Making protein dynamics FAIR: Research platforms for the collection, dissemination, and analysis of molecular dynamics simulations

Mariona Torrens Fontanals

TESI DOCTORAL UPF / 2021

Directors de la tesi

Dr. Ferran Sanz i Dra. Jana Selent

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Acknowledgments

First and foremost, I would like to thank Dr. Jana Selent, co-supervisor of this thesis, for all her support, for giving me the tools and the motivation to keep improving, and for creating an environment where I feel valued and listened to, which is priceless. I would like to thank Dr. Ferran Sanz as well, who also co-supervised this thesis, for all his help and insightful comments during this project.

Thanks also to Dr. Marta Filizola, for hosting me in her lab at the Icahn School of Medicine at Mount Sinai and taking the time to make sure that my project was successful. And to her team, especially Davide and João, for welcoming me and having the patience to teach me so much.

Importantly, this thesis would have never been possible without the incredible support from the GPCRmd community, especially Ramon and Toni, for their valuable insights, ideas, and talent that contributed to shaping our platforms into something that I am proud of. This would also not have been possible without ERNEST, which has provided so much support, as well as opportunities to meet the rest of the community in amazing meetings.

I would like to thank the GPCR Drug Discovery group, starting with the "old generation" members. Tomek, for always being there to help me with his GPCR wisdom, but especially for all the advice, for always making me laugh, for our infinite jokes and conversations about life in science, and for being the other head of a two-headed monster. Juanma, for all your positivity and for helping me feel at home in our group since the first day. Isma, for guiding me since I was a little master student that did not know much about databases and web development and taking the time to solve my doubts and problems. Also Alejandro, who started the bioinformatics journey with me. Likewise, this PhD thesis would have not been possible without David, Adri, Brian, and the new Alejandro, who joined the team after me but now I cannot imagine our group without them. I would like to give special thanks to David, for the help and contributions in GPCRmd but even more for being a friend I can always talk to. Besides, I am grateful for the

contribution of all the colleagues that temporarily worked with us in the lab, including the degree and master students and visiting researchers, particularly Johanna, Alessandro, Gocky, and Nathalie. I have beautiful memories from the time you visited us.

One of the most valuable aspects of my PhD experience has been the support of my fellow early career researchers in the GRIB, who made going to work a pleasure. Thank you for making me take a break and relax in our mandatory one o'clock lunch break, and for all the crazy moments in the beer sessions. I particularly want to thank Judith, for always having my back, for understanding my workaholic self, and for all the valuable advice about science and about life. And also Coté for all encouragement, help, and support.

I have to thank Alfons, Miguel, Carina, and Chus, the fundamental pillars of the GRIB, who always helped me with a smile and made difficult things (server errors in one case, and grant applications in the other) easier. And for having patience with my multiple (potentially annoying) mails and being extremely helpful, I want to thank Neus and Nàtalia, from the UPF PhD Programme in Biomedicine.

I would also like to acknowledge the funding from the Spanish Ministry of Science, Innovation and Universities (FPU16/01209).

Thinking back to when everything started, I would like to thank my friends Alex, Elsa, Raquel, Ainoa, Judit, Maria, Juanjo, Yáiza, Helena, the "Futuros científicos" that somehow made it into this future. With an especial mention to Alex for all her indispensable support and for being there both in the good and the hard times. Also Alberto, for accompanying me through a big part of my PhD, and still doing it in the distance. Paula, for sharing with me not only therapeutic after-work runs but also therapeutic conversations and advice. Julia, for her motivation and help in the fight against the impostor syndrome. And Blanca, for being there in the even in the gloomy seasons.

And a huge huge thank you to Mario, for always believing in me, for inspiring me, and for being the only one that can make me completely forget about my thesis stress.

Finalment, estaré per sempre agraïda a la meva família, que m'ha donat suport i forces des del primer moment. A la Laura, per inspirar-me, per motivar-me a ser millor i per compartir amb mi tant la passió per la ciència i el coneixement com el camí que ens ha portat fins aquí. A l'Anna, la meva altra gran referència, per ajudar-me a aixecar la vista dels llibres i veure més enllà. Als meus pares, Jordi i Núria, pel seu suport incondicional sense el qual res hauria sigut el mateix, i per recordar-me el valor de cada pas que faig. Als meus avis materns, Carmen i Ernest, per ensenyar-nos a tots el valor de l'esforç, l'empatia, la honestedat i la família. I als meus avis paterns, Carmen i Jordi, sobretot a l'àvia per inspirar-me cada vegada que faig una presentació important.

Abstract

Molecular dynamics (MD) simulations are a well-established technique to characterize the structural motions of biological systems at atomic resolution. However, accessing, viewing, and sharing MD trajectories is typically restricted by large file sizes and the need for specialized software, which limits the audience to which this data is available. The aim of this thesis is to extend the outreach of MD simulations by providing online resources that facilitate the dissemination, visual inspection, and analysis of this data. For that, we present GPCRmd and SCoV2-MD, two online resources focused on proteins with high biomedical interest: G protein-coupled receptors (GPCRs) and the proteome of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), respectively. We also showcase the capabilities of GPCRmd and SCoV2-MD for exploring key aspects of protein dynamics. Overall, these platforms have the potential to promote data "Findability, Accessibility, Interoperability, and Reusability" in the MD field, supporting the FAIR principles for scientific data management.

Resum

Les simulacions de dinàmica molecular (MD, per les seves sigles en anglès) són una tècnica ben establerta per caracteritzar els moviments estructurals de sistemes biològics amb una resolució atòmica. No obstant això, l'accés, la visualització i la compartició de trajectòries de MD solen estar restringits per la gran mida dels seus fitxers i la necessitat de programari especialitzat, que limita el públic al qual estan disponibles aquestes dades. L'objectiu d'aquesta tesi és ampliar la difusió de les simulacions de MD proporcionant recursos en línia que facilitin la compartició, inspecció visual i anàlisi d'aquestes dades. Per això, presentem GPCRmd i SCoV2-MD, dos recursos en línia centrats en proteïnes d'alt interès biomèdic: els receptors acoblats a proteïnes G (coneguts com a GPCRs, per les seves sigles en anglès) i el proteoma del coronavirus 2 de la síndrome respiratòria aguda greu (SARS-CoV-2), respectivament. També mostrem les capacitats de GPCRmd i SCoV2-MD per explorar aspectes clau de la dinàmica de proteïnes. En definitiva, aquestes plataformes tenen el potencial de promoure la cercabilitat, l'accessibilitat, la interoperabilitat i la reutilització de dades en l'àmbit de la MD, donant suport als principis FAIR (acrònim de l'anglès *Findable, Accessible, Interoperable and Reusable*) per a la gestió de dades científiques.

Preface

Molecular dynamics (MD) simulations are a widely established method for exploring time-resolved motions of biological systems at an atomic level. This technique has proven useful to incorporate the missing information on protein flexibility into experimentally solved structures and provide high-resolution details of molecular mechanisms that are difficult to capture with experimental techniques. Despite the value of these insights for many research fields, the access, analysis, and even the visual inspection of MD data is restricted by large file sizes and the need to install specialized software. Ultimately, this limits the audience of MD simulations to specialists in this field. Moreover, the publication of studies involving MD simulations in scientific journals is often not accompanied by the generated atomic trajectories, which reduces transparency, reusability, and the understanding of the described molecular processes.

To maximize the potential of MD research, the data generated should be “Findable, Accessible, Interoperable, and Reusable”, as described by the FAIR principles for scientific data management. A promising step towards this goal is the development of online repositories that facilitate the collection and dissemination of MD data and metadata. When paired with online tools for the visual inspection and analysis of the simulations, these repositories promote the study of MD simulations for experts and non-experts likewise. Consequently, this triggers knowledge exchange and multidisciplinarity, opening new avenues for research.

Due to the technical challenges associated with MD data, the number of such online platforms is limited. For this reason, this thesis was devoted to the implementation of two online resources for the collection, dissemination, and analysis of MD simulations. Each of the developed platforms is focused on a group of proteins with particular biomedical interest: G protein-coupled receptors (GPCRs) and the proteome of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2).

On the one hand, GPCRs are a major class of drug targets, and modulating their signaling can produce a wide range of

physiological outcomes. Their functionality is highly determined by their flexibility and ability to transition between distinct conformations. Thus, exploring the structural dynamics of GPCRs is crucial for understanding their molecular mechanisms, as well as for the rational design of new ligands targeting them. For that, we developed GPCRmd (www.gpcrmd.org), a community-driven online resource that provides access to a vast number of GPCR MD simulations, as well as the associated metadata. To simplify the inspection of the data, GPCRmd is equipped with a comprehensive set of intuitive online tools for the interactive visualization and analysis of MD simulations. Moreover, it includes a meta-analysis tool to compare and cluster the available simulations, or a subset of interest, based on their interaction patterns.

On the other hand, the study of the proteome of the SARS-CoV-2, the causative agent of the Coronavirus disease 2019 (COVID-19) pandemic, is crucial for the fight against this disease. Capturing the structural dynamics of the viral proteins can provide a clearer picture of the molecular processes in which they are involved, including cell infection, evasion of the host's immune system, and replication. This information can also be helpful to understand the impact of amino acid substitutions found in emerging viral variants and to find new therapeutic strategies. Thus, our second platform, SCoV2-MD (www.scov2-md.org), aims to integrate, cross-reference, and share MD dynamics data and metadata of the SARS-CoV-2 proteome. As GPCRmd, this resource includes visualization and analysis tools to explore the simulation data. Additionally, an important asset of SCoV2-MD is that it provides tools to analyze the functional impact of variant substitutions that have emerged during the progression of the pandemic, which is achieved based on a combination of static and MD-derived descriptors.

All in all, with the platforms presented in this thesis, we aim to make a step forward towards reproducibility and transparent dissemination in the field of MD simulations. Ultimately, our goal is to broaden the outreach of protein dynamics to researchers of all fields.

Table of Contents

1. INTRODUCTION

1.1. Biological background

This thesis tackles two groups of proteins with particular biomedical interest. First, we introduce G protein-coupled receptors (GPCRs), which are one of the most successful therapeutic targets for a broad spectrum of diseases. Then, we examine the proteome of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the causative agent of the Coronavirus disease 2019 (COVID-19) pandemic, whose understanding is crucial for the fight against this disease. In both cases, a full comprehension of the physiology and pharmacology of these proteins requires the study of their structural dynamics and, consequently, the effective dissemination of this information.

1.1.1. G protein-coupled receptors (GPCRs)

GPCRs are one of the largest families of membrane protein in eukaryotes, with over 800 identified members in the human proteome[1]. Such diversity allows GPCRs to respond to a wide range of extracellular stimuli including light, changes of pressure, odorants, neurotransmitters, hormones, and metabolites, among others. These stimuli are able to trigger GPCRs to initiate intracellular signaling cascades which are involved in many physiological processes. In fact, signal transduction by GPCRs is fundamental for most physiological processes in the human body, spanning from vision, smell, and taste to neurological, cardiovascular, endocrine, and reproductive functions[2].

Because of their abundance, GPCRs are also involved in a wide variety of human diseases. Thus, GPCRs have been of long-standing interest as pharmacological targets, and more than 30% of all drugs approved by the United States Food and Drug Administration (FDA) target a GPCR, accounting for a global sales volume of over 180 billion US dollars annually[3]. Still, the drugs approved so far target only 27% of the human non-olfactory GPCRs, indicating that this protein family still has immense potential for future drug development[4,5].

Human GPCRs can be classified into six main families (or classes) based on sequence homology and phylogenetic analyses: rhodopsin (A), secretin (B1), adhesion (B2), glutamate (C), frizzled (F), and taste 2[6,7]. Among them, class A is by far the largest, including important receptors like the archetypal GPCR rhodopsin, as well as olfactory, dopamine, opioid, adrenergic, chemokine, angiotensin, histamine, and adenosine receptors, along with many others. Although research efforts have focused on humans, GPCRs or GPCR-like proteins are expressed in multiple eukaryotes, including not just animals but also plants[8], fungi[9], and protozoa[10].

1.1.1.1. The conserved structural architecture of GPCRs

Despite the large number of GPCR family members, they all share a common structural architecture (Figure 1): seven transmembrane helices (TM1-7), linked by three extracellular loops (ECL1-3) and three intracellular loops (ICL1-3). The N-terminus of GPCRs is located at the extracellular side, and the C-terminus locates intracellularly. The N-terminus and the extracellular loops are responsible for recognizing ligands and modulating their access into the orthosteric binding pocket, which is located deeper within the receptor bundle and formed primarily by transmembrane helices. The residues of the transmembrane helices constitute the machinery for signal transduction across the membrane[11]. Finally, the intracellular regions interact with cytosolic signal transducers, such as G proteins, arrestins, and G protein-coupled receptor kinases (GRKs).

This conserved scaffold makes it possible to index GPCR residues according to their sequence or structure alignment, establishing generic GPCR residue numbering schemes (Figure 2). Sequence-based generic GPCR residue numbering schemes[12] exist for class A (Ballesteros-Weinstein), B1 and B2 (Wootten[13]), C (Pin[14]), and F (Wang[15]). In these systems, residues are indexed by two numbers separated by a dot (e.g. 2.50), where the first number indicates the helix (1-7) and the second the residue position relative to the most conserved residue in the helix, which is assigned the number 50. The reference conserved residue of the helices differs between the GPCR classes.

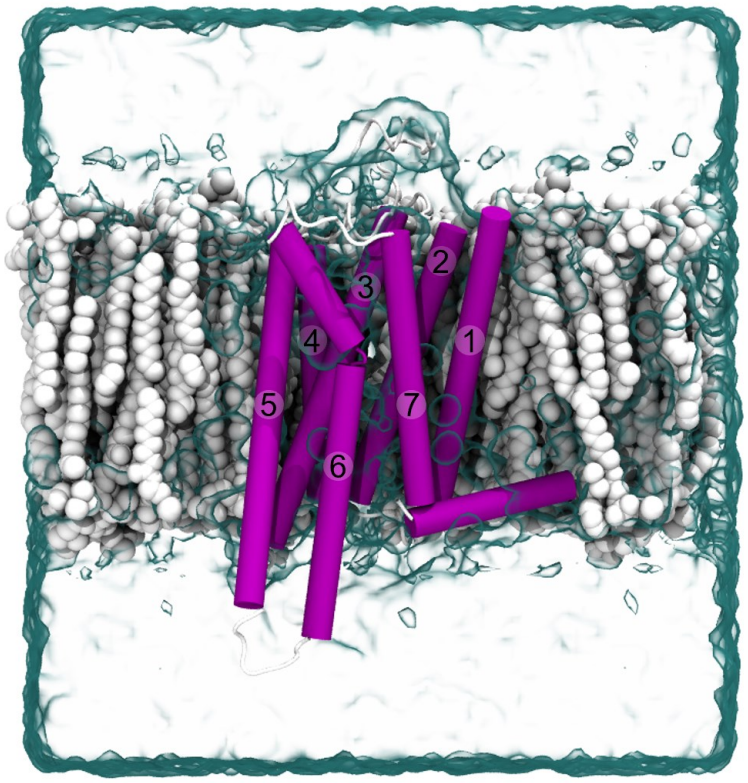


Figure 1. The conserved GPCR structure. GPCRs share a common structural architecture, consisting of seven transmembrane helices (TM1-7) connected by alternating intracellular and extracellular loops, with the amino terminus located at the extracellular side and the carboxyl terminus at the intracellular side. Here, we show the A_{2A} receptor (PDB ID: 3UZA), embedded in a membrane in a water box. Alpha helices are highlighted in magenta, transmembrane helices are numbered according to the conserved GPCR architecture.

GPCR crystal structures revealed that GPCR helices can have distortions such as bulges (one additional residue) and constrictions (one absent residue) in the helical turns[16]. These distortions offset the sequence-based generic residue numbers of the following residues when compared to an undistorted helix, and thus structurally equivalent residues no longer have the same number. To tackle this issue, GPCRdb implemented a complementary scheme based on structure alignments, instead of sequence[12]. GPCRdb numbers are distinguished by the separator "x" and may be used

alone, e.g. 2x56, or together with one of the sequence-based schemes, e.g. 2.57x56 (Figure 2).

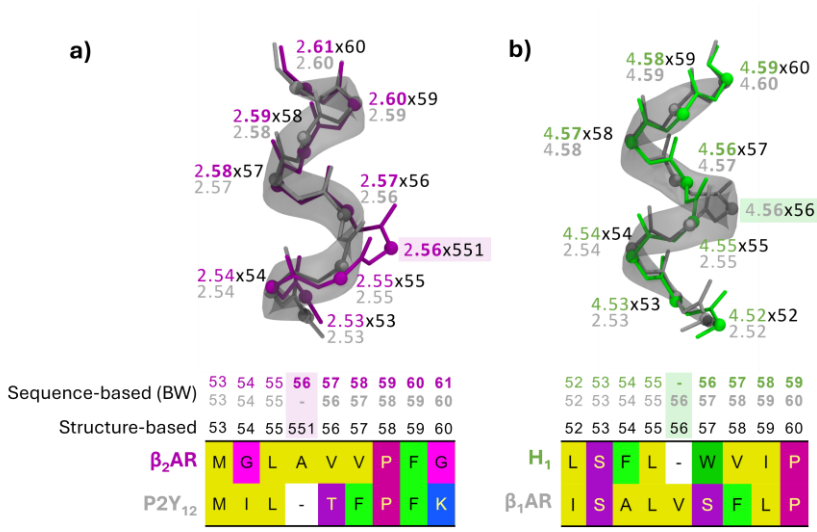


Figure 2. Sequence-based and structure-based GPCR residue numbering schemes. In GPCR residue numbering schemes, residues are indexed by two numbers, indicating the helix (1-7) and the residue position relative to the most conserved residue in the helix, which is assigned the number 50. Sequence-based schemes, such as Ballesteros-Weinstein (BW), are indicated with a "." between the two numbers (e.g. 2.50), while structure-based use a "x" (e.g. 2x50). Bulges (one additional residue) and constrictions (one absent residue) in the helical turns offset the sequence-based generic residue numbers of the following residues when compared to an undistorted helix, while structure-based generic residue numbers are not affected. **(a)** A bulge in helix 2 of the β_2-adrenoceptor (β_2AR, magenta, PDB ID: 2RH1) compared to the $P2Y_{12}$ receptor ($P2Y_{12}$, grey, PDB ID: 4NTJ). **(b)** A constriction in helix 4 of the histamine H_1 receptor (H_1, green, PDB ID: 3RZE), compared to the β_1-adrenoceptor (β_1AR, grey, PDB ID: 4AMJ). Sequence alignments were generated with GPCRdb[17,18].

1.1.1.2. The impact of ligand binding on the GPCR conformational equilibrium

GPCRs exhibit variable basal activities in the absence of bound ligands. For a given GPCR, each ligand has a characteristic capacity to activate or deactivate its target, which is commonly referred to as its efficacy. According to their efficacies, we can differentiate several types of ligands[19]: full agonists, which induce maximal response; partial agonists, which induce submaximal response; and inverse agonists, which decrease basal activity. Furthermore,

antagonists are ligands that do not initiate a signaling response and prevent agonist binding by occupying the binding pocket.

GPCRs are highly flexible proteins that exist in an equilibrium between multiple structural conformations (Figure 3a). Within their conformational landscape, distinct populations are linked to a specific physiological response. The binding of a ligand in the orthosteric site can shift the conformational equilibrium, altering the overall signaling response[20,21]. For instance, an orthosteric agonist would shift the equilibrium towards active conformations. This can be mediated by two mechanisms. On the one hand, the agonist can bind to inactive or intermediate states and induce structural rearrangements towards active conformations (agonist-induced activation, Figure 3b). On the other hand, the agonist can directly sample the active conformations and stabilize them. By doing this, it shifts the conformational equilibrium towards active-like states (conformational selection, Figure 3c). Both mechanisms may contribute to receptor activation to a different extent depending on the ligand and receptor type. Finally, active receptors demonstrate a higher propensity to interact with intracellular coupling partners, known as transducers, leading to the initiation of signaling cascades.

Different GPCR ligands can modulate signaling differently, for instance causing coupling preference to G proteins over arrestin, or one G protein subtype over the others, which leads to the activation of different signaling pathways *(1.1.1.3. Signal transduction through GPCRs)*[22,23]. This phenomenon of preferential signaling is known as “biased signaling” or “functional selectivity”, and the ligands that demonstrate it are “biased ligands" (Figure 4)[22,24]. Biased ligands are remarkably promising for drug design, since they can be used to selectively modulate disease-associated pathways while not engaging non-pathological pathways, avoiding side effects. As an example, β-arrestin biased ligands of the AT_1 receptor show an improved therapeutic profile when compared to conventional AT_1 blockers, which antagonize both G protein and β-arrestin recruitment. This is the case of TRV120027, currently in clinical trial for the treatment of hypertension[25,26], that shows vasodilating effects similar to conventional blockers but, unlike them, avoids side effects such as decreased cardiac performance[27].

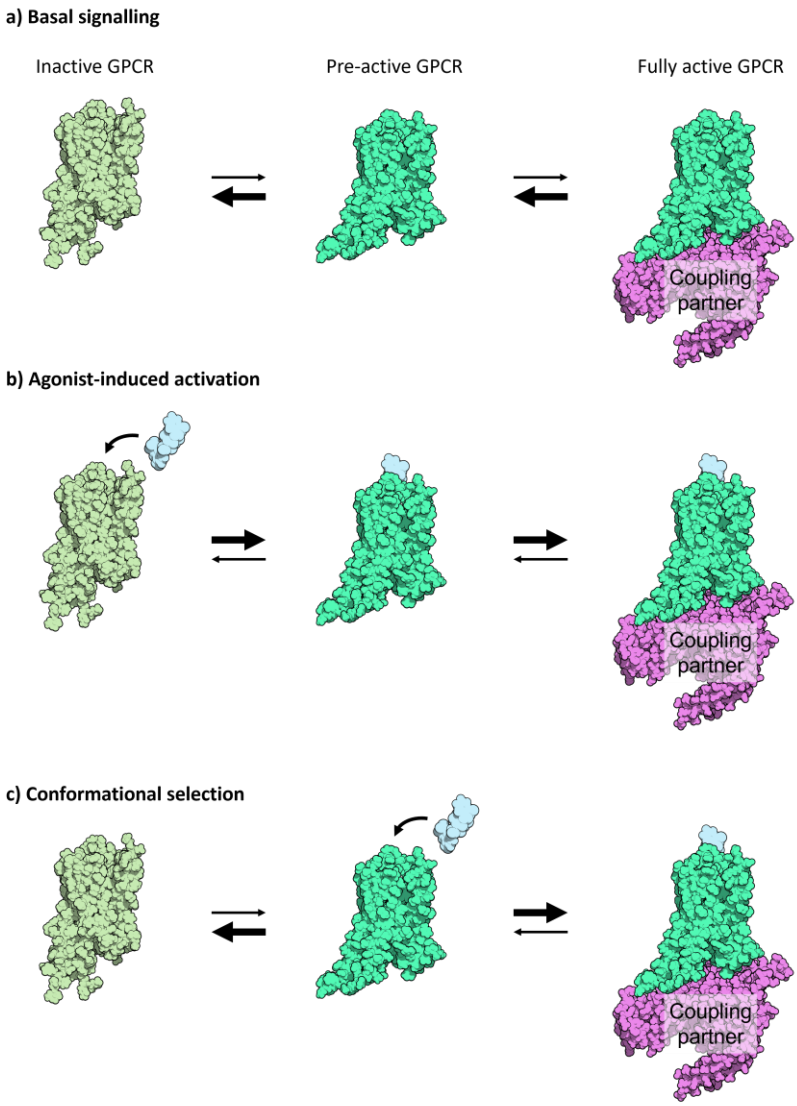


Figure 3. GPCR activation and signaling. (a) GPCRs exist in an equilibrium between multiple conformational states. From a structural perspective, there are two mechanisms by which an agonist (blue) can mediate GPCR activation. **(b)** On the one hand, the agonist can bind to inactive or intermediate states and induce structural rearrangements towards active conformations. **(c)** On the other hand, the agonist can directly sample the active conformations and stabilize them. Both mechanisms may contribute to receptor activation to a different extent depending on the ligand and receptor type. Active receptors interact with intracellular coupling partners, initiating signaling cascades.

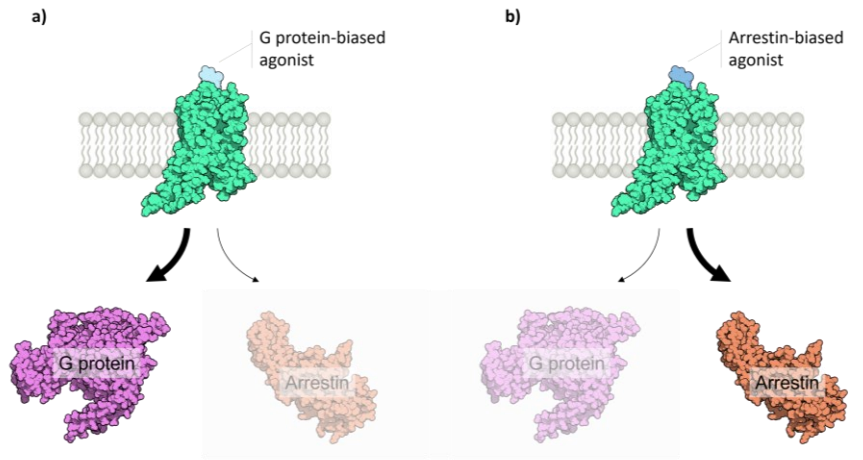


Figure 4. Biased signaling. Biased agonists show preference for certain downstream coupling partners over the others, and thus engage certain signaling pathways over the others. In this respect, A G protein-biased agonist **(a)** shows preference for G proteins, while an arrestin-biased agonist **(b)** shows preference for arrestin.

The phenomenon of biased signaling could be explained by the idea that biased agonists stabilize the receptor in a subset of active conformations ("biased" conformations) with a particular capacity to couple to specific intracellular signal transducers[28–30]. This theory, which is supported by experimental observations[31–33], has opened new questions for structure-based drug discovery. Deciphering how subtle variations in ligand structure can translate into important changes in receptor signaling, as well as understanding how ligands stabilize different GPCR conformational ensembles leading to a specific intracellular coupling, could promote the discovery of safer drugs.

Finally, it is worth noting that GPCRs can not only recognize orthosteric ligands but are also affected by allosteric modulators, including small molecules, lipids, ions, and sterols[34–41]. Allosteric modulators bind to regions spatially distinct from the orthosteric binding pocket and modulate the affinity and/or efficacy of orthosteric ligands[42]. Allosteric binding pockets present greater sequence divergence between receptor subtypes, which makes these modulators promising therapeutic targets with improved selectivity. Moreover, they provide a strategy to fine-tune the response

triggered by the orthosteric ligand, avoiding severe effects on the cell[43,44].

1.1.1.3. Signal transduction through GPCRs

The classical signal transduction through GPCRs is dependent on receptor-mediated activation of G proteins (Figure 5, left), which are composed of three subunits –Gα, Gβ, and Gγ. Based on sequence homology between human Gα isotypes, we can differentiate four major G protein families (G_s, $G_{i/o}$, $G_{q/11}$, and $G_{12/13}$), each with potentially varying signaling properties[45–47].

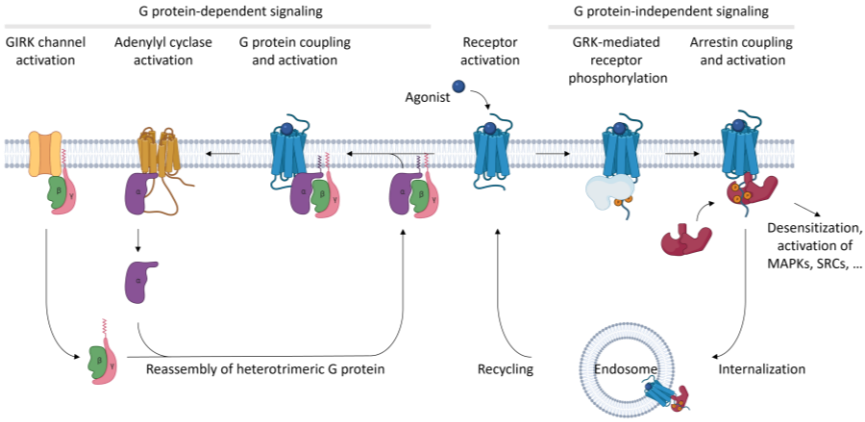


Figure 5. GPCR signal transduction. Simplified diagram of GPCR signaling, including G protein-dependent and G protein-independent signaling[48]. GIRK channel, G-protein-coupled inwardly rectifying potassium channel; GRK, G protein-coupled receptor kinase; MAPK, mitogen-activated protein kinase. Created with BioRender.com.

Receptor activation promotes the engagement, and consequent activation, of the G protein, which causes the dissociation of the Gα and Gβγ subunits[49]. Both dissociated components modulate the activity of different downstream effector proteins. By this, Gα modulates the production of second messengers such as cyclic adenosine monophosphate (cAMP)[50], and Gβγ regulates ion channels and phospholipases, among others[51,52], propagating the signaling cascade. After some time, the Gα subunit returns to an inactive-like conformation, regaining affinity for the Gβγ dimer. Once the heterotrimeric G protein is reformed, it can bind again to

GPCRs in active conformations, completing the G protein activation circle.

Apart from interacting with G proteins, activated GPCRs may be phosphorylated by GRKs[24], which stimulates the coupling of arrestin to the receptor (Figure 5, right)[53]. Arrestin coupling leads to desensitization and internalization of the receptor. Moreover, some researchers propose that arrestin coupling also stimulates G protein-independent signal transduction through activation of downstream effector proteins like mitogen-activated protein kinases (MAPKs) or SRC kinases[48]. However, it is still under debate which of those pathways are independent of G-protein coupling[54]. To date, only four types of arrestin have been identified: arrestin-1 and arrestin-4, which exist exclusively in the visual system[55], and β-arrestin 1 and 2, which regulate GPCRs not involved in sight[56].

1.1.1.4. The importance of three-dimensional structures for the understanding of GPCR biology

Given the complex structure-function relationships involved in GPCR signaling, GPCR structures are key for understanding the details of GPCR functioning, as well as for the identification of new GPCR drugs. Thus, the study of GPCRs has been very linked to the pursuit of novel high-resolution GPCR structures.

The first high-resolution crystal structure of a GPCR was published in the year 2000, consisting of bovine rhodopsin[57]. This achievement was key to triggering the first structural analyses of class A GPCRs. After seven years of extensive research and technology development, the first crystal structure of a human GPCR, β_2-adrenoceptor (β_2AR), was published[58,59]. Importantly, this was the first structure of a GPCR bound to a diffusible (not covalently bound) ligand, providing valuable information on the basis of ligand binding in GPCRs. Another important breakthrough in GPCR structural biology was the first active-state GPCR structure of ligand-free rhodopsin (known as opsin), stabilized with a peptide derived from the G protein Gα subunit[60]. Moreover, shortly after that, the crystal structure of the active state β_2AR in complex with the G_s heterotrimer was published, being the first high-resolution view of the complete signal transduction machinery[61]. Since then, important advances in protein engineering,

X-ray crystallography, and cryo-electron microscopy (cryo-EM) have led to an exponential growth in GPCR structure determination. Consequently, to date, there are over 500 GPCR structures available (Figure 6), spanning all receptor classes, and including complexes with many different ligand types and signal proteins[17].

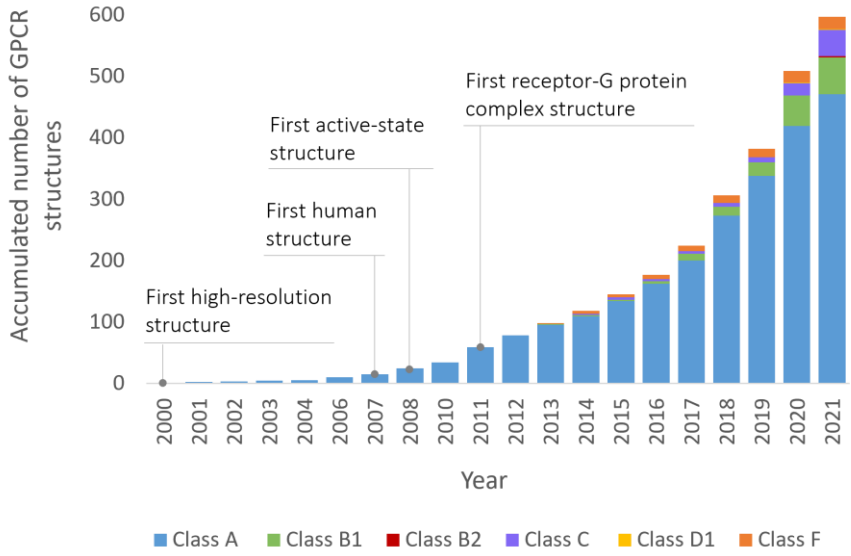


Figure 6. Progression in GPCR crystal structure determination in the past decades. Accumulated number of GPCR structures available in GPCRdb[17].

This large dataset shed unprecedented light on the structural similarity and diversity of the GPCR superfamily, as well as the molecular basis of GPCR ligand recognition, activation, allosteric modulation, and dimerization, among others[62]. Importantly, this has opened new opportunities for structure-based drug design[4,63,64].

The success of high-resolution GPCR structures has also revealed the need to complement this data with information derived from different techniques. The data that can be extracted from structures is limited to the analysis of static protein snapshots isolated from their functional context. Thus, details of GPCR intrinsic flexibility and conformational plasticity can only be partially derived from them. These limitations have stimulated the use of a series of biophysical and computational approaches to study GPCRs from a more dynamic perspective[65–67]. Among them are molecular

dynamics (MD) simulations, which will be discussed in more detail in section *1.2. Molecular dynamics* (MD) simulations.

1.1.2. The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)

The SARS-CoV-2 is a novel virus of the family Coronaviridae and the causative agent of the COVID-19. The COVID-19 was first identified in late December 2019, when a cluster of patients was diagnosed with pneumonia of unknown cause[68]. Soon after, the disease spread globally, leading to the declaration of the COVID-19 pandemic by the World Health Organization (WHO). As of the 16th of September 2021, the disease has spread to 221 countries and territories, with more than 225 million confirmed cases and 4.6 million deaths[69]. As a result, COVID-19 has become a serious threat to global human health and socioeconomic stability.

1.1.2.1. Genome organization and proteome

The SARS-CoV-2 is an enveloped virus with a positive-sense single-stranded RNA (+ssRNA) genome[70,71] with multiple open reading frames (ORFs), which allow the expression of the different viral genes[72] (Figure 7a). The viral +ssRNA can act as a messenger RNA (mRNA) from which coding sequences are directly translated into polyproteins by the cell translation machinery, namely the ribosomes. Particularly, two large polyproteins are translated from ORF1a and ORF1b. After proteolytic cleavage, the polyproteins give rise to 16 proteins known as the nonstructural proteins (nsp1-16), involved in the replication and transcription of viral RNA[73]. They contain multiple enzymatic functions, including proteases (nsp3, nsp5), RNA-dependent RNA polymerases (nsp12), RNA helicase (nsp13), and the proofreading exonucleases (nsp14)[74].

After some viral replication, the virus expresses subgenomic mRNA, encoding other proteins. This gives rise to the four main structural proteins of the virus (the spike glycoprotein, envelope protein, membrane protein, and nucleocapsid protein) as well as several accessory proteins (orf3a, orf3b, orf6, orf7a, orf7b, orf8b, orf9b, orf9c, orf10)[75].

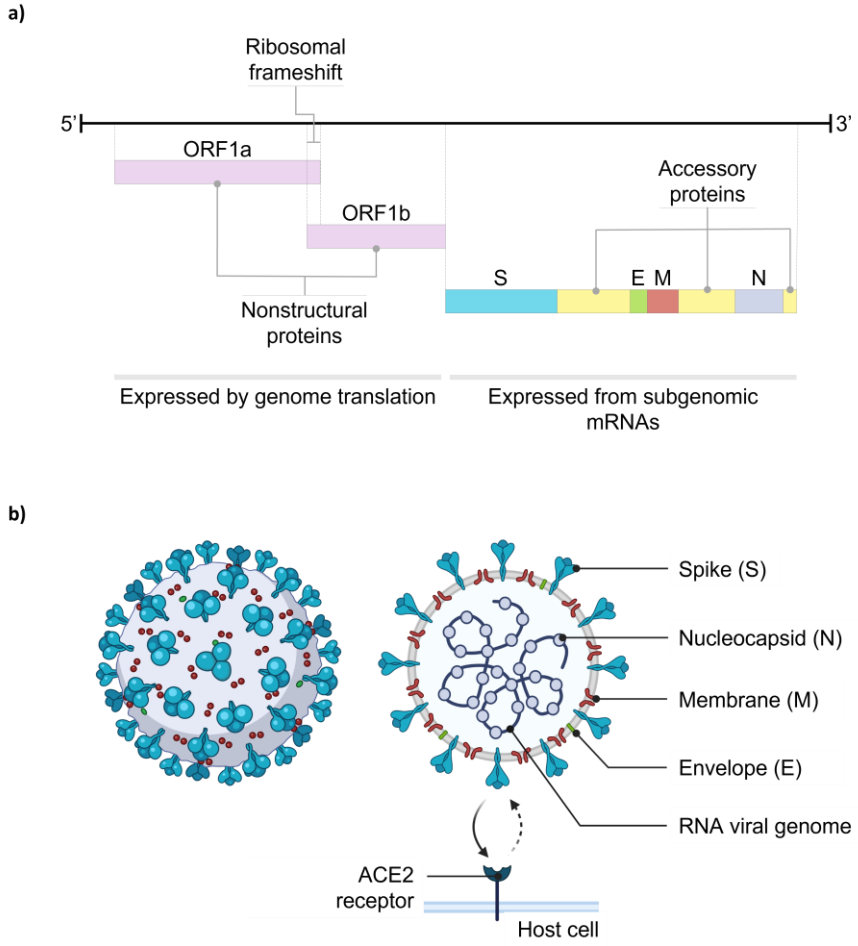


Figure 7. SARS-CoV-2 genome organization and configuration of the structural proteins. (a) The SARS-CoV-2 contains a positive-sense single-stranded RNA (+ssRNA) genome, including the open reading frame (ORF) 1ab (ORF1a and ORF1b) that code for nonstructural proteins, followed by the spike glycoprotein (S), envelope protein (E), membrane protein (M), nucleocapsid protein (N), and non-structural ORFs that code for accessory proteins. ORF1ab is expressed by direct genome translation, while the rest is expressed from subgenomic messenger RNA (mRNA) **(b)** Configuration of the four main structural proteins of SARS-CoV-2. The spike, nucleocapsid, and membrane proteins are embedded in the virus envelope, while the nucleocapsid protein binds to the viral genome. The spike protein recognizes and binds to the human angiotensin-converting enzyme-2 (ACE2) of the host cells. Created with BioRender.com.

Among the structural proteins, the spike, membrane, and envelope proteins are embedded in the membrane and involved in cell recognition and entry (Figure 7b). Particularly, spike proteins are glycosylated trimers that protrude from the virus envelope, giving coronaviruses their characteristic appearance. Their role is to mediate recognition and entry to the host cell by binding to the angiotensin-converting enzyme-2 (ACE2) on the surface of the human cell[76,77]. The spike protein is also the main target of neutralizing antibodies generated following infection by SARS-CoV-2[78,79], and a component of both mRNA and adenovirus-based vaccines currently licensed for use, as well as others awaiting regulatory approval[80]. Also embedded within the virion envelope are membrane proteins, which are dimeric complexes believed to anchor ribonucleoprotein complexes to the envelope and give the virion its spherical shape[81,82]. Similarly, the envelope protein is thought to be a transmembrane protein that forms pentameric ion channels and contributes to viral budding[83]. Unlike the other structural proteins, the nucleocapsid protein is located inside the membrane[84], organizing the RNA into a ribonucleoprotein core[85,86].

Finally, accessory proteins play important roles in viral interaction with host cells, helping the virus to evade the immune system and enhancing its virulence[87].

1.1.2.2. Emergence of variants

As an RNA virus, SARS-CoV-2 has a relatively high mutation rate. Although most mutations are expected to be either deleterious and rapidly purged or relatively neutral, a small proportion will provide a fitness advantage to the virus[88]. Since late 2020, the emergence of sets of phenotype-enhancing mutations has been detected, likely in response to the changing immune profile of the human population[89]. These sets of mutations, showing increased pathogenicity, infectivity, transmissibility, and/or antigenicity, were termed ‘variants of concern’. For example, the B.1.1.7 (Alpha) variant, initially detected in the UK showing augmented pathogenicity and transmissibility, quickly become a dominant strain in many countries[90]. Currently, this variant has de-escalated, with the worldwide spread of other variants such as B.1.351 (Beta)[91–93], P.1 (Gamma)[94,95], and B.1.617.2 (Delta)[96,97], all of which show evidence for impact on severity, transmissibility, and immunity.

Understanding how these variants affect immune recognition and thus can cause an increased risk of reinfection or vaccine failures is of major importance. In this respect, mutations affecting the spike protein, which is a primary antigen, are of particular interest (Figure 8). Similarly, spike mutations affecting the binding to the host cell can affect the severity of the disease. Information on how spike mutations affect antigenic profiles and other functional characteristics of the virus can be derived from structural studies[78,79,98–100], site-directed mutagenesis[101,102], and deep mutational scanning[103–105], among others. For example, the structural study of the spike protein in complex with the antigen-binding fragment of multiple antibodies showed that the spike receptor-binding domain (RBD) is an immunodominant target[78,99]. Moreover, deep mutational scanning experiments showed that the mutations that cause a stronger reduction of antibody binding occur at a relatively small number of RBD residues[103].

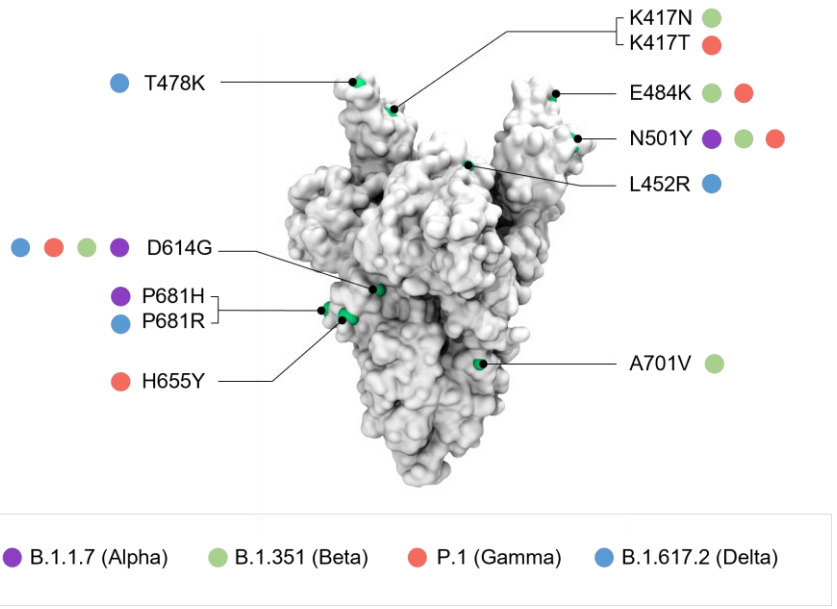


Figure 8. Key spike mutations in SARS-CoV-2 variants[106]. Spike mutations related to increased fitness in previous (B.1.1.7 or Alpha) and current SARS-CoV-2 variants of concern in September 2021 (B.1.351 or Beta, P.1 or Gamma, and B.1.617.2 or Delta). Mutations occur in the three spike subunits but are only highlighted in one for clarity.

The integration of data regarding the functional effect of mutations with the surveillance of emerging variants has the potential to facilitate the automated detection of potential variants of concern before they spread widely. Early detection would help to guide the implementation of targeted control measures and further experimental characterization[107]. Importantly, this strategy strongly benefits from close international collaboration, stressing the importance of rapid and open sharing of data.

1.1.2.3. Protein structures and their relevance in finding pharmacological interventions against SARS-CoV-2

Unveiling the structural basis of SARS-CoV-2 infection has been a key priority since the emergence of the COVID-19 disease. Structural information of the proteins that constitute the virus, as well as proteins with which the virus interacts (e.g. ACE2) helps to further elucidate their mechanisms of action and find new molecular therapeutics that can impact their activity[108]. To this end, the structural biology community has made enormous efforts to rapidly build models of SARS-CoV-2 proteins and the complexes they form using cryo-EM and x-ray crystallographic techniques[109–115]. Consequently, over 1400 structures of over 20 different SARS-CoV-2 proteins have been deposited in the protein data bank (PDB, rcsb.org)[116], as of September 2021. Numerous druggable targets for the inhibition of SARS-CoV-2 have been proposed based on these structures[117].

There are three viral proteins that are particularly promising as targets for pharmacological intervention, and thus their structures have been of special interest: the spike protein, the main viral protease (nsp5, also known as Mpro and 3-CLpro), and the RNA-dependent RNA polymerases (nsp12) (Figure 9).

As mentioned before, the spike protein is responsible for the recognition and entry to the host cell[76,77], and a major target for vaccine development, as well as neutralizing antibodies[78–80]. The spike forms highly glycosylated trimers on the surface of the virion (Figure 9a). Its activation requires the cleave by a host cell protease, which produces two subunits: S1 and S2[109,112,118]. The S1 subunit largely contains the RBD and the amino-terminal domain, and is

responsible for binding to the host cell-surface receptor, ACE2[111] (Figure 9b).

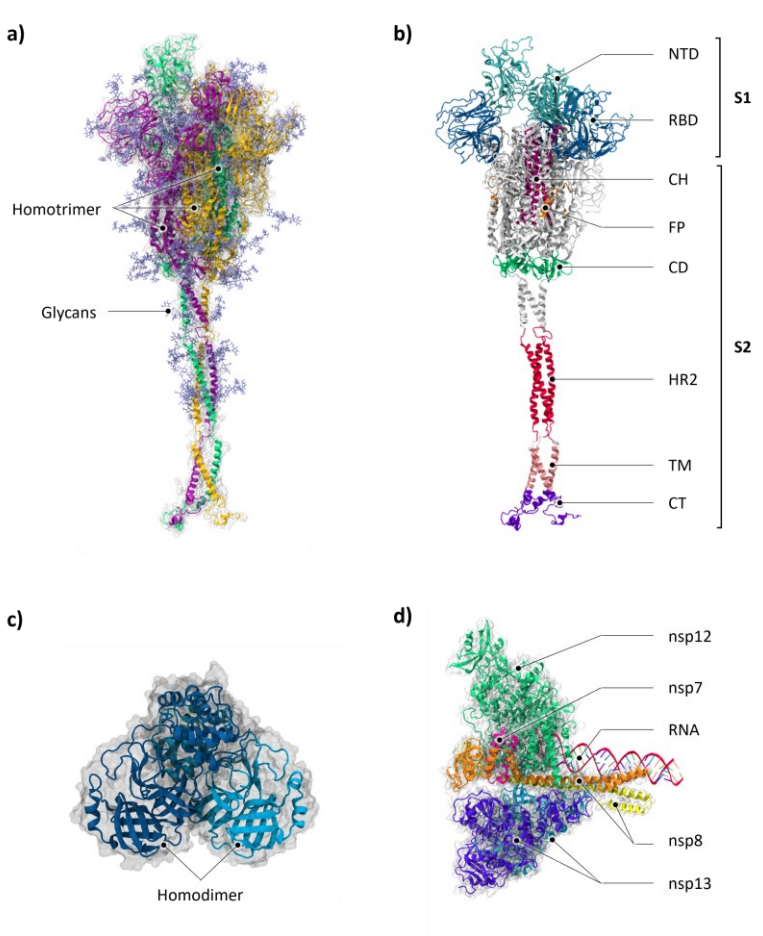


Figure 9. Three-dimensional structure of SARS-CoV-2 proteins with high pharmacological interest. (a) Full-length model of the spike protein[119], where each monomer is represented in a different color (green, orange, or magenta) and glycans are shown in violet licorice. **(b)** Each spike monomer can be divided into two subunits: S1 and S2. The S1 subunit contains the amino-terminal domain (NTD, cyan) and the receptor-binding domain (RBD, blue), while the S2 includes the central helix (CH, magenta), fusion peptide (FP, orange), connecting domain (CD, green), heptad repeat 2 domain (HR2, red), transmembrane domain (TM, pink), and cytoplasmic tail (CT, purple). **(c)** Structure of nsp5 (PDB ID: 7BB2), which is active as a homodimer. Monomers are shown in light and dark blue. **(d)** Structure of the nsp7-nsp8-nsp12-nsp13 replication/transcription complex (RTC, PDB ID: 6XEZ), including nsp12 (green), nsp7 (magenta), double-stranded RNA (red), two copies of nsp8 (orange and yellow), and two copies of nsp13 (purple and light blue).

The S2 is the transmembrane subunit, and it mediates membrane fusion (Figure 9b). Based on the RBD position, the spike protein structure can take two conformational states: closed or open (Figure 10). The closed state is the predominant one, in which the RBDs are buried to evade immunosurveillance mechanisms. Contrarily, the open state exposes the RBD, enabling binding to ACE2. The binding of one of the three RBDs to ACE2 progressively induces the opening of the others, until a fully open, three-ACE2-bound structure is formed[120]. This triggers the activation of the S2 subunit, which undergoes dramatic structural rearrangements. These rearrangements involve shedding the S1 subunit, insertion of an S2 domain known as the fusion peptide into the host membrane, and refolding into a needle-shaped, hairpin-like structure that opens a fusion pore[76].

Nsp5 (Figure 9c), the main protease, is an essential cysteine protease required for cleaving the viral precursor polyproteins, which include the precursors of the SARS-CoV-2 replication and transcription machinery[121]. Viral proteases have been successfully targeted to treat other viral infections, such as those caused by the human immunodeficiency virus (HIV) and hepatitis C virus (HCV). This protease is highly conserved between SARS-CoV-2 and other beta-coronaviruses such as SARS-CoV but differs from host proteases, which makes it a highly attractive drug target. In fact, many of the previously identified inhibitors against SARS-CoV and MERS-CoV nsp5 were found to be also active against SARS- CoV-2[122]. From a structural perspective, nsp5 functions as an active homodimer. Interestingly, the SARS-CoV-2 nsp5 dimer is tighter than in SARS-CoV, which results in a higher catalytic efficiency[115]. Currently, there are over 250 high-resolution structures deposited in the PDB for SARS-CoV-2 nsp5, including complexes with various inhibitors, which help to understand the molecular basis of the interaction[115,123].

Nsp12, the RNA-dependent RNA polymerases, is the main component of the SARS-CoV-2 replication and transcription machinery, known as the replication/transcription complex (RTC, Figure 9d). It is a key target for antiviral inhibitors, mainly nucleotide analogs such as remdesivir, which has been reported to effectively inhibit SARS-CoV-2 proliferation[124]. In fact, viral RNA polymerases have been successfully targeted for the treatment of

numerous viral infections, including those caused by HCV, HIV, and the influenza virus[125–127]. The high-resolution structure of the RTC, consisting of nsp12 and the cofactors nsp7 and nsp8, was determined by several research groups[128–132]. Among them, the model of remdesivir binding to nsp12 within the RTC was revealed, providing insights into its mechanism of action[131]. Several cryo-EM structures also managed to capture the RTC bound to other cofactors –what is known as the extended RTC–[133,134], revealing the importance of the interaction with nsp13 and nsp9 to modulate the complex activity. These structures were critical for understanding the complete architecture of the RTC, as well as potential means of inhibiting nsp12 activity.

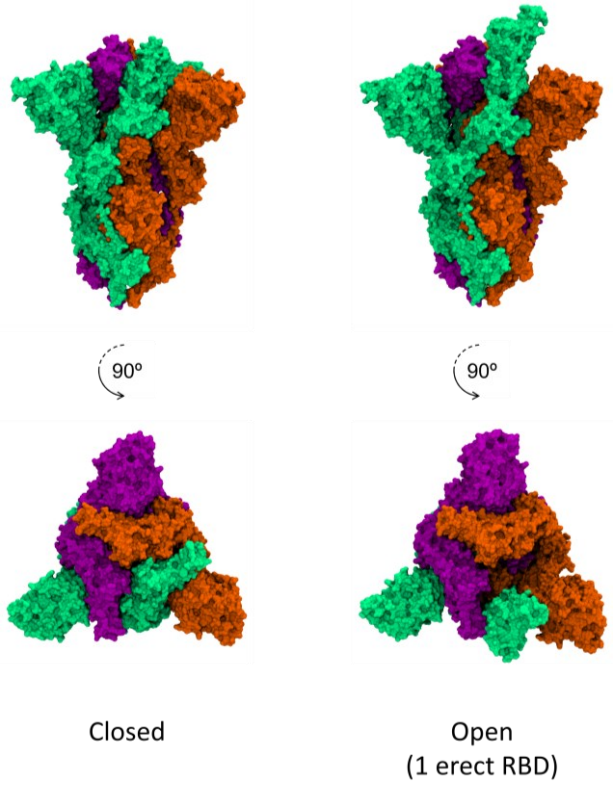


Figure 10. Closed and open conformations of the spike protein. Three-dimensional structures of the spike protein in the closed (PDB ID: 6ZGE) and open (PDB ID: 6ZGG) conformations, shown with a trimer axis vertical view (top) and an orthogonal top-down view along this axis (bottom). Spike monomers are colored in green, orange, and magenta. In the open conformation, one of the three receptor-binding domains (RBDs), in green, is erect, enabling the binding to ACE2.

Despite the undeniable value of structural information of these and other SARS-CoV-2 proteins, there are limits to what they can tell us. As happens with GPCRs (*1.1.1.4. The importance of three-dimensional structures for the understanding of GPCR biology*), it is well established that the function of a protein is dictated by the full range of conformations it can access, many of which remain hidden to experimental static structures. Capturing this range of conformations for SARS-CoV-2 can provide a clearer picture of the molecular mechanism of processes in which they are involved, including cell infection, evasion of host's immune system, and replication. This can also present new therapeutic opportunities, such as cryptic pockets that are absent in experimental snapshots but provide novel targets for drug discovery. Several techniques can be used to capture this conformational plasticity, including MD simulations, which will be discussed in more detail in the next section (*1.2. Molecular dynamics (MD) simulations*).

1.2. Molecular dynamics (MD) simulations

Static three-dimensional structures derived from experiments via X-ray crystallography or cryo-EM provide high-resolution information about specific protein conformational states. However, we need to be aware that these structures represent low energetic conformational states that are obtained under experimental conditions that often deviate from native-like conditions[25]. Several approaches can be used to study the three-dimensional structure of proteins from a more dynamic perspective, incorporating information on their intrinsic flexibility and conformational plasticity. While experimental techniques such as nuclear magnetic resonance (NMR)[67], double electron-electron resonance (DEER)[135], or single-molecule fluorescence energy transfer (smFRET)[66] have provided relevant insights into the dynamics and flexibility of proteins such as GPCRs, MD simulations have emerged as the most promising opportunity to study the complexity of protein conformational dynamics in atomistic detail[65,136].

The first MD simulation of a macromolecule of biological interest was published in 1977, consisting of a 9.2-ps trajectory of a small protein, the bovine pancreatic trypsin inhibitor, in vacuum. This achievement was instrumental in replacing the view of proteins as

relatively rigid structures with the understanding that they are dynamic systems[137]. Interestingly, the first MD simulation of a GPCR was published more than 10 years later, before the first GPCR crystal structure was resolved. It corresponded to an 80 ps long trajectory of a rat dopamine D_2 receptor, modeled from its sequence[138]. From that point forward, thanks to continuous advances in both methodology and computational resources, MD simulations have gradually extended to larger systems and longer timescales. A major breakthrough for the MD field was the development of algorithms optimized for graphical processor units (GPUs), a technology first designed to improve video game performance[139,140]. This enabled researchers to perform on commodity hardware calculations that were previously only possible with the use of supercomputing clusters. High-performance computing has also contributed to making simulations more powerful and accessible[141–143]. Along with these technological advances, the underlying physical models and methods have also improved over the years to address ever more complex biological and chemical questions[144,145]. Moreover, the expansion of free and user-friendly software for the input preparation (e.g., CHARMM-GUI[146], HomolWat[147], HTMD[148]) and analysis (e.g., MDAnalysis[149,150], MDTraj[151]) of MD simulations has greatly contributed to the broad application of this technique. As a result, MD simulations have become a well-established technique (Figure 11), able to resolve mechanistic elements at a spatio-temporal resolution and conditions that are not always accessible with experimental techniques[20,152].

MD simulations provide a prediction of the time-resolved motions of a molecular system (Figure 12). Molecular systems for protein MD simulations include a protein model, obtained by the curation of experimentally solved structures or homology modeling. In the case of membrane proteins, the model is embedded in a lipidic membrane. This system is solvated and ionized to a physiological concentration, and other molecules such as ligands may be included. In classical MD simulations, each atom of the molecular system is represented as a point particle with its corresponding connectivity to other atoms. The movements of all the atoms are predicted by iteratively solving Newton's equations of motion. For that, the forces that act on each atom are calculated based on bonded (bonds, angles, and torsions) and non-bonded (Lennard

Jones and Coulomb) interactions. This is possible thanks to force fields, which are potential energy functions along with the associated parameters that define the interactions of different atoms[153,154]. A variety of force fields have been developed[155–159], based on quantum mechanical calculations and experimental measurements. Based on the obtained forces, the atomic positions and velocities after a defined time interval are updated. This process of calculating atomic forces, solving Newton's equations of motion, and updating atomic velocities and position after a time interval is repeated in an iterative cycle until the final simulation time is reached. With this, the coordinates of each atom over time, so-called trajectories, are obtained, describing the behavior of the simulated molecules. The analysis of these trajectories can lead to the identification of functionally relevant dynamic processes.

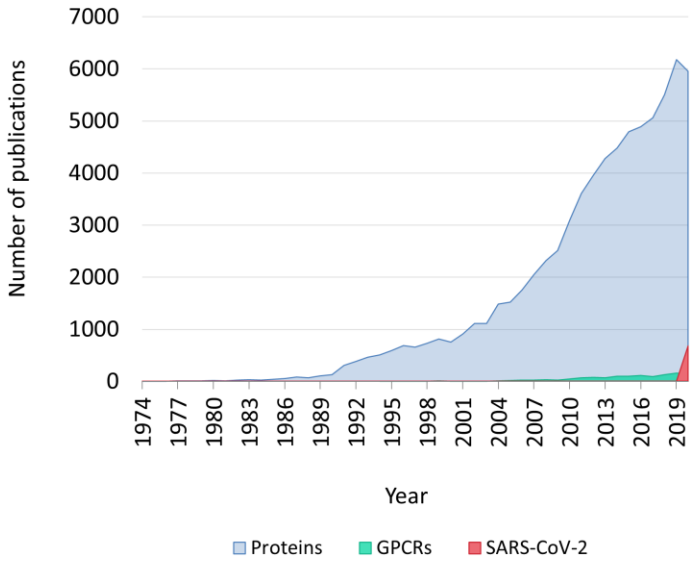


Figure 11. Popularity of MD simulations. The exponential growth of successful MD-based research is evidenced by the rapid upsurge in the number of publications per year indexed at Thomson Reuters' Web of Science with the topics molecular dynamics and proteins (blue). This trend is also found in the field of GPCRs (green, search topics: molecular dynamics, proteins, and GPCR or GPCRs). Moreover, since the COVID-19 outbreak, researchers have also resorted to MD simulations for the study of SARS-CoV-2-related proteins (red, search topics: molecular dynamics, proteins, and SARS-CoV-2).

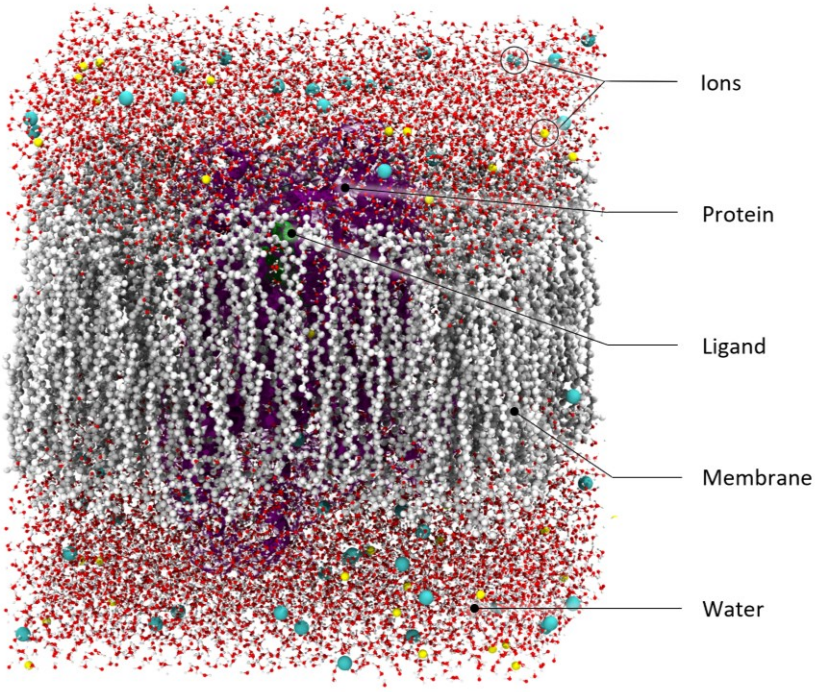


Figure 12. Molecular system for protein molecular dynamics (MD) simulations. Example of a molecular system, including the δ opioid receptor (magenta) with the ligand naltrindole (green) in the orthosteric binding pocket and embedded in a 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) membrane (white licorice). The system is solvated with water (red and white CPK) and ionized with sodium ions (yellow) and chloride (blue).

Despite the broad capabilities of MD simulations, this technique has some limitations that must be taken into account. Firstly, to reduce the computational cost and complexity, force fields contain approximations and thus are imperfect. Studies comparing simulation results with experimental data indicate that force fields have improved significantly over the past decade[160], but more remains to be done to achieve increased accuracy. Moreover, most MD simulations do not consider quantum effects, such as changes in atom charges or the dissolution/formation of covalent bonds. Therefore, in classic MD simulations protonation states of titratable amino acid residues are fixed, as well as disulfide bonds. Thus, they need to be set carefully at the beginning of the simulation[161]. Another limitation of MD techniques is the amount of computational resources necessary to simulate a system for a

biologically relevant amount of time. Simplified models like coarse-grained, where groups of atoms are represented as beads to reduce the level of description, can extend accessible timescales by orders of magnitude, as they are less expensive computationally[162,163]. Still, the problem with long-timescale events is that they often imply the transition between free energy states that are separated by high-energy barriers. In this situation, classical MD simulations tend to get trapped in the local minimum-energy states for a long time, which restrains the sampling process and leads to poor characterization of the protein's dynamic behavior[164]. A useful strategy to tackle this problem is the application of enhanced sampling techniques. Enhanced sampling simulations, including replica-exchange MD, metadynamics, and simulated annealing, are able to efficiently overcome energetic barriers and access additional conformational states by including an external bias[165]. Nevertheless, the capabilities of MD simulations are constantly increased by improvements in simulation algorithms and computer hardware. Thus, it is expected that the timescales accessible to classical, all-atom MD simulations will continue expanding[166–168].

MD simulations have been key for the study of the two types of proteins examined in this thesis –GPCRs and SARS-CoV-2 proteins. In the case of GPCRs, MD simulations have proven useful to complement static structural data, improving our understanding of the physiology and pharmacology of this protein family, as reviewed in[168–170]. In this respect, MD simulations provide valuable information on processes such as the binding of small molecules or drugs to orthosteric[171–175] or allosteric[43,176–179] receptor sites. In the case of biased ligands, MD simulations are useful to explore how they stabilize different GPCR conformational ensembles that lead to a specific intracellular coupling[180–187]. With this technique, we can also determine how a GPCR will respond to perturbations such as mutations[188–191], post-translational modifications[192,193], and the composition of the cell membrane[194–198]. In addition, we can reveal important insights into the activation and inactivation mechanism of GPCRs, such as the conformational rearrangements that occur during receptor (in)activation and the formation of metastable receptor states along the transition pathways[199–203]. We can also explore the interaction with intracellular coupling partners[204,205]. Moreover, MD simulations have the ability to monitor diffusion and binding events of water molecules, and thus can be used to

determine the role of water molecules in receptor functioning and dynamics[206–212]. Similarly, they helped to shed light on the role of ions for GPCR function, such as the molecular mechanism of sodium-induced allosteric modulation[35,36,213,214]. Even larger-scale processes such as receptor dimerization/oligomerization, which has been implicated in fine-tuning GPCR signaling, can be investigated using different MD techniques, typically with coarse-graining[38,215,216].

Regarding the study of SARS-CoV-2-related proteins, the increased availability of structural data has triggered the use of MD simulations to study them, often after immense modeling and computational efforts[217–219], with the goal of supporting research against the pandemic. Obtained MD data have provided highly relevant information that helped to describe the functional dynamics of the viral proteome. Particularly, many MD studies have been directed to the understanding of the dynamics of the spike protein, due to its important role in the SARS-CoV-2 infection mechanism and immune response. These studies shed light on aspects such as the role of the spike glycan shield[119,219–221], which not only acts as a mechanism to evade the host immune system but also is an essential structural element to modulate the conformational dynamics of the RBD. MD simulations can also be used to determine the effect of environmental conditions on the structure of proteins. This revealed that the spike is sensitive to temperature, acquiring an open conformation, which enables receptor binding, at lower temperatures (20-40 ºC) and a closed conformation at higher temperatures (> 40 ºC)[222]. Other important findings were the conformational changes that take place during the closed-to-open transition of the spike's RBD, granting useful information for the design of vaccines and antivirals[217,219,223]. MD simulations also provided relevant insights into the spike-ACE2 binding[224–229] that explain the higher infectivity of the virus in humans. Moreover, they showed how spike interacts with antibodies[230] and with surfaces of different materials[231], as well as the binding of small molecules to the RBD-ACE2 complex[232]. Applied to other viral proteins, MD simulations were able to elucidate cryptic and allosteric pockets that are absent in experimental structures but provide novel targets for drug discovery[217,233]. They also helped to clarify the effect and binding modes of potential drugs[234–237]. Another strength of this technique is that it can help rationalize the

structural/functional impact of sequence variability in the spike[238–241] or other proteins[242]. This is particularly useful when the relationship between mutation location and activity is not obvious, for example when the mutation is distant from the protein's active center. Finally, beyond the simulation of specific proteins, it was possible to construct a coarse-grained model of the SARS-CoV-2 virion from the available structural and atomistic simulation data on SARS-CoV-2 proteins[218]. All this information has helped accelerate COVID-19 research and has improved our knowledge of SARS-CoV-2 biology.

1.3. Sharing MD simulations

The dynamics obtained by MD simulation often contain much more information than what is analyzed in the scope of a single publication or study. However, the publication of MD studies is often not accompanied by the obtained trajectories. Instead, results are frequently shared as text, tables, plots, figures, and, at best, videos. This reduces the dynamic information contained in a simulation to a static or very focused view of the simulated process, losing valuable information on the way. Even in the cases when the simulation data is made available online, these are usually hosted at disparate sites, hardly discoverable, and not amenable to systematic analysis. In practice, this limits the ability of researchers to find these resources, and even to reuse the trajectories in large-scale efforts, e.g. for dynamic docking[243,244], discovering transient pockets[245], or associating variants with phenotypes[246].

There is an urgent need to find effective systems to share MD data[247,248] and to do so following the FAIR principles: Findable, Accessible, Interoperable, and Reusable[249]. Effectively sharing the simulation data not only is key to achieving transparency and reproducibility in the field of MD, but also enables researchers to reuse published data, avoiding the need to duplicate efforts and, in turn, accelerating research[250].

Fortunately, the research community is becoming more and more aware of the benefits of data sharing[251]. In general, we are witnessing a growing effort to make science more open, not only by researchers themselves but also, increasingly, by funders and

journals. Some disciplines, such as protein crystallography or genomics, have achieved to integrate open data into their workflow. However, in the case of MD simulations, these practices still have not become widely adopted. This is partly because of technical difficulties such as the large size of MD output files, but also because best-practice guidelines on how to share MD simulations are still being defined[247,252].

General-purpose data repositories like Zenodo (https://zenodo.org), FigShare (https://figshare.com), and Open Science Framework (https://osf.io), among others, provide an opportunity for researchers to deposit their simulation data. They accept a wide range of data types in a large variety of formats and provide global access to them. However, these resources sometimes do not provide enough space to sustainably store unfiltered MD simulation outputs. Moreover, they do not aim to integrate, harmonize, validate, or standardize the deposited data, and appropriate references and metadata are often not available. Thus, it can be difficult to filter suitable data from the huge variety of deposited files and databases.

In this sense, special-purpose MD sharing platforms that provide indexed and curated data are a promising solution. Particularly, platforms focused on simulations related to a specific research area or protein family, rather than a general database for all MD simulations, have more chance of success. This is because focused resources reduce hurdles like deposition space problems or too general and likely unused analysis and search options. Still, this type of resource comes with challenges as well, such as the difficulty to ensure the maintenance of the resource, usually requiring the support of a research community[247]. Likely due to the technical challenges associated with MD data and the mentioned sustainability limitations, only a modest number of online resources cover MD simulations (Table 1), as reviewed in refs. [247,253].

Moreover, to achieve effective sharing and dissemination of MD data, it is not enough with depositing the simulation files and trajectories in a repository. Even though this provides access to the simulation data, to analyze or even only to visually inspect this data on a local computer requires significant storage resources, as well as the installation and execution of specialized software. This typically increases the barrier for non-experts to extract the

underlying information. This issue can be solved by integrating web-based visualization and analysis tools into MD repositories[254], allowing an easy and interactive study of MD simulations for experts and non-experts likewise. Until recently, interactive visualization of MD simulations on the web has been hindered by the file sizes of trajectories. However, this is now possible thanks to software such as MDsrv[255], HTMoL[256], Mol*[257], and Mol-mil[258], which take advantage of the Web Graphics Library (WebGL) application programming interface (API) to provide fast three-dimensional graphics online[253]. With this, online repositories of MD data can not only contribute to making research more open, but also increase the reliability and understanding of this technique. In the end, this paves the way for data exchange between researchers of different fields, enhancing collaborative efforts and multidisciplinary.

Table 1. Selection of specialized MD databases

Name	Focus	Trajectory visualization	Analysis tools	Refs.
BIGNASim	Nucleic acids	Yes	Yes	[259]
BioExcel-CV19	COVID-19-related proteins	Yes	Yes	[260]
COVID-19 Molecular Structure and Therapeutics Hub	COVID-19-related proteins	No	No	[261]
Cyclo-lib	Cyclodextrins	No	Yes	[262]
GPCRmd	GPCRs	Yes	Yes	[263]
MemProtMD	Membrane proteins	No	Yes	[264]
MoDEL-CNS	Central nervous system proteins	Yes	Yes	[265]
MoDel/ MDWeb	Monomeric soluble proteins	No	Yes	[266]
NMRlipids	Lipid bilayers	No	No	[267]
SCoV2-MD	COVID-19-related proteins and variant data	Yes	Yes	[268]
TMB-iBIOMES	Nucleosome	No	No	[269]

2. OBJECTIVES

Given the relevance of MD simulations for the understanding of protein functionality and the benefits of making this data accessible to the research community, easy-to-use and efficient tools to share and inspect MD data are needed. This PhD thesis aims at the design and development of open-access online resources for the dissemination, visualization, and analysis of MD simulations, focusing on two types of pharmacologically relevant proteins: GPCRs and SARS-CoV-2-related proteins. Ultimately, these resources have the potential to boost transparency, reproducibility, and multidisciplinarity in the MD field.

In order to reach this goal, the following specific objectives were established:

1. To examine state-of-the-art knowledge on the capabilities and limitations of MD simulations for the understanding of the functionality of pharmacologically relevant proteins, focusing on GPCRs.
2. To design and implement the GPCRmd database, a community-driven online resource that provides access to MD simulations of most GPCR structures solved to date together with a set of tools to simplify the visualization and analysis of this data.
3. To design and implement the SCoV2-MD database, an online resource that integrates MD simulations of SARS-CoV-2 proteins from different resources and provides visualization and analysis tools with special emphasis on the prediction of the impact of known mutations on protein functionality.
4. To showcase the potential of GPCRmd and SCoV2-MD for exploring key aspects of protein dynamics related to GPCR and SARS-CoV-2 function.

The first objective of this thesis (objective 1) was accomplished by performing a set of comprehensive reviews (**publications 3.1, 3.2,** and **3.3**) of the applications, strengths, and challenges of protein

MD simulations. With this in mind, objectives 2 and 3 were addressed, respectively, in **publications 3.4** and **3.5**, where the implemented MD platforms are presented. Finally, objective 4 was approached, again, based on both **publications 3.4** and **3.5**.

3. PUBLICATIONS

3.1. Application of Biomolecular Simulations to G Protein-Coupled Receptors (GPCRs)

In this book chapter, we aim to introduce readers to the application of classical MD simulations for the study of GPCR functionality. As described in the *Introduction* section, the functionality of GPCRs is highly determined by their flexibility and ability to transition between distinct conformations. MD simulations can provide a high-resolution view of these structural motions. However, this technique has limitations. Here, we examine some of the challenges of classical MD simulations, focusing on the difficulties of sampling the whole conformational landscape of a GPCR and the limitation of accessible simulation timescales. We also provide some notions on how simulation data can be analyzed. Next, we outline several phenomena related to GPCR functionality that can be clarified with the application of MD simulations, and how this can positively impact the discovery of new and safer drugs. Moreover, we revise the history of MD simulations, discussing how the capabilities of this technique have expanded since the publication of the first protein simulation. Taking this into account, we finally address the future of the field, speculating how the potential of simulations could increase in the following years.

Torrens-Fontanals, M., Stepniewski, T. M., Rodríguez-Espigares, I. & Selent, J. Application of Biomolecular Simulations to G Protein-Coupled Receptors (GPCRs). in *Biomolecular Simulations in Structure-based Drug Discovery* (eds. Gervasio, F. L. & Spiwok, V.) 205–223 (Wiley, 2018). doi:10.1002/9783527806836.ch8

Application of Biomolecular Simulations to G Protein-Coupled Receptors (GPCRs)

Mariona Torrens-Fontanals[1], Tomasz M. Stepniewski[1], Ismael Rodríguez-Espigares[1], and Jana Selent[1]

1. Universitat Pompeu Fabra (UPF) – Hospital del Mar Medical Research Institute (IMIM), GPCR Drug Discovery Group, Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences of Pompeu Fabra University (UPF) – Hospital del Mar Medical Research Institute (IMIM), Dr. Aiguader 88, E-08003, Barcelona, Spain

Abstract

G protein-coupled receptors (GPCRs) are known to be highly dynamic proteins, which can exist in multiple distinct conformational states. This chapter focuses on atomistic unbiased simulations, which provide detailed and reliable structural information but face a series of challenges, such as simulating long molecular events. It discusses the challenges that molecular dynamics (MD)-based studies of GPCRs still face and how MD can provide unique insights into GPCR research. Given the fact that GPCRs are crucial drug targets, the insights yielded by MD simulations are especially relevant for the discovery of new and safer drugs. In this respect, drug designers are interested in understanding phenomena such as drug binding pathways, drug-receptor interaction, receptor subtype selectivity, conformational changes related to receptor (in)activation, and the role of lipids in GPCR modulation. The chapter reviews the advances of MD-based GPCR research since its origin and speculate about its future.

1. Introduction

G protein-coupled receptors (GPCRs) are major targets for the pharmaceutical industry and present an immense potential for future drug development [1]. Although GPCRs have been extensively studied over the past decades, the underlying molecular and structural mechanisms responsible for many critical regulatory

processes of this protein superfamily remain elusive, including signal transduction, allosteric modulation, functional selectivity, and activation [2].

GPCRs are known to be highly dynamic proteins, which can exist in multiple distinct conformational states. In fact, the signaling profile of GPCRs in response to a ligand depends critically on their ability to transition between different states [3]. Because of this, a promising approach to elucidate the molecular basis of GPCR functionality are molecular dynamics (MD) simulations, a potent computational technique capable of generating atomic-resolution simulations of the structural motions of a molecular system [4–6]. Importantly, these motions can be observed over timescales ranging from femtoseconds to milliseconds. As a consequence, MD provides, simultaneously, a temporal and structural resolution greater than what, currently, is achievable by experimental methods [7]. This is especially true for membrane proteins such as GPCRs, for which the experimental characterization of structural dynamics is particularly challenging [2, 8]. For this reason, and thanks to ongoing technology advances, MD simulations are increasingly being applied to the study of GPCR molecular systems, as reflected by the rapid upsurge of the number of publications per year concerning this topic (see Figure 1) [2, 6].

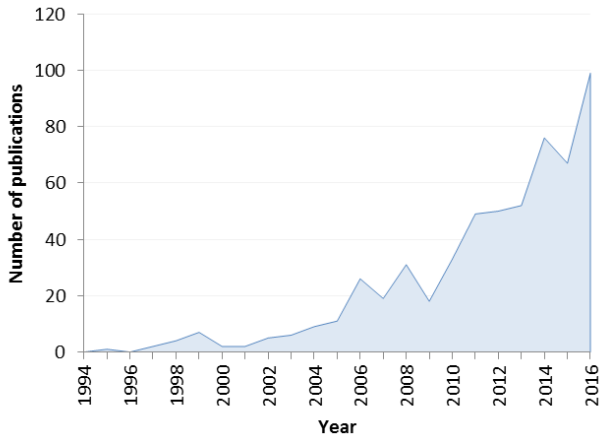


Figure 1. Number of publications per year indexed at Thomson Reuters' Web of Science that contain the topics "molecular dynamics" and ("GPCR" or "GPCRs"). The exponential growth of successful GPCR MD-based research is evidenced by the rapid upsurge in the number of publications per year related to this subject.

Different state-of-the-art simulation methods can be applied for the study of GPCRs, depending on the features of the receptor to be investigated. These methods include quantum chemical descriptors, classical mechanistic models with an atomistic or coarse-grain representation, and phenomenological system-level models. Each of them is characterized by the structural resolution, range of simulation timescales, and physicochemical accuracy they provide, which dictate the suitability of the model.

Considering the three criteria mentioned, for the study of GPCRs mechanistic models are typically the method of choice. These models are based on classical (i.e. Newtonian) mechanics, and thus the molecular system is treated as a set of classical particles. Interactions between these particles are derived from a set of empiric potential energy functions, known as force field, and quantum dynamical effects are not considered. Commonly, each system particle corresponds to a single atom (atomistic representations) or to a rigid set of atoms (coarse-grained representations) [9].

Atomistic representations capture the motion of molecular systems in full atomic detail, since each atom is represented as a point particle with its corresponding connectivity [10]. Consequently, this method leads to a better reproduction of the system's dynamics than coarse-grained representations, albeit at a higher computational cost.

Alternatively, coarse-grain representations consider a set of atoms as a single particle and each molecule as a set of particles, reducing the degrees of freedom of the system. This type of representation implies a substantial reduction of the computational requirements, and therefore can achieve longer simulation timescales. However, this is at the cost of a reduced spatial resolution [11].

In both mechanistic models, we can perform unbiased or biased simulations. Unbiased MD simulations sample the free energy landscape at equilibrium and explore the thermodynamically accessible energy landscape. This technique is limited by the inability to overcome high-energy barriers associated with large-scale motions, such as the receptor activation process. When facing such problems, biased techniques can be used to more efficiently

overcome energetic barriers and access additional conformational states. This is achieved by including an external bias, such as a compensating force or potential. On the other hand, these external forces may drive the simulation along unrealistic deformations, producing structural artifacts [12].

In this chapter, we focus on atomistic unbiased simulations (see Figure 2), which provide detailed and reliable structural information but face a series of challenges, such as simulating long molecular events. With this premise, we discuss the challenges that MD-based studies of GPCRs still face and how MD can provide unique insights into GPCR research. Moreover, we review the advances of MD-based GPCR research since its origin and speculate about its future.

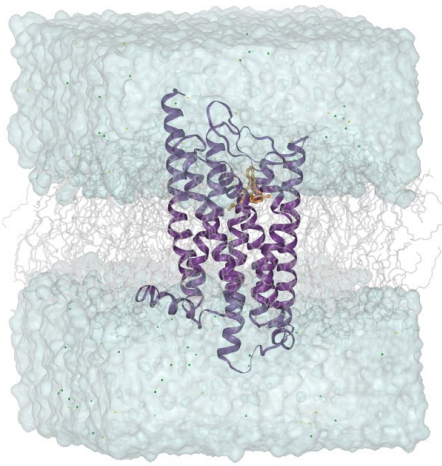


Figure 2. Representation of an atomistic MD simulation including a GPCR (purple) bound to a ligand (orange), the phospholipid bilayer (gray), and intercellular and intracellular water (blue).

2. MD Simulations for Studying the Conformational Plasticity of GPCRs

Due to the inherent structural plasticity of GPCRs, these receptors can acquire an essentially infinite number of conformational states. In fact, several lines of evidence indicate that GPCRs do not present just one single active state and one single inactive state: they can assume multiple distinct states in each case [13–17]. MD

simulations can provide a useful insight on the transitions between conformational states, such as inactive, intermediate, and active states (see Section 3). However, at present, MD-based studies of transition processes are limited in several respects, such as the need for simulations to extend to the millisecond timescale and the difficulty inherent in sufficiently sampling structural fluctuations to properly characterize long molecular events [18]. These constraints pose an obstacle to structure-based drug design for GPCR targets, since understanding the role of intermediate states is crucial for the comprehension of the molecular mechanisms behind receptor activation and inactivation, together with its pharmacological action. Moreover, deciphering the mechanisms by which ligands can differentially induce or stabilize different receptor populations would lead us to the design of drugs that specifically target the conformational state responsible for a desired intracellular response. This would allow, for instance, the rational design of drugs with reduced side effects [3, 19].

2.1. Challenges in GPCR Simulations: The Sampling Problem and Simulation Timescales

The dynamic behavior of a molecular system is intrinsically dictated by its free energy. Complex biomolecules such as GPCRs present a rugged energy landscape (see Figure 3), with many local minima frequently separated by high-energy barriers [20]. The conformational transition between different states requires crossing some of these high-energy barriers, which is a slow process [21]. Indeed, this shape of the free energy landscape makes it easy for systems to get trapped in one of the huge number of local minimum-energy states for a long time, which restrains the sampling process and, in turn, leads to a poor characterization of a protein's dynamic behavior [22, 23]. Therefore, most unbiased simulations explore just a small region around the energy minimum closest to the initial conformation [24]. This inadequate sampling of the conformational landscape limits the ability to analyze and reveal functional properties of the systems being examined [25]. Due to the sampling problem, complex processes such as GPCR activation or inactivation are commonly beyond the capabilities of straightforward unbiased MD simulations [25].

Furthermore, simulations are currently limited to lengths of a few microseconds on publicly available hardware, or a few milliseconds if specialized hardware [26] or distributed computing techniques [27] are used. These simulation times are short compared to many of the events of greatest interest in molecular physiology, which take place on longer timescales [8]. This is the case for large conformational rearrangements in GPCRs. For example, the conformational changes from the inactive to the active state of a GPCR occur at the millisecond to second timescale (see Figure 4) [28]. The time limitation in conventional MD simulation approaches also hampers the routine exploration of drug (un)binding, making *in silico* GPCR-related drug discovery a challenging task. Nevertheless, we experience an exciting era for GPCR research, as we expect to simulate sufficient time frames within the next five years (see Section 4).

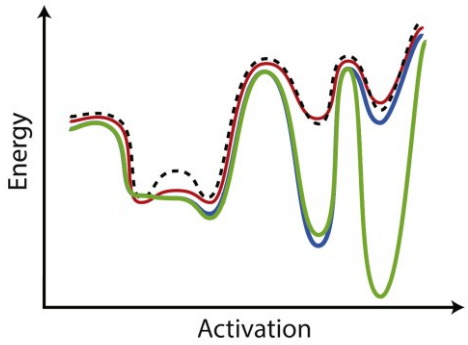


Figure 3. Simplified energy landscape of a GPCR in the absence of ligand (dashed line) and in the presence of an inverse agonist (red), a full agonist (blue), and a full agonist with a G protein (green). Source: Adapted with permission from Ref. [15].

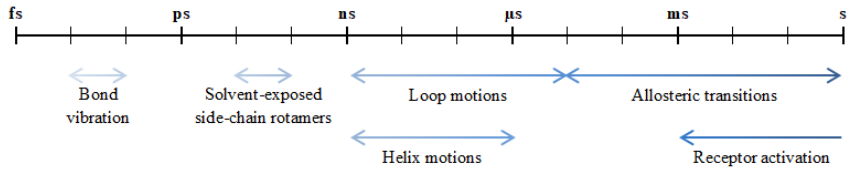


Figure 4. Schema of the characteristic timescales for GPCR motions (time magnitudes in log scale) [2, 9, 28, 29].

2.2. Making Sense Out of Simulation Data

Once sufficient sampling time is obtained from unbiased all-atom MD models, one of the hardest problems is to extract information about the kinetics of relevant biological events (e.g. ligand binding) [27]. An interesting approach to overcome this problem is the application of dimensionality reduction methods. Essentially, these methods are mathematical tools that reduce the complexity of a system while minimizing the loss of information. This is possible due to the interdependence between different variables of the system, imposed by energetic constraints among atom positions [30]. Ultimately, dimensionality reduction helps visualize complex energy landscapes, improving the efficiency of simulation, analysis, and optimization. As a consequence, these methods are considered an important tool for the comprehensive analysis of MD data [31]. Among these techniques, the most popular are the ones that imply linear dimensionality reduction, especially principal component analysis (PCA) and time-dependent independent component analysis (tICA). For instance, PCA has been applied to study simulation data of the adenosine A_{2A} receptor in both the presence and absence of an inverse agonist, providing information on the convergence and reproducibility of the results. The authors found that the addition of the inverse agonist greatly improves the stability of all receptor helices. Moreover, they concluded that PCA provides a more robust assessment of convergence and sampling than other commonly used criteria such as root-mean-square deviation (RMSD), especially in comparative studies [32].

Unfortunately, the above mentioned methods (tICA and PCA) do not provide quantitative information about statistical significance or sampling quality [33]. Still, there exist other approaches that are capable of generating more robust analyses by first partitioning the conformational space into discrete conformational states and subsequently calculating the transition rates or probabilities between them, based on transitions observed in MD trajectories. The analysis of these transitions allows the reconstruction of the global behavior of the system [34, 35]. The resulting models are known as Markov state models (MSMs), where “Markovianity” means that the kinetics are modeled by a memoryless jump process between states [36]. Once the model is obtained, its reliability,

convergence, and sampling quality can be tested by several methods such as Chapman-Kolmogorov tests [37], among others [38–40].

A disadvantage of applying classical MSMs is that the sampling of rare events can still be highly inefficient, which may result in improperly connected models and imprecise residence times. This can typically be solved by extending the simulation time, which is unpractical when computational resources are limited [27]. Furthermore, the sampling of the conformational space and the MSM can be improved by adaptive sampling, a type of biased sampling that identifies underexplored conformational states in MD data and generates new starting points to explore those states by resampling. Therefore, instead of requiring long simulations to model rare events, several shorter MD trajectories (∼10–100 ns) can be used. This facilitates the parallelization of the computational burden of the MD among many processors [41]. A successful application of MSM adaptive sampling has recently contributed to a better understanding of functional selectivity – the phenomena explaining receptor's selectivity for certain signal transduction pathways in front of others due to the interaction with a biased ligand – in the μ-opioid receptor. The μ-opioid receptor was simulated in complex with either a balanced or a G protein-biased agonist, and MSM analysis allowed identifying a differential ligand-specific dynamic behavior of the receptor with kinetically distinct conformational states. Such information can drive the rational design of functionally selective ligands that stabilize the receptor in a specific state, which may eventually be developed into improved drugs [42].

3. Application of MD Simulations to GPCR Drug Design: Why Should We Use MD?

MD simulations provide unique insights that can be critical for the understanding of GPCRs. Given the fact that GPCRs are crucial drug targets, the insights yielded by MD simulations are especially relevant for the discovery of new and safer drugs [1, 43]. In this respect, drug designers are interested in understanding phenomena such as drug binding pathways, drug-receptor interaction, receptor subtype selectivity, conformational changes related to receptor (in)activation, and the role of lipids in GPCR modulation.

The entrance of a drug into the receptor binding site is a multistep process, with diverse metastable and/or intermediate binding sites. The analysis of the crystal structures of GPCRs bound to a ligand grants information on the binding mode of the ligand, but not on the pathway that it follows from the extracellular phase into the binding pocket. MD simulations generate useful insights on such binding pathways, since they enable the study of the development of the molecular system over a given timescale and provide an estimation of the rate at which a ligand associates or dissociates to the receptor. Having knowledge on the ligand binding pathways allows the identification of energetic barriers encountered along those pathways. Energetic barriers can substantially affect the binding and unbinding rates of a ligand into a receptor. Therefore, such information constitutes a foundation for the rational optimization of drug binding and unbinding kinetics [44], which play a critical role in drug efficacy, selectivity, and safety [45–47]. Moreover, the identification of metastable binding sites may be useful for the development of dimeric or bivalent compounds that can bind to different metastable binding sites, which may result in an increased affinity [48]. Some examples of ligands for which the binding pathway was described thanks to MD are (*S*)-alprenolol binding to the β_2-adrenergic receptor (β_2AR), histamine to the histamine H_4 receptor (hH_4R), and clozapine and haloperidol to the dopamine receptor D_3 (D_3R) [44, 49, 50]. Particularly relevant is the case of the β_2AR, which was the first unbiased MD simulation study capturing the full process of ligands spontaneously binding to a GPCR, achieving the final poses of the ligands without incorporation of any prior knowledge of the binding site. Results revealed not only the predominant pathway into the binding site but also the two main energetic barriers that govern drug binding and unbinding kinetics [44].

Concerning drug-receptor interaction, MD studies of this process can provide valuable information that cannot be obtained through rigid docking methods. Such insights are particularly useful for the identification or design of new ligands for a given GPCR. This is possible because MD simulations take into account the flexibility of GPCR binding pockets, which differ from one conformational state of the receptor to another and even within a single global receptor conformational state [51]. Indeed, considering multiple possible receptor structures generally increases the diversity of ligands

identified. Furthermore, the goal of GPCR-related drug discovery is typically to find a ligand that not only binds to the target but also achieves a particular signaling profile. For example, the study of biased ligands, which selectively engage one signaling pathway downstream of the receptor over the others, is remarkably promising for drug design, since they can be used to inhibit the disease-associated pathways while stimulating non-pathological pathways up to their physiological levels, avoiding side effects [52]. The signaling outcome triggered by a GPCR depends on the conformational state of the receptor and thus of the binding pocket that is stabilized by the drug. In this respect, simulations can be used to compare drug-receptor interactions in different conformational states of the receptor or using different types of ligands (e.g. unbiased agonist, biased agonist, inverse agonist or antagonists) [2]. A successful story of identifying new ligands with an unprecedented level of bias was carried out for the serotonin 5-hydroxytryptamine receptor 2A (5-HT_{2A}). Martí-Solano et al. used extensive MD simulations to characterize the dynamics of ligand-receptor interactions of known biased and balanced agonists with the purpose of becoming capable of discriminating the different types of receptor agonists. Thanks to this information, they discovered new biased ligands of outstanding efficacy by tuning the structure of the balanced natural ligand serotonin. These compounds represent valuable tools to promote the design of improved, safer antipsychotic drugs [53]. Another fruitful study concerning biased agonists is the recent development of an approach for designing β-arrestin-biased ligands for the dopamine receptor D_2 (D_2R) based on structural and MD data. Particularly, specific conserved receptor-ligand contacts responsible for biased signaling were identified and modified in order to develop new biased agonists for this GPCR [54].

Another important goal of drug discovery is the creation of drugs that specifically target one receptor subtype over another. Many GPCR subtypes exhibit a highly conserved orthosteric binding site, such that a single ligand can bind to several receptors simultaneously, contributing to off-target side effects [55]. An effective strategy to achieve better selectivity is the design of drugs that target allosteric pockets, which are less conserved among GPCRs [56]. Besides their specificity, allosteric modulators have the advantage of providing a strategy to fine-tune cellular responses

triggered by a ligand. Allosteric modulators can bind to a GPCR concomitantly to the orthosteric ligand, altering its binding affinity and/or cellular-signaling efficacy in a moderate manner avoiding severe effects on the cell [57–59]. For this reason, allostery has become an area of great interest for the discovery of new drugs with reduced side effects [59]. However, allosteric binding sites are not evident from crystal structures, which hamper the structure-based design of allosteric drugs. Moreover, the molecular mechanisms by which such modulators affect GPCR signaling may depend on dynamical properties that would not be evident from a single static structure. MD simulations can provide a means to detect hidden allosteric binding sites and determine the mechanistic basis of allosteric regulation [60]. One representative case of a GPCR family displaying dramatically low subtype selectivity concerning orthosteric ligands are the muscarinic acetylcholine receptors. This family comprises important drug targets for several central nervous system diseases [61]. MD-based approaches have been critical for the efficient identification of allosteric mechanisms in muscarinic receptors [58, 62], and such findings are usually paradigmatic for the entire GPCR superfamily [61]. For example, atomic-level MD simulations by Dror et al. provided a structural basis of allosteric ligand binding and elucidated mechanisms of cooperativity between the allosteric and the orthosteric ligand. These studies enabled the design of chemical modifications that substantially alter the final effect of allosteric modulators [62].

Besides GPCR subtypes, it can also be informative to study the differences between wild-type GPCRs and variants responsible for disease susceptibility or a distinct drug response. Such functional differences have been predicted to be caused by variations in ligand binding site accessibility and structure, which are determined by the dynamics of the receptor. Hence, MD simulations are a promising approach to elucidate the molecular mechanisms responsible for such interindividual differences. This remains an emerging field of study that could provide important clues on the functional single-nucleotide polymorphisms (SNPs) identified in the recent large-scale sequencing initiatives [12]. A well-defined case of SNP in GPCRs is the Arg16Gly variant of the β_2AR, which has been linked to a differential response to albuterol, a β_2AR agonist commonly used in the treatment of asthma. MD analysis of the conformational differences between the two variants revealed that divergences in

the position and dynamics of the N-terminal region, where the SNP is located, lead to long-range effects at the ligand binding site, reducing the accessibility of the ligand at one of the variants [63].

Another phenomenon of interest is GPCR (in)activation, which is still not completely understood. During the process of activation, GPCRs undergo global conformational changes visiting inactive, intermediate, and active states. Currently available GPCR crystal structures provide high-resolution insights into some of these conformational states [64, 65]. However, these structures represent snapshots of GPCRs in a given conformation, and they do not reveal the mechanism by which they transition between different states. A promising approach to tackle this issue is combining high-resolution experimental data with MD simulation studies. MD simulations can provide detailed information on the transitions between conformational states, helping us explore the GPCR conformational landscape in the presence of different ligands. A notorious case of successful MD-based study of GPCR activation was performed by Dror et al. They proposed an activation mechanism for the β_2AR based on atomic-level simulations in which an agonist-bound receptor transitioned spontaneously from the active to the inactive conformation. Interestingly, their simulations highlighted the existence of an intermediate state that was suggested to represent a receptor conformation to which G protein binds during activation [14]. Yuan et al. also used MD to monitor the activation process of the β_2AR, along with the adenosine A_{2A} receptor and rhodopsin. Their simulations allowed the characterization of the formation of an intrinsic water pathway, which, in the receptors' resting state, is interrupted by a hydrophobic layer of amino acid residues. Upon agonist binding, this hydrophobic layer opens to allow the formation of a continuous intrinsic water channel [17]. Another highlight was achieved by Kohlhoff et al. when they captured the mechanism of β_2AR (in)activation in a simulation of unprecedented total length thanks to Google's Exacycle cloud computing platform. Such a study revealed multiple (in)activation pathways, and showed that agonists and inverse agonists act by modulating the receptor dynamics to prefer different pathways [66].

Finally, a lot of attention has recently been devoted to the role of the membrane environment or of specific membrane lipids in GPCR

functioning. Membrane phospholipids have been found to allosterically modulate the activity [67] and oligomerization [68] of GPCRs, while membrane cholesterol can regulate its stability, ligand-binding properties and function [69–71]. Still, the precise nature of cholesterol's implication in GPCR modulation is a matter of debate [71]. Cholesterol modulation could be due to indirect effects (changes in membrane biophysical properties) [72, 73], direct cholesterol-GPCR interactions [74–77] or both. Moreover, some studies suggest a potential allosteric role of cholesterol [78], while others propose that cholesterol molecules can compete with orthosteric ligands by entering the receptor's binding site [78–80]. Understanding the influence of cholesterol on GPCR function would allow us to explore potential therapeutic uses of membrane sterols or sterol-mimetic molecules in GPCR drug discovery. Still, molecular mechanisms behind cholesterol modulation are difficult to ascertain using experimental methods alone, and thus MD simulations are increasingly being applied. For example, Manna et al. used extensive MD simulations to provide detailed insights into how the β_2AR is allosterically modulated by cholesterol. Their simulations showed a decrease in the receptor's conformational flexibility due to cholesterol binding at specific high-affinity sites located near the transmembrane helices 5–7 [81]. On the other side, Guixà-González et al. recently showed that cholesterol can spontaneously enter the adenosine A_{2A} receptor's binding pocket from the membrane milieu, unveiling a new interaction mode between cholesterol and this receptor that could potentially apply to other GPCRs [79].

4. Evolution of MD Timescales

It has been 40 years since the publication of the first MD simulation of a macromolecule of biological interest, which was instrumental in replacing the view of proteins as relatively rigid structures with the understanding that they are dynamic systems. That first simulation consisted of a 9.2-ps trajectory of a small protein, the bovine pancreatic trypsin inhibitor (BPTI, ~500 atoms), in vacuum [82]. From that point forward, thanks to continuous advances in both methodology and computational resources, MD simulations have gradually extended to larger systems and longer timescales. Interestingly, the first published MD simulation of a GPCR was

reported in 1991, with a simulation time of 80 ps [83] (see Figure 5).

A noteworthy event that highly influenced the development of MD capacities was the release of the distributed volunteer computing project Folding@home in the year 2000. Thanks to its distributed computing environment, the release of Folding@home made it possible to break previous computational barriers for MD simulations, reaching timescales of orders of magnitude longer than what had previously been achieved [84, 85]. At its release, it was already able to simulate the folding process of the 23-residue protein BBA5, which has a duration of approximately 10 μs, by producing thousands of independent, short trajectories totaling an accumulated simulation time of hundreds of microseconds [86].

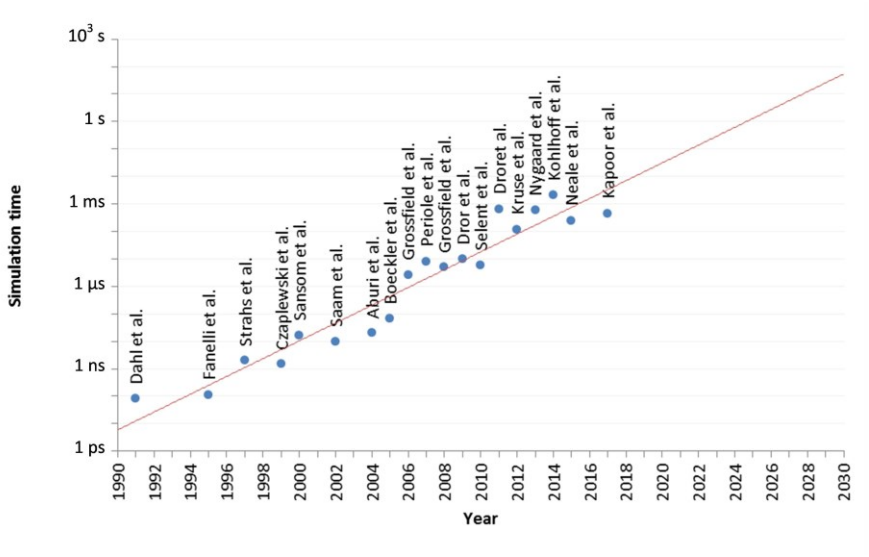


Figure 5. Evolution of the timescales accessible to atomistic MD simulations of GPCRs (in logarithmic scale). Based on the simulation time of relevant publications from the first MD simulation of a GPCR to the present, the accessible timescales have increased exponentially. An exponential function (solid line) was fit to the data points.

The subsequent breakthrough in the MD field occurred at 2008, when the Anton supercomputer was created. What made Anton stand out from previous hardware was the fact that it is a specialized, massively parallel supercomputer particularly designed to perform MD simulations of molecular systems. This specificity

makes Anton highly efficient, being able to increase the speed of MD computations and to generate simulations with longer timescales. For example, large systems such as a GPCR embedded in a lipid bilayer (~105 atoms) could be simulated for hundreds of microseconds [87, 88]. This is the case of the β_2AR activation studies conducted by Dror et al., in which a total simulation time of more than 600 μs was reached thanks to Anton [14]. Moreover, Anton simulations of the muscarinic M_2 and M_3 receptors captured the binding of an important bronchodilator drug at an allosteric site, offering the first structural view of an orthosteric GPCR ligand binding to an allosteric site [89].

The enhancement in the performance of MD simulations has also been heavily influenced by algorithmic advances, including the fine-tuning of energy calculations, parallelization improvements and the use of graphical processing units (GPUs) [24]. Specifically, the development of algorithms exploiting GPUs has become a major breakthrough in simulation codes. During the late 2000s, such progresses in GPU exploitation allowed to use them as performance accelerators for a wide variety of scientific applications, including MD [90, 91]. For example, the use of GPUs has made the simulation of GPCRs in explicit lipid-water environments feasible within reasonable computational times [92]. In 2007, a volunteer-distributed computing project based on GPU resources named GPUGRID was launched, with the capability of simulating thousands of all-atom molecular trajectories at an average of 20 ns/day each (for systems of ~30 000–80 000 atoms) [93]. In 2010, this approach was used to reveal for the first time the molecular details of sodium ion binding to its allosteric binding site for a class A GPCR, accumulating more than 6 μs of simulation time for a system of about 60 000 atoms [94].

Despite the advances in specialized hardware and software for MD, such as supercomputers and distributed computing platforms, most of these tools are not openly available to the whole scientific community and can only be accessed by a set of research groups, which hampers the advance of the study of long-timescale processes. In this respect, cloud computing is a valuable alternative that can bring long-timescale processes, in particular those associated with GPCRs, within reach of a broader community.

All in all, current simulation timescales reach microseconds (conventional hardware [27]) to milliseconds (specialized hardware [26] / distributed computing [96, 97]). Simulations of systems having ~50 000–100 000 atoms are now routine, and simulations of approximately 500 000 atoms are common when appropriate computer facilities are available [24]. Considering the evolution of GPCR-related MD since the publication of the first MD simulation study, it is undeniable that there has been a dramatic increase in the timescales we are able to achieve. This tendency is highly influenced by the technological development, both in terms of computational cost reduction and methodological improvement. Certainly, the evolution of GPCR-related MD is a direct consequence of Moore's law, which describes how the performance of integrated circuits has been increasing exponentially over the past half-century [98]. Based on this trend, we can extrapolate that the second timescale for GPCR simulations will be reached approximately in 2024 (see Figure 5). In fact, several independent authors have reached similar conclusions [92, 99].

5. Sharing MD Data via a Public Database

In view of the growing relevance of MD simulations for the study of GPCRs (see Figures 1 and 5), it is increasingly necessary to find efficient systems for the storage, indexing, and dissemination of all the simulation data being generated. For this reason, initiatives involving the development of MD databases are gaining importance. Regarding GPCRs, the GPCRmd database [100] is being created with the purpose of fostering GPCR MD data from all over the world. Broadly speaking, GPCRmd offers storage of MD data, citable IDs for referencing such data in publications, basic analysis tools, and online visualization of the simulations. Being the first GPCR-specific MD database, GPCRmd sets standards in archiving simulation data of this protein family in a structured and searchable manner with the aim of facilitating data retrieval for other scientists.

One important advantage of GPCRmd is that, by providing public access to the data on which scientific articles are based, it increases the transparency and credibility of MD studies. This allows other researchers to evaluate the quality of the obtained results and

enhances reproducibility, which is critical for the advancement of GPCR research.

Finally, tools such as the GPCRmd database also contribute to the popularization of MD among the research community by offering freely available web resources, including user-friendly representation and analysis options as well as the display of results with easy-to-interpret and interactive graphs. These resources approach GPCR MD data to all researchers studying this protein family, including to those who are not familiarized with computational tools or complex MD analysis software, and contributes to the widespread adoption of MD.

6. Conclusions and Perspectives

MD simulations can provide a high-resolution view of the dynamics of biological systems relevant to human health. They can either capture atomic-level motions within conformational states or structural transitions between different conformational populations, bringing within reach information that is highly difficult to obtain by other methods [2]. For this reason, MD simulations are specially promising for the study of the functionality of complex signaling proteins such as GPCRs, which are drug targets of striking importance in the pharmaceutical industry [5].

However, there are still important drawbacks to the usability of MD simulations for the study of GPCR functionality. One main limitation to this technique is the generation of simulations long enough to allow an adequate sampling of biological events such as receptor activation upon agonist binding [12, 92].

Fortunately, with ongoing advances in both MD hardware and software, GPCR-related MD simulations are being extended to larger systems and longer timescales. If this trend continues, we will soon reach simulation times of the order of a second (see Figure 5), bridging the gap between the timescales of biological processes observed *in vivo* and those accessible *in silico*. This increase of the simulation time will grant us the opportunity to extend the application of MD simulations to the study of processes that were previously difficult to analyze through this method, including global

conformational rearrangements, receptor dimerization, and coupling to intracellular signaling proteins [24].

Given the huge druggability of GPCRs, the predicted progress of MD simulations of this protein superfamily will have a great impact on drug discovery. Despite multiple studies underlining the importance of GPCR flexibility to molecular recognition and signaling, most drug discovery programs currently disregard MD analysis because of their computational expenses. With the forthcoming reduction of the computational costs associated to MD simulations, this technique is expected to be more commonly applied at the pharmaceutical industry and, eventually, to be commonly included in drug discovery pipelines [8, 9, 88]. Particularly, MD can be of great use through the hit discovery, hit-to-lead and lead optimization processes by identifying possible ligands for a given GPCR and assessing the stability and dynamics of the binding poses, along with receptor-ligand binding affinities and even kinetics [101].

Acknowledgments

M.T.-F. acknowledges financial support from the Spanish Ministry of Education, Culture and Sport (FPU16/01209), T.M.S acknowledges financial support from Hospital del Mar Medical Research Institute and help from the National Science Centre of Poland, project number 2017/27/N/NZ2/02571, I.R.-E. acknowledges Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (2015 FI_B00145) for its financial support, J.S. acknowledges financial support from Instituto de Salud Carlos III FEDER (PI15/00460).

References

1. Santos, R., Ursu, O., Gaulton, A. et al. (2017). A comprehensive map of molecular drug targets. Nat. Rev. Drug Discov. 16 (1): 19–34.
2. Latorraca, N.R., Venkatakrishnan, A.J., and Dror, R.O. (2017). GPCR dynamics: structures in motion. Chem. Rev. 117 (1): 139–155.
3. Martí-Solano, M., Schmidt, D., Kolb, P., and Selent, J. (2016). Drugging specific conformational states of GPCRs: challenges and opportunities for computational chemistry. Drug Discov. Today 21 (4): 625–631.

4. Rosenbaum, D.M., Rasmussen, S.G.F., and Kobilka, B.K. (2009). The structure and function of G-protein-coupled receptors. Nature 459 (7245): 356–363.
5. Ciancetta, A., Sabbadin, D., Federico, S. et al. (2015). Advances in computational techniques to study GPCR–ligand recognition. Trends Pharmacol. Sci. 36 (12): 878–890.
6. McRobb, F.M., Negri, A., Beuming, T., and Sherman, W. (2016). Molecular dynamics techniques for modeling G protein-coupled receptors. Curr. Opin. Pharmacol. 30: 69–75.
7. Johnston, J.M. and Filizola, M. (2011). Showcasing modern molecular dynamics simulations of membrane proteins through G protein-coupled receptors. Curr. Opin. Struct. Biol. 21 (4): 552–558.
8. Dror, R.O., Jensen, M.Ø., Borhani, D.W., and Shaw, D.E. (2010). Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. J. Gen. Physiol. 135 (6): 555–562.
9. Adcock, S.A. and McCammon, J.A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. Chem. Rev. 106 (5): 1589–1615.
10. Hug, S. (2013). Classical molecular dynamics in a nutshell. Methods Mol. Biol. (Clifton, N.J.) 924: 127–152.
11. Bond, P.J., Holyoake, J., Ivetac, A. et al. (2007). Coarse-grained molecular dynamics simulations of membrane proteins and peptides. J. Struct. Biol. 157 (3): 593–605.
12. Sengupta, D., Joshi, M., Athale, C.A., and Chattopadhyay, A. (2016). What can simulations tell us about GPCRs: integrating the scales. Methods Cell Biol. 132: 429–452.
13. Nygaard, R., Zou, Y., Dror, R.O. et al. (2013). The dynamic process of β2-adrenergic receptor activation. Cell 152 (3): 532–542.
14. Dror, R.O., Arlow, D.H., Maragakis, P. et al. (2011). Activation mechanism of the β2-adrenergic receptor. Proc. Natl. Acad. Sci. U.S.A. 108 (46): 18684–18689.
15. Manglik, A.I., Kim, T.H., Masureel, M. et al. (2015). Structural insights into the dynamic process of β2-adrenergic receptor signaling. Cell 161 (5): 1101–1111.
16. Staus, D.P., Strachan, R.T., Manglik, A. et al. (2016). Allosteric nanobodies reveal the dynamic range and diverse mechanisms of G-protein-coupled receptor activation. Nature 535 (7612): 448–452.
17. Yuan, S., Filipek, S., Palczewski, K., and Vogel, H. (2014). Activation of G-protein-coupled receptors correlates with the formation of a continuous internal water pathway. Nat. Commun. 5: 4733.
18. Freddolino, P.L., Harrison, C.B., Liu, Y., and Schulten, K. (2010). Challenges in protein folding simulations: timescale, representation, and analysis. Nat. Phys. 6 (10): 751–758.
19. Martí-Solano, M., Guixà-González, R., Sanz, F. et al. (2013). Novel insights into biased agonism at G protein-coupled receptors and their potential for drug design. Curr. Pharm. Des. 19 (28): 5156–5166.

20. Okamoto, Y. (2004). Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. J. Mol.Graph. Model. 22 (5): 425–439.
21. Lei, H. and Duan, Y. (2007). Improved sampling methods for molecular simulation. Curr. Opin. Struct. Biol. 17 (2): 187–191.
22. Liao, C. and Zhou, J. (2014). Replica-exchange molecular dynamics simulation of basic fibroblast growth factor adsorption on hydroxyapatite. J. Phys. Chem. B 118 (22): 5843–5852.
23. Chen, C., Huang, Y., and Xiao, Y. (2013). Enhanced sampling of molecular dynamics simulation of peptides and proteins by double coupling to thermal bath. J. Biomol. Struct. Dyn. 31 (2): 206–214.
24. Hospital, A., Goñi, J.R., Orozco, M., and Gelpí, J.L. (2015). Molecular dynamics simulations: advances and applications. Adv. Appl. Bioinform. Chem. 8: 37–47.
25. Bernardi, R.C., Melo, M.C.R., and Schulten, K. (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. Biochim. Biophys. Acta, Gen. Subj. 1850 (5): 872–877.
26. Shaw, D.E., Maragakis, P., Lindorff-Larsen, K. et al. (2010). Atomic-level characterization of the structural dynamics of proteins. Science 330 (6002): 341–346.
27. Paul, F., Wehmeyer, C., Abualrous, E.T. et al. (2017). Protein–peptide association kinetics beyond the seconds timescale from atomistic simulations. Nat. Commun. 8 (1): 1095.
28. Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. Nature 450 (7172): 964–972.
29. Manglik, A. and Kobilka, B. (2014). The role of protein dynamics in GPCR function: insights from the β2AR and rhodopsin. Curr. Opin. Cell Biol. 27: 136–143.
30. Rodríguez-Espigares, I., Kaczor, A.A., and Selent, J. (2016). In silico exploration of the conformational universe of GPCRs. Mol. Inform. 35 (6–7): 227–237.
31. Brown, W.M., Martin, S., Pollock, S.N. et al. (2008). Algorithmic dimensionality reduction for molecular structure analysis. J. Chem. Phys. 129 (6): 64118.
32. Ng, H.W., Laughton, C.A., and Doughty, S.W. (2013). Molecular Dynamics simulations of the adenosine A2a receptor: structural stability, sampling, and convergence. J. Chem. Inf. Model. 53 (5): 1168–1178.
33. Grossfield, A. and Zuckerman, D.M. (2009). Quantifying uncertainty and sampling quality in biomolecular simulations. Annu. Rep. Comput. Chem. 5: 23–48.
34. Pan, A.C. and Roux, B. (2008). Building Markov state models along pathways to determine free energies and rates of transitions. J. Chem. Phys. 129 (6): 64107.
35. Noé, F. and Fischer, S. (2008). Transition networks for modeling the kinetics of conformational change in macromolecules. Curr. Opin. Struct. Biol. 18 (2): 154–162.
36. Prinz, J.-H., Wu, H., Sarich, M. et al. (2011). Markov models of molecular kinetics: generation and validation. J. Chem. Phys. 134 (17): 174105.

37. Noé, F., Schütte, C., Vanden-Eijnden, E. et al. (2009). Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proc. Natl. Acad. Sci. U.S.A. 106 (45): 19011–19016.
38. Swope, W.C., Pitera, J.W., and Suits, F. (2004). Describing protein folding kinetics by molecular dynamics simulations. J. Phys. Chem. B 108 (21): 6571–6581.
39. Park, S. and Pande, V.S. (2006). Validation of Markov state models using Shannon's entropy. J. Chem. Phys. 124 (5): 54118.
40. Bacallado, S., Chodera, J.D., and Pande, V. (2009). Bayesian comparison of Markov models of molecular dynamics with detailed balance constraint. J. Chem. Phys. 131 (4): 45106.
41. Bowman, G.R., Ensign, D.L., and Pande, V.S. (2010). Enhanced modeling via network theory: adaptive sampling of Markov state models. J. Chem. Theory Comput. 6 (3): 787–794.
42. Kapoor, A., Martinez-Rosell, G., Provasi, D. et al. (2017). Dynamic and kinetic elements of μ-opioid receptor functional selectivity. Sci. Rep. 7 (1): 11255.
43. Lundstrom, K. (2006). Latest development in drug discovery on G protein-coupled receptors. Curr. Protein Pept. Sci. 7 (5): 465–470.
44. Dror, R.O., Pan, A.C., Arlow, D.H. et al. (2011). Pathway and mechanism of drug binding to G-protein-coupled receptors. Proc. Natl. Acad. Sci. U.S.A. 108 (32): 13118–13123.
45. Guo, D., Mulder-Krieger, T., IJzerman, A.P., and Heitman, L.H. (2012). Functional efficacy of adenosine A2A receptor agonists is positively correlated to their receptor residence time. Br. J. Pharmacol. 166 (6): 1846–1859.
46. Swinney, D.C. (2008). Applications of binding kinetics to drug discovery. Pharm. Med. 22 (1): 23–34.
47. Lu, H. and Tonge, P.J. (2010). Drug-target residence time: critical information for lead optimization. Curr. Opin. Chem. Biol. 14 (4): 467–474.
48. Strasser, A., Wittmann, H.-J., and Seifert, R. (2017). Binding kinetics and pathways of ligands to GPCRs. Trends Pharmacol. Sci. 38 (8): 717–732.
49. Wittmann, H.-J. and Strasser, A. (2015). Binding pathway of histamine to the hH4R, observed by unconstrained molecular dynamics. Bioorg. Med. Chem .Lett. 25 (6): 1259–1268.
50. Thomas, T., Fang, Y., Yuriev, E., and Chalmers, D.K. (2016). Ligand binding pathways of clozapine and haloperidol in the dopamine D2 and D3 receptors. J. Chem. Inf. Model. 56 (2): 308–321.
51. Lebon, G., Warne, T., and Tate, C.G. (2012). Agonist-bound structures of G protein-coupled receptors. Curr. Opin. Struct. Biol. 22 (4): 482–490.
52. Shukla, A.K., Singh, G., and Ghosh, E. (2014). Emerging structural insights into biased GPCR signaling. Trends Biochem. Sci. 39 (12): 594–602.
53. Martí-Solano, M., Iglesias, A., de Fabritiis, G. et al. (2015). Detection of new biased agonists for the serotonin 5-HT2A receptor: modeling and experimental validation. Mol. Pharmacol. 87 (4): 740–746.

54. McCorvy, J.D., Butler, K.V., Kelly, B. et al. (2017). Structure-inspired design of β-arrestin-biased ligands for aminergic GPCRs. Nat. Chem. Biol. 14 (2): 126–134.
55. Allen, J.A. and Roth, B.L. (2011). Strategies to discover unexpected targets for drugs active at G protein–coupled receptors. Annu. Rev. Pharmacol. Toxicol. 51 (1): 117–144.
56. Christopoulos, A. (2014). Advances in G protein-coupled receptor allostery: from function to structure. Mol. Pharmacol. 86 (5): 463–478.
57. Jakubík, J. and El-Fakahany, E.E. (2010). Allosteric modulation of muscarínic acetylcholine receptors. Pharmaceuticals (Basel). 3 (9): 2838–2860.
58. Miao, Y., Goldfeld, D.A., Von Moo, E. et al. (2016). Accelerated structure-based design of chemically diverse allosteric modulators of a muscarinic G protein-coupled receptor. Proc. Natl. Acad. Sci. U.S.A. 113 (38): E5675–E5684.
59. Nguyen, T., Li, J.-X., Thomas, B.F. et al. (2017). Allosteric modulation: an alternate approach targeting the cannabinoid CB1 receptor. Med. Res. Rev. 37 (3): 441–474.
60. Hertig, S., Latorraca, N.R., and Dror, R.O. (2016). Revealing atomic-level mechanisms of protein allostery with molecular dynamics simulations. PLoS Comput. Biol. 12 (6): e1004746.
61. Bock, A., Schrage, R., and Mohr, K. (2017). Allosteric modulators targeting CNS muscarinic receptors. Neuropharmacology 136: 427–437.
62. Dror, R.O., Green, H.F., Valant, C. et al. (2013). Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs. Nature 503 (7475): 295.
63. Shahane, G., Parsania, C., Sengupta, D., and Joshi, M. (2014). Molecular insights into the dynamics of pharmacogenetically important N-terminal variants of the human β2-adrenergic receptor. PLoS Comput. Biol. 10 (12): e1004006.
64. Rasmussen, S.G.F., DeVree, B.T., Zou, Y. et al. (2011). Crystal structure of the β2 adrenergic receptor–Gs protein complex. Nature 477 (7366): 549–555.
65. Kang, Y., Zhou, X.E., Gao, X. et al. (2015). Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. Nature 523 (7562): 561–567.
66. Kohlhoff, K.J., Shukla, D., Lawrenz, M. et al. (2014). Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. Nat. Chem. 6 (1): 15–21.
67. Dawaliby, R., Trubbia, C., Delporte, C. et al. (2016). Allosteric regulation of G protein-coupled receptor activity by phospholipids. Nat. Chem. Biol. 12 (1): 35–39.
68. Guixà-González, R., Javanainen, M., Gómez-Soler, M. et al. (2016). Membrane omega-3 fatty acids modulate the oligomerisation kinetics of adenosine A2A and dopamine D2 receptors. Sci. Rep. 6 (1): 19839.
69. Paila, Y.D. and Chattopadhyay, A. (2009). The function of G-protein coupled receptors and membrane cholesterol: specific or general interaction? Glycoconj. J. 26 (6): 711–720.

70. Oates, J. and Watts, A. (2011). Uncovering the intimate relationship between lipids, cholesterol and GPCR activation. Curr. Opin. Struct. Biol. 21 (6): 802–807.
71. Gimpl, G. (2016). Interaction of G protein coupled receptors and cholesterol. Chem. Phys. Lipids 199: 61–73.
72. Zocher, M., Zhang, C., Rasmussen, S.G.F. et al. (2012). Cholesterol increases kinetic, energetic, and mechanical stability of the human 2-adrenergic receptor. Proc. Natl. Acad. Sci. U. S. A. 109 (50): E3463–E3472.
73. Khelashvili, G., Mondal, S., Andersen, O.S., and Weinstein, H. (2010). Cholesterol modulates the membrane effects and spatial organization of membrane-penetrating ligands for G-protein coupled receptors. J. Phys. Chem. B 114 (37): 12046–12057.
74. Cherezov, V., Rosenbaum, D.M., Hanson, M.A. et al. (2007). High-resolution crystal structure of an engineered human 2-adrenergic G protein-coupled receptor. Science 318 (5854): 1258–1265.
75. Hanson, M.A., Cherezov, V., Griffith, M.T. et al. (2008). A specific cholesterol binding site is established by the 2.8 Å structure of the human β2-adrenergic receptor. Structure 16 (6): 897–905.
76. Liu, W., Chun, E., Thompson, A.A. et al. (2012). Structural basis for allosteric regulation of GPCRs by sodium ions. Science 337 (6091): 232–236.
77. Wu, H., Wang, C., Gregory, K.J. et al. (2014). Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. Science 344 (6179): 58–64.
78. Song, Y., Kenworthy, A.K., and Sanders, C.R. (2014). Cholesterol as a co-solvent and a ligand for membrane proteins. Protein Sci. 23 (1): 1–22.
79. Guixà-González, R., Albasanz, J.L., Rodriguez-Espigares, I. et al. (2017). Membrane cholesterol access into a G-protein-coupled receptor. Nat. Commun. 8: 14505.
80. Benned-Jensen, T., Norn, C., Laurent, S. et al. (2012). Molecular characterization of oxysterol binding to the Epstein-Barr virus-induced gene 2 (GPR183). J. Biol. Chem. 287 (42): 35470–35483.
81. Manna, M., Niemelä, M., Tynkkynen, J. et al. (2016). Mechanism of allosteric regulation of β2-adrenergic receptor by cholesterol. Elife 5.
82. McCammon, J.A., Gelin, B.R., and Karplus, M. (1977). Dynamics of folded proteins. Nature 267 (5612): 585–590.
83. Dahl, S.G., Edvardsen, O., and Sylte, I. (1991). Molecular dynamics of dopamine at the D2 receptor. Proc. Natl. Acad. Sci. U.S.A. 88 (18): 8111–8115.
84. Butler, D. (1999). Computing 2010: from black holes to biology. Nature 402 (Suppl 6761): C67–C70.
85. Larson, S.M., Snow, C.D., Shirts, M., and Pande, V.S. (2002). Folding@Home and Genome@Home: Using distributed computing to tackle previously intractable problems in computational biology. arXiv:0901.0866 [physics.bio-ph]
86. Snow, C.D., Nguyen, H., Pande, V.S., and Gruebele, M. (2002). Absolute comparison of simulated and experimental protein-folding dynamics. Nature 420 (6911): 102–106.

87. Shaw, D.E., Bowers, K.J., Chow, E. et al. (2009). Millisecond-scale molecular dynamics simulations on Anton. Proc. Conf. High Perform. Comput. Networking, Storage Anal. (SC '09), 1.
88. Borhani, D.W. and Shaw, D.E. (2012). The future of molecular dynamics simulations in drug discovery. J. Comput.-Aided. Mol. Des. 26 (1): 15–26.
89. Kruse, A.C., Hu, J., Pan, A.C. et al. (2012). Structure and dynamics of the M3 muscarinic acetylcholine receptor. Nature 482 (7386): 552–556.
90. Stone, J.E., Phillips, J.C., Freddolino, P.L. et al. (2007). Accelerating molecular modeling applications with graphics processors. J. Comput. Chem. 28 (16):2618–2640.
91. Anderson, J.A., Lorenz, C.D., and Travesset, A. (2008). General purpose molecular dynamics simulations fully implemented on graphics processing units. J. Comput. Phys. 227 (10): 5342–5359.
92. Martínez-Rosell, G., Giorgino, T., Harvey, M.J., and de Fabritiis, G. (2017). Drug discovery and molecular dynamics: methods, applications and perspective beyond the second timescale. Curr. Top. Med. Chem. 17 (23): 2617–2625.
93. Buch, I., Harvey, M.J., Giorgino, T. et al. (2010). High-throughput all-atom molecular dynamics simulations using distributed computing. J. Chem. Inf. Model. 50 (3): 397–403.
94. Selent, J., Sanz, F., Pastor, M., and De Fabritiis, G. (2010). Induced effects of sodium ions on dopaminergic G-protein coupled receptors. PLoS Comput. Biol. 6 (8): e1000884.
95. Hellerstein, J.L., Kohlhoff, K.J., and Konerding, D.E. (2012). Science in the cloud: accelerating discovery in the 21st century. IEEE Internet Comput. 16 (4): 64–68.
96. Plattner, N., Doerr, S., De Fabritiis, G., and Noé, F. (2017). Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. Nat. Chem. 9 (10): 1005–1011.
97. Sultan, M.M., Denny, R.A., Unwalla, R. et al. (2017). Millisecond dynamics of BTK reveal kinome-wide conformational plasticity within the apo kinase domain. Sci. Rep. 7 (1): 15604.
98. Moore, G.E. (1965). Cramming more components onto integrated circuits. Electronics 38 (8): 114–117.
99. Vendruscolo, M. and Dobson, C.M. (2011). Protein dynamics: Moore's law in molecular biology. Curr. Biol. 21 (2): R68–R70.
100. Rodríguez-Espigares, I., Ramírez-Anguita, J.M., Torrens-Fontanals, M. et al. (2017). GPCRmd web page. http://www.gpcrmd.org/.
101. De Vivo, M., Masetti, M., Bottegoni, G., and Cavalli, A. (2016). Role of molecular dynamics and related methods in drug discovery. J. Med. Chem. 59 (9): 4035–4061.

3.2. How do molecular dynamics data complement static structural data of GPCRs

Due to the dynamic nature of GPCRs, a static snapshot cannot fully explain the complexity of their signal transduction. In this review article, we inspect how MD simulations can incorporate the missing information on protein flexibility into experimentally solved structures. For that, we examine different molecular processes underlying GPCR physiology, including GPCR activation and signaling, orthosteric and allosteric ligand binding, the impact of water molecules and ions, and natural genetic variants, among others. For each of them, we explain how MD simulations have contributed to their understanding. We also discuss how MD can be applied to support different stages of drug discovery. To provide a more technical view of this technique, we describe the workflow of classical MD simulations, and how the resulting data can be analyzed. Finally, we include a detailed discussion of the challenges that still need to be overcome to reach the full potential MD simulations.

Torrens-Fontanals, M. et al. How do molecular dynamics data complement static structural data of GPCRs. *International Journal of Molecular Sciences* 21, 5933 (2020). doi:10.3390/ijms21165933

How Do Molecular Dynamics Data Complement Static Structural Data of GPCRs

Mariona Torrens-Fontanals[1], Tomasz M. Stepniewski[1,2,3], David Aranda-García[1], Adrián Morales-Pastor[1], Brian Medel-Lacruz[1], and Jana Selent[1,*]

1. Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM)—Department of Experimental and Health Sciences, Pompeu Fabra University (UPF), 08003 Barcelona, Spain
2. InterAx Biotech AG, PARK innovAARE, 5234 Villigen, Switzerland
3. Faculty of Chemistry, Biological and Chemical Research Centre, University of Warsaw, 02-093 Warsaw, Poland

* Author to whom correspondence should be addressed.

Abstract

G protein-coupled receptors (GPCRs) are implicated in nearly every physiological process in the human body and therefore represent an important drug targeting class. Advances in X-ray crystallography and cryo-electron microscopy (cryo-EM) have provided multiple static structures of GPCRs in complex with various signaling partners. However, GPCR functionality is largely determined by their flexibility and ability to transition between distinct structural conformations. Due to this dynamic nature, a static snapshot does not fully explain the complexity of GPCR signal transduction. Molecular dynamics (MD) simulations offer the opportunity to simulate the structural motions of biological processes at atomic resolution. Thus, this technique can incorporate the missing information on protein flexibility into experimentally solved structures. Here, we review the contribution of MD simulations to complement static structural data and to improve our understanding of GPCR physiology and pharmacology, as well as the challenges that still need to be overcome to reach the full potential of this technique.

Keywords: GPCRs; molecular dynamics; ligand binding; receptor (in)activation; receptor signaling; drug discovery

1. Introduction

G protein-coupled receptors (GPCRs) are a large and versatile family of transmembrane proteins, encompassing over 800 identified members. These proteins act as receptors for a wide variety of extracellular stimuli including light, changes of pressure, and chemical ligands, odorants, neurotransmitters, chemokines, and metabolites among others, transducing their information into intracellular signaling cascades. Due to their participation in a wide range of pathways and physiological processes, as well as their druggability, GPCRs have become a drug target of major importance in the pharmaceutical industry [1].

As a consequence of their relevance for drug discovery, deciphering the molecular basis of GPCR signaling has become a major research focus. The signaling outcome of GPCRs is determined by their three-dimensional conformation, which is variable and depends on multiple factors, such as the binding of orthosteric and allosteric ligands, the lipidic environment, and post-translational modifications. Understanding how all of these factors contribute to a specific structure, and in turn, a specific signaling response, would not only expand our knowledge of GPCR biology but also provide structural blueprints for the design of novel and better therapeutics. To address this ambitious goal, numerous endeavors have been undertaken to characterize the three-dimensional structure of GPCRs and its changes over time.

Important advances in protein engineering, X-ray crystallography, and cryo-electron microscopy (cryo-EM) during the past decade have led to an exponential growth in the number of known GPCR structures. Since then, the number of available structures has continued increasing (Figure 1a). This large data set has been crucial for advancing our understanding of GPCR function. Moreover, it enabled the application of structure-based drug design approaches, which aid the discovery of novel drug candidates with improved pharmacological profiles [1,2,3].

Despite their enormous utility, high-resolution structures describe proteins mainly as rigid entities, whereas information about their intrinsic flexibility and conformational plasticity cannot be

appreciated. With the goal to incorporate atomic-level dynamic information to static systems, molecular dynamics (MD) simulations were introduced several decades ago. The first MD simulation of a biomolecule was 9.2 ps-long and consisted of the bovine pancreatic trypsin inhibitor (~500 atoms) in vacuum [6]. In the case of GPCRs, the first MD simulation was obtained in 1991, before the first GPCR crystal structure was resolved [7]. It corresponded to an 80 ps-long trajectory of a rat dopamine D_2 receptor, modeled from its sequence with molecular mechanics. Ever since, MD simulations have greatly improved their performance, allowing the simulation of larger systems for longer timescales. A major determinant of these advances has been the development of algorithms optimized for graphical processor units (GPUs), a technology first designed to improve video game performance [8,9]. GPU exploitation was a major breakthrough for the field, enabling researchers to perform on commodity hardware calculations that were previously only possible with the use of supercomputing clusters. Along with these technological advances, the expansion of free and user-friendly software for the input preparation (e.g., CHARMM-GUI [10], HomolWat [11]) and analysis (e.g., MDAnalysis [12,13]) of MD simulations has greatly contributed to the broad application of this technique.

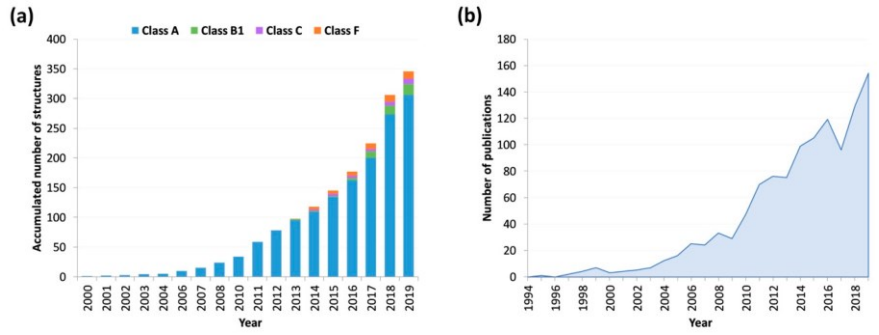


Figure 1. (a) Number of G protein-coupled receptors (GPCRs) structures available in GPCRdb [4,5] over time. **(b)** Number of publications per year indexed at Thomson Reuters' Web of Science that contain the topics "molecular dynamics" and ("GPCR" or "GPCRs"). The exponential growth of successful GPCR research based on molecular dynamics (MD) simulations is evidenced by the rapid upsurge in the number of publications per year related to this subject.

Owing to the aforementioned technical developments, MD simulations currently provide a combination of temporal and

structural resolution greater than what is usually achievable by experimental methods [14]. As a result, MD simulations are widely used for the study of GPCRs, as reflected by the continuous increase in publications per year on this topic (Figure 1b). Moreover, most publications on crystallography now supplement their studies with MD to refine the obtained structure. Here, we review recent developments in the study of GPCR functionality using MD simulations to complement static structural data. We discuss the role of receptor dynamics in several functional processes, outline the applicability of MD simulations for drug discovery, and describe the basis of this technique. We also examine the main challenges that still need to be overcome to reach its full potential. Finally, we discuss the future of the field.

2. Complementing Static Data

Three-dimensional structures derived from experiments via X-ray crystallography or cryo-EM provide high-resolution information about specific conformational states of GPCRs. However, we need to be aware that these structures represent low energetic conformational states that are obtained under experimental conditions that often deviate from native-like conditions. In this scenario, MD simulations are a useful tool to drive these structures to conformational states that are linked to a more native-like environment. Moreover, MD simulations incorporate the missing information on structural motions, yielding insights that can be critical to the understanding of GPCR physiology and pharmacology [14]. In this respect, MD simulations have proven useful to complement static data and expand our knowledge of processes such as binding of small molecules or drugs to orthosteric or allosteric receptor sites. We can also determine how a biomolecular system will respond to perturbations such as mutations, post-translational modifications, and the composition of the cell membrane [15,16]. In addition, we can study the conformational rearrangements that occur during receptor (in)activation, determine metastable receptor states along the transition pathways or explore the interaction with intracellular coupling partners [17,18]. Even processes such as receptor dimerization/oligomerization, which has been implicated in fine-

tuning GPCR signaling, can be investigated using different MD techniques [19].

2.1. Molecular Mechanism of Receptor Activation

From a structural perspective, there are two mechanisms by which a molecule, so-called "agonist", can mediate GPCR activation. On the one hand, an agonist can sample and stabilize a subset of receptor conformations known as "active states", shifting the conformational equilibrium to an active receptor (conformational selection mechanism) [20]. On the other hand, the binding of the agonist can initiate small structural changes in the ligand binding site, which are propagated across the receptor through rearrangements of specific residues. These rearrangements lead to global structural changes towards conformational populations of active receptor states (induced fit mechanism) [21]. Most likely, both mechanisms contribute to a different extent to receptor activation depending on the ligand and receptor type [14,22]. Finally, receptors in an active state have a higher propensity to interact with intracellular partners. This leads to the initiation of signaling cascades which ultimately alter the metabolism of the cell [23].

Experimentally solved structures provide extensive information on the conformation of several active, inactive, and intermediate states [24,25]. Such structures have been an excellent starting point for numerous MD simulation-based studies that clarify the activation/inactivation mechanism. By this means, researchers have been able to probe the flexibility of GPCR-ligand complexes in the initial and final stages of activation and observe structural fundamentals on how ligands stabilize conformational states that are related to specific signaling outcomes [26,27]. Furthermore, extending such simulations it is possible to capture intermediate conformations that are adopted on the transition pathway.

Beyond this, pioneering simulations on the active conformation of the β_2-adrenergic receptor (β_2AR) [28] revealed that the presence of an intracellular coupling partner is crucial to stabilize the receptor in an active state. Without it, the receptor can revert to a fully inactive state, despite the presence of an agonist. The study also highlighted multiple structural features related to activation, which are loosely coupled and do not necessarily occur sequentially. These results

were later supported by NMR data [29]. After these findings, a simulation of unprecedented total length, obtained thanks to Google's Exacycle cloud computing platform, allowed the generation of a complete structural statistical model of GPCR activation [30]. One of the highlights of this study was that GPCRs can follow multiple pathways towards obtaining an active conformation.

On a more detailed level, MD simulation also permits the study of more subtle structural rearrangements related to activation. A notable example includes a comprehensive study carried out by Li et al. [31]. By simulating complexes of the A_{2A} receptor ($A_{2A}R$) with multiple ligands, they were able to obtain a comprehensive view of the activation mechanism of this receptor. Importantly, they observed that the conserved residue W6.48 attained different conformational states in response to agonists. Moreover, by studying receptor-ligand contacts, they were able to identify groups of contacts that lead to a specific signaling response. These interactions promoted local structural changes that led to the increased mobility of the transmembrane helix (TM) 6. Importantly, these results were in line with crystallographic [24] and NMR data [32].

Furthermore, post-translational modifications have been described to be critical for the biological activity of GPCRs. In this respect, Oddi et al. report for instance that the biological activity in terms of the CB_1 receptor is closely linked to palmitoylation of cysteine 415 in helix 8 [33]. MD simulation revealed that this modification stabilizes helix 8 and promotes the binding of cholesterol molecules in the vicinity, which likely facilitates the interaction with lipid rafts and caveolin 1. This goes along with the experimental finding that the C415A mutation impairs the receptor's ability to functionally interact with lipid rafts as well as eliminates agonist-dependent internalization of the CB_1 receptor. In addition, the same group shows that palmitoylation of cysteine 415 fine-tunes CB_1 receptor interaction with the Gαi2 protein, which further highlights the relevance of post-translational modifications for receptor functionality [34].

As integral membrane proteins, GPCRs communicate with the lipid environment, which contributes to the regulation of GPCR function

and dynamics. Membrane phospholipids have been found to allosterically modulate the activity [35,36,37,38] and oligomerization [19] of GPCRs, while membrane cholesterol can regulate its stability, ligand-binding properties and function [16,39,40,41]. Still, the precise nature of lipid implication in GPCR modulation is unclear. Such effects can either be attributed to changes in membrane biophysical properties (including thickness, curvature, and surface tension) [42,43], direct interactions [15,44,45,46], or both. In one of the first MD studies comparing the effects of different single species lipid bilayers on the dynamical behavior of a GPCR, Ng. et al. showed that the structural motions of the $A_{2A}R$ may depend on its phospholipid environment [47]. This could be explained by the physical adaptation of the $A_{2A}R$ to different membrane thicknesses or by molecular interactions of the lipid headgroups and the protein. Similarly, in a recent study Bruzzese et al. examined how much different membranes affect the activation process of the $A_{2A}R$ and the functional effect of their agonists [48]. Based on microsecond-long MD simulations, they revealed an effect of the phospholipid membrane in the intermediate or active receptor conformations observed, which can be attributed to phospholipid-mediated allosteric effects on the intracellular side of the receptor. In addition to identifying potential lipid interaction sites, MD simulations can provide estimates of the free energy of protein-lipid interactions, which permits to quantify their strength. To test the reliability of MD to study the energetics of protein−lipid interactions, Corey et al. compared different MD-based approaches in terms of ease of accuracy and computational cost [49]. They showed that such methods produce estimates of the strength and specificity of lipid-binding sites that are robust and reproducible.

Finally, a relatively recent finding is that GPCRs can couple to diverse intracellular signaling partners, including different G proteins and β-arrestins. An interesting observation is that, in some cases, only a subset of pathways is engaged upon ligand binding, a phenomenon known as “signaling bias” [50,51]. The underlying molecular mechanism of signaling bias is still poorly understood and will be addressed in more detail in a later section (Section 2.2.2).

2.2. Ligand Binding to GPCRs

Typically, GPCRs are able to recognize and bind a variety of ligands that modulate the receptor functional outcome. Deciphering the complex process of receptor modulation, and how specific interactions in the ligand binding site are linked to the final functional outcome, has been a main goal of many scientific endeavors. Such information would help us better understand GPCR physiology and inform the design of molecules with a specific signaling profile [52]. A valuable resource of ligand binding dynamics is found in the recently established GPCRmd server [53], which provides intuitive visualization and analysis tools currently covering 70% of crystallized receptor subtypes (Figure 2).

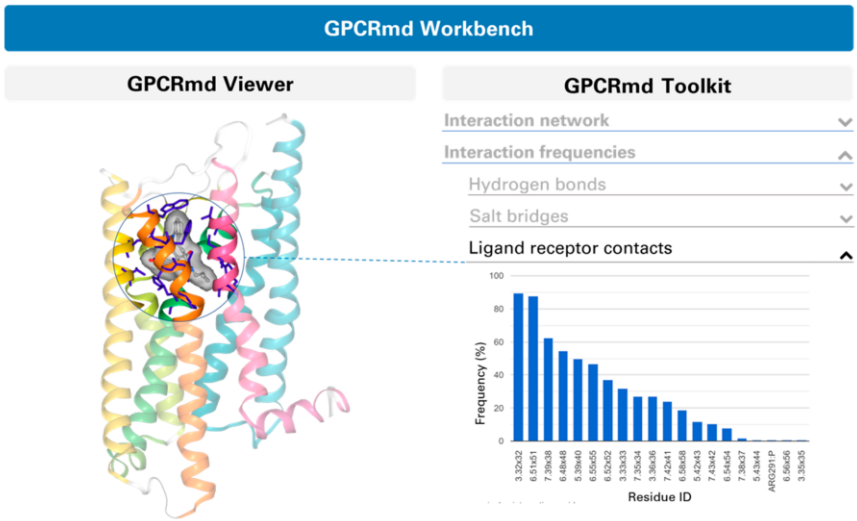


Figure 2. Schematic view of the ligand-protein interaction results that can be obtained with the GPCRmd server [53]. Specifically, the GPCRmd Workbench module of the server enables interactive visualization (GPCRmd Viewer) and analysis (GPCRmd Toolkit) for individual simulations, including ligand-protein interactions among others. Figure obtained from the GPCRmd server [53].

2.2.1. Classical Orthosteric Ligands

The use of static structures to understand ligand binding can lead to incomplete information, especially in receptors with high flexibility. This was highlighted by Ferruz et al. in a study where they compared the binding poses of several dopamine D_3 receptor

antagonists obtained with static docking and with MD simulations [54]. Using large-scale MD simulations and Markov state models (MSMs), they were able to overcome the limitations of docking in the determination of the ligand binding poses and revealed a cryptic binding pocket. Virtual screening protocols considering only static structures would miss compounds binding to this cryptic binding pocket. Thus, the characterization of the intrinsic flexibility of GPCRs is of great value for the identification or design of new ligands [55], as discussed also in Section 3.

Similarly, MD studies provide valuable information on the strength of ligand-receptor interactions in terms of contact frequencies that cannot be obtained by methods that do not account for the flexibility of the binding site. This information facilitates the identification of the key interactions that a ligand establishes in the binding pocket and which likely drives the signaling outcome. For example, a combination of molecular modeling and simulation was used to describe the binding characteristics of the natural agonist and its derivatives in the oxoeicosanoid receptor 1, providing new insights into how this receptor is modulated [56]. Moreover, MD simulation provided information on ligand stability and key interactions that allowed identifying selectivity features of 5-HT_{2B} fluorescent ligands that retain the agonistic functional behavior of the model ligand [57].

The interaction between a ligand and a GPCR, however, is not only determined by the events that happen once in the binding site. The ligand needs to pass through a series of intermediate states between the solution phase and the fully bound pose, known as the ligand binding pathway. Describing this pathway can lead to the identification of energetic barriers that affect the binding and unbinding rates. Ultimately, such rates play a pivotal role in drug efficacy, selectivity, and safety [58,59,60]. The details of the binding pathways are difficult to probe by experimental techniques, but MD simulations generate useful insights on this process [61]. Some highlights in the MD-based characterization of binding pathways include (S)-alprenolol binding to the β_2AR [62], histamine to the histamine H_4 receptor [63], adenosine to the A_{2A}R [64], and clozapine and haloperidol to the dopamine D_2 and D_3 receptors [65]. Importantly, the case of the β_2AR was the first unbiased MD simulation study capturing the full process of ligands spontaneously

binding to a GPCR. Dror et al. were able to achieve final poses matching those determined crystallographically without the incorporation of any prior knowledge of the binding site. Results revealed not only the predominant pathway into the binding site, but also the two main energetic barriers that govern drug binding and unbinding kinetics.

2.2.2. Biased Agonists

Biased agonists are molecules of high interest, as they selectively target a specific signaling pathway in a cell while maintaining other signals in their physiological state. Biased signaling probes are valuable tools to interrogate the involvement of the pathway in physiological processes or in the development of disease symptoms. Furthermore, they are promising starting points for the development of safer drugs, as they potentially allow selective modulation of pathways associated with disease symptoms while not engaging counter-therapeutic pathways or those related to debilitating side-effects.

Several studies demonstrated the usefulness of MD simulations to uncover distinct molecular events that are linked to a biased response [66]. Thus, Martí-Solano et al. characterized the dynamic receptor interaction fingerprint of biased agonists with a specific signaling response [67]. Based on this information, this study succeeded in predicting additional ligands with a tailored signaling profile. Such a strategy has been also successfully applied to ligands targeting the dopamine D_2 [68], M_2 [69], and AT_1 [70] receptors.

Moreover, MD studies can also capture downstream events related to signaling bias. In this respect, novel mechanistic insights revealed the connection between ligand binding, conserved micro-switches, and arrestin bias in serotonin receptors [71]. In particular, simulations showed that interactions of the ligand with the binding pocket determine the rotational freedom of TM6 which, in turn, impacts the conformation of the highly conserved P-I-F motif. Consequently, a hydrophobic connector region between the P-I-F motif and the ionic lock seems to contribute to the formation of a water channel that determines the degree of receptor opening (disrupted ionic lock). This conditions G protein coupling and, thus, whether signaling is biased towards arrestin or not. This work

highlights the capacity of MD to shed light on features that cannot be extracted from static structures. Another relevant example was an extensive study developed by Kapoor et al. aiming to explain the basis of functional selectivity in the μ-opioid receptor [72]. Among other findings, the study identified distinct conformational rearrangements in the receptor bound to a balanced or a G protein-biased agonist. They also highlighted differences in the allosteric communication, with a more pronounced transfer of information triggered by the G protein-biased agonist. Finally, Nivedha et al. developed a computational method to predict ligand bias in GPCRs ahead of experiments [73]. For that, they used MD simulation to calculate the mechanism of allosteric communication from the extracellular region to the intracellular transducer coupling region. Additionally, they were able to identify functional hotspot residues that potentiate the ligand-mediated bias, which can greatly aid in the design of biased ligands for GPCRs.

2.2.3. Allosteric Ligand Binding

When studying ligand binding, traditional efforts have focused on targeting the orthosteric binding site of GPCRs. The orthosteric binding site of many GPCR subtypes is highly conserved. As a consequence, orthosteric ligands often target several receptors simultaneously, leading to off-target side effects. This leads to one important challenge of GPCR drug discovery, which is achieving selectivity, the ability of ligands to specifically target one receptor subtype over another.

Contrarily to orthosteric ligands, allosteric ligands bind to sites topographically distinct from the orthosteric binding site. Such allosteric binding sites are much more variable in terms of the sequence, which gives allosteric ligands the potential to achieve greater selectivity at GPCR subtypes [23,74]. Allosteric ligands modulate the effect of the orthosteric ligand on the target, which provides a strategy to fine-tune cellular responses triggered by the orthosteric ligand. These characteristics of allosteric ligands have invited a growing interest in designing drugs that target allosteric pockets of GPCRs [75,76].

However, targeting allosteric sites comes with some challenges. Allosteric binding sites are not evident from crystal structures.

Moreover, the molecular mechanisms by which these modulators affect GPCR signaling depend on dynamical properties that are not evident from static structures. This makes computational methods such as MD simulations a valuable approach to detect hidden allosteric binding sites and determine the mechanistic basis of allosteric regulation [77]. MD-based studies have been especially helpful for the identification of allosteric mechanisms in muscarinic receptors, which are usually paradigmatic for all GPCRs [78,79,80]. One case is the work from Dror et al. in which they provided a structural basis of allosteric ligand binding and described mechanisms of cooperativity between the allosteric and the orthosteric ligand [80]. In another study, Chan et al. applied long-timescale MD simulations to show that acetylcholine, the endogenous ligand, can go from the orthosteric binding site into a deeper allosteric binding site [81].

2.3. Revealing the Dynamic Behavior of Water Molecules and Ions

Comparative analysis of available crystal structures pointed to the relevance of waters for receptor dynamics and function [82]. The unique ability of MD to monitor diffusion and binding events of all water molecules in a system enabled the further elaboration of this idea. Simulations of the opioid receptors revealed that GPCR activation correlates with the entrance of waters from the extracellular side [83,84]. In line with this finding, further studies demonstrated that activation of the $A_{2A}R$ is linked with the formation of continuous water channels [85,86]. Detailed investigation of the simulation frames revealed that the formation of this channel is mediated by rearrangements of conserved residues W6.48 and Y7.53, the latter of which forms the NPXXY motif.

Importantly, water molecules also have a strong impact on ligand binding and unbinding events, which can be investigated in detail with MD simulations. It is well established that water has a role in ligand-receptor dissociation. For example, Schmidtke et al. showed that shielding ligand-receptor hydrogen bonds from water can contribute to long ligand residence time [87]. Interestingly, Magarkar et al. recently found, based on MD simulations, that shielding of water from intra-protein interactions, not directly involved in ligand−receptor interactions, is also a relevant factor in

ligand binding kinetics, as such interactions confer the rigidity of the binding site [88]. This opened new opportunities for the optimization of the residence time during drug development pipelines.

Similarly, MD simulations helped to shed light on the role of ions for GPCR function. Sodium ions are known to be important allosteric modulators of GPCRs, but the mechanism of this modulation is still not well understood [89]. Using MD simulations, Selent et al. provided structural details on the binding of sodium ions in the D_2 receptor and proposed the molecular mechanism of the allosteric sodium-induced modulation [90]. Several studies have been dedicated to revealing atomistic insights into allosteric sodium ion binding to other class A receptors [91,92]. For example, Selvam et al. elucidated the sodium binding mechanism of 18 GPCRs based on hundreds-of-microsecond long simulations [93]. Their analysis of the kinetics of sodium binding to the allosteric site revealed key residues that act as major barriers for sodium diffusion. Also, they reported that sodium ions can bind to GPCRs from the intracellular side when the allosteric site is inaccessible from the extracellular side. Furthermore, Vickery et al., based on MD simulations and free energy calculations, suggested that the opening of the conserved hydrated channel in the active M_2 muscarinic receptor allows the exchange of a sodium ion from its extracellular binding pocket to the cytoplasm. This exchange of sodium could be a key step in class A GPCR activation [94]. Beyond allosteric ion effects, a recent study has also proposed that sodium ions can stabilize ligand binding in the orthosteric site and by this enhance receptor signaling in the D_2 receptor [95].

2.4. Impact of Natural Genetic Variants

Another factor that impacts GPCR functionality is genetic variants. A huge number of natural genetic variants are observed in GPCRs, as listed in dbSNP [96] and GPCRdb [4]. To name a few, missense variants in rhodopsin are responsible for retinitis pigmentosa due to an alteration in receptor folding and cellular trafficking [97], and missense variants in the C-C chemokine receptor 6 exhibit loss-of-function effect by decreasing G-protein signaling [98].

Understanding the impact of natural variants on GPCRs is critical, as variants can be responsible for disease susceptibility, as well as distinct responses to treatments [99]. Such functional differences can be caused by alterations in dynamic processes of ligand binding pathways, ligand binding interactions, constitutive receptor activity, or recognition of intracellular effector proteins (e.g., G protein binding). Hence, MD simulations are a promising approach to elucidate the molecular mechanisms that explain functional differences between wild type and variant GPCRs, providing genotypic-phenotypic correlations [100,101,102,103]. MD simulations were used, for example, to determine the molecular basis of the effect of a commonly found variant: the Arg16Gly variant of the β_2AR. This variant has been linked to a differential response to albuterol, a β_2AR agonist frequently used in the treatment of asthma. Results revealed that the Arg variant increased the dynamics of the N-terminal region, where this polymorphism is located. This change in dynamics leads to long-range effects at the ligand binding site, altering ligand binding-site accessibility, which is higher in the Gly variant [102]. Similar results were recently found for the Gln27Glu variant of the same receptor, which perturbs the network of electrostatic interactions that connects the N-terminal region with the binding site, altering drug response [103].

2.5. Complementing Experimental Maps

A critical step in X-ray crystallography or cryo-EM of GPCRs is fitting the receptor model to the experimental density map. After the fitting procedure, certain density areas often remain unmatched, a piece of information that can be extracted from the so-called difference maps (fo-fc). The discrepancy between model and experimental density map may arise from the existence of different rotameric states or the binding of water molecules and ions. Thanks to MD it is possible to complement static structures with this information by monitoring the dynamics of sidechain rotations and the diffusion/binding of solvent molecules that can justify unmatched density areas. Moreover, MD allows investigating highly flexible regions that explain low-resolution areas in density maps.

3. Application of MD in Drug Discovery

The drug discovery process implies an immense cost, high risk, and a long time to move from the bench to the market [104]. Computer-aided drug design has the potential to de-risk and accelerate this process [74], and thus it has become an attractive approach for drug discovery targeting GPCRs [105]. Static structures have proven highly effective at aiding drug design [106]. However, due to the high flexibility of GPCRs, especially in druggable regions such as allosteric sites, the full potential of structure-based drug design requires a deeper understanding of GPCR dynamics [80,107].

One of the most widely used structure-based drug design strategies is virtual screening, where libraries of small molecules are screened to identify those structures which most likely bind to the target. Virtual screening is traditionally based on docking the ligands to a static structure of the target protein. This approach has been very successful for the discovery of new ligands. Yet, docking does not consider the flexibility of the binding pocket, thus leading to the identification of only a subset of binders, namely those similar to the crystallized ligand [14]. Using MD to account for the dynamic behavior of the binding pocket generally increases the diversity of ligands identified [55,108]. Moreover, it allows exploring rare conformations that can help define drugs with higher specificity for the receptor [109]. Overall, MD-based methods are more resource-consuming compared with traditional docking, but a higher accuracy can be reached. For example, an interesting approach to include dynamic information in virtual screening protocols is the characterization of the binding site using MD to construct ensembles with structural diversity, where the ligand candidates are docked [110,111,112].

MD can also provide valuable information to guide lead optimization, where the ligand is modified to improve properties, such as potency, selectivity, or pharmacokinetic parameters. Dynamic information can be used to identify the key interactions that the lead ligand establishes with the binding pocket, as well as rearrangements of the binding pocket induced by the ligand [113]. Simulations can further help test and refine potential ligand poses, or even reveal unknown binding sites [77,114]. They are also

valuable to improve selectivity, as they can be used to identify differences in the dynamics of binding pockets of closely related receptor subtypes [113].

Moreover, simulation-based methods were found to provide substantially more accurate estimates of ligand binding affinities (free energies) compared to other computational approaches [115]. For now, it is not possible to sample enough unbinding events to determine rates or affinities by unbiased MD. However, it is possible to combine MD with specialized free-energy techniques to enhance sampling for this purpose. This is the case of the free energy perturbation method, which can be used to evaluate and compare the relative affinity of several compounds, such as derivatives of a particular ligand, on a target receptor. This was shown to be particularly useful, for example, for fragment optimization [116]. Similarly, this technique can be used to characterize and compare the effect of single-point mutations of residues in the binding pocket on the binding affinity of a ligand, which helps to determine its binding mode [117]. Another extended approach is metadynamics simulations. Provasi et al. pioneered the use of metadynamics [13] to study ligand binding to GPCRs [118] and have successfully applied this enhanced MD algorithm to predict the binding pose of several orthosteric and allosteric ligands in opioid receptors [119,120]. Still, automatizing metadynamics protocols in drug discovery workflows is challenging, since they usually require specific testing and optimization, mainly to select adequate collective variables [121]. However, efforts are being made to generate accurate and inexpensive metadynamics protocols that can be applied to a broad range of different GPCRs and ligands. This is the case of Saleh et al., who proposed a generally applicable metadynamics protocol that uses a single, optimal CV to accurately and efficiently explore the entire ligand binding path and predict binding mechanisms and affinities [122].

Another important application of simulation techniques for lead optimization is the optimization of drug binding and unbinding kinetics, which plays a critical role in drug efficacy, selectivity, and safety [58,59,60,61,62,63,65,123]. In fact, the ligand unbinding kinetics (the inverse of its residence time on the protein) is sometimes better correlated with drug efficiency than binding affinities [124]. Successful examples like tiotropium demonstrate

the potential of kinetic optimization. Tiotropium is a well-known M_3 muscarinic receptor antagonist used as treatment for chronic obstructive pulmonary disease. Its very slow dissociation rate from the M_3 receptor is postulated to be the key to its superior pharmacological profile. Interestingly, while tiotropium has a similar affinity for the M_2 and M_3 receptors, it shows kinetic subtype selectivity towards the M_3 [125]. Based on MD simulations, this selectivity was found to be caused by differences in the electrostatics and flexibility of the extracellular surface [126].

To further decipher the molecular basis of binding and unbinding kinetics, MD simulations can be used to obtain the whole binding pathway of the ligand, identify metastable binding sites and detect the energetic barriers that govern drug binding and unbinding kinetics [127]. Ligand dissociation time scales are often much longer than those accessible by unbiased MD, even when specialized hardware is used. Thus, enhanced sampling algorithms such as metadynamics are commonly employed. To increase the applicability of these techniques in drug discovery, several variations of conventional metadynamics protocols are created, for example by combining metadynamics with adiabatic-bias MD [128,129].

When designing a GPCR-targeted drug, one aims to achieve a particular signaling profile. In other words, the drug needs to be able to stabilize certain conformational states of the receptor. This is a complex process that requires an understanding of how subtle changes in the binding pocket lead to different conformations of the intracellular coupling interface and, in turn, different signaling profiles. Achieving the desired signaling profile is especially challenging in the case of biased ligands. The successful design of a biased ligand requires knowledge of the conformations associated with G protein signaling and arrestin signaling. As discussed in previous sections, MD simulations are able to provide detailed information on binding pocket dynamics and allow us to compare the receptor-ligand interactions that occur in different conformational states of the receptor and in complex with different types of ligands (e.g., unbiased agonist, biased agonist, inverse agonist, or antagonists) [14]. This opens the road for a more tailored and fine-tuned drug development.

4. Workflow for MD Simulation

The procedure for conducting MD simulations can be divided into four stages (Figure 3a). In the first stage (stage 1), we create the initial coordinates for our simulation system (Figure 3b). This generally involves the curation of experimentally solved receptor structures (e.g., modeling of missing residues/loops, reverting thermo-stabilizing mutations to the wild-type, etc.) or the application of homology modeling. The obtained GPCR model is then embedded into a specific membrane, solvated, and ionized to a physiological concentration. In this initial stage, one should carefully consider factors such as atomic resolution (atomic scale versus coarse-grained), absence or presence of post-translational modifications (palmitoylation, phosphorylation, glycosylation), and the composition of the membrane environment.

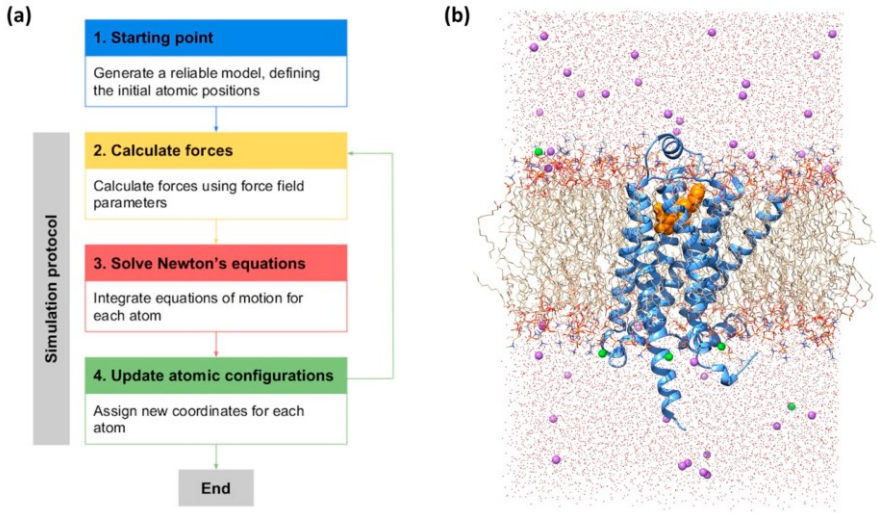


Figure 3. (a) Flowchart summarizing the stages of a MD simulation. **(b)** Example of a GPCR molecular system, including the β_2 adrenergic receptor (β_2AR, blue) with a full agonist in the binding site (orange) in a 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) membrane (tails in light brown, heads colored by heteroatom). The system is solvated with water (red) and ionized with sodium (green) and chloride (purple) ions.

Once the starting structure is obtained, we proceed to simulate the atomic motions of the system. For that, the forces that act on each atom in the system are calculated (stage 2). This is possible thanks

to the so-called force fields, a set of empirical potential energy functions that include all parameters needed to solve both bonded and non-bonded atomic interactions [130,131]. Based on the obtained forces, the atomic positions at the following timestep are predicted by solving the classical (i.e., Newtonian) equations of motion (stage 3). Then, the positions of the atoms are updated accordingly (stage 4). From here, we start an iterative cycle by re-calculating again the forces that act on each atom in the new conformation of the system (stage 2), solving Newton's equations (stage 3), and updating the atomic positions (stage 4). The time length between these iterations, known as the simulation timestep, should be shorter than the fastest process in the system (typically the vibrations of bonds between heavy atoms, as we commonly constrain hydrogen atoms) and usually is around 2 fs.

The timescale of the biological process we are interested in defines the number of iteration steps needed to complete the simulation, which can easily be higher than millions. Knowing when to stop a MD simulation is not trivial, but one has to ensure that the simulation has efficiently sampled the conformational space of the biological process studied.

5. MD Analysis—Extracting Data from the Simulations

5.1. Principles of MD Analysis

Due to the extensive amount of information generated by MD simulations, specific computational tools have become mandatory for their proper analysis. Some of the most popular tools include python modules such as MDAnalysis [12,13] and MDtraj [132], which allow the automatization of analysis pipelines using scripts. The visualization and modeling software Virtual Molecular Dynamics (VMD) [133] also provides a range of analysis tools that can be expanded even further by using plugins. In addition, simulation software like GROMACS [134] and CHARMM [135] include their own build-in sets of analysis tools. Even more, there exist online repositories such as Plumed-nest [136], specifically developed to store scripts used for generating and analyzing MD simulations. Despite the diversity in available tools, certain parameters are frequently analyzed, as they provide relevant information about the simulation.

One of the most important parameters is the root mean square deviation (RMSD), which allows a quantitative evaluation of the structural changes that occur during a simulation. It is based on the distances between the atoms of the protein at a certain frame and the same atoms at a superimposed reference frame (Figure 4d). The RMSD is obtained with the following equation:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|x_i(t_j) - x_i(t_0)\right\|^2}$$

where $x_i(t_j)$ represents the coordinates of atom *i* at frame *j*, $x_i(t_0)$ represents the position of the same atom *i* at the reference frame, and *n* the number of atoms in the system.

RMSD profiles (i.e., RMSD over time) are routinely used to assess the stability of the simulated protein and detect transitions between different conformations (Figure 4a). It is also useful to compare the dynamic behavior of the receptor under different conditions, as done by Ozcan et al. to determine the effect of the intracellular loop 3 in human β_2AR [137].

Another widely used parameter is the root mean square fluctuation (RMSF), which describes the relative mobility of an atom or residue in the simulation. The RSMF is based on the mean square of the residue or atom position in each frame, which can be obtained using the following equation:

$$RMSF = \sqrt{\frac{1}{T}\sum_{j=1}^{T}\left(x_i(t_j) - \bar{x_i}\right)^2}$$

where $x_i(t_j)$ represents the coordinates of atom *i* at frame *j*, x_i the average position of atom *i* in the simulation and *T* the total number of frames in the simulation.

RMSF profiles (i.e., RMSF as a function of atoms/residues) are often employed to describe and compare the relative mobility of

specific regions of the receptor (Figure 4b). For example, Semack et al. were able to detect specific flexibility profiles for the β_2AR and the vasopressin receptor 1A when bound to different sets of peptides derived from the C-terminus of the G alpha subunit [138].

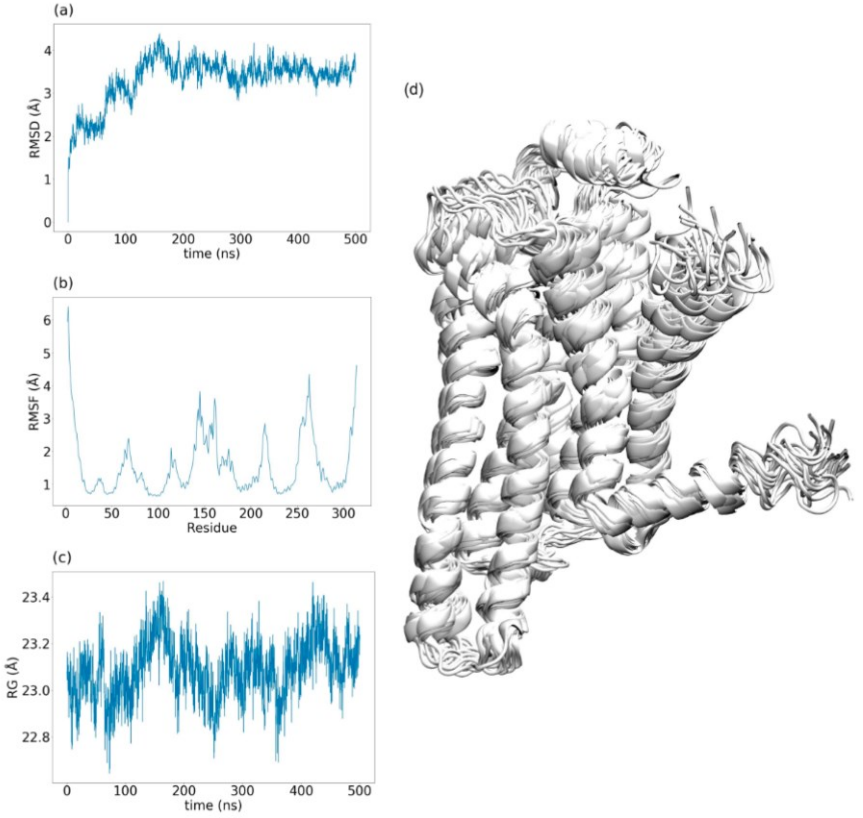


Figure 4. Example of different parameters analyzed in a 500 ns-long MD simulation of the A_{2A} receptor ($A_{2A}R$). **(a)** Root mean square deviation (RMSD) profile taking as reference the first frame of the simulation, which is superimposed to the rest of the frames. RMSD values (i.e., structural differences with respect to the reference frame) increase over the simulation time until the system reaches a stable conformation after 100 ns. **(b)** Root mean square fluctuation (RMSF) profile displaying the values of all the alpha carbons in the protein. Higher RMSF values correspond to flexible loops, while lower ones belong to transmembrane helices, where residues are stabilized by the secondary structure. **(c)** Radius of gyration (RG) profile where the RG fluctuates around the same value during the simulation, indicating that the system does not suffer any big change in compactness. **(d)** Superimposition of 25 representative frames of the simulated receptor. The relative mobility of loop regions contrasts with the rigidness of the transmembrane helices.

Furthermore, the radius of gyration (RG) is a valuable parameter to describe the overall compactness of the protein. Specifically, the RG is defined as the mean square of the distance between each protein atom and the center of mass of the protein:

$$RG = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - r_{cm})^2}$$

where r_i r_{cm} represents the distance between atom *i* and the center of mass of the molecule and *n* the total number of atoms in the system.

RG profiles (i.e., RG over time) can be used to assess the evolution of the protein compactness during a simulation (Figure 4c), as done by Davoudmanesh and Mosaabadi to study the effects of homocysteinylation of the neuropeptide substance P on its binding with the NK1 receptor [139].

In order to obtain a detailed view of the molecular mechanisms that drive general receptor properties such as protein stability (RMSD), conformational flexibility (RMSF), and compactness (RG), one needs to analyze the intramolecular interactions. In this respect, non-covalent interactions between residues play an important role. Also, non-covalent interactions are critical for ligand recognition. In MD simulations, these interactions and their stability can be predicted based on atom distances and angles. Using this methodology, Dror et al. were able to discern the importance of an ionic lock interaction for the conformation change produced during the activation of β_2ARs [140].

Most of the aforementioned MD analysis tools (e.g., MDAnalysis, GROMACS, VMD) focus on hydrogen bond interactions, as this is one of the most abundant and structurally important interaction types in proteins. However, there are many other interaction types that should not be neglected, including van der Waals, salt bridges, π-cation, and π-stacking interactions. To analyze them, more specialized tools have been developed, such as the python module GetContacts [141]. A good example of the capabilities of this module can be found in the Receptor Meta-analysis web tool

(https://submission.gpcrmd.org/contmaps/) included in the GPCRmd platform [53]. This tool analyzes and compares different types of non-covalent interactions obtained from a large GPCR simulation dataset using GetContacts scripts, and displays them into a series of interactive plots (Figure 5). This allows extracting conclusions about the interaction pattern of different GPCRs.

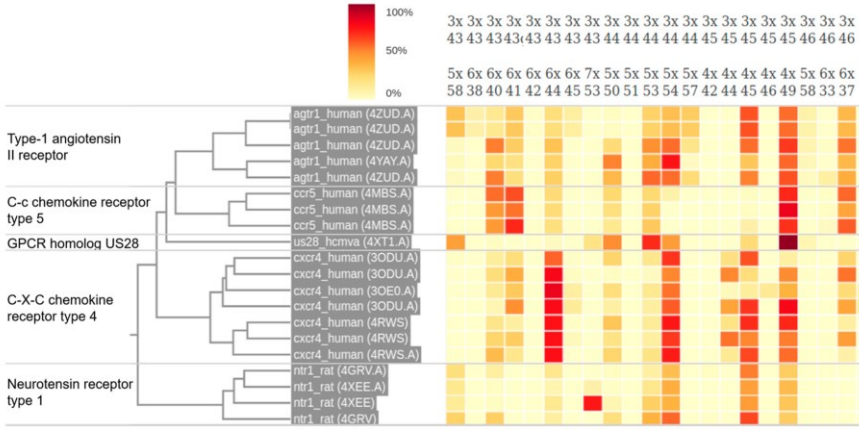


Figure 5. Pattern of total interaction frequency of several MD simulations of GPCRs, extracted from the GPCRmd Receptor Meta-analysis tool (https://submission.gpcrmd.org/contmaps/) of the GPCRmd server [53]. Columns represent interacting residue pairs according to Ballesteros-Weinstein residue numbering [142], whereas rows represent different simulations. The color of each cell shows the frequency in which any type of non-covalent interaction occurs during the simulation. Results are clustered based on the interaction frequencies of the simulations. This clustering is able to separate simulations according to the receptor subtype, showing that different receptor subtypes present differentiated interaction patterns.

5.2. Analysis of the Allosteric Communication

Allostery is a property of a protein by which perturbations that take place in one part of its structure are transmitted to distant parts of it. GPCRs are an excellent example of allosteric proteins. As described in Section 2.1, the binding of a ligand in a GPCR causes local structural changes in the binding pocket, which are transmitted across the receptor leading to a global conformational change and, in turn, a specific signaling response. This is mediated by a complex allosteric network in which multiple nodes (i.e., residues) transmit a specific perturbation through the whole protein. Not only GPCRs

present this phenomenon. Multiple studies have explored allostery in many other proteins like thrombin and PDZ domains among others [143,144].

An accurate model to capture protein allostery is an important focus of current research efforts. For instance, knowing how a drug candidate affects allosteric communication would be of great help to fine-tune its potency and efficacy in drug development programs. Moreover, protein engineering could be considerably improved if we were able to access the repercussion of a mutation in the protein structure and, thus, its functional outcome.

MD simulation has been used in many studies as a tool to analyze protein allostery. However, the way in which researchers look at this data is heterogeneous. Numerous studies rely on the comparison of the structural and dynamic behavior between two or more conditions. Others focus on analyzing the transition between two conformational states. In this case, the role of metadynamics is crucial, given that some of these transitions happen in time scales not accessible for classical (non-biased) simulations [145]. Another approach is to focus on changes in the conformational space, which is commonly studied using principal components analysis [146]. Others base their studies on the correlations in the movement of residues [147]. For this, the use of information theory-based methodologies is the most common approach to measure dependence between residues or groups of residues [148]. Finally, some researchers pay more attention to variables influenced by the chemical context of the residues, such as the contacts with other residues [149].

In many cases, some of these relationships are used to build networks. In these networks, residues are represented as nodes, while edges represent the level of coupling between the residues. Then, centrality and community analysis can be applied to the network to find the residues that contribute the most to communication inside the protein [147].

Some of the most influential works in the field combine several of the approaches mentioned. For example, Dror et al. studied the conformational correlation of β_2AR subdomains in different activation states to propose an activation mechanism of this receptor

[28]. Also, Miao et al. analyzed metadynamics simulations of the M_2 muscarinic receptor using a network representation of the residue cross-correlation [150]. This analysis allowed them to characterize some aspects of the receptor activation. Finally, Bhattacharya and Vaidehi investigated network representations of the inter-residue dihedral correlation of the β_2AR [151]. The resulting model describing allosteric communication was able to identify allosteric pockets and identify residues that affected function upon mutation.

Overall, this field has a great potential for understanding GPCR pharmacology but is still challenging and requires the development of more robust protocols. This robustness might be achieved by integrating the different methodologies that are being used into a more complete analysis.

6. Current Challenges

The capabilities of MD simulations have broadened substantially thanks to the technological advances of the last decades. However, there are still some relevant drawbacks that limit the usability of this technique and must be taken into account.

As described in Section 4, the forces of a MD simulation are calculated based on a force field, which consists of a set of empirical potential energy functions. Force fields are based on quantum mechanical calculations and experimental measurements, and include some approximations. As such, force fields are imperfect. Studies comparing simulation results with experimental data indicate that force fields have improved significantly over the past decade [152], but more remains to be done to achieve increased accuracy. Another limitation of classical MD is that it is not possible to form or break covalent bonds during the simulation. As a consequence, protonation states of titratable amino acid residues are fixed, as well as disulfide bonds. Thus, they have to be set carefully at the beginning of the simulation [113].

An important challenge that needs to be taken into account is the simulation timescale. The simulation timestep, which is the time length between evaluations of the potential, needs to be small

enough to capture the fastest movements in the simulation system. This typically limits the timestep to around 2 fs. Many relevant molecular events, however, take part in the microsecond to millisecond scale, or even longer. This implies the calculation of a vast number of timesteps, each of which involves the calculation of millions of interatomic interactions. As a consequence, reaching long timescales can be challenging for classical MD. Furthermore, the issue with long-timescale events is that they imply the transition between free energy states that are separated by high-energy barriers. In this situation, classical MD simulations tend to get trapped in one of these local minimum-energy states for a long time, which restrains the sampling process. In turn, this leads to a poor characterization of the protein's dynamic behavior [153]. A useful strategy to tackle the sampling problem is the application of enhanced sampling techniques. Enhanced sampling simulations, including replica-exchange MD, metadynamics, and simulated annealing, are able to efficiently overcome energetic barriers and access additional conformational states by including an external bias [22,154]. Simplified models like coarse-graining can also extend accessible timescales by orders of magnitude, as they are less expensive computationally [155]. Nevertheless, the problem of achieving relevant simulation timescales with classical, all-atom MD seems to be within reach of being solved [156]. In recent years, there has been a dramatic increase in achieved timescales. This tendency is expected to continue, thanks to the advances in algorithms [8,9,157,158], software [159,160,161,162], and hardware [163,164] that we are experiencing. In fact, it is expected that all-atom, classical MD simulations will be able to reach the second timescale within the next five years [165,166,167].

Parallel to the limitation of longer timescales accessible to simulations, there is a limitation in the size of the systems that can be studied. As the system size increases, so does the computational power needed to carry out the simulation. In general, the required computational power increases with the square of the number of atoms involved. Moreover, as molecular systems become bigger, the biologically relevant timescales tend to increase too [165]. Overall, this challenges the study of GPCRs in complex with G protein or arrestin. Enhanced sampling techniques are a promising approach for the study of such systems. However, selecting predefined collective variables for the simulation of protein-protein

interactions is a difficult task, as such processes often involve large-scale translations and rotations of the binding partners, as well as complex conformational changes. Thus, methods that do not require predefined collective variables, such as Gaussian accelerated MD, are especially convenient. In fact, recently Miao et al. successfully applied Gaussian accelerated MD to simulate the intracellular association between the M_2 receptor and a G-protein mimetic nanobody [168]. Their simulations revealed important insights into the binding mechanism, despite the fact that the calculated free energies were not converged. Future developments will be needed to achieve converged simulations of such complex systems.

Given this fast evolution of the capabilities of MD simulation, this technique is gaining more and more relevance. In view of this, it is becoming increasingly necessary to define standards and best practices to ensure a reproducible research output [169]. Many challenges remain in order to effectively reach this goal. One issue is the creation of workflows for simulation production and analysis. The file formats and force fields supported by different programs are often incompatible. This limits the combination of software packages that can be used together in a workflow and restricts the choice of algorithms and force fields based on software compatibility rather than scientific-based reasons. Luckily, this can be solved with the development and usage of software that converts molecular information between the different file formats. Still, this does not solve the problem that different programs, or program versions, may implement force fields and features, such as thermostats and integrators, in different ways. Thus, the results of a workflow will be influenced by the combination of programs used [170]. Because of this, it is always important to disclose the version and name of all programs used. In fact, detailed documentation of the entire workflow should always be provided when publishing a simulation. The level of detail in documenting the workflows needs to be enough to ensure the reproducibility of the obtained results. Finally, another challenge that needs to be overcome to achieve reproducibility is data sharing. Data sharing is still not widely adopted in the field of MD simulation, partly because of the technical difficulties derived from the increasing size of the generated trajectories. More efforts should be done to define best practices and guidelines for simulation data sharing [171]. Luckily, many researchers work to promote it [172], and different initiatives

are addressing this issue. Several software packages [173,174,175,176] have been developed to share trajectories by providing online interactive visualization based on the advantages of the WebGL API. Moreover, several community-driven projects provide specialized platforms for deposition and analysis of MD simulations [53,136,177,178,179,180,181]. In the case of GPCRs, GPCRmd [53] is an online resource specialized in the deposition and analysis of GPCR MD simulations.

7. Conclusions and Perspectives

MD simulations are a potent computational technique capable of generating high-resolution simulations of the structural motions of a molecular system. They can either capture atomic-level motions within a specific conformational state or structural transitions between different conformational populations, bringing within reach information that is difficult, or even impossible, to obtain by other methods [14]. This makes MD simulation a promising technique for the study of GPCRs, whose functionality is highly determined by their ability to transition between conformations. In fact, MD simulations have proven their usefulness for the study of important biological processes in GPCRs such as ligand binding, allostery, activation, natural genetic variation, and addition of post-translational modifications, among others. Since GPCRs are drug targets of striking importance in the pharmaceutical industry [182], all this information generated by MD simulations has the potential to accelerate the discovery of new and improved drugs targeting these proteins.

In order for MD simulation to reach its full potential, some difficulties need to be overcome. Fortunately, we are in an era of rapid technological development, which creates great prospects for the advancement of this field in the following years. Computational power is expected to continue increasing following Moore's law, which describes how the performance of integrated circuits has been increasing exponentially over the past half-century [183]. This would imply a reduction in computational costs. At the same time, we expect methodological advances in MD algorithms, including improvements in the fine-tuning of energy calculations, parallelization, GPU exploitation, and algorithmic methods to

increase the sampling of conformational space. Overall, this would cause an increase in the timescales available to simulations. Several authors propose that we may even reach the second timescale within the next five years [165,166,167], bridging the gap between the timescales of biological processes observed in vivo and those accessible in silico. Parallel to timescales would come a growth in the size of the systems that can be studied [165]. This, together with an ever-growing accuracy in the force fields, will grant us the opportunity to extend the application of MD simulations to the study of processes that were previously difficult to capture. This may open the door to significantly advance in the study of macromolecule-macromolecule interactions [184], including GPCR oligomerization, and coupling to intracellular signaling proteins. While coarse-grained MD has been typically used for this type of study [185,186], it is important to capture the effects of macromolecule-macromolecule interactions on the structural dynamics and cell signaling through more detailed MD simulations [107].

Finally, as simulations become faster, cheaper, and more widely accessible, new opportunities will arise for drug discovery. In the past, most drug discovery programs have disregarded MD analysis because of their computational expenses. With the forthcoming reduction of the computational costs associated with MD simulations, this technique is expected to be more commonly applied in the pharmaceutical industry and, eventually, to be commonly included in drug discovery pipelines [184,187,188].

Author Contributions

Conceptualization: J.S. and M.T.-F.; manuscript writing: M.T.-F., T.M.S., D.A.-G., A.M.-P., B.M.-L. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding

MTF acknowledges financial support from the Spanish Ministry of Science, Innovation and Universities (FPU16/01209). TMS would like to acknowledge support from the National Center of Science, Poland (grant number 2017/27/N/NZ2/02571). AMP acknowledges

financial support from the Instituto de Salud Carlos III FEDER (FI19/00037). Finally, JS acknowledges financial support from the Instituto de Salud Carlos III FEDER (PI15/00460 and PI18/00094) and the ERA-NET NEURON & Ministry of Economy, Industry and Competitiveness (AC18/00030).

Conflicts of Interest

The authors declare no conflict of interest.

Abbreviations

GPCR	G protein-coupled receptor
Cryo-EM	Cryo-electron microscopy
MD	Molecular dynamics
GPU	Graphical processor unit
β_2AR	β_2-adrenergic receptor
A_{2A}R	A_{2A} receptor
TM	Transmembrane helix
MSMs	Markov state models
POPC	1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine
VMD	Virtual Molecular Dynamics
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
RG	Radius of gyration

References

1. Hauser, A.S.; Attwood, M.M.; Rask-Andersen, M.; Schiöth, H.B.; Gloriam, D.E. Trends in GPCR drug discovery: New agents, targets and indications. *Nat. Rev. Drug Discov.* **2017**, *16*, 829–842.

2. Congreve, M.; de Graaf, C.; Swain, N.A.; Tate, C.G. Impact of GPCR Structures on Drug Discovery. *Cell* **2020**, *181*, 81–91.
3. Jazayeri, A.; Andrews, S.P.; Marshall, F.H. Structurally enabled discovery of adenosine a2a receptor antagonists. *Chem. Rev.* **2017**, *117*, 21–37.
4. 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*, D440–D446.
5. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
6. McCammon, J.A.; Gelin, B.R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585–590.
7. Dahl, S.G.; Edvardsen, O.; Sylte, I. Molecular dynamics of dopamine at the D2 receptor. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 8111–8115.
8. Stone, J.E.; Phillips, J.C.; Freddolino, P.L.; Hardy, D.J.; Trabuco, L.G.; Schulten, K. Accelerating molecular modeling applications with graphics processors. *J. Comput. Chem.* **2007**, *28*, 2618–2640.
9. Anderson, J.A.; Lorenz, C.D.; Travesset, A. General purpose molecular dynamics simulations fully implemented on graphics processing units. *J. Comput. Phys.* **2008**, *227*, 5342–5359.
10. Jo, S.; Kim, T.; Iyer, V.G.; Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859–1865.
11. Mayol, E.; García-Recio, A.; Tiemann, J.K.S.; Hildebrand, P.W.; Guixà, R.; Guixà-Gonzálezgonz'gonzález, G.; Olivella, M.; Cordomí, A. HomolWat: A web server tool to incorporate "homologous" water molecules into GPCR structures. *Nucleic Acids Res.* **2020**, *1*, 13–14.
12. 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*, 2319–2327.
13. Gowers, R.J.; Linke, M.; Barnoud, J.; Reddy, T.J.E.; Melo, M.N.; Seyler, S.L.; Domański, J.; Dotson, D.L.; Buchoux, S.; Kenney, I.M.; et al. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In Proceedings of the 15th Python in Science Conference, Austin, TX, USA, 11–17 July 2016; pp. 98–105.
14. Latorraca, N.R.; Venkatakrishnan, A.J.; Dror, R.O. GPCR Dynamics: Structures in Motion. *Chem. Rev.* **2017**, *117*, 139–155.
15. 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.; et al. Membrane cholesterol access into a G-protein-coupled receptor. *Nat. Commun.* **2017**, *8*, 14505.
16. Ramírez-Anguita, J.M.; Rodríguez-Espigares, I.; Guixà-González, R.; Bruno, A.; Torrens-Fontanals, M.; Varela-Rial, A.; Selent, J. Membrane cholesterol effect on the 5-HT2A receptor: Insights into the lipid-induced modulation of an antipsychotic drug target. *Biotechnol. Appl. Biochem.* **2018**, *65*, 29–37.

17. Karoussiotis, C.; Marti-Solano, M.; Stepniewski, T.M.; Symeonof, A.; Selent, J.; Georgoussi, Z. A highly conserved δ-opioid receptor region determines RGS4 interaction. *FEBS J.* **2020**, *287*, 736–748.
18. Eichel, K.; Jullié, D.; Barsi-Rhyne, B.; Latorraca, N.R.; Masureel, M.; Sibarita, J.B.; Dror, R.O.; Von Zastrow, M. Catalytic activation of β-Arrestin by GPCRs. *Nature* **2018**, *557*, 381–386.
19. Guixà-González, R.; Javanainen, M.; Gómez-Soler, M.; Cordobilla, B.; Domingo, J.C.; Sanz, F.; Pastor, M.; Ciruela, F.; Martinez-Seara, H.; Selent, J. Membrane omega-3 fatty acids modulate the oligomerisation kinetics of adenosine A2A and dopamine D2 receptors. *Sci. Rep.* **2016**, *6*, 19839.
20. Samama, P.; Cotecchia, S.; Costa, T.; Lefkowitz, R.J. A mutation-induced activated state of the β2-adrenergic receptor. Extending the ternary complex model. *J. Biol. Chem.* **1993**, *268*, 4625–4636.
21. Hunyady, L.; Vauquelin, G.; Vanderheyden, P. Agonist induction and conformational selection during activation of a G-protein-coupled receptor. *Trends Pharmacol. Sci.* **2003**, *24*, 81–86.
22. Rodríguez-Espigares, I.; Kaczor, A.A.; Selent, J. In silico Exploration of the Conformational Universe of GPCRs. *Mol. Inform.* **2016**, *35*, 227–237.
23. Kenakin, T.; Miller, L.J. Seven transmembrane receptors as shapeshifting proteins: The impact of allosteric modulation and functional selectivity on new drug discovery. *Pharmacol. Rev.* **2010**, *62*, 265–304.
24. Rasmussen, S.G.F.; DeVree, B.T.; Zou, Y.; Kruse, A.C.; Chung, K.Y.; Kobilka, T.S.; Thian, F.S.; Chae, P.S.; Pardon, E.; Calinski, D.; et al. Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature* **2011**, *477*, 549–555.
25. Kang, Y.; Zhou, X.E.; Gao, X.; He, Y.; Liu, W.; Ishchenko, A.; Barty, A.; White, T.A.; Yefanov, O.; Han, G.W.; et al. Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. *Nature* **2015**, *523*, 561–567.
26. Martí-Solano, M.; Schmidt, D.; Kolb, P.; Selent, J. Drugging specific conformational states of GPCRs: Challenges and opportunities for computational chemistry. *Drug Discov. Today* **2016**, *21*, 625–631.
27. Martí-Solano, M.; Guixà-González, R.; Sanz, F.; Pastor, M.; Selent, J. Novel insights into biased agonism at G protein-coupled receptors and their potential for drug design. *Curr. Pharm. Des.* **2013**, *19*, 5156–5166.
28. Dror, R.O.; Arlow, D.H.; Maragakis, P.; Mildorf, T.J.; Pan, A.C.; Xu, H.; Borhani, D.W.; Shaw, D.E. Activation mechanism of the β 2-adrenergic receptor. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 18684–18689.
29. Nygaard, R.; Zou, Y.; Dror, R.O.; Mildorf, T.J.; Arlow, D.H.; Manglik, A.; Pan, A.C.; Liu, C.W.; Fung, J.J.; Bokoch, M.P.; et al. The dynamic process of β(2)-adrenergic receptor activation. *Cell* **2013**, *152*, 532–542.
30. Kohlhoff, K.J.; Shukla, D.; Lawrenz, M.; Bowman, G.R.; Konerding, D.E.; Belov, D.; Altman, R.B.; Pande, V.S. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.* **2014**, *6*, 15–21.
31. Li, J.; Jonsson, A.L.; Beuming, T.; Shelley, J.C.; Voth, G.A. Ligand-dependent activation and deactivation of the human adenosine A 2A receptor. *J. Am. Chem. Soc.* **2013**, *135*, 8749–8759.

32. Liu, J.J.; Horst, R.; Katritch, V.; Stevens, R.C.; Wuthrich, K. Biased Signaling Pathways in 2-Adrenergic Receptor Characterized by 19F-NMR. *Science.* **2012**, *335*, 1106–1110.
33. Oddi, S.; Stepniewski, T.M.; Totaro, A.; Selent, J.; Scipioni, L.; Dufrusine, B.; Fezza, F.; Dainese, E.; Maccarrone, M. Palmitoylation of cysteine 415 of CB 1 receptor affects ligand-stimulated internalization and selective interaction with membrane cholesterol and caveolin 1. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* **2017**, *1862*, 523–532.
34. Oddi, S.; Totaro, A.; Scipioni, L.; Dufrusine, B.; Stepniewski, T.M.; Selent, J.; Maccarrone, M.; Dainese, E. Role of palmitoylation of cysteine 415 in functional coupling CB_1 receptor to $G\alpha_{i2}$ protein. *Biotechnol. Appl. Biochem.* **2018**, *65*, 16–20.
35. Dawaliby, R.; Trubbia, C.; Delporte, C.; Masureel, M.; Van Antwerpen, P.; Kobilka, B.K.; Govaerts, C. Allosteric regulation of G protein-coupled receptor activity by phospholipids. *Nat. Chem. Biol.* **2016**, *12*, 35–39.
36. Neale, C.; Herce, H.D.; Pomès, R.; García, A.E. Can Specific Protein-Lipid Interactions Stabilize an Active State of the Beta 2 Adrenergic Receptor? *Biophys. J.* **2015**, *109*, 1652–1662.
37. Yen, H.Y.; Hoi, K.K.; Liko, I.; Hedger, G.; Horrell, M.R.; Song, W.; Wu, D.; Heine, P.; Warne, T.; Lee, Y.; et al. PtdIns(4,5)P2 stabilizes active states of GPCRs and enhances selectivity of G-protein coupling. *Nature* **2018**, *559*, 423–427.
38. Bruzzese, A.; Gil, C.; Dalton, J.A.R.; Giraldo, J. Structural insights into positive and negative allosteric regulation of a G protein-coupled receptor through protein-lipid interactions. *Sci. Rep.* **2018**, *8*, 1–14.
39. Paila, Y.D.; Chattopadhyay, A. The function of G-protein coupled receptors and membrane cholesterol: Specific or general interaction? *Glycoconj. J.* **2009**, *26*, 711–720.
40. Oates, J.; Watts, A. Uncovering the intimate relationship between lipids, cholesterol and GPCR activation. *Curr. Opin. Struct. Biol.* **2011**, *21*, 802–807.
41. Gimpl, G. Interaction of G protein coupled receptors and cholesterol. *Chem. Phys. Lipids* **2016**, *199*, 61–73.
42. Khelashvili, G.; Mondal, S.; Andersen, O.S.; Weinstein, H. Cholesterol Modulates the Membrane Effects and Spatial Organization of Membrane-Penetrating Ligands for G-Protein Coupled Receptors. *J. Phys. Chem. B* **2010**, *114*, 12046–12057.
43. Zocher, M.; Zhang, C.; Rasmussen, S.G.F.; Kobilka, B.K.; Müller, D.J. Cholesterol increases kinetic, energetic, and mechanical stability of the human β2-adrenergic receptor. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 20186.
44. Lee, A.G. How lipids affect the activities of integral membrane proteins. *Biochim. Biophys. Acta Biomembr.* **2004**, *1666*, 62–87.
45. 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 β2-Adrenergic Receptor. *Structure* **2008**, *16*, 897–905.
46. Laganowsky, A.; Reading, E.; Allison, T.M.; Ulmschneider, M.B.; Degiacomi, M.T.; Baldwin, A.J.; Robinson, C.V. Membrane proteins bind

lipids selectively to modulate their structure and function. *Nature* **2014**, *510*, 172–175.
47. Ng, H.W.; Laughton, C.A.; Doughty, S.W. Molecular dynamics simulations of the adenosine A2a receptor in POPC and POPE lipid bilayers: Effects of membrane on protein behavior. *J. Chem. Inf. Model.* **2014**, *54*, 573–581.
48. Bruzzese, A.; Dalton, J.A.R.; Giraldo, J. Insights into adenosine A2A receptor activation through cooperative modulation of agonist and allosteric lipid interactions. *PLoS Comput. Biol.* **2020**, *16*, e1007818.
49. Corey, R.A.; Vickery, O.N.; Sansom, M.S.P.; Stansfeld, P.J. Insights into Membrane Protein-Lipid Interactions from Free Energy Calculations. *J. Chem. Theory Comput.* **2019**, *15*, 5727–5736.
50. Wei, H.; Ahn, S.; Shenoy, S.K.; Karnik, S.S.; Hunyady, L.; Luttrell, L.M.; Lefkowitz, R.J. Independent β-arrestin 2 and G protein-mediated pathways for angiotensin II activation of extracellular signal-regulated kinases 1 and 2. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 10782–10787.
51. Azzi, M.; Charest, P.G.; Angers, S.; Rousseau, G.; Kohout, T.; Bouvier, M.; Piñeyro, G. β-arrestin-mediated activation of MAPK by inverse agonists reveals distinct active conformations for G protein-coupled receptors. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 11406–11411.
52. Wacker, D.; Stevens, R.C.; Roth, B.L. How Ligands Illuminate GPCR Molecular Pharmacology. *Cell* **2017**, *170*, 414–427.
53. Rodríguez-Espigares, I.; Torrens-Fontanals, M.; Tiemann, J.K.S.; Aranda-García, D.; Ramírez-Anguita, J.M.; Stepniewski, T.M.; Worp, N.; Varela-Rial, A.; Morales-Pastor, A.; Medel-Lacruz, B.; et al. GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods* **2020**, 1–11.
54. Ferruz, N.; Doerr, S.; Vanase-Frawley, M.A.; Zou, Y.; Chen, X.; Marr, E.S.; Nelson, R.T.; Kormos, B.L.; Wager, T.T.; Hou, X.; et al. Dopamine D3 receptor antagonist reveals a cryptic pocket in aminergic GPCRs. *Sci. Rep.* **2018**, *8*, 1–10.
55. Amaro, R.E.; Baron, R.; McCammon, J.A. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput. Aided. Mol. Des.* **2008**, *22*, 693–705.
56. Stepniewski, T.M.; Torrens-Fontanals, M.; Rodríguez-Espigares, I.; Giorgino, T.; Primdahl, K.G.; Vik, A.; Stenstrøm, Y.; Selent, J.; Hansen, T.V. Synthesis, molecular modelling studies and biological evaluation of new oxoeicosanoid receptor 1 agonists. *Bioorganic Med. Chem.* **2018**, *26*, 3580–3587.
57. Azuaje, J.; López, P.; Iglesias, A.; De La Fuente, R.A.; Pérez-Rubio, J.M.; García, D.; Stępniewski, T.M.; García-Mera, X.; Brea, J.M.; Selent, J.; et al. Development of Fluorescent Probes that Target Serotonin 5-HT2B Receptors. *Sci. Rep.* **2017**, *7*, 1–16.
58. Guo, D.; Mulder-Krieger, T.; IJzerman, A.P.; Heitman, L.H. Functional efficacy of adenosine A_2A receptor agonists is positively correlated to their receptor residence time. *Br. J. Pharmacol.* **2012**, *166*, 1846–1859.
59. Swinney, D.C. Applications of Binding Kinetics to Drug Discovery. *Pharmaceut. Med.* **2008**, *22*, 23–34.
60. Lu, H.; Tonge, P.J. Drug-target residence time: Critical information for lead optimization. *Curr. Opin. Chem. Biol.* **2010**, *14*, 467–474.

61. Pan, A.C.; Borhani, D.W.; Dror, R.O.; Shaw, D.E. Molecular determinants of drug-receptor binding kinetics. *Drug Discov. Today* **2013**, *18*, 667–673.
62. Dror, R.O.; Pan, A.C.; Arlow, D.H.; Borhani, D.W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D.E. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 13118–13123.
63. Wittmann, H.-J.J.; Strasser, A. Binding pathway of histamine to the hH4R, observed by unconstrained molecular dynamics. *Bioorganic Med. Chem. Lett.* **2015**, *25*, 1259–1268.
64. Sabbadin, D.; Ciancetta, A.; Deganutti, G.; Cuzzolin, A.; Moro, S. Exploring the recognition pathway at the human A2A adenosine receptor of the endogenous agonist adenosine using supervised molecular dynamics simulations. *Medchemcomm* **2015**, *6*, 1081–1085.
65. Thomas, T.; Fang, Y.; Yuriev, E.; Chalmers, D.K. Ligand Binding Pathways of Clozapine and Haloperidol in the Dopamine D_2 and D_3 Receptors. *J. Chem. Inf. Model.* **2016**, *56*, 308–321.
66. Rodríguez-Espigares, I.; Kaczor, A.A.; Stepniewski, T.M.; Selent, J. Challenges and opportunities in drug discovery of biased ligands. In *Computational Methods for GPCR Drug Discovery*; Heifetz, A., Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, USA, 2018; Volume 1705, pp. 321–334. ISBN 9781493974641.
67. Martí-Solano, M.; Iglesias, A.; de Fabritiis, G.; Sanz, F.; Brea, J.; Loza, M.I.; Pastor, M.; Selent, J. Detection of New Biased Agonists for the Serotonin 5-HT2A Receptor: Modeling and Experimental Validation. *Mol. Pharmacol.* **2015**, *87*, 740–746.
68. McCorvy, J.D.; Butler, K.V.; Kelly, B.; Rechsteiner, K.; Karpiak, J.; Betz, R.M.; Kormos, B.L.; Shoichet, B.K.; Dror, R.O.; Jin, J.; et al. Structure-inspired design of β-arrestin-biased ligands for aminergic GPCRs. *Nat. Chem. Biol.* **2018**, *14*, 126–134.
69. Bermudez, M.; Bock, A.; Krebs, F.; Holzgrabe, U.; Mohr, K.; Lohse, M.J.; Wolber, G. Ligand-Specific Restriction of Extracellular Conformational Dynamics Constrains Signaling of the M2 Muscarinic Receptor. *ACS Chem. Biol.* **2017**, *12*, 1743–1748.
70. Suomivuori, C.M.; Latorraca, N.R.; Wingler, L.M.; Eismann, S.; King, M.C.; Kleinhenz, A.L.W.; Skiba, M.A.; Staus, D.P.; Kruse, A.C.; Lefkowitz, R.J.; et al. Molecular mechanism of biased signaling in a prototypical G protein-coupled receptor. *Science* **2020**, *367*, 881–887.
71. Martí-Solano, M.; Sanz, F.; Pastor, M.; Selent, J. A dynamic view of molecular switch behavior at serotonin receptors: Implications for Functional selectivity. *PLoS ONE* **2014**, *9*, e109312.
72. Kapoor, A.; Martinez-Rosell, G.; Provasi, D.; de Fabritiis, G.; Filizola, M. Dynamic and Kinetic Elements of μ-Opioid Receptor Functional Selectivity. *Sci. Rep.* **2017**, *7*, 11255.
73. Nivedha, A.K.; Tautermann, C.S.; Bhattacharya, S.; Lee, S.; Casarosa, P.; Kollak, I.; Kiechle, T.; Vaidehi, N. Identifying functional hotspot residues for biased ligand design in G-protein-coupled receptors. *Mol. Pharmacol.* **2018**, *93*, 288–296.

74. Zou, Y.; Ewalt, J.; Ng, H.L. Recent insights from molecular dynamics simulations for g protein-coupled receptor drug discovery. *Int. J. Mol. Sci.* **2019**, *20*, 4237.
75. Wootten, D.; Christopoulos, A.; Sexton, P.M. Emerging paradigms in GPCR allostery: Implications for drug discovery. *Nat. Publ. Gr.* **2013**, *12*, 630–644.
76. Allen, J.A.; Roth, B.L. Strategies to Discover Unexpected Targets for Drugs Active at G Protein–Coupled Receptors. *Annu. Rev. Pharmacol. Toxicol.* **2011**, *51*, 117–144.
77. Hertig, S.; Latorraca, N.R.; Dror, R.O. Revealing Atomic-Level Mechanisms of Protein Allostery with Molecular Dynamics Simulations. *PLoS Comput. Biol.* **2016**, *12*, e1004746.
78. Miao, Y.; Goldfeld, D.A.; Von Moo, E.; Sexton, P.M.; Christopoulos, A.; McCammon, J.A.; Valant, C. Accelerated structure-based design of chemically diverse allosteric modulators of a muscarinic G protein-coupled receptor. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E5675–E5684.
79. Bock, A.; Schrage, R.; Mohr, K. Allosteric modulators targeting CNS muscarinic receptors. *Neuropharmacology* **2017**, *136*, 427–437.
80. Dror, R.O.; Green, H.F.; Valant, C.; Borhani, D.W.; Valcourt, J.R.; Pan, A.C.; Arlow, D.H.; Canals, M.; Lane, J.R.; Rahmani, R.; et al. Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs. *Nature* **2013**, *503*, 295–299.
81. Chan, H.C.S.; Wang, J.; Palczewski, K.; Filipek, S.; Vogel, H.; Liu, Z.J.; Yuan, S. Exploring a new ligand binding site of G protein-coupled receptors. *Chem. Sci.* **2018**, *9*, 6480–6489.
82. Angel, T.E.; Chance, M.R.; Palczewski, K. Conserved waters mediate structural and functional activation of family A (rhodopsin-like) G protein-coupled receptors. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 8555–8560.
83. Yuan, S.; Vogel, H.; Filipek, S. The Role of Water and Sodium Ions in the Activation of the μ-Opioid Receptor. *Angew. Chemie Int. Ed.* **2013**, *52*, 10112–10115.
84. Yuan, S.; Palczewski, K.; Peng, Q.; Kolinski, M.; Vogel, H.; Filipek, S. The mechanism of ligand-induced activation or inhibition of μ- And κ-opioid receptors. *Angew. Chemie Int. Ed.* **2015**, *54*, 7560–7563.
85. Yuan, S.; Hu, Z.; Filipek, S.; Vogel, H. W2466.48 opens a gate for a continuous intrinsic water pathway during activation of the adenosine A2A receptor. *Angew. Chemie Int. Ed.* **2015**, *54*, 556–559.
86. Yuan, S.; Filipek, S.; Palczewski, K.; Vogel, H. Activation of G-protein-coupled receptors correlates with the formation of a continuous internal water pathway. *Nat. Commun.* **2014**, *5*, 4733.
87. Schmidtke, P.; Javier Luque, F.; Murray, J.B.; Barril, X. Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. *J. Am. Chem. Soc.* **2011**, *133*, 18903–18910.
88. Magarkar, A.; Schnapp, G.; Apel, A.K.; Seeliger, D.; Tautermann, C.S. Enhancing Drug Residence Time by Shielding of Intra-Protein Hydrogen Bonds: A Case Study on CCR2 Antagonists. *ACS Med. Chem. Lett.* **2019**, *10*, 324–328.
89. Zarzycka, B.; Zaidi, S.A.; Roth, B.L.; Katritch, V. Harnessing ion-binding sites for GPCR pharmacology. *Pharmacol. Rev.* **2019**, *71*, 571–595.

90. Selent, J.; Sanz, F.; Pastor, M.; de Fabritiis, G. Induced effects of sodium ions on dopaminergic G-protein coupled receptors. *PLoS Comput. Biol.* **2010**, *6*, e1000884.
91. Hu, X.; Wang, Y.; Hunkele, A.; Provasi, D.; Pasternak, G.W.; Filizola, M. Kinetic and thermodynamic insights into sodium ion translocation through the μ-opioid receptor from molecular dynamics and machine learning analysis. *PLOS Comput. Biol.* **2019**, *15*, e1006689.
92. Gutiérrez-De-Terán, H.; Massink, A.; Rodríguez, D.; Liu, W.; Han, G.W.; Joseph, J.S.; Katritch, I.; Heitman, L.H.; Xia, L.; Ijzerman, A.P.; et al. The role of a sodium ion binding site in the allosteric modulation of the A2A adenosine G protein-coupled receptor. *Structure* **2013**, *21*, 2175–2185.
93. Selvam, B.; Shamsi, Z.; Shukla, D. Universality of the Sodium Ion Binding Mechanism in Class A G-Protein-Coupled Receptors. *Angew. Chemie Int. Ed.* **2018**, *57*, 3048–3053.
94. Vickery, O.N.; Carvalheda, C.A.; Zaidi, S.A.; Pisliakov, A.V.; Katritch, V.; Zachariae, U. Intracellular Transfer of Na+ in an Active-State G-Protein-Coupled Receptor. *Structure* **2018**, *26*, 171–180.e2.
95. Chan, H.C.S.; Xu, Y.; Tan, L.; Vogel, H.; Cheng, J.; Wu, D.; Yuan, S. Enhancing the Signaling of GPCRs via Orthosteric Ions. *ACS Cent. Sci.* **2020**, *6*, 274–282.
96. Sherry, S.T.; Ward, M.-H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E.M.; Sirotkin, K. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29*, 308–311.
97. Mallory, D.P.; Gutierrez, E.; Pinkevitch, M.; Klinginsmith, C.; Comar, W.D.; Roushar, F.J.; Schlebach, J.P.; Smith, A.W.; Jastrzebska, B. The Retinitis Pigmentosa-Linked Mutations in Transmembrane Helix 5 of Rhodopsin Disrupt Cellular Trafficking Regardless of Oligomerization State. *Biochemistry* **2018**, *57*, 5188–5201.
98. Julian, B.; Gao, K.; Harwood, B.N.; Beinborn, M.; Kopin, A.S. Mutation-induced functional alterations of CCR6. *J. Pharmacol. Exp. Ther.* **2017**, *360*, 106–116.
99. Hauser, A.S.; Chavali, S.; Masuho, I.; Jahn, L.J.; Martemyanov, K.A.; Gloriam, D.E.; Babu, M.M. Pharmacogenomics of GPCR Drug Targets. *Cell* **2018**, *172*, 41–54.e19.
100. Sengupta, D.; Sonar, K.; Joshi, M. Characterizing clinically relevant natural variants of GPCRs using computational approaches. *Methods Cell Biol.* **2017**, *142*, 187–204.
101. Tandale, A.; Joshi, M.; Sengupta, D. Structural insights and functional implications of inter-individual variability in β 2 -adrenergic receptor. *Sci. Rep.* **2016**, *6*, 1–11.
102. Shahane, G.; Parsania, C.; Sengupta, D.; Joshi, M. Molecular insights into the dynamics of pharmacogenetically important N-terminal variants of the human β2-adrenergic receptor. *PLoS Comput. Biol.* **2014**, *10*, e1004006.
103. Bhosale, S.; Nikte, S.V.; Sengupta, D.; Joshi, M. Differential Dynamics Underlying the Gln27Glu Population Variant of the β2-Adrenergic Receptor. *J. Membr. Biol.* **2019**, *252*, 499–507.
104. Paul, S.M.; Mytelka, D.S.; Dunwiddie, C.T.; Persinger, C.C.; Munos, B.H.; Lindborg, S.R.; Schacht, A.L. How to improve RD productivity:

The pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **2010**, *9*, 203–214.
105. Shoichet, B.K.; Kobilka, B.K. Structure-based drug screening for G-protein-coupled receptors. *Trends Pharmacol. Sci.* **2012**, *33*, 268–272.
106. Jazayeri, A.; Dias, J.M.; Marshall, F.H. From G protein-coupled receptor structure resolution to rational drug design. *J. Biol. Chem.* **2015**, *290*, 19489–19495.
107. Miao, Y.; McCammon, J.A. G-Protein Coupled Receptors: Advances in Simulation and Drug Discovery. *Curr. Opin. Struct. Biol.* **2016**, *41*, 83–89.
108. Lin, J.H.; Perryman, A.L.; Schames, J.R.; McCammon, J.A. Computational drug design accommodating receptor flexibility: The relaxed complex scheme. *J. Am. Chem. Soc.* **2002**, *124*, 5632–5633.
109. Velgy, N.; Hedger, G.; Biggin, P.C. GPCRs: What Can We Learn from Molecular Dynamics Simulations. In *Computational Methods for GPCR Drug Discovery*; Heifetz, A., Ed.; Humana Press: Totowa, NJ, USA, 2018; pp. 133–158. ISBN 978-1-4939-7464-1.
110. Osguthorpe, D.J.; Sherman, W.; Hagler, A.T. Generation of Receptor Structural Ensembles for Virtual Screening Using Binding Site Shape Analysis and Clustering. *Chem. Biol. Drug Des.* **2012**, *80*, 182–193.
111. Jaiteh, M.; Rodríguez-Espigares, I.; Selent, J.; Carlsson, J. Performance of virtual screening against GPCR homology models: Impact of template selection and treatment of binding site plasticity. *PLoS Comput. Biol.* **2020**, *16*, e1007680.
112. Friesner, R.A.; Murphy, R.B.; Repasky, M.P.; Frye, L.L.; Greenwood, J.R.; Halgren, T.A.; Sanschagrin, P.C.; Mainz, D.T. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.
113. Hollingsworth, S.A.; Dror, R.O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129–1143.
114. Kappel, K.; Miao, Y.; Andrew McCammon, J. Accelerated molecular dynamics simulations of ligand binding to a muscarinic G-protein-coupled receptor. *Q. Rev. Biophys.* **2015**, *48*, 479–487.
115. Perez, A.; Morrone, J.A.; Simmerling, C.; Dill, K.A. Advances in free-energy-based simulations of protein folding and ligand binding. *Curr. Opin. Struct. Biol.* **2016**, *36*, 25–31.
116. Matricon, P.; Ranganathan, A.; Warnick, E.; Gao, Z.G.; Rudling, A.; Lambertucci, C.; Marucci, G.; Ezzati, A.; Jaiteh, M.; Dal Ben, D.; et al. Fragment optimization for GPCRs by molecular dynamics free energy calculations: Probing druggable subpockets of the A2A adenosine receptor binding site. *Sci. Rep.* **2017**, *7*, 1–12.
117. Jespers, W.; Verdon, G.; Azuaje, J.; Majellaro, M.; Keränen, H.; GArcía-Mera, X.; Congreve, M.; DeFlorian, F.; de Graaf, C.; Zhukov, A.; et al. X-Ray Crystallography and Free Energy Calculations Reveal the Binding Mechanism of A2A Adenosine Receptor Antagonists. *Angew. Chemie* **2020**, ange.202003788.
118. Provasi, D.; Bortolato, A.; Filizola, M. Exploring Molecular Mechanisms of Ligand Recognition by Opioid Receptors with Metadynamics. *Biochemistry* **2009**, *48*, 10020.

119. Shang, Y.; Yeatman, H.R.; Provasi, D.; Alt, A.; Christopoulos, A.; Canals, M.; Filizola, M. Proposed Mode of Binding and Action of Positive Allosteric Modulators at Opioid Receptors. *ACS Chem. Biol.* **2016**, *11*, 1220–1229.
120. Crowley, R.S.; Riley, A.P.; Sherwood, A.M.; Groer, C.E.; Shivaperumal, N.; Biscaia, M.; Paton, K.; Schneider, S.; Provasi, D.; Kivell, B.M.; et al. Synthetic Studies of Neoclerodane Diterpenes from Salvia divinorum: Identification of a Potent and Centrally Acting μ Opioid Analgesic with Reduced Abuse Liability. *J. Med. Chem.* **2016**, *59*, 11027–11038.
121. Cavalli, A.; Spitaleri, A.; Saladino, G.; Gervasio, F.L. Investigating Drug–Target Association and Dissociation Mechanisms Using Metadynamics-Based Algorithms. *Acc. Chem. Res.* **2015**, *48*, 277–285.
122. Saleh, N.; Ibrahim, P.; Saladino, G.; Gervasio, F.L.; Clark, T. An Efficient Metadynamics-Based Protocol to Model the Binding Affinity and the Transition State Ensemble of G-Protein-Coupled Receptor Ligands. *J. Chem. Inf. Model.* **2017**, *57*, 5.
123. Sánchez-Melgar, A.; Albasanz, J.L.; Guixà-González, R.; Saleh, N.; Selent, J.; Martín, M. The antioxidant resveratrol acts as a non-selective adenosine receptor agonist. *Free Radic. Biol. Med.* **2019**, *135*, 261–273.
124. Vilums, M.; Zweemer, A.J.M.; Yu, Z.; De Vries, H.; Hillger, J.M.; Wapenaar, H.; Bollen, I.A.E.; Barmare, F.; Gross, R.; Clemens, J.; et al. Structure-kinetic relationships - An overlooked parameter in hit-to-lead optimization: A case of cyclopentylamines as chemokine receptor 2 antagonists. *J. Med. Chem.* **2013**, *56*, 7706–7714.
125. Casarosa, P.; Bouyssou, T.; Germeyer, S.; Schnapp, A.; Gantner, F.; Pieper, M. Preclinical evaluation of long-acting muscarinic antagonists: Comparison of tiotropium and investigational drugs. *J. Pharmacol. Exp. Ther.* **2009**, *330*, 660–668.
126. Tautermann, C.S.; Kiechle, T.; Seeliger, D.; Diehl, S.; Wex, E.; Banholzer, R.; Gantner, F.; Pieper, M.P.; Casarosa, P. Molecular basis for the long duration of action and kinetic selectivity of tiotropium for the muscarinic M3 receptor. *J. Med. Chem.* **2013**, *56*, 8746–8756.
127. Strasser, A.; Wittmann, H.-J.J.; Seifert, R. Binding Kinetics and Pathways of Ligands to GPCRs. *Trends Pharmacol. Sci.* **2017**, *38*, 717–732.
128. Bortolato, A.; Deflorian, F.; Weiss, D.R.; Mason, J.S. Decoding the Role of Water Dynamics in Ligand−Protein Unbinding: CRF 1 R as a Test Case. *J. Chem. Inf. Model.* **2015**, *55*.
129. Mason, J.S.; Bortolato, A.; Weiss, D.R.; Deflorian, F.; Tehan, B.; Marshall, F.H. High end GPCR design: Crafted ligand design and druggability analysis using protein structure, lipophilic hotspots and explicit water networks. *Silico Pharmacol.* **2013**, *1*, 23.
130. MacKerell, A.D.; Bashford, D.; Bellott, M.; Dunbrack, R.L.; Evanseck, J.D.; Field, M.J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
131. Cornell, W.D.; Cieplak, P.; Bayly, C.I.; Gould, I.R.; Merz, K.M.; Ferguson, D.M.; Spellmeyer, D.C.; Fox, T.; Caldwell, J.W.; Kollman, P.A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. SOC* **1995**, *117*, 5179–5197.

132. McGibbon, R.T.T.; Beauchamp, K.A.A.; Harrigan, M.P.P.; Klein, C.; Swails, J.M.M.; Hernández, C.X.X.; Schwantes, C.R.R.; Wang, L.-P.; Lane, T.J.J.; Pande, V.S.S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532.
133. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 27–28.
134. Berendsen, H.J.C.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–56.
135. Brooks, B.R.; Brooks, C.L.; Mackerell, A.D.; Nilsson, L.; Petrella, R.J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
136. Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G.A.; Banáš, P.; Barducci, A.; Bernetti, M.; Bolhuis, P.G.; Bottaro, S.; Branduardi, D.; et al. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **2019**, *16*, 670–673.
137. Ozcan, O.; Uyar, A.; Doruker, P.; Akten, E.D. Effect of intracellular loop 3 on intrinsic dynamics of human β2-adrenergic receptor. *BMC Struct. Biol.* **2013**, *13*, 29.
138. Semack, A.; Sandhu, M.; Malik, R.U.; Vaidehi, N.; Sivaramakrishnan, S. Structural elements in the Gαs and Gβq C termini that mediate selective G Protein-coupled Receptor (GPCR) signaling. *J. Biol. Chem.* **2016**, *291*, 17929–17940.
139. Davoudmanesh, S.; Mosaabadi, J.M. Investigation of the effect of homocysteinylation of substance P on its binding to the NK1 receptor using molecular dynamics simulation. *J. Mol. Model.* **2018**, *24*, 1–15.
140. Dror, R.O.; Arlow, D.H.; Borhani, D.W.; Jensen, M.O.; Piana, S.; Shaw, D.E. Identification of two distinct inactive conformations of the 2-adrenergic receptor reconciles structural and biochemical observations. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 4689–4694.
141. GetContacts. Available online: https://getcontacts.github.io/ (accessed on 22 June 2020).
142. Ballesteros, J.A.; Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* **1995**, *25*, 366–428.
143. Gautier, C.; Laursen, L.; Jemth, P.; Gianni, S. Seeking allosteric networks in PDZ domains. *Protein Eng. Des. Sel.* **2018**, *31*, 367–373.
144. Bowerman, S.; Wereszczynski, J. Detecting Allosteric Networks Using Molecular Dynamics Simulation. *Methods Enzymol.* **2016**, *578*, 429–447.
145. Meral, D.; Provasi, D.; Filizola, M. An efficient strategy to estimate thermodynamics and kinetics of G protein-coupled receptor activation using metadynamics and maximum caliber. *J. Chem. Phys.* **2018**, *149*, 224101.
146. Chopra, N.; Wales, T.E.; Joseph, R.E.; Boyken, S.E.; Engen, J.R.; Jernigan, R.L.; Andreotti, A.H. Dynamic Allostery Mediated by a

Conserved Tryptophan in the Tec Family Kinases. *PLOS Comput. Biol.* **2016**, *12*, e1004826.
147. Wang, W.; Jiang, C.; Zhang, J.; Ye, W.; Luo, R.; Chen, H.F. Dynamics Correlation Network for Allosteric Switching of PreQ 1 Riboswitch. *Sci. Rep.* **2016**, *6*, 1–10.
148. LeVine, M.V.; Weinstein, H. NbIT - A New Information Theory-Based Analysis of Allosteric Mechanisms Reveals Residues that Underlie Function in the Leucine Transporter LeuT. *PLoS Comput. Biol.* **2014**, *10*, e1003603.
149. Kumawat, A.; Chakrabarty, S. Hidden electrostatic basis of dynamic allostery in a PDZ domain. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E5825–E5834.
150. Miao, Y.; Nichols, S.E.; Gasper, P.M.; Metzger, V.T.; McCammon, J.A. Activation and dynamic network of the M2 muscarinic receptor. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 10982–10987.
151. Bhattacharya, S.; Vaidehi, N. Differences in allosteric communication pipelines in the inactive and active states of a GPCR. *Biophys. J.* **2014**, *107*, 422–434.
152. Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M.P.; Dror, R.O.; Shaw, D.E. Systematic validation of protein force fields against experimental data. *PLoS ONE* **2012**, *7*, e32131.
153. Hospital, A.; Goñi, J.R.; Orozco, M.; Gelpí, J.L. Molecular dynamics simulations: Advances and applications. *Adv. Appl. Bioinform. Chem.* **2015**, *8*, 37–47.
154. Bernardi, R.C.; Melo, M.C.R.; Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta - Gen. Subj.* **2015**, *1850*, 872–877.
155. Marrink, S.J.; Tieleman, D.P. Perspective on the martini model. *Chem. Soc. Rev.* **2013**, *42*, 6801–6822.
156. Srivastava, A.; Tiwari, S.P.; Miyashita, O.; Tama, F. Integrative/Hybrid Modeling Approaches for Studying Biomolecules. *J. Mol. Biol.* **2020**, *432*, 2846–2860.
157. 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. *Chem. Rev.* **2019**, *119*, 6184–6226.
158. Bottaro, S.; Lindorff-Larsen, K. Biophysical experiments and biomolecular simulations: A perfect match? *Science* **2018**, *361*, 355–360.
159. Jung, J.; Nishima, W.; Daniels, M.; Bascom, G.; Kobayashi, C.; Adedoyin, A.; Wall, M.; Lappala, A.; Phillips, D.; Fischer, W.; et al. Scaling molecular dynamics beyond 100,000 processor cores for large-scale biophysical simulations. *J. Comput. Chem.* **2019**, *40*, 1919–1930.
160. Lagardère, L.; Jolly, L.H.; Lipparini, F.; Aviat, F.; Stamm, B.; Jing, Z.F.; Harger, M.; Torabifard, H.; Cisneros, G.A.; Schnieders, M.J.; et al. Tinker-HP: A massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chem. Sci.* **2018**, *9*, 956–972.
161. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindah, E.; Lindahl, E. Gromacs: High performance molecular simulations

through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.
162. Phillips, J.C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
163. Shaw, D.E.; Deneroff, M.M.; Dror, R.O.; Kuskin, J.S.; Larson, R.H.; Salmon, J.K.; Young, C.; Batson, B.; Bowers, K.J.; Chao, J.C.; et al. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **2008**, *51*, 91–97.
164. Shaw, D.E.; Grossman, J.P.; Bank, J.A.; Batson, B.; Butts, J.A.; Chao, J.C.; Deneroff, M.M.; Dror, R.O.; Even, A.; Fenton, C.H.; et al. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC, New Orleans, LA, USA, 16—21 November 2014; IEEE Computer Society: Washington, DC, USA, 2014; Volume 2015, pp. 41–53.
165. Vendruscolo, M.; Dobson, C.M. Protein Dynamics: Moore's Law in Molecular Biology. *Curr. Biol.* **2011**, *21*, R68–R70.
166. Martínez-Rosell, G.; Giorgino, T.; Harvey, M.J.; de Fabritiis, G. Drug Discovery and Molecular Dynamics: Methods, Applications and Perspective Beyond the Second Timescale. *Curr. Top. Med. Chem.* **2017**, *17*, 2617–2625.
167. Torrens-Fontanals, M.; Stepniewski, T.M.; Rodríguez-Espigares, I.; Selent, J. Application of Biomolecular Simulations to G Protein-Coupled Receptors (GPCRs). In *Biomolecular Simulations in Structure-Based Drug Discovery*; Gervasio, F.L., Spiwok, V., Eds.; Wiley: Hoboken, NJ, USA, 2018; pp. 205–223. ISBN 9783527806836.
168. Miao, Y.; McCammon, J.A. Mechanism of the G-protein mimetic nanobody binding to a muscarinic G-protein-coupled receptor. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 3036–3041.
169. Elofsson, A.; Hess, B.; Lindahl, E.; Onufriev, A.; van der Spoel, D.; Wallqvist, A. Ten simple rules on how to create open access and reproducible molecular simulations of biological systems. *PLOS Comput. Biol.* **2019**, *15*, e1006649.
170. Loeffler, H.H.; Bosisio, S.; Duarte Ramos Matos, G.; Suh, D.; Roux, B.; Mobley, D.L.; Michel, J. Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages. *J. Chem. Theory Comput.* **2018**, *14*, 5567–5582.
171. Abraham, M.J.; Apostolov, R.P.; Barnoud, J.; Bauer, P.; Blau, C.; Bonvin, A.M.J.J.; Chavent, M.; Chodera, J.D.; Čondić-Jurkić, K.; Delemotte, L.; et al. Sharing Data from Molecular Simulations. *J. Chem. Inf. Model.* **2019**, *59*, 4093–4099.
172. Sommer, M.E.; Selent, J.; Carlsson, J.; De Graaf, C.; Gloriam, D.E.; Keseru, G.M.; Kosloff, M.; Mordalski, S.; Rizk, A.; Rosenkilde, M.M.; et al. The European Research Network on Signal Transduction (ERNEST): Toward a Multidimensional Holistic Understanding of G Protein-Coupled Receptor Signaling. *ACS Pharmacol. Transl. Sci.* **2020**, *3*, 361–370.

173. Tiemann, J.K.S.; Guixà-González, R.; Hildebrand, P.W.; Rose, A.S. MDsrv: Viewing and sharing molecular dynamics simulations on the web. *Nat. Methods* **2017**, *14*, 1123–1124.
174. Carrillo-Tripp, M.; Alvarez-Rivera, L.; Lara-Ramírez, O.I.; Becerra-Toledo, F.J.; Vega-Ramírez, A.; Quijas-Valades, E.; González-Zavala, E.; González-Vázquez, J.C.; García-Vieyra, J.; Santoyo-Rivera, N.B.; et al. HTMoL: Full-stack solution for remote access, visualization, and analysis of molecular dynamics trajectory data. *J. Comput. Aided. Mol. Des.* **2018**, *32*, 869–876.
175. Bekker, G.J.; Nakamura, H.; Kinjo, A.R. Molmil: A molecular viewer for the PDB and beyond. *J. Cheminform.* **2016**, *8*, 42.
176. Mol*. Available online: https://molstar.org/ (accessed on 22 June 2020).
177. Botan, A.; Favela-Rosales, F.; Fuchs, P.F.J.; Javanainen, M.; Kanduč, M.; Kulig, W.; Lamberg, A.; Loison, C.; Lyubartsev, A.; Miettinen, M.S.; et al. Toward Atomistic Resolution Structure of Phosphatidylcholine Headgroup and Glycerol Backbone at Different Ambient Conditions. *J. Phys. Chem. B* **2015**, *119*, 15075–15088.
178. Domański, J.; Stansfeld, P.J.; Sansom, M.S.P.; Beckstein, O. Lipidbook: A public repository for force-field parameters used in membrane simulations. *J. Membr. Biol.* **2010**, *236*, 255–258.
179. Doma Nski, J.; Beckstein, O.; Iorga, B.I. Ligandbook: An online repository for small and drug-like molecule force field parameters. *Bioinformatics* **2017**, *33*, 1747–1749.
180. Meyer, T.; D'Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Pérez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; et al. MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories. *Structure* **2010**, *18*, 1399–1409.
181. Hospital, A.; Andrio, P.; Cugnasco, C.; Codo, L.; Becerra, Y.; Dans, P.D.; Battistini, F.; Torres, J.; Gõni, R.; Orozco, M.; et al. BIGNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.* **2016**, *44*, D272–D278.
182. Santos, R.; Ursu, O.; Gaulton, A.; Bento, A.P.; Donadi, R.S.; Bologa, C.G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T.I.; et al. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **2017**, *16*, 19–34.
183. Moore, G.E. Cramming More Components onto Integrated Circuits. *Electronics* **1965**, *38*, 114–117.
184. Borhani, D.W.; Shaw, D.E. The future of molecular dynamics simulations in drug discovery. *J. Comput. Aided. Mol. Des.* **2012**, *26*, 15–26.
185. Johnston, J.M.; Wang, H.; Provasi, D.; Filizola, M. Assessing the Relative Stability of Dimer Interfaces in G Protein-Coupled Receptors. *PLoS Comput. Biol.* **2012**, *8*, e1002649.
186. Periole, X.; Knepp, A.M.; Sakmar, T.P.; Marrink, S.J.; Huber, T. Structural determinants of the supramolecular organization of G protein-coupled receptors in bilayers. *J. Am. Chem. Soc.* **2012**, *134*, 10959–10965.
187. Dror, R.O.; Jensen, M.Ø.; Borhani, D.W.; Shaw, D.E. Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. *J. Gen. Physiol.* **2010**, *135*, 555–562.

188. Adcock, S.A.; McCammon, J.A. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem. Rev.* **2006**, *106*, 1589–1615.

3.3. Structural dynamics bridge the gap between the genetic and functional levels of GPCRs

The specific functionality of different GPCR types or subtypes is determined by their genetic sequence. Even within a single GPCR, we can find variability in the genetic sequence due to natural genetic variation, which can result in specific disease phenotypes or altered drug responses. We can also find variability in the protein sequence due to isoforms. Deciphering the molecular link between sequence diversity and its functional consequences is challenging but critical for the full comprehension of the signaling properties of GPCRs. In this review, we explore how genetic or protein sequence translates into structure, how this impacts the structural motions of the protein, and, finally, how all these factors determine the receptor functionality. For that, we revise the available online resources and state-of-the-art computational approaches that help to address these questions by providing information on GPCRs at different biological levels.

Torrens-Fontanals, M., Stepniewski, T. M., Gloriam, D. E. & Selent, J. Structural dynamics bridge the gap between the genetic and functional levels of GPCRs. *Curr. Opin. Struct. Biol*. 69, 150–159 (2021). doi:10.1016/J.SBI.2021.04.005

Structural dynamics bridge the gap between the genetic and functional levels of GPCRs

MarionaTorrens-Fontanals[1], Tomasz M. Stepniewski[1,2,3,] David E.Gloriam[4], and JanaSelent[1,*]

1. Research Programme on Biomedical Informatics (GRIB), Hospital Del Mar Medical Research Institute (IMIM)—Department of Experimental and Health Sciences, Pompeu Fabra University (UPF), 08003 Barcelona, Spain
2. InterAx Biotech AG, PARK InnovAARE, 5234 Villigen, Switzerland
3. Faculty of Chemistry, Biological and Chemical Research Centre, University of Warsaw, 02-093 Warsaw, Poland
4. Department of Drug Design and Pharmacology, University of Copenhagen, 2100 Copenhagen, Denmark

* Corresponding author

Abstract

G protein–coupled receptors (GPCRs) are implicated in nearly all physiological processes in the human body and represent an important drug targeting class. The genes encoding the different GPCR (sub)types determine their specific functionality, which can be altered by natural genetic variants and isoforms. Deciphering the molecular link between sequence diversity and its functional consequences is a current challenge and critical for the comprehension of the physiological response of GPCRs. It requires a global understanding of how protein sequence translates into protein structure, how this impacts the structural motions of the protein, and, finally, how all these factors determine the receptor functionality. Here, we discuss available resources and state-of-the-art computational approaches to address this question.

Keywords: Receptor signaling; GPCRs; structural dynamics; web resources; computational biology

1. Introduction

G protein–coupled receptors (GPCRs) form a large and versatile family of membrane proteins. Because of their versatility, GPCRs

are involved in nearly all physiological processes and have become a major target for the pharmaceutical industry [1]. GPCRs are complex biological microprocessors that have evolved to recognize specific types of exogenous and endogenous stimuli (light, pressure, odorants, neurotransmitters, hormones, etc.) with a ligand and signaling specificity that is genetically encoded. Based on this, the human GPCRs are commonly divided into families (or classes) [2,3]: rhodopsin (A), secretin (B1), glutamate (C), adhesion (B2), frizzled (F), and taste 2. Apart from recognizing different signaling stimuli, the genetic variability drives other global receptor properties such as receptor dynamics, which is tightly linked to established intramolecular networks of direct (covalent and noncovalent) and indirect (water-/ion-mediated) interactions. Protein dynamics is a key feature that enables a receptor to sample functionally relevant receptor states (e.g. inactive state, G_s coupling state, G_q coupling state, …). This typically occurs to some extent even in the absence of an extracellular stimulus, determining its constitutive (basal) activity. Genetic variability also modulates coupling specificity/promiscuity to intracellular signaling proteins—a phenomenon linked to functional selectivity or signaling bias [4]. Another important functional consequence is the impact on the allosteric modulation of GPCRs by the membrane environment such as membrane lipids [5, 6, 7] or other membrane proteins (GPCR dimerization/oligomerization) [5,8].

Receptor function can be affected by genetic polymorphisms resulting in specific disease phenotypes or altered drug response [9]. However, the link between the genetic and protein sequence level and receptor function is complex and often poorly understood. Tackling this challenge requires a global understanding of how protein sequence translates into structural dynamics and ultimately into a functional response (Figure 1). Here, we outline and discuss current resources and state-of-the-art computational approaches (Table 1) to address this question.

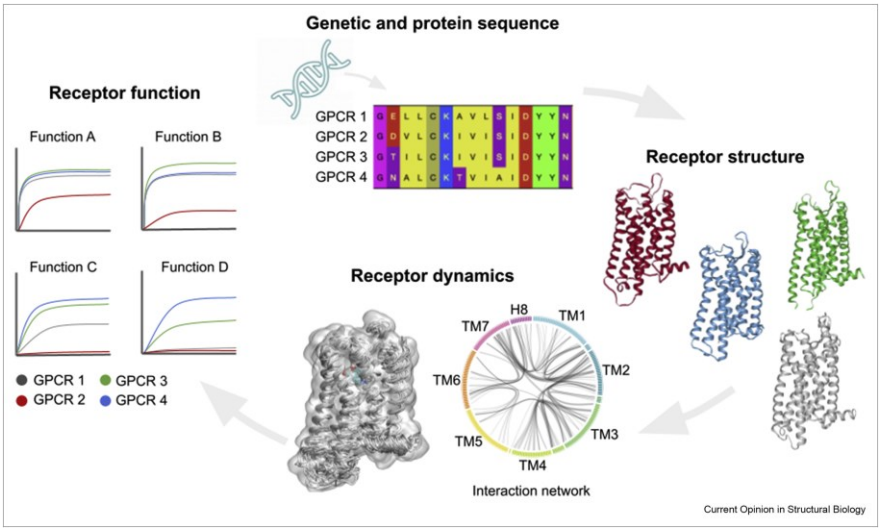


Figure 1. Different levels of GPCR biology. Bridging the gap between the genetic and protein sequence level and receptor function using structural dynamics.

Table 1. Selection of platforms and tools for the study of GPCRs at different biological levels.

Name	Focus	Main biological level	GPCR specific	Refs	Links
GnomAD	Aggregation of exome and genome sequencing data	Genetic sequence	No	[10]	https://gnomad.broadinstitute.org/
1000 Genomes Project	Genetic variation from whole-genome sequencing	Genetic sequence	No	[11]	https://www.internationalgenome.org/
DisGeNET	Genes and variants associated with diseases	Genetic sequence	No	[12]	https://www.disgenet.org/
UniProt	Protein sequences and annotations	Protein sequence	No	[13]	https://www.uniprot.org/
PDB	Structural data of biological macromolecules	Structure	No	[14]	https://www.rcsb.org/
GPCRdb	Reference data, interactive	Protein	Yes	[15]	https://gpcrdb.

	visualization, and experiment design tools for GPCRs	sequence Structure Function			org/
GPCRmd	MD simulations of GPCRs with interactive analysis and visualization	Structural dynamics	Yes	[16]	http://gpcrmd.org/
MemProtMD	MD simulations of membrane-embedded proteins	Structural dynamics	No	[17]	http://memprotmd.bioch.ox.ac.uk/
CHARMM-GUI	Built of molecular systems and preparation of inputs for simulation	Structural dynamics	No	[18]	http://www.charmm-gui.org/
GPCR-ModSim	Structural modeling and MD simulation of GPCRs	Structural dynamics	Yes	[19]	http://gpcr-modsim.org/
MERMAID	Preparation and run of coarse-grained MD simulations	Structural dynamics	No	[20]	https://molsim.sci.univr.it/mermaid/main.php
PACKmol	Built of initial configurations for MD simulations	Structural dynamics	No	[21]	http://m3g.iqm.unicamp.br/packmol/
Hybrid MM/CG Webserver	Setup of molecular mechanics/coarse-grained simulations for human GPCR/ligand complexes	Structural dynamics	Yes	[22]	https://mmcg.grs.kfa-juelich.de/
IUPHAR/BPS Guide to Pharmacology	Ligand-activity-target relationships	Functional	No	[23]	https://www.guidetopharmacology.org/
PRECOG	Prediction of coupling probability of GPCRs to	Functional	Yes	[24]	http://precog.russelllab.org/

	individual G proteins				
BiasDB	Biased GPCR ligands	Functional	Yes	[25]	https://biasdb.drug-design.de/
FigShare	Scientific general-purpose data repository	General	No	-	https://figshare.com/
Zenodo	Scientific general-purpose data repository	General	No	-	https://zenodo.org/

2. Genetic and protein sequence variability of GPCRs

2.1. General considerations

GPCRs constitute the largest family of human cell surface receptors, with more than 800 known members [26]. Despite the fact that GPCRs share a common architecture, consisting of seven transmembrane (TM) helices bridged by three loops at each side of the membrane [27], they are characterized by a relatively low overall sequence identity across classes. A small fraction of highly conserved regions is primarily related to structural motifs that are relevant for the overall receptor function. In contrast, the sequence differences between GPCR (sub)types account for differences in receptor structure, dynamics, and functionality. The primary resource for sequence information is UniProt, which provides free access to protein sequences and functional information [13]. Specifically for GPCRs, structure-based alignments are a common approach to determine sequence variability between receptor (sub)types. Curated alignments are available in GPCRdb [15], which includes additional tools for performing similarity searches and creating similarity matrices and phylogenetic trees (Figure 2a and b).

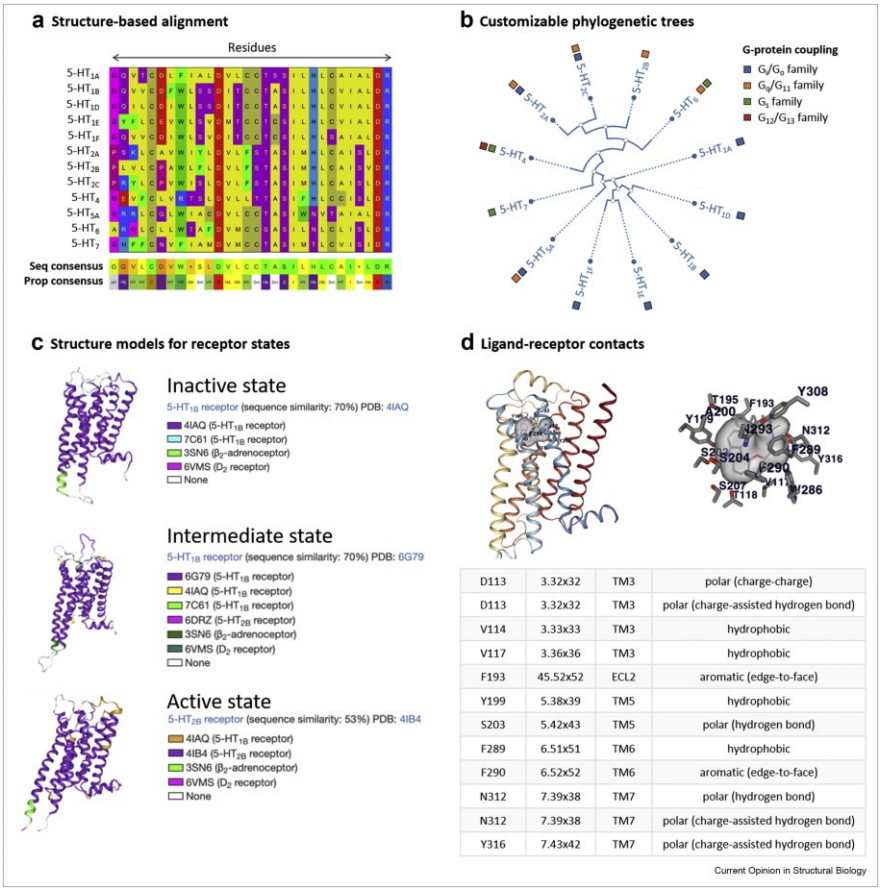


Figure 2. Sequence and structural information for GPCRs as implemented in GPCRdb. **(a)** Structure-based alignment, **(b)** phylogenetic trees, **(c)** structural models for different receptor states, and **(d)** ligand–receptor interactions implemented in GPCRdb.

2.2. Sources for sequence variations

Even within a single GPCR, we can find variability in the protein sequence because of isoforms. Protein isoforms are originated from the same gene but have distinct amino acid sequences. They are generated by tissue-specific alternative splicing and via alternative transcription start and termination sites. Thus, a GPCR gene can diversify into several isoforms with different patterns of expression across different tissues. Isoforms may have distinct structures and signaling properties, influencing tissue-specific physiology and drug response. In fact, a recent study shows how GPCRs diversify

in key structural segments that determine receptor function [28]. It also describes how different combinations of isoform expressions can contribute to system-specific signaling, highlighting the need to consider isoform composition to successfully understand physiological responses.

Another source of differences in GPCR functionality is genetic variants. Natural human genetic variants may cause the GPCRs of different individuals to respond differently to the same stimuli (e.g. natural ligands and drugs). Large-scale reference data sets of human genetic variation are critical for the functional interpretation of such genetic variations. One of the most widely known databases of human genetic variation is The Genome Aggregation Database (gnomAD) [10], which includes exome and genome sequencing data from a variety of large-scale sequencing projects. Another important resource is the 1000 Genomes Project, which provides the largest public catalog of genotype data [11].

2.3. Altered drug response and disease association

Based on data from resources such as gnomAD and the 1000 Genomes Project, a recent study structurally mapped variant information for GPCRs that are targeted by Food and Drug Administration–approved drugs and discussed its implication for drug response [9]. They reported that individual differences in responses to medications are an underappreciated burden on public health. In addition to altered drug response, variants can be also associated with disease phenotypes (Table 2). An excellent resource for tracking disease-associated variants is DisGeNET [12]. It represents one of the largest publicly available collections of genes and variants associated with human diseases, as it integrates data from the most popular repositories in the field, as well as information extracted from the scientific literature by text mining.

Table 2. Translation of sequence variability into an altered GPCR functionality.

Receptor	Variant	Location	Impact	Clinical relevance	Reference
D_2 receptor	Ser311Cys	IL	Decreased agonist binding affinity	Schizophrenia and other mental disorders	[29]

V_2 receptor	Arg113Trp	TM	Increases Kd for the agonist Arg-vasopressin	Nephrogenic Diabetes Insipidus, Type I	[30]
	Tyr128Ser	TM	Abolished agonist binding		
ET_B receptor	Trp276Cys	TM	Altered receptor coupling to G_q	Hirschsprung disease	[31]
Rhodopsin	Arg135Gly	TM	Defect to mediate G protein release	Retinitis pigmentosa	[32]

IL, intracellular loop; TM, transmembrane domain.

3. Translation of sequence variability into three-dimensional structures

3.1. High-resolution structural data for GPCRs

Thanks to recent advances in protein engineering, X-ray crystallography, and cryo-electron microscopy (cryo-EM), we have experienced an exponential increase in the number of available GPCR structures (3D-GPCRome). The obtained structures have provided opportunities to understand how sequence translates into a specific structural architecture of inactive but also active receptor states. To date, high-resolution structures of about 90 unique GPCRs are available in at least one functional state (see http://gpcrdb.org).

The inactive state is typically the resting/default and most stable conformation of GPCRs. It was first structurally characterized for rhodopsin [33] and β_2-adrenoceptor [34], which revealed the basic 7TM architecture of GPCRs. The active receptor state represents a high energetic state and is only accessible in a G protein–bound state by current experimental approaches. A first breakthrough has been the crystal structure of the active β_2-adrenoceptor in complex with the heterotrimeric G protein, G_s [35]. The largest difference to the inactive state is a receptor opening near the cytoplasmic side that accommodates the C-terminal helix of the G protein α subunit. Since then, numerous active structures coupled to G_s, G_i, and G_o have been obtained, thanks to major advances in cryo-EM [36]. In addition to G proteins, GPCRs also bind arrestins, firstly observed for rhodopsin [37]. This event primarily desensitizes the receptor but can also modulate downstream signaling pathways [38].

3.2. Web resources for GPCR structures

The general source for structural data of proteins is the widely known Protein Data Bank (PDB) [14]. The PDB provides three-dimensional data of large molecules typically obtained by X-ray crystallography, nuclear magnetic resonance spectroscopy, or, increasingly, cryo-EM. A reference database for GPCR structure and coupling partners is GPCRdb [15], which not only collects but also curates experimental structures by remodeling missing/distorted regions and reverting mutations. A special highlight of GPCRdb is structure models of inactive, intermediate, and active states for all human nonolfactory receptors (Figure 2c). In addition, GPCRdb exposes a wide range of tools for structural analysis including ligand–receptor contacts (Figure 2d). An important asset is the integration and cross-reference of functional data (e.g. G protein coupling, mutagenesis data, variant information, etc.), which has converted GPCRdb into a highly used and cited platform over the last decade.

3.3. Sequence–structure relationships

An intriguing discovery is related to the recent observation that sequence variability can induce different architectures of GPCR–G protein/arrestin complexes with functional consequences. For instance, the solved GPR97–G_o complex presents a so far unknown GPCR–G protein architecture which is not observed in other solved GPCR complex structures [39]. This architecture is the result of a post-translational modification of the G_o protein and seems to be linked to G_o protein specificity. In addition, substantial differences in arrestin binding have been observed when comparing rhodopsin [37,40] and the NTS_1 receptor [41]. Combined computational and biochemical studies suggest that the structural determinants of arrestin binding are driven by the phosphorylation pattern at the receptor C-tail and have important implications for the functional response [38]. Obviously, the C-tail phosphorylation pattern depends on the protein sequence and the presence of residues (e.g. Ser, Tyr) that can be phosphorylated by G protein-coupled receptor kinases (GRKs) [40]. This is a structural property of each receptor type that translates into distinct receptor responses.

At times, the impact of sequence variability on the structure is very subtle, as seen for the serotonin subtype receptors 5-HT_{1B} and 5-HT_{2B} (5-HT_{1B}R and 5-HT_{2B}R) (Figure 3a, $RMSD_{backbone}$ 1.7 Å). This is surprising considering their differences in G protein selectivity (G_i versus G_q) and their functional selectivity on ergotamine binding (balanced versus β-arrestin bias) [42]. Mechanistically, small variability in the protein sequence is amplified at the level of receptor dynamics which translates into the favorable sampling of conformational receptor states that are linked to a specific receptor response.

4. Receptor dynamics governs GPCR function

4.1. The conformational landscape

GPCR function is largely determined by the flexibility and ability to sample a wide range of structural conformations. Within their conformational landscape, distinct populations are linked to a specific physiological response (e.g. inactive state, G_s coupling state, G_q coupling state, etc.). Because of this dynamic nature, a complete understanding of the structural background of GPCR signaling and pharmacology requires an in-depth knowledge of receptor dynamics [43]. Molecular dynamics (MD) simulations are a well-established technique to investigate time-resolved motions of biological macromolecules at atomic resolution, thus incorporating information on protein flexibility into experimentally solved structures [44,45].

4.2. Time-resolved fluctuations of specific receptor states

Applied to GPCRs, MD allows us to sample in detail the dynamic fluctuations of specific functional states [45]. This provides information on the stabilized intramolecular network that drives the functional response. Such intramolecular networks consist of direct (covalent and noncovalent) and indirect (water-/ion-mediated) interactions. In particular, water-mediated interactions are often poorly resolved in available high-resolution structures but accessible in time-resolved simulations [16,46]. In addition, the relevance of ion binding in the inactive receptor states has been extensively investigated in several studies using MD [16,47, 48, 49].

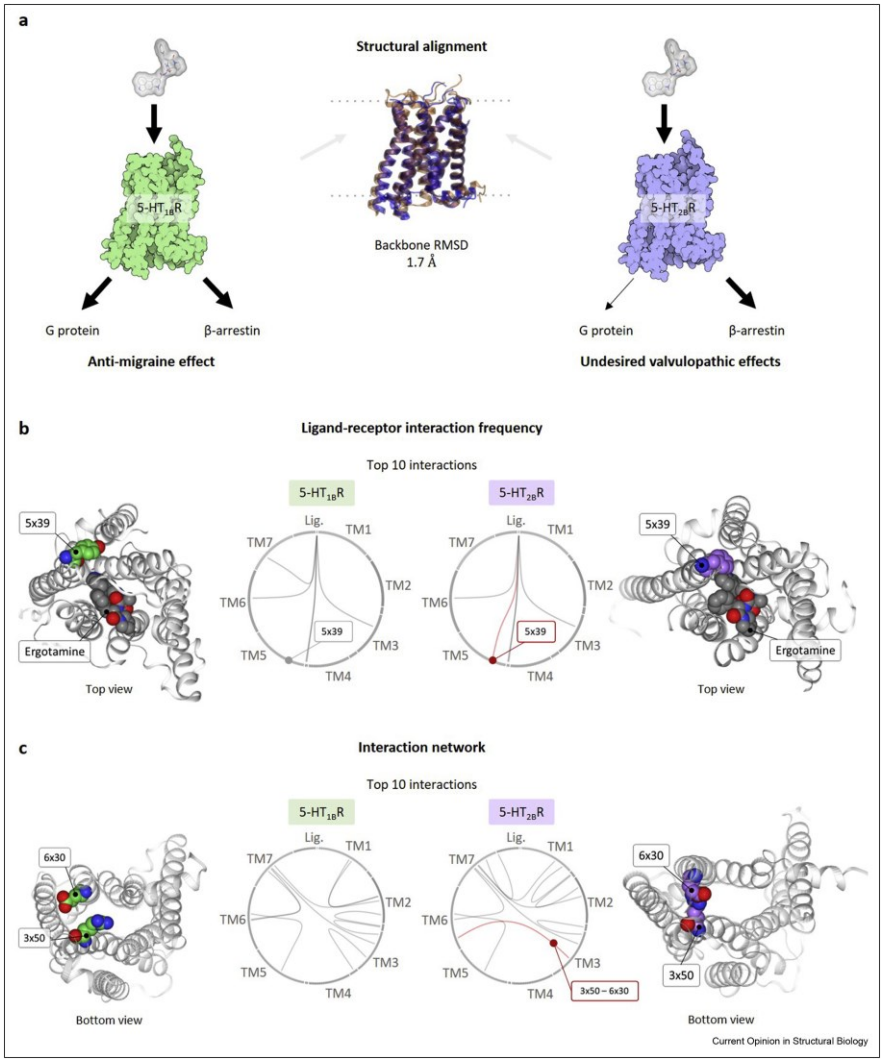


Figure 3. Receptor dynamics drive the functional outcome. 5-HT_{1B}R and 5-HT_{2B}R differ in their functional response on ergotamine binding. A structural alignment of the static structures shows minimal differences in the overall 7TM architecture (RMSD: 1.7 Å). Time-resolved simulations are able to capture differences in the stability of specific ergotamine–receptor interactions which translate into a similar G protein and β-arrestin coupling response (5-HT_{1B}R) or β-arrestin bias (5-HT_{2B}R).

4.3. Transitions between different receptor states

Beyond the study of individual states, MD allows approximations of the transition pathways between different states and the detection of

metastable states along the pathway. Recent highlights are insights into the activation pathway of the μ receptor (data link 1) as well as dynamic and kinetic elements underlying functional selectivity [50]. Authors applied an adaptive sampling regime and Markov state models combined with the information theory to build the transition pathways. Still, such an approach requires a total simulation time of ~240 μs to achieve convergence of the free-energy landscape, which is not achievable in routine simulation work. Enhanced sampling approaches such as metadynamics combined with the principle of maximum caliber proved to be a more efficient strategy for estimating the thermodynamic and kinetic properties of GPCR activation at an affordable computational cost [51]. Another recent work used MD simulations in an optimized string-of-swarms framework to calculate free-energy landscapes of β_2-adrenoceptor activation [52]. This work revealed new insights into how molecular microswitches govern the equilibrium between conformational states.

4.4. Web resources for time-resolved simulation data

The generation, analysis, and visualization of MD simulations require specialized software and efficient storage resources, which traditionally limited the audience to which this technology is available. Fortunately, several initiatives are currently working to provide online resources for MD data sharing, approaching the valuable information of simulations to both computational and nonexpert scientists [53]. A very recent development is GPCRmd—the first platform that features interactive visualization of time-resolved data [16]. GPCRmd is a community-driven database and web platform that provides free access to a vast number of GPCR MD simulations. It also includes a comprehensive set of tools to easily analyze the protein motions and interactions formed during the simulation. Moreover, it provides the possibility to map experimental data on the structure, such as genetic variants, mutations, and X-ray crystallography density maps.

The strength of GPCRmd is that it allows monitoring how subtle differences in protein sequence translate into dynamics of ligand–receptor interaction, receptor flexibility, and function. For instance, the serotonin receptors 5-HT_{1B} (data link 2) and 5-HT_{2B} (data link 3) expose differences in functional selectivity

on ergotamine binding. By comparing the time-resolved ligand interaction pattern of the two receptors, we can pick up important differences that are not evident from the static structure (Figure 3b). In particular, highly frequent ergotamine interactions with TM5 are found in the 5-HT_{2B}R but lack in the 5-HT_{1B}R. As previously described by Martí-Solano et al. [54], this may cause a difference in TM5 and TM6 stabilization and the conformational sampling of diverse functional receptor states. In fact, this is evident when monitoring the most frequent contacts in the entire intramolecular networks of sampled receptor conformations (Figure 3c). A functionally relevant contact between residues 3×50 and 6×30 (ionic lock) is frequently formed in 5-HT_{2B}R, thus reducing the opening of the G protein binding site. This goes along with the experimental observation that the 5-HT_{2B}R shows decreased ability to engage the G protein [42].

Apart from GPCRmd, another useful database for the study of GPCR dynamics is MemProtMD [17]. It incorporates coarse-grained simulations of membrane-embedded protein structures, describing protein–lipid interactions. The simulation setup is performed using their automated pipeline, which allows the lipid bilayer assembly around membrane protein structures released from the PDB [55]. The obtained simulations and the results of subsequent analysis are freely accessible online, including interactive 3D visualizations of the assembled bilayer and 2D visualizations of lipid contact data and membrane protein topology. In addition, they provide ensemble analyses to describe conserved lipid interaction information across proteins, families, and for the entire database of PDB entries. It is also worth mentioning the general-purpose data repositories such as FigShare (https://figshare.com/) or Zenodo (http://zenodo.org/), which contain large collections of heterogeneous MD data. However, in general, such resources do not aim to integrate, harmonize, validate, or standardize the deposited data, which makes it complicated to filter suitable data from the huge variety of deposited files.

Finally, there exist several web tools to guide and facilitate the setup of realistic GPCR/membrane simulations (CHARMM-GUI [18], GPCR-ModSim [19], MERMAID [20], PACKmol [21], etc.). A very recent development is the hybrid molecular mechanics/coarse-grained (MM/CG) Webserver, which automates

the setup of MM/CG simulations for human GPCR/ligand complexes [22]. Such hybrid MM/CG setup allows for an improvement of simulation speed.

4.5. Resources for the GPCR functional outcome

As highlighted throughout this article, genetic and protein sequence variability combined with induced structural dynamics determines the final functional outcome. A long-standing, open-access, and expert-curated database for GPCR functionality is the IUPHAR/BPS Guide to Pharmacology [23]. Presently, it contains functional data for about 400 GPCRs compiled from a multitude of research studies including information about ligands, canonical G protein coupling properties, and downstream signaling. The value of this database arises from their approach to document selection, data curation, and annotation. Interestingly, information from the IUPHAR/BPS Guide to Pharmacology, among other sources, has recently been incorporated in GPCRdb to provide a browser of G protein coupling (https://gpcrdb.org/signprot/couplings). However, interpretation of data across different studies is often complicated because of large variability in experimental setups. Thanks to advances in high-throughput screening, functional data are rapidly increasing and available for large data sets (e.g. > 100 receptors) using consistent experimental protocols [56,57]. Systematic data sets have been exploited for creating the PRECOG web server, which allows predicting GPCR coupling probability of specific G proteins [24]. Another strong research focus is to systemize and harmonize information for so-called biased (ant)agonists (BiasDB [25] and GPCRdb). Although there is an imprinted signaling profile for each receptor, biased ligands are able to shift the signaling profile to specific pathway(s)—information of high value for drug development programs toward more efficient and safer drugs.

5. Outlook

Recent innovations in sequencing techniques, protein structure determination, and high-throughput signaling profiling have substantially improved our understanding of the complex signaling properties for GPCRs. Combined with computer simulation of time-resolved receptor motion, we have obtained a first glimpse into the

underlying molecular mechanisms that drive receptor functionality. Yet, a complete comprehension of the molecular GPCR machinery will require far more investigation. Structurally, we are just beginning to understand the mechanistic basis of the dynamics of engaged intramolecular networks that are linked to signaling bias or allosteric modulation. Many other questions remain completely unsolved and require the integration of information across multiple spatial and temporal scales. Tackling this challenge will bring us closer to link GPCR sequence, structure, and dynamics to cellular physiology, which is a critical milestone for the rational design of improved drugs.

Conflict of interest statement

Nothing declared.

Acknowledgements

M.T-F. acknowledges financial support from the Spanish Ministry of Science, Innovation and Universities (FPU16/01209). T.M.S. would like to acknowledge support from the National Center of Science, Poland (grant number 2017/27/N/NZ2/02571). D.E.G. acknowledges financial support from the Lundbeck foundation (R163-2013-16327) and Novo Nordisk Foundation (NNF18OC0031226). J.S. acknowledges financial support from the Instituto de Salud Carlos III FEDER (PI15/00460 and PI18/00094) and the ERA-NET NEURON & Ministry of Economy, Industry and Competitiveness (AC18/00030).

References

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest

1. A.S. Hauser, M.M. Attwood, M. Rask Andersen, H.B. Schiöth, D.E. Gloriam. Trends in GPCR drug discovery: new agents, targets and indications. Nat Rev Drug Discov, 16 (2017), pp. 829-842
2. R. Fredriksson, M.C. Lagerström, L.-G. Lundin, H.B. Schiöth. The G-protein-coupled receptors in the human genome form five main families.

Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol, 63 (2003), pp. 1256-1272

3. K.J.V. Nordström, M. Sällman Almén, M.M. Edstam, R. Fredriksson, H.B. Schiöth. Independent HHsearch, Needleman–Wunsch-based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families. Mol Biol Evol, 28 (2011), pp. 2471-2480
4. M. Martí-Solano, A. Iglesias, G. de Fabritiis, F. Sanz, J. Brea, M.I. Loza, M. Pastor, J. Selent. Detection of new biased agonists for the serotonin 5-HT2A receptor: modeling and experimental validation. Mol Pharmacol, 87 (2015), pp. 740-746
5. R. Guixà-González, M. Javanainen, M. Gómez-Soler, B. Cordobilla, J.C. Domingo, F. Sanz, M. Pastor, F. Ciruela, H. Martinez-Seara, J. Selent. Membrane omega-3 fatty acids modulate the oligomerisation kinetics of adenosine A2A and dopamine D2 receptors. Sci Rep, 6 (2016), p. 19839
6. Md Jafurulla, G. Aditya Kumar, B.D. Rao, A. Chattopadhyay A critical analysis of molecular mechanisms underlying membrane cholesterol sensitivity of GPCRs. Adv Exp Med Biol, 1115 (2019), pp. 21-52
7. P. Sarkar, S. Mozumder, A. Bej, S. Mukherjee, J. Sengupta, A. Chattopadhyay. Structure, dynamics and lipid interactions of serotonin receptors: excitements and challenges. Biophys Rev (2020), 10.1007/s12551-020-00772-8
8. D. Meral, D. Provasi, D. Prada-Gracia, J. Möller, K. Marino, M.J. Lohse, M. Filizola Molecular details of dimerization kinetics reveal negligible populations of transient μ-opioid receptor homodimers at physiological concentrations. Sci Rep, 8 (2018), p. 7705
9. A.S. Hauser, S. Chavali, I. Masuho, L.J. Jahn, K.A. Martemyanov, D.E. Gloriam, M.M. Babu. Pharmacogenomics of GPCR drug targets. Cell, 172 (2018), pp. 41-54. e19
10. K.J. Karczewski, L.C. Francioli, G. Tiao, B.B. Cummings, J. Alföldi, Q. Wang, R.L. Collins, K.M. Laricchia, A. Ganna, D.P. Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581 (2020), pp. 434-443
11. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature, 526 (2015), pp. 68-74
12. •J. Piñero, J.M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, L.I. Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res, 48 (2020), pp. D845-D855

    DisGeNET is a knowledge management platform that integrates and standardizes data about disease-associated genes and variants from multiple sources, including the scientific literature. In the latest release, it includes new sources of data, data attributes, and prioritization metrics, as well as a redesigned web interface and recently launched APIs.

13. The Uniprot Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res, 47 (2019), pp. D506-D515

14. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The protein Data Bank. Nucleic Acids Res, 28 (2000), pp. 235-242
15. •A.J. Kooistra, S. Mordalski, G. Pándy-Szekeres, M. Esguerra, A. Mamyrbekov, C. Munk, G.M. Keserű, D.E. Gloriam. GPCRdb in 2021: integrating GPCR sequence, structure and function. Nucleic Acids Res, 49 (2020), pp. D335-D343

    GPCRdb offers reference data, analysis tools, experiment design and dissemination of published datasets for GPCR research. Here, they present an updated GPCRdb resource with a particular focus on integration of sequence, structure, and function data at the receptor and residue levels.

16. •I. Rodríguez-Espigares, M. Torrens-Fontanals, J.K.S. Tiemann, D. Aranda-García, J.M. Ramírez-Anguita, T.M. Stepniewski, N. Worp, A. Varela-Rial, A. Morales-Pastor, B. Medel-Lacruz, et al. GPCRmd uncovers the dynamics of the 3D-GPCRome. Nat Methods, 17 (2020), pp. 777-787

    GPCRmd is an open access and community-driven research resource that fosters transparent and easy dissemination of GPCR MD simulations. It hosts simulations of most GPCR crystal structures solved to date, as well as web-based visualization capabilities and a comprehensive and user-friendly analysis toolbox.

17. T.D. Newport, M.S.P. Sansom, P.J. Stansfeld. The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. Nucleic Acids Res, 47 (2019), pp. D390-D397
18. S. Jo, T. Kim, V.G. Iyer, W. Im. CHARMM-GUI: a web-based graphical user interface for CHARMM. J Comput Chem, 29 (2008), pp. 1859-1865
19. M. Esguerra, A. Siretskiy, X. Bello, J. Sallander, H. Gutiérrez-de-Terán, GPCR-ModSim. A comprehensive web based solution for modeling G-protein coupled receptors. Nucleic Acids Res, 44 (2016), pp. W455-W462
20. M. Damre, A. Marchetto, A. Giorgetti. MERMAID: dedicated web server to prepare and run coarse-grained membrane protein dynamics. Nucleic Acids Res, 47 (2019), pp. W456-W461
21. L. Martínez, R. Andrade, E.G. Birgin, J.M. Martínez. PACKMOL: a package for building initial configurations for molecular dynamics simulations. J Comput Chem, 30 (2009), pp. 2157-2164
22. J. Schneider, R. Ribeiro, M. Alfonso-Prieto, P. Carloni, A. Giorgetti. Hybrid MM/CG webserver: automatic set up of molecular mechanics/coarse-grained simulations for human G protein-coupled receptor/ligand complexes. Front Mol Biosci, 7 (2020), p. 576689
23. J.F. Armstrong, E. Faccenda, S.D. Harding, A.J. Pawson, C. Southan, J.L. Sharman, B. Campo, D.R. Cavanagh, S.P.H. Alexander, A.P. Davenport, et al.. The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. Nucleic Acids Res, 48 (2020), pp. D1006-D1021

24. G. Singh, A. Inoue, J.S. Gutkind, R.B. Russell, F. Raimondi. PRECOG: PREdicting COupling probabilities of G-protein coupled receptors. Nucleic Acids Res, 47 (2019), pp. W395-W401
25. C. Omieczynski, T.N. Nguyen, D. Sribar, L. Deng, D. Stepanov, D. Schaller, G. Wolber, M. Bermudez. BiasDB: a comprehensive database for biased GPCR ligands. bioRxiv (2019), 10.1101/742643
26. J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al. The sequence of the human genome. Science, 291 (2001), pp. 1304-1351.
27. K.L. Pierce, R.T. Premont, R.J. Lefkowitz. Seven-transmembrane receptors. Nat Rev Mol Cell Biol, 3 (2002), pp. 639-650
28. •M. Marti-Solano, S.E. Crilly, D. Malinverni, C. Munk, M. Harris, A. Pearce, T. Quon, A.E. Mackenzie, X. Wang, J. Peng, et al. Combinatorial expression of GPCR isoforms affects signalling and drug responses. Nature, 587 (2020), pp. 650-656

Using a data science approach, this study highlights how a single GPCR gene can diversify into several isoforms with distinct signaling properties. It also shows how unique isoform combinations diversify signaling responses in human tissues. This highlights the need to consider isoform composition in order to successfully understand GPCR signaling and drug responses.

29. A. Cravchik, D.R. Sibley, P.V. Gejman. Functional analysis of the human D2 dopamine receptor missense variants. J Biol Chem, 271 (1996), pp. 26013-26017
30. Oksche, W. Rosenthal. The molecular basis of nephrogenic diabetes insipidus. J Mol Med, 76 (1998), pp. 326-337
31. E.G. Puffenberger, K. Hosoda, S.S. Washington, K. Nakao, D. deWit, M. Yanagisawa, A. Chakravarti. A missense mutation of the endothelin-B receptor gene in multigenic hirschsprung's disease. Cell, 79 (1994), pp. 1257-1266
32. S. Acharya, S.S. Karnik. Modulation of GDP release from transducin by the conserved Glu134-Arg135 sequence in rhodopsin. J Biol Chem, 271 (1996), pp. 25406-25411
33. K. Palczewski, T. Kumasaka, T. Hori, C.A. Behnke, H. Motoshima, B.A. Fox, I. Le Trong, D.C. Teller, T. Okada, R.E. Stenkamp, et al. Crystal structure of rhodopsin: a G protein-coupled receptor. Science, 289 (2000), pp. 739-745
34. V. Cherezov, D.M. Rosenbaum, M.A. Hanson, S.G.F. Rasmussen, F.S. Thian, T.S. Kobilka, H.-J. Choi, P. Kuhn, W.I. Weis, B.K. Kobilka, et al. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science, 318 (2007), pp. 1258-1265
35. S.G.F. Rasmussen, B.T. DeVree, Y. Zou, A.C. Kruse, K.Y. Chung, T.S. Kobilka, F.S. Thian, P.S. Chae, E. Pardon, D. Calinski, et al. Crystal structure of the β 2 adrenergic receptor–Gs protein complex. Nature, 477 (2011), pp. 549-555
36. J. García-Nafría, C.G. Tate. Cryo-EM structures of GPCRs coupled to gs, Gi and Go. Mol Cell Endocrinol, 488 (2019), pp. 1-13

37. Y. Kang, X.E. Zhou, X. Gao, Y. He, W. Liu, A. Ishchenko, A. Barty, T.A. White, O. Yefanov, G.W. Han, et al. Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. Nature, 523 (2015), pp. 561-567
38. H. Dwivedi-Agnihotri, M. Chaturvedi, M. Baidya, T.M. Stepniewski, S. Pandey, J. Maharana, A. Srivastava, N. Caengprasath, A.C. Hanyaloglu, J. Selent, et al. Distinct phosphorylation sites in a prototypical GPCR differently orchestrate β-arrestin interaction, trafficking, and signaling. Sci Adv, 6 (2020), Article eabb8368
39. Y.-Q. Ping, C. Mao, P. Xiao, R.-J. Zhao, Y. Jiang, Z. Yang, W.-T. An, D.-D. Shen, F. Yang, H. Zhang, et al. Structures of the glucocorticoid-bound adhesion receptor GPR97–G o complex. Nature, 589 (2021), pp. 620-626
40. X.E. Zhou, Y. He, P.W. de Waal, X. Gao, Y. Kang, N. Van Eps, Y. Yin, K. Pal, D. Goswami, T.A. White, et al. Identification of phosphorylation codes for arrestin recruitment by G protein-coupled receptors. Cell, 170 (2017), pp. 457-469. e13
41. W. Huang, M. Masureel, Q. Qu, J. Janetzko, A. Inoue, H.E. Kato, M.J. Robertson, K.C. Nguyen, J.S. Glenn, G. Skiniotis, et al. Structure of the neurotensin receptor 1 in complex with β-arrestin 1. Nature, 579 (2020), pp. 303-308
42. C. Wang, Y. Jiang, J. Ma, H. Wu, D. Wacker, V. Katritch, G.W. Han, W. Liu, X.-P. Huang, E. Vardy, et al. Structural basis for molecular recognition at serotonin receptors. Science, 340 (2013), pp. 610-614
43. N. Velgy, G. Hedger, P.C. Biggin. GPCRs: what can we learn from molecular dynamics simulations? Methods Mol Biol, 1705 (2018), pp. 133-158
44. N.R. Latorraca, A.J. Venkatakrishnan, R.O. Dror. GPCR dynamics: structures in motion. Chem Rev, 117 (2017), pp. 139-155
45. M. Torrens-Fontanals, T.M. Stepniewski, D. Aranda-García, A. Morales-Pastor, B. Medel-Lacruz, J. Selent. How do molecular dynamics data complement static structural data of GPCRs. Int J Mol Sci, 21 (2020), p. 5933
46. A.J. Venkatakrishnan, A.K. Ma, R. Fonseca, N.R. Latorraca, B. Kelly, R.M. Betz, C. Asawa, B.K. Kobilka, R.O. Dror. Diverse GPCRs exhibit conserved water networks for stabilization and activation. Proc Natl Acad Sci, 116 (2019), pp. 3288-3293
47. J. Selent, F. Sanz, M. Pastor, G.D. Fabritiis. Induced effects of sodium ions on dopaminergic G-protein coupled receptors. PLoS Comput Biol, 6 (2010), Article e1000884
48. B. Selvam, Z. Shamsi, D. Shukla. Universality of the sodium ion binding mechanism in class A G-protein-coupled receptors. Angew Chem Int Ed, 57 (2018), pp. 3048-3053
49. X. Hu, Y. Wang, A. Hunkele, D. Provasi, G.W. Pasternak, M. Filizola. Kinetic and thermodynamic insights into sodium ion translocation through the μ-opioid receptor from molecular dynamics and machine learning analysis. PLoS Comput Biol, 15 (2019), Article e1006689
50. A. Kapoor, G. Martinez-Rosell, D. Provasi, G. de Fabritiis, M. Filizola. Dynamic and kinetic elements of μ-opioid receptor functional selectivity. Sci Rep, 7 (2017), p. 11255

51. D. Meral, D. Provasi, M. Filizola. An efficient strategy to estimate thermodynamics and kinetics of G protein-coupled receptor activation using metadynamics and maximum caliber. J Chem Phys, 149 (2018), p. 224101
52. O. Fleetwood, P. Matricon, J. Carlsson, L. Delemotte. Energy landscapes reveal agonist control of G protein-coupled receptor activation via microswitches. Biochemistry, 59 (2020), pp. 880-891
53. •P.W. Hildebrand, A.S. Rose, J.K.S. Tiemann. Bringing molecular dynamics simulation data into view. Trends Biochem Sci, 44 (2019), pp. 902-913

Interactive visualization and sharing of MD trajectories can increase the understanding, reliability, and reusability of MD data. Thanks to recent technological developments, centralized, special-purpose MD platforms are being created. This will facilitate access and ultimately broaden the outreach of MD simulations.

54. M. Martí-Solano, F. Sanz, M. Pastor, J. Selent. A dynamic view of molecular switch behavior at serotonin receptors: implications for Functional selectivity. PloS One, 9 (2014), Article e109312
55. P.J. Stansfeld, J.E. Goose, M. Caffrey, E.P. Carpenter, J.L. Parker, S. Newstead, M.S.P. Sansom. MemProtMD: automated insertion of membrane protein structures into explicit lipid membranes. Structure, 23 (2015), pp. 1350-1361
56. •A.Inoue, F. Raimondi, F.M.N. Kadji, G. Singh, T. Kishi, A. Uwamizu, Y. Ono, Y. Shinjo, S. Ishida, N. Arang, et al. Illuminating G-protein-coupling selectivity of GPCRs. Cell, 177 (2019), pp. 1933-1947. e25

Based on a large-scale functional interaction study, authors reveal sequence-based features of GPCRs underlying G-protein selectivity. They then apply this knowledge to develop a highly accurate predictor for scoring G-protein coupling from a given GPCR sequence. The obtained dataset is a valuable resource for research in GPCR signaling.

57. •C. Avet, A. Mancini, B. Breton, C.L. Gouill, A.S. Hauser, C. Normand, H. Kobayashi, F. Gross, M. Hogue, V. Lukasheva, et al. Selectivity landscape of 100 therapeutically relevant GPCR profiled by an effector translocation-based BRET platform. bioRxiv (2020), 10.1101/2020.04.20.052027

Authors use a novel BRET-based effector membrane translocation assay to monitor the recruitment of downstream effectors upon G protein activation and β-arrestin engagement. Based on that, they profile 100 therapeutically relevant GPCRs, revealing a great diversity in coupling selectivity. The obtained data provides a rich source of information to explore the principles underlying receptor/G protein coupling selectivity relationships.

3.4. GPCRmd uncovers the dynamics of the 3D-GPCRome

MD simulation is a well-established method for the characterization of the structural motions of GPCRs, which highly determine their functionality. However, the study of MD simulations typically requires efficient storage resources and specialized software, limiting the dissemination of this data to specialists in the field. To bridge this gap, we have created GPCRmd (www.gpcrmd.org), a community-driven online resource that aims to approach the information of MD simulations to all researchers interested in GPCRs. As such, GPCRmd provides access to simulations of most GPCR structures solved to date, together with the necessary metadata to ensure transparency and reproducibility. Moreover, it facilitates the analysis of this data, as it is equipped with a comprehensive set of web-based tools for the interactive visualization and analysis of the MD simulations. This includes multiple selection options to create representations of the simulated systems, as well as tools to study the interaction networks, root-mean-square deviation (RMSD), formation of tunnels and channels, and mapping of GPCR variants on the protein structure, among many others. We also implemented a meta-analysis of GPCR simulations to compare and cluster the available simulations, or a subset of interest, based on their interaction patterns. We demonstrate the use of GPCRmd and its data by performing comparative analyses across multiple receptors to shed light on two important aspects of GPCR biology, namely the dynamics of water networks and the binding of allosteric sodium. Overall, being an open, intuitive, and standardized resource, GPCRmd has the potential to foster interdisciplinary research, data reproducibility, and transparent dissemination of GPCR MD simulations.

The PhD candidate was responsible for the design and development of multiple sections of the GPCRmd platform, with special focus on the visualization and analysis tools. She was also in charge of the maintenance of the server and was involved in writing the manuscript.

Rodríguez-Espigares, I.*, Torrens-Fontanals, M.* et al. GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods* 17, 777–787 (2020). doi:10.1038/s41592-020-0884-y

* Both authors contributed equally to this work.

GPCRmd uncovers the dynamics of the 3D-GPCRome

Ismael Rodríguez Espigares[#,1], Mariona Torrens Fontanals[#,1], Johanna K. S. Tiemann[2,3], David Aranda García[1], Juan Manuel Ramírez-Anguita[1], Tomasz M. Stepniewski[1], Nathalie Worp[1], Alejandro Varela-Rial[4,5], Adrián Morales-Pastor[1], Brian Medel-Lacruz[1], Gáspár Pándy-Szekeres[6], Eduardo Mayol[7], Toni Giorgino[8,9], Jens Carlsson[10], Xavier Deupi[11,12], Slawomir Filipek[13], Marta Filizola[14], José Carlos Gómez-Tamayo[7], Angel Gonzalez[7], Hugo Gutiérrez-de-Terán[15], Mireia Jiménez-Rosés[7], Willem Jespers[15], Jon Kapla[10], George Khelashvili[16,17], Peter Kolb[18], Dorota Latek[13], Maria Marti-Solano[18,19], Pierre Matricon[10], Minos-Timotheos Matsoukas[7,20], Przemyslaw Miszta[13], Mireia Olivella[7], Laura Perez-Benito[7], Davide Provasi[14], Santiago Ríos[7], Iván R. Torrecillas[7], Jessica Sallander[15], Agnieszka Sztyler[13], Silvana Vasile[15], Harel Weinstein[16,17], Ulrich Zachariae[21,22], Peter W. Hildebrand[2,3,23], Gianni De Fabritiis[4,5], Ferran Sanz[1], David E. Gloriam[6], Arnau Cordomi[7], Ramon Guixà-González[24,25,*], Jana Selent[26,*]

1. Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute-Department of Experimental and Health Sciences, Pompeu Fabra University, Barcelona, Spain.
2. Institute of Medical Physics and Biophysics, Charite University Medicine Berlin, Berlin, Germany.
3. Institute of Medical Physics and Biophysics, Medical University Leipzig, Leipzig, Sachsen, Germany.
4. Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park, Barcelona, Spain.
5. Acellera, Barcelona, Spain.
6. Department of Drug Design and Pharmacology, University of Copenhagen, Copenhagen, Denmark.
7. Laboratori de Medicina Computacional, Unitat de Bioestadistica, Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra, Spain.
8. Biophysics Institute, National Research Council of Italy, Milan, Italy.
9. Department of Biosciences, University of Milan, Milan, Italy.
10. Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden.
11. Laboratory of Biomolecular Research, Paul Scherrer Institute (PSI), Villigen PSI, Switzerland.
12. Condensed Matter Theory Group, PSI, Villigen PSI, Switzerland.
13. Faculty of Chemistry, Biological and Chemical Research Centre, University of Warsaw, Warsaw, Poland.
14. Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

15. Department of Cell and Molecular Biology, Uppsala University, Biomedical Center, Uppsala, Sweden.
16. Department of Physiology and Biophysics, Weill Cornell Medical College of Cornell University, New York, NY, USA.
17. Institute for Computational Biomedicine, Weill Cornell Medical College of Cornell University, New York, NY, USA.
18. Department of Pharmaceutical Chemistry, Philipps-University Marburg, Marburg, Germany.
19. MRC Laboratory of Molecular Biology, Cambridge, UK.
20. Department of Pharmacy, University of Patras, Patras, Greece.
21. Computational Biology, School of Life Sciences, University of Dundee, Dundee, UK.
22. Physics, School of Science and Engineering, University of Dundee, Dundee, UK.
23. Berlin Institute of Health, Berlin, Germany.
24. Laboratory of Biomolecular Research, Paul Scherrer Institute (PSI), Villigen PSI, Switzerland. ramon.guixa@psi.ch.
25. Condensed Matter Theory Group, PSI, Villigen PSI, Switzerland. ramon.guixa@psi.ch.
26. Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute-Department of Experimental and Health Sciences, Pompeu Fabra University, Barcelona, Spain. jana.selent@upf.edu.

# Contributed equally.

* Corresponding author

Abstract

G-protein-coupled receptors (GPCRs) are involved in numerous physiological processes and are the most frequent targets of approved drugs. The explosion in the number of new three-dimensional (3D) molecular structures of GPCRs (3D-GPCRome) over the last decade has greatly advanced the mechanistic understanding and drug design opportunities for this protein family. Molecular dynamics (MD) simulations have become a widely established technique for exploring the conformational landscape of proteins at an atomic level. However, the analysis and visualization of MD simulations require efficient storage resources and specialized software. Here we present GPCRmd (http://gpcrmd.org/), an online platform that incorporates web-based visualization capabilities as well as a comprehensive and user-friendly analysis toolbox that allows scientists from different disciplines to visualize, analyze and share GPCR MD data. GPCRmd originates from a community-driven effort to create an open, interactive and standardized database of GPCR MD simulations.

1. Introduction

GPCRs are abundant cell surface receptors accounting for ~4% (800) of all human genes. They play a vital role in signal transduction by regulating numerous aspects of human physiology and are the targets of 34% of the drugs approved by the US Food and Drug Administration[1]. Important advances in protein engineering, X-ray crystallography and cryo-electron microscopy (cryo-EM) over the past decade have led to an exponential increase in the number of available GPCR structures (3D-GPCRome) deposited in the Protein Data Bank (PDB) (GPCRdb http://gpcrdb.org/structure/statistics, 2019). This rapid growth has fueled the development of the GPCRdb[2], an online resource for GPCR reference data, analysis, visualization and data-driven experiment design. This resource provides a wide range of tools including a knowledge-based resource for GPCR crystal and cryo-EM structure determination[3].

However, static high-resolution structures provide little information on the intrinsic flexibility of GPCRs, a key aspect to fully understand their function. Important advances in the computer science field have transformed computer simulations into a very powerful technique to explore protein conformational landscapes. In particular, all-atom molecular dynamics (MD) simulations have proved useful to complement experiments and characterize GPCR fluctuations at the atomic level[4]. Likely due to technical and sustainability limitations, only a modest number of online resources cover MD simulations (reviewed in ref. [5]). Recent large improvements of internet bandwidth, compression of simulation data and storage capacities now enable faster and larger online repositories that host atom trajectories from MD simulations. Moreover, new visualization[6] and online file-sharing[7,8] tools have opened the door to streaming and remotely inspecting MD trajectories online, thereby removing the need for specialized MD software[5].

Here we present the GPCRmd platform (Fig. 1), an open access and community-driven research resource for sharing GPCR MD simulations with the aim of mapping the entire 3D-GPCRome. This new resource paves the way for GPCR scientists from very different

disciplines to perform comparative studies on universal aspects of GPCR dynamics. We showcase the potential of GPCRmd for exploring key aspects of GPCR dynamics by performing comparative analyses of internal water molecules and sodium ion binding in multiple GPCR MD simulations. The open and intuitive design of the GPCRmd platform will not only foster interdisciplinary research and data reproducibility, but also transparent and easy dissemination of GPCR MD simulations.

2. Results

2.1. MD simulations from all GPCR classes structurally solved to date

GPCRmd is a community-driven resource that provides direct and interactive visualization of MD trajectories, and that is only contingent on a web browser. As a result, the GPCRmd platform grants easy access for both computational and nonexpert scientists. Moreover, we equipped it with a comprehensive set of tools to easily analyze molecular interactions and protein motions involving conserved, pharmacologically relevant or disease-related residues and structural motifs potentially involved in GPCR function (Fig. 1b,c). In adherence to the findable, accessible, interoperable and reusable principles for scientific data management[9], GPCRmd provides open access to all of its data and simulations protocols (Fig. 1a). Corresponding data are deposited either by individual contributions or biannual updates from the GPCRmd community.

We initiated the GPCRmd database by creating a comprehensive MD dataset including at least one representative structure from each of the four structurally characterized GPCR classes. To allow for comparison of ligand-induced effects across receptors, this first set comprises 98 PDB identifiers from 38 different receptor subtypes (Fig. 2) in their apo form or bound to a natural ligand, surrogate agonist or antagonist (see Methods). To generate reproducible data, we carefully designed a common protocol for the collective set-up and simulation of all structures listed in Fig. 2 (see Methods) and made it publicly available at https://github.com/GPCRmd/MD-protocol. Each system was simulated for 500 ns in three replicates (total time 1.5 μs) allowing for structure relaxation as well as sampling of receptor flexibility. At the time of writing, the

GPCRmd platform holds 588 GPCR MD simulations from the GPCRmd community plus 28 simulations from individual contributions totaling to an aggregated simulation time of ~400 µs.

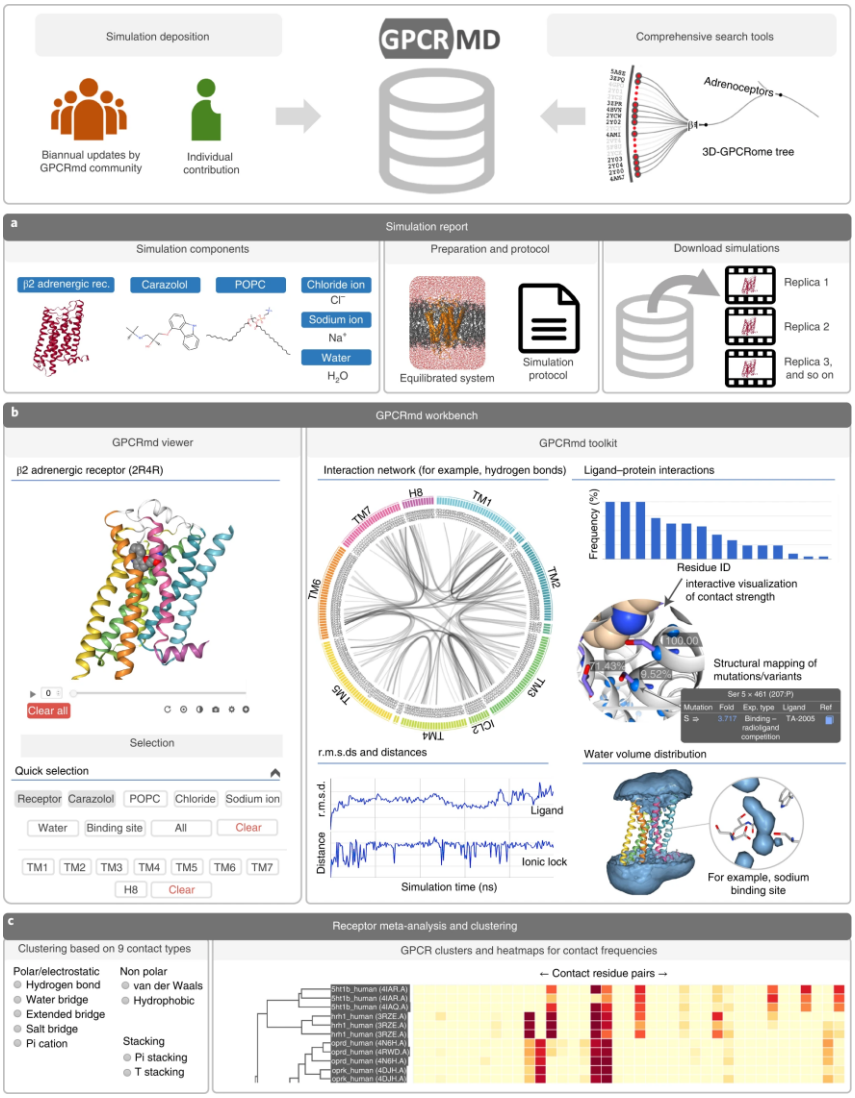


Fig. 1. GPCRmd framework. GPCRmd is an online resource for storage, streaming and analysis of GPCR MD simulation data from individual contributions and biannual collective updates. An intuitive search algorithm allows for comprehensive screening of the database. **a**, The user obtains detailed information about the simulation data via the simulation report. **b**, A GPCR-specific workbench enables interactive visualization (GPCRmd viewer) and analysis (GPCRmd toolkit) for individual simulations. **c**, Finally, the comparative analysis and clustering of multiple MD simulations helps finding relationships between receptors based on nine different molecular interaction types.

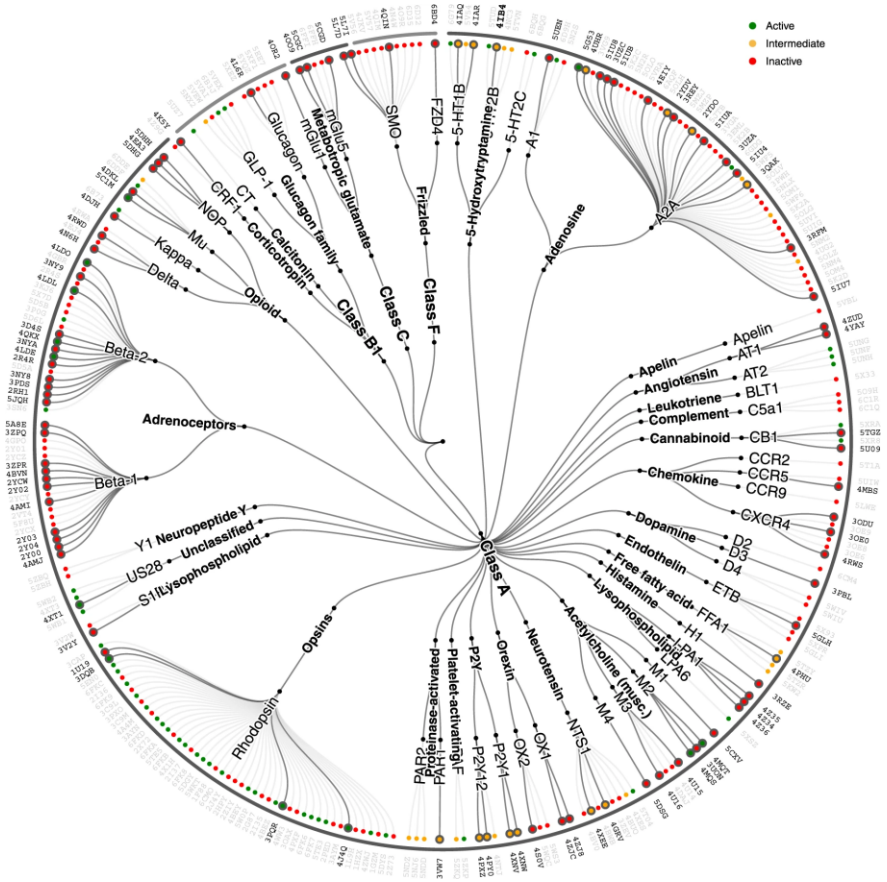


Fig. 2. The 3D-GPCRome. Mapping the GPCR structures contained in the first GPCRmd release onto the 3D-GPCRome tree. The first GPCRmd dataset of simulated structures (191 systems at the time of manuscript preparation) covers 100% of GPCR classes, 71% of receptors subtypes and 80% of GPCR families with solved structure at the time of writing, and accounts for approximately 35% of all GPCR structures deposited in the PDB (black PDB identifiers). Colored circles differentiate between active (green), intermediate (yellow) or inactive (red) receptor states.

2.2. GPCRmd viewer: sharing and interactive visualization of GPCRs in motion

To provide easy sharing and interactive visualization of GPCR MD simulations within the 3D-GPCRome, we created the GPCRmd viewer (Fig. 3). This viewer builds on MDsrv[6], a recently published tool that allows easy trajectory sharing and makes use of the

interactive capabilities of the popular web-based structure viewer NGL[5].

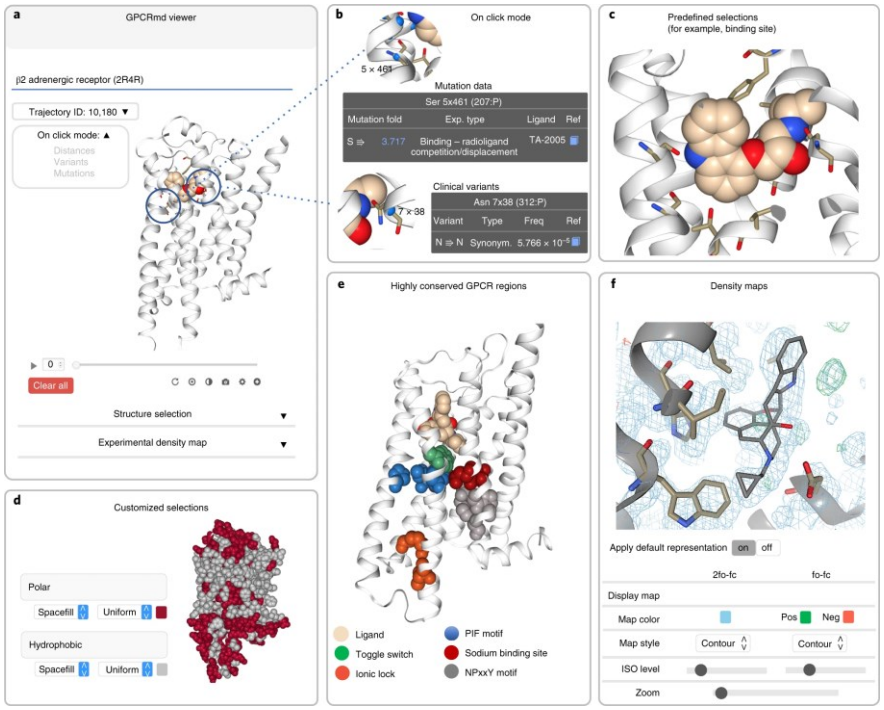


Fig. 3. The GPCRmd viewer. **a–f**, Interactive visualization of GPCR MD simulations allows for streaming simulations **(a)**, structural mapping of mutation data and clinical variants **(b)**, predefined selections of simulation components and ligand binding sites **(c)**, customized selections that enable tailored visualization of trajectories **(d)**, knowledge-based selections for visualization of GPCR conserved regions **(e)** and density maps allows for comparison between experiments and MD simulations **(f)**. A set of predefined, custom and knowledge-based selections enables quick exploration of particular regions of the map such as the ligand binding pocket. Flexible options allow users to change the color of the classical difference (Fo – Fc) and composite (2Fo – Fc) map, style (for example, wireframe or contour) or the surface and zoom levels.

The GPCRmd viewer provides interactive structural analysis of the simulations through on-click actions (Fig. 3b). To account for the fact that almost 25% of the GPCR functional sites show an average of at least one polymorphism, we mapped all GPCR variants[10] and site-directed mutations[11] from the GPCRdb[2] to each GPCR structure. Activation of the modes 'Show variants' or 'Show mutations' displays, respectively, each variant or mutation as small

beads (Fig. 3b). A click on a bead reveals further information on the variant/mutation, including a link to experimental data and the original publication. A separate on-click mode, 'Show distances', exploits NGL[5] to measure atom pair distances.

The powerful selection capabilities of the viewer (Fig. 3c–e) enable fast inspection of trajectories. Standard selections quickly visualize any molecule type in the simulation, neighboring molecules at a custom distance of each other or specific positions along the protein sequence. It is worth noting that the GPCRmd viewer makes use of GPCRdb generic residue numbering[12] by automatically linking each residue to its respective index position. Predefined knowledge-based selections enable more specific displays such as residues within 2.5 Å of the ligand (Fig. 3c), individual GPCR helices or highly conserved positions and functional motifs (Fig. 3e). In addition, the NGL selection language (see Documentation) enables the use of custom selection keywords to create tailored representations of any atom or part of the trajectory loaded in the GPCRmd viewer (Fig. 3d). Since several of these keywords stand for the chemical nature or secondary structure of proteins, they are particularly helpful for visual analysis of GPCR dynamics.

Furthermore, the GPCRmd viewer provides visualization of X-ray and electron microscopy density maps from the PDB. This allows for atomic-level comparison of the GPCR conformational landscape inferred from experimentally determined structures and observed in MD simulations (Fig. 3f).

2.3. GPCRmd toolkit: investigation of GPCR dynamics through interactive analysis

The GPCRmd toolkit provides intuitive analysis of the MD simulations by complementing and directly interacting with the GPCRmd viewer (Fig. 1b, left). The toolkit allows to compute custom distances, root mean square deviation (r.m.s.d.), and averaged water density maps for individual simulations (Fig. 1b, right). In addition, it provides interactive tools to qualitatively and quantitatively compare the noncovalent landscape of contacts for the entire GPCRmd dataset (Fig. 1b, right).

2.3.1. Interaction network tool

To easily identify relevant noncovalent contacts in GPCRmd simulations, the GPCRmd toolkit uses Flareplots[13], an interactive circular representation of contact networks that can be displayed per frame or summarized for the complete trajectory (Fig. 4, right). The interaction network tool automatically integrates the GPCRmd viewer with the GPCRmd toolkit, making it straightforward to detect, for instance, differences in the hydrogen bonding network dynamics between active and inactive receptor simulations. The current version of the interaction network tool focuses on intra- and inter-helical interactions including nine different types of noncovalent interaction (see Methods).

2.3.2. Interaction frequency tools

The GPCRmd toolkit provides two dedicated tools to study key electrostatic interactions, namely hydrogen bonds and salt bridges. The hydrogen bonds tool identifies GPCR intra- and intermolecular hydrogen bonds formed during the simulation, whereas the salt bridges tool identifies GPCR intramolecular salt bridges. Moreover, these tools allow studying the interplay between the receptor and the membrane by computing protein–lipid interactions. Furthermore, it can identify protein residues involved in ligand binding through the ligand–receptor contacts tool. The tool outputs the interaction strength at each residue by computing its contact frequency (Fig. 1b, right). All three contact tools provide interactive visualization of the results in the GPCRmd viewer.

2.3.3. The r.m.s.d. and distance tools

The GPCRmd toolkit can monitor a change in distance between any pair of atoms during the simulation. Alternatively, per-frame atom distances can be measured and displayed in the viewer via on-click actions. While distance measurements can provide relevant information on protein structure (for example, functionally relevant protein motions, bond formation/breaking and so on), r.m.s.d. calculations are more suited to quantify structural stability and conformational changes. The r.m.s.d. tool measures the structural difference of protein and ligand atoms at any point in the simulation with respect to the initial frame. Therefore, it can be used to monitor

simulation integrity or structural deviations throughout the simulation. Both tools generate time course plots (Fig. 1b, right) that can interactively link each data point to its respective frame in the viewer.

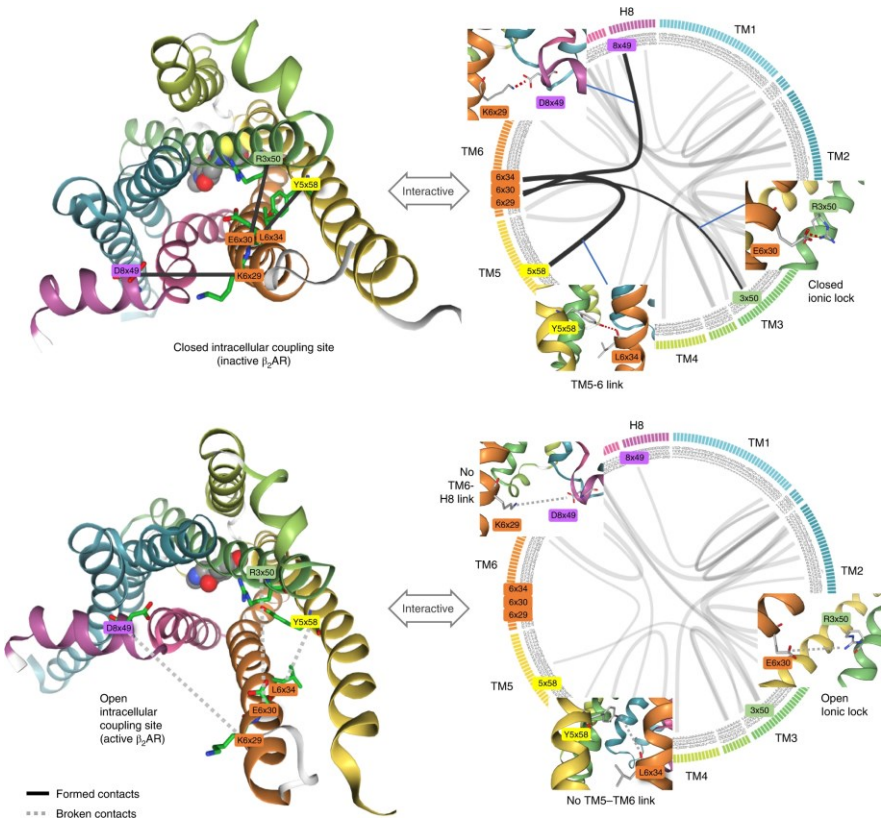


Fig. 4. Interaction network tool. Interactive visualization and analysis of intramolecular contacts. Summary plot for the hydrogen bonding network (that is, average over the entire trajectory) obtained by selecting hydrogen bonds as interaction type. Circular plots (right) for the inactive β_2AR in complex with timolol (PDB ID 3D4S, upper panel) and the active β_2AR in complex with adrenaline (PDB 4LDO, lower panel), where line thickness represents contact frequency. For clarity, only contact frequencies over 10% are shown. Comparison of these plots reveals important differences specifically at the intracellular coupling site. The inactive receptor displays contacts that help maintain the receptor in a closed state, such as the characteristic ionic lock between R3x50 and E6x30 (ref. [30]), a TM5–TM6 linkage established by Y5x58 to the backbone of L6x34, and a TM6-Helix8 connection between K6x29 and D8x49. Such contacts are missing in the active β_2AR conformation. The user can interactively explore the dynamics of the plotted contacts in the circular plot (right panel) in a structural context (left panel). Residues are numbered according to their GPCRdb generic numbering scheme[12].

2.3.4. Water volume distribution tool

Due to the vital role of internal water molecules in GPCRs[14], we equipped the GPCRmd toolkit with a water density map tool. This tool can quickly display an averaged water density map of the MD trajectory under study in the GPCRmd viewer (Fig. 1b, right), thus allowing to monitor, for example, the formation of the continuous internal water channel known to be essential for GPCR activation[15].

2.3.5. Tunnels and channels tool

Just like all proteins, GPCRs hold an intricate system of tunnels and channels that can facilitate the access of water, ions, lipids and ligands by connecting the outside environment to the receptor core[16,17]. Since these pathways may change substantially over time, we provided the GPCRmd toolkit with a tool to analyze and display tunnels and channels. The new widget allows users to select among the list of computed tunnels and channels to immediately display them in the GPCRmd viewer using different visualization schemes.

2.4. Functional hotspots discovered through meta-analysis of GPCR simulations

The GPCRmd platform can uniquely compare GPCR simulations within the 3D-GPCRome (Figs. 1c and 2). We developed a module specifically comparing multifold GPCR simulations to uncover universal or distinct mechanisms governing the structural dynamics of these receptors. This module computes the contact frequency of each residue pair for multiple simulations and displays a global comparative analysis via an interactive heatmap plot (Fig. 5a, left). The tool also performs clustering analysis of the contact frequency data to hierarchically classify each receptor and display the resulting tree alongside the heatmap plot (Fig. 5a, left). To further facilitate the interpretation of large heatmaps, we added interactive analysis and visualization capabilities of selected clusters using Flareplots[13] and NGL[5].

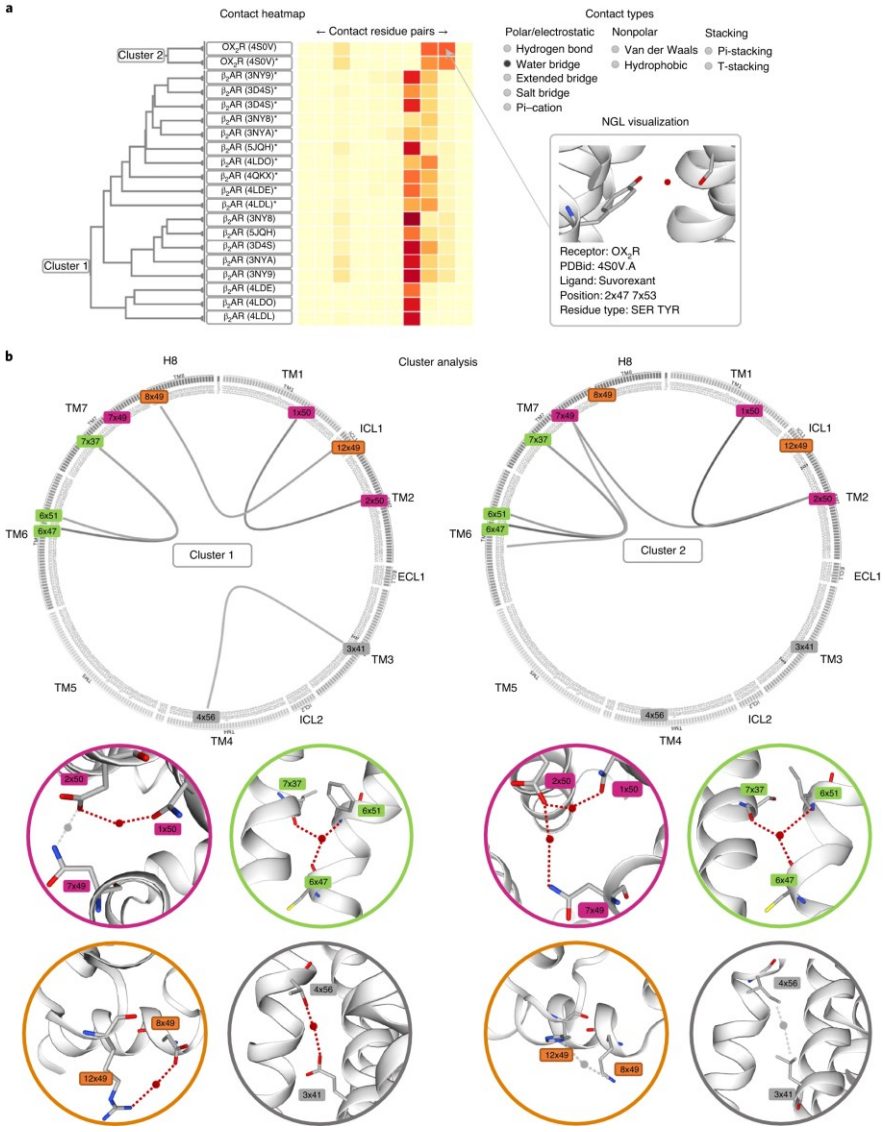


Fig. 5. A water bridge signature revealed by comparative analysis using the GPCRmd. a, Heatmap of water-mediated interactions of clusters belonging to the β_2AR and OX_2R. The plot displays each residue pair (columns) for each GPCR (rows). Yellow to red color scale stands for low to high contact frequency. Users can select up to nine different noncovalent interaction types to perform the analysis across the complete GPCRmd database or just using a custom subset of simulations. On-click actions provide detailed information on the specific interaction and system involved for each cell of the heatmap. **b**, Representative water-mediated interactions for the investigated clusters are shown in circular plots. Corresponding structural depictions of interactions are found below the circular plots. Residues are numbered according to their GPCRdb generic numbering scheme[12].

To demonstrate the use of the meta-analysis tool, and due to their critical role in receptor function[14,15], we investigated the interaction fingerprint of water molecules in GPCRs. Along with previously described[18] conserved water networks, this analysis revealed other water-mediated interactions that are conserved among different receptor subtypes and reported here. For example, in line with Venkatakrishnan et al.[18], the β_2-adrenoceptor (β_2AR) and OX_2-receptor (OX_2R) display a common water network that links TM1 (N1x50) and TM2 (D2x50) (Fig. 5b, highlighted in purple). This water network is extended from TM2 (2x50) to TM7 (7x49) in the OX_2R cluster. Such a water network extension is not observed in the β_2AR cluster due to closer proximity of residues 2x50 and 7x49, which enables direct, unmediated, contacts. Another conserved water-mediated feature is a bifurcated polar network linking TM6 (6x47, 6x51) and TM7 (7x37) via helix backbones in the β_2AR and the OX_2R clusters (Fig. 5b, highlighted in green). Our study shows that this bifurcated network is less prominent in active structures (Supplementary Fig. 1). Taking into account that TM6 undergoes large conformational changes on receptor activation, it is tempting to speculate that uncoupling the interactions between individual water molecules in this bifurcated network represents a step during receptor activation.

Likewise, our analysis reveals important differences between both clusters. A water bridge between intracellular loop 1 (ICL1, 12x49) and helix 8 (H8, 8x49) is found to be only present in the β_2AR (Fig. 5b, highlighted in orange). Further studies (for example, site-directed mutagenesis) could be used in the future to investigate whether this water bridge contributes to the distinct coupling efficacy and/or specificity shown by the β_2AR (principal signaling pathway, G_s family[19]) and OX_2R (principal signaling pathways; G_s family, G_i/G_o family and G_q/G_{11} family[19]). Finally, our collective analysis reveals a water bridge between TM3 (3x41) and TM4 (4x56) only observed in the β_2AR (Fig. 5b, highlighted in gray) and likely related to the striking change in receptor stability observed on mutation of residue E3x41 in experiments[20].

2.5. Exploiting the entire GPCRmd dataset: custom analysis of sodium ion interactions across class A GPCRs

We made the entire GPCRmd dataset available for download (see Methods), thus opening the door for the scientific community to perform comparative analyses of multiple simulations across several receptor structures, families, subtypes and classes. To demonstrate the value of such a comprehensive dataset, we studied sodium ion (Na^+) interaction in GPCRs[21], an almost universal, albeit poorly understood, mechanism of allosteric modulation of these receptors[22]. We analyzed Na^+ interaction to conserved orthosteric (3x32) and allosteric (2x50) residues in 183 simulations (61 different apo structures × three replicas) covering 26 different class A receptor subtypes. The markedly different frequencies of Na^+ interaction with these two residues enable receptors to be clustered in three groups (I, II and III, Fig. 6a,c). Note that our dataset (3×0.5 μs) provides valuable insights into sodium interaction sites but it is not sufficient to conclude about binding kinetics.

In line with previous studies using multiple simulations[23], our analysis shows that Na^+ binds to D2x50 and/or position 3x32 in most of the receptor subtypes (Fig. 6a). Group I (serotonin, dopamine and nociception receptors) shows high sodium interaction frequencies to positions 3x32 and 2x50, the latter being stabilized by a hydrogen bonding network often composed of D2x50, S3x39, N7x45 and S7x46 (so-called DSNS motif) (Fig. 6b). The high interaction frequency to both positions implicates that at times Na^+ ions bind simultaneously to position 3x32 and the allosteric site at D2x50. This seems to be a consequence of a higher negative net charge at the extracellular side (Fig. 6c,d), which increases the local concentration of positively charged Na^+ around the receptor entrance, and likely facilitates the simultaneous entrance of a second ion. Notably, despite a completely conserved DSNS motif, group II (β-adrenergic and muscarinic receptors) shows marginal interaction frequency at D2x50, while still exhibiting a high interaction frequency at 3x32. Visual inspection of the simulation reveals hydrophobic barriers that hamper Na^+ passage from 3x32 to 2x50 (Fig. 6c), in line with previous MD simulation studies[23,24]. In contrast to group II, we find high interaction frequencies at position 2x50 for group III and none or only marginal contacts with 3x32.

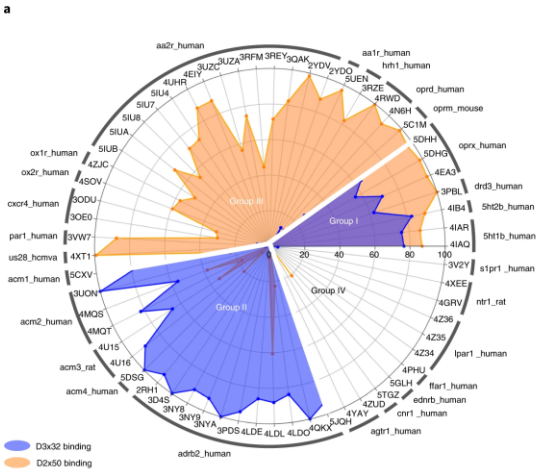
Most receptors of this group (for example, adenosine A_{1A} and A_{2A}, or the chemokine receptor CXCR4) lack an aspartate in position 3x32 allowing for direct diffusion to position 2x50 (Fig. 6c). Despite having a D3x32 (Fig. 6b), only low binding is observed during the simulated time frame at this position for a small subgroup of receptors including histamine H_1 and opioid μ and δ receptors. More simulation time would be required to improve the sampling of ion binding to D3x32. Finally, in a particular subset of receptors, Na^+ binds neither to D2x50 nor to position 3x32 within the studied time frame (group IV, Fig. 6a). In fact, slower Na^+ binding kinetics has previously been reported[23] and could be the consequence of blocked access to the binding site from the extracellular side (for example, receptors taking up ligands from the lipid bilayer).

While our results confirm the essential role of D2x50 for allosteric sodium binding[14,25] in class A GPCRs, they also reveal that the presence or absence of D3x32 in the orthosteric binding site determine distinct Na^+ binding profiles. This analysis exemplifies the potential of the comprehensive GPCRmd dataset to investigate how GPCR sequence, structure and dynamics can jointly contribute to receptor allosteric modulation.

3. Discussion

In the last decade, static structures in the 3D-GPCRome have predominantly been described as active, intermediate or inactive states. However, a growing body of research suggests that GPCRs are not two- or three-state systems but exhibit a wide range of conformational states with sometimes subtle yet important differences. While several experimental techniques such as nuclear magnetic resonance (NMR)[26], double electron-electron resonance (DEER)[27] or single-molecule fluorescence energy transfer (smFRET)[28] have provided relevant insights into the dynamics and flexibility of GPCRs, MD simulations have emerged as the most promising opportunity to study the complexity of GPCR conformational dynamics in atomistic detail[4]. Moreover, MD simulations can resolve mechanistic elements at time scales and conditions that are not always accessible with experimental techniques.

We have demonstrated the use of the GPCRmd platform by performing comparative analyses across multiple receptors of two important aspects of GPCR biology, namely water network and allosteric Na^+ interaction analysis. We were able to pinpoint relevant structural features that help improve our current understanding of the diversity of GPCR function.

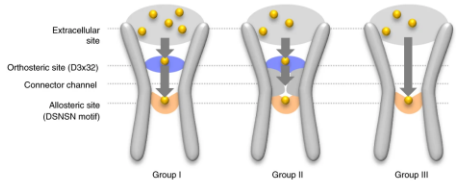


b

UniprotKB entry name	Orthosteric binding site	Allosteric binding site (DSNS motif)				
	3x32	2x50	3x39	7x45	7x46	Group
drd3_human	D	D	S	N	S	I
oprx_human	D	D	S	N	S	I
5ht1b_human	D	D	S	N	S	I
5ht2b_human	D	D	S	S	S	I
adrb2_human	D	D	S	N	S	II
acm1_human	D	D	S	N	S	II
acm2_human	D	D	S	N	S	II
acm3_rat	D	D	S	N	S	II
acm4_human	D	D	S	N	S	II
oprd_human	D	D	S	N	S	III
oprm_mouse	D	D	S	N	S	III
hrh1_human	D	D	S	N	S	III
aa1r_human	V	D	S	N	S	III
aa2ar_human	V	D	S	N	S	III
ox2r_human	Q	D	S	N	S	III
ox1r_human	Q	D	A	N	S	III
us28_hcmva	F	D	S	H	C	III
cxcr4_human	Y	D	S	H	C	III
par1_human	F	D	S	S	C	III
lpar1_human	I	D	S	N	S	IV
cnr1_human	V	D	S	N	S	IV
s1pr1_human	M	D	S	N	S	IV
ednrb_human	Q	D	T	N	S	IV
agtr1_human	V	D	S	N	N	IV
ntr1_rat	R	D	T	S	S	IV
ffar1_human	H	D	G	S	V	IV

100 — DSNS motif conservation (%) — 45

c



d

UniprotKB entry name	PDB ID	Number of (+) residues	Number of (–) residues	Net charge	Second ion entry*	Group
5ht2b_human	4IB4	4	9	–5	Yes	I
oprx_human	5DHH	3	8	–5	Yes	I
5ht1b_human	4IAR	5	9	–4	Yes	I
drd3_human	3PBL	2	5	–3	Yes	I
acm1_human	5CXV	4	5	–1	No	II
aa2ar_human	2YDO	3	6	–3	No	III
aa1r_human	5UEN	4	6	–2	No	III
oprm_mouse	5C1M	5	3	+2	No	III
oprd_human	4RWD	7	7	0	No	III

*ion entrance computed over an accumulated time of 1.5 µs (3 × 0.5 µs)

Fig. 6. Allosteric Na^+ ion interaction in class A GPCRs. a, Na^+ interaction frequency at D3x32 (green) and D2x50 (orange) in class A GPCRs across 61 structures including 26 different receptor subtypes. Receptor subtypes and 3D structures are identified by UniprotKB and PDB identifiers, respectively. The radar plot shows the prevalence of sodium interactions (0–100%) over the total accumulated simulation time of 1.5 µs (3 × 0.5 µs). **b**, Sequence alignment of sodium binding sites for the GPCR subtypes included in the simulated dataset. Allosteric binding consists of a multi-step binding process typically initiated with accumulation at the extracellular receptor side followed by receptor penetration through the orthosteric binding site (visiting D3x32, if present) before progressing to the allosteric site D2x50. **c**, GPCRs can be classified into three groups based on the sodium interaction profile. The interaction profile is driven by the structural features of the sodium entrance channel. **d**, Extracellular net charge and receptor entrance of a second ion.

3.1. A platform for interdisciplinary investigation of the 3D-GPCRome

The GPCRmd is designed to facilitate interactions and data exchange between GPCR scientists of different disciplines including structural and evolutionary biologists, computational and medicinal chemists and protein engineers (Table 1). Our tool will become a useful asset for experimental laboratories by providing open access to the dynamic context of specific GPCRs, hence directing or assisting functionally relevant experiments such as cross-linking or mutagenesis studies. Similarly, protein engineers and structural biologists will now be able to employ the GPCRmd workbench to quickly identify specific flexible regions that potentially require protein stabilization.

Table 1. Examples of how researchers from different scientific disciplines can make use of the GPCRmd database

User	Usage	GPCRmd features	Added value
Protein engineers	Stabilizing proteins for crystallization. Detection of flexible receptor regions that require stabilization to improve crystallization success.	GPCRmd workbench including the GPCRmd viewer and toolkit	Flexible receptor regions are poorly captured in experimental density maps
Crystallographers	Retrospective refinement of experimental density maps		
Crystallographers Biophysicists	(1) Detection of highly flexible regions that explain low-resolution regions in experimental density maps.	GPCRmd viewer. MD streaming with overlaid experimental density maps	Flexible receptor regions are poorly captured in experimental density maps
	(2) Detection of stable water or ion binding sites that can explain unmatched electron	GPCRmd workbench including the GPCRmd viewer for simulation	Water and ion binding are poorly captured in experimental density maps

	density areas.	streaming and the water map tool in the GPCRmd toolkit	
	(3) Rotamers and protonation states.	GPCRmd viewer for simulation streaming	Rotation states and corresponding protonation states are difficult to obtain from experimental density maps
	Study of interaction networks critical for receptor functionality. Search for receptor regions to implement linkers or signaling probes (FRET, NMR and so on) to study receptor functionality.	GPCRmd workbench including the GPCRmd viewer and toolkit	Receptor dynamics are not available in experimental density maps
Evolutionary biologists	Structural relationships and diversity across different GPCRs. How does evolution (that is, small sequence differences) affect receptor dynamics?	Receptor meta-analysis and clustering tool	Receptor dynamics are not available in experimental density maps
Medicinal chemists and drug designers	Improvement of drug–receptor interactions and design of new drugs.		
Medicinal chemists and drug designers	(1) Exploration of ligand–receptor contacts.	Ligand–receptor contacts tool	Stability of ligand–receptor interactions cannot be deduced from experimental density maps
Biomedical researcher and clinicians for personalized medicine	(2) Detection of indirect interactions (that is water or ion-mediated) that are crucial for ligand binding.	Ligand-receptor contacts and water volume distribution tools	Stability of ligand–receptor interactions cannot be deduced from experimental density maps

	(3) Flexible and transient switches.	Interaction network and interaction frequency tools	Receptor dynamics are not available in experimental density maps
	Design of treatment strategies for personalized medicine. Estimation of the impact of polymorphisms/variants on drug response through their ability to alter drug–receptor interactions or receptor dynamics in regions relevant for receptor functionality (for example, PIF motif, G-protein coupling site).	Cross-linked mutation and variant information	The impact of polymorphism/variants on the strength of ligand–receptor contacts or receptor dynamics cannot be deduced from static structures
Computational biologists (MD novices and experts, bioinformaticians)	Aid in experimental design and comparison of set-up and results in terms of force field performance, impact of ligands or mutations. Support for modeling dynamic regions able to adopt distinct conformations. Guide docking experiments on the basis of the sampled conformational space.	Simulation protocol and input structures, GPCRmd viewer for simulation streaming, GPCRmd workbench including the GPCRmd viewer and toolkit	Receptor dynamics are not available in experimental density maps
Students and teachers	Visually learning about protein dynamics: for example, receptor inactivation, water channel formation in active receptor structures, allosteric binding of sodium	GPCRmd workbench including the GPCRmd viewer and toolkit	Dynamics cannot be visualized in printed form and trajectories are not part of additional teaching materials

	ions.		
Reviewers and publishers	Data made available for scrutiny of MD articles.	GPCRmd platform	Transparency and reproducibility

Moreover, the GPCRmd will be of great benefit to medicinal chemists and drug designers. They will be able to quickly use atomic-level information on the stability/strength of specific ligand–receptor interactions, and the binding of water molecules or ions using the ligand–receptor contacts or water volume distribution tools (see Fig. 1 and Table 1). In addition, drug design scientists can use GPCRmd to investigate potential ligand binding and unbinding pathways based on the dynamics of specific structural elements such as loops, hence aiding the design of new or improved compounds. Furthermore, the GPCRmd can provide valuable structural insights into the location of natural variants and its potential impact on drug binding or receptor functionality. Our cross-referenced data allows easy mapping of variants and site-directed mutations onto the receptor structure and investigation of their dynamics during the simulations (Fig. 1b, right, and Fig. 3b). This could guide further investigations to predict drug efficacy or adverse reactions in individuals with a specific variant and in turn support the selection of more efficacious and safer drug treatments.

Beyond wet-laboratory applications, GPCRmd is an important dissemination resource for computational biologists, ranging from students and MD novices to MD experts and bioinformaticians from related fields. Our platform offers a harmonized database to perform future comparative studies across different MD setups, force fields, ligands, lipid compositions or GPCR variants, which offers a substantial advantage over currently available archives or data repositories such as FigShare (https://figshare.com/) or Zenodo (http://zenodo.org/).

3.2. The GPCRmd consortium: reproducible and sustainable research in GPCR MD simulations

This community-driven effort has laid the foundation of the GPCRmd consortium, an open community of GPCR computational

researchers driving the centralization, dissemination, and development of open source and reproducible analysis of massive amounts of GPCR MD data. We believe that GPCRmd will enhance the dissemination of scientific results by offering a platform to make published protocols and simulation data publicly available. This will promote transparency, consistency and reproducibility in the field of GPCR dynamics. On the other hand, community engagement will overcome one of the most important challenges faced by this kind of resource, namely sustainability (Supplementary Note 1). The implementation of the GPCRmd consortium under the umbrella of the active European Research Network on Signal Transduction (ERNEST, https://ernest-gpcr.eu)[29] will provide support to (1) foster the development of new analysis tools (for example, conformational analyses and dynamic pharmacophore models), and (2) increase the coverage of the 3D-GPCRome with future releases of the GPCRmd platform. While the first GPCRmd dataset from the consortium already maps more than 70% of GPCR subtypes within the 3D-GPCRome, future biannual releases as well as individual contributions from the scientific community will further increase this coverage bridging the gap between solved and simulated structures.

4. Methods

4.1. MD simulations

The first GPCRmd includes 98 different GPCR structures bound to their natural ligand (for example, sphingosine-bound $S1P_1R$), an agonist (for example, ergotamine-bound $5HT_{2B}R$) or an antagonist (for example, alprenolol-bound β_2AR). In addition to ligand–bound structures, we included an apo form of each receptor by removing the ligand from its binding pocket. We carefully designed a common protocol for the collective set-up and simulation (Supplementary Note 2) phases of all structures. During the set-up phase, different expert members of the GPCR MD community individually prepared each family of GPCR structures by refining/remodeling PDB structures (for example, missing residues, disulfide bridges, cocrystallization molecules, loop remodeling and so on), placing missing water molecules[31] and sodium ions or assigning relevant protonation states (Supplementary Note 2). Next, each protein was prepared for simulation by embedding it in a lipid

bilayer and adding water and ions to the system. Each system was equilibrated following a standard procedure previously outlined and discussed within the GPCR MD community (Supplementary Note 2). Finally, the distributed computing platform GPUGRID[32] was used to simulate three replicas of each system for 500 ns (that is, accumulated 1.5 μs). We made all set-up and simulation protocols openly available at https://github.com/GPCRmd/MD-protocol.

4.2. Database structure

The GPCRmd database and web interface were developed using Django Web Framework (v.1.9), Python (v.3.4), JavaScript libraries, jQuery 1.9, jQuery UI 1.11.2 and PostgreSQL 9. The structure of the database (Supplementary Figs. 2–8) is based on five main objects: protein objects identified by their sequence and their relationship with UniprotKB entries (Supplementary Fig. 2), molecular entities (molecule object) identified by an InChI[33] generated with forced hydrogen connectivity (Supplementary Fig. 3), crystalized assembly (model) (Supplementary Fig. 4), MD simulations (dynamics) objects (Supplementary Fig. 5) and chemical species (compound) identified by standard InChI (Supplementary Fig. 3). Furthermore, we incorporated experimental data from IUPHAR[34] and BindingDB[35] (Supplementary Fig. 6) and linked each main object to bibliographic references. GPCRdb[2] tables were used to add standard nomenclatures to GPCR sequence residue numbers.

4.3. Custom analysis

The whole GPCRmd repository is released as open source under the Creative Commons Attribution 4.0 International License hence enabling downloading and custom analysis of the comprehensive dataset. Each trajectory can be downloaded from its respective link at the simulation report page (see Documentation). We exemplified this usage by studying sodium ion binding across a selection of class A GPCRs within the GPCRmd dataset. The frequency of sodium ion binding to the closest oxygen atom of the carboxylic group ($2 \times O^{\epsilon}$) of residues 3x32 and 2x50 were computed using a cutoff distance of 5 Å. Both highly conserved positions are normally aspartate residues. For nonconserved residues we used the

following atoms: Gln (N^{ϵ}, O^{ϵ}), His (N^{δ}), Arg (N^{ϵ}), Ala (C^{γ}), Val ($2 \times C^{\gamma}$ Hydrogen), Ile ($2 \times C^{\gamma}$ Hydrogen), Met (S^{δ}), Phe ($2 \times C^{\delta}$) and Tyr ($2 \times C^{\delta}$).

4.4. GPCRmd viewer

The GPCRmd viewer uses builds on NGL 2.00 (ref. [5]) and MDsrv 0.3.5 (ref. [6]) and uses data from the PDB (rcsb.org[36]), the GPCRdb[2] and the Genome Aggregation Database (gnomAD)[37]. The data for on-click modes, variants and site-directed mutagenesis annotations are taken from the GPCRdb[2,10,11] and include generic GPCR numbers[12], original and mutated residues, effect of the mutation in ligand binding (fold change), experiment type, ligand used for the experiment and bibliographic reference. Variant data is obtained from the gnomAD[37], and includes amino acid substitutions (canonical and variant), allele frequencies and link to the gnomAD entry describing the variant. On-click selection capabilities build on NGL 2.0.0 (ref. [5]) web viewer, which allow the creation of different representation objects using the NGL selection language. GPCRmd selection capabilities also feature the GPCR generic numbering scheme[12]. In this case, GPCRdb numbers are adapted to the NGL selection language through regular expressions. Experimental density maps are loaded from PDB and aligned to the first frame of the simulation displayed using NGL 2.0.0 (ref. [5]). The transformation matrix applied to the density map to perform the alignment is precomputed using the Python library MDAnalysis v.0.20.1 (ref. [38]).

4.5. GPCRmd toolkit

4.5.1. Interaction networks

Noncovalent residue–residue interactions formed in the simulation are displayed using Flareplots[13]. To precompute interactions during the simulation, we used GetContacts[13] in all interaction types except for hydrogen bonds, where we used the definition of Wernet and Nilsson. We manually integrated Flareplots and NGL to allow for interactivity between the GPCRmd toolkit and the GPCRmd viewer.

4.5.2. Interaction frequencies

Hydrogen bonds are calculated using the 'wernet_nilsson' module of MDtraj[39]. A hydrogen bond is defined using distance and angle cut offs between hydrogen donor (NH or OH) and acceptor (N or O) atoms as follows:

$$r_{DA} < 3.3\,\text{Å} - 0.00044\,\text{Å} \cdot \delta HDA2$$

where r_{DA} is the distance (Å) between donor and acceptor heavy atoms and δ_{HDA} is the angle (degrees) formed between the hydrogen atoms of donor and acceptor atoms. By default, the analysis does not consider hydrogen bonds between neighboring residues and includes side chains as well as backbone atoms. Ligand–receptor contacts are computed using the compute_contacts module of MDtraj[39]. Salt bridge frequency is computed using the 'compute_distances' module of MDtraj[39]. Salt bridges are defined as any combination between the sets {Arg-NH1, Arg-NH2, Lys-NZ, His-NE2, His-ND1} and {Glu-OE1, Glu-OE2, Asp-OD1, Asp-OD2} with atoms closer than 4 Å. Histidine atoms are only considered if the residue is protonated. The distance between atom pairs through the entire or strided trajectories is computed using the 'compute_distances' module of MDtraj[39]. Atom pairs can be defined either using the 'Show distances' on-click mode and imported to the tool, or NGL selection language instances.

4.5.3. The r.m.s.d.

The r.m.s.d. is computed using the rmsd module of MDtraj[39]. The first frame of the trajectory is used as a reference structure by default. The atoms used for r.m.s.d. computation can be defined using the provided preselection in the GPCRmd toolkit (for example, protein alpha carbons, nonhydrogen protein atoms, ligand and so on). The r.m.s.d. is computed after optimal alignment according to the following equation:

$$r.m.s.d.(t) = \sqrt{\frac{1}{N_{atoms}} \sum_{i}^{N_{atoms}} [r_i(1) - r_i(t)]^2}$$

where N_{atoms} is the number of atoms for structure comparison, $r_i(1)$ is the position of atom i in the reference frame (that is, trajectory frame 1) and $r_i(t)$ is the position of atom i at time t of the trajectory.

4.5.4. Water volume distribution

Water occupancy maps are precomputed and stored on the server side using the VolMap tool of VMD[40]. Maps are generated only for oxygen atoms of a water molecule using a cutoff distance of 10 Å to the protein and a resolution of 1 Å. Atoms are treated as spheres using their atomic radii. The resulting isosurface is displayed in the GPCRmd viewer.

4.5.5. Tunnels and channels

Tunnels and channels are precomputed using the CAVER 3.0 software[41] and stored on the server side. We used as starting point coordinates for apo forms and receptor-ligand structures the center of mass of ligand–interacting residues in the respective PDB structure. Computations were carried out using a shell radius 3 Å, shell depth 4 Å and a probe radius of 1.4 Å. Selected results are displayed in the GPCRmd viewer.

4.6. Meta-analysis tool

Contacts are computed using GetContacts[13] and results plotted as interactive heatmaps using the Bokeh visualization library (https://docs.bokeh.org/en/latest/). Contact frequencies per system are averaged over simulation replicas. For accurate comparison, residue contact pairs are aligned using the GPCRdb generic numbering scheme[12]. Hierarchical clustering uses the 'linkage' function of the SciPy v.0.18.1 (ref. [42]) library with default parameters. Dendrogram plots use the Plotly library (https://plot.ly/python/).

4.7. Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The MD data have been deposited in the GPCRmd database (http://gpcrmd.org/).

Code availability

Set-up, simulation and analysis protocols are openly available at https://github.com/GPCRmd.

References

1. Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B. & Gloriam, D. E. Trends in GPCR drug discovery: mew agents, targets and indications. Nat. Rev. Drug Discov. 16, 829–842 (2017).
2. Munk, C. et al. GPCRdb in 2018: adding GPCR structure models and ligands. Nucleic Acids Res. 46, 440–446 (2017).
3. Munk, C. et al. An online resource for GPCR structure determination and analysis. Nat. Methods 16, 151–162 (2019).
4. Latorraca, N. R., Venkatakrishnan, A. J. & Dror, R. O. GPCR dynamics: structures in motion. Chem. Rev. 117, 139–155 (2017).
5. Hildebrand, P. W., Rose, A. S. & Tiemann, J. K. S. Bringing molecular dynamics simulation data into view. Trends Biochem. Sci. 44, 902–913 (2019).
6. Rose, A. S. & Hildebrand, P. W. NGL Viewer: a web application for molecular visualization. Nucleic Acids Res. 43, W576–W579 (2015).
7. Tiemann, J. K. S., Guixà-González, R., Hildebrand, P. W. P. W. & Rose, A. S. MDsrv: viewing and sharing molecular dynamics simulations on the web. Nat. Methods 14, 1123–1124 (2017).
8. Carrillo-Tripp, M. et al. HTMoL: full-stack solution for remote access, visualization, and analysis of molecular dynamics trajectory data. J. Comput. Aided Mol. Des. 32, 869–876 (2018).
9. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3, 160018 (2016).
10. Hauser, A. S. et al. Pharmacogenomics of GPCR drug targets. Cell 172, 41–54 (2018).
11. Munk, C., Harpsøe, K., Hauser, A. S., Isberg, V. & Gloriam, D. E. Integrating structural and mutagenesis data to elucidate GPCR ligand binding. Curr. Opin. Pharmacol. 30, 51–58 (2016).
12. Isberg, V. et al. Generic GPCR residue numbers - aligning topology maps while minding the gaps. Trends Pharmacol. Sci. 36, 22–31 (2015).
13. Venkatakrishnan, A. J. et al. Uncovering patterns of atomic interactions in static and dynamic structures of proteins. Preprint at bioRxiv https://doi-org.sare.upf.edu/10.1101/840694 (2019).
14. Liu, W. et al. Structural basis for allosteric regulation of GPCRs by sodium ions. Science 337, 232–236 (2012).

15. Yuan, S., Filipek, S., Palczewski, K. & Vogel, H. Activation of G-protein-coupled receptors correlates with the formation of a continuous internal water pathway. Nat. Commun. 5, 4733 (2014).
16. Hildebrand, P. W. et al. A ligand channel through the G protein coupled receptor opsin. PloS ONE 4, e4382 (2009).
17. Guixà-González, R. et al. Membrane cholesterol access into a G-protein-coupled receptor. Nat. Commun. 8, 14505 (2017).
18. Venkatakrishnan, A. J. et al. Diverse GPCRs exhibit conserved water networks for stabilization and activation. Proc. Natl Acad. Sci. USA 116, 3288–3293 (2019).
19. Alexander, S. P. et al. The concise guide to pharmacology 2017/18: G protein-coupled receptors. Br. J. Pharmacol. 174, S17–S129 (2017).
20. Roth, C. B., Hanson, M. A. & Stevens, R. C. Stabilization of the human β2-adrenergic receptor TM4-TM3-TM5 helix interface by mutagenesis of Glu1223.41, a critical residue in GPCR structure. J. Mol. Biol. 376, 1305–1319 (2008).
21. Selent, J., Sanz, F., Pastor, M. & De Fabritiis, G. Induced effects of sodium ions on dopaminergic G-protein coupled receptors. PLoS Comput. Biol. 6, e1000884 (2010).
22. Zarzycka, B., Zaidi, S. A., Roth, B. L. & Katritch, V. Harnessing ion-binding sites for GPCR pharmacology. Pharmacol. Rev. 71, 571–595 (2019).
23. Selvam, B., Shamsi, Z. & Shukla, D. Universality of the sodium ion binding mechanism in class A G-protein-coupled receptors. Angew. Chem. 130, 3102–3107 (2018).
24. Yuan, S., Vogel, H. & Filipek, S. The role of water and sodium ions in the activation of the μ-Opioid receptor. Angew. Chem. 52, 1–5 (2013).
25. Gutiérrez-De-Terán, H. et al. The role of a sodium ion binding site in the allosteric modulation of the A2A adenosine G protein-coupled receptor. Structure 21, 2175–2185 (2013).
26. Bostock, M. J., Solt, A. S. & Nietlispach, D. The role of NMR spectroscopy in mapping the conformational landscape of GPCRs. Curr. Opin. Struct. Biol. 57, 145–156 (2019).
27. Wingler, L. M. et al. Angiotensin analogs with divergent bias stabilize distinct receptor conformations. Cell 176, 468–478 (2019).
28. Gregorio, G. G. et al. Single-molecule analysis of ligand efficacy in β2AR–G-protein activation. Nature 547, 68–73 (2017).
29. Sommer, M. E. et al. The European Research Network on Signal Transduction (ERNEST): toward a multidimensional holistic understanding of G protein-coupled receptor signaling. ACS Pharmacol. Transl. Sci. 3, 361–370 (2020).
30. Ballesteros, J. A. et al. Activation of the β2-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. J. Biol. Chem. 276, 29171–29177 (2001).
31. Mayol, E. et al. HomolWat: a web server tool to incorporate 'homologous' water molecules into GPCR structures. Nucleic Acids Res. (in the press); https://doi-org.sare.upf.edu/10.1093/nar/gkaa440

32. Buch, I., Harvey, M. J., Giorgino, T., Anderson, D. P. & De Fabritiis, G. High-throughput all-atom molecular dynamics simulations using distributed computing. J. Chem. Inf. Model. 50, 397–403 (2010).
33. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. J. Cheminformatics 7, 23 (2015).
34. Southan, C. et al. The IUPHAR/BPS Guide to pharmacology in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic Acids Res. 44, D1054–D1068 (2016).
35. Gilson, M. K. et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 44, D1045–D1053 (2016).
36. Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000).
37. Karczewski, K. J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. Preprint at bioRxiv 531210 (2019).
38. Gowers, R. J. et al. MDAnalysis: a python package for the rapid analysis of molecular dynamics simulations. In Proc. 15th Python Sci. Conference 98–105 (2016).
39. McGibbon, R. T. et al. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. Biophys. J. 109, 1528–1532 (2015).
40. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. J. Mol. Graph. 14, 33–38 (1996).
41. Chovancova, E. et al. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. PLoS Comput. Biol. 8, 23–30 (2012).
42. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272 (2020).

Acknowledgements

The GPCRmd consortium acknowledges the support of COST Action CA18133, the European Research Network on Signal Transduction (https://ernest-gpcr.eu) and COST Action CM1207 GLISTEN. We thank R. Fonseca and A.J. Venkatakrishnan for their help implementing Flareplots into the GPCRmd toolkit. We also thank the volunteers of GPUGRID for donating their computing time for the simulations. M.T.-F. acknowledges financial support from the Spanish Ministry of Science, Innovation and Universities (FPU16/01209). T.M.S. acknowledges support from the National Center of Science, Poland (grant no. 2017/27/N/NZ2/02571). I.R.-E. acknowledges Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (2015 FI_B00145) for its financial support. X.D. and

R.G.-G. acknowledge support from the Swiss National Science Foundation (grant no. 192780). P.K. thanks the German Research Foundation DFG for the Heisenberg professorship grant nos. KO4095/4-1 and KO4095/5-1 as well as project KO4095/3-1 (funding M.M.-S.). G.D.F. acknowledges support from MINECO (Unidad de Excelencia María de Maeztu, funded by the AEI (CEX2018-000782-M) and BIO2017-82628-P) and FEDER and from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 823712 (CompBioMed2 Project). D.L. acknowledges support from the National Centre of Science in Poland (DEC-2012/07/D/NZ1/04244). P.W.H. thanks the DFG (Hi 1502, project number 168703014; SFB1423, project number 421152132, subproject Z04), the Stiftung Charité and the Einstein Foundation. S.F. thanks the National Science Centre Poland grant no. 2017/25/B/NZ7/02788. M.F. acknowledges support by the Office of Research Infrastructure of the National Institutes of Health under award numbers S10OD018522 and S10OD026880, as well as the Extreme Science and Engineering Discovery Environment (XSEDE) under MCB080077, which is supported by National Science Foundation grant number ACI-1548562. J.K.S.T. acknowledges support from HPC-EUROPA3 (INFRAIA-2016-1-730897) and the EC Research Innovation Action under the H2020 Programme. The work was supported by grants from the Swedish Research Council (2017-4676), the Swedish strategic research program eSSENCE and the Science for Life Laboratory to J.C. H.W. and G.K. acknowledge support from NSF grant no. 1740990 for In Situ Data Analytics for Next Generation Molecular Dynamics Workflows, and the 1923 Fund. D.E.G. acknowledges the Lundbeck Foundation (R163-2013-16327) and European Research Council (639125) for support. F.S. received support from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement number 802750 (FAIRplus) with the support of the European Union's Horizon 2020 Research and Innovation Programme and EFPIA Companies. The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), funded by ISCIII and FEDER (PT17/0009/0014). The DCEXS is a 'Unidad de Excelencia María de Maeztu', funded by the AEI (CEX2018-000782-M). The GRIB is also supported by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya (2017 SGR 00519). Finally, J.S. acknowledges financial

support from the Instituto de Salud Carlos III FEDER (PI15/00460 and PI18/00094) and the ERA-NET NEURON and Ministry of Economy, Industry and Competitiveness (AC18/00030).

Contributions

Conceptualization came from J.S., R.G.-G., I.R.-E. and M.T.-F. Database structure was designed by I.R.-E. GPCRmd workbench was by M.T.-F. with support from N.W., A.V.-R., J.K.S.T. and F.S. Meta-analysis tool was developed by D.A.-G. with support from M.T.-F. and I.R.-E. Submission system was developed by J.M.A.R. with support from I.R.-E. Query system was developed by A.V.-R. with support from I.R.-E. Server maintenance was performed by M.T.-F. Simulation standard protocol, original draft was written by R.G.-G. and J.S.; Revised simulation standard protocol was written by G.D.F., A.C., I.R.-E., J.C., H.G.-d.-T., W.J., M.M.-S., P.K., J.K.S.T., P.W.H., T.M.S., S.F., T.G. and M.J.-R. Protein curation and modeling of missing loops was done by G.P.-S. and D.E.G. Protein curation, placement of internal water molecules was done by E.M., J.K.S.T., P.W.H., R.G.-G., M.O. and A.C. Expert knowledge for final protein curation (for example, protonation states, disulfide bridges and so on) was done by I.R.-E., M.T.-F., J.K.S.T., D.A.G., J.M.R.-A., T.M.S., N.W., A.V.-R., A.M.-P., B.M.-L., G.P.-S., E.M., T.G., J.C., X.D., S.F., J.C.G.-T., A.G., H.G.-d.-T., M.J.-R., W.J., J.K., P.K., D.L., M.M.-S., P. Matricon, M.-T.M., P. Miszta, M.O., L.P.-B., S.R., I.R.T., J.S., A.S., S.V., P.W.H., G.D.F., F.S., D.E.G., A.C., R.G.-G. and J.S. Coordination of data exchange was done by T.M.S. Preparation of solvated receptor-membrane systems was done by I.R.-E. with support from T.M.S. MD simulation was done by G.D.F., I.R.-E. and B.M.-L. MD data curation and submission was done by I.R.-E., A.M.-P., M.T.-F., D.A.-G., T.M.S. and J.S. Individual contribution of MD data was given by G.K., H.W., U.Z., NV, D.P. and M.F. The original draft of the manuscript was written by R.G.-G., J.S. with input from I.R.-E., M.T.-F. and J.K.S.T. The manuscript was reviewed and edited by all authors with important contributions from D.E.G., T.G. and P.K. Project supervision and administration was done by J.S.

Ethics declarations

Competing interests

The authors declare no competing interests.

Supplementary information

Supplementary Note 1. Sustainability

Projects such as the GPCRmd face huge challenges to overcome in order to be accepted, used and sustained in the future. As previously shown, structural biology or (proteo-)genomics projects like the PDB (rcsb.org[1]) or the Galaxy project[2], respectively, have demonstrated the need for strong community support rather than individual laboratories to preserve adequate sustainability. Therefore, we created the GPCRmd consortium under the umbrella of the newly granted ERNEST network (GLISTEN COST action, https://ernest-gpcr.eu), with the support of the GPCR community. Additionally, the focus on a specific research area such as GPCRs rather than a general database for all MD simulations reduces hurdles like deposition space problems, too general and likely unused analysis tools and problems in the findability due to too broad keywords/labels. An important asset promoting sustainability is the automated update of GPCRmd with new MD simulations for newly published GPCR structures. These updates will follow a standardized protocol developed by the GPCRmd community. On the other hand, GPCRmd is designed to serve as a revision platform in the near future allowing the editor and reviewers to evaluate dynamics data on the fly, making them more transparent and trustful and allowing upon acceptance for an easy deposition into the database.

Supplementary Note 2. System set-up and simulation protocol

Protein structure preparation. The structures of simulated GPCRs were obtained from the Protein Data Bank (PDB) (rcsb.org[1]). Auxiliary proteins and long unresolved or truncated N- and C-terminal regions were removed and stabilizing mutations were reverted back to the native sequence. In receptors where ICL3 was longer than 10 residues, a chain break was introduced in the middle and five residues at each end were modeled using MODELLER[3]. Modelling and refinement of the structures was performed using the methodology described by Pándy-Szekeres et al[4]. Steric clashes were energy minimized using the MOE software[5]. Crystallographic waters and lipid modifications present in the X-ray structure were

preserved. The rest of the co-crystallized molecules were discarded. Receptor residue protonation and tautomeric states were assigned using PROPKA[6,7] as implemented in PDB2PQR[8] at pH 7 and subsequently curated by members of the GPCRmd consortium. Residue D2x50 was either kept deprotonated or protonated for antagonist- or agonist-bound receptor complexes, respectively.

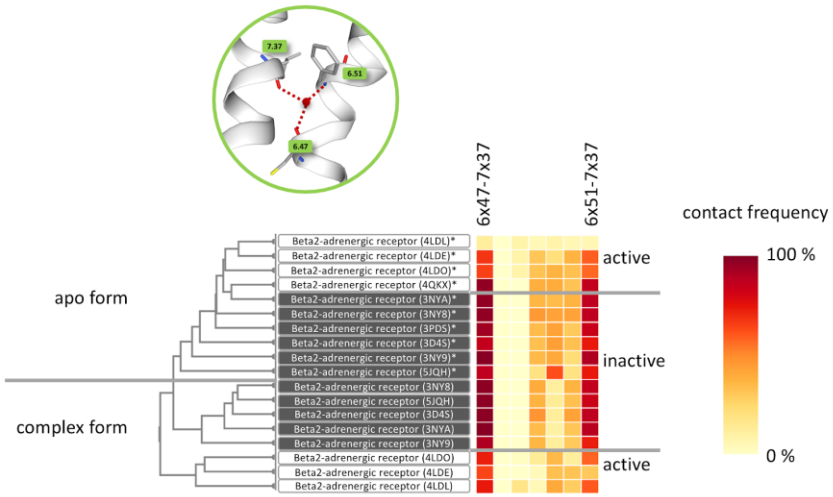
Ligand parameterization. Tripos Mol2 File Format files were taken from the PDB using the MOE software[5]. Protonation and tautomeric states at pH 7 were predicted and assigned using the Marvin Calculator Plugin[9]. The obtained Mol2 files were used to generate parameters by analogy using the ParamChem server 1.0.0 (https://cgenff.umaryland.edu) and CGenFF 3.0.1[10–13]. Parameters for inorganic phosphate (PO_4^{2-}) were obtained from CGenFF and SwissParam[14] web server. Finally, adenosine parameters where taken from the RNA CHARMM 36 force-field[15]

Placement of additional internal waters: For each structure we used HomolWat[16] (http://lmc.uab.cat/homolwat/) to incorporate internal water molecules not determined in other structures of the same or parent receptors. To this end, we used water molecules with a circular variance[19] > 0.6 measured using vectors from the oxygen atom of a water molecule to the surrounding atoms up to 10 Å. The algorithm uses *blastp* (ncbi-blast v2.6.0+)[18] to generate an ordered list of receptors with determined structures that contain resolved water molecules and subsequently tries to incorporate into the model all water molecules that do not clash (> 2.4 Å) with receptor atoms or previously introduced water molecules, starting with 1) receptors with the largest sequence identity, 2) structures with the best resolution and 3) water molecules with the smallest B-factors. Water hydrogens are added using the PDB2PQR software[8]. Non-coincident water molecules (distance > 2 Å) predicted by Dowser+ software[20] were also incorporated into the structure.

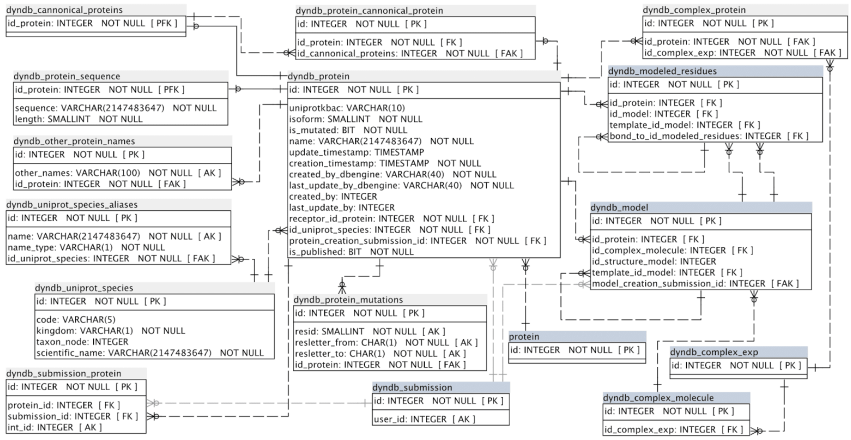
System set-up. Structures generated by the GPCRmd consortium were checked for protonation consistency. Acetylated and charged N terminus were used for incomplete and complete N terminus capping, respectively. Amidated and charged C terminus were used for incomplete and complete C terminus capping, respectively. Each GPCR model was aligned to its respective orientation taken from the Orientations of Proteins in Membranes database[21] using STAMP

4.4[22,23] . Receptors were then embedded into a POPC bilayer and solvated ensuring a 20 Å distance between protein periodic distances, considering also receptor diffusional rotation. Furthermore, the system was checked for lipids inserted into aromatic rings. Finally, systems were solvated with TIP3 water molecules and the ionic strength of the solution was adjusted to 0.15 M using NaCl ions. Parameters for the simulation were obtained from the CHARMM36m force field[24,25].

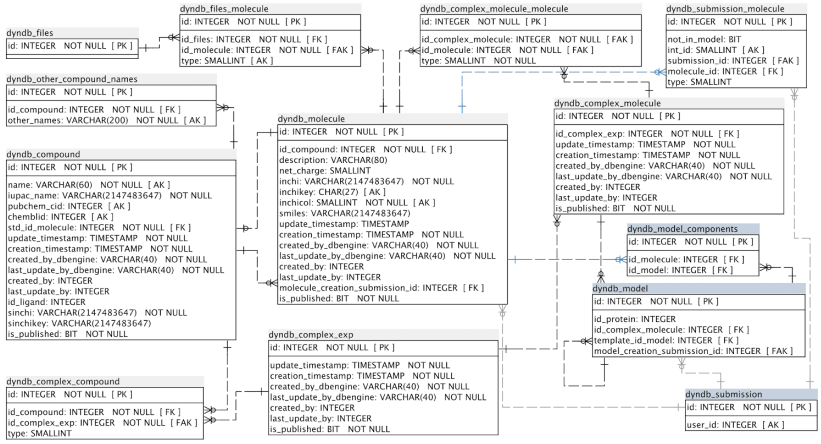
Molecular dynamics (MD) simulations. Systems were first energy minimized during 5000 step and then equilibrated at constant pressure (NPT, 1.01325 bar) using the Berendsen barostat[26] with a pressure relaxation time of 800 fs and a compressibility factor of $4.57x10^{-5}$ bar^{-1} during 30 ns. In a first step, harmonic restraints of 1.0 kcal/(mol·Å^2) were set on protein backbone and water oxygen atoms during 10 ns. Then, restrains were progressively released in a ramp of -0.095 kcal/(mol·Å^2·ns) during 10 ns followed by a restraints-free equilibration step of 10 ns. Production simulations were performed at constant volume (NVT) in 3 replicates of 500 ns per system using ACEMD[27] and GPUGRID[28]. Time-step of 2 and 4 fs were used during the equilibration and production runs, respectively. Non-bonded interactions were cut-off at 9 Å. A smooth switching function for the cut-off was applied, starting at 7.5 Å. Long-distance electrostatic forces were calculated using the Particle Mesh Ewald algorithm[29] with a grid spacing of 1 Å. Bond lengths of hydrogen atoms were kept constrained using the RATTLE algorithm[30]. All simulations were carried out at a temperature of 310K using the Langevin thermostat[31] with damping constants γ of 1 ps^{-1} and 0.1 ps^{-1} for NPT and NVT simulations, respectively.



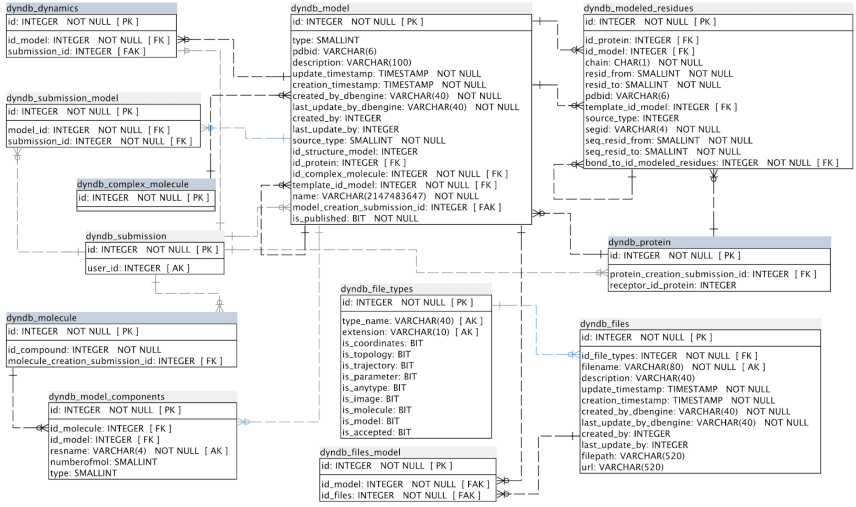
Supplementary Figure 1. Water-mediated contact map of the β_2AR highlights the bifurcated network between TM6 and TM7. Systems marked with an asterisk indicate apo-form simulations which are nicely separated from the ligand-receptor complex simulations by the clustering analysis. Within the clustered groups for apo- and complex forms, we see also a separation of active and inactive structures. The bifurcated network which links TM6 to TM7 can be seen in inactive as well as active structures. Interestingly, contact frequencies are reduced in active structures indicating a network loosening. The water-mediated contact map is computed over the accumulated simulation time of 1.5 µs (3 x 0.5 µs) per system.



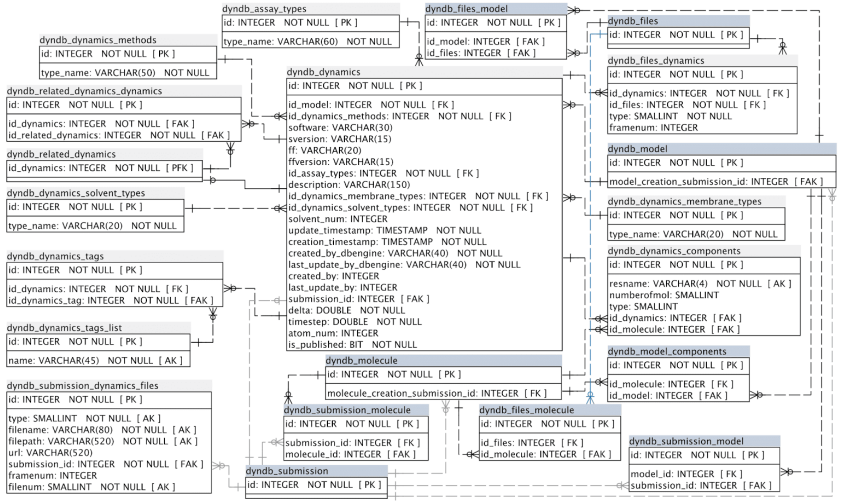
Supplementary Figure 2. GPCRmd entity-relationship (ER) diagram of entities related to protein objects. Tables in blue only display fields taking part in the relationships.



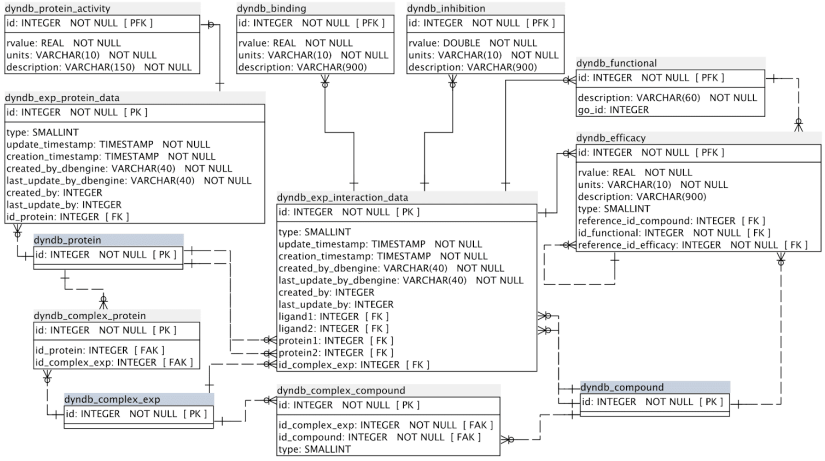
Supplementary Figure 3. GPCRmd ER diagram of entities related to molecule objects. Tables in blue only display fields taking part in the relationships.



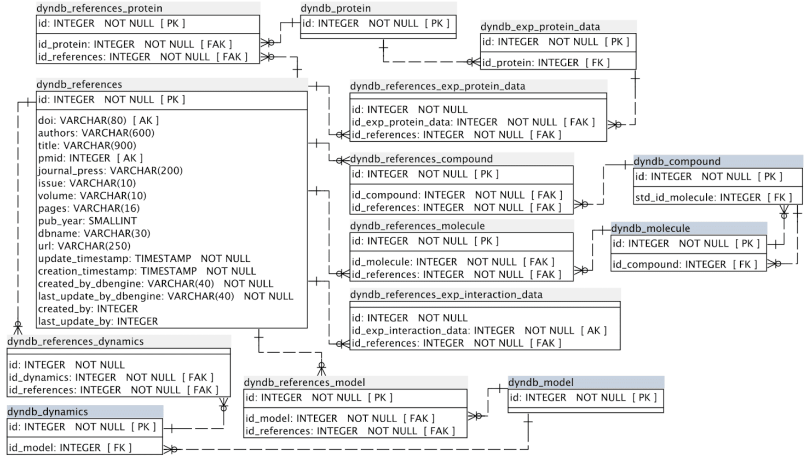
Supplementary Figure 4. GPCRmd ER diagram of entities related to model objects. Tables in blue only display fields taking part in the relationships.



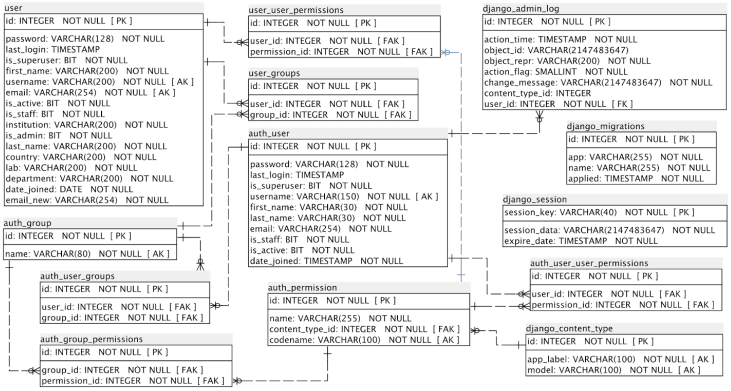
Supplementary Figure 5. GPCRmd ER diagram of entities related to dynamics objects. Tables in blue only display fields taking part in the relationships.



Supplementary Figure 6. GPCRmd ER diagram of entities related to experimental data. Tables in blue only display fields taking part in the relationships.



Supplementary Figure 7. GPCRmd ER diagram of entities related to reference objects. Tables in blue only display fields taking part in the relationships.



Supplementary Figure 8. GPCRmd ER diagram of user and Django tables.

References

1. Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000).
2. Boekel, J. et al. Multi-omic data analysis using Galaxy. Nat. Biotechnol. 33, 137–139 (2015).
3. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. Curr. Protoc. Bioinforma. 54, 5.6.1-5.6.37 (2016).
4. Munk, C. et al. GPCRdb in 2018: adding GPCR structure models and ligands. Nucleic Acids Res. 46, 440–446 (2017).
5. Chemical Computing Group ULC. Molecular Operating nvironment (MOE). (2014).
6. Søndergaard, C. R., Olsson, M. H. M., Rostkowski, M. & Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. J. Chem. Theory Comput. 7, 2284–2295 (2011).
7. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J. Chem. Theory Comput. 7, 525–537 (2011).
8. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. a. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. 32, W665–7 (2004).
9. ChemAxon. Marvin 14.11.24.0. (2014).
10. Vanommeslaeghe, K. et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J. Comput. Chem. 31, 671–690 (2010).
11. Yu, W., He, X., Vanommeslaeghe, K. & MacKerell, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. J. Comput. Chem. 33, 2451–2468 (2012).

12. Vanommeslaeghe, K. & MacKerell, a D. Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. J. Chem. Inf. Model. 52, 3144–54 (2012).
13. Vanommeslaeghe, K., Raman, E. P. & MacKerell, a D. Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. J. Chem. Inf. Model. 52, 3155–68 (2012).
14. Zoete, V., Cuendet, M. A., Grosdidier, A. & Michielin, O. SwissParam: A fast force field generation tool for small organic molecules. J. Comput. Chem. 32, 2359–2368 (2011).
15. Denning, E. J., Priyakumar, U. D., Nilsson, L. & Mackerell, A. D. Impact of 2′-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA. J. Comput. Chem. 32, 1929–1943 (2011).
16. Mayol, E. et al. HomolWat: a web server tool to incorporate "homologous" water molecules into GPCR structures. Nucleic Acids Res. (in press), doi: 10.1093/nar/gkaa440.
17. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 2.0.5.
18. Camacho, C. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009).
19. Mezei, M. A new method for mapping macromolecular topography. J. Mol. Graph. Model. 21, 463–472 (2003).
20. Morozenko, A., Leontyev, I. V. & Stuchebrukhov, A. A. Dipole Moment and Binding Energy of Water in Proteins from Crystallographic Analysis. J. Chem. Theory Comput. 10, 4618–4623 (2014).
21. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: Orientations of Proteins in Membranes database. Bioinformatics 22, 623–625 (2006).
22. Russell, R. B., Walsh, T. & Barton, G. STAMP Structural Alignment of Multiple Proteins Version 4.4.
23. Russell, R. B. & Barton, G. J. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. Proteins Struct. Funct. Bioinforma. 14, 309–323 (1992).
24. Huang, J. et al. Charmm36M: an Improved Force Field for Folded and Intrinsically Disordered Proteins. Nat. Methods (2016) doi:10.1038/nMeth.4067.
25. Klauda, J. B. et al. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. J. Phys. Chem. B 114, 7830–7843 (2010).
26. 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. 81, 3684 (1984).
27. Harvey, M. J., Giupponi, G. & Fabritiis, G. D. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. J. Chem. Theory Comput. 5, 1–9 (2009).
28. Buch, I., Harvey, M. J., Giorgino, T., Anderson, D. P. & De Fabritiis, G. High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing. J. Chem. Inf. Model. 50, 397–403 (2010).

29. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. J. Chem. Phys. 98, 10089 (1993).
30. Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. J. Comput. Chem. 13, 952–962 (1992).
31. Loncharich, R. J., Brooks, B. R. & Pastor, R. W. Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N′-methylamide. Biopolymers 32, 523–535 (1992).

3.5. SCoV2-MD: a database for the dynamics of the SARS-CoV-2 proteome and variant impact predictions

Since the emergence of the COVID-19 disease, unveiling the structural basis of SARS-CoV-2 infection has been a key priority. Thus, numerous groups have generated MD simulations to study the structural dynamics of the viral proteins. However, these data are usually hosted at disparate sites, hardly discoverable, and not amenable to systematic analyses. Here, we present SCoV2-MD (www.scov2-md.org), an online resource with the objective to organize, cross-reference, and share MD dynamics data and metadata of the SARS-CoV-2 proteome. It includes interactive visualization and analysis tools that provide a rapid and intuitive way to explore the simulation data. An important asset of SCoV2-MD is that it also provides tools to interrogate the functional impact of variant-associated mutations. For that, it integrates the MD data with available information on SARS-CoV-2 variants sequenced during the pandemic. The impact of each mutation can then be analyzed interactively by combining static (e.g. a variety of amino acid substitution penalties) and dynamic (time-dependent data derived from the MD simulations) descriptors. All in all, SCoV2-MD is a cross-disciplinary database that not only promotes the reproducibility and transparent dissemination of SARS-CoV-2-related simulations, but also allows the investigation of questions on the interplay between the structural dynamics of the viral proteome and viral variants' phenotypes.

Torrens-Fontanals, M. et al. SCoV2-MD: a database for the dynamics of the SARS-CoV-2 proteome and variant impact predictions. *Nucleic Acids Res.* 1, 13–14 (2021). doi: 10.1093/nar/gkab977

SCoV2-MD: a database for the dynamics of the SARS-CoV-2 proteome and variant impact predictions

Mariona Torrens-Fontanals[1], Alejandro Peralta-García[1], Carmine Talarico[2], Ramon Guixà-González[3,4], Toni Giorgino[5,6,*] and Jana Selent[1,*]

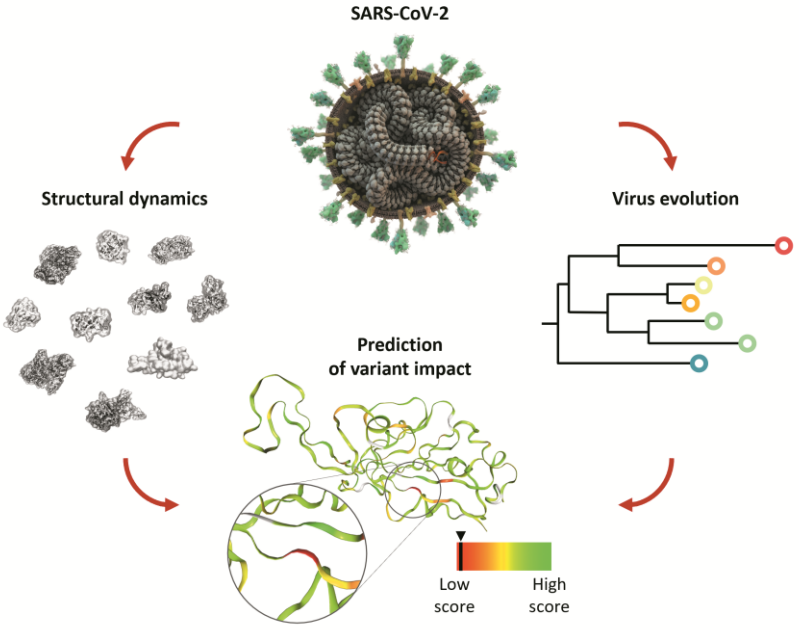
1. Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute—Department of Experimental and Health Sciences, Pompeu Fabra University, Barcelona, 08003, Spain
2. Dompé Farmaceutici SpA, Via Campo di Pile, 67100 L'Aquila, Italy
3. Laboratory of Biomolecular Research, Paul Scherrer Institute, CH-5232 Villigen PSI, Switzerland
4. Condensed Matter Theory Group, Paul Scherrer Institute, CH-5232 Villigen PSI, Switzerland
5. Biophysics Institute, National Research Council of Italy, Milan, 20133, Italy
6. Department of Biosciences, University of Milan, Milan, 20133, Italy

* To whom correspondence should be addressed.

Abstract

SCoV2-MD (www.scov2-md.org) is a new online resource that systematically organizes atomistic simulations of the SARS-CoV-2 proteome. The database includes simulations produced by leading groups using molecular dynamics (MD) methods to investigate the structure-dynamics-function relationships of viral proteins. SCoV2-MD cross-references the molecular data with the pandemic evolution by tracking all available variants sequenced during the pandemic and deposited in the GISAID resource. SCoV2-MD enables the interactive analysis of the deposited trajectories through a web interface, which enables users to search by viral protein, isolate, phylogenetic attributes, or specific point mutation. Each mutation can then be analyzed interactively combining static (e.g. a variety of amino acid substitution penalties) and dynamic (time-dependent data derived from the dynamics of the local geometry) scores. Dynamic scores can be computed on the basis of nine non-covalent interaction types, including steric properties, solvent accessibility, hydrogen bonding, and other types of chemical interactions. Where available, experimental data such as antibody

escape and change in binding affinities from deep mutational scanning experiments are also made available. All metrics can be combined to build predefined or custom scores to interrogate the impact of evolving variants on protein structure and function.



1. Introduction

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the causative agent of the Coronavirus disease 2019 (COVID-19), which already accounts for more than 4.2 million deaths globally, as of 10th August 2021 (WHO, Coronavirus (COVID-19) Dashboard, covid19.who.int). The diffusion of the COVID-19 pandemic has produced emerging variants (1), which have been tracked through massive sequencing efforts at an unprecedented rate soon surpassing that of any other pathogen and phylogenetic analysis (2–4). Thus, as of July 2021, almost 2 million full genomes are available via the Global Initiative on Sharing All Influenza Data (GISAID), one of the main pandemic genome databases (3, 5).

SARS-CoV-2 is a single-stranded RNA beta-coronavirus enveloped by an outer membrane and expressing 16 non-structural, 9 accessory, and 4 structural proteins (Figure 1). While the spike, membrane, and envelope structural proteins are embedded in the membrane and involved in cell recognition and entry, one structural protein, the nucleocapsid, interacts inside the membrane with viral RNA to form a ribonucleoprotein complex that works as a scaffold for genome replication and virion assembly. The four structural proteins make up approximately one third of the viral genome (6). The remaining two-thirds of the viral genome encodes for the non-structural proteins (nsp) 1 to 16 (Figure 1). Some nsps are critical enzymes for virus replication such as proteases (nsp3, nsp5), RNA-dependent RNA polymerases (consisting of nsp7, two copies of nsp8, and nsp12), the RNA helicase (nsp13), and the proofreading exonuclease (nsp14) (7).

Unveiling the structural basis of SARS-CoV-2 infection has been a key priority since the emergence of the COVID-19 disease. In the wake of the increased availability of structural information of SARS-CoV-2 proteins, numerous groups have tackled the study of SARS-CoV-2 proteins using molecular dynamics (MD) simulations, often after Herculean modeling and computational efforts, with the goal of supporting pandemic response efforts (8–10). Obtained MD data are highly relevant to understand the functional dynamics of the viral proteome which cannot often be deduced from the static structure that has been experimentally solved. In addition, it can help rationalize the structural/functional impact of sequence variability in the viral proteome. This is particularly useful when the relationship between mutation location and activity is not obvious (e.g. the mutation is distant from the protein's active center).

However, while computational scientists are urgently aware of the need to share the resulting data (11), these are usually hosted at disparate sites, hardly discoverable, and not amenable to systematic analysis. In practice, this limits the ability of computational structural biologists to reuse these trajectories in large-scale efforts, e.g. for dynamic docking (12, 13), discovering transient pockets (14), or associating them with phenotypes (15).

Here, we present SCoV2-MD (www.scov2-md.org), a cross-disciplinary database developed to investigate diverse questions on the interplay between the structural biology of the viral 3D proteome, its dynamics, and viral variants' phenotypes. The platform focuses on the dynamics of protein non-covalent contacts, supporting the interpretation of allostery, the exploration of individual subunits, interfaces, and protein-ligand contacts, and the mapping of external information. An important asset of SCoV2-MD is that it provides tools to interrogate the impact of variant substitutions using a combination of static and time-resolved structural descriptors.

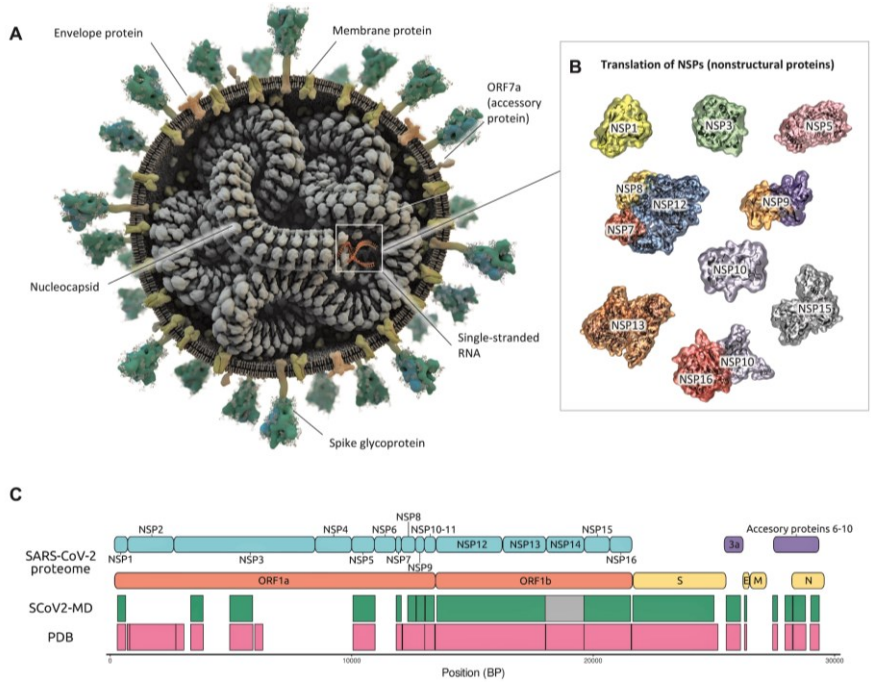


Figure 1. Structure-based overview of the SARS-CoV-2 proteome and protein-based entry point to the database. (a) Three-dimensional model of the virion, displaying the structural proteins' assembly; **(b)** Three-dimensional models of the available nsp proteins. The diagrams provide a unified overview of the available viral structures and provide one of the entry points for browsing the proteome. Selecting the proteins leads to a list of the related simulation data. **(c)** Coverage of SARS-CoV-2 proteome plotting experimentally available structures (pink) and simulated proteins in SCoV2-MD (green). Nsp14 (grey) was simulated based on a theoretical model using a structure of SARS-CoV-1 nsp14 (PDB ID 5C8S) as template.

So far, numerous resources have been dedicated to track, monitor, and classify the phylogenetic diversity of SARS-CoV-2, such as GISAID (16), ViruSurf (17), or PANGO lineage (18). Other efforts aimed at predicting potential functional effects of SARS-CoV-2 variants based on evolutionary considerations (19), static structures of proteins and complexes, or experimental data (e.g. MutFunc (20), COV3D (21), and COVID-3D (22)). Our database builds upon previous approaches, extending them through the integration of time-resolved MD data with available information on variants to improve the prediction of their potential functional impact.

2. Materials and methods

2.1. Data source overview

The SCoV2-MD platform includes, at the time of manuscript preparation, simulations of more than 250 different systems, covering all SARS-CoV-2 proteins with known structure. A large part of the simulations is collated from public databases, mainly BioExcel-CV19 (https://bioexcel-cv19.bsc.es/), COVID-19 Molecular Structure and Therapeutics Hub (https://covid.molssi.org/), the CHARMM-GUI simulation archive (23), and the Exscalate4Cov project (https://www.exscalate4cov.eu/). Additionally, we generated in-house simulations to achieve complete coverage of all SARS-CoV-2 proteins with known structures (Supplementary Note S1 and Supplementary Table S1). Individual researchers can contribute their simulation data upon request. The whole dataset is accessible for free and without registration at www.scov2-md.org.

2.2. Database schema and infrastructure

The data model of the database (Supplementary Figure S1) is based on five main entities, namely: *protein* objects, identified by their sequence and their relationship with UniprotKB entries; *final protein* objects, representing the viral proteins after the transcribed poly-proteins are cleaved by its proteinases; *model* objects, describing the three-dimensional structures identified by the Protein Data Bank (PDB, rcsb.org) (24) identifier; *dynamics* objects representing the MD simulations; and *dynamics components* with details of the molecules in the simulated systems. The database

integrates experimental data from GISAID (16) and Mutfunc: SARS-CoV-2 (20).

Similar to the GPCRmd platform (25), the Workbench page of SCoV2-MD builds on a WebGL-based structure viewer, NGL version 2.0.0, (26, 27) with the MDsrv 0.3.5 (28) backend, which allows efficient streaming and sharing of trajectories online. Intuitive selection capabilities enable the creation of various 3D representations using the NGL selection language (27). The Workbench integrates annotation data from UniprotKB (29) and variant data from GISAID (16). UniprotKB data is used to extract domains and annotations of the protein represented and map them to the structure, while GISAID (16) data is referenced to display and cross-reference known protein variants on the structure. The SCoV2-MD database and web interface are based upon the Django Web Framework (v.1.9), PostgreSQL (v.9), Python (v.3.4), and JavaScript libraries jQuery 1.9, jQuery UI 1.11.2.

2.3. Variant impact scoring

The database's web portal provides the *Variant Impact toolkit* for interactively referencing the MD simulations with SARS-CoV-2 sequences obtained from GISAID (16), which can be found in the Workbench page. Each of the variants is annotated with the corresponding static (mutation-dependent) and time-dependent (computed on the basis of the simulation dynamics) descriptors of their impact on multiple aspects of the protein's structure and dynamics and the viral function (Figure 4 and Supplementary Note S2). The descriptors are further combined in an *impact score,* defined as a weighted sum of descriptors:

$$impact\ score = \sum_{i}^{N_{desc}} v_i w_i$$

where N_{desc} is the number of descriptors, v_i the value of descriptor i, and w_i the corresponding weight. Users may assign either predefined (see next section) or custom weight combinations to the descriptors to reflect various aspects of the structural impact of the variant. The obtained score is presented together with a q value, showing its normalized rank in the distribution of impact scores of

all the variant-associated substitutions occurring in residues modeled in the simulation considered (e.g. in a simulation of the receptor-binding domain (RBD) of the spike protein, an histogram of impact scores is built on the basis of the amino-acid substitutions associated to known variants located in the RBD). In other words, q=0, 0.5, and 1 respectively mean that the selected amino acid variant achieves the minimum, median and maximum effect score with respect to the other variants observed in the sequence of the simulated protein.

2.4. Model-based predictions of variant impact for the spike protein's receptor-binding domain (RBD)

We developed simple predictive models able to qualitatively estimate the impact of each variant, in terms of (a) the change in *binding affinity* between SARS-CoV-2's spike RBD and the angiotensin-converting enzyme 2 (ACE2) receptor (30); (b) the change in *expression* of the RBD on yeast cells (30) and (c) the potential for *antibody evasion* (31). The three models are based on regularized regression models (LASSO) (Supplementary Note S3, Supplementary Table S2) trained on 23 per-variant covariates, used as predictors, computed based on the 12 MD trajectories containing the RBD available at the time of writing. Each of the models was trained to fit the corresponding quantity, measured experimentally per-variant in deep scanning mutagenesis experiments. The three pre-computed models can be enabled via the web interface, in the *Variant Impact* section of the Workbench page, with buttons that load the corresponding sets of coefficients in the "weight" sliders.

3. Results & discussion

3.1. Tracking of structural virus evolution

A key aspect in understanding the diversity and impact of SARS-CoV-2 is monitoring the emergence of variants. Since the outbreak in 2019, many viral mutations have occurred. Some of them reach high regional frequencies with the ability to rapidly spread worldwide and potentially evade immunization and antibody treatment (e.g. B.1.1.7 (Alpha), B.1.351 (Beta), B.1.617.2 (Delta), and P.1 (Gamma) variants). Our database provides a visualization tool for viral phylogenetic data from the worldwide collaborative

effort hosted at GISAID (16). This tool allows viral infections to be tracked by region, gender, age, etc. as well as following the emergence of mutations as a function of time (Figure 2). Each small circle is interactive and represents one sequenced viral genome. When clicking on a viral genome that includes novel variants/mutations, the user can directly investigate the variant location on the 3D structure of implicated viral proteins as well as stream time-resolved dynamics on the fly (see next section).

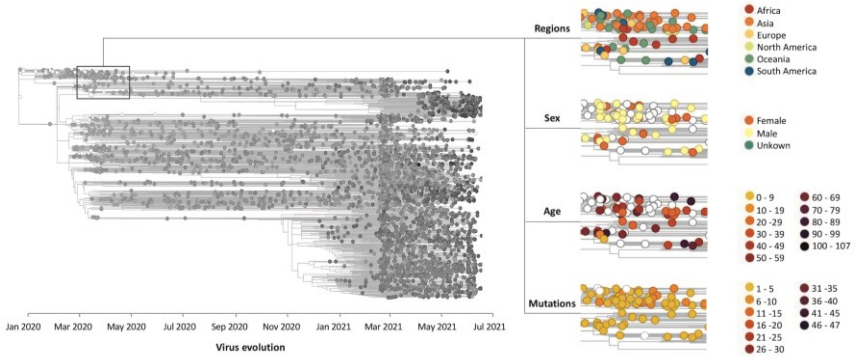


Figure 2. Phylogenetic tree of SARS-CoV-2 viral evolution. Sequenced samples (isolates) are mapped onto the tree. Each interactive circle represents one sequenced viral genome and is linked to the simulations of the proteins mutated with respect to the reference sequence. The rectangular inset (left) tags a subset of isolates with some of the available descriptors (regions, sex, age and mutations).

3.2. Structural dynamics of the viral 3D proteome

3.2.1. Visualize and stream viral proteins' dynamics

The database of time-resolved dynamics of SARS-CoV-2 currently consists of over 250 simulations covering the entire viral 3D proteome (i.e. experimentally solved 3D structures) with at least one simulation entry for each protein (Figure 3A). Data entries have been either simulated by us or collected from public resources. The user can intuitively select a simulation of interest from an interactive graphical representation of the SARS-CoV-2 virus (Figure 3B) or from a dedicated search tool (Figure 3C). Once a simulation has been selected, one can easily view (Figure 3D) and modify its graphical representation using either the quick or customized selection (Figure 3E). In addition, we implemented the option to highlight domains relevant for protein function (e.g.

binding motif of the spike protein to the human ACE2 host cell receptor) which have been retrieved from the Uniprot database (Figure 3E). The most important merit of the SCoV2-MD resource is that viral proteins are not static, instead one can stream the time-resolved dynamics at atomistic resolution on-the-fly using the simulation viewer (Figure 3D).

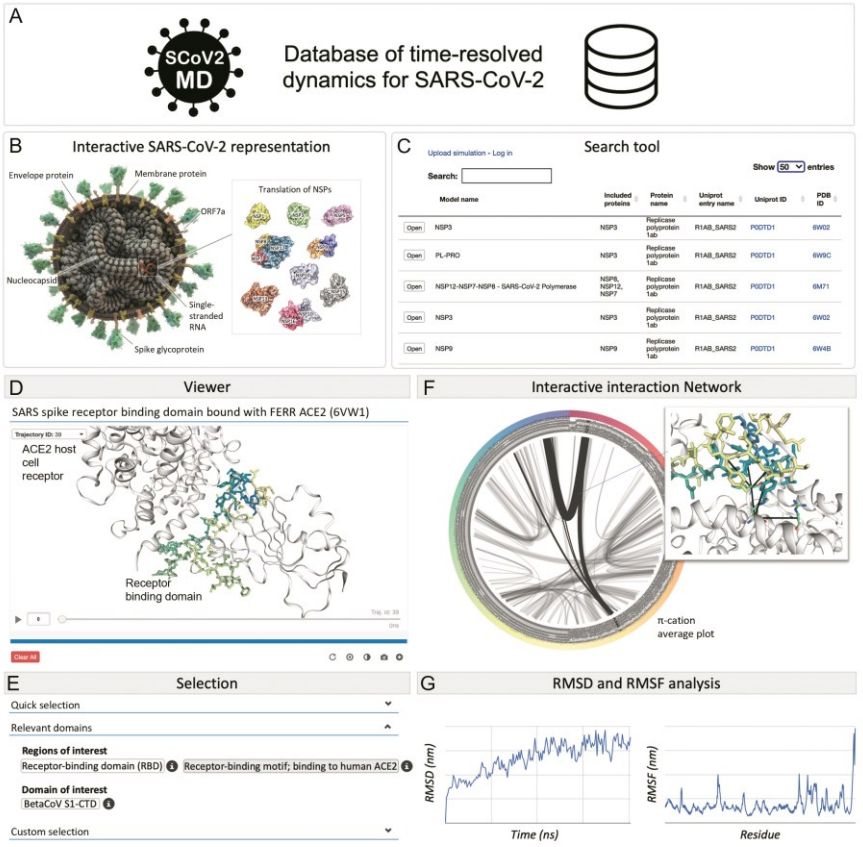


Figure 3. Structural dynamics of the SARS-CoV-2 proteome. (a) SCoV2-MD is a database of time-resolved dynamics for SARS-CoV-2 proteins. **(b)** The user can intuitively select a simulation of interest from an interactive graphical representation of the SARS-CoV-2 virus, or **(c)** use our dedicated search tool. **(d)** The viewer module enables interactive visualization and streaming of the MD simulations. **(e)** For that, we provide a selection panel including quick and custom selection. **(f)** Our platform includes interactive analysis tools such as an interaction network displaying intermolecular and intramolecular contacts. **(g)** The user can also check common structural stability metrics such as RMSD and RMSF.

3.2.2. General analysis of the wild-type (WT) simulation

SCoV2-MD provides analysis tools that allow for an overall understanding of the structural dynamics for a specific protein (Supplementary Note S4). This includes intramolecular interaction networks including hydrogen bonds, π-stacking, and T-stacking, among others, which can be shown for the current frame or as an average over the entire trajectory. In the average mode, line thickness indicates the contact strength, which enables quick identification of relevant contacts. For instance, one can easily identify the non-covalent contacts (electrostatics, dispersion effects, etc.) corresponding to π-cation interactions in the binding interface between the RBD from the spike protein and the ACE2 host cell receptor (Figure 3F). Moreover, a RMSD (root mean squared deviation) plot shows the evolution of the viral protein along the trajectory with respect to the initial structure (Figure 3G). Furthermore, a RMSF (root mean square fluctuation) plot provides a quick overview of stable and highly flexible regions in the protein (Figure 3G).

3.3. Variant descriptors and phenotype analysis tool

An important goal of SCoV2-MD is the ability to predict the impact of variant substitution on the viral proteome based on static and time-resolved descriptors via an "impact score" (Figure 4). The database integrates over 30 static and time-dependent descriptors (Supplementary Note S2). Static descriptors reflect substitution (e.g. BLOSUM, charge differences, and so on), conservation (frequency, SIFT score (32), etc.), structural impacts (post-translational modification, surface accessibility, etc.) and, if available, experimental observations (e.g. antibody escape, binding affinity, expression changes). Conservation, structural and experimental descriptors are collected from the Mutfunc database (20). Time-dependent descriptors are extracted from atomistic simulations and include RMSF and a large variety of non-covalent contact types (such as Van der Waals, hydrogen bonding, etc.) among others. They are computed on the basis of the MDtraj (33) and GetContacts libraries (34). The list of descriptors provides the user with a rich repertoire to score and interrogate variant substitutions on-the-fly.

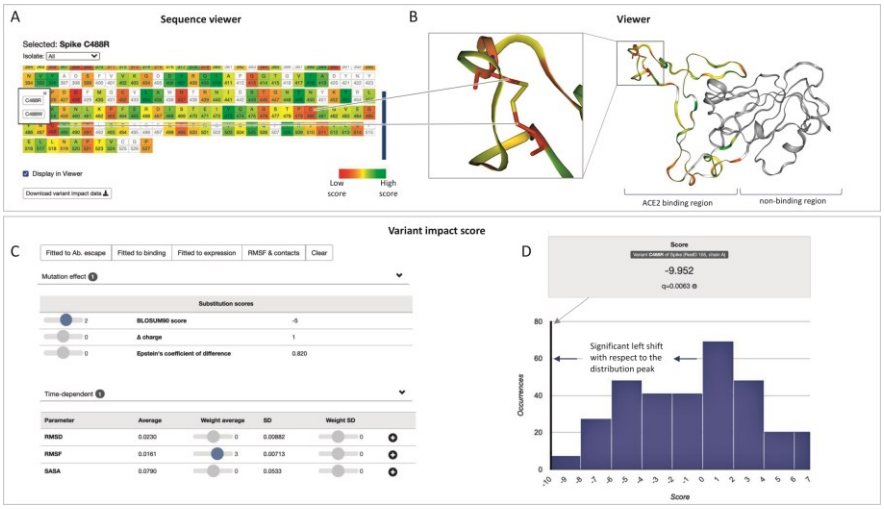


Figure 4. Prediction of variant impact. (a) Sequence viewer with color-coded impact score across the protein sequence and information on reported variants. The sequence viewer is interactive in two ways: (i) the color-coded scale updates on the fly with the defined impact score, and (ii) selected residues are automatically shown in **(b)** the 3D viewer of the structural dynamics. **(c)** Variant impact score. The user can combine over 30 different descriptors into a user-defined score. We also provide default scores, such as RMSD, contacts, or scores fitted to experimental data. **(d)** The impact score for a specific variant is highlighted within the score distribution for the entire protein. A significant shift of the variant score to the right or left from the distribution peak reflects a variant with a high propensity to disturb protein function based on the selected descriptor combination.

3.4. Search for variant substitutions with impact on protein function

A basic search for critical regions impacted by variant substitutions can be as follows. Regions of high structural stability are expected to be crucial for the overall function of the viral protein. Unfavorable variant substitutions (e.g. Cys to Arg) in these regions can significantly disturb protein stability. To detect a combination of such events in the protein, the user needs to 'turn on' RMSF in addition to any of the provided substitution scores (e.g. BLOSUM) in the impact score panel (Figure 4C). The user-defined impact score is plotted across the entire protein in the sequence highlighting regions that are predicted to be affected by reported variant substitutions (orange to red, Figure 4A). In our example (https://submission.gpcrmd.org/covid19/29/), we observe hotspots

of high stability with unfavorable substitutions in C488 in the RBD of the spike protein. Interestingly, structural visualization in the MD viewer reveals that C488 forms a disulfide bridge with C480. Without a doubt, unfavorable variant substitutions such as C488R will disrupt the disulfide bridge introducing flexibility into this region. This in turn can alter the propensity of the SARS-CoV-2 virus to attach to the ACE2 host cell receptor.

Finally, the user can validate the customized impact score for a C488R substitution in the context of all reported substitutions in the viral protein of interest (the score distribution across the entire protein is shown on Figure 4D). Overall, high or low impact scores that are significantly shifted from the distribution peak can be expected to significantly alter protein function. Once such variants have been detected, experimental validation is required to determine if the function of the viral protein is enhanced or diminished by the structural alterations.

3.5. Case study: assessing the impact of SARS-CoV-2 variability on drug binding

An important viral threat is that newly emerging variants can develop resistance against antiviral agents or antibodies. Sites of high mutational frequency 'in' or 'adjacent to' the binding sites of antibodies/antiviral agents in the viral proteome can impact the therapeutic efficacy. In a case study, we interrogated one of these highly variable positions (T24) located in the SARS-CoV-2 main protease, Mpro (3CLpro), with around 1000 detected cases to date (35). Position 24 is adjacent to the binding site of the protease inhibitor ML188 (Figure 5) with an antiviral SARS-CoV-2 inhibition activity at micromolar range (2.5 μM) (36, 37). T24A (T3287A in orf1a) is also a characteristic mutation of lineage B.1.524 (Malaysian strain, variant of concern, which peaked around November 2020) (B.1.524 Lineage Report, outbreak.info). We have simulated the WT (https://submission.gpcrmd.org/covid19/255/) and the T24A mutant (https://submission.gpcrmd.org/covid19/257/) in complex with the protease inhibitor ML188 for 1 μs in three replicates (Supplementary Note S1 and Supplementary Table S1). Interestingly, we observe that the ML188 inhibitor unbinds in both the WT as well as the T24A mutant (Supplementary Note S5, Supplementary Table S3). One important finding is that the ML188

inhibitor visits along its unbinding pathway an intermediate state that is in contact with position 24 (Figure 5). Therefore, we can expect that structural alteration of position 24 from a threonine to an alanine will alter (un)binding kinetics. In fact, we observe that T24A tends to unbind at shorter time scales compared to the WT (Supplementary Table S3). Of note, another polar-to-hydrophobic mutation at the same site, T24I, is present in ~43% of samples in the C.1.2 strain, which peaked around July 2021, hinting at a selective advantage (C.1.2 Lineage Report, outbreak.info). This example highlights the relevance of variability within the viral proteome for drug action but can also serve as a guide for the rational design of antiviral drugs/antibodies that are more resistant to virus evolution by avoiding these regions.

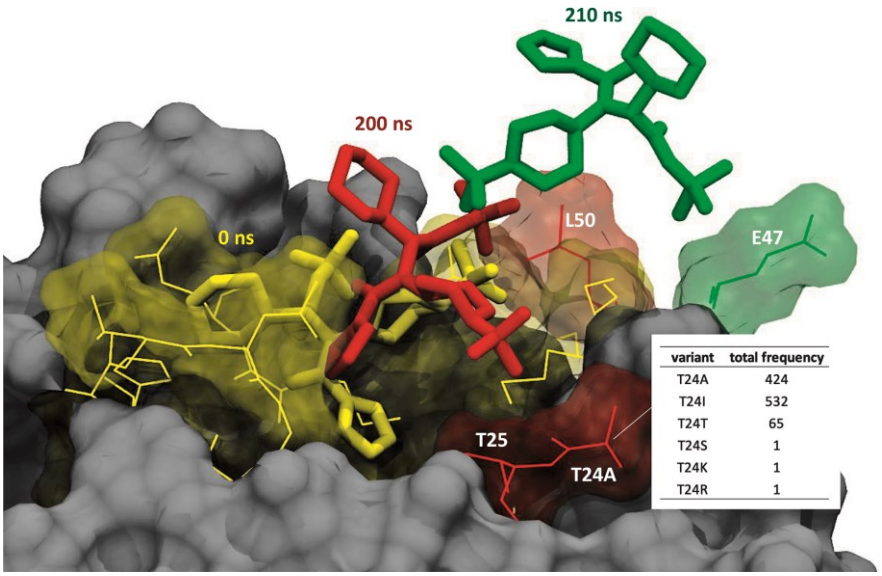


Figure 5. Unbinding pathway of the ML188 inhibitor from the SARS-CoV-2 main protease, Mpro (3CLpro). The inhibitor is in its crystallized binding pose at time T=0 ns (yellow). The inhibitor leaves its original binding pose towards an intermediate state which is in contact with the mutated position T24A at T=200 ns (red). From here, it moves to a second intermediate state with contacts to E47 at T=201 ns (green) before it completely unbinds.

4. Conclusion

Enormous research efforts have resulted in high-resolution structural information on most of the SARS-CoV-2 3D-proteome,

widely accessible via the PDB (rcsb.org) (24). The experimental techniques employed, namely X-ray crystallography and cryogenic electron microscopy, provide structural data which has been an excellent starting point for further efforts launched using MD simulations to gather time-resolved information about the functional dynamics of viral proteins. The SCoV2-MD database has the objective to organize, cross-reference, and share MD dynamics data for the entire viral proteome. In particular, interactive streaming and analysis tools provide a rapid and intuitive way to explore viral protein flexibility, and enable checking of hypotheses on the fly (19). Importantly, such dynamics data is not only of high value for understanding protein function but also allows for an improved insight on the structural determinants of variant impact as demonstrated in this work. We expect to periodically update the SCoV2-MD with new simulation data for multimeric complexes and missing regions in the viral proteome when they are experimentally solved and released.

Data availability

SCoV2-MD is an open-source collaborative initiative. The database is freely accessible without registration at www.scov2-md.org. Its source code is available in the GitHub repository https://github.com/GPCRmd/SCoV2-md.

Acknowledgements

We are grateful to the groups who shared and made reusable their simulations, including Vito Genna (IRB Barcelona), Adam Hospital (IRB Barcelona), the Chang group (University of California, Riverside), D. E. Shaw Research, Dmitry Morozov (University of Jyväskylä, Finland), Dr. Im's research team (Lehigh University, Bethlehem), the Amaro Lab (University of California San Diego), the Onufriev group (California State University, Los Angeles and Virginia Tech). We are grateful to the Coronavirus Structural Task Force (www.insidecorona.net) for making available curated protein structures, which we used in our simulations. We would like to thank the GISAID Initiative and are grateful to all of the data contributors, i.e. the Authors, the Originating laboratories responsible for obtaining the specimens, and the Submitting

laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We thank scistyle.com for the SARS-CoV-2 virion rendering. We are grateful to the anonymous reviewers of NAR for their extensive and thoughtful suggestions.

Funding

This work was supported by the Spanish Ministry of Science, Innovation and Universities [FPU16/01209 to M.T.-F.]; the Health department of the Government of Catalonia [DGRIS to A.P.G. and J.S.]; the EU H2020-SC1-PHE-CORONAVIRUS-2020 call [101003551 "EXaSCale smArt pLatform Against paThogEns for Corona Virus-Exscalate4CoV" to C.T.]; CINECA for the availability of high performance computing resources and support under the ISCRA initiative [GSNSDr3 and GSNSDr2 to T.G.]; and the Swiss National Science Foundation [192780 to R.G.-G.].

Conflict of interest

None to declare.

References

1. Harvey,W.T., Carabelli,A.M., Jackson,B., Gupta,R.K., Thomson,E.C., Harrison,E.M., Ludden,C., Reeve,R., Rambaut,A., Peacock,S.J., *et al.* (2021) SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.*, **19**, 409–424.
2. Hadfield,J., Megill,C., Bell,S.M., Huddleston,J., Potter,B., Callender,C., Sagulenko,P., Bedford,T. and Neher,R.A. (2018) Nextstrain: real-time tracking of pathogen evolution. *Bioinforma. Oxf. Engl.*, **34**, 4121–4123.
3. Plessis,L. du, McCrone,J.T., Zarebski,A.E., Hill,V., Ruis,C., Gutierrez,B., Raghwani,J., Ashworth,J., Colquhoun,R., Connor,T.R., *et al.* (2021) Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*, **371**, 708–712.
4. Hodcroft,E.B., Zuber,M., Nadeau,S., Vaughan,T.G., Crawford,K.H.D., Althaus,C.L., Reichmuth,M.L., Bowen,J.E., Walls,A.C., Corti,D., *et al.* (2021) Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature*, **595**, 707–712.
5. Shu,Y. and McCauley,J. (2017) GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, **22**, 30494.
6. Payne,S. (2017) Family Coronaviridae. In *Viruses*. Elsevier, pp. 149–158.

7. Denison,M.R., Graham,R.L., Donaldson,E.F., Eckerle,L.D. and Baric,R.S. (2011) Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol.*, **8**, 270–279.
8. Casalino,L., Dommer,A.C., Gaieb,Z., Barros,E.P., Sztain,T., Ahn,S.-H., Trifan,A., Brace,A., Bogetti,A.T., Clyde,A. et al. (2021) AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics. *Int. J. High Perform. Comput. Appl.,***35**, 432-451.
9. Yu,A., Pak,A.J., He,P., Monje-Galvan,V., Casalino,L., Gaieb,Z., Dommer,A.C., Amaro,R.E. and Voth,G.A. (2021) A multiscale coarse-grained model of the SARS-CoV-2 virion. *Biophys. J.*, **120**, 1097–1104.
10. Zimmerman,M.I., Porter,J.R., Ward,M.D., Singh,S., Vithani,N., Meller,A., Mallimadugula,U.L., Kuhn,C.E., Borowsky,J.H., Wiewiora,R.P., *et al.* (2021) SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nat. Chem.*, **13**, 651–659.
11. Amaro,R.E. and Mulholland,A.J. (2020) A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19. *J. Chem. Inf. Model.*, **60**, 2653–2656.
12. Gioia,D., Bertazzo,M., Recanatini,M., Masetti,M. and Cavalli,A. (2017) Dynamic Docking: A Paradigm Shift in Computational Drug Discovery. *Mol. Basel Switz.*, **22**, E2029.
13. Basciu,A., Malloci,G., Pietrucci,F., Bonvin,A.M.J.J. and Vargiu,A.V. (2019) Holo-like and Druggable Protein Conformations from Enhanced Sampling of Binding Pocket Volume and Shape. *J. Chem. Inf. Model.*, **59**, 1515–1528.
14. Yuan,J.-H., Han,S.B., Richter,S., Wade,R.C. and Kokh,D.B. (2020) Druggability Assessment in TRAPP Using Machine Learning Approaches. *J. Chem. Inf. Model.*, **60**, 1685–1699.
15. Cagiada,M., Johansson,K.E., Valanciute,A., Nielsen,S.V., Hartmann-Petersen,R., Yang,J.J., Fowler,D.M., Stein,A. and Lindorff-Larsen,K. (2021) Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance. *Mol. Biol. Evol.*, **38**, 3235–3246.
16. Elbe,S. and Buckland-Merrett,G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.*, **1**, 33–46.
17. Canakoglu,A., Pinoli,P., Bernasconi,A., Alfonsi,T., Melidis,D.P. and Ceri,S. (2021) ViruSurf: an integrated database to investigate viral sequences. *Nucleic Acids Res.*, **49**, D817–D824.
18. Rambaut,A., Holmes,E.C., O'Toole,Á., Hill,V., McCrone,J.T., Ruis,C., du Plessis,L. and Pybus,O.G. (2020) A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.*, **5**, 1403–1407.
19. Lubin,J.H., Zardecki,C., Dolan,E.M., Lu,C., Shen,Z., Dutta,S., Westbrook,J.D., Hudson,B.P., Goodsell,D.S., Williams,J.K., *et al.* (2020) Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first six months of the COVID-19 pandemic. *bioRxiv*, 10.1101/2020.12.01.406637.

20. Dunham,A., Jang,G.M., Muralidharan,M., Swaney,D. and Beltrao,P. (2021) A missense variant effect prediction and annotation resource for SARS-CoV-2. *bioRxiv*, 10.1101/2021.02.24.432721.
21. Gowthaman,R., Guest,J.D., Yin,R., Adolf-Bryfogle,J., Schief,W.R. and Pierce,B.G. (2021) CoV3D: a database of high resolution coronavirus protein structures. *Nucleic Acids Res.*, **49**, D282–D287.
22. Portelli,S., Olshansky,M., Rodrigues,C.H.M., D'Souza,E.N., Myung,Y., Silk,M., Alavi,A., Pires,D.E.V. and Ascher,D.B. (2020) Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. *Nat. Genet.*, **52**, 999–1001.
23. Jo,S., Kim,T., Iyer,V.G. and Im,W. (2008) CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.*, **29**, 1859–1865.
24. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
25. Rodríguez-Espigares,I., Torrens-Fontanals,M., Tiemann,J.K.S., Aranda-García,D., Ramírez-Anguita,J.M., Stepniewski,T.M., Worp,N., Varela-Rial,A., Morales-Pastor,A., Medel-Lacruz,B., *et al.* (2020) GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods*, **17**, 777–787.
26. Rose,A.S., Bradley,A.R., Valasatava,Y., Duarte,J.M., Prlic,A. and Rose,P.W. (2016) Web-based molecular graphics for large complexes. In *Proceedings of the 21st International Conference on Web3D Technology - Web3D '16*. ACM Press, New York, New York, USA, pp. 185–186.
27. Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: A web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
28. Tiemann,J.K.S., Guixà-González,R., Hildebrand,P.W. and Rose,A.S. (2017) MDsrv: Viewing and sharing molecular dynamics simulations on the web. *Nat. Methods*, **14**, 1123–1124.
29. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
30. Starr,T.N., Greaney,A.J., Hilton,S.K., Ellis,D., Crawford,K.H.D., Dingens,A.S., Navarro,M.J., Bowen,J.E., Tortorici,M.A., Walls,A.C., *et al.* (2020) Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, **182**, 1295-1310.e20.
31. Greaney,A.J., Starr,T.N., Gilchuk,P., Zost,S.J., Binshtein,E., Loes,A.N., Hilton,S.K., Huddleston,J., Eguia,R., Crawford,K.H.D., *et al.* (2021) Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe*, **29**, 44-57.e9.
32. Vaser,R., Adusumalli,S., Leng,S.N., Sikic,M. and Ng,P.C. (2016) SIFT missense predictions for genomes. *Nat. Protoc.*, **11**, 1–9.
33. McGibbon,R.T.T., Beauchamp,K.A.A., Harrigan,M.P.P., Klein,C., Swails,J.M.M., Hernández,C.X.X., Schwantes,C.R.R., Wang,L.-P., Lane,T.J.J. and Pande,V.S.S. (2015) MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.*, **109**, 1528–1532.
34. Venkatakrishnan,A.J., Fonseca,R., Ma,A.K., Hollingsworth,S.A., Chemparathy,A., Hilger,D., Kooistra,A.J., Ahmari,R., Babu,M.M.,

Kobilka,B.K., *et al.* (2019) Uncovering patterns of atomic interactions in static and dynamic structures of proteins. *bioRxiv*, 10.1101/840694.
35. Singer,J., Gifford,R., Cotten,M. and Robertson,D. (2020) CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. 10.20944/preprints202006.0225.v1.
36. Jacobs,J., Grum-Tokars,V., Zhou,Y., Turlington,M., Saldanha,S.A., Chase,P., Eggler,A., Dawson,E.S., Baez-Santos,Y.M., Tomar,S., *et al.* (2013) Discovery, Synthesis, And Structure-Based Optimization of a Series of N-(tert-Butyl)-2-(N-arylamido)-2-(pyridin-3-yl) Acetamides (ML188) as Potent Noncovalent Small Molecule Inhibitors of the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) 3CL Protease. *J. Med. Chem.*, **56**, 534–546.
37. Lockbaum,G.J., Reyes,A.C., Lee,J.M., Tilvawala,R., Nalivaika,E.A., Ali,A., Kurt Yilmaz,N., Thompson,P.R. and Schiffer,C.A. (2021) Crystal Structure of SARS-CoV-2 Main Protease in Complex with the Non-Covalent Inhibitor ML188. *Viruses*, **13**, 174.

Supplementary data

Supplementary Note S1. System set-up and simulation protocol

To generate the starting systems, the structures were obtained either from the Protein Data Bank (PDB, rcsb.org) (1), or from the curated versions provided by the Coronavirus Structural Task Force (www.insidecorona.net). All proteins were left in their dominant protonation state at pH 7.4 using the MOE software (https://www.chemcomp.com). The CHARMM-GUI Solution Builder (2) was used to solvate each protein in a water box, where the distance between protein atoms and box edges was kept at 10 Å, and to adjust the ionic strength using 0.15 M NaCl. Protein N- and C-terminal ends were capped using ACE and CT3 patches, respectively. Protein and ligand parameters were obtained from the CHARMM36m (3, 4) and CGenFF (5, 6) forcefields, respectively.

Systems were first energy minimized for 20 ps using the NAMD simulation engine (7). Then, a first equilibration step was run at constant volume and temperature (NVT, 303.15 K) for 250 ps using a timestep of 2 fs. Long-distance electrostatic forces were calculated using the Particle Mesh Ewald (PME) algorithm with a switch distance of 10 Å and a cutoff of 12 Å. Restraints were applied to every atom in the system. A second equilibration step was run at constant pressure and temperature (NPT, 1.01325 bar, 310 K) for 20 ns with a timestep of 2 fs. During the equilibration phase, PME was used with a switch distance of 7.5 Å and a cutoff distance of 9.5 Å, and restraints were applied to the backbone of the protein. Finally, in the production phase, 3 replicas of each system were run (with different random seeds to assign Maxwell-distributed initial velocities) at 310 K for 1 μs each using a timestep of 4 fs, the PME algorithm with a switch distance of 7.5 Å and a cutoff distance of 9 Å. During the second equilibration step, the pressure was kept constant using the Berendsen barostat (8). In both the equilibration and the production phase, the temperature was kept constant using the Langevin thermostat (9). Both equilibration and production steps were run using ACEMD3 (10). Coordinates were wrapped to the closest periodic image using HTMD (11) and analyzed with

VMD (12). Supplementary Table S1 below provides a list of the simulations performed in this work.

Supplementary Table S1. List of simulations performed in this work to achieve complete coverage of the SARS-CoV-2 proteome with known structures.

ID	Model name	Included proteins	Uniprot ID	PDB ID	Replicates	Software	Force field	Simulation time
243	NS8 in aqueous solution	ORF8	P0DTC8	7JTL	1	AceMD 3.2.3	CHARMM36 2019	1 x 1 µs
244	Envelope small membrane protein in aqueous solution	E	P0DTC4	7K3G	1	AceMD 3.2.3	CHARMM36 2019	1 x 1 µs
245	NSP1 in aqueous solution	NSP1	P0DTD1	7k3n	1	AceMD 3.2.3	CHARMM36 2019	1 x 1 µs
246	NSP10 in aqueous solution	NSP10	P0DTD1	6ZCT	1	AceMD 3.2.3	CHARMM36 2019	1 x 1 µs
247	Nucleocapsid dimerization domain P21 form in aqueous solution	N	P0DTC9	6WZQ	1	AceMD 3.2.3	CHARMM36 2019	1 x 1 µs
248	ORF7a in aqueous solution	ORF7a	P0DTC7	6W37	1	AceMD 3.2.3	CHARMM36 2019	1 x 1 µs
249	Helicase NSP13 in aqueous solution	NSP13	P0DTD1	6ZSL	1	AceMD 3.2.3	CHARMM36 2019	1 x 1 µs
250	NendoU NSP15 in aqueous solution	NSP15	P0DTD1	6VWW	1	AceMD 3.2.3	CHARMM36 2019	1 x 1 µs
251	NSP16 in aqueous solution	NSP16	P0DTD1	6W4H	1	AceMD 3.2.3	CHARMM36 2019	1 x 1 µs

252	ORF3A in aqueous solution	ORF3a	P0DTC3	6XDC	1	AceMD 3.2.3	CHARMM36 2019	1 x 1 µs
253	ORF9B in aqueous solution (missing loops modelled)	ORF9b	P0DTD2	6Z4U	1	AceMD 3.2.3	CHARMM36 2019	1 x 1 µs
255	Main Protease (Mpro) in Complex with ML188 (no mutation)	NSP5	P0DTD1	7L0D	2	AceMD 3.2.3	CHARMM36 2019	3 x 1 µs
256	Main Protease (Mpro) in Complex with ML188 (A191V mutation)	NSP5	P0DTD1	7L0D	3	AceMD 3.2.3	CHARMM36 2019	3 x 1 µs
257	Main Protease (Mpro) in Complex with ML188 (T24A mutation)	NSP5	P0DTD1	7L0D	3	AceMD 3.2.3	CHARMM36 2019	3 x 1 µs
258	Macrodomain (NSP3) in complex with ADP-ribose	NSP3	P0DTD1	6WOJ	3	AceMD 3.2.3	CHARMM36 2019	3 x 1 µs
259	Macrodomain (NSP3) in complex with ADP-ribose (S167L mutation)	NSP3	P0DTD1	6WOJ	3	AceMD 3.2.3	CHARMM36 2019	3 x 1 µs

Supplementary Note S2. Descriptors of variant impact

Our descriptors include:

- **Mutation effect descriptors**, representing the impact of the amino acid change. Unless indicated otherwise, they were obtained with Biopython (13) version 1.76
    - BLOSUM90 score
    - Charge difference, based on the charge of the wild type (WT) and mutated amino acid
    - Epstein's coefficient of difference (14)
    - Experimental exchangeability (15)
    - Grantham's distance (16)
    - Miyata's distance (17)
    - Sneath's index (18)
    - Several scales of change in hydrophobicity: Kyte-Doolittle (19), Eisenberg-Weiss (20), Engelman (21), Hessa (22, 23), Hopp-Woods (24), Janin (25), Moon-Fleming (26), Wimley-White (27, 28), Zhao-London (29)
    - Variant effect predictions extracted from the database *Mutfunc: SARS-CoV-2* (30), including conservation, structural consequences, and experimental antibody escape data (31).
- **Time-dependent predictors**, consisting of parameters extracted from the WT simulation for the residue affected by the variant. They allow us to evaluate the structural impact of each variant. These predictors were obtained from the trajectory strided so that the delta is 0.1 ns or, in longer simulations, 1 ns.
    - RMSD of the residue with relation to the first trajectory frame, obtained using the *rmsd* module of MDtraj (32).
    - RMSF of the residue atoms, obtained with the *rmsf* module of MDtraj (32).

- Solvent-accessible surface (SASA) of the residue, based on the Shrake and Rupley algorithm (33), calculated with the *shrake_rupley* algorithm of MDtraj (32).
- Chi1 angle of the residue, which corresponds to the first side chain torsion angle formed between the 4 atoms over the CA-CB axis, if available. The Chi1 angle is calculated using the *compute_chi1* module of MDtraj (32).
- Number of contacts that the residue makes with other protein residues, obtained with GetContacts (34). This includes hydrogen bonds, salt bridges, hydrophobic contacts, π-cation contacts, π-stacking contacts, T-stacking contacts, Van der Waals, water bridges, extended water bridges, and total contacts.

- **User-provided descriptors**, which can be provided by the user as a CSV-formatted file and uploaded in the "User-provided" section of the Workbench. Custom descriptors enable the inclusion of arbitrary external data, as well as non-linearities and interactions, in the prediction model.

Supplementary Note S3. Regularized regression models of variant impact

Three linear regression models used least absolute shrinkage and selection operator (LASSO) regularization, a method to minimize the number of predictors for the sake of interpretability and avoid overfitting. In short, LASSO finds the optimal set of coefficients $\beta_i \in \mathbb{R}^p$ to predict a response variable y on the basis of a training set (x_i, y_i) of N examples, namely:

$$min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} ||y - X\beta||_2^2 + \lambda ||\beta||_1 \right\}$$

where λ is a regularization parameter that is usually chosen by cross-validation and $||\ldots||_q$ is the q-norm of a vector.

Here, for the sake of clarity, we modeled the effects of predictors as linear and additive (hence, no interaction terms or non-linearities). We chose the value of λ which provided a cross-validated mean-squared prediction error within one standard error of the minimum

(obtained for λ=0, i.e. including all of the coefficients). The N=3,684 examples used are the RBD variants reported in GISAID (35) for which experimental data from deep mutational scanning experiments, i.e. differences in binding (36), expression (36), and escape (31) with respect to the WT (*y*), and static and time-dependent covariates computed from MD (*x*) were available. The calculations were performed with the *glmnet* package (37) version 4.1. Supplementary Table S2 reports the predictors selected after the LASSO procedure and the corresponding coefficients of each model, as well as the reference from which the experimental data were retrieved.

Supplementary Table S2. Descriptors selected as predictors based on the regression models (LASSO) and corresponding coefficients of each model: differences in binding (36), expression (36), and escape (33) with respect to the WT. "--" means that the corresponding covariate was discarded by the LASSO procedure at the chosen value of λ.

	Binding (36)	**Expression (36)**	**Escape (31)**
(Intercept)	-0.573	-0.673	0.009
BLOSUM90	0.110	0.162	-0.001
Δ Hydrophobicity (Kyte Doolittle)	0.012	--	--
Δ Charge	--	0.007	0.004
At RBD-ACE2 interface?	--	0.360	0.014
In receptor binding motif?	-0.062	0.090	0.005
RMSD average (time-dependent)	--	-1.729	--
RMSD SD (time-dependent)	--	-6.196	--
RMSF average (time-dependent)	0.791	4.367	-0.027
RMSF SD (time-dependent)	0.113	4.157	--
SASA average (time-dependent)	0.243	0.168	-0.002
SASA SD (time-dependent)	0.854	1.787	--
Contacts H bond average (time-dependent)	--	-0.103	--

Contacts H bond SD (time-dependent)	0.246	0.272	--
Contacts salt bridge average (time-dependent)	--	-0.159	0.005
Contacts salt bridge SD (time-dependent)	--	0.125	-0.007
Contacts π-cation average (time-dependent)	--	--	-0.010
Contacts π-cation SD (time-dependent)	--	-0.342	--
Contacts π-stacking average (time-dependent)	--	-0.494	--
Contacts π-stacking SD (time-dependent)	--	0.953	--
Contacts T-stacking average (time-dependent)	--	-0.996	-0.015
Contacts T-stacking SD (time-dependent)	--	--	--
Contacts Van der Waals average (time-dependent)	-0.063	--	--
Contacts van der waals SD (time-dependent)	--	-0.321	0.007

Supplementary Note S4. Analyses of the molecular dynamics (MD) simulations presented in the General section of the Toolkit, in the Workbench page

Interaction networks

Noncovalent residue-residue interactions formed in the simulation are displayed using Flareplots (34). To precompute interactions during the simulation, we used GetContacts (34) for all interaction types. We manually integrated Flareplots and NGL to allow for interactivity between the Toolkit and the Viewer sections of the Workbench page.

Root mean square deviation (RMSD)

The RMSD is computed using the rmsd module of MDtraj (32). The first frame of the trajectory is used as a reference structure by default. The atoms used for RMSD computation can be defined using the provided preselection options (for example, protein alpha

carbons, non-hydrogen protein atoms, ligand, etc.). The RMSD is computed after optimal alignment according to the following equation:

$$RMSD(t) = \sqrt{\frac{1}{N_{atoms}} \sum_{i}^{N_{atoms}} |r_i(1) - r_i(t)|^2}$$

where N_{atoms} is the number of atoms for structure comparison, $r_i(1)$ is the position of atom i in the reference frame (that is, trajectory frame 1) and $r_i(t)$ is the position of atom i at time t of the trajectory.

Root mean square fluctuation (RMSF).

The RMSF is computed for all the alpha carbons of the protein using the rmsf module of MDtraj (32). It is calculated based on the average structure of the simulation, obtained by averaging the coordinates of each atom during the trajectory. The trajectory is aligned to the obtained average structure, and then the RMSF is obtained according to the following equation:

$$RMSF(i) = \sqrt{\frac{1}{T} \sum_{j=1}^{T} (r_i(t_j) - \bar{r}_i)^2}$$

where $r_i(t_j)$ represents the coordinates of atom i at frame j, $\bar{r}_i$ the average position of atom i, and T the total number of frames in the trajectory.

Supplementary Note S5. Custom analysis: 3CLpro case study

The target structure of the SARS-CoV-2 main protease, Mpro (3CLpro) (PDB ID 7L0D) was obtained from the PDB database (1). The corresponding mutation for the T24A variant was introduced using the HTMD software (11). The WT and the T24A variant system were further prepared and simulated as described above (Supplementary Note S1). ML188 inhibitor unbinding was typically observed for the WT as well as the T24A mutant within 1 μs as reported in Supplementary Table S3.

Supplementary Table S3. ML188 inhibitor unbinding from the SARS-CoV-2 main protease WT and T24A mutant. Ligand unbinding from its original pocket was monitored as follows. We first tracked the "native contact set" of residues which were within 3 Å from the equilibrated crystallographic ligand binding pose (a total of 20 residues). We defined a state as unbound if the ligand was at more than 3 Å of distance from all of the residues in the "native contact" set. Times to first unbinding were rounded to 10 ns. Simulation in which the ligand did not escape from the binding pocket but shows a substantial destabilization are highlighted with an asterisk (*).

Simulation	Replicate	ML188 unbinding time (ns)
WT	1	>1000*
WT	2	>1000*
WT	3	780
T24A	1	70
T24A	2	930
T24A	3	710

Supplementary Figure S1. Entity-relationship (ER) diagram of the SCoV2-MD database.The primary keys in the entities are bolded, and the foreign keys in the entities are highlighted. Each of the entities correspond to a table in the database. For instance, we have a table with information on the different molecular dynamics (MD) simulations (CovidDynamics), where each row is a different simulation entry with an unique identifier (ID) known as primary key, and columns hold attributes of the simulation (delta, force field, software, etc.). The tables can be related in different ways so that it is possible to connect the data held in them. In some cases, each record of a table can be related to one or more records of a second table. For instance, each simulation entry can be related to one or more simulation files. This is known as a one-to-many relationship. It can also be the case that each record of a table can relate to more than one record of a second table while, at the same time, each record of the second table can also relate to more than one record of the first . For example, we can have cases of

multiple PDB structures (different records in the CovidModel table) for each protein record (CovidProtein) and, at the same time, a PDB structure can include more than one protein, so we can also have multiple protein records for each model record. In these cases, we establish a many-to-many relationship between the tables. For that, we create a junction table between them (e.g. CovidModelProtein table) to save each occurrence of the relationship. For example, if we add to the database a protein with ID *p1* (e.g. spike protein) that corresponds to models with ID *m1* and *m2* (e.g. 6VXX and 6M0J), but model m2 (6M0J) corresponds to both protein *p1* and *p2* (spike and ACE2), we create three entries in the CovidModelProtein table: one relating *p1* to *m1*, another relating *p1* to *m2*, and a third relating *p2* to *m2*. All table records can then be mapped into Python objects via the Python-based web framework Django.

References

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
2. Jo,S., Kim,T., Iyer,V.G. and Im,W. (2008) CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.*, **29**, 1859–1865.
3. Huang,J., Rauscher,S., Nawrocki,G., Ran,T., Feig,M., de Groot,B.L., Grubmüller,H. and MacKerell,A.D. (2016) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, **14**, 71–73.
4. Lee,J., Cheng,X., Swails,J.M., Yeom,M.S., Eastman,P.K., Lemkul,J.A., Wei,S., Buckner,J., Jeong,J.C., Qi,Y., *et al.* (2016) CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.*, **12**, 405–413.
5. Vanommeslaeghe,K., Hatcher,E., Acharya,C., Kundu,S., Zhong,S., Shim,J., Darian,E., Guvench,O., Lopes,P., Vorobyov,I., *et al.* (2010) CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, **31**, 671–690.
6. Yu,W., He,X., Vanommeslaeghe,K. and MacKerell,A.D. (2012) Extension of the CHARMM General Force Field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J. Comput. Chem.*, **33**, 2451–2468.
7. Phillips,J.C., Hardy,D.J., Maia,J.D.C., Stone,J.E., Ribeiro,J.V., Bernardi,R.C., Buch,R., Fiorin,G., Hénin,J., Jiang,W., *et al.* (2020) Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.*, **153**, 044130.
8. Berendsen,H.J.C., Postma,J.P.M., van Gunsteren,W.F., DiNola,A. and Haak,J.R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**, 3684–3690.
9. Loncharich,R.J., Brooks,B.R. and Pastor,R.W. (1992) Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers*, **32**, 523–535.

10. Harvey,M.J., Giupponi,G. and De Fabritiis,G. (2009) ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.*, **5**, 1632–1639.
11. Doerr,S., Harvey,M.J., Noé,F. and De Fabritiis,G. (2016) HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.*, **12**, 1845–52.
12. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: Visual molecular dynamics. *J. Mol. Graph.*, **14**, 27–28.
13. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B., *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
14. Epstein,C.J. (1967) Non-randomness of Ammo-acid Changes in the Evolution of Homologous Proteins. *Nature*, **215**, 355–359.
15. Yampolsky,L.Y. and Stoltzfus,A. (2005) The Exchangeability of Amino Acids in Proteins. *Genetics*, **170**, 1459–1472.
16. Grantham,R. (1974) Amino Acid Difference Formula to Help Explain Protein Evolution. *Science*, **185**, 862–864.
17. Miyata,T., Miyazawa,S. and Yasunaga,T. (1979) Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.*, **12**, 219–236.
18. Sneath,P.H.A. (1966) Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.*, **12**, 157–195.
19. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
20. Eisenberg,D., Schwarz,E., Komaromy,M. and Wall,R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, **179**, 125–142.
21. Engelman,D.M., Steitz,T.A. and Goldman,A. (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 321–353.
22. Hessa,T., Meindl-Beinker,N.M., Bernsel,A., Kim,H., Sato,Y., Lerch-Bader,M., Nilsson,I., White,S.H. and von Heijne,G. (2007) Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature*, **450**, 1026–1030.
23. Hessa,T., Kim,H., Bihlmaier,K., Lundin,C., Boekel,J., Andersson,H., Nilsson,I., White,S.H. and von Heijne,G. (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, **433**, 377–381.
24. Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U. S. A.*, **78**, 3824–3828.
25. Janin,J. (1979) Surface and inside volumes in globular proteins. *Nature*, **277**, 491–492.
26. Moon,C.P. and Fleming,K.G. (2011) Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proc. Natl. Acad. Sci.*, **108**, 10174–10177.
27. Wimley,W.C., Creamer,T.P. and White,S.H. (1996) Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry*, **35**, 5109–5124.

28. Wimley,W.C. and White,S.H. (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.*, **3**, 842–848.
29. Zhao,G. and London,E. (2006) An amino acid 'transmembrane tendency' scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. *Protein Sci. Publ. Protein Soc.*, **15**, 1987–2001.
30. Dunham,A., Jang,G.M., Muralidharan,M., Swaney,D. and Beltrao,P. (2021) A missense variant effect prediction and annotation resource for SARS-CoV-2. *bioRxiv*, 10.1101/2021.02.24.432721.
31. Greaney,A.J., Starr,T.N., Gilchuk,P., Zost,S.J., Binshtein,E., Loes,A.N., Hilton,S.K., Huddleston,J., Eguia,R., Crawford,K.H.D., *et al.* (2021) Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe*, **29**, 44-57.e9.
32. McGibbon,R.T.T., Beauchamp,K.A.A., Harrigan,M.P.P., Klein,C., Swails,J.M.M., Hernández,C.X.X., Schwantes,C.R.R., Wang,L.-P., Lane,T.J.J. and Pande,V.S.S. (2015) MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.*, **109**, 1528–1532.
33. Shrake,A. and Rupley,J.A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.*, **79**, 351–371.
34. Venkatakrishnan,A.J., Fonseca,R., Ma,A.K., Hollingsworth,S.A., Chemparathy,A., Hilger,D., Kooistra,A.J., Ahmari,R., Babu,M.M., Kobilka,B.K., *et al.* (2019) Uncovering patterns of atomic interactions in static and dynamic structures of proteins. *bioRxiv*, 10.1101/840694.
35. Elbe,S. and Buckland-Merrett,G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.*, **1**, 33–46.
36. Starr,T.N., Greaney,A.J., Hilton,S.K., Ellis,D., Crawford,K.H.D., Dingens,A.S., Navarro,M.J., Bowen,J.E., Tortorici,M.A., Walls,A.C., *et al.* (2020) Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, **182**, 1295-1310.e20.
37. Friedman,J.H., Hastie,T. and Tibshirani,R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.*, **33**, 1–22.

4. DISCUSSION

4.1. The capabilities of MD simulations

Traditionally, proteins have been perceived as rigid entities, mainly because experimental structures revealed static snapshots of protein conformations. However, when taken together, these static structures show that proteins exist in multiple conformational states, and therefore that there are dynamic fluctuations between them. A clear example is GPCRs, which exhibit a wide range of conformational states with different signaling properties[270]. Thus, a complete understanding of the structural mechanisms underlying protein functionality requires the exploration of conformational dynamics. MD simulations have emerged as one of the most promising approaches to study the complexity of such dynamics[20,271], being able to capture molecular events at a spatio-temporal resolution and conditions that are not always accessible with experimental techniques[20,152].

In this thesis, we first built a solid knowledge base of the capabilities of MD simulations and how this technique can be applied to advance our understanding of pharmacologically relevant proteins, focusing on GPCRs (**publications 3.1, 3.2,** and **3.3**). As reviewed in these publications, MD simulations have proven their usefulness for the study of important biological processes in GPCRs such as ligand binding, allostery, activation/inactivation, post-translational modifications, and sequence variability, among others. Given the importance of GPCRs as pharmacological targets, this has contributed to accelerating the discovery of new and safer therapeutic strategies. Yet, GPCRs are complex signaling machines, and a complete understanding of the dynamic molecular processes that guide their functionality will require far more investigation. Moreover, MD simulations have technical limitations, which restrict the processes that can be examined with them. Fortunately, considering the fast and continuous technological developments that we are experiencing, the capabilities of MD simulations are expected to expand with time as well[166,167]. As accessible computational power increases and MD algorithms are improved, the costs of simulating are reduced, allowing to carry out simulations for longer lengths of time, and of systems of increasing

size. Coupled with an ever-growing accuracy in the force fields, this should progressively push the limitations of MD simulations and extend their application to the study of molecular mechanisms that were previously difficult to capture. This could include, for instance, global conformational rearrangements, receptor dimerization, and coupling to intracellular signaling proteins. In the end, this will enable us to address biological questions that are currently unsolved.

4.2. GPCRmd and SCoV2-MD: collection, dissemination, and analysis of MD simulations

Considering the expanding popularity and capabilities of MD simulations, researchers are becoming aware of the importance of sharing this data, as previously discussed (*1.3. Sharing MD simulations*). To maximize the potential of MD research, the data generated should follow the FAIR principles, being 'findable' by anyone in a searchable resource; 'accessible' using an open, free, and universally implementable protocol; 'interoperable' so that comparable data from non-cooperating resources can be easily integrated; and 'reusable' thanks to robust metadata, provenance information, and clear usage licenses[249,272]. A very important step to align with these principles is the development of online platforms that facilitate data sharing. Thus, this thesis was devoted to the design and development of open-access repositories for the dissemination of MD simulations, including online tools to easily analyze and visually inspect the trajectories. Particularly, we focused on resources dedicated to two groups of proteins with pharmacological relevance: GPCRs and the SARS-CoV-2 proteome. These repositories are GPCRmd (**publication 3.4**) and SCoV2-MD (**publication 3.5**).

GPCRmd (www.gpcrmd.org, **publication 3.4**) is a community-driven online resource that provides access to MD simulations of most GPCR structures solved to date, together with the necessary metadata (e.g. force field, simulation software, integration time-step) to ensure transparency and reproducibility. Its final aim is to map the entire set of solved GPCR structures. For that, data are deposited either by individual contributions or by periodic updates

from the GPCRmd community. GPCRmd also provides a comprehensive set of interactive tools specifically designed for the visualization and analysis of GPCR simulations. Moreover, it includes a meta-analysis tool to compare and cluster the available simulations.

In response to the COVID-19 pandemic, we also created SCoV2-MD (www.scov2-md.org, **publication 3.5**) with the goal to integrate, cross-reference, and share MD dynamics data and metadata of the SARS-CoV-2 proteome. We collated and organized simulation data from several public databases, which was complemented with in-house simulations to complete the coverage of SARS-CoV-2 proteins with known structures. As GPCRmd, this resource includes visualization and analysis tools that provide a rapid and intuitive way to explore the simulation data. However, in this case, we focused on the analysis of known variant substitutions, incorporating tools to predict their functional impact based on a combination of descriptors.

4.3. GPCRmd and SCoV2-MD in the context of currently available MD repositories

As mentioned previously (*1.3. Sharing MD simulations*), due to the technical challenges of sharing MD data (i.e. the large file sizes), the number of online resources specialized in this type of data is not high (Table 1). Even fewer resources include options for the interactive analyses and visualization of the trajectories, which has only recently become feasible thanks to advances in browser technology. Such platforms are typically benefitted from focusing on one specific group of proteins. This reduces the required storage space, but also enables the implementation of analysis, visualization, or search options that are specific for the selected proteins. To our knowledge, GPCRmd is the first GPCR-specific MD platform. Another platform that includes MD simulations of GPCRs together with a comprehensive set of interactive visualization and analysis tools is MoDEL-CNS[265]. However, this platform focuses on all proteins involved in processes of the central nervous systems. Thus, unlike GPCRmd, it does not include GPCR-specific tools that take into account the generic GPCR residue numbering schemes, such as user-defined selections based on these

schemes, interaction networks between the different TMs, or our meta-analysis tool for the comparison of the interaction pattern of different receptors based on GPCR sequence alignments. In the case of SARS-CoV-2-related proteins, the platform BioExcel-CV19[260] includes simulations of the viral proteome together with multiple tools to display and analyze the trajectories online. However, it does not share with SCoV2-MD the focus on the study of the functional impact of variants sequenced during the pandemic. As mentioned, SCoV2-MD links the structural dynamics of the viral proteome with information on detected SARS-CoV-2 variants. For that, we integrated to the platform experimental data from the GISAID[273] resource, one of the main pandemic genome databases. This allows to extensively evaluate the structural and functional impact of variant-associated mutations taking into account protein dynamics. Our platform also includes variant effect predictors from the Mutfunc[274] resource, among other impact descriptors. Moreover, since SCoV2-MD collates and organizes simulation data from available open repositories, BioExcel-CV19 simulations were included in the dataset.

4.4. The potential impact of MD resources on future research

4.4.1. Democratization of the access to MD data, a key step for multidisciplinarity

One of the main benefits of online platforms such as GPCRmd and SCoV2-MD is their capacity to democratize access to scientific data – in this case, MD data – to both experts and non-experts[254]. They only require a web browser to access the data, without the need of installing external software or plugins. Thus, there is no technical barrier to using them, and they are compatible with any device with a web browser. Moreover, the tools for the visualization and analysis of MD simulations in GPCRmd and SCoV2-MD are designed to be interactive and easy to use, which allows an instant and intuitive understanding of the dynamic processes captured in the simulations. This can be especially useful, for instance, to inspect the simulation data generated in a publication to observe the phenomena described by the authors. This not only strengthens the understanding of the findings described in the publication, but also

generates transparency, trust, and reliability[253]. Consequently, this triggers discussion, knowledge exchange, and even the discovery of additional findings. All these factors are key to supporting multidisciplinarity, which ultimately provides new perspectives and opens new avenues for research. For instance, in **publication 3.4** we discuss how the open access to the structural dynamics of proteins, in this case, GPCRs, can support the research of scientists of different disciplines, including structural and evolutionary biologists, computational and medicinal chemists, and protein engineers, among others.

4.4.2. The re-utilization of previously published datasets enable new discoveries

By making large and difficult-to-collect datasets available, data sharing can enable scientific progress that is far beyond the resources of a single research group[251]. This is the case of large-scale studies across different MD setups, force fields, ligands, lipid compositions, GPCR subtypes, or variants. For instance, Anila et al were able to perform a benchmarking of available force fields completely based on open-access MD trajectories[275]. In **publication 3.4** we demonstrate the potential of performing comparative studies based on the GPCRmd dataset using tools implemented in the platform (*4.4.3.1. Functional hotspots revealed by comparative analysis implemented in GPCRmd*) or custom analyses (*4.4.3.2. Exploiting the GPCRmd dataset to examine sodium ion interactions across class A GPCRs*).

Importantly, in comparative studies such as those exposed here, one needs to be aware of the subtle differences in the simulation parameters, software, and force field used, as they could have an impact on the results. Thus, it is important that all the simulation data and metadata in the repository are harmonized and standardized so that they can be found easily, and simulations can be filtered accordingly. What is more, structured and well-defined metadata is essential to ensure the reproducibility of the MD simulations. For these reasons, the online platforms developed in this thesis include simulation report sections, where all the available metadata and simulation files are organized. Moreover, both platforms include search pages that allow to easily filter the

available data. In the case of GPCRmd, reproducibility and comparative studies are also simplified by the fact that a large part of the dataset is generated by the GPCRmd community with a specific and consistent protocol, carefully designed by the community members. Thus, it is easy to perform comparative analyses using this data.

Apart from large-scale initiatives, smaller studies based on public datasets can also provide important contributions to research. Thanks to open repositories, MD data can be reused for analyses that go beyond the original purpose of the studies that generated them, avoiding the need to repeat the simulation effort. Despite GPCRmd being a fairly new resource, there are already several published studies that use some of the simulations deposited in it. For instance, some authors used GPCRmd simulations to demonstrate the use of new computational approaches for the large-scale detection of protein binding sites[276], the generation of interaction fingerprints in molecular complexes[277], and the prediction of ligand poses in human GPCRs based on a hybrid molecular mechanics/coarse-grained approach[278]. Moreover, Denzinger et al. studied the conformational impact of biased ligands on 5-HT_{2B} receptors and compared the structural movements captured in their results with those observed in a GPCRmd simulation[279]. In the case of SCoV2-MD, it has just been accepted for publication at the time of writing, so this resource has not yet been disseminated for its use in further studies. However, we present an example of a possible analysis that can be achieved with this platform (*4.4.3.3. Using SCoV2-MD to pinpoint variant substitutions with impact on protein function*).

4.4.3. Case studies: showcasing the potential of the obtained resources

To demonstrate the capabilities of the platforms developed in this thesis, we provided several case studies, exposed in **publications 3.4** and **3.5**. These case studies highlight the possible applications of some of the implemented online tools, as well as the potential of exploiting the accumulated dataset to shed light on important aspects of GPCR and SARS-CoV-2 biology.

4.4.3.1. Functional hotspots revealed by comparative analysis implemented in GPCRmd

Among the tools implemented in GPCRmd, we included the receptor meta-analysis to perform comparative studies within the dataset. This module provides the possibility to compare and cluster the available simulations, or a subset of interest, based on their interaction patterns. To exemplify the applicability of this tool, in **publication 3.4** we used it to investigate water networks, which have been proposed to play an important role for GPCR function[209,280]. For that, we clustered simulations based on water-mediated intra-receptor interactions, which revealed interesting similarities and differences between β_2AR and OX_2-receptor.

Along with previously described conserved water networks[206], our analysis captured other water networks conserved among these subtypes. For instance, we detected a conserved network linking TM6 (6x47, 6x51) and TM7 (7x37), which is less prominent in active structures. Considering the link between TM6 conformational changes and receptor activation, we could speculate that uncoupling the interactions in this network represents a step during receptor activation. Contrarily, some relevant differences were found between the two receptor subtypes, such as a water bridge between intracellular loop 1 (ICL1, 12x49) and helix 8 (H8, 8x49) only found in the β_2AR. These differences could potentially be related to the distinct coupling profile to intracellular partners shown by these receptors.

All in all, this case study shows how comparative analyses performed with the meta-analysis tool could hint at both universal and distinct mechanisms governing the structural dynamics of GPCRs.

4.4.3.2. Exploiting the GPCRmd dataset to examine sodium ion interactions across class A GPCRs

In order to open the door to the scientific community to exploit the data accumulated in our platforms, the entire dataset of GPCRmd and SCoV2-MD is available for download. As exposed in **publication 3.4**, to demonstrate the value of such a comprehensive dataset, we used the simulation data included in GPCRmd to

perform a comparative analysis of sodium ion binding in class A GPCRs.

Sodium interaction is an almost universal mechanism of allosteric modulation for GPCRs, but the details of this mechanism are still poorly understood[34]. To shed light on this process, we analyzed sodium interaction to conserved orthosteric (3x32) and allosteric (2x50) residues in 183 simulations (61 different apo structures, with three replicas each) covering 26 different class A receptor subtypes. The obtained interaction frequencies allowed us to classify receptors in four groups with notably different patterns: group I (high interaction frequency to both positions), group II (interaction frequency marginal at D2x50 but high at 3x32), group III (interaction frequency high at D2x50 but marginal at 3x32), and group IV (no binding at neither position). Further inspection of the simulation data revealed how differences in sequence, structure and dynamics between the groups could explain such distinct sodium binding profiles. Interestingly, the differences observed between the obtained groups may serve as an evolutionary mechanism to differentially modulate allosteric sodium ion binding in different receptors.

Overall, while our results confirm the essential role of D2x50 for allosteric sodium binding in class A GPCRs[36,280], they also show that the presence or absence of D3x32 in the orthosteric binding site determine the sodium interaction profile. This analysis exemplifies the potential of the comprehensive GPCRmd dataset to elucidate the molecular mechanisms underlying GPCR physiology.

4.4.3.3. Using SCoV2-MD to pinpoint variant substitutions with impact on protein function

As described in **publication 3.5**, the simulations of viral proteins collected in SCoV2-MD are annotated with known variant-associated mutations. For each mutation, we provide a wide set of descriptors including static and MD-derived parameters. These descriptors can be combined in an impact score to interrogate the impact of each mutation.

To illustrate the application of the variant impact prediction tool, in our publication we analyze the amino acid substitutions that have

been detected so far in the spike RBD. Understanding the impact of RBD mutations is particularly relevant, as they could have an effect on immune recognition or on the binding to the host cell. For our impact score, we took into account the RMSD of the simulation in order to detect regions of high structural stability, which tend to be important for the overall function of the viral protein. We combined the RMSD with a substitution score such as BLOSUM to capture unfavorable substitutions as well. The obtained score highlighted position C488 as high-impact. Interestingly, visualization of the protein structure revealed that C488 forms a disulfide bridge with C480. Unfavorable substitutions such as C488R disrupt the disulfide bridge, introducing flexibility into this region. This in turn can alter the propensity of the virus to attach to the host cell receptor, causing an impact on viral infectivity.

This analysis exemplifies the capability of this tool to evaluate the structural and functional impact of variant-associated mutations taking into account protein dynamics. This could be a first step to stimulate the early detection of emerging variants of concern, guiding further experimental characterization.

4.5. Limitations, challenges, and future improvements

4.5.1. Sustainability, a main challenge for research resources

Regardless of the capabilities of MD platforms, some challenges need to be overcome before they reach their full potential. One of the most important challenges is maintenance. First, we need to ensure the continuous update of the MD data and related information included in the repositories. GPCRmd aims to map the entire 3D GPCRome (i.e. the available GPCR structures), and thus, as new structures are published, these will need to be simulated according to our standardized protocol and uploaded. As mentioned, this will be tackled in periodic updates with the help of the community supporting GPCRmd. Similarly, SCoV2-MD needs to continuously incorporate MD simulations newly published in open-access repositories or generate in-house simulations of new structures. Also, it is important to update the information of

detected amino acid substitution in the viral genome, obtained from GISAID[273], as well as experimental impact predictions, obtained from Mutfunc[274]. The extraction of data from external resources, including other MD repositories, GISAID, and Mutfunc is automatized whenever possible. However, automatized protocols need to be maintained as well, as these external resources may incorporate changes in the future. Moreover, as simulation time scales increase and systems become bigger, an increase in the deposited data is expected for both platforms. This will require an increase in data storage space, internet bandwidth, and computational power. Finally, another important point for the maintenance of the online platforms is the file formats of MD data. The MD simulations field has a tendency to produce a multitude of formats depending on the MD software used. As the field evolves, and new software and force fields are developed, new formats may appear[247]. Thus, it is essential that the supported formats in our platforms are updated accordingly. In practice, the supported formats are defined by the software used to display, interact and analyze the simulations, such as MDsrv[255], MDTraj[151], and MDAnalysis[149,150].

As previously mentioned, the best strategy to face challenges such as those exposed here and preserve adequate platform sustainability is counting on a strong community support[247]. Community engagement ensures the continuous maintenance and update of the resource, avoiding issues that individual laboratories may face, such as the termination of a grant or the graduation of students. This has been shown, for instance, by structural biology and (proteo)genomics projects like the PDB (rcsb.org)[116] or the Galaxy project[281], respectively. For this reason, GPCRmd has been created under the umbrella of a strong research community, as explained in **publication 3.4**. Since SCoV2-MD was created as a spin-off of GPCRmd, and thus shares many resources with it, it can also benefit from this community support.

4.5.2. Ensuring the quality of the submitted trajectories

Another important issue of online repositories is controlling the quality of the submitted data. Convincing users of the validity of the data generated by other users can be challenging, so in order to gain the trust of the community, it is critical to provide some guarantees.

At the moment, the strategy followed in GPCRmd and SCoV2-MD is to only accept submissions of published data. We also accept submissions of data that is under the peer review process, with restricted access only for the authors and reviewers. In SCoV2-MD, as indicated before, apart from individual submissions we include simulations produced by leading groups in the field extracted from well-known resources, mainly BioExcel-CV19[260], COVID-19 Molecular Structure and Therapeutics, the CHARMM-GUI simulation archive[282], and the Exscalate4Cov project[283].

Quality control could be improved by implementing an automatized quality analysis pipeline to test for potential problems in the deposited simulations. However, due to the extensive diversity of MD simulations, the implementation of automatized pipelines able to accurately detect artifacts is not an easy task. Often, these are limited to tagging the simulations as suspicious, leaving the responsibility to decide whether the simulation data should be trusted or not to the user. In this sense, a common effort is still needed from the MD community to define rules on the validation of trajectories[284].

4.5.3. Integrating open data practices in the workflow of MD

It is also worth mentioning that a relevant challenge of data sharing platforms is overcoming resistance and reluctance in publishing data sets. Unlike disciplines such as protein crystallography or genomics, which have open data practices well integrated into their workflow, data sharing in the MD community still has not become widely adopted. To establish an efficient sharing culture, a community effort is needed to continue developing and improving tools and repositories that support the FAIR principles, as well as defining best practice guidelines on how MD simulations should be shared in order to fully comply with these principles[247,253,284,285]. In the long run, the MD field may reach the situation in which, prior to publication, authors deposit the MD data to a global repository equivalent to the PDB for experimental structures. Special-purpose MD platforms such as those described in this thesis may accordingly be considered a first step on the way towards this goal.

4.6. Final remarks

To conclude, repositories such as GPCRmd and SCoV2-MD have the potential to promote transparency and reproducibility in the field of MD simulations. They provide dissemination of MD simulation data, broadening the outreach of protein dynamics to researchers of different fields, and thus enhancing collaboration and multidisciplinarity. By facilitating the inspection of MD simulations, it is possible to increase the understanding of the studied mechanisms, detect or explain unresolved issues, and complement previous findings. Consequently, novel ideas can be generated, and new lines of analysis may be triggered, contributing to the acceleration of research. However, the development of MD repositories is bound to several challenges, such as ensuring its sustainability, controlling the quality of the deposited data, and motivating researchers to upload their simulations. With the effort of the MD community, we will continue improving our strategies to effectively share simulation data, with the final goal of making MD data FAIR.

5. CONCLUSIONS

The main conclusions arising from the work presented in this thesis are the following:

- An extensive review of the literature shows that MD simulations can provide unique insights into the structural dynamics of proteins, and in particular GPCRs. As such, this technique has been key to advancing our understanding of important phenomena in GPCR physiology such as ligand binding, allostery, activation/inactivation, post-translational modifications, and sequence variability, among others. This has contributed to accelerating the discovery of new and safer drugs targeting GPCRs.
- We present GPCRmd, a community-driven online resource that makes MD simulations of GPCRs available to the whole community of researchers interested in this protein family. It includes simulations of most GPCR structures solved to date, as well as the corresponding metadata. It also provides access to a comprehensive set of tools to simplify the visualization and analysis of the simulations.
- In response to the COVID-19 pandemic, we developed SCoV2-MD, an online resource with the objective to organize, cross-reference, and disseminate MD simulations of the SARS-CoV-2 proteome, together with the corresponding metadata. This resource also includes visualization and analysis tools, with special emphasis on the prediction of the impact of known viral mutations on protein functionality.
- The simulation data accumulated in GPCRmd and SCoV2-MD, together with the implemented online tools, simplify the exploration of protein dynamics and can help to shed light on the molecular mechanisms underlying GPCR and SARS-CoV-2 biology.
- Repositories such as GPCRmd and SCoV2-MD have the potential to promote reproducibility and transparent dissemination in the field of MD simulations. These are key ingredients to enhance collaborative and multidisciplinary

research. A community effort is still needed to completely align MD data to the FAIR principles, but open-access repositories such as those presented in this thesis may be an important step towards this goal.

6. LIST OF COMMUNICATIONS

6.1. Articles

1. Torrens-Fontanals, M., Peralta-García, A., Talarico, C., Guixà-González, R., Giorgino, T. & Selent, J. SCoV2-MD: a database for the dynamics of the SARS-CoV-2 proteome and variant impact predictions. *Nucleic Acids Res.* **1**, 13-14 (2021).
2. Stepniewski, T. M., Mancini, A., Ågren, R., Torrens-Fontanals, M., Semache, M., Bouvier, M., Sahlholm, K., Breton, B. & Selent, J. Mechanistic insights into dopaminergic and serotonergic neurotransmission – concerted interactions with helices 5 and 6 drive the functional outcome. *Chem. Sci.* **12**, 10990–11003 (2021).
3. Torrens-Fontanals, M., Stepniewski, T. M., Gloriam, D. E. & Selent, J. Structural dynamics bridge the gap between the genetic and functional levels of GPCRs. *Curr. Opin. Struct. Biol.* **69**, 150–159 (2021).
4. Torrens-Fontanals, M., Stepniewski, T. M., Aranda-García, D., Morales-Pastor, A., Medel-Lacruz, B. & Selent, J. How do molecular dynamics data complement static structural data of GPCRs. *Int. J. Mol. Sci.* **21**, 5933 (2020).
5. Rodríguez-Espigares, I.*, Torrens-Fontanals, M.*, Tiemann, J. K. S., Aranda-García, D., Ramírez-Anguita, J. M., Stepniewski, T. M., Worp, N., Varela-Rial, A., Morales-Pastor, A., Medel-Lacruz, B. *et al.* GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods* **17**, 777–787 (2020). *Both authors contributed equally to this work.
6. Stepniewski, T. M.*, Torrens-Fontanals, M.*, Rodríguez-Espigares, I., Giorgino, T., Primdahl, K. G., Vik, A., Stenstrøm, Y., Selent, J. & Hansen, T. V. Synthesis, molecular modelling studies and biological evaluation of new oxoeicosanoid receptor 1 agonists. *Bioorganic Med. Chem.* **26**, 3580–3587 (2018). *Both authors contributed equally to this work.
7. Ramírez-Anguita, J. M., Rodríguez-Espigares, I., Guixà-González, R., Bruno, A., Torrens-Fontanals, M., Varela-

Rial, A. & Selent, J. Membrane cholesterol effect on the 5-HT_{2A} receptor: Insights into the lipid-induced modulation of an antipsychotic drug target. *Biotechnol. Appl. Biochem.* **65**, 29–37 (2018).

6.2. Book chapters

1. Torrens-Fontanals, M., Stepniewski, T. M., Rodríguez-Espigares, I. & Selent, J. Application of Biomolecular Simulations to G Protein-Coupled Receptors (GPCRs). in *Biomolecular Simulations in Structure-based Drug Discovery* (eds. Gervasio, F. L. & Spiwok, V.) 205–223 (Wiley, 2018). doi:10.1002/9783527806836.ch8.

6.3. Oral communications

1. *GPCRmd: a multidisciplinary platform to explore GPCR dynamics.* Fourth ERNEST meeting: Insights from structures of signalling complexes and computational modelling on GPCR function, online. April 2021.
2. *GPCRmd is growing: new tools and data coverage.* GDR3545 Virtual Workshop, online. November 2020.
3. *GPCRmd - molecular dynamics of GPCRs for everybody.* Third ERNEST meeting: Signal transduction –From the genomic to the systems level (and everything in between), online. October 2020.
4. *GPCRmd update: Bringing transparency and GPCR MD data within reach.* Second ERNEST meeting: New Perspectives in Signal Transduction: GPCRs and Beyond, online. March 2020.
5. *GPCRmd – Dynamics meets multidisciplinary research.* First ERNEST meeting: GPCR Pharmacology – Activation, Signalling and Drug Design, Belfast (Ireland). October 2019.
6. *GPCRmd: GPCR-specialized storage, dissemination and analysis of molecular dynamics data.* G Protein-Coupled Receptors: New Techniques, Tools and Concepts Symposium, Copenhagen (Denmark). October 2019.

7. *Brainstorm@home: Fighting brain diseases with volunteer distributed computing.* Interdisciplinary Meeting of Predoctoral Researchers (JIPI), Barcelona (Spain). February 2019.

6.4. Posters

1. *GPCRmd Workbench: a GPCR-specialized platform for sharing, visualization and analysis of GPCR dynamics.* GDR3545-GPCR international meeting, Montpellier (France). October 2019.
2. *Synthesis, molecular modelling studies and biological evaluation of new oxoeicosanoid receptor 1 agonists.* 2018 DCEXS-UPF Symposium, Barcelona (Spain). November 2018.
3. *Synthesis, molecular modelling studies and biological evaluation of new oxoeicosanoid receptor 1 agonists.* Early Career Scientist Forum on GPCR Signal Transduction, Berlin (Germany). July 2018.

7. BIBLIOGRAPHY

1. Pierce, K. L., Premont, R. T. & Lefkowitz, R. J. Seven-transmembrane receptors. *Nat. Rev. Mol. Cell Biol.* **3**, 639–650 (2002).
2. Mendoza, A. de, Sebé-Pedrós, A. & Ruiz-Trillo, I. The Evolution of the GPCR Signaling System in Eukaryotes: Modularity, Conservation, and the Transition to Metazoan Multicellularity. *Genome Biol. Evol.* **6**, 606–619 (2014).
3. Hauser, A. S. *et al.* Pharmacogenomics of GPCR Drug Targets. *Cell* **172**, 41–54 (2018).
4. Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B. & Gloriam, D. E. Trends in GPCR drug discovery: New agents, targets and indications. *Nat. Rev. Drug Discov.* **16**, 829–842 (2017).
5. Rask-Andersen, M., Almén, M. S. & Schiöth, H. B. Trends in the exploitation of novel drug targets. *Nat. Rev. Drug Discov.* **10**, 579–590 (2011).
6. Fredriksson, R., Lagerström, M. C., Lundin, L.-G. & Schiöth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **63**, 1256–1272 (2003).
7. Nordstrom, K. J. V., Sallman Almen, M., Edstam, M. M., Fredriksson, R. & Schioth, H. B. Independent HHsearch, Needleman-Wunsch-Based, and Motif Analyses Reveal the Overall Hierarchy for Most of the G Protein-Coupled Receptor Families. *Mol. Biol. Evol.* **28**, 2471–2480 (2011).
8. Taddese, B. *et al.* Do Plants Contain G Protein-Coupled Receptors? *Plant Physiol.* **164**, 287–307 (2014).
9. Brown, N. A., Schrevens, S., van Dijck, P. & Goldman, G. H. Fungal G-protein-coupled receptors: mediators of pathogenesis and targets for disease control. *Nat. Microbiol.* **3**, 402–414 (2018).
10. Kim, J.-Y., Haastert, P. Van & Devreotesl, P. N. Social senses: G-protein-coupled receptor signaling pathways in Dictyostelium discoideum. *Chem. Biol.* **3**, 239–243 (1996).
11. Katritch, V., Cherezov, V. & Stevens, R. C. Structure-Function of the G Protein–Coupled Receptor Superfamily. *Annu. Rev. Pharmacol. Toxicol.* **53**, 531–556 (2013).
12. Isberg, V. *et al.* Generic GPCR residue numbers - aligning topology maps while minding the gaps. *Trends Pharmacol. Sci.* **36**, 22–31 (2015).
13. Wootten, D., Simms, J., Miller, L. J., Christopoulos, A. & Sexton, P. M. Polar transmembrane interactions drive formation of ligand-specific and signal pathway-biased family B G protein-coupled receptor conformations. *Proc. Natl. Acad. Sci.* **110**, 5211–5216 (2013).
14. Pin, J.-P. P., Galvez, T. & Prézeau, L. Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacol. Ther.* **98**, 325–354 (2003).
15. Wang, C. *et al.* Structural basis for Smoothened receptor modulation and chemoresistance to anticancer drugs. *Nat. Commun.* **5**, 4355 (2014).
16. Kant, R. Van der & Vriend, G. Alpha-Bulges in G Protein-Coupled Receptors. *Int. J. Mol. Sci.* **15**, 7841–7864 (2014).

17. Kooistra, A. J. *et al.* GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Res.* **49**, D335–D343 (2021).
18. Isberg, V. *et al.* GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **42**, D422–D425 (2014).
19. Wacker, D., Stevens, R. C. & Roth, B. L. How Ligands Illuminate GPCR Molecular Pharmacology. *Cell* **170**, 414–427 (2017).
20. Latorraca, N. R., Venkatakrishnan, A. J. & Dror, R. O. GPCR Dynamics: Structures in Motion. *Chem. Rev.* **117**, 139–155 (2017).
21. Ye, L., Van Eps, N., Zimmer, M., Ernst, O. P. & Scott Prosser, R. Activation of the A2A adenosine G-protein-coupled receptor by conformational selection. *Nature* **533**, 265–268 (2016).
22. Wei, H. *et al.* Independent β-arrestin 2 and G protein-mediated pathways for angiotensin II activation of extracellular signal-regulated kinases 1 and 2. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 10782–10787 (2003).
23. Azzi, M. *et al.* β-arrestin-mediated activation of MAPK by inverse agonists reveals distinct active conformations for G protein-coupled receptors. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 11406–11411 (2003).
24. Smith, J. S., Lefkowitz, R. J. & Rajagopal, S. Biased signalling: from simple switches to allosteric microprocessors. *Nat. Rev. Drug Discov.* **17**, 243–260 (2018).
25. Xiang, J. *et al.* Successful Strategies to Determine High-Resolution Structures of GPCRs. *Trends Pharmacol. Sci.* **37**, 1055–1069 (2016).
26. Zanaty, M. *et al.* β-Arrestin-Biased Agonist Targeting the Brain AT1R (Angiotensin II Type 1 Receptor) Increases Aversion to Saline and Lowers Blood Pressure in Deoxycorticosterone Acetate-Salt Hypertension. *Hypertension* **77**, 420–431 (2021).
27. Violin, J. D. *et al.* Selectively Engaging β-Arrestins at the Angiotensin II Type 1 Receptor Reduces Blood Pressure and Increases Cardiac Performance. *J. Pharmacol. Exp. Ther.* **335**, 572–579 (2010).
28. Kobilka, B. & Deupi, X. Conformational complexity of G-protein-coupled receptors. *Trends Pharmacol. Sci.* **28**, 397–406 (2007).
29. Park, P. S.-H. Ensemble of G protein-coupled receptor active states. *Curr. Med. Chem.* **19**, 1146–54 (2012).
30. Rose, A. S. *et al.* Position of Transmembrane Helix 6 Determines Receptor G Protein Coupling Specificity. *J. Am. Chem. Soc.* **136**, 11244–11247 (2014).
31. Michel, M. & Charlton, S. Biased Agonism in Drug Discovery-Is It Too Soon to Choose a Path? *Mol. Pharmacol.* **93**, 259–265 (2018).
32. Ambrosio, M., Zürn, A. & Lohse, M. Sensing G protein-coupled receptor activation. *Neuropharmacology* **60**, 45–51 (2011).
33. Deupi, X., Li, X. & Schertler, G. Ligands stabilize specific GPCR conformations: but how? *Structure* **20**, 1289–1290 (2012).
34. Zarzycka, B., Zaidi, S. A., Roth, B. L. & Katritch, V. Harnessing Ion-Binding Sites for GPCR Pharmacology. *Pharmacol. Rev.* **71**, 571–595 (2019).
35. Selent, J., Sanz, F., Pastor, M. & de Fabritiis, G. Induced effects of sodium ions on dopaminergic G-protein coupled receptors. *PLoS Comput. Biol.* **6**, e1000884 (2010).
36. Gutiérrez-De-Terán, H. *et al.* The role of a sodium ion binding site in the

allosteric modulation of the A2A adenosine G protein-coupled receptor. *Structure* **21**, 2175–2185 (2013).
37. Gimpl, G. Interaction of G protein coupled receptors and cholesterol. *Chem. Phys. Lipids* **199**, 61–73 (2016).
38. Guixà-González, R. *et al.* Membrane omega-3 fatty acids modulate the oligomerisation kinetics of adenosine A2A and dopamine D2 receptors. *Sci. Rep.* **6**, 19839 (2016).
39. Dawaliby, R. *et al.* Allosteric regulation of G protein–coupled receptor activity by phospholipids. *Nat. Chem. Biol.* **12**, 35–39 (2016).
40. Bruzzese, A., Gil, C., Dalton, J. A. R. & Giraldo, J. Structural insights into positive and negative allosteric regulation of a G protein-coupled receptor through protein-lipid interactions. *Sci. Rep.* **8**, 4456 (2018).
41. Wootten, D., Christopoulos, A. & Sexton, P. M. Emerging paradigms in GPCR allostery: implications for drug discovery. *Nat. Publ. Gr.* **12**, 630–644 (2013).
42. Christopoulos, A. Advances in G protein-coupled receptor allostery: from function to structure. *Mol. Pharmacol.* **86**, 463–78 (2014).
43. Miao, Y. *et al.* Accelerated structure-based design of chemically diverse allosteric modulators of a muscarinic G protein-coupled receptor. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E5675–E5684 (2016).
44. Nguyen, T. *et al.* Rational design of cannabinoid type-1 receptor allosteric modulators: Org27569 and PSNCBAM-1 hybrids. *Bioorg. Med. Chem.* **41**, 116215 (2021).
45. Downes, G. & Gautam, N. The G protein subunit gene families. *Genomics* **62**, 544–552 (1999).
46. Robishaw, J. D. & Berlot, C. H. Translating G protein subunit diversity into functional specificity. *Curr. Opin. Cell Biol.* **16**, 206–209 (2004).
47. Neves, S. R., Ram, P. T. & Iyengar, R. G protein pathways. *Science.* **296**, 1636–1639 (2002).
48. Hilger, D., Masureel, M. & Kobilka, B. K. Structure and dynamics of GPCR signaling complexes. *Nat. Struct. Mol. Biol.* **25**, 4–12 (2018).
49. Kristiansen, K. Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. *Pharmacol. Ther.* **103**, 21–80 (2004).
50. Milligan, G. & Kostenis, E. Heterotrimeric G-proteins: a short history. *Br. J. Pharmacol.* **147**, S46–S55 (2006).
51. Khan, S. M. *et al.* The Expanding Roles of Gβγ Subunits in G Protein–Coupled Receptor Signaling and Drug Action. *Pharmacol. Rev.* **65**, 545–577 (2013).
52. Smrcka, A. V. G protein βγ subunits: central mediators of G protein-coupled receptor signaling. *Cell. Mol. Life Sci.* **65**, 2191–2214 (2008).
53. Gurevich, V. V. & Gurevich, E. V. The structural basis of arrestin-mediated regulation of G-protein-coupled receptors. *Pharmacol. Ther.* **110**, 465–502 (2006).
54. Grundmann, M. *et al.* Lack of beta-arrestin signaling in the absence of active G proteins. *Nat. Commun.* **9**, 341 (2018).
55. Craft, C. M., Whitmore$, D. H. & Wiechmannll, A. F. Cone Arrestin Identified by Targeting Expression of a Functional Family. *J Biol Chem*

269, 4613–4619 (1994).
56. Smith, J. & Rajagopal, S. The β-Arrestins: Multifunctional Regulators of G Protein-coupled Receptors. *J. Biol. Chem.* **291**, 8969–8977 (2016).
57. Palczewski, K. *et al.* Crystal structure of rhodopsin: A G protein-coupled receptor. *Science.* **289**, 739–745 (2000).
58. Cherezov, V. *et al.* High Resolution Crystal Structure of an Engineered Human β2-Adrenergic G protein-Coupled Receptor. *Science* **318**, 1258–1265 (2007).
59. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into β2-adrenergic receptor function. *Science.* **318**, 1266–1273 (2007).
60. Scheerer, P. *et al.* Crystal structure of opsin in its G-protein-interacting conformation. *Nature* **455**, 497–502 (2008).
61. Rasmussen, S. G. F. *et al.* Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549–55 (2011).
62. Zhang, D., Zhao, Q. & Wu, B. Structural Studies of G Protein-Coupled Receptors. *Mol. Cells* **38**, 836–842 (2015).
63. Congreve, M., de Graaf, C., Swain, N. A. & Tate, C. G. Impact of GPCR Structures on Drug Discovery. *Cell* **181**, 81–91 (2020).
64. Jazayeri, A., Andrews, S. P. & Marshall, F. H. Structurally enabled discovery of adenosine a2a receptor antagonists. *Chem. Rev.* **117**, 21–37 (2017).
65. Miao, Y., Caliman, A. D. & McCammon, J. A. Allosteric effects of sodium ion binding on activation of the m3 muscarinic g-protein-coupled receptor. *Biophys. J.* **108**, 1796–806 (2015).
66. Gregorio, G. G. *et al.* Single-molecule analysis of ligand efficacy in β2AR-G protein activation. *Nature* **547**, 68 (2017).
67. Bostock, M. J., Solt, A. S. & Nietlispach, D. The role of NMR spectroscopy in mapping the conformational landscape of GPCRs. *Curr. Opin. Struct. Biol.* **57**, 145–156 (2019).
68. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
69. WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. https://covid19.who.int/.
70. Fehr, A. R. & Perlman, S. Coronaviruses: An Overview of Their Replication and Pathogenesis. *Coronaviruses* **1282**, 1 (2015).
71. Alshammary, A. F. & Al-Sulaiman, A. M. The journey of SARS-CoV-2 in human hosts: a review of immune responses, immunosuppression, and their consequences. *Virulence* **12**, 1771–1794 (2021).
72. de Wilde, A. H., Snijder, E. J., Kikkert, M. & van Hemert, M. J. Host Factors in Coronavirus Replication. *Curr. Top. Microbiol. Immunol.* **419**, 1–42 (2018).
73. Snijder, E. J., Decroly, E. & Ziebuhr, J. The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. *Adv. Virus Res.* **96**, 59–126 (2016).
74. Denison, M. R., Graham, R. L., Donaldson, E. F., Eckerle, L. D. & Baric, R. S. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol.* **8**, 215–218 (2011).
75. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell*

181, 914-921.e10 (2020).
76. Cai, Y. *et al.* Distinct conformational states of SARS-CoV-2 spike protein. *Science.* **369**, 1586–1592 (2020).
77. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562–569 (2020).
78. Piccoli, L. *et al.* Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell* **183**, 1024–1042 (2020).
79. Liu, L. *et al.* Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**, 450–456 (2020).
80. Dai, L. & Gao, G. F. Viral targets for vaccines against COVID-19. *Nat. Rev. Immunol.* **21**, 73–82 (2020).
81. Siu, Y. L. *et al.* The M, E, and N Structural Proteins of the Severe Acute Respiratory Syndrome Coronavirus Are Required for Efficient Assembly, Trafficking, and Release of Virus-Like Particles. *J. Virol.* **82**, 11318–11330 (2008).
82. Neuman, B. W. *et al.* Supramolecular Architecture of Severe Acute Respiratory Syndrome Coronavirus Revealed by Electron Cryomicroscopy. *J. Virol.* **80**, 7918–7928 (2006).
83. Schoeman, D. & Fielding, B. C. Coronavirus envelope protein: current knowledge. *Virol. J.* **16**, 69 (2019).
84. Payne, S. Family Coronaviridae. in *Viruses* 149 (Elsevier, 2017). doi:10.1016/B978-0-12-803109-4.00017-9.
85. Yao, H. *et al.* Molecular Architecture of the SARS-CoV-2 Virus. *Cell* **183**, 730–738 (2020).
86. Chang, C., Hou, M.-H., Chang, C.-F., Hsiao, C.-D. & Huang, T. The SARS coronavirus nucleocapsid protein – Forms and functions. *Antiviral Res.* **103**, 39–50 (2014).
87. Liu, D. X., Fung, T. S., Chong, K. K.-L., Shukla, A. & Hilgenfeld, R. Accessory proteins of SARS-CoV and other coronaviruses. *Antiviral Res.* **109**, 97–109 (2014).
88. Frost, S. D. W., Magalis, B. R. & Pond, S. L. K. Neutral Theory and Rapidly Evolving Viral Pathogens. *Mol. Biol. Evol.* **35**, 1348–1354 (2018).
89. Martin, D. P. *et al.* The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. *medRxiv* (2021) doi:10.1101/2021.02.23.21252268.
90. Washington, N. L. *et al.* Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *Cell* **184**, 2587-2594.e7 (2021).
91. Madhi, S. A. *et al.* Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N. Engl. J. Med.* **384**, 1885–1898 (2021).
92. Cele, S. *et al.* Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. *Nature* **593**, 142–146 (2021).
93. Funk, T. *et al.* Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. *Eurosurveillance* **26**, 2100348 (2021).
94. Dejnirattisai, W. *et al.* Antibody evasion by the P.1 strain of SARS-CoV-

2. *Cell* **184**, 2939 (2021).
95. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021).
96. Bernal, J. L. *et al.* Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *N. Engl. J. Med.* **385**, 585–594 (2021).
97. Sheikh, A., McMenamin, J., Taylor, B. & Robertson, C. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. *Lancet* **397**, 2461–2462 (2021).
98. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332–2347 (2021).
99. Barnes, C. O. *et al.* SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588**, 682–687 (2020).
100. Chi, X. *et al.* A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* **369**, 650 (2020).
101. McCarthy, K. R. *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* **371**, 1139 (2021).
102. Thomson, E. C. *et al.* Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* **184**, 1171 (2021).
103. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463-476.e6 (2021).
104. Starr, T. N. *et al.* Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **371**, 850–854 (2021).
105. Greaney, A. J. *et al.* Antibodies elicited by mRNA-1273 vaccination bind more broadly to the receptor binding domain than do those from SARS-CoV-2 infection. *Sci. Transl. Med.* **13**, 9915 (2021).
106. SARS-CoV-2 variants of concern as of 23 September 2021. https://www.ecdc.europa.eu/en/covid-19/variants-concern.
107. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 1 (2021).
108. Estrada, E. COVID-19 and SARS-CoV-2. Modeling the present, looking at the future. *Phys. Rep.* **869**, 1–51 (2020).
109. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260 (2020).
110. Yuan, M. *et al.* A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science.* **368**, 630–633 (2020).
111. Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science.* **367**, 1444–1448 (2020).
112. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).
113. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
114. Hillen, H. S. *et al.* Structure of replicating SARS-CoV-2 polymerase. *Nature* **584**, 154–156 (2020).
115. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science.* **368**, 409–412 (2020).

116. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
117. Cox, R. M. & Plemper, R. K. The impact of high-resolution structural data on stemming the COVID-19 pandemic. *Curr. Opin. Virol.* **49**, 127–138 (2021).
118. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
119. Casalino, L. *et al.* Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* **6**, 1722–1734 (2020).
120. Benton, D. J. *et al.* Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* **588**, 327–330 (2020).
121. Rut, W. *et al.* SARS-CoV-2 Mpro inhibitors and activity-based probes for patient-sample imaging. *Nat. Chem. Biol.* **17**, 222–228 (2020).
122. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020).
123. Hoffman, R. L. *et al.* Discovery of Ketone-Based Covalent Inhibitors of Coronavirus 3CL Proteases for the Potential Therapeutic Treatment of COVID-19. *J. Med. Chem.* **63**, 12725–12747 (2020).
124. Beigel, J. H. *et al.* Remdesivir for the Treatment of Covid-19 — Final Report. *N. Engl. J. Med.* **383**, 1813–1826 (2020).
125. Sofia, M. J. *et al.* Discovery of a β-d-2′-Deoxy-2′-α-fluoro-2′-β-C-methyluridine Nucleotide Prodrug (PSI-7977) for the Treatment of Hepatitis C Virus. *J. Med. Chem.* **53**, 7202–7218 (2010).
126. Noshi, T. *et al.* In vitro characterization of baloxavir acid, a first-in-class cap-dependent endonuclease inhibitor of the influenza virus polymerase PA subunit. *Antiviral Res.* **160**, 109–117 (2018).
127. Holec, A. D., Mandal, S., Prathipati, P. K. & Destache, C. J. Nucleotide Reverse Transcriptase Inhibitors: A Thorough Review, Present Status and Future Perspective as HIV Therapeutics. *Curr. HIV Res.* **15**, 411–421 (2017).
128. Kokic, G. *et al.* Mechanism of SARS-CoV-2 polymerase stalling by remdesivir. *Nat. Commun.* **12**, 279 (2021).
129. Wang, Q. *et al.* Structural Basis for RNA Replication by the SARS-CoV-2 Polymerase. *Cell* **182**, 417–428 (2020).
130. Bravo, J. P. K., Dangerfield, T. L., Taylor, D. W. & Johnson, K. A. Remdesivir is a delayed translocation inhibitor of SARS-CoV-2 replication. *Mol. Cell* **81**, 1548–1552 (2021).
131. Yin, W. *et al.* Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science* **368**, 1499–1504 (2020).
132. Peng, Q. *et al.* Structural Basis of SARS-CoV-2 Polymerase Inhibition by Favipiravir. *Innov.* **2**, 100080 (2021).
133. Chen, J. *et al.* Structural Basis for Helicase-Polymerase Coupling in the SARS-CoV-2 Replication-Transcription Complex. *Cell* **182**, 1560-1573.e13 (2020).
134. Yan, L. *et al.* Cryo-EM Structure of an Extended SARS-CoV-2 Replication and Transcription Complex Reveals an Intermediate State in Cap Synthesis. *Cell* **184**, 184-193.e10 (2021).

135. LM, W. *et al.* Angiotensin Analogs with Divergent Bias Stabilize Distinct Receptor Conformations. *Cell* **176**, 468-478.e11 (2019).
136. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **99**, 1129–1143 (2018).
137. McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267**, 585–590 (1977).
138. Dahl, S. G., Edvardsen, O. & Sylte, I. Molecular dynamics of dopamine at the D2 receptor. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 8111–8115 (1991).
139. Stone, J. E. *et al.* Accelerating molecular modeling applications with graphics processors. *J. Comput. Chem.* **28**, 2618–2640 (2007).
140. Anderson, J. A., Lorenz, C. D. & Travesset, A. General purpose molecular dynamics simulations fully implemented on graphics processing units. *J. Comput. Phys.* **227**, 5342–5359 (2008).
141. Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**–**2**, 19–25 (2015).
142. Lagardère, L. *et al.* Tinker-HP: A massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chem. Sci.* **9**, 956–972 (2018).
143. Jung, J. *et al.* Scaling molecular dynamics beyond 100,000 processor cores for large-scale biophysical simulations. *J. Comput. Chem.* **40**, 1919–1930 (2019).
144. Marrink, S. J. *et al.* Computational Modeling of Realistic Cell Membranes. *Chem. Rev.* **119**, 6184–6226 (2019).
145. Bottaro, S. & Lindorff-Larsen, K. Biophysical experiments and biomolecular simulations: A perfect match? *Science.* **361**, 355–360 (2018).
146. Jo, S., Lim, J. B., Klauda, J. B. & Im, W. CHARMM-GUI Membrane Builder for Mixed Bilayers and Its Application to Yeast Membranes. *Biophys. J.* **97**, 50–58 (2009).
147. Mayol, E. *et al.* HomolWat: a web server tool to incorporate 'homologous' water molecules into GPCR structures. *Nucleic Acids Res.* **1**, 13–14 (2020).
148. Doerr, S., Harvey, M. J., Noé, F. & De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **12**, 1845–52 (2016).
149. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32**, 2319–2327 (2011).
150. Gowers, R. J. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. in *Proceedings of the 15th Python in Science Conference* (eds. Benthall, S. & Rostrup, S.) 98–105 (2016).
151. McGibbon, R. T. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. **109**, 1528–1532 (2015).
152. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
153. MacKerell, A. D. *et al.* All-Atom Empirical Potential for Molecular

Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
154. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. SOC* **117**, 5179–5197 (1995).
155. Klauda, J. B. *et al.* Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *J. Phys. Chem. B* **114**, 7830–7843 (2010).
156. Lim, J. B., Rogaski, B. & Klauda, J. B. Update of the cholesterol force field parameters in CHARMM. *J. Phys. Chem. B* **116**, 203–210 (2012).
157. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2016).
158. Christen, M. *et al.* The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **26**, 1719–1751 (2005).
159. 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.* **118**, 11225–11236 (1996).
160. Lindorff-Larsen, K. *et al.* Systematic validation of protein force fields against experimental data. *PLoS One* **7**, e32131 (2012).
161. Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L. & Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19011–19016 (2009).
162. Bond, P. J., Holyoake, J., Ivetac, A., Khalid, S. & Sansom, M. S. P. Coarse-grained molecular dynamics simulations of membrane proteins and peptides. *J. Struct. Biol.* **157**, 593–605 (2007).
163. Marrink, S. J. & Tieleman, D. P. Perspective on the martini model. *Chem. Soc. Rev.* **42**, 6801–6822 (2013).
164. Hospital, A., Goñi, J. R., Orozco, M. & Gelpí, J. L. Molecular dynamics simulations: advances and applications. *Adv. Appl. Bioinform. Chem.* **8**, 37–47 (2015).
165. Bernardi, R. C., Melo, M. C. R. & Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta* **1850**, 872–877 (2015).
166. Vendruscolo, M. & Dobson, C. M. Protein Dynamics: Moore's Law in Molecular Biology. *Curr. Biol.* **21**, R68–R70 (2011).
167. Martínez-Rosell, G., Giorgino, T., Harvey, M. J. & de Fabritiis, G. Drug Discovery and Molecular Dynamics: Methods, Applications and Perspective Beyond the Second Timescale. *Curr. Top. Med. Chem.* **17**, 2617–2625 (2017).
168. Torrens-Fontanals, M., Stepniewski, T. M., Rodríguez-Espigares, I. & Selent, J. Application of Biomolecular Simulations to G Protein-Coupled Receptors (GPCRs). in *Biomolecular Simulations in Structure-based Drug Discovery* (eds. Gervasio, F. L. & Spiwok, V.) 205–223 (Wiley, 2018). doi:10.1002/9783527806836.ch8.
169. Torrens-Fontanals, M. *et al.* How do molecular dynamics data complement static structural data of GPCRs. *Int. J. Mol. Sci.* **21**, 5933 (2020).

170. Torrens-Fontanals, M., Stepniewski, T. M., Gloriam, D. E. & Selent, J. Structural dynamics bridge the gap between the genetic and functional levels of GPCRs. *Curr. Opin. Struct. Biol.* **69**, 150–159 (2021).
171. Ferruz, N. *et al.* Dopamine D3 receptor antagonist reveals a cryptic pocket in aminergic GPCRs. *Sci. Rep.* **8**, 897 (2018).
172. Dror, R. O. *et al.* Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 13118–13123 (2011).
173. Wittmann, H.-J. J. & Strasser, A. Binding pathway of histamine to the hH4R, observed by unconstrained molecular dynamics. *Bioorganic Med. Chem. Lett.* **25**, 1259–1268 (2015).
174. Sabbadin, D., Ciancetta, A., Deganutti, G., Cuzzolin, A. & Moro, S. Exploring the recognition pathway at the human A2A adenosine receptor of the endogenous agonist adenosine using supervised molecular dynamics simulations. *Medchemcomm* **6**, 1081–1085 (2015).
175. Thomas, T., Fang, Y., Yuriev, E. & Chalmers, D. K. Ligand Binding Pathways of Clozapine and Haloperidol in the Dopamine D $_2$ and D $_3$ Receptors. *J. Chem. Inf. Model.* **56**, 308–321 (2016).
176. Hertig, S., Latorraca, N. R. & Dror, R. O. Revealing Atomic-Level Mechanisms of Protein Allostery with Molecular Dynamics Simulations. *PLoS Comput. Biol.* **12**, e1004746 (2016).
177. Bock, A., Schrage, R. & Mohr, K. Allosteric modulators targeting CNS muscarinic receptors. *Neuropharmacology* **136**, 427–437 (2017).
178. Taylor, B. C., Lee, C. T. & Amaro, R. E. Structural basis for ligand modulation of the CCR2 conformational landscape. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8131–8136 (2019).
179. Chan, H. C. S. *et al.* Exploring a new ligand binding site of G protein-coupled receptors. *Chem. Sci.* **9**, 6480–6489 (2018).
180. Martí-Solano, M., Schmidt, D., Kolb, P. & Selent, J. Drugging specific conformational states of GPCRs: challenges and opportunities for computational chemistry. *Drug Discov. Today* **21**, 625–631 (2016).
181. Rodríguez-Espigares, I., Kaczor, A. A., Stepniewski, T. M. & Selent, J. Challenges and opportunities in drug discovery of biased ligands. in *Computational Methods for GPCR Drug Discovery* (ed. Heifetz, A.) vol. 1705 321–334 (Humana Press, 2018).
182. Martí-Solano, M. *et al.* Detection of New Biased Agonists for the Serotonin 5-HT2A Receptor: Modeling and Experimental Validation. *Mol. Pharmacol.* **87**, 740–746 (2015).
183. McCorvy, J. D. *et al.* Structure-inspired design of β-arrestin-biased ligands for aminergic GPCRs. *Nat. Chem. Biol.* **14**, 126–134 (2018).
184. Bermudez, M. *et al.* Ligand-Specific Restriction of Extracellular Conformational Dynamics Constrains Signaling of the M2 Muscarinic Receptor. *ACS Chem. Biol.* **12**, 1743–1748 (2017).
185. Suomivuori, C. M. *et al.* Molecular mechanism of biased signaling in a prototypical G protein-coupled receptor. *Science.* **367**, 881–887 (2020).
186. Kapoor, A., Martinez-Rosell, G., Provasi, D., de Fabritiis, G. & Filizola, M. Dynamic and Kinetic Elements of μ-Opioid Receptor Functional Selectivity. *Sci. Rep.* **7**, 11255 (2017).
187. Nivedha, A. K. *et al.* Identifying functional hotspot residues for biased

ligand design in G-protein-coupled receptors. *Mol. Pharmacol.* **93**, 288–296 (2018).
188. Sengupta, D., Sonar, K. & Joshi, M. Characterizing clinically relevant natural variants of GPCRs using computational approaches. *Methods Cell Biol.* **142**, 187–204 (2017).
189. Shahane, G., Parsania, C., Sengupta, D. & Joshi, M. Molecular insights into the dynamics of pharmacogenetically important N-terminal variants of the human β2-adrenergic receptor. *PLoS Comput. Biol.* **10**, e1004006 (2014).
190. Bhosale, S., Nikte, S. V., Sengupta, D. & Joshi, M. Differential Dynamics Underlying the Gln27Glu Population Variant of the β2-Adrenergic Receptor. *J. Membr. Biol.* **252**, 499–507 (2019).
191. Tandale, A., Joshi, M. & Sengupta, D. Structural insights and functional implications of inter-individual variability in β 2 -adrenergic receptor. *Sci. Rep.* **6**, 24379 (2016).
192. Oddi, S. *et al.* Palmitoylation of cysteine 415 of CB 1 receptor affects ligand-stimulated internalization and selective interaction with membrane cholesterol and caveolin 1. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* **1862**, 523–532 (2017).
193. Oddi, S. *et al.* Role of palmitoylation of cysteine 415 in functional coupling CB1 receptor to Gαi2 protein. *Biotechnol. Appl. Biochem.* **65**, 16–20 (2018).
194. Guixà-González, R. *et al.* Membrane cholesterol access into a G-protein-coupled receptor. *Nat. Commun.* **8**, 14505 (2017).
195. Ramírez-Anguita, J. M. *et al.* Membrane cholesterol effect on the 5-HT2A receptor: Insights into the lipid-induced modulation of an antipsychotic drug target. *Biotechnol. Appl. Biochem.* **65**, 29–37 (2018).
196. Ng, H. W., Laughton, C. A. & Doughty, S. W. Molecular dynamics simulations of the adenosine A2a receptor in POPC and POPE lipid bilayers: Effects of membrane on protein behavior. *J. Chem. Inf. Model.* **54**, 573–581 (2014).
197. Bruzzese, A., Dalton, J. A. R. & Giraldo, J. Insights into adenosine A2A receptor activation through cooperative modulation of agonist and allosteric lipid interactions. *PLoS Comput. Biol.* **16**, e1007818 (2020).
198. Corey, R. A., Vickery, O. N., Sansom, M. S. P. & Stansfeld, P. J. Insights into Membrane Protein-Lipid Interactions from Free Energy Calculations. *J. Chem. Theory Comput.* **15**, 5727–5736 (2019).
199. Dror, R. O. *et al.* Activation mechanism of the β 2-adrenergic receptor. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18684–18689 (2011).
200. Kohlhoff, K. J. *et al.* Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.* **6**, 15–21 (2014).
201. Li, J., Jonsson, A. L., Beuming, T., Shelley, J. C. & Voth, G. A. Ligand-dependent activation and deactivation of the human adenosine A 2A receptor. *J. Am. Chem. Soc.* **135**, 8749–8759 (2013).
202. Meral, D., Provasi, D. & Filizola, M. An efficient strategy to estimate thermodynamics and kinetics of G protein-coupled receptor activation using metadynamics and maximum caliber. *J. Chem. Phys.* **149**, 224101 (2018).

203. Fleetwood, O., Matricon, P., Carlsson, J. & Delemotte, L. Energy Landscapes Reveal Agonist Control of G Protein-Coupled Receptor Activation via Microswitches. *Biochemistry* **59**, 880–891.
204. Karoussiotis, C. *et al.* A highly conserved δ-opioid receptor region determines RGS4 interaction. *FEBS J.* **287**, 736–748 (2020).
205. Eichel, K. *et al.* Catalytic activation of β-Arrestin by GPCRs. *Nature* **557**, 381–386 (2018).
206. Venkatakrishnan, A. J. *et al.* Diverse GPCRs exhibit conserved water networks for stabilization and activation. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 3288–3293 (2019).
207. Yuan, S., Vogel, H. & Filipek, S. The Role of Water and Sodium Ions in the Activation of the μ-Opioid Receptor. *Angew. Chemie Int. Ed.* **52**, 10112–10115 (2013).
208. Yuan, S., Hu, Z., Filipek, S. & Vogel, H. W2466.48 opens a gate for a continuous intrinsic water pathway during activation of the adenosine A2A receptor. *Angew. Chemie - Int. Ed.* **54**, 556–559 (2015).
209. Yuan, S., Filipek, S., Palczewski, K. & Vogel, H. Activation of G-protein-coupled receptors correlates with the formation of a continuous internal water pathway. *Nat. Commun.* **5**, 4733 (2014).
210. Yuan, S. *et al.* The mechanism of ligand-induced activation or inhibition of μ- And κ-opioid receptors. *Angew. Chemie - Int. Ed.* **54**, 7560–7563 (2015).
211. Schmidtke, P., Javier Luque, F., Murray, J. B. & Barril, X. Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. *J. Am. Chem. Soc.* **133**, 18903–18910 (2011).
212. Magarkar, A., Schnapp, G., Apel, A. K., Seeliger, D. & Tautermann, C. S. Enhancing Drug Residence Time by Shielding of Intra-Protein Hydrogen Bonds: A Case Study on CCR2 Antagonists. *ACS Med. Chem. Lett.* **10**, 324–328 (2019).
213. Selvam, B., Shamsi, Z. & Shukla, D. Universality of the Sodium Ion Binding Mechanism in Class A G-Protein-Coupled Receptors. *Angew. Chemie Int. Ed.* **57**, 3048–3053 (2018).
214. Hu, X. *et al.* Kinetic and thermodynamic insights into sodium ion translocation through the μ-opioid receptor from molecular dynamics and machine learning analysis. *PLOS Comput. Biol.* **15**, e1006689 (2019).
215. Johnston, J. M., Wang, H., Provasi, D. & Filizola, M. Assessing the Relative Stability of Dimer Interfaces in G Protein-Coupled Receptors. *PLoS Comput. Biol.* **8**, e1002649 (2012).
216. Periole, X., Knepp, A. M., Sakmar, T. P., Marrink, S. J. & Huber, T. Structural determinants of the supramolecular organization of G protein-coupled receptors in bilayers. *J. Am. Chem. Soc.* **134**, 10959–10965 (2012).
217. Zimmerman, M. I. *et al.* SARS-CoV-2 Simulations Go Exascale to Capture Spike Opening and Reveal Cryptic Pockets Across the Proteome. *Nat. Chem.* **13**, 651–659 (2021).
218. Yu, A. *et al.* A multiscale coarse-grained model of the SARS-CoV-2 virion. *Biophys. J.* **120**, 1097–1104 (2021).
219. Casalino, L. *et al.* AI-Driven Multiscale Simulations Illuminate Mechanisms of SARS-CoV-2 Spike Dynamics. *Int. J. High Perform.*

Comput. Appl. **35**, 432–451 (2021).
220. Grant, O. C., Montgomery, D., Ito, K. & Woods, R. J. Analysis of the SARS-CoV-2 spike protein glycan shield reveals implications for immune recognition. *Sci. Rep.* **10**, 14991 (2020).
221. Ghorbani, M., Brooks, B. R. & Klauda, J. B. Exploring dynamics and network analysis of spike glycoprotein of SARS-COV-2. *Biophys. J.* **120**, 2902–2913 (2021).
222. Rath, S. L. & Kumar, K. Investigation of the Effect of Temperature on the Structure of SARS-CoV-2 Spike Protein by Molecular Dynamics Simulations. *Front. Mol. Biosci.* **7**, 583523 (2020).
223. Gur, M. *et al.* Conformational transition of SARS-CoV-2 spike glycoprotein between its closed and open states. *J. Chem. Phys.* **153**, 75101 (2020).
224. Piplani, S., Singh, P. K., Winkler, D. A. & Petrovsky, N. In silico comparison of SARS-CoV-2 spike protein-ACE2 binding affinities across species and implications for virus origin. *Sci. Rep.* **11**, 13063 (2021).
225. Nelson, G. *et al.* Millisecond-scale molecular dynamics simulation of spike RBD structure reveals evolutionary adaption of SARS-CoV-2 to stably bind ACE2. *bioRxiv* (2020) doi:10.1101/2020.12.11.422055.
226. Spinello, A., Saltalamacchia, A. & Magistrato, A. Is the Rigidity of SARS-CoV-2 Spike Receptor-Binding Motif the Hallmark for Its Enhanced Infectivity? Insights from All-Atom Simulations. *J. Phys. Chem. Lett.* **11**, 4785–4790 (2020).
227. He, J., Tao, H., Yan, Y., Huang, S.-Y. & Xiao, Y. Molecular Mechanism of Evolution and Human Infection with SARS-CoV-2. *Viruses* **12**, 428 (2020).
228. Peng, C. *et al.* Computational study of the strong binding mechanism of SARS-CoV-2 spike and ACE2. *ChemRxiv* (2021) doi:10.26434/CHEMRXIV.11877492.V2.
229. Wang, Y., Liu, M. & Gao, J. Enhanced receptor binding of SARS-CoV-2 through networks of hydrogen-bonding and hydrophobic interactions. *Proc. Natl. Acad. Sci.* **117**, 13967–13974 (2020).
230. Cao, Y. *et al.* Dynamic Interactions of Fully Glycosylated SARS-CoV-2 Spike Protein with Various Antibodies. *J. Chem. Theory Comput.* **17**, 6559–6569 (2021).
231. Malaspina, D. C. & Faraudo, J. Computer simulations of the interaction between SARS-CoV-2 spike glycoprotein and different surfaces. *Biointerphases* **15**, 51008 (2020).
232. Deganutti, G., Prischi, F. & Reynolds, C. A. Supervised molecular dynamics for exploring the druggability of the SARS-CoV-2 spike protein. *J. Comput. Aided. Mol. Des.* **35**, 195–207 (2020).
233. Sztain, T., Amaro, R. & McCammon, J. A. Elucidation of Cryptic and Allosteric Pockets within the SARS-CoV-2 Main Protease. *J. Chem. Inf. Model.* **61**, 3495–3501 (2021).
234. Sundar, S. *et al.* Screening of FDA-approved compound library identifies potential small-molecule inhibitors of SARS-CoV-2 non-structural proteins NSP1, NSP4, NSP6 and NSP13: molecular modeling and molecular dynamics studies. *J. Proteins Proteomics* **12**, 161–175 (2021).

235. Tazikeh-Lemeski, E. *et al.* Targeting SARS-COV-2 non-structural protein 16: a virtual drug repurposing study. *J. Biomol. Struct. Dyn.* **39**, 4633–4646 (2021).
236. Menezes, G. de L. & Silva, R. A. da. Identification of potential drugs against SARS-CoV-2 non-structural protein 1 (nsp1). *J. Biomol. Struct. Dyn.* **39**, 5657–5667 (2021).
237. Sharma, J. *et al.* An in-silico evaluation of different bioactive molecules of tea for their inhibition potency against non structural protein-15 of SARS-CoV-2. *Food Chem.* **346**, 128933 (2021).
238. Choi, K.-E. *et al.* Molecular Dynamics Studies on the Structural Characteristics for the Stability Prediction of SARS-CoV-2. *Int. J. Mol. Sci. Artic. Int. J. Mol. Sci* **22**, 8714 (2021).
239. Verkhivker, G. M. & Paola, L. Di. Dynamic Network Modeling of Allosteric Interactions and Communication Pathways in the SARS-CoV-2 Spike Trimer Mutants: Differential Modulation of Conformational Landscapes and Signal Transmission via Cascades of Regulatory Switches. *J. Phys. Chem. B* **125**, 850–873 (2021).
240. Mohammad, A. *et al.* Higher binding affinity of furin for SARS-CoV-2 spike (S) protein D614G mutant could be associated with higher SARS-CoV-2 infectivity. *Int. J. Infect. Dis.* **103**, 611–616 (2021).
241. Qiao, B. & Cruz, M. O. de la. Enhanced Binding of SARS-CoV-2 Spike Protein to Receptor by Distal Polybasic Cleavage Sites. *ACS Nano* **14**, 10616–10623 (2020).
242. Amamuddy, O. S., Verkhivker, G. M. & Bishop, Ö. T. Impact of Early Pandemic Stage Mutations on Molecular Dynamics of SARS-CoV-2 Mpro. *J. Chem. Inf. Model.* **60**, 5080–5102 (2020).
243. Gioia, D., Bertazzo, M., Recanatini, M., Masetti, M. & Cavalli, A. Dynamic Docking: A Paradigm Shift in Computational Drug Discovery. *Mol. A J. Synth. Chem. Nat. Prod. Chem.* **22**, 2029 (2017).
244. Basciu, A., Malloci, G., Pietrucci, F., Bonvin, A. M. J. J. & Vargiu, A. V. Holo-like and Druggable Protein Conformations from Enhanced Sampling of Binding Pocket Volume and Shape. *J. Chem. Inf. Model.* **59**, 1515–1528 (2019).
245. Yuan, J.-H., Han, S. B., Richter, S., Wade, R. C. & Kokh, D. B. Druggability Assessment in TRAPP Using Machine Learning Approaches. *J. Chem. Inf. Model.* **60**, 1685–1699 (2020).
246. Cagiada, M. *et al.* Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance. *Mol. Biol. Evol.* **38**, 3235–3246 (2021).
247. Abraham, M. J. *et al.* Sharing Data from Molecular Simulations. *J. Chem. Inf. Model.* **59**, 4093–4099 (2019).
248. Amaro, R. E. & Mulholland, A. J. A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19. *J. Chem. Inf. Model.* **60**, 2653–2656 (2020).
249. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
250. Woelfle, M., Olliaro, P. & Todd, M. H. Open science is a research accelerator. *Nat. Chem.* **3**, 745–748 (2011).
251. Data sharing and the future of science. *Nat. Commun.* **9**, 2817 (2018).

252. Elofsson, A. *et al.* Ten simple rules on how to create open access and reproducible molecular simulations of biological systems. *PLOS Comput. Biol.* **15**, e1006649 (2019).
253. Hildebrand, P. W., Rose, A. S. & Tiemann, J. K. S. Bringing Molecular Dynamics Simulation Data into View. *Trends Biochem. Sci.* **44**, 902–913 (2019).
254. Abriata, L. A., Rodrigues, J. P. G. L. M., Salathé, M. & Patiny, L. Augmenting research, education and outreach with client-side web programming. *Trends Biotechnol.* **36**, 473–476 (2018).
255. Tiemann, J. K. S., Guixà-González, R., Hildebrand, P. W. & Rose, A. S. MDsrv: Viewing and sharing molecular dynamics simulations on the web. *Nat. Methods* **14**, 1123–1124 (2017).
256. Carrillo-Tripp, M. *et al.* HTMoL: full-stack solution for remote access, visualization, and analysis of molecular dynamics trajectory data. *J. Comput. Aided. Mol. Des.* **32**, 869–876 (2018).
257. Sehnal, D. *et al.* Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.* **49**, W431–W437 (2021).
258. Bekker, G. J., Nakamura, H. & Kinjo, A. R. Molmil: A molecular viewer for the PDB and beyond. *J. Cheminform.* **8**, 42 (2016).
259. Hospital, A. *et al.* BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.* **44**, D272–D278 (2016).
260. BioExcel - COVID-19. https://bioexcel-cv19.bsc.es/.
261. COVID-19 Molecular Structure and Therapeutics Hub. https://covid.bioexcel.eu/.
262. Mixcoha, E., Rosende, R., Garcia-Fandino, R. & Piñeiro, Á. Cyclo-lib: a database of computational molecular dynamics simulations of cyclodextrins. *Bioinformatics* **32**, 3371–3373 (2016).
263. Rodríguez-Espigares, I. *et al.* GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods* **17**, 777–787 (2020).
264. Newport, T. D., Sansom, M. S. P. & Stansfeld, P. J. The MemProtMD database: A resource for membrane-embedded protein structures and their lipid interactions. *Nucleic Acids Res.* **47**, D390–D397 (2019).
265. MoDEL - Central Nervous System. https://mmb.irbbarcelona.org/MoDEL-CNS/.
266. Meyer, T. *et al.* MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories. *Structure* **18**, 1399–1409 (2010).
267. NMRlipids - High precision lipid models. http://www.nmrlipids.fi/.
268. Torrens-Fontanals, M. *et al.* SCoV2-MD: a database for the dynamics of the SARS-CoV-2 proteome and variant impact predictions. *Nucleic Acids Res.* **1**, 13–14 (2021).
269. Sun, R., Li, Z. & Bishop, T. C. TMB-iBIOMES: An iBIOMES-Lite Database of Nucleosome Trajectories and Meta-Analysis. *ChemRxiv* (2019) doi:10.26434/CHEMRXIV.7793939.V1.
270. Manglik, A. & Kobilka, B. The role of protein dynamics in GPCR function: insights from the β2AR and rhodopsin. *Curr. Opin. Cell Biol.* **27**, 136–143 (2014).

271. Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **100**, 20603 (2008).
272. Stall, S. *et al.* Make scientific data FAIR. *Nature* **570**, 27–29 (2019).
273. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science.* **185**, 862–864 (1974).
274. Dunham, A., Jang, G. M., Muralidharan, M., Swaney, D. & Beltrao, P. A missense variant effect prediction and annotation resource for SARS-CoV-2. *bioRxiv* (2021) doi:10.1101/2021.02.24.432721.
275. Antila, H. S., Ferreira, T. M., Ollila, O. H. S. & Miettinen, M. S. Using Open Data to Rapidly Benchmark Biomolecular Simulations: Phospholipid Conformational Dynamics. *J. Chem. Inf. Model.* **61**, 938–949 (2021).
276. Kozlovskii, I. & Popov, P. Spatiotemporal identification of druggable binding sites using deep learning. *Commun. Biol.* **3**, 618 (2020).
277. Bouysset, C. & Fiorucci, S. ProLIF: a library to encode molecular interactions as fingerprints. *J. Cheminform.* **13**, 72 (2021).
278. Schneider, J. *et al.* Ligand Pose Predictions for Human G Protein-Coupled Receptors: Insights from the Amber-Based Hybrid Molecular Mechanics/Coarse-Grained Approach. *J. Chem. Inf. Model.* **60**, 5103–5116 (2020).
279. Denzinger, K., Nguyen, T. N., Noonan, T., Wolber, G. & Bermudez, M. Biased Ligands Differentially Shape the Conformation of the Extracellular Loop Region in 5-HT2B Receptors. *Int. J. Mol. Sci.* **21**, 9728 (2020).
280. Liu, W. *et al.* Structural basis for allosteric regulation of GPCRs by sodium ions. *Science* **337**, 232–236 (2012).
281. Boekel, J. *et al.* Multi-omic data analysis using Galaxy. *Nat. Biotechnol.* **33**, 137–139 (2015).
282. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).
283. Consortium, E. Exscalate4CoV. https://www.exscalate4cov.eu/.
284. Hospital, A., Battistini, F., Soliva, R., Gelpí, J. L. & Orozco, M. Surviving the deluge of biosimulation data. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **10**, e1449 (2020).
285. Abriata, L. A., Lepore, R. & Dal Peraro, M. About the need to make computational models of biological macromolecules available and discoverable. *Bioinformatics* **36**, 2952–2954 (2020).