Protein-cholesterol contacts**: Relevant and persistent contacts across all trajectory frames are extracted – *maximum occupancy time, or* t_{max} – and the statistical outliers in the distribution are reported.

b) **Spatial Density Function**: The probability density map of cholesterol around the protein is calculated and the top 0.01% of the data – *cartesian XYZ coordinates where cholesterol is most likely to be found* – is reported.

4. **MERGE REPLICAS**

This step simply iterates over all produced replicas and merges the results, creating a CSV file for all of the statistical outliers found and an XYZ file with the total cholesterol density distribution.

5. **RANK POCKETS**

By providing a formatted file with pockets detected in the simulated structure, CHAMP uses the outliers found and the density points to score and rank these pockets based on their propensity to bind cholesterol. For each outlier found in the residues describing the pocket, the scoring system assigns the pocket 1000 points. For each density point within 6Å of any atom in the pocket the scoring system assigns the pocket 1 point.

Example Run

In example_outputs/ the files created on a mock run with short MD simulations can be found. This example was conducted on the **Cannabinoid 1 Receptor** (PDB: *7FEE*) using a POPC:75,CHOL:25 membrane composition. To avoid overloading the repository with large files, the final production trajectories have been downsampled by keeping only every second frame (i.e., 50% of the original frames). Additionally, the cube file generated by *gmx spatia*l is not included in the example for the same reason.

After the analysis step, the cube file (grid.cube) should be located in {output_dir}/sims/rep_x/analysis/. This file is used to extract the top 0.01% of the volumetric data (corresponding to the highest probability regions) and convert it to Cartesian coordinates. The resulting coordinates are saved in {output_dir}/sims/rep_x/analysis/{system_name}_all.xyz.

Usage

The correct usage of this tool is quite straightforward, as once it is installed it only requires to run

CHAMP prm_file

With CHAMP -h or CHAMP --help a help message with example parameters and its usage will be printed to STDOUT. The installed version of the CHAMP can be checked by running CHAMP -v or CHAMP --version.

CHAMP tools

Together with the main functionality of the package two additional tools are provided that will help on running the protocol.

· CHAMP_tools generate_prm will generate a template **.prm** file that can then be adapted to user's needs. Default name is *CHAMP_template.prm* and will be created where the script is executed. However, custom output name can be given by running CHAMP_tools generate_prm output_name.

· CHAMP_tools orient_protein input_structure.pdb accepts a pdb file and orients it along the Z-axis. This tool provides a rough starting point for simulations when a more refined alternative is not available. Note that this script was designed for GPCRs and it assumes that only the receptor part of the protein (mostly transmembrane) is present in the PDB. Having other extra or intracellular parts of the protein present will cause the orientation of the principal axis of the protein pore to be incorrectly calculated and the resulting orientation will most likely to be shifted.

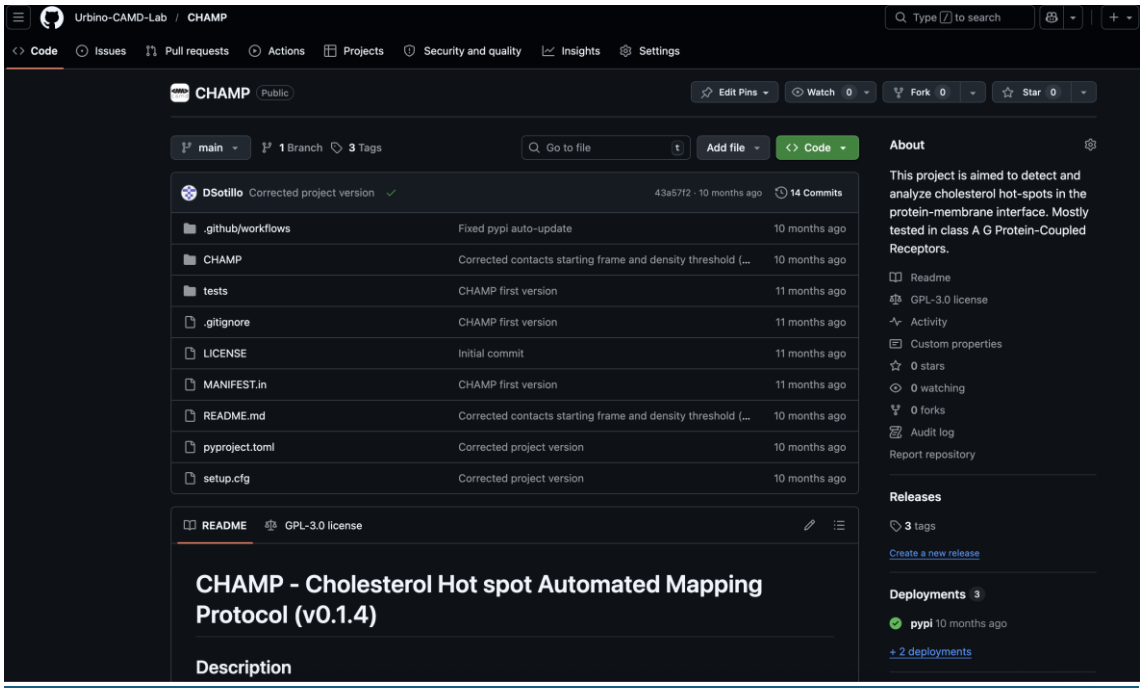
References

[1] Tsjerk. (n.d.). Tsjerk/insane: Insert membrane - a simple, versatile tool for building coarse-grained simulation systems. GitHub. https://github.com/Tsjerk/Insane

[2] Marrink-Lab. (n.d.). Marrink-Lab/Vermouth-Martinize: Describe and apply transformation on molecular structures and Topologies. GitHub. https://github.com/marrink-lab/vermouth-martinize

[3] Generic residue numbering - GPCRdb 1 documentation. (n.d.). https://docs.gpcrdb.org/generic_numbering.html#:~:text=GPCRdb%20numbers%20are%20distinguished%20by,a%20bulge%20following%20position%2055.

Manual and source code available at: https://github.com/Urbino-CAMD-Lab/CHAMP



Appendix III. List of Figures and Tables

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

So many people have had an impact in my life during these years, and I would like to include so many names and personal dedications in this section that if I did it would probably be titled Chapter 8. So, I will try to keep it as short as my feelings allow me.

It is only fitting that I start by personally giving my thanks to my supervisor, Professor Giovanni Bottegoni, the biggest contributor to the success of this work and to my graduation. I lack words of gratitude towards your attitude, closeness and implication with my both my professional and personal situation during these past years. If I come out of this PhD experience any better as a scientist and a professional, it is mostly due to your guidance and to the lessons I will take with me from now on. I hope you are okay with me thinking of you not only as my boss or my supervisor, but also as my mentor, because that is truly how I feel.

Next, I want to address the people that entered my life since I first moved to Urbino. If living abroad and away from my loved ones has been an enjoyable and unforgettable experience it is because of all of you. I am, of course, referring to my lab mates, to whom I would like to dedicate a few words. Vitto, aside from the Prof., you were my first meaningful interaction here in Urbino and I will always be grateful for it. Our common hobbies and your easy-going personality made it extremely easy for me to very quickly consider you a close friend, and our jokes and laughs in the lab is something I will treasure forever. Adriana, you were the next to join us and, despite you being remote working for most of the time, every time you came to Urbino it really felt as if you were actually spending every single day with us. Gian Marco, you know how much I appreciate you, your company and the LOTR references despite your CIA-level skills of tracking and I promise you I was never ignoring your messages (in fact, I was ignoring everyone's because I hardly check on my phone some days). My friendly neighbor Federico, do not worry too much about bothering me my guy, you know I am always glad to help, especially someone as chill and friendly as you, keep that up and be more confident with your scripts. Matilda, even if we had blackout every week or two, we knew we could count on your cheerful and lively presence to lighten up our moods. I wish you the best of luck for the remaining of your PhD and whatever lies ahead, but I know how good of a scientist you are, so I am sure you will not need it. Mattia, you have been an amazing senior and guiding figure. Your closeness and your ability to explain complex things is something I aim to achieve someday, and I cannot state how much I have learned working with you. Lara, you joined towards the end of my period in Urbino, but we got to share some time, even if short. Keep your bright and smiling nature, you will probably need it surrounded by our computational guys. I Finally, I want to take some time for the master students with whom I also shared part of my time. Anna Lisa, Sabrina, Irene, Duyen, Matteo, Vincenzo; thank you guys for the moments we shared and I wish you all the best.

I also want to extend my gratitude words towards every single person involved in the ALLODD consortium, with special thanks to the people in the institutions who hosted and guided me during my periods abroad. Thank you to Maria (the point of origin of all my work), Francesca and Ijen for their help during my time in Cambridge. And thank you to Sigrid and Léxane for letting me join them and for showing and teaching me their own work. Here I also want to include Sonja and Varbina, with whom I shared a long time and have some of the most vibrant and outgoing personalities I have ever encountered, so thank you for being how you are.

On a more personal note, and to conclude, I want to thank all my friends and family back home. Since these words are meant for them, what better way to do it than in my native languages.

A todos mis amigos. Todos sabéis quienes sois y cuánto os quiero, pero aun así quiero que quede por escrito, que no se diga. A la peñita: Joel, Laia, Canek, Marc, Albert, Alberto, Cristina, Juanra, Laura, Marina, Sara(s), Andrea, Carla. Aunque la mayoría lleváis conmigo toda una vida, algunos otros llegasteis más tarde y aun así habéis tenido y tenéis un gran impacto en mi vida. Los más veteranos nos hicimos amigos por simple coincidencia geográfica cuando nuestros cerebros no estaban desarrollados y teníamos la personalidad de una goma de borrar, así que la amistad tan fuerte que aún mantenemos tantos años después es de mis mayores orgullos y de las mejores cosas de mi vida. A mis amigos desde el instituto: Ulises, Edu, Rubén, Carlos, Alberto, Andrés, Óscar y Javi. Nos fuimos conociendo a medida que nos íbamos desarrollando como personas y hemos crecido de niños a adultos juntos. Que sigamos teniendo el mismo humor roto que hace 16 años y que seamos capaces de compartir neuronas en una mente colmena es algo que no dejará de fascinarme jamás y que espero no perder nunca. Als meus bioquímics: Sergi, David, Guillem(s) i Xavi(s). El 2014 vaig decidir entrar a la UAB sense conèixer ningú i sense saber què em trobaria. El que tampoc sabia era que allà trobaria amics per tota la vida, d'aquells que tot i veure'ns cada cent anys mai se sent que hi hagi cap distància quan ens trobem. A mis amigos desde el máster (7 años ya, no veas): Alex, Miguel, Gerard, Cristina, Ana, Armand. Sois todas unas personas tan maravillosas que ni me acuerdo de cuándo exactamente empecé a consideraros tan buenos amigos. Siento que fue conoceros y de repente ya erais una parte enorme de mi vida y que os conozco desde siempre. Y para todos mis grupos de amigos: Durante estos años que he pasado fuera, y a pesar de lo contento y cómodo que estaba en Urbino, contaba los días para la siguiente vuelta a casa para poder veros, aunque tuviese que hacer malabares de horarios para estar con todos. Ha merecido la pena cada segundo de sueño que he perdido, ya que vuestra compañía siempre era el chute de motivación que necesitaba para volver y seguir con mi trabajo. Hay veces que se vuelve muy difícil diferenciar entre amigos y familia, y con amigos como vosotros siento que simplemente tengo una familia muy grande.

Finalmente, a mi familia. A todos mis tíos y a mis primos. Aunque las reuniones familiares sean un absoluto caos a veces siempre disfruto mucho de juntarnos y saber qué hace cada uno. A mis iaios, que siempre se alegraban tanto de verme volver unos días y sonreían como si se hubiesen solucionado todos los problemas del mundo. Seguramente ya lo sabéis, pero es mutuo. Reservar tiempo para veros es lo primero que siempre hacía al volver. A mis avis: mi avi, la mejor persona que he conocido y a mi àvia, siempre fuerte y optimista; os quiero muchísimo a los dos y espero ser algún día la mitad de lo que sois.

A mi hermano Àngel. A pesar de las peleas de pequeños, crecer con un hermano es lo mejor que me pudo pasar. Más aún si además acaba siendo una persona tan trabajadora y que disfruta tanto lo que hace. Estoy muy orgulloso de ti y me alegro mucho de que nos entendamos y nos llevemos tan bien.

Papa y mama, las gracias más grandes son para vosotros, que me habéis dado todo. Si he llegado aquí y si consigo llegar más lejos es por los sacrificios que habéis hecho para que jamás me faltase de nada, para que mi única obligación durante mis primeros años fuera ser feliz y pensar en qué quería para mi futuro. Estoy completamente seguro de que todo lo bueno que hay en mí es la suma de lo mejor de vosotros dos, y deberíais sentir como vuestro cada cumplido que me hagan por cómo soy. Lo que me hace más feliz de llegar hasta aquí es haceros sentir orgullosos.

Os quiero mucho a todos. GRACIAS.